STAT 302 INFORMATION BOOK

R. J. Carroll *

* Copyright, R. J. Carroll, 1994.
Contents

1 FIRST CLASS OVERVIEW 1

2 COURSE SCHEDULE 3

3 MECHANICS AND RULES 7

4 FORMULAE BY EXAM # 13
  4.1 FORMULAE FOR EXAM #1 13
  4.2 FORMULAE FOR EXAM #2 14
  4.3 FORMULAE FOR EXAM #3 14

5 CLASS PROJECT RULES 17
  5.1 GRADING OF PROJECTS 17
  5.2 GETTING STARTED ON THE TERM PROJECT 18
    5.2.1 CREATING A TEAM 18
    5.2.2 THE LAZY PLAYER 18
    5.2.3 THE QUESTION 18
    5.2.4 TYPING OF PROJECTS 18
  5.3 YOUR RESPONSIBILITIES 19
  5.4 POTENTIAL TOPICS ASSIGNMENT (MAXIMUM OF 2 PAGES). 140 points. 20
    5.4.1 Basic Idea of the Assignment 20
    5.4.2 The Specific Assignment 20
  5.5 TEAM WORKING AGREEMENT ASSIGNMENT: 22
    5.5.1 Purpose of the Assignment 22
    5.5.2 Work to be Done 22
  5.6 WRITTEN PROPOSAL (460 points) 23
    5.6.1 The Purpose of this Assignment 23
    5.6.2 Some Notes About Grades for Projects 24
    5.6.3 Some Notes about Lurking Variables 24
    5.6.4 (40 points) The question to be studied 24
    5.6.5 (150 points) The statistical questions 24
5.6.6  (100 points) Data collection plans. [DO NOT START COLLECTING DATA UNTIL YOUR PROJECT HAS BEEN APPROVED.] 24
5.6.7  (70 points) Data analysis plans 25
5.6.8  (100 points) Thoroughness and Creativity 25
5.6.9  GROUP EVALUATION 25
5.6.10 Other Notes Relating to the Proposal 26
5.7  WRITTEN FINAL REPORT (980 points) 27
    5.7.1  (150 points) PART I. Executive summary. 28
    5.7.2  (400 points) PART II. Statistical summary 28

6  TYPICAL PROJECT BLUNDERS 33

7  ADVANCED STATAQUEST FOR PROJECTS 35

8  HOMEWORK ASSIGNMENTS 39

9  EXAM PREPARATION & RULES 73

10 TEXAS A&M EXAMS, 1992 75

11 TEXAS A&M EXAMS, 1993 97

12 TEXAS A&M EXAMS, 1994 125

13 UNC EXAMS 173

14 VARIOUS PROBLEMS 179

15 COURSE DATASETS 189
CHAPTER 1

FIRST CLASS OVERVIEW

STAT 302
Spring 1995

Professor Raymond J. Carroll
Laboratory for Biostatistics and Biosystems Modeling
Department of Statistics
Blocker Building, Room 430
845-3141
e-mail: carroll@stat.tamu.edu

Course Packet: You must purchase a course packet containing the critical information you will need to fulfill the course requirements. The packet includes homeworks, project instructions, old exams, sample problems and past class projects.

Lecture Notes: You must purchase the lecture notes for the class.

Critical Dates are:
- Exam #1: Lecture #11
- Exam #2: Lecture #19
- Exam #3: Lecture #27
- Course Project: December 12, 1993

Exams will be held from 6:00–7:30PM on the dates given above. Exam rooms will be announced in class. The exams will be predominantly multiple choice. Review sessions will be held the night before each exam.
CHAPTER 2

COURSE SCHEDULE
<table>
<thead>
<tr>
<th>Lecture #</th>
<th>Date</th>
<th>Special Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8-30</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9-1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9-6</td>
<td>HW 1</td>
</tr>
<tr>
<td>4</td>
<td>9-8</td>
<td>HW #2</td>
</tr>
<tr>
<td>5</td>
<td>9-13</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9-15</td>
<td>HW #3</td>
</tr>
<tr>
<td>7</td>
<td>9-20</td>
<td>HW #4, Voluntary project teams formed</td>
</tr>
<tr>
<td>8</td>
<td>9-22</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9-27</td>
<td>HW #5, Project teams assigned</td>
</tr>
<tr>
<td>10</td>
<td>9-29</td>
<td>HW #6</td>
</tr>
<tr>
<td>11</td>
<td>10-4</td>
<td>Exam #1</td>
</tr>
<tr>
<td>12</td>
<td>10-6</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>10-11</td>
<td>HW #7, Potential topics assignments</td>
</tr>
<tr>
<td>14</td>
<td>10-13</td>
<td>HW #8</td>
</tr>
<tr>
<td>15</td>
<td>10-18</td>
<td>Team working agreement due</td>
</tr>
<tr>
<td>16</td>
<td>10-20</td>
<td>HW #9</td>
</tr>
<tr>
<td>17</td>
<td>10-25</td>
<td>HW #10</td>
</tr>
<tr>
<td>18</td>
<td>10-29</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>11-1</td>
<td>Exam #2</td>
</tr>
<tr>
<td>20</td>
<td>11-3</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>11-8</td>
<td>Written proposal due</td>
</tr>
<tr>
<td>22</td>
<td>11-10</td>
<td>HW #11</td>
</tr>
<tr>
<td>23</td>
<td>11-15</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>11-17</td>
<td>HW #12</td>
</tr>
<tr>
<td>25</td>
<td>11-22</td>
<td>HW #13</td>
</tr>
<tr>
<td>26</td>
<td>11-29</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>12-1</td>
<td>Exam #3</td>
</tr>
<tr>
<td>28</td>
<td>12-6</td>
<td>Project discussion and project checks</td>
</tr>
<tr>
<td>Finals</td>
<td>12-12</td>
<td>Final report due</td>
</tr>
</tbody>
</table>

Table 2.1. This table gives the homework and project assignment due dates throughout the course of the semester.
<table>
<thead>
<tr>
<th>Lecture #</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–17</td>
</tr>
<tr>
<td>2</td>
<td>32–42, 68–74</td>
</tr>
<tr>
<td>3</td>
<td>STATQUEST Demo</td>
</tr>
<tr>
<td>4</td>
<td>80–88, 94–98</td>
</tr>
<tr>
<td>5</td>
<td>98–101, notes</td>
</tr>
<tr>
<td>6</td>
<td>114–122, 141–153</td>
</tr>
<tr>
<td>7</td>
<td>Class notes</td>
</tr>
<tr>
<td>8</td>
<td>206–214</td>
</tr>
<tr>
<td>9</td>
<td>291–294, 302–317</td>
</tr>
<tr>
<td>10</td>
<td>340–368</td>
</tr>
<tr>
<td>11</td>
<td>Exam #1</td>
</tr>
<tr>
<td>12</td>
<td>368–373</td>
</tr>
<tr>
<td>13</td>
<td>390–402</td>
</tr>
<tr>
<td>14</td>
<td>403–409, 410–415</td>
</tr>
<tr>
<td>15</td>
<td>426–434</td>
</tr>
<tr>
<td>16</td>
<td>Class Notes, 487–493</td>
</tr>
<tr>
<td>17</td>
<td>487–493</td>
</tr>
<tr>
<td>18</td>
<td>448–453</td>
</tr>
<tr>
<td>19</td>
<td>Exam #2</td>
</tr>
<tr>
<td>20</td>
<td>520–521</td>
</tr>
<tr>
<td>21</td>
<td>598–618</td>
</tr>
<tr>
<td>22</td>
<td>626–627</td>
</tr>
<tr>
<td>23</td>
<td>626–627</td>
</tr>
<tr>
<td>24</td>
<td>123–132</td>
</tr>
<tr>
<td>25</td>
<td>405–406</td>
</tr>
<tr>
<td>26</td>
<td>No new material</td>
</tr>
<tr>
<td>27</td>
<td>Exam #3</td>
</tr>
<tr>
<td>28</td>
<td>No new material</td>
</tr>
</tbody>
</table>

Table 2.2. Text pages by lecture number.
CHAPTER 3

MECHANICS AND RULES

INSTRUCTOR: Professor Raymond J. Carroll

MY OFFICE HOURS

My office hours are

- Tuesdays, 9:45–10:30, 3:00–5:00
- Wednesdays, 3:00–5:00
- Thursdays, 9:45–10:30

If you need to speak with me otherwise, please make an appointment at 845–3141; ask for Ms. Holmes. Because I have many other teaching, research and administrative duties, dropping by my office generally won’t work.

I will also be available on Wednesdays from 6:00–7:00PM to answer any questions you might have. The room will be announced in class. There is no requirement that you attend these sessions.

Office hours are one of the few chances you get to talk with me about anything, including but not limited to class stuff. I hope you will all try to drop by at least once in the course of the semester. If nothing else, you can look over my collection of wildlife photos taken during my recent sabbatical (wolf, loon, grizzly bear, heron, kangaroo, coyote, antelope, koala, caribou).

EXAMS

There will be three exams, which will count 22%, 22% and 23% of your final grades. You are responsible for bringing into the class an operating calculator and all the statistical tables used in the class. I suggest that to be safe, you bring in 2 calculators. The exams are closed book, with the exception that for each exam, you are allowed THREE 8 × 11 inch pages of notes on which you may write anything you want, including examples, formulae, etc.

The night prior to each exam, I will hold a review session.

I do not believe that makeup exams are fair to those doing the makeup exams, because such students almost always do poorly. Ordinarily, excused absences from exams will thus be handled by
placing more emphasis on the other two exams.

Please note that you are responsible for getting an excuse. In case of illness, make sure that your doctor does not merely write the Quackshack nostrum “this student was seen by me”, but instead does state that you were not able to take the exam. Note the next paragraph about missing either more than one exam or the last exam.

In case you are asking to miss an exam because of university related activities, please note that you must first get my permission. I will at my discretion give a very hard make-up exam, especially if you miss more than one exam or if you miss the last exam. I understand that there are university related activities, but on the other hand the exam dates are available to you from the first class, and if you anticipate problems with attending exams, you should transfer to another section.

Exams will be given from 7:00PM–8:30PM. Rooms will be announced in class.

Please note that the registration packet gave slightly different times for the exams. We were not able to get rooms at the original times, and have had to switch to the times given above.

**HOMEWORK**

Homework will make up 8% of your final grade. No late homeworks will be accepted under any circumstances. However, your three lowest homework scores will be dropped and your average homework score computed from the rest. You may work together if you wish. The best way to do this is to first work the problems out yourself, and then compare solutions. Merely copying the correct answers will yield a good homework average but terrible test scores. You may ask the T.A.’s or the instructor for help in solving problems. The T.A.’s have been instructed to help you, but they may forget. If you run into problems of this type, either show them this piece of paper or tell me.

Because of lack of staffing, not all homework problems will be graded. For essentially all assignments, I will designate only a few questions for grading.

**PROJECT**

A group project will count for 25% of your final grade. The project requires a substantial amount of work, but you will find that it helps make statistics real to you. If you are not willing to put in the work, you may wish to consider changing to another section of the course. The course packet contains full details about the project. Most project groups do put in a substantial amount of effort, and the project thus forms a nice “fourth exam” where reasonably high grades can be expected.

My major concern about group projects is the possibility that some group members will “carry”
along one student who does no work, but who tries to get full credit anyway. I have set up a fairly elaborate system to try to prevent this. Not doing one’s fair share of the work but yet claiming full credit is equivalent to cheating on a written exam.

**FINAL GRADES**

Students scoring above a 90% average are guaranteed an A, students scoring above 80% are guaranteed a B, etc. Sometimes, grades are “curved” to take into account random differences in exam difficulty, although this has been unnecessary in some years. The “curve” will be based to adjust the grades both within this semester’s students and also my own perception of how well you have performed as a class relative to previous classes. If class performance overall is better than average, and if there are minimal disruptions, the grades will reflect these facts.

**E-MAIL AND PHONE**

My e-mail address is carroll@stat.tamu.edu. My office phone number is 845-3141. Please don’t call me at home.

**COMPLAINTS ABOUT GRADING**

Complaints about homework and exam grading should be taken up with me. The mechanics are as follows. You should write a separate note explaining why you think there was an error in the grade, attach the homework/exam in question, and leave the material with the statistics department (Blocker #447). I’d would appreciate your not giving me the material after class, because I have lost such stuff in the past.

**T.A.’s**

The name and office hours of the T.A.’s will be announced early in the semester. Their primary duty is to help you with homework and computing. If you experience any problems with this, please tell me. In the past, the T.A.’s have been both helpful and competent.

**COMPUTING**

We will use the PC package **STATAQUEST**, which you can purchase at the bookstore along with the textbook. In your student packet, with each homework assignment you will find a list of instructions for **STATAQUEST**. Both the T.A.’s and I will help you with **STATAQUEST** questions during our office hours.

You will be required to learn this package, what the output looks like, etc. The standard system is
on PC’s. Please remember that graduation requirements here assume that you are computer literate.

I apologize, but those of you with MAC’s will still be required to learn STATAQUEST on the PC. The MAC version of STATAQUEST is somewhat different from the PC version.

You should buy STATAQUEST and then bring your disk and your STATQUEST manual to class on the first 2 days and otherwise to the Statistics Department in Blocker #447. We will mark the book and disk, and then give you a 3.5 inch floppy disk containing STATAQUEST and all the examples we will use for homework.

The assignments become increasingly computer intensive as the semester wears on.

To use STATQUEST on your home computer, you will have to have a high density floppy disk drive. The third class lecture will include extensive discussion of this software.

SOMETHING ABOUT THE INSTRUCTOR

I was born in Japan of military parents and grew up in D.C., Germany and Wichita Falls, got a B.A. from U.T. Austin in 1971 and a Ph.D. in Statistics from Purdue in 1974. From 1974-1987 I was on the faculty at the University of North Carolina at Chapel Hill, also spending time at the Universities of Heidelberg and Wisconsin, as well as at the National Heart, Lung & Blood Institute. From 1987 to the present I have been a full professor here at A&M, along with having visiting appointments at the Australian National University and the National Cancer Institute. I was head of the department from 1987-1990. I currently work as a consultant to the National Cancer Institute, and hold two research grants from that institute. I also work with the U.S. EPA on risk assessment to inhalation of chemicals (such as those in dry cleaning establishments) and with the Texas Natural Resources Conservation Commission on population exposure to harmful airborne chemicals such as ozone.

My research interests lie in the general area of regression, which is covered in Chapters 11 and 12 of this course. I wrote a book, Transformation and Weighting in Regression (Chapman & Hall, 1988), which is essentially concerned with the material in section 11.5. My special interest is nonlinear modeling under unusual error situations. What I try to do is to develop new statistical techniques for solving important practical problems arising out of my consulting. The techniques I’ve developed have been applied to cancer epidemiology, marine biology, chemometrics, pharmacokinetics, marketing, and image processing, among others. The work has won my research team a number of awards and honors. I’m currently very interested in the problems of regression when some of the predictors are measured with error, and my newest techniques are being used by cancer epidemiologists in their study of the relationship between breast cancer and nutrition.

My hobbies are trout fishing, golf and cycling. I’ve also been doing some bass fishing lately, with mixed success.

DATA SETS
Most of the homework assignments will be based on my own data, as will most of the class examples.

**LEAVING CLASS EARLY**

I understand that there will be circumstances which require you to leave early from class. The rules for doing so are simple: please do so with a minimal amount of fuss and noise.
4.1 FORMULAE FOR EXAM #1

Empirical Rule for the Values in a population:
\[ \bar{X} \pm (\text{empirical critical value})s, \]
where the empirical critical values are 1.645, 1.96, 2.58 and 3.00 for 90%, 95%, 99% and 99.7% ranges.

Empirical Rule Interval for the population mean \( \mu \),
\[ \bar{X} \pm (\text{empirical critical value})\frac{s}{\sqrt{n}} \]

Empirical Rule Interval for the difference in two population means \( \mu_1 - \mu_2 \).
\[ \left( \bar{X}_1 - \bar{X}_2 \right) \pm (\text{empirical critical value})\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]

Empirical rule interval for the slope of a single regression line:
\[ b \pm (\text{empirical critical value})s_b, \]

Empirical rule interval for the intercept of a single regression line:
\[ a \pm (\text{empirical critical value})s_a, \]

Empirical rule interval for the difference in slopes of two populations:
\[ b_1 - b_2 \pm (\text{empirical critical value})\sqrt{s_{b_1}^2 + s_{b_2}^2}, \]
4.2 FORMULAE FOR EXAM #2

Computing probabilities about the sample mean:

\[
Pr(\bar{X} \leq c) = Pr\left( Z \leq \frac{c - \mu}{\sigma/\sqrt{n}} \right) \\
Pr(\bar{X} \geq c) = Pr\left( Z \geq \frac{c - \mu}{\sigma/\sqrt{n}} \right) = 1 - Pr\left( Z < \frac{c - \mu}{\sigma/\sqrt{n}} \right)
\]

Computing probabilities about the sample fraction:

\[
Pr(p \leq c) = Pr\left( Z \leq \frac{c - \pi}{\sqrt{\pi(1 - \pi)}/n} \right) \\
Pr(p \geq c) = Pr\left( Z \geq \frac{c - \pi}{\sqrt{\pi(1 - \pi)}/n} \right) = 1 - Pr\left( Z < \frac{c - \pi}{\sqrt{\pi(1 - \pi)}/n} \right)
\]

Confidence Interval for the population mean \( \mu \) if \( n > 30 \) and, as usually happens, the population standard deviation is unknown.

\[ \bar{X} \pm (z \text{ critical value}) \frac{s}{\sqrt{n}} \]

Confidence Interval for the population proportion \( \pi \).

\[ p \pm (z \text{ critical value}) \sqrt{\frac{p(1-p)}{n}} \]

Confidence Interval for the population mean \( \mu \) if \( n \leq 30 \) and, as usually happens, the population standard deviation is unknown.

\[ \bar{X} \pm (t \text{ critical value}) \frac{s}{\sqrt{n}} \]

4.3 FORMULAE FOR EXAM #3

Large Sample Confidence Interval for the difference in population means \( \mu_1 - \mu_2 \) if \( n_1 + n_2 > 30 \).

\[ \left( \bar{X}_1 - \bar{X}_2 \right) \pm (z \text{ critical value}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]

Small Sample Confidence Interval for the difference in population means \( \mu_1 - \mu_2 \) if
First define the pooled sample variance:

\[ s_p^2 = \left( \frac{n_1 - 1}{n_1 + n_2 - 2} \right) s_1^2 + \left( \frac{n_2 - 1}{n_1 + n_2 - 2} \right) s_2^2. \]

The degrees of freedom are \( n_1 + n_2 - 2 \). The confidence interval is

\[ \left( \bar{x}_1 - \bar{x}_2 \right) \pm (t \text{ critical value}) \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}. \]

**Large Sample Confidence Interval for the difference in proportions \( \pi_1 - \pi_2 \):**

\[ (p_1 - p_2) \pm (z \text{ critical value}) \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}. \]

**Confidence interval for the slope of a single regression line:**

\[ b \pm (\text{critical value})s_b, \]

where the critical value is the \( z \)-critical value if \( n > 30 \) and is the \( t \)-critical value with \( n-2 \) degrees of freedom if \( n \leq 30 \).

**Confidence interval for the intercept of a single regression line:**

\[ a \pm (\text{critical value})s_a, \]

where the critical value is the \( z \)-critical value if \( n > 30 \) and is the \( t \)-critical value with \( n-2 \) degrees of freedom if \( n \leq 30 \).

**Confidence interval for the difference in slopes of two populations:**

\[ b_1 - b_2 \pm (\text{critical value}) \sqrt{s^2_{b_1} + s^2_{b_2}}, \]

where the critical value is the \( z \)-critical value if \( n_1 + n_2 > 30 \) and is the \( t \)-critical value with \( n_1 + n_2 - 4 \) degrees of freedom if \( n_1 + n_2 \leq 30 \).

**Confidence interval for the difference in intercepts of two populations:**

\[ a_1 - a_2 \pm (\text{critical value}) \sqrt{s^2_{a_1} + s^2_{a_2}}, \]

where the critical value is the \( z \)-critical value if \( n_1 + n_2 > 30 \) and is the \( t \)-critical value with \( n_1 + n_2 - 4 \) degrees of freedom if \( n_1 + n_2 \leq 30 \).

**Prediction interval for an actual response in a population (normal range):**

\[ a + bX^* \pm (\text{critical value})s_e \sqrt{1 + \frac{1}{n} + \frac{(X^* - \bar{x})^2}{(n - 1)s_x^2}}. \]
where the critical value is the $z$-critical value if $n > 30$ and is the $t$-critical value with $n-2$ degrees of freedom if $n \leq 30$. 
CHAPTER 5

CLASS PROJECT RULES

Outline of Term Project

(1) Roster of students by majors (from university)  Lecture #6
(2) Initial Formation of Project Teams:  Lecture #7
(3) Assignment of Project Teams:  Lecture #9
(4) Potential Topics Assignment:  Lecture #13
(5) Team Working Agreement Assignment:  Lecture #15
(6) Written Proposal:  Lecture #21
(7) Final Report:  See schedule

5.1 GRADING OF PROJECTS

It is absolutely required that you follow the directions closely. Failure to present report in exactly
the order and form specified will cause substantial loss of credit.
5.2 GETTING STARTED ON THE TERM PROJECT

5.2.1 CREATING A TEAM

Create a team of three to five individuals. You have until Lecture #7 to do this by yourself, and report the team members to me. Team members may be from either section that I am teaching. For those of you who are not able to form a team, I will make the assignments for you on Lecture #9. I will grade projects, and will not “curve” on the basis of number of team members.

How well one’s team works together is often one of the more satisfying or dissatisfying aspects of the project. Here are some ideas and questions which some students in previous terms found helpful in clarifying their expectations of each other as they were forming their teams:

* Make sure that there are times in the week when you can get together as a team.
* What do you expect each other to contribute?
* How will you contact each other about meetings?
* How much advance notice will you give before meetings?
* Be specific. Write down your agreements. The Team Working Agreement Assignment will give you a chance to earn some points for doing this.
* Identify a team “contact author.” This person submits works, receives the graded work, and is my contact if I have questions about your project.

5.2.2 THE LAZY PLAYER

Inevitably, there will be a few cases that one person in a group does little or no work, but yet wants to claim a full grade. I consider this cheating. To try to cut down on the problem, there are two places in the semester where I will ask you to evaluate the work of the other members in their group, and in particular whether one member is not doing his or her fair share.

5.2.3 THE QUESTION

Make sure that the question your team chooses to investigate has these properties:

A. It is of interest to you.
B. This question can be answered using methods discussed in this course.
C. The entire project can be done in approximately 60 total hours for the team.

5.2.4 TYPING OF PROJECTS

Typing instructions for all work (Potential Topics Assignment, Team Working Agreement Assignment, Proposal, and Final Report) are as follows.

A. All work is to be typed, double spaced on 8 1/2” X 11” paper using type size at least as big as elite type.

B. If you use a computer for data analyses and graph construction, all relevant computer output is to be presented on 8 1/2” X 11” paper. Label it as “Figure 1”, “Figure 2”, etc. and refer to it in the text as “Figure 1”, “Figure 2”, etc. Submission of computer output alone on 11” X 17” pages is NOT acceptable. Treat all hand-constructed graphs and tables in this way also.

C. Organize your answers in numbered and lettered paragraphs corresponding to the number and letter of the questions being answered. It is critical that you do
this so that your work can be accurately graded.

D. Please staple your pages together. Paperclipped, folded or loose papers run the risk of being lost or misplaced.

E. Number your pages.

5.3 YOUR RESPONSIBILITIES

It is your responsibility to contact your other project team members and be available to meet with them. In the past, students have tried to avoid working (while still claiming full credit) by citing conflicts such as student government, drill team, etc. I do not want to give full credit to those who use such excuses.
5.4 POTENTIAL TOPICS ASSIGNMENT (MAXIMUM OF 2 PAGES). 140 points.

5.4.1 Basic Idea of the Assignment

Part of the purpose of this assignment is to give you a chance to form a team and see if you can work together.

You are to identify two projects:

(1) A comparison of two population means

(2) A project in which you identify regression relationships between two continuous or discrete variables (height and weight, petal and sepal length, etc.). Please note that full credit will be lost if any of the variables in the relationship are categorical. You might want to compare the relationship between two continuous or discrete variables across two populations, e.g., GPA and SAT scores for men and women.

In identifying (1) and (2), you may consult with your advisor or faculty/graduate students in your department of major interest. The problems do not have to be biologically oriented. Use your imagination! Here are a few examples of previous projects:

- Comparing the reported family income of women in dorms who have and have not joined a sorority.
- Comparing left handers and right handers to see who is more ambidextrous in shooting nerf-ball baskets in a dorm room.
- Studying the time from entrance into a bar to first restroom break for men and women on the basis of height, body mass (big, medium, small) and number of beers consumed.

5.4.2 The Specific Assignment

For each of projects (1) and (2), answer the following questions.

a. What question do you propose to study? Give me a brief answer that would have been understandable by my mother (who was not a statistician)

b. What population or populations will be sampled? This is the question most of you will miss. I want to see the population or populations that are sampled to answer the question you have posed.

c. What variables do you propose to measure?

d. For project (2), I want to see a mock-up of a scatterplot, with $Y$ and $X$ clearly labelled, and your best guess at what the line or lines will look like.

Notes:

1. You must submit this assignment as a team, and it will be graded as such.

2. Each question (there are a total of 6) will count 20 points.

3. In your term project, you will either make a comparison or identify relationships, but not both. The purpose of this assignment is to encourage you to think about projects before you pick one to study in depth.

4. Please label your report as follows:

   Project (1) [Comparison]

   a.

   b.

   c.

   Project (2) [Relationship]
I am particularly concerned that you be able to distinguish between a sample and a population. Not doing so carefully will cause you to lose points.
5.5 TEAM WORKING AGREEMENT ASSIGNMENT:

This assignment is worth 100 points.

5.5.1 Purpose of the Assignment

At later stages of the project, I will be asking you as a group to evaluate the effort/time put in by the other project team members. Your grade will depend on whether the other project members basically agree that you were a full partner.

In this assignment, I want you to think ahead to what each of you like to do. Someone might like computers, another writing and a third be willing to do major league grunt work. Try to think how you will make sure that everyone does their fair share.

5.5.2 Work to be Done

Construct your Team Working Agreement in which you address the issues discussed in Section I. and related ones, sign it, give a copy to everyone, and have the contact author turn in a copy. In particular, I want to see that you have thought about how you will warn me that someone is not doing a reasonable amount of work.
5.6 WRITTEN PROPOSAL (460 points)

5.6.1 The Purpose of this Assignment

The purpose of this proposal is to convince me to approve your project and to approve your team as competent to do it. Both goals will be achieved by a clear, complete description of all aspects of the project.

Proposal Table of Contents: (460 points possible)

Cover Page: Signed original work statement (from the box below). (Put this page on the very front of your proposal; no blank pages on top of it.)

Pages 1–3: (a maximum of 3 pages) Answers to Tasks 1–4. [Make sure you first state and then answer the question.]

Page 4 on: Appendix: Questionnaire (if appropriate)
Appendix: Verbatim script (if appropriate)
Appendix: Mock-up of graphs

Evaluation: Each of you are to include a sealed envelope evaluating the effort of your fellow team members. I am particularly concerned to know whether someone has not been available, and if so their grade will be adjusted.

ORIGINAL WORK: Students in this course may discuss with each other and with the course staff the selection of a project topic and various questions they encounter while doing the term project. The final choice of a project topic, resolution of difficult questions, interpretations, and composition of project documents are to be done only by team members.

SIGN THE FOLLOWING STATEMENT AND INCLUDE IT ON THE COVER PAGE OF YOUR PROPOSAL:

ORIGINAL WORK STATEMENT

Project Title:

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

<table>
<thead>
<tr>
<th>Signatures</th>
<th>Typed Names</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Author:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.6.2 Some Notes About Grades for Projects

The easiest projects invariable involve the comparison of two population means. The next easiest is the comparison of two population proportions. The next most difficult involve regression relationships. Finally, the hardest projects involve doing both regression relationships and the comparison of two population means.

It seems to me not unfair to in effect give more credit for those of you who are willing to do the hardest type of project. You will see later on in this section that I have specifically allowed for this possibility by means of evaluating the proposal on the basis of thoroughness. However, I also give credit for originality. Thus, a project which is really dull and boring but does hard statistics cannot be expected to get higher marks than a project which is easier statistically but involves a clever problem or one of some intrinsic interest in the life sciences.

5.6.3 Some Notes about Lurking Variables

In almost any statistical problem of comparing two populations, there will be lurking variables, i.e., variables which if not measured will cause a distortion of the results. I will mention a few of these in class, and the former students who present in class will also mention lurking variables. I expect that your project will consider and control for lurking variables, and thus that very few projects will consist only of a simple comparison of two populations.

5.6.4 (40 points) The question to be studied

What question do you propose to study?

5.6.5 (150 points) The statistical questions

A. (60 points) What is the population or populations which you will study? Be careful here. I am a real stickler for precision about populations, so don’t merely say something like “students at Texas A&M University”. You will want to say something like “the populations are males who are regular drinkers, males who don’t drink, females who are regular drinkers and females who don’t drink”.

B. (30 points) What variables (at least two) will you measure?

C. (60 points) What parameters of the population or populations are of interest to you? Note: you should use the phrase “population parameter” in its statistical sense as it has been used in class. The statistical definition specifically does NOT mean limits on observations. It means things such as the population mean, the population fraction, the population intercept and the population slope. Absolutely nothing should be said about the sample.

5.6.6 (100 points) Data collection plans. [DO NOT START COLLECTING DATA UNTIL YOUR PROJECT HAS BEEN APPROVED.]

A. (50 points) Sampling Process:

   HOW will you select your sample?

B. (50 points) Measurement Process:

   How will you measure your variables? How will you obtain as much reliability and as little
bias as possible? If data are collected by more than one observer, how will you verify that all observers are measuring the same way throughout the study?

If you plan to use a questionnaire, describe how you will get it to your subjects. If you propose to gather data directly from people by telephone interview, face-to-face interview, or some other method, describe explicitly how you will approach these individuals and what you will say to them. Give a verbatim script in an appendix. Include a copy of your questionnaire in an appendix.

5.6.7 (70 points) Data analysis plans

A. (40 points) Give exactly two graphs that you plan to use in describing your results. I have a fairly flexible definition of a graph. For example, two box plots comparing two populations counts as one graph. Present in an appendix a mock-up (drawn by hand is OK) of each graph you propose to use, with the axes labelled. If you find that you are about to suggest a graph that has not been used in class, you are almost certainly confused. If you do a regression analysis, label the X and Y axes and draw in the least squares line that you expect will result.

B. (30 points) What numerical descriptive statistics do you plan to compute for what variables? Basic descriptive statistics that measure central tendency and variability may be useful for your quantitative variables. Regression slopes and intercepts may also be useful. For qualitative variables, present proportions in categories. Don’t just tell me you will measure sample means. Be more specific, tell me on what variables, etc.

5.6.8 (100 points) Thoroughness and Creativity

I will give a subjective evaluation of your proposal, based on two factors. First, is the idea really clever or of scientific interest, and second is whether you seem to have done a reasonably thorough job of things. I am hoping that this will induce you to think of a project which is both interesting and leaves scope for some statistical analysis.

For example, if you state that you plan to compare GPA’s of males and females, you are being neither original or thorough. For example, there are lurking variables (classification), and there are also possible regression relationships (with SAT).

Most groups will get around 60 points for this part. A few will get 30 or less, and a few will get 90 or more.

5.6.9 GROUP EVALUATION

Do not forget to hand in sealed letters to me in which I can tell whether the group is functioning adequately.
5.6.10 Other Notes Relating to the Proposal

1. I suggest that you spend approximately 60 hours as a team on this project. You may spend additional time if you are convinced that the benefits of spending additional hours on this project are greater than the costs.

2. Keep a diary (lab book) during the project, recording expenditures of time, money, and other resources, as well as any notable events which may influence the interpretation of your data analysis. Keep track of which team members attend each meeting. **Turn this diary in with your final report.** (It may be neatly handwritten.) Credit will be awarded for this effort.

3. You will invest a certain amount of time, energy, money, and other resources in doing this project. As in the case with all investments, the amount of return you will receive on this project both now and later in your career depends on the amount you invest in it and the risk you take. In the past, some students have done a minimal amount of effort, gotten a “C” on the marginal project, and regarded the whole effort as a waste of time.

Other students have invested time and energy necessary to identify a problem of interest to them (**a KEY step**), worked vigorously on the project, gotten an “A” on an excellent report, used the report on job interviews, and regarded the project as the activity that really made statistics “come alive” for them. The choice is yours.
5.7 WRITTEN FINAL REPORT (980 points)

(HAND IN 2 COPIES) MAXIMUM 4 PAGE REPORT. IN ADDITION, YOU ARE ALLOWED 4 PAGES OF SUMMARY STATISTICS AND CONFIDENCE INTERVAL CALCULATIONS, TWO PAGES OF GRAPHS, DATA LISTING, DIARY AND EVALUATION LETTERS

Observe the space limit!!!

Late Projects will not be Accepted

ORIGINAL WORK: Students may discuss with each other and with the course staff various questions they encounter while doing the term project. The final computer work, resolution of difficult questions, interpretations, and composition of the final report is to be done only by team members.

SIGN THE FOLLOWING STATEMENT AND INCLUDE IT ON THE COVER PAGE OF YOUR FINAL REPORT:

ORIGINAL WORK STATEMENT

Project Title:

We, the undersigned, certify that we did gather, in the manner we have described, the data analyzed in this report. We also certify that the actual composition of this report and any associated computer work were done by us and are original work.

<table>
<thead>
<tr>
<th>Signatures</th>
<th>Typed Names</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Author:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cover Page: Signed Original Work Statement (from above). (Put this page on the very front of your final report; no blank pages on top of it.)

Page 1: Part I: Executive Summary.

Page 2 on: Part II: Statistical summary.
Observe the space limit!!!

5.7.1 (150 points) PART I. Executive summary.

This is to be a report to an intelligent person who has NOT taken a statistics course. Thus, do not use technical terms like sample, hypothesis test, confidence interval, etc. The best way to think of this is that you are writing the summary so that your parents can understand it.

Prepare a summary (NO MORE THAN ONE PAGE) of your project and its findings for the person in charge of the process you studied, answering the following questions, in numbered paragraphs:

1. (50 points) What question did you study?
2. (20 points) How did you study it? (Briefly sketch your sampling and measurement processes.)
3. (80 points) What are your major findings? How do they compare to what you expected to find?

5.7.2 (400 points) PART II. Statistical summary

This is to be a report to your statistician supervisor who has taken STAT 302. Be concise. Three page maximum of text excluding graphs and tables. Make this report in a question and answer format.

1. (70 points) What populations were studied? Be precise, stating all the populations you studied. Obviously, depending on the question you asked, you will have different populations. Please indicate this.
2. (10 points) What variables were studied?
3. (45 points) What were the population parameters of interest?
4. (30 points) Describe your sampling process.
5. (25 points) Present your complete sample of values in an appendix (may be neatly handwritten).
6. (150 points) Present your data analyses, calculations of inferential statistics, graphs, and figures (may be neatly handwritten). You can simply place this material in the appendix.
   A. Present numerical descriptive statistics.
   B. Present at most two pages of graphs, using the same definition of a graph as in the written proposal.
   C. Present confidence intervals. You should note that you are allowed an appendix where you give confidence intervals, but you must show your work (computer output neatly labelled and annotated counts)! I suggest that you do 90%, 95% and 99% confidence intervals. More simply, just do a 95% C.I. and state and interpret the associated p-value.
7. (250 points) What are your findings? Using technical statistical terminology, interpret the results of your experiment. You should interweave the numerical summaries
such as means with the inferential summaries such as confidence intervals, using both to describe your results and conclusions. The most common error people make here is in overinterpreting confidence intervals. If you find a statistically significant difference between two populations with a 90% confidence interval, but not at 95% or 99%, the correct conclusion is that there is a 90% chance that the two populations differ. An incorrect conclusion is “there is clearly no evidence that the populations are anything but the same”.

8. (50 points) Present the actual project diary, which is discussed on the Project Proposal Assignment. (May be neatly handwritten.) The diary should include who attended each meeting, so that I can get a clear sense of the relative contributions of each member.

9. (200 points) I will give a subjective grade of the quality, originality and effort you have shown in the project writeup. See the grading sheet I used in 1993.

10. Each team member must turn in a sealed letter with the project indicating whether any one member has not been a fully functioning member of the team.
PROJECT GRADE SHEET: 1994

EXECUTIVE SUMMARY

TOTAL POINTS POSSIBLE: 880
TOTAL POINTS SUBTRACTED:

SCORE ON THIS PART OF THE PROJECT:

Questions to be Studied:
- (10) Ambiguous
- (20) No distinction between population and sample

Sampling and Measurement Process
- (10) Mostly clear but with some ambiguities
- (20) Mostly clear but with many ambiguities
- (30) Ambiguous

Major Findings
- (30) Leaving the impression of accepting a null hypothesis
- (40) “Proving” a hypothesis, “definitely” an effect, etc.
- (20) Stated results don’t match the analysis
- (20) Ambiguous, so that mom would be confused
- (20) Mom doesn’t understand confidence or correlations (read the directions!)

POPULATIONS

Populations
- (20) Any errors (by this time, you know I insist you get this right.

Population Parameters
- (30) Any errors
- (20) Stating one set of population parameters, but then doing an analysis with others, e.g., not mentioning regression slopes, but doing a regression analysis

SAMPLE

How you got your samples
- (15) Mostly correct but with some ambiguities.
- (25) Mostly correct but with major ambiguities.
- (45) Noninformative statements.

Data
- (25) Either not given or not reproducible by me.
DATA ANALYSIS

Numerical descriptive statistics
(-30) Not available, or not easily readable by me

Graphics
My feeling is that by this time in the semester you should know how to do graphs.
(-40) Any error
(-15) When comparing population means and proportions for more than two different
variables (e.g., sex and age category), not plotting them simultaneously as I did many
times in class (Mean Plot).

Inference
(-80) Misapplication of formulae
(-80) Using the wrong test to get at what you want as described in the executive summary.
(-80) Calculation errors in applying confidence interval formulae.
(-20) Not doing 90%, 95% and 99% confidence intervals

RESEARCH FINDINGS

(-70) Accepted a null hypothesis, with no indication that you understand that this is a
dangerous business because you control sample size and confidence level.
(-40) Accepted a null hypothesis because you used sloppy terminology.
(-50) Found a difference among populations (rejected a null hypothesis) without saying
how much difference exists (without a CI – I’ll accept the mere statement of the interval
itself.)
(-60) Did not state the level of confidence you used.
(-40) Where appropriate, you should have used different confidence levels to see how strong
any differences found would be. For example, you may have found “statistical significance”
at the 90% confidence level, but what about 95% and 99%? That’s what p-values are for.
(+40) Quality points for a great job of describing your findings.
(-20) Minor ambiguities in description of the findings.
(-40) Major ambiguities in description of the findings.
(-80) Incomprehensible description of the findings.
(-100) Major blunder in the description, even based on your calculations.
(-70) Over- or misinterpret confidence intervals.
(-50) Misconception about sample size, e.g., that differences in sample sizes causes differ-
ences in means.
(-50) Mistakenly (sloppily) say that you do not reject a null hypothesis when you have, or
vice-versa.
(-150) Tried to see if two populations were different by computing confidence intervals on
each and then comparing the confidence intervals. This is wrong, wrong, wrong.
**PROJECT DIARY**

(-50) Not included.

**QUALITY, ORIGINALITY AND EFFORT**

**THE PROBLEM**

(0) Top 15% project ideas.

(-20) A good project idea but not in the top 10%, while at the same time considering lurking variables.

(-40) A good project idea but not in the top 30%, while at the same time considering lurking variables.

(-50) Not considering obvious lurking variables

(-90) A project which is of little interest and has to my mind been chosen to be as simple as possible.

(-40) A problem which could have been improved by slightly expanding the study, e.g., to study gender differences or time of measurement differences, and this fact was not brought up in the project itself.

**THOROUGHNESS**

(0) Incredibly thorough as well as being selective in presenting results.

(-15) Thorough, very good project, but not in top 10%.

(-30) Thorough, very good project, but not in top 30%.

(-40) Nothing special on thoroughness but clearly substantial work put in.

(-75) In my opinion, did the minimum possible to get the project done.

(-90) Many missed opportunities for further analyses, opportunities which would have vastly improved the project.

**FOLLOWING DIRECTIONS**

(-50) Exceeding the page limit by one.

(-75) Exceeding the page limit by two.

(-100) Exceeding the page limit by more than two.

(-25) Only 1 copy of the project submitted
CHAPTER 6

TYPICAL PROJECT BLUNDE...
In the instance, what you really want to know is whether the percentage (fraction) or men who wear maroon backpacks exceeds the percentage of women who wear maroon backpacks.

- Here is another classic error, again having to deal with categorical data. Many of you will have made a statement like “we want to compare those who use Q-tips after showering to those who do not”.
  - This makes no sense at all, and certainly does not answer part 1(a)–1(c).
  - The problem here is that there is exactly one population (those who take showers), the experiment is binomial (use or not) and the outcome variable is the percentage who use Q-tips.

- Now consider problem 2. The directions are completely explicit that you should identify two numeric variables (continuous or discrete). Categorical variables will not do.

- The directions even say that full credit will be lost if any of the variables in the relationship are categorical. Read the directions!

- Remember, we talked about regression. I want you to identify regression relationships. The intent of the question is to see if you know what a regression relationship is.

- You should pick two variables (i.e., numbers remember) which might be related and which could be graphed against one another.
  - I want to know what $X$ is and what $Y$ is.
  - You might ask the question: is there a relationship between $X$ and $Y$.
  - Or you might want to know whether the relationship (i.e., population slope) between $X$ and $Y$ is the same in two different populations.

- If you have not been able to contact a member of your team, they are history. This is part of the team working agreement.

- Don’t begin to gather data until I have approved the written proposal.

- In general, read the directions carefully, and follow them to the letter.
CHAPTER 7

ADVANCED STATAQUEST FOR PROJECTS

I am going to show you here how to do some advanced Stataquest calculations that will be needed for the project.

PUTTING TWO LINES ON A PLOT:

This is done in homework #12.

TESTING WHETHER TWO SLOPES OR INTERCEPTS ARE THE SAME

In homework #12, you had two populations (gender) and you wanted to regress $Y = \text{loga}$ on $X = \text{logy}$. You defined a new variable $\text{logygen}$. After you have done this, hit Statistics then Multiple Regression and accept $\text{logy}$ as the $Y$-variable. Type for the $X$-variables $\text{logygen}$ gender logv. Answer “no” for “stepwise”. The confidence interval for the difference in two slopes is the row for logygen, while for the difference in two intercepts it is the row for gender. Hit the escape key to get back to the main menu.

CONSTRUCTING 4 BOX PLOTS BASED ON TWO CATEGORICAL VARIABLES

In the suicide data, there are two binary variables of interest: sex and treatment assignment. This forms four populations. I want to construct box plots for all four populations, simultaneously. The variable “sex” has values 0 and 1, as does the variable “treatment”. To get four box plots, hit Enter spreadsheet then F10 key, then Add then Variable. Call the new variable “group”. Hit F10, Replace then Formula and type sex + (2* treatment).

Now hit F10, Files Save and answer yes twice. Hit escape key to get back to main menu. Then hit Graphs One Variable by Groups, Box Plot Comparisons, select the variable (“iq” in my example) to graph and then the variable identifying the groups (“group” in my example). You’ll have 4 box plots (save the page). To get means and s.d.’s, get back to main menu, hit Statistics, ANOVA, One Way, iq, group and this gives the list you need.

CONSTRUCTING A MEAN PLOT BASED UPON TWO CATEGORICAL VARIABLES
Note that you have a table

<table>
<thead>
<tr>
<th>Sex</th>
<th>treatment</th>
<th>$\bar{X}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>100.5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>96.8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>92.0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>98.1</td>
</tr>
</tbody>
</table>

From the main menu, hit Edit, Spreadsheet, Files, New, answer no to save data, answer 4 for “Number of observations”, answer 3 for “Number of variables”. Hit any key. Hit F10, Replace, Rename Variable and give it a name, say sex. Move cursor to x2, hit F10, Rename Variable and give it a name, say treatment. Move cursor to x3, hit F10, Rename Variable and give it a name, say Mean IQ. Now type in the data. Then move cursor to sex, hit F10, Label Values and give it labels, say “sex==0” becomes Females and “sex==1” becomes Males. Move cursor to treatment, hit F10, Label Values and “treatment==0” is Control while “treatment==1” is Treat. Hit F10, Files, Save and give the dataset a name, like mplota. Hit any key, then escape key. Go to Graphs, hit Scatter Plots, Plot Y vs X, naming points. Y-axis is Mean IQ and X-axis is the one you want on the X-axis, say sex, while “labeling variable” is the other one. Save graph for printing.
PROBLEM #1: Are there any applications of statistics in your major? To find out, talk to a professor or graduate student in your major and report briefly (≤ 1 page) on a project of theirs in which they use statistics as part of the analysis. You may be surprised.

Some Notes about Homework Grading: As discussed in the introductory notes, on any given assignment not all homework problems will be graded. The papers will come back to you with almost nothing on them except your score, and possibly a red line through mistakes in the problems that were graded.

Since not all problems will be graded, this means that there may be situations that you have worked all the problems correct except one. If I happen to pick that one to be graded, and only two problems will be graded, this means that your score for the homework will only be 50%. In order to help cut down on the variability inherent in such a system, I will drop your 3 lowest homework scores.

Please do not think that just because you did not lose credit for a problem that this means you did it correctly. For one thing, not all problems are graded. For another, the graders make mistakes. You should consult the answer sheet to see how you did.

Finally, in some circumstances the answer sheet and what I do will not coincide, because the answer sheet is made up by graduate students. The correct method is always the one I use in class.

Special Note About Computing: As you work through these assignments, you will see that I have given you commands that will enable you to do all the computing for the homework assignments. I have checked the typing over pretty carefully, but I cannot guarantee that all typos have been removed. If you find that there is a problem, please tell me.
HOMEWORK ASSIGNMENTS

HOMEWORK #2, due Lecture #4

This is your first homework assignment using StataQuest. Because many of you are not used to using the computer, you can expect that it will take longer than the usual assignment. Please note that I will spend a considerable portion of one lecture introducing you to StataQuest, so I suggest you wait until that lecture is over before attempting to work this assignment.

Each of you has from me a 3.5 inch floppy disk containing the datasets that will be used in the course. You will need to find a machine which accepts such diskettes. If for any reason you have troubles with your diskette, please tell me or one of the graders and we will provide a replacement.

The diskette drive you put the floppy disk into will be called either the A or B drive, depending on the machine. You have to figure out which one, but this can be done by trial and error. First place the diskette into the drive. Then type

a: [RETURN]

where [RETURN] means to hit the carriage return key, sometimes labelled by “enter”. If the 3.5-inch drive is the A drive, then your machine should read something like A:\> . If your diskette is really in drive B, then you’ll get a nasty message; just type a (for abort), then type

b: [RETURN]

Some of you will have your own PC’s with a hard disk. You can install StataQuest onto your hard drive as follows. Assuming your 3.5 inch floppy is a: and the hard drive is c:, just type

a: [RETURN]

and then type

install from a: to c: [RETURN]

The prompts you will see are very easy to answer.

Now you can enter StataQuest by typing

   go [RETURN]

You will see a menu with 7 items, with the first item [Starting Stata Quest] highlighted. Choose it by hitting [RETURN]. StataQuest will then be loaded. There will be some messages and the line “--more--” on the screen. Hit the space bar and you will be in StataQuest’s command line.

When StataQuest prints output on the screen, it stop at one screen-full and display the line “--more--” at the bottom of the screen. Whenever you see this, just hit the space bar on the keyboard to go on.

You should now see a “dot prompt” toward the bottom of the screen. Now hit the F10 key. You should see the StataQuest menu on top of the screen.
Throughout these notes, you will see boxed commands. What this means is that you should move the cursor keys (the ↑, ↓, ←, and → on your keyboard) to the label in the box, and then hit the enter key.

Some students in the past have become very upset with the need to use a computer in this course. The reasons for doing so are many. First of all, your using a computer allows me to assign serious problems for homework assignments, based on examples with real data. Secondly, using the computer will help you in your projects at the end of the semester. Thirdly, this course is an introduction to Statistics, and I feel that you cannot get a feeling for statistics without analyzing data.

I will happily help you with StataQuest during my office hours, as will the T.A.’s for the course.

StataQuest is “case-sensitive”. What this means is that the names “cats” and “Cats” are different from one another. You should make a habit of typing everything in lower case letters.

**PROBLEM #1:** This first problem is meant to give you practice at inputting data into StataQuest, and doing some simple manipulations. You may want to save this example for later use in your project.

The following data are from a physiology lab experiment performed by a student in the spring 1992 STAT 302 class. The response is the reaction time to stimulus in 10th’s of a second. For males, the auditory reaction times are

2.35, 2.19, 2.57, 2.67, 2.11, 1.71, 2.23, 2.92, 2.81, 2.04, 2.02, 3.95

while for females the times are

1.67, 2.16, 1.96, 1.68, 1.92, 2.11, 2.23, 2.63.

I would like you to input this data and perform a few simple manipulations. Here is how you should input the data. First get into StataQuest and start the menu mode as described above.

There are many places where StataQuest asks you to “Hit any key” to continue. Please go slowly through these instructions and look for this prompt. The answer is, of course, to hit any key! I will try to indicate most of these by the term HAK, but it is your responsibility to find them all.

Then do Files → Session logging (which means choosing “Session logging” from the “File” menu). You will be prompted for a file name. Type

```
hw21
```

which will record the commands and output to a file called hw21.log, which you can print later (see below). It will then require you to HAK.
Now move the cursor keys to Edit and do Edit → Spreadsheet, which will get you into the StataQuest spreadsheet. It will tell you to “open a file”, and ask you to press a key to go on. Hit any key. After that the Files menu will be dropped down. Choose New from that menu. StataQuest will ask you how many observations you have. Type in 20
It will next ask for number of variables. Type in 2

Now it asks you to “hit any key to continue”. Do so. Then you will be in the StataQuest spreadsheet. Press the F10 key to get to the menu. Do Replace → Rename variable. It will ask you for a new name for “x1”. Type in time
You will be back in the spreadsheet with “x1” changed to “time”. Move the cursor keys so that the black bar is in the column labelled “x2”. Repeat the steps to change “x2” to “sex”. Now you should be in the spreadsheet with the first column labelled “sex” and the second labelled “time”.

You can now start typing in the data. Use 0 for male and 1 for female in the sex column. After typing in the number in a cell, you can move to the next row by either hitting ← or ↓. To move to the other column, use either ← or →. After you have enter all the data, press F10 to get to the menu and do Files → Save. StataQuest will ask you for a name of the data set. Enter aud
Note that you don’t need to specify an extension. The extension .dta will be automatically attached to the file (i.e. the data will be saved to the file aud.dta). After saving the data you can get back to StataQuest by doing HAK then F10 and then do Files → Quit editor. If you have problems, you can always hit the escape key (usually “Esc” on your keyboard) to get back to the main menu.

The next thing I want you to do is set up a data set for males and a data set for females. To do this, get back into the editor by Edit → Spreadsheet. You will see the data you just typed in the spreadsheet. Move to a cell for male (0 in the sex column). Now hit the F10 key and do Drop → Observations. Note the plural. It is not Observation. StataQuest will ask you if it is OK to delete observations with sex==0. Answer “y”. The data for males will be deleted from the spreadsheet. Now hit the F10 key and then do Files → Save. You will be asked if you want to overwrite the current file. Hit ← to not overwrite the original data file. You will be prompted for a new file name. Type audf

To call back the original data set, HAK and then the F10 key and do Files → Open. StataQuest will ask you if you want to save the data currently in memory. Answer “no”. You will then be prompted will name of file to open. Type
HOMEWORK ASSIGNMENTS

It will ask you to hit any key: do so.

Move to a cell for female (1 in the sex column). Again hit the F10 key and then do Drop → Observations. You will be asked if it’s OK to drop the 8 observations with sex==1. Answer “yes”. Now save this data set to “audm” by the same steps as for the female data, namely F10, then Files → Save. You have to hit the ← key to not overwrite the file, and then you answer with

audm ←

Finish with HAK.

Finally, in some cases it might be useful to have the logarithm of the reaction time. Load the original data (in “aud”) as mentioned above. The easiest way to do this is to hit the escape key until you get to the StataQuest menu, then move the cursor keys and do Edit → Spreadsheet. Then hit the F10 key and do Files → Open. StataQuest will ask you if you want to save the data currently in memory. Answer “no”. You will then be prompted will name of file to open. Type

aud ←

It will ask you to hit any key: do so.

Now hit the F10 key and then do Add → Variable. You will be prompted for a name of the new variable. Type

logtime ←

A new column will be added with the heading “logtime”. Hit the F10 key and do Replace → Formula. You will be prompted for a formula for logtime. Type

log(time) ←

The column should be filled with the natural logarithm of the reaction times. Save the file by the following commands. Hit the F10 key and do Files → Save and answer “yes” twice when asked if you want to overwrite the file. Finish by HAK.

Your last part of this problem is to get a stem and leaf plot of all the data. To do this, quit the spreadsheet if have not already done so by hitting the escape key until the StataQuest menu shows up. Then move the cursor keys and do Graphs → One Variable → Stem and leaf. You will be prompted for the name of the variable. Type

time ←

and the stem and leaf plot will show up on the screen. Hit any key to get back to the menu.

Note: The stem and leaf plot you will see is not very informative. StataQuest seems not to give good stem and leaf plots. This is a design flaw they are working on.
Now you can quit StataQuest by doing `Files → Quit`. You will be asked if you want to save the data currently in memory. Answer “no”. You will be back to the initial startup menu.

You can print the resulting stem-leaf plot by choosing `Print a StataQuest log file` from the main menu. You will be shown a list of file names. Choose `hw21.log`. You will then see another menu which lets you print to “lpt1”, “lpt2”, view the file, etc. Choose the first item (printing to lpt1). The log file will be printed and you will be back to the main menu again. Choose `Quit` to get back to the DOS prompt. For your assignment, I want you to simply turn this plot in.

You may get a message like “I can’t find any log files in this directory”. If this is the case, select

| Change to a different directory → Change Drive → a: and then hw21.log will show up. |

Printing will vary from printer to printer. On my printer (an HP laserJet printer), when I follow these commands my “form feed” light comes on and I have to take the printer offline by hitting the “on line” key, hit the “form feed” button, and then put the printer “on line”.

If you are having trouble with my directions, get out of StataQuest entirely and then type

```bash
print hw21.log
```

This works fine on my printer. Other computers do not respond to this instruction, in which case you should try

```bash
copy hw21.log prn:
```

This is a pretty foolproof command for printing a file.
PROBLEM #2: On your floppy disk you will find three data sets, one called LDL.DTA, the other called LDLCHD.DTA and the third LDLNOCHD.DTA. These are data from a study undertaken by the National Heart, Lung & Blood Institute, where they measured whether one had coronary heart disease (CHD), one’s LDL, total serum cholesterol, and smoking status. CHD=0 for healthy, =1 for diseased. Smoking =0 for non-smoker, =1 for smoker. The logarithms of LDL and total cholesterol are called LOGLDL and LOGCHOL.

The data set LDL contains all the data, LDLCHD contains those who had CHD, and LDLNOCHD are those with no CHD.

I want you to use what techniques you have learned to see if the healthy individuals have “significantly” lower LDL than the sick ones, where “significantly” is something you have to define.

There are many places where StataQuest asks you to “Hit any key” to continue. Please go slowly through these instructions and look for this prompt. The answer is, of course, to hit any key! I will try to indicate most of these by the term HAK, but it is your responsibility to find them all.

To analyze these data, first get into StataQuest. Then do Files → Session logging. Type hw22 when prompted for a file name. You can get the healthy individual data set by HAK → Files → Open and then type

1dlnochd

when prompted for a file name, then HAK. You can get a stem-leaf plot of LOGLDL for healthy people by Graphs → One Variable → Stem and leaf. It will ask you for which variable to graph. Answer logldl. Then HAK.

You can get the stem-leaf plot for the sick people by Files → Open and answer “no” when asked if you want to save the data currently in memory. When prompted for a file name, type

1dlchd

Now do HAK → Graphs → One variable → Stem and leaf. Answer “logldl” (or its variable number) when asked for a variable name. Then HAK.

To get histograms, box plots, and summary statistics for both group, first do Files → Open. Answer “no” to a question about saving. Then type

1dl

to load the data for both groups. To get histograms, do HAK → Graphs → One Variable by groups → Histogram comparison. You will be prompted for name of a variable. Type

1ogld1
(You can type the variable number instead of its name.) You will then be asked for a variable identifying the groups. Type

```
chd 1
```

You will also be prompted for number of divisions. Answer 10. You will next be asked if you want a normal curve overlaid. Answer yes. The histograms will then be shown on the screen. Hit → to get back to StataQuest. You will then see a menu with three items:

- Show the graph again
- Save the graph for printing
- Continue

If the graph looks right, choose Save the graph for printing. You will be prompted for a file name. Type, say,

```
hw22g1
```

You will also be asked for a label of the graph. You can type in a label, say something like

```
log-ldl for CHD No-0 Yes-1
```

which will be printed at the bottom of the graph. The graph will be shown on screen again with the label added. HAK to get back to StataQuest.

To get box plots and summary statistics for later use in the course, Choose Graphs → One variable by groups → Box plot comparisons. You will be asked for the name of variable to graph and the variable identifying the groups. Answer the variable numbers for logdl and chd, respectively. Two box plots will be shown on the screen. You return to the menu by HAK. Save the graph to the file name hw22g2, and use the same title as before; the commands are exactly the same, but call the graph hw22g2.

If you don’t change the name of the graph each time, you’ll be overwriting the work you did before. StataQuest is careful to make sure that you realize you are doing this.

Now to get summary statistics, we’ll first do them for those without CHD. This is Files → Open. Answer “no” about saving the file, then “ldlnochd”, then HAK. Now hit Summaries → Data detail (median...) and select the right variable. Finish by HAK.

We’ll now do the summary statistics for those with CHD. This is Files → Open. Answer “no” about saving the file, then “ldlchd”, then HAK. Now hit Summaries → Data detail (median...) and select the right variable. Finish by HAK.

After finishing the above steps, you can quit StataQuest and get to the main menu by Files → Quit and answer no to prompt. There is an item Print a StataQuest graph in the main menu. You can print the graphs you saved (histograms and box plots) by choosing that item. A menu will pop up showing a list of file names. Choose HW22G1.GPH. You will then see a menu that lets you choose the printer. If you are printing to a laser printer in one of the computing centers on
campus, choose the one that lets you print to a HP LaserJet printer. Note that you have to [HAK]
to get back to the main menu after each graph is printed.

From the main menu, do [Print a StataQuest log file] → [hw22.log]. Select lpt1 as printer. Then
choose [Quit] to get back to DOS.

Your assignment is to answer the question I have posed. In order to get credit on the assignment,
place the two stem and leaf plots on the same page, the two histograms on the same page, and use
≤ 1 page for your writeup.

Naturally, I would like your answer to be semi–coherent, so organize it as follows. First define the two
populations precisely. Then point out any unusual features gleamed by the stem–leaf plot, e.g., the
percentage whose logldl is above 5.0 in each of the two data sets. Then compare the two histograms
on a qualitative basis, giving me some idea you know where the “center” of each population is.
The purposes of this assignment are as follows:

- Learn how to use StataQuest to compute summary statistics.
- Develop a feel for the differences between measures of the center and measures of variability.

You should come out of this assignment with an understanding that summary measures are not enough to make precise decisions. Suggestions on how to run StataQuest are given at the end of this assignment. As usual, I will go through an example in class, and you may wish to hold off starting the computing parts of this assignment until then.

Please keep an extra copy of all work on this and all other assignments, as they will come in handy for future assignments.

**Problem #1:** The Laboratory for Biostatistics and Biosystems Modeling here at A&M is involved in a program to treat patients who are inclined towards suicide. The patient population consists of active military at Ft. Hood who have either attempted suicide or expressed an attempt to commit suicide (ideation). The study is funded by the National Institute of Mental Health and run here at A&M by Professor Fred Dahm. It is an attempt to determine whether the behaviors of a **CONTROL** group (those who are given standard care) can be improved by a more aggressive **TREATMENT** program. The point is, how can we answer such questions using good statistical principles? The data set that you will use are actual data values from the clinical trial.

The best device for testing whether a **TREATMENT** is better than a **CONTROL** is to take a patient base and randomly assign some of the patients to the **CONTROL** group, and others to the **TREATMENT** group. This is being done in the Ft. Hood study, although by design more patients are being assigned to the **TREATMENT** group than to the **CONTROL** group. Having randomly assigned patients to one of the groups, they will then be treated and followed for 2 years. At the end of the 2 years, their status will be assessed.

The reasoning behind randomized clinical trials is that

- The randomization insures that there should be no large differences between the **CONTROL** and **TREATMENT** groups at the start of the study; and hence
- Any differences we see between the two groups after two years is due to the different modes of treatment.

The logic here is perfectly sound, as long as the randomization “worked”, i.e., the **TREATMENT** and **CONTROL** groups do not differ significantly on any variables measured at the start of the study.

The following variables were measured at the beginning of the study (known as the baseline):

- **TREATMENT** treatment assignment, 1==**TREATMENT** group, 0==**CONTROL** group.
- **AGE** age of the patient.
- **SEX** sex of the patient, 0==female, 1==male.
- **MARRIED** marital status, 1==married, 0==other.
- **ATTEMPT** entry criterion, 1==have made a previous suicide attempt, 0==ideation.
- **PSISCORE** score for the problem solving inventory, a psychological test.
LESSCORE score on the life experiences survey.
HOPELESS Score on the Beck depression inventory.
SUICIDE Suicide probability scale.
IQ Standard IQ score.

The data for the entire sample are in your diskette as the file SUICIDE.DTA. The CONTROL group is SUICONT.DTA, while the TREATMENT group is SUICASE.DTA.

YOUR ASSIGNMENT: The goal of this assignment is to see whether the TREATMENT and the CONTROL groups differ in any significant way in terms of age, IQ, or the 4 psychological testing scores. If this is not the case, then the study can proceed, but if there are differences, we may need to change the study. Hence, if you find differences, we’ve got problems. I want you to focus on the variable IQ, to see if there are any noticeable differences in the two groups.

You should turn in the following material:

• A ≤ 1 page maximum summary of what you found.
• A single page containing histograms of IQ by treatment assignment.
• A single page containing stem-leaf plots of IQ by treatment assignment.
• A single page with the following tabulated by treatment assignment group: mean, median, 25th percentile, 75th percentile, interquartile range, standard deviation, and median absolute deviation from the median (MAD).

PROBLEM #2: Compute the same statistics as in Problem #1 for problem 3.2 on page 75 in the text, only do so by using a calculator. Remember, you may have to reproduce these calculations on your exams. Just as important, I want you to develop a feeling for these summary statistics.

PROBLEM #3: The average playing time of a collection of records is 35 minutes with a standard deviation of 5 minutes. Assuming that the distribution of playing times is normal, approximately what percentage of times are between 25 and 45 minutes? Less than 20 minutes or greater than 20 minutes?

PROBLEM #4: In a study to investigate the effect of car speed on accident severity, a study was done of fatal car accidents. The sample mean speed was 42 miles per hour, with sample standard deviation of 15 MPH. Assuming the data are approximately normally distributed, what percentage of fatal arguments are over 57 MPH?

PROBLEM #5: Interpret what it means when a person says that they scored at the 83rd percentile on the SAT Verbal test, and at the 92nd percentile on the SAT Math test.

Studying for Exams: The following questions are not homework problems. They can be used
to help you study for exams. The graders will provide answers to these questions. They all concern
the empirical rule.

**Study #1:** Suppose a cheap scale is claimed to be accurate for bags up to 100 pounds. You decide
to test the scale by getting a 50 pound bag (measured by the National Institute of Standard and
Technology so that it really does weigh 50 pounds), and you take a sample and find that the mean
in the sample is 49.5 with a standard deviation of .20. Assuming normality, do you think that the
scale is accurate? Why or why not? Use the empirical rule.

**Study #2:** Five randomly selected rats were treated with an injection of HRP (horseradish perox-
idase). The total number of injured neurons in the 4th nerve nucleus was recorded: the data were
209, 187, 123, 184, 194. Compute the sample mean. Suppose the population standard deviation 
\( \sigma = 20 \). Use the empirical rule to determine if the data show conclusively that the population mean
is different from 200.

**Study #3:** Suppose that an experienced smallmouth bass fisherperson using dry flies catches an
average of 22 bass per day, with a standard deviation of 8 bass. Use the empirical rule to determine
the percentage of times an experience fisherperson will catch between 6 and 38 bass per day.

**StataQuest Directions**

The following instructions will enable you to work the homework assignment. I am assuming
you now know how to get into StataQuest, and that you are at the StataQuest menu.

Once again, there are many places in a StataQuest session where you are asked to “hit any
key”. I have tried to indicate as many of these as possible with the notation \[HAK\], but this is the
last assignment I will do so. Practice going through the session carefully since in the future you’ll
be expected to look for these prompts.

It’s also important to remember that if all else fails, you can hit the escape key a few times
and get back into the main menu.

First start StataQuest and do \[Files \rightarrow Session logging\] Then type

\[hw31, \rightarrow\]

You’ll see the prompt \[HAK\].

\[\star\] To get the stem and leaf for the controls:

Do \[Files \rightarrow Open\] and type

\[suicont \rightarrow\]

Do \[HAK \rightarrow Graphs \rightarrow One variable \rightarrow Stem and leaf\] and answer iq when asked for variable
name. At end, \[HAK\].

\[\star\] To get stem and leaf for the treated group:
Do **Files → Open** and answer **no** to clear data from memory. When asked for file name, type **suicide**. 

Do **HAK → Graphs → One variable → Stem and leaf** and answer **iq** when asked for variable name. At end, **HAK**.

To get histograms for both groups:

Do **Files → Open** and answer **no** to clear data from memory. When asked for file name, type **suicide**. 

Do **HAK → Graphs → One variable by groups → Histogram comparisons**. Answer **iq** when asked for variable to graph. Answer **treatment** when asked for variable identifying groups. Answer 10 when asked for number of divisions. Answer **yes** when asked if you want normal curves overlaid.

The graph will appear, and then you **HAK**. Then do **Save graph for printing**.

Save the graph to **hw3t1g1**. When it asks for a title, use **IQ’s Control-0 Treat-1**. At the end, **HAK**

To get the basic summary statistics for the controls:

Do **Files → Open** and answer **no** to clear data in memory, then answer **suicont** as file name to load. Then do **HAK → Summaries → Data detail (medians...)**. Answer **iq** when asked for variable(s) to summarize. At end, **HAK**

To get the basic summary statistics for the treated group:

Do **Files → Open** and answer **no** to clear data in memory, then answer **suicase** as file name to load. Note that it asks you to **HAK**. Then do **Summaries → Data detail (medians...)**. Answer **iq** when asked for variable(s) to summarize. At end, **HAK**

To compute the MAD for the controls, you first have to look up the median, namely 98.

Do **Files → Open** and answer **no** to clear data in memory, then answer **suicont** as file name to load. There will be a prompt for **HAK**. Do **Edit → Spreadsheet** to get into the spreadsheet. Press **F10** to get the menu. Do **Add → Variable** and then answer **medd** when asked for the name of the new variable. Then do **F10 → Replace → Formula**. Type

```
abs(iq-98) ←1
```

when asked for a formula for **medd**. Then do **F10 → Files → Save** and answer **yes** twice to overwrite the data file. This constructs the variable **medd** = |X - median|. Again, you’ll be prompted with **HAK**. Quit the spreadsheet by **F10 → Files → Quit editor**. To find the median
of the variable \texttt{medd}, i.e., the MAD, do \texttt{Summaries} \rightarrow \texttt{Data detail (medians...)} Answer the variable \texttt{medd}. At the end, \texttt{HAK}.

\begin{itemize}
\item To compute the MAD for the treated group, you first have to look up the median, namely 99.
\item Do \texttt{Files} \rightarrow \texttt{Open} and answer \texttt{no} to clear data in memory, then answer \texttt{suicide} as file name to load. There will be a prompt for \texttt{HAK}. Do \texttt{Edit} \rightarrow \texttt{Spreadsheet} to get into the spreadsheet. Press \texttt{F10} to get the menu. Do \texttt{Add} \rightarrow \texttt{Variable} and then answer \texttt{medd} when asked for the name of the new variable. Then do \texttt{F10} \rightarrow \texttt{Replace} \rightarrow \texttt{Formula}. Type
\begin{verbatim}
abs(iq-99) <-
\end{verbatim}
when asked for a formula for \texttt{medd}. Then do \texttt{F10} \rightarrow \texttt{Files} \rightarrow \texttt{Save} and answer \texttt{yes} twice to overwrite the data file. This constructs the variable \texttt{medd} = |X - median|. Again, you’ll be prompted with \texttt{HAK}. Quit the spreadsheet by \texttt{F10} \rightarrow \texttt{Files} \rightarrow \texttt{Quit editor}. To find the median of the variable \texttt{medd}, i.e., the MAD, do \texttt{Summaries} \rightarrow \texttt{Data detail (medians...)} Answer the variable \texttt{medd}. At the end, \texttt{HAK}.
\item Now leave StataQuest and print the log file by \texttt{Files} \rightarrow \texttt{Quit} and answer \texttt{no} to the prompt. Then choose \texttt{Printing a StataQuest log file}, choose the file name \texttt{hw31.log}, and \texttt{lpt1} as printer. See Homework #2 for some details about printing.
\item Now you have to print your graph. Do \texttt{Print a StataQuest graph} \rightarrow \texttt{hw31g1.gph} and then select your printer and hit \texttt{-.} After it prints, you should \texttt{HAK} (it does not give you a prompt, but remember that if all else fails, do \texttt{Escape}). Leave StataQuest by choosing \texttt{Quit} from the main menu.
\end{itemize}
PROBLEM #1: Consider the Suicide data from the previous homework. You are allowed no more than two full pages for this problem, including all graphs. Using StataQuest, construct box plots of IQ by treatment assignment. This can be done as follows. Start StataQuest and get to the StataQuest menu.

1. Do **Files** → **Open** and type **suicide** when asked for a file name to load. Remember to **HAK**.

2. To get box plots, do **Graphs** → **One variable by groups** → **Box plot comparisons**. Answer **iq** when asked for variable to graph, and **treatment** when asked for variable to identify groups. The graph will show up. Hit **<**. Then **Save graph for printing** and respond with **hw41g1**. Then label the graph IQ’s **Control**–0 **Treat**–1. Remember to **HAK**.

3. Quit StataQuest by **Files** → **Quit** and answer **no** to prompt. Then choose **Print a StataQuest graph** from the main menu, choose **hw41g1.gph** as the file to print, and select the appropriate printer. Remember that after it prints, you have to **HAK** to get back to the main menu.

4. Choose **Quit** from the main menu to get back to DOS.

Once again you may get a message like “I can’t find any (whatever) files in this directory”. If this is the case, select

- **Change to a different directory** → **Change Drive** → **a:** and then **hw41g1.GPH** will show up.

---

**Do not try to compute box plots on each data set individually**, because you will then have the unequal scale problem I have talked about in class for histograms.

(a) Display the plots in your homework using cut and paste.

(b) Are the two groups basically the same in terms of their centers? Why or why not? Use all the pictures from this and the previous assignment.

(c) Now answer the previous problem using the empirical rule. You have already computed the sample means and standard deviations of IQ in the previous assignment. We know that the empirical rule for the difference in population means is that this difference should be within 2 sample standard deviations of the difference in sample means, where the sample standard deviation is \( \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^{1/2} \). If there really is no difference in the population means, then their difference is the number 0, and so the number 0 should be within 2 sample standard deviations of the difference in sample means. Is it?

(d) Are the two groups basically the same in terms of their variability? Why or why not? Use all the pictures from this and the previous assignment.

**PROBLEM #2:** Do Problem 3.47 in the text (page 101) by hand. Replace question (b) by

- **Calculate the median**, the MAD, the quartiles and the interquartile range.
**PROBLEM #3:** The 12/15/91 article from the New York Times magazine on cancer statistics (if you can find it) will give you some idea of how the same statistics can be interpreted by different people in different ways, depending on their prior beliefs. It is a common problem that researchers cite only the statistics that back up their arguments, and ignore the ones that don’t (most of you probably do this in your political discussions!).

In any event, everyone agrees that cancer incidence and mortality is up, and this is particularly so with breast and skin cancer. A key factor is: are there more cancers, or are we seeing the effects of better and hence earlier diagnosis? I want you to use your knowledge of summary statistics to show why this might be an area of controversy, as follows.

Suppose that we studied 100 people in 1960, and a separate 100 people in 1990. In each case, unbeknownst to us the populations consisted of 50 people who were free from cancer, 25 people who had the disease but for less than 6 months, and 25 people who had had the disease for more than 6 months. Thus, in actuality, 50% of this hypothetical population actually has cancer. Let’s suppose that in 1960, 20% of those who had cancer for less than 6 months were diagnosed correctly, while 60% of those who had it for more than 6 months were diagnosed correctly. The corresponding percentages for the 1990 population were 60% and 80%. When cancer statistics are reported, what is the reported percentage of cancer cases in 1960 and in 1990 in the whole population of 100 people?
**Problem #1:** The Skeena River sockeye salmon data set is called SKEENA.DTA on your StataQuest disk. Here, \( X = \) spawnerx is the number of mature fish who spawn (and then die, of course), while \( Y = \) recruit is the number of fish which are born and recruited into the fishery. Use StataQuest to construct (and turn in) a plot of the data. Draw in the least squares line, labeling the line in the form \( Y = a + bX \) (StataQuest can do this for you) with the least squares fit to the data.

Here are the StataQuest instructions that will enable you to work this problem. First start StataQuest and get to the StataQuest menu.

1. Do **Files** → **Session logging**, using hw51 as file name. **HAK**
2. Do **Files** → **Open**, using skeena as file name. Then **HAK**
3. Do **Summaries** → **Means and SDs**, hit ← when asked for variables to summarize (which means summarize all variables in the data set). Then **HAK**
4. Do **Statistics** → **Simple regression**, Answer recruit when asked for the \( Y \) variable. Answer spawnerx when asked for the \( X \) variable. The regression summaries will be shown. After you **HAK**, a menu will pop up. Choose **Plot fitted model** Hit ← twice to answer no to the prompts. Hit ← again and the graph will be shown. Hit ← again and another menu will pop up. Choose **Save graph for printing**, use hw51g1 as file name and Skeena River Sockeye Salmon as label. The graph will be shown again. Hit ← again and choose **Continue** from the menu.
5. Quit StataQuest by **Files** → **Quit**, and answer no to prompt. From the menu, do **Print a StataQuest graph** → hw51g1.gph and select the appropriate printer. Remember **HAK** to get back to main menu after printing.
6. Choose **Print a StataQuest log file** → hw51.log Choose the system printer(lpt1) to print the file.
7. Choose **Quit** from the main menu to get back to DOS.

Once again you may get a message like "I can't find any (whatever) files in this directory". If this is the case, select

**Change to a different directory** → **Change Drive** → a: and then hw51g1.GPH will show up.

**Problem #2:** In problem #1, I computed the following numbers:

\[
\sum_{i=1}^{n} Y_iX_i = 28 \times 787910.107; \quad \sum_{i=1}^{n} X_i^2 = 28 \times 378075.107; \\
\bar{Y} = 1239.107; \quad \bar{X} = 557.107.
\]
If I did this correctly, you should be able to reproduce the least square line computed by StataQuest from these numbers. Did I do it correctly?

**PROBLEM #3:** In the Skeena River data set, give the predicted number of fish recruited into the fishery when there are 1000 spawners.

**PROBLEM #4:** Use the empirical rule of class to give a range for the population slope in this problem.

**PROBLEM #5:** In the Skeena River data set, the main goal is to try to predict next year's recruits after having observed the number of spawners. The empirical rule for predicting a new observation works like this. If we want to predict at $X^* = 1000$ spawners, the prediction is $a + bX^*$ with a range of plus or minus $2s_e$, where $s_e^2$ can be found in the analysis of variance table as the mean squared error. Looking at class notes for the stenosis example, for normal kids $s_e^2 = .155$, while for stenotic kids $s_e^2 = .326$. Give a range for the number of recruits if there are 1000 spawners.
1. Let $Z$ be the standard normal distribution, i.e., have mean zero and standard deviation 1. Compute
   
   (a) $\Pr(0 \leq Z \leq 2.2)$
   (b) $\Pr(-0.5 \leq Z \leq 2.2)$
   (c) $\Pr(Z \leq 2.2)$
   (d) $\Pr(1.5 \geq Z)$

2. The distribution of the lengths of pregnancies is approximately normally distributed with population mean 266 and population standard deviation 16.
   
   (a) What is the chance that a pregnancy lasts between 250 and 300 days?
   (b) What is the chance that a pregnancy lasts at most 240 days?
   (c) What is the chance that a pregnancy lasts at least 310 days?
   (d) Some insurance companies will pay medical expenses associated with childbirth only if the insurance has been in effect for more than 275 days (9 months). Suppose conception occurs two weeks after coverage begins. What is the chance that the insurance company will pay the medical expenses?

3. Suppose that the height of adult women has a mean of 66 inches and a population standard deviation of 2 inches.
   
   (a) What percentage of women are shorter than 5 foot 7 inches?
   (b) What percentage of women exceed 5 foot 8.5 inches?

4. If grades on a test are normally distributed with population mean 78 and population standard deviation 7, and the top 15% get A’s, do you get an A if your exam score is 89?

5. My old UNC exams, problem #1.
7. My old UNC exams, problem #17.
(1) My old UNC exams, problem #4.
(2) My old UNC exams, problem #13, part(a) only.
(3) My old UNC exams, problem #19.
(4) The following data are from a physiology lab experiment performed by a student in the spring 1992 STAT302 class. The response is the reaction time to stimulus in 10\textsuperscript{th}s of a second. For males, the auditory reaction times are
\[
2.35, 2.19, 2.57, 2.67, 2.11, 1.71, 2.23, 2.92, 2.81, 2.04, 2.02, 3.95
\]
while for females the times are
\[
1.67, 2.16, 1.96, 1.68, 1.92, 2.11, 2.23, 2.63
\]
The visual reaction times for males are
\[
2.27, 2.05, 2.10, 2.62, 3.10, 2.05, 2.27, 2.56, 2.38, 1.97, 2.33, 2.50
\]
while for females they are
\[
2.54, 2.04, 2.17, 3.24, 1.93, 1.92, 3.29, 3.26
\]
The interesting part about this example is that there appears to be an outlier in the auditory reaction times for males. We’ll try to follow these data throughout the semester to see how various techniques are affected by outliers.

(a) Compute the sample mean and standard deviation for the auditory response times for males.
(b) Pretend that the sample mean and standard deviation equals the population mean and standard deviation, and assume that the response times are normally distributed. Compute the probability that an individual response time is more than 3.50.
(c) Repeat (a), but delete the largest response time.
(d) Repeat (b) using the results you found in (c).
(e) Turn in your output.

The complete data are already in your StataQuest file called FULLPHSI.DTA, while the data with the outlier deleted is in the StataQuest file called REDPHSI.DTA. The variables are sex (0=females, 1=males), exper (0=auditory, 1=visual), time (response time) and logtime (the logarithm of response time). You can do this assignment with the following StataQuest instructions:
1. Do Files → Session logging  Use hw71 when prompted for file name.
2. Do Files → Open  Type the file name fullphsi.
3. Next you need to go to the command mode: Files → Command mode  You should be at the “dot prompt”.
4. Type
\[
\text{sort sex exper by sex exper: summ time,detail}
\]
**Note 1:** Type the command exactly as shown above. Don’t leave out or add any spaces.

**Note 2:** After you type in the second command above, the output will be several screens long. At each screen, the line “--more--” will appear at the bottom of the screen. Just hit the space bar to see the next screen.

5. Type

```plaintext
use redphi
sort sex exper
by sex exper: summ time,detail
exit
```

6. Now you should be in the main menu. Choose `Print a StataQuest log file`, select the file name `hw71.log` and `lpt1` as printer. Choose `Quit` to get back to DOS.

Here is how I constructed the data (you don’t have to do this, but it might come in handy when you do your project).

1. Start StataQuest and do `Edit → Spreadsheet`. A message will pop up telling you to load some data. Hit `→`. Choose `New` from the `Files` menu. Answer 40 for number of observations and 3 for number of variables.

2. Using the sequence `F10→Replace→Rename variable` to change the column headings to `sex`, `exper`, and `time`.

3. Type in the full data.

4. Do `F10→Add→Variable`, answer `logtime` to the prompt.

5. Do `F10→Replace→Formula` Type

   ```plaintext
   log(time)
   ```

6. Save the data by `F10→Files→Save` when prompted for a file name, use `fullphi`.

7. Quit the spreadsheet by `F10→Files→Quit editor`.

8. Get into the command mode by `Files→Command mode`.

9. Type

   ```plaintext
   sort sex exper
   ```

   (This command sorts the data in terms of sex and type of experiment.)

10. Go to the menu and save the data by `F10→Files→Save` Answer yes twice to the prompt.

This constructs the full data set. To construct the reduced data set, go into the spreadsheet and delete the case corresponding to the outlier. Suppose the case is number 12:

1. From the StataQuest menu, do `Files→Open` Answer no if prompted about saving file. Type the file name `fullphi`.

2. Do `Edit→Spreadsheet` You should see the full data set in the spreadsheet.

3. Move down to the 12th row and do `Drop→Observation` Hit the return at the prompt which asks if you really want to delete the observation.
4. Save the data by F10→Files→Save Answer no. Answer redphsi when prompted for file name.

5. Quit spreadsheet by pressing Escape (maybe twice). Get into the command mode by Files→Command mode.

6. Type
   \[ \text{sort sex exper} \]

7. Go to the menu and save the data by F10→Files→Save Answer yes to the prompt.

You can read off the output to get the means and standard deviations which you need. You should find that the deletion of a single point out of 12 has a large impact on the sample standard deviation.
1. My old UNC exams, problem #3.
3. You have been called in to design a sampling plan for the quality control of the production of boots by Carroll Brothers Boot Shop, whose motto is “We skin ’em, you wear ’em, no sissies welcome”. They produce boots in lots of size 10,000, and they are willing to ship the lot if 6% or less of the boots are defective. Note that in this case, the POPULATION is the current set of 10,000 boots. The population percentage \( \pi \) is the percentage of defective boots in the population. The population is labelled “bad” if \( \pi > 0.06 \).

You recommend a sampling plan in which 500 boots are selected; brother Joe and his son Brent try them out, and if more than 20 are defective, the shipment is inspected item by item; otherwise, you ship. Note that in this case, the sample is 500 boots out of the population of 10,000 boots. The percentage of defective boots in the sample is called \( p \). We are concerned about whether \( p > 20/500 \), and how often that occurs.

What is the chance that a shipment with 6% defectiveness is shipped, i.e., compute the probability that \( p \) exceeds 20/500 if the shipment has 6% defectives.

4. Suppose that we are going to decide if a set of records has to be completely audited. A good set of records has 5% error. The records are fully audited if 10% or more of them from a sample are found to be in error. A sample size of \( n = 200 \) is taken. What is the chance of fully auditing a set of good records?

Note: the wording here is tricky but logical. The population is the set of all the records, and \( \pi \) is the percentage of all the records which are in error. A good set of records has \( \pi = 0.05 \). I’m taking a sample of size 200, and observing the sample percentage \( p \) of records which are in error.
HOMEWORK #9, due Lecture #16

1. In the Skeena River sockeye salmon data set you have, the StataQuest data set is called SKEENA.DTA and the number of new recruits into the fishery is call recruity. Assuming the number of recruits per year is a random sample of normally distributed observations, compute a 90% confidence interval for the true mean number of recruits per year. Here are the StataQuest instructions. First start StataQuest and get to the StataQuest menu.

(a) Do Files → Session logging use hw91 as file name.
(b) Do Files → Open answer skeena when asked for a file name.
(c) Do Summaries → Confidence intervals Answer 90 for confidence level, and type recruity when asked for variable to be summarized.
(d) Do Files → Quit answer no to the prompt.
(e) From the main menu, choose Print a StataQuest log file Choose hw91.log and lpt1 as printer.
(f) Choose Quit from the main menu to get back to DOS.

Why is answering this question basically irrelevant for managing a fishery?

2. Consider the physiology experiment from Homework #2. Deleting the outlier, construct a 99% confidence interval for the true mean auditory reaction times for males. Here are the StataQuest commands. First remember that all the data are in the data set you called aud, while the males are in audm and the females are in audf. Presumably, the outlier is observation #12 for both aud and audm.

This problems requires that you have done Homework #2, since it is in that assignment that you constructed the data set.

(a) Do Files → Session logging Use the file name hw92.
(b) Do Files → Open and answer aud when asked for a file name.
(c) Do Edit → Spreadsheet to get into the spreadsheet.
(d) Move to the 12th row, hit F10, and do Drop → Observation Hit ← when asked to confirm the deletion.
(e) Do F10 → Files → Save Answer no and use the file name redaud.
(f) Do F10 → Files → Open Answer no and use the file name audm.
(g) Move to the 12th row and do F10 → Drop → Observation Hit ← when asked to confirm the deletion.
(h) Do F10 → Files → Save Answer no to the prompt and use the file name redaudm.
(i) Hit Escape (maybe twice) to quit the spreadsheet. Do Summaries → Confidence intervals answer 99 for confidence level and time for variable.
(j) Do Files → Quit Answer no.
(k) From the main menu, choose **Print a StataQuest log file**. Choose hw92.1og and lpt1 as printer. Choose **Quit** from main menu to get back to DOS.

3. You have been called in to design a sampling plan for the quality control of the production of boots by Carroll Brothers Boot Shop, whose motto is “We skin ’em, you wear ’em, no sissies welcome”. They produce boots in lots of size 10,000, and they are willing to ship the lot if 6% or less of the boots are defective. You recommend a sampling plan in which 500 boots are selected. Suppose that in a sample, they observe 25 defectives. Find a 95% confidence interval for the true percentage of defectives.

4. Suppose I sample 21 normally distributed variables and observe a sample mean of 13.3 and a sample standard deviation of 4.4. Construct a 95% confidence interval for the population mean.

5. Consider the following data which are the weight losses by people enrolled in an experimental diet program.

   6 12 8 10

   Assume that weight losses are normally distributed, but you know neither the mean nor the standard deviation in the population.

   (a) What is the sample mean of these numbers?
   (b) What is the sample standard deviation of these numbers?
   (c) Pretend that the sample mean is 6.0 and the sample standard deviation is 2.0. These may be wrong but pretend they are correct! Construct a 99% confidence interval for the mean weight loss.

6. Carroll Sisters Accountants (motto: no boots allowed) is based in Dallas, Texas. They are auditing a rather large firm for its sales, and they look at the error

   \[ X = \text{reported sale} - \text{audited sale} \]

   They believe that these errors are normally distributed with mean $0 and standard deviation $120.

   (a) They take a sample of size \( N = 400 \). What is the probability that the sample mean error exceeds $9.00?
   (b) If the sample mean is actually $9.00, find a 99% confidence interval for the true mean.

7. My old UNC exams, problem #15(b).
1. In the physiology data set, consider the auditory reaction times for males. Delete the outlier and test whether the population mean equals 2.8. Use 90%, 95%, and 99% confidence intervals to make this test.

You can use the information from an earlier assignment, or the following StataQuest commands. First start StataQuest and get to the StataQuest menu.

(a) Do Files → Session logging
Use the file name hw101.

(b) Do Files → Open
Answer redphsi when prompted for a file name.

(c) Do Edit → Spreadsheet
Move to a cell with sex=0 (female). Do F10→ Drop → Observations
Hit $\rightarrow$ to confirm the deletion. This will delete all the female data.

(d) Now delete the visual reaction data for males. Do this by moving the cursor to a cell with exper=1. Then do F10→ Drop → Observations
Hit $\rightarrow$ to confirm the deletion. Now the data contains only auditory times for males.

(e) Hit Escape to quit the spreadsheet. Do Summaries → Confidence intervals
Answer 90 for confidence level and time for variable(s) to summarize.

(f) Repeat the preceding step, using 95 and 99 for levels of confidence instead.

(g) Do Files → Quit
Answer no to the prompt. From the main menu, choose Print a StataQuest log file
Choose hw101.1og and lpt1 as printer. Choose Quit from main menu to get back to DOS.

2. In the Skeena River sockeye salmon fishery, it has been hypothesized that the mean number of spawners is 1,000. Is this reasonable if you use an $\alpha = 0.05$ test? As usual, you should do a confidence interval.

Here are the StataQuest commands. Start StataQuest and get to the StataQuest menu.

(a) Do Files → Session logging
Use the file name hw102.

(b) Do Files → Open
Use the file name skeena.

(c) Do Summaries → Confidence intervals
Answer 95 for confidence level and spawnerx for variable name.

(d) Do Files → Quit and answer no to the prompt.

(e) Choose Print a StataQuest log file from the main menu. Choose the file hw102.1og and lpt1 as printer. Choose Quit from the main menu to get back to DOS.

3. Scientists think that robots will play a crucial role in factories in the next 20 years. Consider an experiment to determine whether robots should be used to weave computer cables. It is known that the population of humans make defective cables 3% of the time. You use a robot to assemble a sample of 300 cables, and find that 14 of them are defective. Does this support the hypothesis that the proportion of defectives is the same for robots as it is for humans, or different? Use confidence intervals to answer the question. Use $\alpha = 0.01$. 

4. You are working as a quality control supervisor for Carroll Brothers Boots of Wichita Falls. You are in charge of the size nine men’s boots. No boot can be made to be exactly size nine, and as part of your job you measure the difference between the actual size and size nine. When the process is in control, this difference is a normally distributed random variable with mean 0mm and a standard deviation of 3mm. You decide that each hour you will take a sample of 16 boots and measure their sample mean. Your decision rule is to shut down the operation and call it out of control if the sample mean you observe exceeds 0.75mm.

(a) Assume the process is in control and the true mean is really 0mm. Having taken your quality assurance sample of size 16, what is the probability that you will shut the operation down?

(b) Now suppose that the true mean has shifted to 2mm. What is the probability that your decision rule will shut the operation down?

5. Problem 8.77(a) of the text, page 349. Change the phrase “does exceed” to “is different from”. Use confidence intervals to answer this question. Skip p-values questions.
Problem #1: We have previously discussed the result from a physiology lab experiment performed by a student in the spring 1992 STAT302 class. The response is the reaction time to stimulus in $10^4$s of a second. The auditory data set is given in aud and the reduced data set is given in redaud. The interesting part about this example is that there appears to be an outlier in the auditory reaction times for males.

(a) Find a 90% confidence interval for the difference in the mean auditory reaction times for males and females using all the data.

(b) Does the result in (a) give evidence that males and females have different reaction times?

(c) Repeat (a) but remove the outlier for the males.

(d) Repeat (b) but remove the outlier for males.

(e) How do these results compare to the informal results you see in box plots? (obviously, you should attach the plots)

Here are the StataQuest commands that are required. As usual, start up StataQuest and get to the StataQuest menu.

1. Do Files → Session logging. Use the file name hw111.
2. Do Files → Open. Use the file name aud.
3. Do Statistics → Parametric Tests → 2-sample t-test. You will see three choices. Enter 1. Answer time when asked for variable to be tested. Answer sex when asked for variable identifying the two groups. Answer yes when asked if assume equality of variance. Answer 90 when asked for confidence level. The result will then be shown on the screen.
4. Do Graphs → One variable by groups → Box plot comparisons. Answer time when asked for variable to graph. Answer sex when asked for variable identifying groups. After viewing the graph, do HAK and choose Save graph for printing and use hw111g1 as file name and Aud time Male-0 Female-1 as label.
5. Do Files → Open. Answer no to the prompt and use reduced when asked for file name to load.
6. Do Statistics → Parametric Tests → 2-sample t-test. You will see three choices. Enter 1. Answer time when asked for variable to be tested. Answer sex when asked for variable identifying the two groups. Answer yes when asked if assume equality of variance. Answer 90 when asked for confidence level. The result will then be shown on the screen.
7. Do Graphs → One variable by groups → Box plot comparisons. Answer time when asked for variable to graph. Answer sex when asked for variable identifying groups. After viewing the graph, do HAK and choose Save graph for printing and use hw111g2 as file name and Aud time Male-0 Female-1 reduced as label.
8. Do Files → Quit and answer no to prompt.
9. From the main menu, choose Print a StataQuest graph. Choose the file hw111g1.gph and the appropriate printer. Remember to HAK. Repeat the step to print the file hw111g2.gph.
10. From the main menu, choose Print a StataQuest log file. Choose the file hw111.log and lpt1 as printer. Choose Quit from main menu to get back to DOS.
Problem #2: Consider the data from problem 9.43 of the text, page 391.
(a) Find a 95% confidence interval for the difference in the two proportions.
(b) Test with a type I error rate of 5%, whether the population percentage of major malfunctions is the same or not.

This assignment concerns the response time data we have been using in class. For both males and females, let \( X = \log(\text{visual response time}) \), and let \( Y = \log(\text{auditory response time}) \).

**Problem #1:** Among the males, using a 95% confidence interval, is there evidence of a relationship between the two variables? Does your formal statistical analysis agree with a plot of the data? (attach the plot)

I have placed the full data into a data set called `audvreg.dta`, with the males in `audvregm` and the females in `audvregf.dta`. To do this assignment, here are the StataQuest instructions. Start StataQuest and get to the StataQuest menu as usual.

1. Do **Files** → **Session logging**. Use the file name hw121.
2. Do **Files** → **Open**. Answer `audvregm` when asked for the file name to load.
3. Do **Statistics** → **Correlation** → **Regular (Pearson)**. When asked for variables, type `loga logv` ←
4. Do **Statistics** → **Simple regression**. Answer `loga` when asked for the \( Y \) variable. Answer `logv` when asked for the \( X \) variable.

The regression summaries will be shown. After you [HAK] a menu will pop up. Choose [Plot fitted model] Hit ← again to answer no to the first prompt, then yes to the second. The graph will be shown. Hit ← again and another menu will pop up. Choose [Save graph for printing] use hw121g1 as file name and log time for male Y and X-vis as label. The graph will be shown on screen again. Hit ← again and choose [Continue] from the menu.

5. Quit StataQuest by **Files** → **Quit** , and answer no to prompt. Choose [Print a StataQuest graph] from the main menu and [Print a StataQuest log file] choose hw1211g1.gph. Select the appropriate printer. Also Choose from the main menu. Choose the file hw121, log and lpt1 as printer. Choose [Quit] to get back to DOS.

**PROBLEM #2:** Redo the previous problem, but delete the outlying observation (data point #5). To do this in StataQuest, do the following (first start StataQuest and get to the StataQuest menu):

1. Do **Files** → **Session logging** Use the file name hw122.
2. Do **Files** → **Open**. Use the file name `audvregm`.
3. Do **Edit** → **Spreadsheet**
4. Move down to the fifth row and do F10→ **Drop** → **Observation**. Hit ← when prompted for confirmation.
5. Hit Escape to quit the spreadsheet.
6. Do **Statistics** → **Correlation** → **Regular (Pearson)** when asked for variable names, type `loga logv` ←
7. Do Statistics \rightarrow Simple regression \rightarrow. Answer \textit{loga} when asked for the Y variable. Answer \textit{logy} when asked for the X variable.
   The regression summaries will be shown. After you H
   A K a menu will pop up. Choose \textit{Plot fitted model} Hit \rightarrow to answer no to the first prompt, then yes to the second. The graph will be shown. Hit \rightarrow again and another menu will pop up. Choose \textit{Save graph for printing}, use \textit{hw122g1} as file name and \textit{log-time for male Y-aud X-vis reduced} as label. The graph will be shown again. Hit \rightarrow again and choose \textit{Continue} from the menu.
8. Quit StataQuest by Files \rightarrow Quit, and answer no to prompt. Choose \textit{Print a StataQuest graph} from the main menu and choose the file \textit{hw122g1.gph}. Select the appropriate printer. Also Choose \textit{Print a StataQuest log file} from the main menu. Choose the file \textit{hw122.log} and \textit{lpt1} as printer. Choose \textit{Quit} to get back to DOS.

Problem \#3: If you did things correctly, you should have noticed a large difference in your analyses, based on whether the outlier was deleted or not. Which analysis do you believe, and why?

Problem \#4: Consider the response data we have been using in class. For both males and females, let \( X = \log(\text{visual response time}) \), and let \( Y = \log(\text{auditory response time}) \). Among the females, using a 95\% confidence interval, is there evidence of a relationship between the two variables? Does your formal statistical analysis agree with a plot of the data (attach the plot)? To do this assignment, here are the StataQuest commands.

1. Do Do Files \rightarrow Session logging. Use the file name \textit{hw124}.
2. Do Files \rightarrow Open. Answer \textit{audregf} when asked for file name to load.
3. Do Statistics \rightarrow Correlation \rightarrow Regular (Pearson). When asked for variables, type
   \( \textit{loga logy} \rightarrow \)
4. Do Statistics \rightarrow Simple regression \rightarrow. Answer \textit{loga} when asked for the Y variable. Answer \textit{logy} when asked for the X variable.
   The regression summaries will be shown. After you HAK a menu will pop up. Choose \textit{Plot fitted model} Hit \rightarrow to answer no to the first prompt, then yes to the second. The graph will be shown. Hit \rightarrow again and another menu will pop up. Choose \textit{Save graph for printing}, use \textit{hw124g1} as file name and \textit{log-time for female Y-aud X-vis as label} as label. The graph will be shown again. Hit \rightarrow again and choose \textit{Continue} from the menu.
5. Quit StataQuest by Files \rightarrow Quit, and answer no to prompt. Choose \textit{Print a StataQuest graph} from the main menu and choose the file \textit{hw124g1.gph}. Select the appropriate printer. Also Choose \textit{Print a StataQuest log file} from the main menu. Choose the file \textit{hw124.log} and \textit{lpt1} as printer. Choose \textit{Quit} to get back to DOS.

Problem \#5: Compare the slopes for males and females using a 90\% confidence interval. Also, I want you to construct a plot of the regressions for males and females, but put on the same scale.

What do I mean by this? If you go back through your plots, you’ll notice that the x-axis and the y-axis are different in each of the three plots you’ve done. This makes it hard to compare the pictures for males and females. Especially in project, I’ll want you to have the axes be the same.
This can be done in Stataquest, using the following instructions.

1. Do Files → Session logging. Use the file name hw125.
2. Do Files → Open. Answer audvreg when asked for file name to load.
3. Do Edit → Spreadsheet
4. Hit the F10 key and then do Add → Variable. Answer logvgen to the prompt.
5. Hit the F10 key and then do Replace → Formula. Answer logv*gender to the prompt.
6. Hit the F10 key and then do Files → Save and answer yes twice to the prompt.
7. Return to the StataQuest menu (hit the escape key). Then do Files → Command mode
8. You’ll now have left the menu.
9. Type the following exactly:
   \begin{verbatim}
   regress loga logv gender logvgen
   predict hat
   graph loga hat logv, connect(.1) by(gender) saving(hw125g1) title(Males-0 Females-1)
   \end{verbatim}
   Note: In the last command, the character in the parentheses after connect is lower case “l” instead of the number “1”!
10. Do HAK, hit the F10 key and HAK to return to the StataQuest menu.
11. Do Files → Quit and answer no to the prompts to save the data set
12. You can now print the graph, which is in the file hw125g1.gph

Please note that now the plots for the males and the females are on the same scale, and the differences are obvious.
Consider the Skeena River sockeye salmon data set of previous homework assignments. I'll want you to work with these data to try to understand some of the basic concepts of regression and prediction.

This is CLEARLY a prediction problem. We want to know how many recruits \( Y \) there will be having observed the number of spawners \( X \). One of the more common confusions made in prediction is between prediction the mean response (the mean number of recruits) and predicting an individual response.

**Problem #1**: Evaluate the correlation between spawners and recruits and the \( R^2 \) for the line.

**Problem #2**: Find a 90% confidence interval for the slope of the line.

**Problem #3**: Give a formula for a 90% confidence interval for the mean number of recruits given the number of spawners. Evaluate this interval when the number of spawners is 1,000.

**Problem #4**: Give a formula for a 90% confidence interval for the actual number of recruits given the number of spawners. Evaluate this interval when the number of spawners is 1,000.

Here are the StataQuest instructions (first start StataQuest and get to the StataQuest menu):

1. Do [Files] → [Session logging] when prompted for a file name, answer `hw131`.
2. Do [Files] → [Open] Use the file name `skeena`.
3. Do [Statistics] → [Correlation] → [Regular (Pearson)] When asked for variables, hit \( \Rightarrow \) to use all variables.
4. Do [Statistics] → [Simple regression] Answer `recruit` when asked for the \( Y \) variable. Answer `spawner` when asked for the \( X \) variable.
   The regression summaries will be shown. After you [HAK], a menu will pop up. Choose [Plot fitted model].
   Hit \( \Rightarrow \) to answer no to the first prompt, then yes to the second. The graph will be shown.
   Hit \( \Rightarrow \) again and another menu will pop up. Choose [Save graph for printing], use `hw131g1` as file name and Skeena River Sockeye Salmon as label. The graph will be shown again. Hit \( \Rightarrow \) again and choose [Continue] from the menu.
5. Quit StataQuest by [Files] → [Quit], and answer no to prompt. Choose [Print a StataQuest graph] from the main menu and choose the file `hw131g1.gph`. Select the appropriate printer. Also Choose [Print a StataQuest log file] from the main menu. Choose the file `hw131.log` and `lpt1` as printer.
6. Choose [Quit] to get back to DOS.
CHAPTER 9

EXAM PREPARATION & RULES

I suggest that in preparing for the exam, you first prepare your 3 pages of formulae. This will help you organize the material. Work problems using your three sheets and nothing else. This will help you to know if you have put the right material on the formula pages.

- Make sure to bring a functioning calculator and the standard normal probability tables. I suggest avoiding the use of solar calculators (exams are at night!).
- The test will be primarily, probably only, multiple choice, so please bring a Scantron and the appropriate pencil.
- You will be expected to read the computer output.
- My 1992 exams were given in a different format than you will see this semester. In my 1992 Exam #1, practice by working problems #1--#6 and #9--#12. In my 1992 Exam #2, you can work Problem #6 (change “wrong only 10% of the time” to “with 95% confidence” and use the empirical rule) and Problem #7b, both using the empirical rule. In my 1992 Exam #3, you can use the two-sample empirical rule to work problem #14 (change 90% to 95%).

My 1993 exams are in the same format that you will see. I have provided answers (correct, I hope).

Here are a few extra supplementary exercises which you might find useful in your studying.

- If the sample slope is b = .65, and the sample intercept is a = -1.01, and the standard error of the slope is .25, what is the value of the least squares line at X = 1? What is the error bound on the sample slope? Is the population slope clearly different from zero? Is it clearly greater than 0.2? Is there a relationship between Y and X?
- Suppose that the sample size is n = 9, the sample mean is 1.00 and the sample standard deviation is 0.60. Find a range which covers 90% of the values in the population. Find an error bound for the sample mean. Is the population mean clearly different from 0? from 0.9?
- Suppose that I take two samples, each of size 25, with sample means 10 and 6 and sample standard deviations 4 and 6. Find an error bound on the difference in the sample means. Are the sample means different with probability 95%? Are the populations means different with probability 95%?
CHAPTER 10

TEXAS A&M EXAMS, 1992

EXAMS FROM TEXAS A&M
EXAM #1, STAT 302, SPRING 1992

PROBLEM #1 : ( 15 Points) Consider the following set of data; these data are a sample, not the population!!!

0 0 2 3 4 6 7 10 10 13 13 16 18 19 20 20 21 23 24 29 31 31 33 33 35 38 39 40 40 41 43 45
46 50 54 57 59

Compute

• Sample mean.
• Sample Median.
• Sample standard Deviation.
• Mean absolute error.
• Median Absolute Deviation.
• 75th percentile.
• 25th percentile.
• Interquartile range.
• What percentage of any population is between the 50th and 80th percentiles?
• What percentage of any population be above the 80th percentile?

PROBLEM #2 : ( 15 Points) Consider the following set of data:

0 0 2 3 4 6 7 10 10 13 13 16 18 19 20 20 21 23 24 29 31 31 33 33 35 38 39 40 40 41 43 45
46 50 54 57 59

• What is the sample size?
• What is the sample median?
• Compute a stem–leaf plot of these data with the stems being 0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59.

PROBLEM #3 : ( 15 Points) One of the important components of reporting scientific data is to identify observations in which the experiment has gone awry. Of course, one should report such instances in scientific papers, because the difficulty of doing an experiment is often a good clue as to whether it can be reproduced. In labs, it is common to use a screening technique, i.e., an automated method for deleting observations from further analyses.

The most common method is to say that a point is an outlier if it is more than 3× IQR below the lower quartile or 3× IQR above the upper quartile. If the lower quartile is 9 and the IQR is 18, when will a point be labelled an outlier?

PROBLEM #4: ( ?? Points) In the NHANES data, identify each of the following variables as either discrete, categorical or continuous.

• age in years
• Family Income in dollars
• body mass index
• alcohol usage (yes–no)
• Number of previous live births
• family history of breast cancer (yes–no)
• age at menarche (< 12 or ≥ 12)

PROBLEM #5 : ( 15 Points) Suppose that I have a sample of size 10, and I observe the
following summary statistics.

\[
\sum_{i=1}^{n} Y_i X_i = 100; \quad \sum_{i=1}^{n} X_i = 50;
\]

\[
\sum_{i=1}^{n} Y_i = 30; \quad \sum_{i=1}^{n} X_i = 0.
\]

(a): What is the least squares line \(a + b x\)?
(b): What is its value at \(x = 5\) if \(a = 0\) and \(b = 3\) (these are not correct, but pretend they are anyway)?
(a): Suppose I tell you that the value of a least squares line is 10 when \(x = 5\) and 20 when \(x = 10\). What is the formula for the least squares line?

**PROBLEM #6**: (Points) Fifty students took an exam. Ten of them received an A, and their average grade was 95. The average grade of the other 40 was 75.
(a): What is the sum of the grades for the A students?
(b): What is the sum of the grades for the other students?
(c): What is the average grade?

**PROBLEM #9**: (Points) What is the single most important step that people forget to do when drawing histograms to compare populations, a step that can cause people to come to the wrong conclusion. You have 10 words or less to answer this question.

**PROBLEM #10**: (Points) In 20 words or less, name two major items that histograms and box plots can be used for when comparing two populations.
ANSWERS, EXAM #1, STAT 302, SPRING 1992

PROBLEM #1
(a) 16.18
(b) 16
(c) 12.62
(d) 13.18
(e) 13
(f) 27
(g) 4.5
(h) 22.5
(i) 30%
(j) 20%

PROBLEM #2
(a) 39
(b) 24
(c) NA

PROBLEM #3 25th percentile = 9, 75th percentile = 27. 3× IQR = 54. The rule is an outlier if ≤ 9 – 54 or ≥ 27 + 54.

PROBLEM #4
(a) numerical (discrete)
(b) numerical (continuous)
(c) numerical (continuous)
(d) categorical
(e) categorical
(f) categorical
(g) categorical

PROBLEM #5
(a) a = 3, b = 2
(b) 15
(c) a + 5b = 10 and a + 10b = 20. Thus, b = 2, a = 0 and the line is 2X.

PROBLEM #6
\[ \sum_{i=1}^{4} 0 = 75 \times 40 = 3000 \] 
\[ \sum_{i=1}^{5} 0 = 95 \times 10 = 950 \] 
\[ \bar{X} = \frac{3950}{50} = 79. \]

PROBLEM #9 Not to have common scales on the histograms

PROBLEM #10 Centers and variability
Problem #1: 16 points
(a) As the sample size gets larger, the population variance becomes smaller.
(b) As the sample size becomes very large, \( \bar{X} \) begins to look more and more like the population mean.
(c) The population variance of \( p \) is smaller when \( p = .20 \) than when \( p = .30 \).
(d) The \( z \)-critical value gets smaller as \( n \) increases.
(e) If I compute a 90% confidence interval for a population mean in 10 experiments, then once in the 10 times the mean will not be in the interval.
(f) The length of confidence intervals gets longer as the sample size gets smaller.
(g) The sample variance is always greater than the sample standard deviation.

Problem #2: 12 points Assume \( X_1, \cdots, X_9 \) are a sample from a population which is normally distributed with mean \( \mu_x = 2 \) and variance \( \sigma_x^2 = 36 \). Let \( \bar{X} \) be the sample mean based on these \( n = 9 \) observations. Compute \( P_{\bar{X}}(0 \leq \bar{X}) \).

Problem #3: 12 points Suppose \( X_1, X_2, X_3 \) are a random sample from a normal distributed population with observed values are 4, 5, and 6. Compute a 90% confidence interval for the population mean.

Problem #6: 15 points The mean response time to an auditory stimulus among the general nonathletic population is known to be 28. You are interested in knowing whether the population of athletes has a different response time, so you take a sample of size 11 athletes, and record that their sample mean and variance are 25 and 16, respectively. You’d like to announce a conclusion but be wrong only 10% of the time, at worst. Do the data give sufficient evidence to conclude that athletes have a different mean response time? Why or why not?

Problem #7: 14 points The disease brucelosis (OK, I can’t spell) has infected the Yellowstone National Park bison herd. The disease spreads rapidly among cattle herds, and for this reason ranchers on the edge of the park have been able to convince the Park Service to shoot buffalo which leave the park during winter. A key question is whether the disease is as prevalent now as it was in 1980, when it was known that 40% of the herd was infected. To find out, the park service took a sample of 49 adult bison, and found that 29 of them were infected. If I want to be right 90% of the time, do you think the percentage of infected bison has changed since 1980? Why or why not?

Problem #8: 5 points You are an experimenter, and you are asked by your employer to do a test about whether the mean has changed or not. The employer says that you will be fired if you find a change. The employer also says that you must do a fully valid statistical experiment, with random samples, legitimate data, confidence intervals, etc. In 25 words or less, name the two things you can manipulate to help make sure you find no change. (1 right = 2 points)

Problem #1b: 16 points Each wrong answer loses 4 points. Maximum to be lost = 16 points. Answer each of the following as true or false. No partial credit.
(a) If 30% of males are born left-handed, then in a random sample of 100,000 males, 30,000 will be left-handed.
(b) If the population variance is 2, the population standard deviation is 4.
(c) The \( t \)-critical value gets smaller as \( n \) increases.
(d) If the sample size is \( n = 10 \) and \( \sigma_x \) is unknown, we use \( t \)-critical values.
(e) The empirical rule states that 95% of observations are within 1.96 sample standard deviations of the population mean.
(f) The length of confidence intervals gets shorter as the confidence level increases.
(g) The effect of deleting the auditory stimulus outlier in the homework was to make the sample standard deviation larger.
Problem #4b  You are setting up a practice in dermatology. Patients are either male or female, with 60% being male. Independently, patients either have solar keratoses or basal cell carcinoma; 80% have the former. The probability distribution of your patients is

<table>
<thead>
<tr>
<th>Type</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male &amp; solar</td>
<td>.48</td>
</tr>
<tr>
<td>Male &amp; basal</td>
<td>.12</td>
</tr>
<tr>
<td>Female &amp; solar</td>
<td>.32</td>
</tr>
<tr>
<td>Female &amp; basal</td>
<td>.08</td>
</tr>
</tbody>
</table>

(a) (7 points) You get paid $100 for each solar keratosis, and $400 for each basal cell carcinoma. If it costs you $150 per patient in fixed overhead costs, assuming you'll see lots of patients will your practice turn a profit? Why or why not?

(b) (2 points) What is the chance that a randomly selected patient has a solar keratosis?

(c) (10 points) Assume that the answer to (b) is 70% (this may not be right, but don’t worry about it). Assuming a random sample of size 100, what is the chance that the sample proportion of patients with solar keratoses is greater than .764156?

Problem #5b: 12 points  Suppose I want to know the % of gum chompers who are smokers. I observe 64 chompers, and note that 28 of them smoke. Find a 95% confidence interval for the percentage of chompers who smoke.

Problem #6b  A consistent and unusual phenomenon is that as women age, they are less likely to get a mammogram, even though their risk of breast cancer increases with age. Scott & White clinic is doing a study, and they have found in the past that \( \pi = .30 \) of women age 65+ in the Temple area have had a mammogram in the last two years. In the last few years, they have implemented a new education program, and found that in a random sample of 81 women, 34 of them had received a mammogram in the last two years.

(a) (10 points) Using a 95% confidence interval, is it reasonable for Scott & White to conclude that their education program has raised mammography rates?

(b) (9 points) If the new rate is actually 35%, what is the chance that 34 or more of the 81 women would have received a mammogram in the last two years? Why or why not?

Problem #7b: 10 points  In the past, (pre 1980), it has been thought that the typical American diet has 37 grams of fat per day. You are the statistician on a current study in which a sample of 21 observations is taken, with the sample mean and variance of the grams of fat per day being 30 and 100, respectively. Assuming that the random variable \( X = \text{grams of fat per day} \) is normally distributed, if you are willing to make an error only 5% of the time, do the data contain sufficient evidence to conclude that the population mean grams of fat per day has changed since 1980? Why or why not?
ANSWERS TO EXAM PROBLEMS FOR 1992, EXAM #2

PROBLEM #1 (a) N  (b) Y  (c) N/A  (d) N  (e) N  (f) Y  (g) N

PROBLEM #2

\[ Pr(\bar{X} \geq 0) = Pr \left( Z \geq \frac{0 - 2}{\frac{6}{\sqrt{6}}} \right) = Pr(Z \geq -1) = .8413 \]

PROBLEM #3 \( \bar{X} = 5, s = 1, n = 3, \text{df}=2, 90\% \text{ confidence}, \alpha = .10, t\text{-critical value} = 2.92 \). The C.I. is

\[ \bar{X} \pm (t\text{-c.v.})s/\sqrt{n} = 5 \pm 1.7 = [3.3, 6.7] \]

PROBLEM #6 \( H_0: \mu = 28, \bar{X} = 25, s^2 = 16, s = 4, n = 11, \text{df}=10, 90\% \text{ confidence}, \alpha = .10, t\text{-critical value} = 1.812 \). The C.I. is

\[ \bar{X} \pm (t\text{-c.v.})s/\sqrt{n} = 25 \pm 2.185 = [22.815, 27.185] \]
We reject the hypothesis since with 90\% confidence, the population mean is no more than 27.185.

PROBLEM #7 \( H_0: \pi = 0.4, n = 49, p = 29/49 = .592, 90\% \text{ confidence}, \alpha = .10, z\text{-critical value} = 1.645 \). The C.I. is

\[ p \pm 1.645\sqrt{\frac{p(1-p)}{n}} = .392 \pm .115 = [.477, .707] \]
We reject the hypothesis because with 90\% confidence, the infection rate is at least 47.7\%.

PROBLEM #8 Small sample size, large confidence %

PROBLEM #1b (a) N  (b) N  (c) Y  (d) Y  (e) N/A  (f) N  (g) N

PROBLEM #5b \( n = 64, p = 28/64 = .4375, 95\% \text{ confidence}, \alpha = .05, z\text{-critical value} = 1.96. \) The C.I. is

\[ p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = .4375 \pm .1215 = [.316, .559] \]

PROBLEM #6b, part a: \( H_0: \pi = 0.3, n = 81, p = 34/81 = .420, 95\% \text{ confidence}, \alpha = .05, z\text{-critical value} = 1.96. \) The C.I. is

\[ p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = .420 \pm .107 = [.313, .527] \]
We reject the hypothesis because with 95\% confidence, the rate is at least 31.3\%. Scott & White has done what they wanted to do.

PROBLEM #6b, part b: We are setting \( \pi = .35 \) and trying to compute the chance that the sample fraction is at least 34/81 = .42. This is given by

\[ Pr(p \geq .42) = Pr \left( Z \geq \frac{.42 - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \right) \]
$$= Pr \left( Z \geq \frac{42 - .35}{\sqrt{\frac{35(1 - .35)}{81}}} \right) = Pr(Z \geq 1.32) = .0934.$$ 

**PROBLEM #7b**

$H_0: \mu = 37$, $\bar{X} = 30$, $s^2 = 100$, $s = 10$, $n = 21$, df=20, 95% confidence, $\alpha = .05$, t-critical value = 2.086. The C.I. is

$$\bar{X} \pm (t-c.v.)s/\sqrt{n} = 30 \pm 4.55 = [25.45, 34.55]$$

We reject the hypothesis since with 95% confidence, the population mean is no more than 34.55.
Problem #4: Homework #7 Solution

Problem #5:
\[ n = 25 \quad \bar{X} = 13.3 \quad s = 4.4 \quad n < 30 \quad n - 1 \quad df \quad t.c.v. = 2.06 \]
\[ \bar{X} \pm (t.c.v.)s/\sqrt{n} = 11.49 \text{ to } 15.11 \]

Problem #8:
(a) \( \bar{X} = 11, \quad n = 3, \quad s = .5 \quad n < 30 \quad n - 1 = 2 \quad t.c.v. = 9.93 \)
\[ \bar{X} \pm 9.93s/\sqrt{n} = 11 \pm 2.87 = 8.13 \text{ to } 13.87 \]
(b) The chance is 99\% that population mean is between 8.13 and 13.87
(c) 95\% t.c.v. = 4.30 \( \bar{X} \pm 4.30 \times .5/\sqrt{3} = 11 \pm 1.24 = 9.76 \text{ to } 12.24 \). Reject \( H_0 \) that \( \mu = 9 \)

Problem #10:
\[ n = 18 \quad \bar{X} = 35 \quad s = \sqrt{300} \quad \alpha = .05 \quad 95\% \text{ confidence} \quad H_0 : \mu = 30 \quad n < 30 \quad t.c.v. = 2.11 \quad (17 df) \quad \bar{X} \pm (t.c.v.)s/\sqrt{n} = 35 \pm (2.11)\sqrt{300}/\sqrt{18} = 26.39 \text{ to } 43.6. \text{ Cannot reject } H_0; \text{ Big Smoke may be in compliance.} \]

Problem #12:
\[ \pi = .06 \quad n = 500 \quad \Pr \left( p > \frac{20}{500} \right) = \Pr (p > .04) \]
\[ = 1 - \Pr \left( Z \leq \frac{.04 - .06}{\sqrt{.06(1 -.06)/500}} \right) = 1 - \Pr (Z \leq -1.88) = 1 - .0301 \]

Problem #13
(a) Old homework problem
(b) Omit

Problem #15:
(a) Omit
(b) \( n = 200 \quad p = 42/200 = .21 \quad 95\% \quad p \pm 1.96 \sqrt{p(1-p)/n} \text{ see homework #9} \)

Problems 28, 29, and 33 do not exist.
ANSWER TO VARIOUS PROBLEMS FOR EXAM #2

Problem #4:

\[ n = 2,000 \quad p = .08 \quad 99\% \quad p \pm 2.58\sqrt{p(1-p)/n} = .08 \pm .0156 = .0644 \text{ to } .0956. \]

This is for the population fraction. For total demand, multiply by population size.

Problem #6:

\[ n = 9, \quad \bar{X} = 26, \quad s = 1.87, \quad 95\%, \quad n < 30, \quad n - 1 = 8df \quad t.c.v. = 2.31 \]

C.I. = 26 ± (2.31)(1.87)/sqrt9 = 26 ± 1.44 = 24.86 to 27.44. The mean \( \mu \neq 30 \Rightarrow \) don’t trust salesperson.

Problem #8:

\[ n = 9, \bar{X} = 50, s = 20 \quad 99\% \quad n < 30 \quad n - 1 = 8df \quad t.c.v. = 3.36 \quad \text{confidence interval} = 50 \pm (3.36)(20)/\sqrt{n} = 27.6 \text{ to } 72.4. \]

Problem #10:

\[ \alpha = .05 \quad 95\% \text{ confidence,} \quad n = 100, \bar{X} = 770, s = 120 \quad n > 30 \quad z.c.u. = 1.96 \quad 770 \pm (1.96)(12)/\sqrt{100} = 766.48 \text{ to } 773.52. \]

Cannot be said to be different from $750.

Problem #11:

\[ n = 16 \quad n - 1 = 15 \quad \text{Use } t.c.v. \bar{X} : 3.87 \quad s = .7 \]

(a) t.c.v. for 95\% = 2.13 \quad 3.87 \pm (2.13)(.7)/\sqrt{16} = 3.50 \text{ to } 4.24

(b) t.c.v. for 90\% = 1.75 \quad 3.87 \pm (1.75)(.7)/\sqrt{16} = 3.56 \text{ to } 4.18.

Cannot conclude that they are scrimping.

Problem #12:

(a) decreases

(b) decreases

Problem #14:

\[ H_0 : \mu = 33 \quad n = 50 \quad \bar{X} = 29.5 \quad s = \sqrt{128} \quad \alpha = .01 \quad 99\% \quad z.c.u. = 2.58 \quad 29.5 \pm (2.58)\sqrt{128}/\sqrt{50} = 25.37 \text{ to } 33.63. \]

Claim could be true.

Problem #19:

\[ n = 100 \quad \bar{X} = 1285 \quad s = 150 \quad n > 30 \quad 95\% \quad \bar{X} \pm 1.96 s/\sqrt{n} = 1285 \pm 29.4 \]

Problem #20:

\[ 95\% \quad n = 625 \quad p = 58/625 = .0928 \quad p \pm 1.96 \sqrt{p(1-p)/n} = .0701 \text{ to } .1156 \]

Problem #26:

\[ H_0 : \mu = 120 \quad n = 16 \quad \bar{X} = 117 \quad s = 8 \quad \alpha = .05 \quad 95\% \quad n < 30 \quad n - 1 = 15 \quad df \quad t.c.v. = 2.13 \quad \bar{X} \pm 2.13 s/\sqrt{n} = 117 \pm (2.13)(8)/\sqrt{16} = 117 \pm 4.26 = 112.74 \text{ to } 121.26. \]

Don’t reject \( H_0. \)

Problem #34:

\[ n = 100 \quad 90\% \quad p = .36 \quad p \pm 1.645 \sqrt{p(1-p)/n} = .28 \text{ to } .44 \]

Problem #38:

\[ H_0 : \mu = 12 \quad n = 60 \quad \bar{X} = 11.95 \quad s = .09 \quad \alpha = .05 \quad 95\% \quad \bar{X} \pm 1.96 s/\sqrt{n} = 11.95 \pm (1.96)(.09)/\sqrt{60} = 11.93 \text{ to } 11.97. \]

Pop mean is < 12
Problem #40:

\[ H_0 : \mu = 120 \quad n = 36 \quad \bar{X} = 122.8 \quad s = 10.9 \quad \alpha = .05 \quad 95\% \quad n > 30 \quad z.c.u. = 1.96 \quad \bar{X} \pm 1.96s/\sqrt{n} = 122.8 \pm (1.96)(10.9)/\sqrt{36} = 119.2 \text{ to } 126.4. \] Population mean could equal 120.

Problem #41:

(a) \[ H_0 : \pi = .60 \quad n = 200 \quad p = .52 \quad (a)? \quad p \pm 1.96\sqrt{\frac{p(1-p)}{n}} \quad (\alpha = .05 = 95\%) = .52 \pm 1.96\sqrt{(.52)(.48)/200} = .451 \text{ to } .589. \] Reject \( H_0. \)
QUESTION #1, 15 POINTS

Circle the single correct answer. No partial credit. Each mistake will cost you three points.

1. If a null hypothesis is rejected at the level \( \alpha = 0.05 \) then:
   A. the null hypothesis will also be rejected at the level \( \alpha = 0.01 \).
   B. the null hypothesis will also be rejected at the level \( \alpha = 0.10 \).
   C. the \( p \)-value associated with the test statistic is greater than 0.05.
   D. the computed test statistic is a \( z \)-statistic.
   E. the alternative hypothesis would also be rejected at the level \( \alpha = 0.05 \).

2. Select the statement below which best represents the primary objective of statistics.
   A. To use formulas to manipulate data.
   B. To make inferences about a population based on information contained in a sample.
   C. To make a definite conclusion about some phenomenon.
   D. To remove all mathematical assumptions from techniques of data analysis.
   E. To make inferences about a sample with a high degree of reliability.

3. If the 95% confidence interval for a population mean \( \mu \) is (-1,1) and you are testing
   \( H_0: \mu = 1.3 \) against \( H_a: \mu \neq 1.3 \),
   A. you fail to reject \( H_0 \) at 1% level.
   B. you fail to reject \( H_0 \) at 5% level.
   C. you reject \( H_0 \) at 5% level.
   D. you reject \( H_0 \) at 1% level.
   E. there is insufficient information to make a decision.

(4.) We have had a bewildering array of formulas and definitions, especially concerning
\( s_b \) and \( s_e \). Which is correct?
   A. \( s_b \) measures the average error about the line.
   B. Knowing the sample size, the slope \( b \) and \( s_b \), and nothing else enables you to construct
      a confidence interval for the population slope.
   C. Knowing the slope \( b \) and \( s_b \) and the sample size and nothing else enables you to
      construct a confidence interval for the population slope.
   D. Knowing the slope \( b \) and \( s_b \) and the sample size and nothing else enables you to
      construct a confidence interval for the population mean.

(5.) If the Pearson correlation is -0.80, and there are no outliers, which is most accurate:
   A. The slope of the line is positive, and there is little variation about the line.
   B. The slope of the line is negative, and there is little variation about the line.
   C. The slope of the line is positive, and there is lots of variation about the line.
   D. The slope of the line is negative, and there is lots of variation about the line.

QUESTION #2, 7 POINTS

Suppose you do a confidence interval for a slope, and find
that zero is in the interval. Which statement is correct, and why is the other one incorrect
(30 words or less, and I’ll take off points for anything wrong).

A. We are 90% confident that there is no correlation between \( X \) and \( Y \).
B. At 90% confidence, we do not have enough information to conclude that a relation-
ship between $X$ and $Y$ exists.

**QUESTION #3, 15 POINTS**  
Be 90\% confident in this problem. It has been hypothesized that male undergraduates at A&M are four inches taller than females on average. To test this hypothesis, a random sample of 12 males and 12 females was taken. The sample mean and sample standard deviation of the heights of men in inches was 71 and 3, respectively, while the sample mean and sample standard deviation of women was 63 and 2, respectively.

(a) Does the hypothesis seem reasonable, why or why not?

(b) Give a precise description of what the data tell you about how much taller men are than women on average in the populations?

(c) Now suppose I tell you that the $p$-value for the hypothesis is 6\% [it is not, but do not worry about it]. What decision would you make about the hypothesis if you want to be 90\% certain about being right, and why?

**QUESTION #4, 14 POINTS**  
Be 90\% confident in this problem. A recent completely random survey of 825 women and 583 men revealed that 273 of the women gambled regularly whereas 39.28\% of the men were regular gamblers. Do you think that there is enough evidence to support the conclusion that men and women differ in their gambling habits (in terms of regularity)? Why or why not?

**QUESTION #5, 6 POINTS**  
In trying to decide whether to hold the third exam before or after the project due date, I decide to let majority rule. Hence, I am planning to take a survey of the students in this and next year's class, and see what they prefer. What sample size should I take if I want to be 90\% confident that the observed sample fraction of students who wish to take the exam before the project is within .20 of the true sample fraction.

**QUESTION #6, 18 POINTS**  
Be 90\% confident in this problem. Suppose I am interested in comparing the height/weight relationship of men to women, where height is in inches over 5 feet and weight is in pounds. I take a sample of 10 men and 14 women. The information gathered from the sample is as follows:

- For Men: mean and s.d. for height = 10 and 3. Mean and s.d. for weight = 180 and 30. Slope and standard error of slope for predicting weight from height = 10 and 0.5. Intercept and standard error of intercept for predicting weight from height = 90 and 3.0.
- The value of $s_\varepsilon = 1$ for both men and women.
- For Women: mean and s.d. for height = 6 and 3. Mean and s.d. for weight = 150 and 20. Slope and standard error of slope for predicting weight from height = 7 and 0.8. Intercept and standard error of intercept for predicting weight from height = 110 and 4.0.

(a) What are the predicted weights for men and women who are 5 feet 5 inches tall?

(b) Is there evidence that the population intercept for men is different from that of women? How different?

(c) In part (a), give a range for the mean weight in the population of women who are 5 feet 5 inches tall.
QUESTION #7, 15 POINTS

A group of students in the 1992 class was interested in the crown ratio of Loblolly Pine trees in open and dense stands. They want to know whether or not the mean crown ratios of the two populations are the same, and if not, by how much are they different. They are also particularly interested in knowing whether the p-value for the comparison is less than 0.01.

I will now give you the information from their project, and you should answer the questions posed in the previous paragraph.

The students took two independent random samples. In the open stand, they took a sample of size 30 and found that the sum (not the mean) of the crown ratios for all observations was 1450, while the standard deviation was 10.28. In the dense stand, they found that the sum of the 30 crown ratios for all the observations was 900, while the sample standard deviation was 11.93.

QUESTION #8, 10 POINTS

Consider a generic problem involving continuous random variables. Suppose that I know that the population standard deviation $\sigma$ is 3.0 and that the population mean is 1.5. I take a sample of size 9. If my decision rule is to reject the hypothesis that $\mu = 1.0$ whenever the sample mean exceeds 3.5 or is less than -0.5, what is the Type I error?

QUESTION #9, 16 POINTS

Circle the single correct answer. No partial credit. Each mistake will cost you three points.

1. A two sample $t$-test was performed using independent samples from two populations to test the null hypothesis $\mu_1 - \mu_2 = 0$. What conclusion can be made if a very small $p$-value ($< .001$) is obtained?
   A. We have evidence that $\bar{X}_1$ and $\bar{X}_2$ are the same.
   B. We have evidence that $\mu_1$ and $\mu_2$ are the same.
   C. We have evidence that $\bar{X}_1$ and $\bar{X}_2$ are not the same.
   D. We have evidence that $\mu_1$ and $\mu_2$ are not the same.
   E. The sample sizes were very large.

2. If the $p$-value of a hypothesis test is 0.02 and you are performing a 5% level test, then
   A. the null hypothesis will not be rejected at 5% level.
   B. the null hypothesis will be rejected at 5% level.
   C. we don't have enough information to make a decision.

3. Suppose I want to compare the population slopes for men and women in a physiology experiment. The men's slope is .13, the women's is .07, and a 99% confidence interval for the difference of one to the other (which one you have to figure out) is [-.11, -.01]. What is the most appropriate conclusion, i.e., what would I conclude?
   A. Men have a slope at least .03 greater than women.
   B. Women have a slope no more than .11 greater than men.
   C. We can't reject the hypothesis that men have a greater slope than women.
   D. Men have a slope no more than .11 greater than women.

4. I'm trying to compare two slopes to see if they are different. The 90%, 95% and 99% confidence intervals for the difference in slopes are [-1.0, -2.0], [-1.15, -0.5] and [-1.30, 1.0], respectively. What is the most accurate statement about the $p$-value for the test that the
slopes are the same:
A. less than 1%
B. between 1% and 5%
C. between 5% and 10%
D. greater than 10%.

QUESTION #10, 14 POINTS

U. S. A. Today (Jan. 15, 1986) described a study of automobile accident victims at Newton-Wellesley Hospital in Massachusetts. The mean cost for medical treatment for a motorist wearing a seatbelt at the time of the accident was $465, while the mean cost for a motorist who was not wearing a seatbelt was $1200. Suppose the sample sizes and sample standard deviations were as given in the accompanying table

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>average</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>15</td>
<td>465</td>
<td>25</td>
</tr>
<tr>
<td>Y</td>
<td>9</td>
<td>1200</td>
<td>54</td>
</tr>
</tbody>
</table>

(a) Test whether the population mean costs differ, using a confidence interval and a Type I error rate of 5%. No credit unless you compute a confidence interval.

(b) Suppose the difference between not wearing and wearing seatbelts has a 95% confidence interval of [300, 800]. Interpret this.

QUESTION #11, 12 POINTS

One of my T.A.'s thinks that, over the course of a number of semesters, a higher percentage of the students coming to his office hours are women than to my office hours. Over the course of the last month, he had 98 students come to his office, of whom 79 were women, while of my 136 visitors, 65% were women. I of course believe that there isn't any real difference, and that his hypothesis is just based on subjective impressions. I'll give him the benefit of the doubt and let the final conclusion be wrong 5% of the time. Do you think his conclusion is correct, and can you give a range on how his percentage differs from mine?

QUESTION #12, 14 POINTS

Suppose you want to find out whether men or women are bigger tippers, and by how much, being 90% confident of being correct.

I will now describe an experiment and state the data. Use this information to answer the questions posed in the previous paragraph.

A group went to a number of restaurants and got the waiters and waitresses to write down the size of the tip, expressed as a percentage of the total bill. To discern the sex effect, they considered only all males groups and all female groups of from 2–5 people eating at a table. There were 14 groups of males and 14 groups of females. The males had a sample mean tip of 30, while the females had a sample mean tip of 16. The sample standard deviations for males and females were 12 and 10, respectively. Use this information to answer the questions posed in the first paragraph.

QUESTION #13, 5 POINTS

Consider the two box plots given below. Pretend that, as drawn, there is a statistically significant difference in the means of the two populations. Show with a "*" the location of an outlier which is most likely to change the result to be not statistically
question #14, 21 points

on the next two pages i am attaching mystat output
for the normal and abnormal kids in the aortic stenosis experiment. recall that \( x = \log(1 + \text{body surface area}) \) and \( y = \log(1 + \text{aortic value area}) \). answer the following:

(a) what is the least squares line for the normal kids?

(b) which population shows a stronger linear relationship between \( x \) and \( y \)? justify your answer with numbers. remember, just because one slope is larger than the other does not mean there is a stronger linear relationship. \([ \leq 15 \text{ words}]\).

(c) if i want to be right 90\% of the time, how much smaller is the population slope for the abnormal kids than the population slope for the normal kids?

(d) if i observe a kid with \( x = 0.5 \), what is my prediction for \( y \) if the kid is abnormal and hence has aortic stenosis?

(e) suppose i have observed a kid who has the values \( x = 0.5 \) and \( y = 0.20 \). if i want to label a normal kid as normal 90\% of the time, should i label this kid as normal? justify your answer.

question #15, 8 points

suppose that you are in charge of managing a large game reserve in africa, and you are concerned about maintaining the population of leopards in a time of drought. based on past experience, you know that the natural mortality of leopards is 20\% per year. because of an intensive program, you have been able to radio-collar 300 leopards. if the drought is seriously affecting the leopard population, you are going to allow the importation of a few american cattle, who will be easy prey. your decision rule is that the program will begin at the end of the year if more than 100 of the collared leopards have died.

(a) what is the type i error of this decision rule?

(b) describe is \( \leq 25 \text{ words} \) what it means here to make a type ii error.
ANSWERS, EXAM #3, STAT 302, SPRING 1992

PROBLEM #1
(1) B  
(2) B  
(3) C  
(4) C  
(5) B

PROBLEM #2
   B (A is wrong because we can never “conclude” that there is no relation, i.e. $H_0$ is true, but only “fail to reject” $H_0$.

PROBLEM #3
   (a) $n_1 + n_2 = 24 < 30$, so we use t-critical values with 22 degrees of freedom, i.e., $1.72$. The pooled sample variance is $s^2_p = 6.5$, so a 90% confidence interval for the difference in population mean heights is $8 \pm 1.72\sqrt{(6.5/12) + (6.5/12)} = 8 \pm 1.8$.
   (b) Men are thus between 6.2 and 9.8 inches taller, on average. The hypothesis that they are 4 inches taller on average is not true, at 90% confidence.
   (c) Since the $p$-value is 0.06, which is less than $0.1$, we would reject $H_0$ at $0.1$.

PROBLEM #4 $p_{women} = \frac{271}{827} = 0.3309$, $p_{men} = 0.3928$. We are testing $H_0: p_w = p_m$ vs. $H_A: p_w \neq p_m$ at $0.10$. The 90% confidence interval is
   \[ 0.3928 - 0.3309 \pm 1.645 \sqrt{\frac{0.3928(1 - 0.3928)}{383} + \frac{0.3309(1 - 0.3309)}{825}} = 0.0619 \pm 0.0428 = 0.0191 \text{ to } 0.1047 \]
   Men are more likely to gamble.

PROBLEM #5 Skip. This is a sample size problem.

PROBLEM #6 According to the problem, you are regressing weight $Y$ on height $X$.
   (a) For men, height = 5, $\bar{Y} = 90 + 10 \cdot 5 = 140$ pounds. For women, height = 5, $\bar{Y} = 110 + 7 \cdot 5 = 145$ pounds.
   (b) A 90% confidence interval for $(\alpha_1 - \alpha_2)$ would be
   \[
   (\alpha_1 - \alpha_2) \pm t_{0.05, n_1+n_2-4} \sqrt{\frac{s^2_{\alpha_1}}{n_1} + \frac{s^2_{\alpha_2}}{n_2}}
   \]
   \[
   \Rightarrow (110 - 90) \pm 1.73 \sqrt{9 + 16}
   \]
   \[
   \Rightarrow 20 \pm 1.73 \cdot 5
   \]
   \[
   \Rightarrow (11.35, 28.65)
   \]
   Since 0 is not in this interval, we reject the hypothesis that the two intercepts are equal.
   (c) For women, the number of observations is $n = 14 < 30$, so you use a t-critical value with 12 degrees of freedom, which at 90% confidence is 1.78. The standard deviation of heights is $s_x = 3$. The 90% confidence interval for the mean at $X^* = 5$ is
   \[
   (a + b \cdot X^*) \pm 1.78 \cdot s_x \sqrt{1 + \frac{1}{n} \frac{(X^* - \bar{X})^2}{(n-1)s_x^2}}
   \]
\[ 145 \pm 1.78 \cdot 1 \cdot \sqrt{1 + \frac{1}{14} + \frac{(5 - 6)^2}{13 \cdot 3^2}} \]
\[ 145 \pm 1.85 \]
\[ (143.15, 146.85) \]

**Problem #7** The total sample size exceeds 30, so we use \( z \)-critical values, i.e., since we want 99\% confidence, the critical value is 2.58.

\[
\begin{align*}
\sum n_1 &= 30 \\
\bar{X}_1 &= \frac{1460}{30} = 48.33 \\
s_1 &= 10.28 \\
\sum n_2 &= 30 \\
\bar{X}_2 &= \frac{900}{30} = 30 \\
s_2 &= 11.93 \\
\end{align*}
\]

First, we'll find the 99\% confidence interval for \((\mu_1 - \mu_2)\):

\[ (\bar{X}_1 - \bar{X}_2) \pm 2.58 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]
\[ (48.33 - 30) \pm 2.58 \sqrt{\frac{10.28^2}{30} + \frac{11.93^2}{30}} \]
\[ 18.33 \pm 7.418 \]
\[ (10.915, 25.75) \]

Therefore we conclude that the two population means differ by somewhere between 10.915 and 25.75 with 99\% confidence. Also, since this is a 99\% confidence interval, and zero is not included, we conclude that the \( p \)-value is less than 0.01.

**Problem #8** We are not computing Type I errors this year.

**Problem #9**

1. D
2. B
3. D
4. B

**Problem #10**

(a) Since the total sample size is less than 30, we will use \( t \)-critical values. The confidence level is 95\%, with 22 degrees of freedom, so the \( t \)-critical value is 2.07. We first calculate the pooled sample variance, namely

\[
s_p^2 = \left( \frac{n_1 - 1}{n_1 + n_2 - 2} \right) s_1^2 + \left( \frac{n_2 - 1}{n_1 + n_2 - 2} \right) s_2^2 \\
= \frac{14}{22} \cdot 25^2 + \frac{8}{22} \cdot 54^2 \\
= 1458 \\
\]

The 95\% confidence interval for \((\mu_1 - \mu_2)\) is

\[ (\bar{X}_1 - \bar{X}_2) \pm 2.07 \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \]
\[ (1200 - 465) \pm 2.07 \sqrt{\frac{1458}{15} + \frac{1458}{9}} \]
Since zero is not contained in this interval, we conclude that the mean costs of the two populations are different, with a Type I error rate of 5%.

(b) We are 95% confident that the true difference between population means is between 300 and 800.

**PROBLEM #11** This is from homework #11. The T.A.’s worked it, and so check to see if the following is correct. The sample proportions are \( p_1 = \frac{70}{58} = .806 \) and \( p_2 = .65 \). A 95% confidence interval for the difference is

\[
(p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}
\]

\[= (.806 - .65) \pm 1.96 \sqrt{\frac{.806(1-.806)}{98} + \frac{.65(1-.65)}{136}}
\]

\[= .156 \pm .1121
\]

\[= (.0441, .2682)
\]

Since zero is not contained in this interval, we can conclude with 95% confidence that the percentages are different.

**PROBLEM #12** The total sample size is 28 < 30, so we have to use \( t \)-critical values. It asks you to be 90% confident, so the critical value is 1.71.

\[
\bar{X}_m = 30 \quad s_m = 12 \quad n_m = 14
\]

\[
\bar{X}_f = 16 \quad s_f = 10 \quad n_f = 14
\]

The pooled variance \( s_p^2 \) is

\[
s_p^2 = \left( \frac{n_m - 1}{n_m + n_f - 2} \right) d_m^2 + \left( \frac{n_f - 1}{n_m + n_f - 2} \right) d_f^2
\]

\[= \frac{13}{26} \cdot 12^2 + \frac{13}{26} \cdot 10^2
\]

\[= 122
\]

The 90% confidence interval for \((\mu_m - \mu_f)\) is

\[
(\bar{X}_m - \bar{X}_f) \pm 1.71 \sqrt{\frac{s_p^2}{n_m} + \frac{s_p^2}{n_f}}
\]

\[= (30 - 16) \pm 1.71 \sqrt{\frac{122}{14} + \frac{122}{14}}
\]

\[= 14 \pm 7.139
\]

\[= (6.86, 21.14)
\]

Men are bigger tippers.

**PROBLEM #13** Whoops!

**PROBLEM #14**

(a) For normal kids, \( Y = .0471 + 1.115 \cdot \log\text{bsa} \).
(b) By comparing $r^2$s, the data for the normal kids have stronger linear relationship ($r^2_{\text{normal}} = .7642$ vs. $R^2_{\text{abnormal}} = .3567$).

(c) A 90% confidence interval for the difference in slopes is

$$ (b_1 - b_2) \pm 1.645 \sqrt{s_b^2 + s_b^2} $$

$$ \implies (1.115 - .670) \pm 1.645 \sqrt{.0751^2 + .1222^2} $$

$$ \implies .445 \pm .236 $$

$$ \implies (.209, .6815) $$

(d) For an abnormal kid, $Y = .045 + .670 \cdot (.5) = .380$.

(e) A 90% confidence interval for the logit for normal kid with $X^* = .5$ is

$$ (a + b \cdot X^*) \pm 1.645 \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n - 1)s_e^2}} $$

$$ \implies (.047 + 1.115(.5)) \pm 1.645 \cdot 1.1757 \sqrt{1 + \frac{1}{70} + \frac{(.5 - .5455)^2}{69(.2816)^2}} $$

$$ \implies .6045 \pm .29117 $$

$$ \implies (.313, .896) $$

Since $Y = .2$ is not in this interval, I should not label this kid as normal (with 90% confidence).

**PROBLEM #15** Not doing this problem this year.
### KIDS WITH AORTIC STENOSIS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOGMA</td>
<td>56</td>
<td>0.7613</td>
<td>0.2729</td>
<td>0.182</td>
<td>1.2</td>
</tr>
<tr>
<td>LOGMA</td>
<td>56</td>
<td>0.5650</td>
<td>0.3069</td>
<td>0.068</td>
<td>1.386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1.83616515</td>
<td>1</td>
<td>1.83616515</td>
<td>64</td>
</tr>
<tr>
<td>Residual</td>
<td>3.31199372</td>
<td>64</td>
<td>0.061333217</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.14815888</td>
<td>65</td>
<td>0.093602888</td>
<td></td>
</tr>
</tbody>
</table>

| LOGMA | Coef. | Std. Err. | t     | P>|t| | 95% Conf. Interval |
|-------|-------|-----------|-------|-----|-------------------|
|       |       |           |       |     |                   |

### KIDS WITH HEALTHY HEARTS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>logbsa</td>
<td>70</td>
<td>0.6458</td>
<td>0.2816</td>
<td>0.077</td>
<td>1.131</td>
</tr>
<tr>
<td>logava</td>
<td>70</td>
<td>0.6564</td>
<td>0.3592</td>
<td>0</td>
<td>1.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6.80464674</td>
<td>1</td>
<td>6.80464674</td>
<td>68</td>
</tr>
<tr>
<td>Residual</td>
<td>2.09067039</td>
<td>68</td>
<td>0.030877506</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8.90431714</td>
<td>69</td>
<td>0.129048074</td>
<td></td>
</tr>
</tbody>
</table>

| LOGMA | Coef. | Std. Err. | t     | P>|t| | 95% Conf. Interval |
|-------|-------|-----------|-------|-----|-------------------|
|       |       |           |       |     |                   |

| logbsa | Coef. | Std. Err. | t     | P>|t| | 95% Conf. Interval |
|--------|-------|-----------|-------|-----|-------------------|
|        |       |           |       |     |                   |
CHAPTER 11

TEXAS A&M EXAMS, 1993

Question #1  How do you spell the instructor's surname? (see top of exam)

Question #2  Refer to the box plot pictures. In comparing plots (c) and (e), which is most nearly correct?

Question #3  Refer to the box plot pictures. In box plot (a), approximately what % of the sample lies between 70 and 80?

Question #4  Which of the following cannot be read off from a MYSTAT stem-leaf plot? See the MYSTAT output at the end of the exam if you are in doubt.

(a) The former has bigger mean, bigger variance.
(b) The former has bigger mean, smaller variance.
(c) The former has smaller mean, bigger variance.
(d) The former has smaller mean, smaller variance.
(e) The former has smaller mean, about equal variance.

Questions #6-#7 relate to the following sample:

9  -2  -10  5  7  -6  4

Question #6  What is the sample median?

(a) 4  
(b) 5  
(c) 7  
(d) 1  
(e) None of the above.

Question #7  What is the MAD?

(a) 0  
(b) 1  
(c) 3  
(d) 6  
(e) None of the above.

Question #8  Using Z-tables, what is Pr[Z ≤ 1.2]?

(a) .1151  
(b) .8849  
(c) .587  
(d) .8413  
(e) None of the above

Question #9  If the population mean of X is 1.0 and the population standard deviation is 2.0,
what fraction of the population is between 1 and 3.0?
(a) .8413  
(b) .4987  
(c) .3413  
(d) .1587  
(e) None of the above

Questions #10—#11 relate to the following data:
\[
\begin{array}{cc}
X & Y \\
-2 & 0 \\
1 & 3 \\
1 & 4 \\
\end{array}
\]

**Question #10** What is the slope of the least squares line? Use rounding if necessary.
(a) 1.17  
(b) 2.33  
(c) 1.35  
(d) 4.04  
(e) None of the above

**Question #11** Suppose the sample slope is 1.0 and the sample intercept is 3.0. What do you predict \(Y\) should be if \(X = 0\)?
(a) 1.0  
(b) 2.0  
(c) 3.0  
(d) 4.0  
(e) None of the above

**Question #12** Suppose you are interested in the mean weight loss from a 3-month weight watchers (WW) diet. You take a sample of 81 and observe a sample mean of 7 and a sample variance of 9. Use the empirical rule at 95% to describe an interval for the population mean. Please note, I have done some rounding to make the answers pretty. You should pick the answer closest to the actual interval.
(a) 7 ± (1/3)  
(b) 7 ± 1  
(c) 7 ± 2  
(d) 7 ± (2/3)  
(e) 7 ± (2/9)

**Question #13** Suppose your interval in the Weight Watchers program is from 6 to 9. They claim the population mean is 10. What do you conclude?
(a) Who can tell with statistics?  
(b) Statistics can’t help answer the question.  
(c) Weight Watchers is probably wrong.  
(d) Weight Watchers is probably right.  
(e) Not enough information.

Questions #14—#19 relate to the MYSTAT output I’ve attached. The variables are the quarters (3-month periods) since 1983, which is called DATE, and the number of AIDS cases diagnosed in that quarter (AIDSCASE), in hundreds. Hence, the number 40 for AIDSCASE means that there were 4,000 cases in that quarter. This is the regression of AIDSCASE on DATE.

**Question #14** Which variable is \(X\), and which is \(Y\)?
(a) \(Y = \text{Date}, \ X = \text{AIDSCASE}\).  
(b) \(X = \text{Date}, \ Y = \text{AIDSCASE}\).

**Question #15** What % of the observations are strictly greater than 31?
(a) 62.1%
(b) 58.6%
(c) 55.2%
(d) 41.4%
(e) None of the above

Question #16: Use the empirical rule to describe a 95% interval for the number of cases per quarter. Please note, I have done some rounding to make the answers pretty. You should pick the answer closest to the actual interval.
(a) 30.4 to 41.6
(b) 33.2 to 38.8
(c) 21 to 51
(d) 6 to 66
(e) None of the above

Question #17 What is the least squares slope of the regression of AIDSCASE on DATE?
(a) 0.11
(b) 1.25
(c) 1.9
(d) 94
(e) None of the above.

Question #18 Suppose you are at quarter #14. What is your best estimate of the number of new AIDS cases for that quarter, in hundreds?
(a) 33.5
(b) 12.5
(c) 1.5
(d) 176.5
(e) None of the above.

Question #19 Suppose that you use the empirical rule and conclude that the estimated slope is 1.0 with an error bound of 0.7. Which statement is most correct?
(a) Since the slope is so large, there is clearly a relationship between the two variables.
(b) The range makes it incorrect to come to the conclusion that there is a relationship.
(c) Since the population slope is less than 1.7, there is clearly a relationship.
(d) Since the population slope is at least 0.3, there is clearly a relationship.
(e) Since the sample slope is at least 0.3, there is clearly a relationship.

Question #20 In fermentation experiments to make a new drug, chemists often use quality limits to determine whether a current batch is any good. Generally, it is thought that batch reactivities in the log scale are normally distributed with a population standard deviation of 2 (the scale of the numbers is irrelevant, but because of the log transformation any number is possible). Batches are accepted when their sample reactivity is less than 5.0 or exceeds -1. If batches have a true population mean reactivity of 1.0, what fraction of batches are accepted?
(a) .8413
(b) .8185
(c) .9772
(d) .6915
(e) None of the above. Questions #21-#22 refer to the following scenario. In the Concho Water Snake class project, we are interested in whether females or males have longer tails. We took a sample of size 20 from each. The males had a sample mean of 6 and a sample variance of 10 (sample variance is the square of the sample standard deviation). The females had a sample mean tail length of 9 and a sample variance of 10.

Question #21 What is the 95% error bound for the difference in the sample means? Please note, I have done some rounding to make the answers pretty. You should pick the answer closest to the actual interval.
(a) 3.16
(b) 6.32
(c) 1.0
(d) 2.0
(e) None of the above

**Question #22** Suppose that the error bound is 2.0. What does this mean?
(a) Substantial evidence for a difference in sample tail lengths.
(b) Substantial evidence for a difference in population tail lengths.
(c) Little evidence for a difference in sample tail lengths.
(d) Little evidence for a difference in population tail lengths.

Questions #23–#24 refer to the following scenario. You are interested in the GPA’s of graduating seniors at A&M and take a sample of size \( n = 4 \), observing a sample mean of 3.0 and a sample standard deviation of 0.25. Using empirical rules, answer the following questions. Please note, I have done some rounding to make the answers pretty. You should pick the answer closest to the actual interval.

**Question #23** Where do 95% of the population values lie?
(a) from 2.75 to 3.25.
(b) from 2.50 to 3.25
(c) from 2.50 to 3.50
(d) from 2.75 to 3.50
(e) Empirical rules are not relevant for answering this question.

**Question #24** Suppose that the error bound on the sample mean is 0.30. What does this mean?
(a) The sample mean is between 2.7 and 3.3.
(b) The population mean is between 2.4 and 3.6.
(c) The sample mean is between 2.4 and 3.6.
(d) The population mean is between 2.7 and 3.3.

**Question #25** Refer to the plots at the end of the exam. Ignoring the labels on the plots, and suppose I tell you that the intercept is 3 and the slope is -1.5. Which of plots (a) and (b) is most likely to have been associated with these data?
(a) Least squares plot (a).
(b) Least squares plot (b).
(c) Not enough information to determine.
Variables:
DATE = Quarters since 1983
AIDSCASE = Numbers of AIDS cases, in hundreds

TOTAL OBSERVATIONS: 29

<table>
<thead>
<tr>
<th>DATE</th>
<th>AIDSCASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>29</td>
</tr>
</tbody>
</table>

MINIMUM: 1.000  15.000
MAXIMUM: 29.000  51.000
MEAN: 15.000  36.000
STANDARD DEV: 8.515  15.000

STEM AND LEAF PLOT OF VARIABLE: AIDSCASE , N = 29
MINIMUM IS: 15.000
LOWER HINGE IS: 20.000
MEDIAN IS: 46.000
UPPER HINGE IS: 48.000
MAXIMUM IS: 51.000

DEP VAR: AIDSCASE N: 29 MULTIPLE R: .940 SQUARED MULTIPLE R: .884
ADJUSTED SQUARED MULTIPLE R: .879 STANDARD ERROR OF ESTIMATE: 4.880

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>STD ERROR</th>
<th>STD COEF</th>
<th>T</th>
<th>P(2 TAIL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>12.5</td>
<td>1.900</td>
<td>0.000</td>
<td>6.743</td>
<td>0.000</td>
</tr>
<tr>
<td>DATE</td>
<td>1.5</td>
<td>0.110</td>
<td>0.940</td>
<td>14.330</td>
<td>0.000</td>
</tr>
</tbody>
</table>

ANALYSIS OF VARIANCE

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM-OF-SQUARES</th>
<th>DF</th>
<th>MEAN-SQUARE</th>
<th>F-RATIO</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGRESSION</td>
<td>4891.035</td>
<td>1</td>
<td>4891.035</td>
<td>205.345</td>
<td>0.000</td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>643.103</td>
<td>27</td>
<td>23.819</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ANSWER KEY FOR EXAM #1, 1993

1 = (a)
2 = (c)
3 = (d)
4 = (b)
5 = (a)
6 = (a)
7 = (e)
8 = (b)
9 = (c)
10 = (a)
11 = (c)
12 = (d)
13 = (c)
14 = (b)
15 = (b)
16 = (d)
17 = (e)
18 = (a)
19 = (d)
20 = (b)
21 = (d)
22 = (b)
23 = (c)
24 = (d)
25 = (a)
Figure 11.1. Sample histograms, 1993 exam #1.
Figure 11.2. Sample box plots, 1993 exam #1.
Figure 11.3. Sample regression plots, 1993 exam #1.
QUESTION #1: If I want to form a 99% confidence interval for the population mean and \( n = 20 \), what is the critical value?
   (a) 2.86  
   (b) 2.54  
   (c) 2.85  
   (d) 2.38  
   (e) 1.96

QUESTION #2: The formula \( \bar{X} \pm 1.96s/\sqrt{n} \) most closely describes which of the following \( (n = 100) \):
   (a) A 90% C. I. for the sample mean  
   (b) A 95% C. I. for the sample mean  
   (c) A 90% C. I. for the population mean  
   (d) A 95% C. I. for the population mean  

QUESTION #3: Suppose I tell you that a 90% confidence interval for the population mean is from -4 to 8. What is the sample mean? Hint: the sample mean is in the center of the confidence interval.
   (a) 3  
   (b) 2  
   (c) 0  
   (d) Can’t determine from the information, but it is between -4 and 8.

QUESTION #4: When people say that they tested a hypothesis at an \( \alpha \) of 0.05, what do they mean?
   (a) They used a 5% confidence interval  
   (b) They had a 5% chance of accepting the hypothesis if it is true.  
   (c) They had a 5% chance of rejecting the hypothesis if it is true  
   (d) They had a 95% chance of rejecting the hypothesis if it is false  
   (e) They had a 5% chance of rejecting the hypothesis if it is false.

QUESTION #5: Suppose I have a sample of size \( n = 9 \) and I found that the sample mean is 3 and the sample standard deviation is 4. Now, suppose that a computer hacker got into my system and changed one of the observations from 3 to 27. Which of the following is most likely to happen to my numbers?
   (a) \( \bar{X} \) increases, \( s \) increases.  
   (b) \( \bar{X} \) decreases, \( s \) increases.  
   (c) \( \bar{X} \) increases, \( s \) decreases.  
   (d) \( \bar{X} \) decreases, \( s \) decreases.

Questions #6–#11 refer to the following scenario. I am working with Professor Taylor of the Biology Department on discovering the number of initial or precursor cells in the MERISTEM of tomato roots. Here is a grossly simplified description of the experiment. In each experiment, Professor Taylor marks cells, which after marking eventually causes part of the root to turn blue (expresses GUS). What he does is to look at each root and measure \( X \), the inverse of the fraction of the root which is blue.

Now, the key choice is whether there are 2 or 3 initial cells in the meristem. If there are 2 initial cells then we expect \( X \) to have a population mean of 2, while if there are 3 initial cells, we expect \( X \) to have a population mean of 3. The random variation in the observed values is caused by experimental variation and nonuniform cell division.

The idea is to use the data and a statistical analysis to discriminate between the two possibilities. In particular, I am interested in testing whether the hypothesis that there are two precursor cells is supported by the data.
While the experiment has yet to be performed, I'm attaching MYSTAT output which describes what I think will happen. You will need to use this MYSTAT output to answer Questions #6–#11.

**QUESTION #6:** What percentage of the sample has $X$ less than 2.1?

(a) .241  
(b) .207  
(c) .233  
(d) .200

**QUESTION #7:** If you were to form a 90% confidence interval for the population mean, what critical value would you use?

(a) 2.05  
(b) 1.31  
(c) 1.96  
(d) 1.70  
(e) 1.645

**QUESTION #8:** Suppose the answer to Question #7 is 2.0. What would be your 90% confidence interval for the population mean of $X$.

(a) [2.9701, 3.3817]  
(b) [9.7774, 2.650]  
(c) [2.9737, 3.3781]  
(d) [3.1383, 3.2135]  
(e) [2.9701, 3.2135]

**QUESTION #9:** Suppose your 95% confidence interval for the population mean is (2.2, 3.7). What conclusion is most correct?

(a) There is at most a 5% chance that there are 2 initial cells.  
(b) There is at least a 5% chance that there are 2 initial cells.  
(c) There is at most a 5% chance that there are 3 initial cells.  
(d) There is at most a 95% chance that there are 3 initial cells.  
(e) There is exactly a 10% chance that there are 2 initial cells.

**QUESTION #10:** Consider the interval in Question #9. Suppose I add to these data a root which is all blue, so that $X = 1$. This is an outlier, being much too low. Which is most correct, if you were to redo the confidence interval?

(a) You are more likely to conclude that the data are able to reject the hypothesis that there are 2 initial cells.  
(b) You are less likely to conclude that the data are able to reject the hypothesis that there are 2 initial cells.  
(c) The sample mean of $X$ and the sample standard deviation of $X$ will increase.  
(d) The sample mean of $X$ and the sample standard deviation of $X$ will decrease.

**QUESTION #11:** Suppose now that the 95% confidence interval is from (2.1, 4.1). If you wanted to conclude that the hypothesis of 2 initial cells is probably wrong, what should you do?

(a) Change to a 90% interval and increase $n$.  
(b) Change to a 99% interval and increase $n$.  
(c) Change to a 90% interval and decrease $n$.  
(d) Change to a 99% interval and decrease $n$.  
(e) You already have enough information to make this conclusion.

**QUESTION #12:** Suppose you take a sample of size $n = 36$ and observe that in your sample, 30% are left handed. Find a 99% confidence interval for the fraction of the population which is left handed.

(a) [.103, .430]  
(b) [.183, .783]  
(c) [.255, .315]  
(d) [.150, .450]
QUESTION #13: Take a sample of size 25 and suppose that the population proportion equals .5. What is the chance that the sample fraction exceeds .395?
(a) .1587
(b) .1469
(c) .8531
(d) 8413
(e) .8508

QUESTION #14: Suppose I'm interested in the percentage of passengers on American Airlines who are under 25 years old. American claims that 40% of its passengers are under 25. You've taken a completely random sample at DFW of 100 people, and observed that 30 of them were under 25 years of age. Suppose that a 90% C.I. for the population percentage of American Airlines passengers who are under 25 is from [.23, .37]. What conclusion do you make? READ CAREFULLY!
(a) The % of passengers younger than 25 years old is between 23% and 37%.
(b) The chance is only 10% that between 23% and 37% of American's passengers are younger than 25 years old.
(c) American's claim is wrong because a smaller percentage in the sample are under 25 years old than they have claimed.
(d) The chance is 10% (or less) that American's claim is correct.

QUESTION #15: Assuming that the balance in my checking account is approximately normally distributed, use the following information to construct a 90% confidence interval for the true average balance in my account. The sample mean is 104.1667, the sample variance is 810.1395 and the sample size is 6.
(a) [46.6715, 161.6619]
(b) [80.0944, 127.6390]
(c) [81.3331, 126.8003]
(d) [85.0044, 123.3289]
(e) [80.6944, 126.8003]

QUESTION #16: I am worried about pesticide levels in the Brazos River, and any decision I make should be right with 90% confidence. A sample of 200 fish from the Brazos River has been taken. Eighty-two of the fish have an unacceptable pesticide level in their blood. Compute a 90% confidence interval for the true proportion of fish with excessive pesticide levels.
(a) (-.3991, 1.2191)
(b) (.3528, .4672)
(c) (.4080, .4120)
(d) (.3418, .4782)

QUESTION #17: Suppose that my interval in the above problem is [.42, .63]. What do you conclude?
(a) The sample proportion exceeds 42%
(b) The population proportion exceeds 42%
(c) The sample proportion exceeds 42% with 90% confidence.
(d) The population proportion exceeds 42% with 90% confidence.

QUESTION #18: Suppose that you are interested in knowing whether men and women are equally likely to handle family finances if both work outside the home. You take a sample of size 100, and find that the sample proportion of men who handle family finances is 63%, with a 95% error bound of .11. What do you conclude?
(a) You can reject the null hypothesis that \( \pi = .63 \), and conclude with 95% confidence that men are more likely to handle family finances.
(b) You can reject the null hypothesis that \( \pi = .50 \), and conclude with 95% confidence that at least 63% of men handle family finances.
(c) You can reject the null hypothesis that $\pi = .50$, and conclude with 95% confidence that at least 52% of men handle family finances.
(d) You can reject the null hypothesis that $\pi = .63$, and conclude with 95% confidence that at least 52% of men handle family finances.

**QUESTION #19:** Suppose that I have a normally distributed population with values $X$ which are normally distributed with population mean 10, and you are told that $\Pr(X > 20) = .30$. Now suppose that I take a random sample of size 16 and want to compute $\Pr(\bar{X} > 20)$. Which is true? You can do this without trying to compute the probability, and even though I have not told you the value of the population standard deviation $\sigma$.

(a) $\Pr(\bar{X} > 20) = .30$
(b) $\Pr(\bar{X} > 20) < .30$
(c) $\Pr(\bar{X} > 20) > .30$

Questions #20–#23 relate to the following problem. You are working for Dumpstuf, Inc., a maker of chemicals for the pharmaceutical industry. The company’s processes leave a residue which is released into the Wichita River. The EPA has issued a regulation that Dumpstuf’s residue is allowed to be .5 parts per billion. They insist on a significance level of $\alpha = 0.05$. You have gone out and taken a sample, and found that the sample residues you took average .7 parts per billion, based on a sample of size 18.

**QUESTION #20:** What is the type I error here?

(a) .05
(b) .95
(c) Saying that you have met the EPA regulation when you haven’t
(d) Saying that you have met the EPA regulation when you have
(e) Saying that you have not met EPA the regulation when you have

**QUESTION #21:** Suppose that your whole career depends on your experiment providing no evidence that Dumpstuf is clearly exceeding the EPA requirement. What should you do?

(a) Take a larger sample size and hire a lawyer to convince the EPA to use $\alpha = .01$.
(b) Take a smaller sample size and hire a lawyer to convince the EPA to use $\alpha = .01$.
(c) Take a larger sample size and hire a lawyer to convince the EPA to use $\alpha = .10$.
(d) Take a smaller sample size and hire a lawyer to convince the EPA to use $\alpha = .10$.

**QUESTION #22:** Remember that your sample mean was $\bar{X} = .7$, and you are using $\alpha = .05$. Suppose that your sample mean has a 95% error bound of .17. What do you conclude?

(a) The sample mean exceeds .53, and hence Dumpstuf is exceeding the EPA limit.
(b) The population mean exceeds .53, and hence Dumpstuf is exceeding the EPA limit.
(c) The population mean exceeds .53 with 95% probability, and hence Dumpstuf is exceeding the EPA limit
(d) The sample mean exceeds .36 with 95% probability, and hence Dumpstuf is exceeding the EPA limit.
(e) The population mean exceeds .36 with 95% probability, and hence Dumpstuf has not been shown to be exceeding the EPA limit.

**QUESTION #23:** Now look at the Dumpstuf problem from the EPA’s perspective. They want to design a decision rule so that the Type I error rate is at an acceptable level. They are testing $H_0: \mu = .5$. Their decision rule is to reject the hypothesis and fine Dumpstuf if in a sample of size 4, the sample mean exceeds 0.65. The population standard deviation is known to be $\sigma = .20$. What is the Type I error rate?

(a) .05
(b) Rejecting $H_0$ when it is true
(c) .0382
(d) .0087
(e) .0668
QUESTION #24: In the New York State rate case, I found a sample mean overcharge by New York State of $\bar{X} = $200, and a 95% confidence interval of [141.20, 258.8]. What did the judge conclude?

(a) New York State owed $141.20, because the federal government had to eat the cost of sampling, since the federal government chose the sample size.

(b) New York State owed the sample mean, $200, because that is the best estimate that the sample can provide.

(c) The Federal government owed the sample mean, $200, because that is the best estimate that the sample can provide.

(d) The Federal government owed $141.20, because it had to eat the cost of sampling, since the federal government chose the sample size.

QUESTION #25: In trying to solve problems, what is the very first step in the flow chart, and what is the second step?

(a) Step #1 = Look up the right formula. Step #2 = is it a question about a proportion or mean?

(b) Step #1 = is it a question about a proportion or mean? Step #2 = Look up the right formula.

(c) Step #1 = is it a question about a probability or data? Step #2 = is it a question about a proportion or mean?

(d) Step #1 = is it a question about a proportion or mean? Step #2 = is it a question about a probability or data?
MYSTAT OUTPUT FOR CELL DATA

COUNT
N OF CASES 29
MINIMUM 1.2000
MAXIMUM 5.7000
MEAN 3.1759
STANDARD DEV 1.0891

BOX PLOT OF VARIABLE: COUNT, N = 29

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
</tr>
</tbody>
</table>

STEM AND LEAF PLOT OF VARIABLE: COUNT, N = 29
MINIMUM IS: 1.2
LOWER HINGE IS: 2.5
MEDIAN IS: 3.0
UPPER HINGE IS: 4.0
MAXIMUM IS: 5.7

1 | 24
1 | 7789
2 | 1
2 H 5556899
3 M 001444
3 | 6
4 H 011233
4 | 9
5 | 
5 | 7
1 = (a)
2 = (d)
3 = (b)
4 = (c)
5 = (a)
6 = (b)
7 = (d)
8 = (c)
9 = (a)
10 = (b)
11 = (e)
12 = (e)
13 = (c)
14 = (d)
15 = (b)
16 = (b)
17 = (d)
18 = (c)
19 = (b)
20 = (c)
21 = (b)
22 = (c)
23 = (e)
24 = (a)
25 = (d)
QUESTION #1: If I want to form a 99% confidence interval for the population slope in a regression line and \( n = 21 \), what is the critical value?
(a) 2.86  
(b) 2.54  
(c) 2.85  
(d) 2.58  
(e) 1.96

QUESTION #2: The formula \( b \pm 1.96 s_b / \sqrt{n} \) most closely describes which of the following (\( n = 100 \)):
(a) A 90% C.I. for the sample mean  
(b) A 95% C.I. for the sample mean  
(c) A 90% C.I. for the population mean  
(d) A 95% C.I. for the population mean  
(e) Incorrect use of a formula to compute a confidence interval.

QUESTION #3: When people say that they tested a hypothesis at an \( \alpha \) of 0.05, and did not reject it, what do they mean?
(a) The \( p \)-value is less than 0.05.  
(b) The \( p \)-value exceeds 0.05.  
(c) No range can be given.

Questions #4–#10 refer to the least squares plots in Figure 11.4 labelled as SCHEMATIC PLOTS and the MYSTAT output labelled SCHEMATIC OUTPUT.

QUESTION #4: Which of the plots is most in keeping with the schematic output?
(a) Plot (a)  
(b) Plot (b)  
(c) Plot (c)

QUESTION #5: What is the most informative correct interpretation of plot (a)?
(a) For each population, the higher the value of X, the higher the value of Y.  
(b) The rates of change are the same in each population, and hence the differences between the two populations depend on the value of X.  
(c) The rates of change are the same in each population, and hence the differences between the two populations do not depend on the value of X.  
(d) The rates of change are different between the populations, and one has uniformly higher values of the population least squares line, but the differences depend on the value of X.  
(e) The rates of change are different between the populations, but neither population has uniformly larger values of the population least squares line.

QUESTION #6: What is the most informative correct interpretation of plot (c)?
(a) For each population, the higher the Value of X, the higher the value of Y.  
(b) The rates of change are the same in each population, and hence the differences between the two populations depend on the value of X.  
(c) The rates of change are the same in each population, and hence the differences
between the two populations do not depend on the value of \( X \).

(d) The rates of change are different between the populations, and one has uniformly higher values of the population least squares line, but the differences depend on the value of \( X \).

(e) The rates of change are different between the populations, but neither population has uniformly larger values of the population least squares line.

**QUESTION #7:** For POPULATION #2 in the schematic output, what do the p-values mean?

(a) No matter which of the three kinds of confidence intervals you do, you would reject the hypothesis that the intercept equals zero and fail to reject the hypothesis that the slope equals zero.

(b) No matter which of the three kinds of confidence intervals you do, you would fail to reject the hypothesis that the intercept equals zero and fail to reject the hypothesis that the slope equals zero.

(c) No matter which of the three kinds of confidence intervals you do, you would fail to reject the hypothesis that the intercept equals zero and reject the hypothesis that the slope equals zero.

(d) No matter which of the three kinds of confidence intervals you do, you would accept the hypothesis that the intercept equals zero and reject the hypothesis that the slope equals zero.

(e) Without computing all the confidence intervals, you are unable to determine the results of hypothesis tests from the information given.

**QUESTION #8:** For POPULATION #1 in the schematic output, find a confidence interval for the population slope which has a 1% chance of being wrong.

(a) 3 \pm 1.29
(b) 3 \pm 0.98
(c) 3 \pm 1.44
(d) 3 \pm 1.05
(e) 3 \pm 0.32

**QUESTION #9:** In POPULATION #2, in computing a prediction interval for a new response at \( X^* = 15 \), you have to compute \( \sum_{i=1}^{n} (X_i - \bar{X})^2 \). What is its value?

(a) 323
(b) 684
(c) 17
(d) 25
(e) 5491

**QUESTION #10:** For POPULATION #2, find a 95% prediction interval for a new response when \( X^* = 15 \).

(a) 45.001 \pm .707
(b) 45.001 \pm .692
(c) 45.001 \pm .239
(d) 45.001 \pm 2.394
(e) 15.000 \pm .757

Questions #11–#19 refer to the following scenario. As part of your Ph.D. thesis in Public Health, you are investigating whether regions with an oversupply of doctors have a higher rate of C-sections for pregnant women. To investigate this, you study two cities, TOONTOWN and MEDTOWN.

In Toontown, 25% of children are delivered by C-section, based on a sample of 100. Of the C-sections, the average cost was $3,000 with a sample standard deviation of $300.
In Medtown, 50% of children are delivered by C-section, based on a sample of 100. Of the C-sections, the average cost was $3,400 with a sample standard deviation of $400.

**QUESTION #11:** What question can the dollar figure data answer?
(a) Whether medical costs are about the same in Toontown and Medtown.
(b) Whether delivery costs are about the same in Toontown and Medtown.
(c) Whether the rates of C-sections are about the same in Toontown and Medtown.
(d) Whether delivery costs for C-sections are about the same in Toontown and Medtown.

**QUESTION #12:** When comparing rates of C-sections in the two towns, what is a Type I error?
(a) 5%
(b) Saying that the rates are nearly the same when they are the same.
(c) Saying that the rates are nearly the same when they are different.
(d) Saying that the rates are different when they are the same.
(e) Saying that the rates are different when they are different.

**QUESTION #13:** If you want to make an error only 1% of the time, find an error bound on the difference of sample fractions of C-sections in the two towns.
(a) .171
(b) .130
(c) .134
(d) .177

**QUESTION #14:** Suppose your answer to Question #13 is 0.10. What do you conclude?
(a) At 99% confidence, you can conclude that Medtown has at least 25% more C-sections, and that the p-value is less than 0.01.
(b) At 99% confidence, you can conclude that Toontown has at least 15% more C-sections, and that the p-value exceeds 0.01.
(c) At 99% confidence, you can conclude that Medtown has at least 15% more C-sections, and that the p-value is no more than 0.01.
(d) At 99% confidence, you can conclude that Medtown has at least 15% more C-sections, and that the p-value exceeds 0.01.
(e) At 99% confidence, you can conclude that Medtown has at least 15% more C-sections, and that the p-value is no more than 0.01.

**QUESTION #15:** Suppose your answer to Question #13 is 0.30. What do you conclude, at 90% confidence?
(a) Medtown has more C-sections.
(b) Toontown has more C-sections.
(c) Toontown and Medtown have the same number of C-sections.
(d) Medtown has not been shown to have more C-sections.

**QUESTION #16:** Suppose your answer to Question #13 is 0.30. Your advisor is convinced that there really is a difference in the population rates and tells you that you cannot graduate until you find it, i.e., use statistics to find a statistically significant effect. What do you do?
(a) Do a 10% level test and take smaller sample sizes.
(b) Do a 10% level test and take larger sample sizes.
(c) Do a 1% level test and take larger sample sizes.
(d) Do a 1% level test and take smaller sample sizes.

**QUESTION #17:** Now take up the question of whether the costs of C-sections are the same in the two towns.

**NOTE NOTE NOTE**

The experiment changes, so that the sample size in Toontown is now 10 C-sections, and the sample size in Medtown is 15 C-sections.
If you are willing to be wrong 5% of the time, what is your error bound on the difference in sample means?

(a) 274.87  
(b) 307.74  
(c) 297.93  
(d) 392.18  
(e) 314.65

QUESTION #18: In Question #17, suppose that your error bound is $250. What do you conclude?

(a) At 95% confidence, you cannot conclude that the costs are different, Medtown costs at least $150 more, and the p-value is less than 5%.  
(b) At 95% confidence, you can conclude that the costs are different, Medtown costs at least $400 more, and the p-value is less than 5%.  
(c) At 95% confidence, you cannot conclude that the costs are different, Medtown costs at least $150 more, and the p-value is greater than 5%.  
(d) At 95% confidence, you can conclude that the costs are different, Medtown costs at least $150 more, and the p-value is greater than 5%.  
(e) At 95% confidence, you can conclude that the costs are different, Medtown costs at least $150 more, and the p-value is less than 5%.

QUESTION #19: In Question #17, suppose that your error bound is $650. What do you conclude?

(a) At 95% confidence, you can conclude that the costs are the same.  
(b) At 95% confidence, you can conclude that the costs are different.  
(c) At 95% confidence, you cannot conclude that the costs are different.

Questions #21–#27 refer to the MYSTAT output and a regression plot of the logarithm of LDL (LOGLDL) on the logarithm of Cholesterol (LOGCHOL), for those with and without Coronary Heart Disease (CHD).

QUESTION #21: Concerning those patients without CHD, which is correct?

(a) The p-value for the intercept is 0.000, and all three confidence intervals would conclude that the population intercept is nonzero.  
(b) The p-value for the intercept is 0.000, and at least one of the confidence intervals would conclude that the population intercept being zero cannot be ruled out.  
(c) The p-value for the intercept is 0.018, and all three confidence intervals would conclude that the population intercept is nonzero.  
(d) The p-value for the intercept is 0.018, and at least one of the confidence intervals would conclude that the population intercept is zero.
would conclude that the population intercept being zero cannot be ruled out.

QUESTION #22: If I want to decide whether the two population slopes are the different, and I want to make a Type I error only 10% of the time, which is the correct procedure?
(a) Note that the difference in the sample slopes is 0.174, the error bound is .153, hence we can reject the hypothesis that the two populations have equal slopes.
(b) Note that the difference in the sample slopes is 0.174, the error bound is .595, hence we can reject the hypothesis that the two populations have equal slopes.
(c) Note that the difference in the sample slopes is 0.174, the error bound is .183, hence we cannot reject the hypothesis that the two populations have equal slopes.
(d) Note that the difference in the sample slopes is 0.919, the error bound is .153, hence we can reject the hypothesis that the two populations have equal slopes.
(e) Note that the difference in the sample slopes is 0.033, the error bound is .153, hence we cannot reject the hypothesis that the two populations have equal slopes.

QUESTION #23: Continuing on with the previous question, suppose that the difference in the sample slopes is 0.20, and that the 90%, 95% and 99% error bounds are .17, .23 and .28, respectively. Which is true?
(a) The p-value is less than 0.01.
(b) The p-value is between 0.01 and 0.05.
(c) The p-value is between 0.05 and 0.10.
(d) The p-value exceeds 0.10.

QUESTION #24: I want to compare two persons each of whom have a LOGCHOL value of 5.4, on the basis of what I predict their LOGLDL will be. Which is most appropriate?
(a) The predicted value for LOGLDL is higher for a person with CHD than for a person without CHD.
(b) The predicted value for LOGLDL is lower for a person with CHD than for a person without CHD.

QUESTION #25: The purpose of a prediction interval as described in class is to:
(a) Be able to predict a person's LOGLDL from his LOGCHOL
(b) Give a range for the actual LOGLDL in the population with a given LOGCHOL
(c) Be able to predict a person's LOGCHOL from his LOGLDL
(d) Give a range for the actual LOGCHOL in the population with a given LOGLDL.

QUESTION #26: In a comparison of LDL between those patients with and without CHD, we note that those with CHD have a higher sample mean than those who do not have CHD. What is the p-value for testing the hypothesis that the two population means are equal?
(a) The p-value is less than 0.01.
(b) The p-value is between 0.01 and 0.05.
(c) The p-value is between 0.05 and 0.10.
(d) The p-value exceeds 0.10.

QUESTION #27: Consider the two regression plots, and notice that in the top plot, there is an observation which appears to be an outlier, while in the bottom plot, there is an observation which has a very large value of LOGCHOL but seems to fit the data pretty well. What would happen if I were to delete these observations:
(a) The slope in the top plot (without CHD) would decrease, and the standard deviation of LOGCHOL for the bottom plot (with CHD) would decrease.
(b) The slope in the top plot (without CHD) would decrease, and the standard deviation of LOGCHOL for the bottom plot (with CHD) would increase.
(c) The slope in the top plot (without CHD) would increase, and the standard deviation of LOGCHOL for the bottom plot (with CHD) would increase.
(d) The slope in the top plot (without CHD) would increase, and the standard deviation of LOGCHOL for the bottom plot (with CHD) would decrease.
## MYSTAT OUTPUT FOR SCHEMATIC PLOTS, QUESTIONS #4–#10

### POPULATION #1

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>N OF CASES</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>MEAN</td>
<td>10</td>
<td>51</td>
</tr>
<tr>
<td>STANDARD DEV</td>
<td>6</td>
<td>17</td>
</tr>
</tbody>
</table>

DEP VAR: Y N: 20 MULTIPLE R: 1.000 SQUARED MULTIPLE R: 1.000
ADJUSTED SQUARED MULTIPLE R: 1.000 STANDARD ERROR OF ESTIMATE: 0.009

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>STD ERROR</th>
<th>STD COEF</th>
<th>TOLERANCE</th>
<th>T</th>
<th>P(2 TAIL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>19.992</td>
<td>0.004</td>
<td>0.000</td>
<td>.49E+04</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>3.000</td>
<td>0.500</td>
<td>1.000</td>
<td>.89E+04</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

ANALYSIS OF VARIANCE

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM-OF-SQUARES</th>
<th>DF</th>
<th>MEAN-SQUARE</th>
<th>F-RATIO</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGRESSION</td>
<td>5985.741</td>
<td>1</td>
<td>5985.741</td>
<td>.788603</td>
<td>0.000</td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>0.001</td>
<td>18</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### POPULATION #2

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>N OF CASES</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>MEAN</td>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>STANDARD DEV</td>
<td>6</td>
<td>18</td>
</tr>
</tbody>
</table>

DEP VAR: Y N: 20 MULTIPLE R: 1.000 SQUARED MULTIPLE R: 1.000
ADJUSTED SQUARED MULTIPLE R: 1.000 STANDARD ERROR OF ESTIMATE: 0.010

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>STD ERROR</th>
<th>STD COEF</th>
<th>TOLERANCE</th>
<th>T</th>
<th>P(2 TAIL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>0.001</td>
<td>0.004</td>
<td>0.000</td>
<td>.165</td>
<td>0.871</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>3.000</td>
<td>0.300</td>
<td>1.000</td>
<td>.89E+04</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

ANALYSIS OF VARIANCE

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>SUM-OF-SQUARES</th>
<th>DF</th>
<th>MEAN-SQUARE</th>
<th>F-RATIO</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGRESSION</td>
<td>5986.186</td>
<td>1</td>
<td>5986.186</td>
<td>620.544</td>
<td>0.000</td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>1.80</td>
<td>18</td>
<td>0.100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 11.1. Patients With CHD

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Err</th>
<th>Std. Coef</th>
<th>T</th>
<th>P(2 Tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.853</td>
<td>0.314</td>
<td>0.000</td>
<td>-5.908</td>
<td>0.000</td>
</tr>
<tr>
<td>LogChol</td>
<td>1.268</td>
<td>0.058</td>
<td>0.874</td>
<td>21.998</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-Squares</th>
<th>DF</th>
<th>Mean-Square</th>
<th>F-Ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>8.678</td>
<td>1</td>
<td>8.678</td>
<td>483.894</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>2.672</td>
<td>149</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 11.2. Patients Without CHD

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Err</th>
<th>Std. Coef</th>
<th>T</th>
<th>P(2 Tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.934</td>
<td>0.393</td>
<td>0.000</td>
<td>-2.377</td>
<td>0.018</td>
</tr>
<tr>
<td>LogChol</td>
<td>1.094</td>
<td>0.073</td>
<td>0.740</td>
<td>14.998</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-Squares</th>
<th>DF</th>
<th>Mean-Square</th>
<th>F-Ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>7.629</td>
<td>1</td>
<td>7.629</td>
<td>224.949</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>6.308</td>
<td>186</td>
<td>0.034</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Answer Key for Exam #3, 1993**

1=(a)  
2=(e)  
3=(b)  
4=(a)  
5=(c)  
6=(e)
Figure 11.4. SCHEMATIC PLOTS FOR QUESTIONS #4–#10
Figure 11.5. \textit{PLOTS FOR QUESTION \#20}
Figure 11.6. SCATTER PLOTS FOR QUESTIONS #21-#27

LOGCHOL for Patients without CHD

LOGCHOL for Patients with CHD
7=(c) 
8=(c) 
9=(b) 
10=(b) 
11=(d) 
12=(d) 
13=(a) 
14=(e) 
15=(d) 
16=(b) 
17=(b) 
18=(e) 
19=(c) 
20=(c) 
21=(d) 
22=(a) 
23=(c) 
24=(a) 
25=(b) 
26=(a) 
27=(d)
Questions #1–#3 relate to the following sample:

-1 3 5 8 12 6 10 13

**Question #1**  What is the sample median?
(a) 10
(b) 6
(c) 8
(d) 7

**Question #2**  What is the sample MAD?
(a) 3
(b) 3.5
(c) 1
(d) 4

**Question #3**  What is the sample 75th percentile?
(a) 6
(b) 8
(c) 10
(d) 11

**Question #4**  Suppose I have a sample from a population, and I observe $\overline{X} = 5$, $s = 6$. I am going to add ONE outlier to the data set. Consider the following three scenarios:

(i) $\overline{X} = 9$, $s = 9$
(ii) $\overline{X} = 2$, $s = 8$
(iii) $\overline{X} = 2$, $s = 4$

Describe what happens to cause these scenarios.
(a) Large positive outlier in (i), large negative outlier in (ii), large negative outlier in (iii).
(b) Large positive outlier in (i), (ii) is implausible, and (iii) is implausible.
(c) Large positive outlier in (i), large negative outlier in (ii), and (iii) is implausible.
(d) (i) is implausible, large negative outlier in (ii), large negative outlier in (iii).

**Question #5**  In comparing histograms (d) and (e) in Figure 12.1, which is most correct:
(a) The former has bigger mean, bigger variance.
(b) The former has bigger mean, about equal variance.
(c) The former has biggerer mean, smaller variance.
(d) The former has smaller mean, bigger variance.
(e) The former has smaller mean, about equal variance.
**Question #6**  Refer to the box plot pictures in Figure 12.2. In comparing plots (b) and (d), which is most nearly correct?
(a) The former has bigger mean, bigger variance.
(b) The former has about the same median, smaller variance.
(c) The former has bigger median, about equal variance.
(d) The former has about the same mean, smaller variance.
(e) The former has smaller median, about equal variance.

**Question #7**  Refer to the box plots in Figure 12.2. In plot (c), what percent of the data in the sample are less than 12?
(a) 0%
(b) This box plot cannot tell you this information.
(c) 25%
(d) 10%

**Question #8**  In least squares, we are fitting lines to data. What is the most correct meaning of these lines?
(a) They estimate the mean of \( Y \) in the population.
(b) They estimate the mean of \( Y \) for the data.
(c) They estimate the mean of \( Y \) in the data for given values of \( X \).
(d) They estimate the value of \( Y \) in the data for a given value of \( X \).
(e) They estimate the mean value of \( Y \) in the population for a given value of \( X \).

**Question #9**  Using Z-scores, what is the probability that \( Z \) exceeds \(-2.13\) and is less than \(-0.11\)?
(a) .4728
(b) .4562
(c) .0166
(d) .4396

**Question #10**  It is known that, when a standard artificial heart valve is implanted in a sheep and removed a year later, on average (among all sheep) it will be 30% calcified. A new heart valve has been developed, and we want to know whether it prevents calcification, i.e., is better than the old model. A sample of 30 sheep were given the new type of valve, and after a year of use the valves were removed. The mean percent calcification was 25, with a standard deviation of 6. If I want to be 90% confident, which statement is most accurate? Note that I have rounded the numbers; choose the answer closest to the actual values. The answers below refer to conclusions about the population.
(a) One can conclude that the new valve is better, since it has mean percent of calcification ranging from 23.2 to 26.8.
(b) One can conclude that the new valve is better, since it has percent of calcification ranging from 15.1 to 34.9.
(c) One cannot conclude that the new valve is better, since it has percent of calcification ranging from 15.1 to 34.9.
(d) One can conclude that the new valve is better, since it has mean percent of calcification ranging from 15.1 to 34.9.
(e) One cannot conclude that the new valve is better, since it has mean percent of calcification ranging from 23.2 to 26.8.

**Question #11**  Consider the heart valve example once again. We have 30 sheep on the new valve as before. However, in statistical science it is always a good idea to get a set of controls, i.e., in this case, a set of sheep who are concurrently given the old type of valve. There were 20 such control sheep in the experiment, and we observed that their standard deviation was 5 and their mean was 29. What can we conclude from these data? Be 95% confident.
(a) The new valves are still better than the old ones, because they are less calcified on average by at least 3.7. However, one should be worried that the the mean of the 20 sheep given the old valve is less than 30.
The new valves are still better than the old ones, because they are less calcified on average by at least 3.7.
(c) The new valves are still better than the old ones, because they are less calcified on average by at least 0.9.
(d) The new valves are still better than the old ones, because they are less calcified on average by at least 0.9. However, one should be worried that the the mean of the 20 sheep given the old valve is less than 30.

Question #12 Refer to the plots in Figure 12.3. Ignoring the labels on the plots, and suppose I tell you that the intercept is 3 and the slope is −1.0. Which of the plots is most likely to have been associated with these data?

(a) Least squares plot (a).
(b) Least squares plot (b).
(c) Least squares plot (c).
(d) Least squares plot (d).
(e) Not enough information to determine.

Questions #13–#24 refer to a project from the spring of 1993. Students went to fields located both near (NEARFAR=1) and far (NEARFAR=2) from a highway, and studied the heights (HEIGHT) and number of red petals (REDPETAL) of bluebonnets. They sampled 100 plants from each type of location. I have attached a fair amount of Stataquest output and graphs to the end of the exam.

PLEASE NOTE: directly at the end of the questions for the exam you will find a display of stem-leaf plots.

Question #13 Consider the plants near the highways. What is the slope and standard error of the slope for the regression of number of red petals on plant height?

(a) slope = 3.68, standard error of slope = 1.07.
(b) slope = 0.39, standard error of slope = 0.035.
(c) slope = 4.62, standard error of slope = 0.860.
(d) slope = −0.52, standard error of slope = 0.025.
(e) slope = −0.52, standard error of slope = 1.74.

Question #14 Consider the fields which are far from the highway. What is your percentile estimate of the percentage of plants in the population which have heights which are greater than or equal to 40?

(a) 11%
(b) 3%
(c) 5%
(d) 10%

Question #15 Now suppose that the heights are approximately normally distributed. For fields which are far from a highway, using the Stataquest output, estimate the percentage of plants in the population which have heights which are greater than or equal to 40.

(a) .1100
(b) .0823
(c) .9515
(d) .0485
(e) .9177
Question #16  There are four box plots. What do they tell you?
(a) Heights are comparable across the two populations, but the number of red petals is greater and more variable for fields near a highway.
(b) Heights are clearly greater in fields far from a highway, but the number of red petals is smaller and more variable for fields far away from a highway.
(c) Heights are comparable across the two populations, but the number of red petals is greater and more variable for fields far away from a highway.
(d) Heights are comparable across the two populations, but the number of red petals is greater and less variable for fields far away from a highway.

Question #17  To one decimal place, the average heights are the same. What does this tell you?
(a) The numbers are meaningless.
(b) Bluebonnets near highways grow taller when they are far from highways.
(c) The population of all bluebonnets near highways has the same mean heights as the population of bluebonnets far from highways.
(d) The population of all bluebonnets near highways could have the same or could plausibly have different mean heights from the population of bluebonnets far from highways.

Question #18  In this problem, you’ll want to be 90% confident about being correct. Which statement is most correct?
(a) For plants far from highways, the population mean number of red petals ranges from 4.4 to 5.2.
(b) For plants far from highways, the sample mean number of red petals ranges from 4.4 to 5.2.
(c) For plants far from highways, the population mean number of red petals ranges from 4.3 to 5.3.
(d) For plants far from highways, the population mean number of red petals ranges from 5 to 9.1.

Question #19  Study the histograms and box plots for the number of red petals. You’ll note that there appear to be some outliers. Which statement is more correct?
(a) There are possibly as many as three outliers. Deleting them would tend to increase the difference between the two estimated population means. Thus, any reasonable analysis probably will not be affected by deleting these observations.
(b) There are possibly as many as three outliers. Deleting them would tend to lower the difference between the two estimated population means. However, the differences are so great anyway in the sample medians that any reasonable analysis probably will not be affected by deleting these observations.
(c) There are possibly as many as two outliers. Deleting them would tend to increase the difference between the two estimated population means. Thus, any reasonable analysis probably will not be affected by deleting these observations.
(d) There are possibly as many as two outliers. Deleting them would tend to lower the difference between the two estimated population means. However, the differences are so great anyway in the sample medians that any reasonable analysis probably will not be affected by deleting these observations.

Question #20  There are four scatterplots, with least squares lines added in. Using only the scatterplots, and not the regression output, which statement is most correct?
(a) For the regression of heights on the number of red petals, the population slope is positive near highways but negative far from highways.
(b) For the regression of heights on the number of red petals, the sample slope is positive near highways but negative far from highways.
(c) For the regression of the number of red petals on heights, the sample slope is negative near highways but positive far from highways.
(d) For the regression of the number of red petals on heights, the population slope is negative near highways but positive far from highways.

Question #21  Using the regression output, which statement is most correct? Assume you want
to be 95\% confident. This concerns the regression of the number of red petals on plant height, among plants which are far from a highway.

(a) The population slope lies between $-0.24$ to $0.86$, so there is an indication of a positive relationship.
(b) The population slope lies between $-0.03$ to $0.108$, so there is no indication of a positive relationship.
(c) The population slope lies between $-0.03$ to $0.108$, so there is an indication of a positive relationship.
(d) The population slope lies between $-0.24$ to $0.86$, so there is no indication of a positive relationship.
(e) The slope of the least squares line is positive, so there is a positive relationship.

**Question #22** Assume you want a 95\% chance of being correct. Let’s consider the number of red petals. Which statement is most correct.

(a) Plants far from highways have, on average, between 1.4 and 2.0 more red petals.
(b) Plants far from highways have, on average, between 1.3 and 2.1 more red petals.
(c) Plants far from highways have, on average, between 1.5 and 1.9 more red petals.
(d) Plants far from highways have, on average, between 1.1 and 2.3 more red petals.

**Question #23** Consider plants near the highway, and the regression of the number of red petals on heights. Which statement is most correct, if you want to be 95\% certain?

(a) The sample slope is negative, being no larger than $-0.003$.
(b) The population slope is negative, being no larger than $-0.003$.
(c) The population slope is negative, being no larger than $0.003$.
(d) The population slope is negative, being no larger than $-0.049$.

**Question #24** Consider plants near the highway, and the regression of the number of red petals on heights. Which statement is most correct, for a plant which has a height of 10?

(a) Our best guess is that the plant will have 31.54 petals.
(b) Our best guess is that the plant will have 25.2 petals.
(c) Our best guess is that the plant will have 46.1 petals.
(d) Our best guess is that the plant will have 4.1 petals.
1 = (d)
2 = (b)
3 = (d)
4 = (c)
5 = (c)
6 = (b)
7 = (a)
8 = (e)
9 = (d)
10 = (a)
11 = (c)
12 = (c)
13 = (d)
14 = (a)
15 = (b)
16 = (c)
17 = (d)
18 = (a)
19 = (b)
20 = (c)
21 = (b)
22 = (d)
23 = (b)
24 = (d)
Here are the stem-leaf plots for the bluebonnet data.

**Stem and leaf plot for HEIGHT, NEARFAR = 1**
1 | 56689
2 | 0001122234444555555555666666777778888999999
3 | 00000000001111111112223333333334445555666677777778888889
4 | 00134

**Stem and leaf plot for HEIGHT, NEARFAR = 2**
1 | 246678
2 | 000111222233333444444555556666667777788889999999
3 | 000000001111111112222233333333334445555555566668899
4 | 01111223789
This is a junk page, due to the computer system I use. You can use it as a scratch pad.
Figure 12.1. Sample histograms, 1994 exam #1.
Figure 12.2. Sample box plots, 1994 exam #1.
Figure 12.3. Sample regression plots, 1994 exam #1.
Bluebonnets Near a Highway

REDPETAL = Number of Red Petals
HEIGHT = Height of the plants

### Number of Red Petals

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Obs</th>
<th>Sum of Wgt.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>.5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>3</td>
<td></td>
<td></td>
<td>3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td>3.2</td>
<td></td>
</tr>
</tbody>
</table>

### Plant Height

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Obs</th>
<th>Sum of Wgt.</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>21.5</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>25</td>
<td>18</td>
<td></td>
<td></td>
<td>29.7</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>30</td>
<td></td>
<td></td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>33.5</td>
<td>40</td>
<td></td>
<td></td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>38</td>
<td>41</td>
<td></td>
<td></td>
<td>38.4</td>
<td></td>
</tr>
</tbody>
</table>

### REGRESSION

**Number of obs = 100**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F(1, 98)</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10.1672495</td>
<td>1</td>
<td>10.1</td>
<td>3.35</td>
<td>0.07</td>
</tr>
<tr>
<td>Residual</td>
<td>297.19275</td>
<td>98</td>
<td>3.03</td>
<td>Adj R-square = 0.03</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>307.36</td>
<td>99</td>
<td>3.1066</td>
<td>Root MSE = 1.74</td>
<td></td>
</tr>
</tbody>
</table>

| redpetal | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------|-------|-----------|-------|------|----------------------|
| height   | -0.052 | 0.025     | -5.37 | 0.000 | 2.91                 |
| _cons    | 4.62   | 0.860     | 5.37  | 0.000 | 2.91                 |

### ANOTHER REGRESSION

**Number of obs = 100**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F(1, 98)</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>124.477356</td>
<td>1</td>
<td>124.47</td>
<td>3.35</td>
<td>0.07</td>
</tr>
<tr>
<td>Residual</td>
<td>3638.52264</td>
<td>98</td>
<td>37.12</td>
<td>R-square = 0.03</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3763.00</td>
<td>99</td>
<td>38.01</td>
<td>Root MSE = 6.09</td>
<td></td>
</tr>
</tbody>
</table>

| redpetal | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------|-------|-----------|-------|------|----------------------|
| height   | -0.636 | 0.347     | -1.83 | 0.070 | -1.32               |
| _cons    | 31.6   | 1.23      | 25.703| 0.000 | 29.21               |
Bluebonnets Far from a Highway

REDPETAL = Number of Red Petals
HEIGHT = Height of the plants

Number of Red Petals

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>2</td>
<td>0</td>
<td>Obs</td>
</tr>
<tr>
<td>25%</td>
<td>3</td>
<td>1</td>
<td>Sum of Wgt.</td>
</tr>
<tr>
<td>50%</td>
<td>5</td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>8.5</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Plant Height

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>21</td>
<td>16</td>
<td>Obs</td>
</tr>
<tr>
<td>25%</td>
<td>24</td>
<td>16</td>
<td>Sum of Wgt.</td>
</tr>
<tr>
<td>50%</td>
<td>30</td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>33</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>40.5</td>
<td>47</td>
<td></td>
</tr>
</tbody>
</table>

REGRESSION

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>8.07165506</td>
<td>1</td>
<td>8.07</td>
<td>F( 1, 98) = 1.21</td>
</tr>
<tr>
<td>Residual</td>
<td>656.038345</td>
<td>98</td>
<td>6.69</td>
<td>Prob &gt; F = 0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R-square = 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-square = 0.00</td>
</tr>
<tr>
<td>Total</td>
<td>664.11</td>
<td>99</td>
<td>6.70</td>
<td>Root MSE = 2.58</td>
</tr>
</tbody>
</table>

redpetal | Coef. | Std. Err. | t  | P>|t| | [95% Conf. Interval] |
-----------|-------|------------|----|------|---------------------|
height     | .039  | .035       | 1.098 | 0.275 |                     |
_cons      | 3.68  | 1.07       | 3.414 | 0.001 | 1.54                | 5.82                  |

ANOTHER REGRESSION

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>65.3903602</td>
<td>1</td>
<td>65.39</td>
<td>F( 1, 98) = 1.21</td>
</tr>
<tr>
<td>Residual</td>
<td>5314.71964</td>
<td>98</td>
<td>54.23</td>
<td>Prob &gt; F = 0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R-square = 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-square = 0.0021</td>
</tr>
<tr>
<td>Total</td>
<td>5380.11</td>
<td>99</td>
<td>54.34</td>
<td>Root MSE = 7.36</td>
</tr>
</tbody>
</table>

redpetal | Coef. | Std. Err. | t  | P>|t| | [95% Conf. Interval] |
-----------|-------|------------|----|------|---------------------|
height     | .31   | .28        | 1.098 | 0.275 | -.25                | .88                   |
_cons      | 28.15 | 1.56       | 17.997 | 0.000 | 25.04               | 31.25                 |
This is a junk page, due to the computer system I use. You can use it as a scratch pad.
This is a junk page, due to the computer system I use. You can use it as a scratch pad.
Question #1 Suppose $n = 64$, $\pi = .6$. What’s $\Pr(.30 \leq p \leq .65)$?

(a) .9836  
(b) .9911  
(c) .2698  
(d) .7939  
(e) 1.000
Question #2 The mean height of the population of male Martians is 100 inches. Which is correct?
(a) \( \Pr(100 < X < 120) > \Pr(100 < \bar{X} < 120) \)
(b) \( \Pr(100 < X < 120) < \Pr(100 < \bar{X} < 120) \)
(c) \( \Pr(100 < X < 120) = \Pr(100 < \bar{X} < 120) \)
(d) Cannot determine because the population standard deviation is unknown.

Question #3 Suppose \( n = 4, \mu = 10, \sigma = 10 \). Which is correct?
(a) \( \Pr(\bar{X} \geq 20) = .0228 \) and \( \Pr(X \geq 20) > .0228 \)
Question #4 Suppose you observe the following data: -2, -1, 1, 2, 5. Find a 90% confidence interval for the population mean. Pick the answer closest to the correct one.

(a) -1 to 3
(b) -3.5 to 5.5
(c) -1.6 to 3.6
(d) -1.1 to 3.1
(e) The question makes no sense. We know that the population mean equals 1.

**Question #5** I have spent most of the semester and the project emphasizing the difference between samples and population. Why?

(a) I’ve run out of ideas, and this one is easy to talk about.
(b) It’s a great way to ask tricky multiple choice problems.
(c) I can’t stand the term ‘sample population.’
(d) Who can tell with the Irish Catholics? They’re all irrational anyway, especially those whose families were gun runners during the aftermath of the 1916 Rising.
(e) You don’t know what you’re doing if you don’t know what you are making conclusions about.

**Question #6**  The formula

\[ \bar{X} \pm 3.25s/\sqrt{n} \]

most closely describes which of the following?

(a) A 99% confidence interval for the population mean with \( n = 11 \).
(b) A 99% confidence interval for the sample mean with \( n = 11 \).
(c) A 99% confidence interval for the population mean with \( n = 10 \).
(d) A 99% confidence interval for the sample mean with $n = 10$.
(e) A 99% confidence interval for the population mean with $n = 9$.

**Question #7**  Suppose you take a sample of size $n = 36$ and observe that in your sample, 30% are left handed. Find a 99% confidence interval for the fraction of the population which is left handed.

(a) [0.103, 0.450]
(b) [-0.183, 0.783]
(c) [0.285, 0.315]
(d) [0.150, 0.460]
Question #8  You had a homework assignment in which I asked you to compute a 95% confidence interval for the population mean number of recruits in the Skkena River Sockeye Salmon fishery (there were 28 years of data). You were asked: why is this QUESTION irrelevant for managing the fishery?. Please answer the homework problem, after reading the wording carefully.

(a) Very little can be learned from 28 observations, especially when the numbers are in the 1000's.
(b) If you are managing a fishery, you really want a 99% chance of being right.
(c) The mean does not tell you where the individual observations lie. For this you need regression.
(d) You need to know where both the population mean and median are, just in case the data are skew.

Question #9  A competitor of the McDonald’s Fast Food chain wished to show that the Quarter Pounder does not live up to its name: i.e. to show that it weights less than the asserted 4 ounces. A random sample of 16 Quarter Pounders had an average weight of 3.870 ounces with a standard deviation for the sample of .30 ounces. Did the sample present sufficient evidence to conclude at the 5% level of significance that MacDonalds is skimping?

(a) Yes, the confidence interval is from 3.723 to 4.017.
(b) No, the confidence interval is from 3.710 to 4.030.
(c) No, the confidence interval is from 3.723 to 4.017.
(d) Yes, the confidence interval is from 3.710 to 4.030.
(e) No, the confidence interval is from 3.674 to 4.066.

Question #10  Suppose we want to test \( H_0: \mu = 0 \). As sample size increases, what happens:

(a) The variance of the sample mean \( \bar{X} \) decreases, the length of a 95% confidence interval decreases, and the chance of finding a statistically significant effect decreases.
(b) The variance of the sample mean \( \bar{X} \) increases, the length of a 95% confidence interval increases, and the chance of finding a statistically significant effect increases.
(c) The variance of the sample mean \( \bar{X} \) decreases, the length of a 95% confidence interval decreases, and the chance of finding a statistically significant effect increases.
(d) The variance of the sample mean \( \bar{X} \) decreases, the length of a 95% confidence interval increases, and the chance of finding a statistically significant effect decreases.
(e) The variance of the sample mean \( \bar{X} \) increases, the length of a 95% confidence interval decreases, and the chance of finding a statistically significant effect increases.

Question #11  It's a general folklore result that says that statisticians tend to be much more lefthanded than the general population. In fact, in my previous and current jobs, those statisticians who were born in the U.S. were overwhelmingly left-handed. I'm really interested in knowing whether statisticians are equally likely to be left or right handed. To check on this, I surveyed all the major statistics departments in the U.S., and found that of the 230 statisticians born in the U.S. (not counting me, since I was born in Japan), 90 are left-handed. What do you conclude?

(a) You can reject the null hypothesis that \( \pi = .39 \), and conclude with 95% confidence that statisticians are more likely to be left-handed.
(b) You fail to reject the null hypothesis that \( \pi = .39 \), and conclude with 95% confidence that statisticians are less likely to be left-handed.
(c) You can reject the null hypothesis that \( \pi = .30 \), and conclude with 95% confidence statisticians are less likely to be left-handed.
(d) You can reject the null hypothesis that $\pi = .50$, and conclude with 95% confidence that statisticians are more likely to be left-handed.
(e) You fail to reject the null hypothesis that $\pi = .50$, and conclude with 95% confidence that statisticians could be equally likely to be left-handed.

**Question #12** Suppose I own a 10 acre private bass pond which I rent out for $600 per day. It’s obviously in my best interest to advertise big fish, because that will bring in more customers. Suppose I advertise that the bass average 16 inches in my pond. Since I’m a business, I’m regulated by Texas Fish and Wildlife (TFW), and they ask me to prove my claim. So, having taken STAT 302, I go out and take a random sample of 10 fish, find that the mean is 13 and the 95% confidence interval is from 8 to 18. I tell TFW that my claim is correct, since 16 is in the confidence interval. They object: why?

(a) Actually, there’s nothing to object about. I’ve done everything properly, and the conclusion is correct.
(b) The conclusion is wrong. The average fish length in my pond is 13 inches.
(c) The statistics were OK, but the conclusion is incorrect. Your interval is too large because you took too large a sample size, thereby helping to make sure that 16 is in the confidence interval.
(d) The statistics were OK, but the conclusion is incorrect. Your interval is too large because you took too small a sample size, thereby helping to make sure that 16 is in the confidence interval.

**Question #13** Suppose I own a 10 acre private bass pond which I rent out for $600 per day. Since I’m a business, I’m regulated by Texas Fish and Wildlife (TFW). In my bass pond, I claim that 20% of all the fish are 18 inches or longer; these are called big fish. Suppose I’m correct. TFW forces me to take a sample of 100 fish (pretend they are random) to verify my claim. I’m nervous about this, because my company will go out of business if 15% or less of the sample are big fish. My lawyer tells me not to worry, saying that ‘it’s really unlikely that 15% or less of the sample will be big fish.’ Should I fire the lawyer, and get one who knows a little statistics?

(a) Fire the lawyer on general principles (Shakespeare was right after all, although his solution was a bit more drastic).
(b) Fire the lawyer. There’s a 10.56% chance that your business will go bankrupt.
(c) Fire the lawyer. There’s a 89.44% chance that your business will go bankrupt.
(d) Fire the lawyer. There’s a 90.92% chance that your business will go bankrupt.
(e) Fire the lawyer. There’s a 8.08% chance that your business will go bankrupt.

**Question #14** There is a Stataquest table in the back called *Salinity Values in the Pamlico Sound*. I have measured the daily amount of river water discharged into the Pamlico sound. The intervals are listed there 90%, 95% and 99% confidence intervals. Which order are they in (from bottom to top)?

(a) 90%, 95%, 99%
(b) 95%, 99%, 90%
(c) 90%, 99%, 95%
(d) 99%, 90%, 95%
(e) 99%, 95%, 90%

**Question #15** In the salinity data set, what is the sample standard deviation?

(a) 2.96
(b) .56
(c) 15.68
(d) .02
(e) .11

**Question #16** In the salinity data set, suppose that the middle interval is the 95% interval. What does this mean?

(a) Discharges range from 9.58 to 11.52.
(b) The average discharge is between 9.58 and 11.52.
(c) The chance is 95% that discharges range from 9.58 to 11.52.
(d) The chance is 95% that the average discharge ranges from 9.58 to 11.52.
(e) If you are testing the hypothesis that the average discharge is 10, at the level \( \alpha = 0.05 \), then you would conclude that the average discharge is indeed 10.

**Question #17** There is a Stataquest table in the back called Concho Water Snakes. This was explained in class that previous students took a random sample of 3-year old Concho water snakes. Find a 95% confidence interval for the population proportion of water snakes which are male.

(a) .466 to .820
(b) .180 to .534
(c) .635 to .651
(d) .340 to .365
(e) .164 to 1.122

**Question #18** In the Concho water snake data, it has been conjectured that the average ‘Snout-to-Vent Length’ equals the average of ‘3 times the Tail Length’. This means that I want to know whether the population mean of the difference equals 0. To test this hypothesis, I computed the statistics \( X = \text{Snout-to-Vent Length} - (3 \times \text{Tail Length}) \), and called the variable ‘svlm3t’. At \( \alpha = 0.05 \), what do you think about the claim?

(a) Claim is true, because the confidence intervals includes 0.
(b) Claim could be true, because the confidence intervals includes 0.
(c) Claim is false, because the confidence intervals includes 0.
(d) Cannot tell. It is not clear which hypothesis is being tested.
(e) The relevant information is not given to test this hypothesis.

**Question #19** In the Concho water snake data, it has been conjectured that the average 3-year old snake has a ‘Snout-to-Vent Length’ equal to 475mm. If I want to test this claim with a type I error rate of 1%, what do I conclude? Please note that the computer output lists 95% confidence intervals, which will not solve this problem. I want you to read the output, and then do confidence intervals by hand. A special note: this question is about the variable labelled “svl”.

(a) Claim is false, since the mean is from 487.3 to 520.9.
(b) Claim is false, since the mean is from 491.6 to 516.5.
(c) Claim is false, since the mean is from 488.4 to 519.7.
(d) Claim is true, since the mean is from 487.3 to 520.9.
(e) Claim could be true, since the mean is from 487.3 to 520.9.

Questions #20 to #24 revolve around the following example. Please note that there are two pages of stataquest output.

In the medical products industry, when developing a new hormone assay it is customary to compare your new assay, called the TEST, to the usual assay, called the REFERENCE. The usual method is to take logarithms of the results and study the difference: diffllog = log(TEST) - log(REFERENCE). At the very least, one has to show that the two methods have the same mean, i.e., diffllog has a (population) mean of zero. If this is not true, the TEST method is not allowed.

**Question #20** What do I mean when I say that I want to test the company’s claim that the TEST and REFERENCE methods are the same, with \( \alpha = 0.01 \)?

(a) I’ll use a 1% confidence interval.
(b) I’ll use a 95% confidence interval and reject the hypothesis if 0 is in the interval.
(c) I’ll use a 99% confidence interval and reject the hypothesis if 0 is in the interval.
(d) I’ll use a 99% confidence interval and fail to reject the hypothesis if 0 is in the interval.
(e) I’ll use a 99% confidence interval and accept the hypothesis if 0 is in the interval.

Question #21 Looking at the output, which is most nearly correct?
(a) If we test $H_0: \mu = 0$ at level $\alpha = 0.05$, we fail to reject $H_0$, and conclude that the TEST method can be allowed.
(b) If we test $H_0: \mu = 0$ at level $\alpha = 0.05$, we fail to reject $H_0$, and conclude that the TEST method is exactly the same as the REFERENCE method.
(c) If we test $H_0: \mu = 0$ at level $\alpha = 0.05$, we reject $H_0$, and conclude that the TEST method cannot be allowed.
(d) If we test $H_0: \mu = 0$ at level $\alpha = 0.05$, we reject $H_0$, and conclude that the TEST method cannot be allowed.
(e) If we test $H_0: \mu = 0$ at level $\alpha = 0.05$, we fail to reject $H_0$, and conclude that the TEST method cannot be allowed.

Question #22 Looking at the output for the hormone assay data (where I have cleverly hidden the value of the sample median) and the box plot, which is most nearly correct? (Note to future students: There was a box plot which had the median clearly < 0. The idea was to be able to look at the box plot and tell me that the median was < 0.)
(a) Sample median < 0, sample mean = −.108, sample standard deviation = 0.276, and a 99% confidence interval is from −.187 to −.029.
(b) Sample median cannot be determined to be < 0, sample mean = −.108, sample standard deviation = 0.276, and a 99% confidence interval is from −.187 to −.029.
(c) Sample median < 0, sample mean = −.108, sample standard deviation = 0.276, and a 99% confidence interval is from −.167 to −.048.
(d) Sample median cannot be determined to be < 0, sample mean = −.108, sample standard deviation = 0.276, and a 99% confidence interval is from −.167 to −.048.

Question #23 Looking at the output for the hormone assay data, suppose I add a large positive outlier. What is most likely to occur?
(a) $\bar{x}$ will increase, $s$ will decrease, the confidence interval is more likely to include 0, in which case we would conclude that it’s possible that $\mu = 0$.
(b) $\bar{x}$ will increase, $s$ will increase, the confidence interval is more likely to include 0, in which case we would conclude that it’s possible that $\mu = 0$.
(c) $\bar{x}$ will increase, $s$ will increase, the confidence interval is less likely to include 0, in which case we would conclude that it’s definitely true that $\mu = 0$.
(d) $\bar{x}$ will increase, $s$ will increase, the confidence interval is more likely to include 0, in which case we would conclude that it’s definitely true that $\mu = 0$.
(e) $\bar{x}$ will decrease, $s$ will increase, the confidence interval is more likely to include 0, in which case we would conclude that it’s possible that $\mu = 0$.

Question #24 Consider the hormone assay once again. Suppose you work for the company trying to market the new assay, so it’s in your interest to show that $\mu = 0$, i.e., to have $\mu$ in the confidence interval. Suppose further that the initial results from the experiment were unfavorable to your case. If you are unethical, what would you do?
(a) Throw away the first experiment, redo it with a larger $n$ and set $\alpha = 0.01$.
(b) Throw away the first experiment, redo it with a larger $n$ and set $\alpha = 0.10$.
(c) Throw away the first experiment, redo it with a smaller $n$ and set $\alpha = 0.10$.
(d) Throw away the first experiment, redo it with a smaller $n$ and set $\alpha = 0.01$.
(e) Just say that ‘‘Statistics can lie’’ and hide behind your idiocy.
Question #25  Which of the following is correct? Consult the front page of the exam if you are in doubt.
(a) It seems sort of silly to have to know how to spell Dr. Carol’s last name, but probably he is giving us a gift question as part of a prearranged curve. Of course, his brain was fried by going to t.u.
(b) It seems sort of silly to have to know how to spell Dr. Carrol’s last name, but probably he is giving us a gift question as part of a prearranged curve. Of course, his brain was fried by going to t.u.
(c) It seems sort of silly to have to know how to spell Dr. Caroll’s last name, but probably he is giving us a gift question as part of a prearranged curve. Of course, his brain was fried by going to t.u.
(d) It seems sort of silly to have to know how to spell Dr. Carroll’s last name, but probably he is giving us a gift question as part of a prearranged curve. Of course, his brain was fried by going to t.u.
(1) d
(2) b
(3) e
(4) c
(5) e
(6) c
(7) e
(8) c
(9) b
(10) c
(11) c
(12) d
(13) b
(14) b
(15) a
(16) d
(17) a
(18) b
(19) a
(20) d
(21) d
(22) a
(23) b
(24) d
(25) d
This is a junk page, due to the computer system I use. You can use it as a scratch pad.
CONCHO WATER SNAKES

$svlm_3t = \text{Snout-to-Vent Length} - (3 \times \text{Tail Length})$

$svl = \text{Snout to Vent Length}$

$tail = \text{Tail Length}$

$sex = \text{male}=0, \text{female}=1$

Stem-Leaf Plot for Gender, Males = 0, Females = 1

0 | 0000000000000000
1 | 0000000000

Some Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$svlm_3t$</td>
<td>28</td>
<td>3.607143</td>
<td>52.52889</td>
<td>-62</td>
<td>84</td>
</tr>
<tr>
<td>number</td>
<td>28</td>
<td>70.92857</td>
<td>29.63948</td>
<td>41</td>
<td>114</td>
</tr>
<tr>
<td>snake</td>
<td>28</td>
<td>76.39286</td>
<td>35.38025</td>
<td>3</td>
<td>114</td>
</tr>
<tr>
<td>sex</td>
<td>28</td>
<td>.3571429</td>
<td>.48795</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$svl$</td>
<td>28</td>
<td>504.0714</td>
<td>32.10392</td>
<td>454</td>
<td>586</td>
</tr>
<tr>
<td>age</td>
<td>28</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>tail</td>
<td>28</td>
<td>166.8214</td>
<td>13.40275</td>
<td>138</td>
<td>186</td>
</tr>
</tbody>
</table>

Some Inferential Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$svlm_3t$</td>
<td>28</td>
<td>3.607143</td>
<td>9.927027</td>
<td>-16.76143, 23.97572</td>
</tr>
<tr>
<td>number</td>
<td>28</td>
<td>70.92857</td>
<td>5.6002</td>
<td>59.43791, 82.41923</td>
</tr>
<tr>
<td>snake</td>
<td>28</td>
<td>76.39286</td>
<td>6.688128</td>
<td>62.66995, 90.11576</td>
</tr>
<tr>
<td>sex</td>
<td>28</td>
<td>.3571429</td>
<td>.0922139</td>
<td>.1679356, .5463501</td>
</tr>
<tr>
<td>$svl$</td>
<td>28</td>
<td>504.0714</td>
<td>6.06707</td>
<td>491.6228, 516.52</td>
</tr>
<tr>
<td>age</td>
<td>28</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>tail</td>
<td>28</td>
<td>166.8214</td>
<td>2.532881</td>
<td>161.6244, 172.0185</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$svlm_3t$</td>
<td>28</td>
<td>3.607143</td>
<td>52.52889</td>
</tr>
</tbody>
</table>

Ho: mean - 0 = 0

$t = 0.36$ with 27 d.f.

$Pr > |t| = 0.7192$

95% conf. interval = (-16.7614, 23.9757)
### SALINITY VALUES IN THE PAMLICO SOUND

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[??% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>salinity</td>
<td>28</td>
<td>10.55000</td>
<td>0.5600000</td>
<td>9.380000  11.72000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[??% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>salinity</td>
<td>28</td>
<td>10.55000</td>
<td>0.5600000</td>
<td>9.580000  11.52000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[??% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>salinity</td>
<td>28</td>
<td>10.55000</td>
<td>0.5600000</td>
<td>8.970000  12.12000</td>
</tr>
</tbody>
</table>
HORMONE ASSAY: DOES THE TEST EQUAL THE REFERENCE METHOD?

difflog = Log(Test) - Log(Reference)

Percentiles Smallest
1%  -0.821  -0.821
5%  -0.5470001 -0.7099999
10% -0.4449999 -0.636 Obs 60
25% -0.272  -0.573 Sum of Wgt. 60
50%  %  Mean -0.1080000
75%  .043  .457 Largest Std. Dev. .2760000
90%  .22  .531 Variance .0760000
95%  .405  .536 Skewness .1440000
99%  .588  .588 Kurtosis 3.243000

Variable | Obs  Mean Std. Err. [90% Conf. Interval]
-----------|------|-------|---------------------
difflog | 60  -0.108000  .0290000 -0.1580000 -0.0580000

Variable | Obs  Mean Std. Err. [95% Conf. Interval]
-----------|------|-------|---------------------
difflog | 60  -0.108000  .0290000 -0.1670000 -0.0480000

Variable | Obs  Mean Std. Err. [99% Conf. Interval]
-----------|------|-------|---------------------
difflog | 60  -0.108000  .0290000 -0.1870000 -0.0290000

Variable | Obs  Mean Std. Dev.
-----------|------|---------------------
difflog | 60  -0.1080000 .2760000

Ho: mean - 0 = 0

  t = -3.62 with 59 d.f.
  Pr > |t| = 0.0005
95% conf. interval = (-0.1600, -0.0480)
This is a junk page, due to the computer system I use. You can use it as a scratch pad.
This is a junk page, due to the computer system I use. You can use it as a scratch pad.
Question #1  Suppose that the sample slope is $b = .50$, the sample size is $n = 16$ and the standard error of the slope is $s_b = 0.15$. Which is a 95% confidence interval for the population slope?
(a) .206 to .794  
(b) .178 to .822  
(c) .426 to .574  
(d) .420 to .580  
(e) .330 to .670

Question #2  Suppose your answer to the previous question is [0.01, 0.99]. What do you conclude?
(a) The population slope is positive, the sample slope is positive, and there could be no relationship between $Y$ and $X$. 
(b) The population slope is positive, the sample slope is negative, and there could be no relationship between $Y$ and $X$.  
(c) The population slope is possibly zero, the sample slope is negative, and there is a relationship between $Y$ and $X$.  
(d) The population slope is possibly zero, the sample slope is positive, and there is a relationship between $Y$ and $X$.  
(e) The population slope is positive, the sample slope is positive, and there is a relationship between $Y$ and $X$. 

Question #3  I want to know whether two population proportions are the same. From each population I take a sample of size 20. The sample fractions are .60 and .30. Find a 90% confidence interval for the difference in population proportions.
(a) .0775 to .3225  
(b) .15 to .45  
(c) .054 to .547  
(d) .006 to .594  
(e) .263 to .337

Question #4  I want to know whether two population intercepts are the same or not. I take samples of size 10 from each population, and observe sample intercepts of .350 and -.150, with standard errors of .90 and 1.2. Find a 99% confidence interval for the difference in population intercepts.
(a) -0.72 to 1.72  
(b) -2.44 to 3.44  
(c) -0.89 to 1.89  
(d) -3.88 to 4.88  
(e) -3.73 to 4.73

Question #5  Suppose I am testing whether two population proportions are the same or not, and I find that the $p$-value equals 0.012. Which is the correct conclusion about these proportions?
(a) 90% C.I.: different 95% C.I.: different 99% C.I.: could be same  
(b) 90% C.I.: could be same 95% C.I.: different 99% C.I.: different  
(c) 90% C.I.: different 95% C.I.: different 99% C.I.: different  
(d) 90% C.I.: could be same 95% C.I.: different 99% C.I.: different  
(e) 90% C.I.: different 95% C.I.: could be same 99% C.I.: could be same

Question #6  I have stressed many concepts throughout the course of the semester. Here is a list of possible concepts.
(i) The differences between samples and populations
(ii) The use of the $t$-test statistic is absolutely critical
(iii) The reason one makes probability statements is because of sample-to-sample variability
(iv) One can be 100% confident and still make useful conclusions.
(v) A $p$-value is the chance that a hypothesis is exactly true.
(vi) $p$-values are often cited in scientific journals.
(vii) One needs to employ graphics and statistical confidence intervals together. Graphs help see patterns and outliers, but may be interpreted differently by different people. Statistical confidence intervals may be mislead by outliers.
(viii) Because we have computers, we can avoid consulting with a statistician.

Which concepts have I really stressed?
(a) (i), (ii), (iii), (vi), (vii)
(b) (i), (iii), (iv), (vi), (vii)
(c) (i), (ii), (iii), (vi)
(d) (i), (iii), (vi), (vii)
(e) (i), (iii), (v), (vi), (vii)

Question #7  I fly Continental Airlines a lot to Washington D.C., and I’m absolutely convinced that their planes are more empty going from Washington than going there. To test this, I observed three flights on Boeing 737’s each way. The number of empty seats were as follows:

<table>
<thead>
<tr>
<th>To DC</th>
<th>From DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>10, 20, 30</td>
<td>30, 40, 50</td>
</tr>
</tbody>
</table>

Find a 90% confidence interval for the difference in population means.
(a) -2.67 to 42.67
(b) 2.59 to 37.40
(c) 6.57 to 33.43
(d) 9.09 to 30.96
(e) 5.79 to 34.20

Question #8  Suppose your interval in the previous problem is from -5 to 45. Which is the most correct?
(a) There is not sufficient evidence that my claim is correct, and the $p$-value is < 0.10.
(b) There is sufficient evidence that my claim is correct, and the $p$-value is < 0.10.
(c) There is not sufficient evidence that my claim is correct, and the $p$-value is > 0.10.
(d) There is sufficient evidence that my claim is correct, and the $p$-value is > 0.10.

Question #9  I am interested in knowing whether the population of women in sororities has a higher percentage of very high GPA’s then the population of women who are not members of sororities. To test this, I took a sample of 80 sorority women, of whom 36 had GPA’s which exceed 3.7. I also took a sample of 50 non-sorority women, of whom 23% had GPA’s exceeding 3.7. What conclusion do you reach, at 95% confidence?
(a) Sororities have between 6% and 38% more women with high GPA’s.
(b) Sororities have between 14% and 30% more women with high GPA’s.
(c) Sororities have between -3% and 29% more women with high GPA’s.
(d) Sororities have between 5% and 21% more women with high GPA’s.
(e) Sororities have between -43% and 87% more women with high GPA’s.

Question #10  It has been hypothesized that male undergraduates at A&M are four inches taller than females on average. To test this hypothesis, a random sample of 8 males and 12 females was taken. The sample mean and sample standard deviation of the heights of men in inches was 71 and 3, respectively, while the sample mean and sample
standard deviation of women was 63 and 2, respectively. Does the hypothesis seem reasonable at 90% confidence), why or why not?
(a) Men are taller by between 6.07 to 9.93 inches, on average, and the p-value < .10.
(b) Men are taller by between 6.07 to 9.93 inches, on average, and the p-value > .10.
(c) Men are taller by between 6.17 to 9.83 inches, on average, and the p-value < .10.
(d) Men are taller by between 5.91 to 10.09 inches, on average, and the p-value < .10.
(e) Men are taller by between 5.91 to 10.09 inches, on average, and the p-value > .10.

Question #11  Attached is an aortic stenosis plot (Figure 12.4). There are observed log(aortic value area) (logava) and log(body surface area) (logbsa) for kids with aortic stenosis. There are three lines, all of which are computed from healthy kids (least squares line, along with 95% prediction intervals for a new response). Please note: the lines are computed from the healthy kids, but the data are for kids with aortic stenosis. Which of the following statements is most correct?
(a) An observation falling inside the lines has been correctly classified as having aortic stenosis. A healthy kid with logbsa = 1.0 has a 95% chance of having logava between 0.8 and 1.5.
(b) An observation falling outside the lines has been correctly classified as having aortic stenosis. A healthy kid with logbsa = 1.0 has a 95% chance of having logava between 0.2 and 1.4.
(c) An observation falling inside the lines has been correctly classified as having aortic stenosis. A healthy kid with logbsa = 1.0 has a 95% chance of having logava between 0.2 and 1.4.
(d) An observation falling outside the lines has been correctly classified as having aortic stenosis. A healthy kid with logbsa = 1.0 has a 95% chance of having logava between 0.8 and 1.5.
(e) There is no information in these lines, because the observations are on ill children, while the lines are for normal kids.

Question #12  Refer to the previous example. Which is more correct, if I were to switch from the 95% prediction interval to a 90% prediction interval.
(a) more healthy kids would be called stenotic, and less ill kids would be called stenotic.
(b) less healthy kids would be called stenotic, and more ill kids would be called stenotic.
(c) more healthy kids would be called stenotic, and more ill kids would be called stenotic.
(d) less healthy kids would be called stenotic, and less ill kids would be called stenotic.

Question #13  Figure 12.5 gives you schemetics of a number of different lines. I want to be 95% confident. At this level of confidence, if the p-value for whether the intercept is zero equals 0.025 and the p-value for whether the slope is zero equals 0.078, which of the lines is most correct?
(a) Line a
(b) Line b
(c) Line c
(d) Line d

Question #14  Consider the schematic plots given in Figure 12.6. In each plot, one line is labelled ‘‘Population 1’’ and the other is labelled ‘‘Population 2’’. The plots are labelled as ‘‘Plot (a)’’, ‘‘Plot (b)’’, ‘‘Plot (c)’’, ‘‘Plot (d)’’. Suppose that a 95% confidence interval for the difference in intercepts between population 2
and population 1 \((\alpha_2 - \alpha_1)\) is from 0.6 to 1.3, and that a 95% confidence interval for the difference in slopes between population 2 and population 1 \((\beta_2 - \beta_1)\) is from \(-2\) to \(-1\). Please read this carefully, because the order matters. Which schematic plot is the best representation of this data?

(a) a
(b) b
(c) c
(d) d

**Question #15** Consider Figure 12.6 once again. What is the most informative correct interpretation of plot (b)?

(a) For each population, the higher the value of \(X\), the higher the value of \(Y\).
(b) The rates of change are different between the populations, and one has uniformly higher values of the population least squares line, but the differences depend on the value of \(X\).
(c) The rates of change are different between the populations, but neither population has uniformly larger values of the population least squares line.
(d) The rates of change are the same in each population, and hence the differences between the two populations depend on the value of \(X\).
(e) The rates of change are the same in each population, and hence the differences between the two populations do not depend on the value of \(X\).

**Question #16** Consider Figure 12.7. These are the boxplots of the tail lengths of male and female concho water snakes. Note that there is a single outlier among the males. What will happen if this outlier is removed?

(a) The mean of the males will increase, the standard deviation of the males will decrease, the length of the confidence interval for the difference of two population means will decrease and the net effect will be to make the two populations appear more different.
(b) The mean of the males will increase, the standard deviation of the males will decrease, the length of the confidence interval for the difference of two population means will increase and the net effect will be to make the two populations appear more different.
(c) The mean of the males will decrease, the standard deviation of the males will decrease, the length of the confidence interval for the difference of two population means will decrease and the net effect will be to make the two populations appear more different.
(d) The mean of the males will increase, the standard deviation of the males will decrease, the length of the confidence interval for the difference of two population means will decrease and the net effect will be to make the two populations appear less different.
(e) The mean of the males will increase, the standard deviation of the males will decrease, the length of the confidence interval for the difference of two population means will increase and the net effect will be to make the two populations appear less different.

**Question #17** Consider Figure 12.8. These are the outcomes of the analyses of the endometrial cancer data set, where I am comparing women with cancer to women without cancer with respect to their body mass index, bread intake, total calories and percentage of calories from fat. Note that I have left the results for total calories blank. Which is more correct, if I want to be 99% certain of my results? *Note: the computer output lists 90% confidence intervals.*

(a) I should conclude that the population of cancer victims have lower body mass, and that no statistically significant differences between cancer victims and non-cancers are seen for breads and % calories from fat.
(b) I should conclude that the population of cancer victims have higher body mass, and that no statistically significant differences between cancer victims
and non-cancers are seen for breads and % calories from fat.
(c) I should conclude that cancer victims have higher body mass in terms of sample means, but that no statistically significant differences between cancer victims and non-cancers are seen for body mass, breads and % calories from fat.
(d) I should conclude that cancer victims have lower body mass in terms of sample means, but that no statistically significant differences between cancer victims and non-cancers are seen for body mass, breads and % calories from fat.
(e) Cannot determine, since the 99% confidence intervals are not listed.

Question #18 Consider Figure 12.8 once again. These are the outcomes of the analyses of the endometrial cancer data set, where I am comparing women with cancer to women without cancer with respect to their body mass index, bread intake, total calories and percentage of calories from fat. Note that I have left the results for total calories blank. With respect to total calories, at the 90% level of confidence, which is more correct?
(a) The population of cancer victims has higher total calories on average, because the confidence interval for the difference in population means ranges from 23.4 to 107.5.
(b) The population of cancer victims has lower total calories on average, because the confidence interval for the difference in population means ranges from 23.4 to 107.5.
(c) The population of cancer victims has higher total calories on average, because the confidence interval for the difference in population means ranges from -3.8 to 134.7.
(d) The population of cancer victims has higher total calories on average, because body mass index is statistically significant, and body mass is closely related to total caloric intake.
(e) The population of cancer victims cannot be said to have higher total calories on average, because the confidence interval for the difference in population means ranges from -3.8 to 134.7.

Questions #19-#25 concern the endometrial cancer data set covered in class on 11/22/94. The variables that are measured are body mass index (called bmi in the computer output), total calories (calsall in the computer output) and a binary variable “smoking”, which equals 0 if the person is a non-smoker or ex-smoker, and equals 1 if the person currently smokes. Also, I have data for endometrial cancer status. Summary statistics for body mass index, total calories and endometrial cancer status are given in Figure 12.9, where they are split up between Non/Ex Smokers and Current Smokers. Regression output is given in Figure 12.10.

Question #19 It has been hypothesized that smokers are thinner than non-smokers. This is of course a factor that induces people to smoke, in a culture which values thinness. Using body mass index, which is smaller for thin people, is there evidence that smokers are thinner?

(a) At 95% confidence, there is a statistically significant difference in body masses, with current smokers being between 2.20 and 5.8 smaller on this scale. In addition, even if we did a 99% confidence interval, we would still conclude that current smokers are thinner.
(b) At 95% confidence, there is a statistically significant difference in body masses, with current smokers being between 2.20 and 5.8 smaller on this scale. However, if we did a 99% confidence interval, we would not be able to conclude that current smokers are thinner.
(c) At 95% confidence, there is a statistically significant difference in body masses, but contrary to the common wisdom, current smokers are between 2.20 and 5.8 bigger on this scale. In addition, even if we did a 99% confidence interval, we would still conclude that current smokers are bigger.
(d) At 95% confidence, there is no discernible difference in the two groups.
The difference in population means is between -85.7 and 163.7. The p--value exceeds 0.01.
(e) At 95% confidence, there is no discernible difference in the two groups. The difference in population means is between -85.7 and 163.7. The p--value is less than 0.01.

**Question #20**

Suppose in the previous question that you computed 90%, 95% and 99% confidence intervals for the differences in population means, and these became $5 \pm 4$, $5 \pm 6$ and $5 \pm 8$, respectively. Which is correct?
(a) The p--value is less than 0.01.
(b) The p--value is between 0.01 and 0.05.
(c) The p--value is between 0.05 and 0.10.
(d) The p--value exceeds 0.10.

**Question #21**

For most types of heart disease or cancer, smoking is a very strong risk factor. Heart disease, lung cancer and breast cancer are the major killers, and smoking has been demonstrated to be a cause of these diseases. Endometrial cancer is fairly rare. Studying Figure 12.9, what is the effect of smoking on endometrial cancer?

(a) With 95% confidence, smokers have at least 10.7% higher percentages of endometrial cancer. At 99% confidence, there is still a statistically significant difference between the two populations.
(b) With 95% confidence, smokers have at least 33.5% lower percentages of endometrial cancer. At 99% confidence, there is still a statistically significant difference between the two populations.
(c) With 95% confidence, smokers have at least 33.5% higher percentages of endometrial cancer. At 99% confidence, there is still a statistically significant difference between the two populations.
(d) With 95% confidence, smokers have at least 10.7% lower percentages of endometrial cancer. At 99% confidence, there is still a statistically significant difference between the two populations.
(e) With 95% confidence, smokers have at least 10.7% lower percentages of endometrial cancer. At 99% confidence, there is still a statistically significant difference between the two populations.

**Question #22**

In studying the regressions of body mass index on total calories, which statement is most nearly correct, at 95% confidence?

(a) Total calories are clearly related to body mass index among non/ex--smokers, but there is no evidence of such a relationship among smokers. Among non/ex--smokers, the relationship is fairly weak, because the regression slope is so small.
(b) Total calories are clearly related to body mass index among non/ex--smokers, but there is no evidence of such a relationship among smokers. Among non/ex--smokers, the relationship is fairly weak, because the correlation is 0.187.
(c) Total calories are clearly related to body mass index among non/ex--smokers, but there is no evidence of such a relationship among smokers. Among non/ex--smokers, the relationship is fairly weak, because the correlation is 0.035.
(d) Total calories are clearly related to body mass index among non/ex--smokers and also among current smokers. Among non/ex--smokers, the relationship is fairly weak, because the correlation is 0.187.
(e) Total calories are clearly related to body mass index among non/ex--smokers, but there is no evidence of such a relationship among smokers. Among non/ex--smokers, the relationship is pretty strong, because the correlation is 0.187.

**Question #23**

Suppose I compare the average non/ex-smoking women who eats 2,000 calories per day with the average smoker who eats 2,000 calories per day. What are my best estimates of how these women will differ in terms of average body mass?

(a) The average non/ex--smoker has a body mass of 29.222, average smoker has 25.206
(b) The average non/ex--smoker has a body mass of 25.008, average smoker has
22.780.
(c) The average non/ex-smoker has a body mass of 50017.043, average smoker has 45561.621.
(d) The average non/ex-smoker has a body mass of 31.107, average smoker has 26.393.
(e) The average non/ex-smoker has a body mass of 2.449, average smoker has 5.447.

**Question #24**  Give a 90% normal range for body mass if a women is a non/ex-smoker who eats 3,000 calories per day.
(a) 32.57 to 35.89
(b) 28.81 to 39.66
(c) -83.09 to 151.55
(d) 31.65 to 36.81
(e) 32.25 to 36.21

**Question #25**  Is there evidence that the population slopes of the regressions of body mass on total calories are different between the two populations? Be 90% confident. Pick the answers which is most in keeping with the analyses we have been doing in STAT--302.
(a) Yes. The slope for the non/ex-smokers is different from zero, and the slope for smokers may equal zero.
(b) Yes. The slopes differ by between 0.00095 to 0.00153.
(c) No. The slopes differ by between -0.00215 to 0.00463.
(d) No. The slopes differ by between -0.07742 to 0.07991.
(e) No. The slopes differ by between -0.00161 to 0.00409.
ANSWER KEY FOR EXAM #3, 1994

1= (b)
2= (e)
3= (c)
4= (d)
5= (a)
6= (d)
7= (b)
8= (c)
9= (a)
10= (a)
11= (d)
12= (c)
13= (c)
14= (c)
15= (e)
16= (a)
17= (b)
18= (e)
19= (a)
20= (c)
21= (e)
22= (b)
23= (d)
24= (a)
25= (e)
Figure 12.4. Aortic Stenosis Data.
Figure 12.5. Schematics of four regression lines.
Figure 12.6. Schematic plots for two regression lines.
Figure 12.7. *Concho Water Snake Tail Lengths.*
ENDOMETRIAL CANCER: NUTRIENT VARIABLES AS CAUSES OF CANCER

********************
BODY MASS INDEX
********************

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonCancer</td>
<td>273</td>
<td>26.42606</td>
<td>5.9657</td>
</tr>
<tr>
<td>Cancer</td>
<td>327</td>
<td>30.54949</td>
<td>9.69331</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>28.67333</td>
<td></td>
</tr>
</tbody>
</table>

Ho: mean(x) - mean(y) = 0 (assuming unequal variances)
Pr > |t| = 0.0000
90% conf. interval = (-6.1806, -3.0063)

********************
BREADS
********************

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonCancer</td>
<td>273</td>
<td>13.92711</td>
<td>6.821976</td>
</tr>
<tr>
<td>Cancer</td>
<td>327</td>
<td>14.363</td>
<td>7.213678</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>14.1667</td>
<td></td>
</tr>
</tbody>
</table>

Ho: mean(x) - mean(y) = 0 (assuming unequal variances)
Pr > |t| = 0.4450
90% conf. interval = (-1.3817, 0.5099)

********************
TOTAL CALORIES
********************

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonCancer</td>
<td>273</td>
<td>1340.919</td>
<td>498.1018</td>
</tr>
<tr>
<td>Cancer</td>
<td>327</td>
<td>1406.38</td>
<td>530.9756</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>1376.596</td>
<td></td>
</tr>
</tbody>
</table>

Ho: mean(x) - mean(y) = 0 (assuming unequal variances)
Pr > |t| = ????
90% conf. interval = (????, ????)

********************
PERCENT CALORIES FROM FAT
********************

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonCancer</td>
<td>273</td>
<td>37.44212</td>
<td>8.642712</td>
</tr>
<tr>
<td>Cancer</td>
<td>327</td>
<td>38.0833</td>
<td>7.613704</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>37.79702</td>
<td></td>
</tr>
</tbody>
</table>

Ho: mean(x) - mean(y) = 0 (assuming unequal variances)
Pr > |t| = 0.3326
90% conf. interval = (-1.7676, 0.4652)

Figure 12.8. Endometrial Cancer: Nutrient Intake Comparisons.
# Endometrial Cancer: Smoking and Nutrient Variables

## Comparing Body Mass Index

<table>
<thead>
<tr>
<th>Smoking?</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non/Ex</td>
<td>518</td>
<td>29.2212</td>
<td>8.39248</td>
</tr>
<tr>
<td>Current</td>
<td>82</td>
<td>26.20686</td>
<td>7.63206</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>28.67333</td>
<td></td>
</tr>
</tbody>
</table>

\[ H_0: \text{mean}(x) - \text{mean}(y) = 0 \] (assuming unequal variances)

\[ t = 4.39 \text{ with } 116 \text{ d.f.} \]

\[ Pr > |t| = 0.0000 \]

95% confidence interval = (2.2635, 5.6276)

## Comparing Total Calories

<table>
<thead>
<tr>
<th>Smoking?</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non/Ex</td>
<td>518</td>
<td>1311.924</td>
<td>514.849</td>
</tr>
<tr>
<td>Current</td>
<td>82</td>
<td>1342.003</td>
<td>531.6306</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>1376.696</td>
<td></td>
</tr>
</tbody>
</table>

\[ H_0: \text{mean}(x) - \text{mean}(y) = 0 \] (assuming unequal variances)

\[ t = 0.62 \text{ with } 107 \text{ d.f.} \]

\[ Pr > |t| = 0.5367 \]

95% confidence interval = (-66.7096, 163.6973)

## Comparing Proportions of Endometrial Cancer

<table>
<thead>
<tr>
<th>Smoking?</th>
<th>Obs</th>
<th>Proportion</th>
<th>Std. Error of Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non/Ex</td>
<td>518</td>
<td>.5762886</td>
<td>.0497508</td>
</tr>
<tr>
<td>Current</td>
<td>82</td>
<td>.3636888</td>
<td>.04810470</td>
</tr>
<tr>
<td>combined</td>
<td>600</td>
<td>.446</td>
<td></td>
</tr>
</tbody>
</table>

\[ H_0: \text{mean}(x) - \text{mean}(y) = 0 \] (assuming unequal variances)

\[ t = 3.86 \text{ with } 111 \text{ d.f.} \]

\[ Pr > |t| = 0.0002 \]

95% confidence interval = (0.1079, 0.3364)

---

Figure 12.9. Endometrial Cancer: Comparisons for body mass index and total calories for the two smoking groups.

\{fig:wuwq5\}
ENDOMETRIAL CANCER:
SMOKING AND NUTRIENT VARIABLES REGRESSIONS

---------------------------------------------
NONSMOKERS OR EX-SMOKERS
---------------------------------------------

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>518</td>
<td>29.2212</td>
<td>8.39248</td>
<td>15.662</td>
<td>64.92</td>
</tr>
<tr>
<td>calall</td>
<td>518</td>
<td>1381.924</td>
<td>514.840</td>
<td>480.1</td>
<td>3427.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 518</th>
<th>F(1, 516) = 18.71</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1274.05732</td>
<td>1</td>
<td>1274.05732</td>
<td>Prob &gt; F = 0.0000</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>36429.6373</td>
<td>516</td>
<td>69.1000722</td>
<td>R-square = 0.0350</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>36442.6946</td>
<td>517</td>
<td>70.4326781</td>
<td>Adj R-square = 0.0331</td>
<td></td>
</tr>
</tbody>
</table>

---------------------------------------------
CURRENT SMOKERS
---------------------------------------------

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>82</td>
<td>26.20666</td>
<td>7.68206</td>
<td>14.042</td>
<td>78.593</td>
</tr>
<tr>
<td>calall</td>
<td>82</td>
<td>1342.93</td>
<td>631.6306</td>
<td>259.7</td>
<td>2775.1</td>
</tr>
</tbody>
</table>

| Source | SS  | df  | MS  | Number of obs = 82 | F(1, 80) = 1.30 |
|--------|-----|-----|-----|-------------------|-----------------
| Model  | 74.6667019 | 1 | 74.6667019 | Prob > F = 0.2669 |
| Residual | 4666.49798 | 80 | 57.278035 | R-square = 0.0160 |
| Total  | 4666.49798 | 81 | 67.4876294 | Adj R-square = 0.0027 |

---------------------------------------------
| bmi | Coef. | Std. Err. | t | P>|t| | 95% Conf. Interval |
|-----|-------|-----------|---|------|-------------------------------|
| calall | 0.062063 | 0.01852 | 1.142 | 0.267 | -0.01362 | 0.00466 |
| cons  | 23.78084 | 2.382973 | 9.978 | 0.000 | 23.23756 | 23.32407 |

Figure 12.10. Endometrial Cancer: Regression output for the two smoking groups. {fig:wuoq6}
CHAPTER 13

UNC EXAMS

These are exam questions and examples from when I was at the University of North Carolina. Most of the questions were irrelevant, so I have picked only those that are of direct interest.
(1) Suppose that nails are an average of 3' long, with a standard deviation of 1/10''. If nails are normally distributed, what is the chance that one of them is more than 3 1/5' long?

(2) Suppose that we are going to decide if a set of records has to be completely audited. A good set of records has 5% error, while a bad set of records has 15% error. The records are fully audited if 10% or more of them from a sample are found to be in error.

a) A sample size of \( n = 20 \) is taken. What is the chance of fully auditing a set of good records?

(3) I am a pollster working for Lou Harris and it is October 31. I want to look at the Texas gubernatorial race between Clements and White. If I take a sample of size \( n = 100 \) and Clements actually has 55% of the vote in the total population, what is the chance that Clements will receive more than 50% of the votes in the poll?

(4) Suppose I have 25 normally distributed observations, each with mean 2 and variance 1. What is the chance that their sample mean is between 1.85 and 1.95?

(5) Suppose I sample 25 normally distributed variables and observe a sample mean of 13.3 and a sample standard deviation of 4.4. Construct a 95% confidence interval for the population mean.

(6) Suppose you are a marketing executive for a firm that sells a regional brand of beer (e.g., Iron City), and you are considering expanding into North Carolina. However, you are convinced that you can only make an adequate return on your investment if you can capture more than 5% of the market and you are particularly concerned that you might expand and lose lots of money. You decide to sample 200 individuals and see how many of them say they will switch from their present brand to yours. Take \( \alpha = .05 \). State the null hypothesis, the alternative hypothesis, and the decision rule you would use.

(7) Suppose I am interested in the percentage of ag econ majors who enjoy STAT 303. I would like my estimate of this percentage to be in error by no more than 6%, with 93% confidence. If I have no idea what the true percentage is, what sample size should I use?

(8) Suppose the diameter of bolts is normally distributed with mean \( \mu \) and unknown standard deviation. You take a sample of size 3 bolts, the outcomes of which are 10.5, 11, 11.5

a) Construct a 99% confidence interval for the true mean.
b) Explain in 20 words or less what your answer to part (a) means.
c) Test at \( \alpha = .05 \):

\[ H_0 : \mu = 9 \]
\[ H_A : \mu \neq 9 \]
(9) Suppose I am interested in trying to estimate the percentage of right-handed golfers in the golf population. I want to construct an 90% confidence interval for this percentage, and I want the sample proportion to be within 0.04 of the correct figure, with probability at least 90%. If my best guess is that at least 70% of all golfers are right-handed, what sample size should I take?

(10) You are working as an environmental quality consultant to the EPA. They ask you to evaluate Big Smoke Industries’ plant in Waxahachie, Texas. Big Smoke has claimed that their plant falls within acceptable limits, putting no more per day than 30 mg of benzene into the river, on average. You have been asked to see if Big Smoke is meeting this upper limit, and you are of course interested in detecting whether they are equalling the limit. You take a sample of eighteen days, and observe an average of 35 mg, with a sample variance of 300. Assume benzene amounts are normally distributed. Test Big Smoke’s claim with a Type I error of five percent.

(11) Let $X$ be normally distributed. Compute
   a) $P(X \geq 3)$ if the mean is 3 and the standard deviation is 4.
   b) $P(X \leq -3)$ if the mean is 3 and the variance is 9.

(12) You have been called in to design a sampling plan for the quality control of the production of boots by Carroll Brothers Boot Shop, whose motto is ‘‘We skin ’em, you wear ’em, no sissies welcome’’. They produce boots in lots of size 10,000, and they are willing to ship the lot if 6% or less of the boots are defective. You recommend a sampling plan in which 500 boots are selected; brother Joe and his son Brent try them out, and if more than 20 are defective, the shipment is inspected item by item; otherwise, you ship. What is the chance that a shipment with 6% defectiveness is shipped?

(13) Carroll Sisters Accountants (motto: no boots allowed) is based in Dallas, Texas. They are auditing a rather large firm for its sales, and they look at the error

$$ X = \text{reported sale} - \text{audited sale} $$

They believe that these errors are normally distributed with mean $0$ and standard deviation $120$.
   a) They take a sample of size $N = 400$. What is the probability that the sample mean error exceeds $9.00$?
   b) What sample size should they take so that the chance the sample mean error exceeds $9.00$ is 0.025?

(14) Consider the following data which are the weight losses by people enrolled in an experimental diet program.

$$ 6 \quad 12 \quad 8 \quad 10 $$

Assume that weight losses are normally distributed, but you know neither the mean nor the standard deviation in the population.
a) What is the sample mean of these numbers?
b) What is the sample standard deviation of these numbers?
c) Pretend that the sample mean is 6.0 and the sample standard deviation is 2.0. These may be wrong but pretend they are correct! Construct a 99% confidence interval for the mean weight loss.
d) Using the figures in part (c), test with $\alpha = 0.05$ the null hypothesis that the average weight loss was 5.0 pounds versus the alternative that the average weight loss was different from 5.0 pounds.

\[ \alpha \] (15) Suppose I am interested in studying the fraction of American statisticians who are left handed. I have a prior guess that this value is no more than 25%.

a) What sample size should I take if I want to construct a 90% confidence interval and have the sample fraction of left handers be in error by no more than 5%?
b) Suppose I take a sample size of $N = 200$ and observe in my sample that 42 are left handed. Construct a 95% confidence interval for the true fraction of left handers in the population.

\[ \alpha \] (16) You are developing a new method to prevent the calcification of artificial heart valves, and you will be experimenting with sheep. In the past, the average calcification score has been $\mu_0 = 46$ with a standard deviation of $\sigma = 5$. You wish to test the hypothesis $H_0 : \mu = 46$

$H_A : \mu \neq 46$

using $\alpha = 0.05$. It will be an important advance in the treatment of children needing artificial heart valves if you can lower the calcification score to an average of $\mu_1 = 42$. You may assume calcification scores are normally distributed.

a) Suppose I take a sample of size 10 and find that the Type II error probability $\beta$ would be 0.18. Explain what this means in English. I will only read your first 25 words!

b) If I want the sample mean to be in error by no more than 4 with 95% confidence, what sample size should I take?

\[ \alpha \] (17) Let $X$ be normally distributed with mean $\mu = 3$ and standard deviation $\sigma = 2$.

Compute

a) $\Pr(X < 3)$.
b) $\Pr(1 < X < 4)$.
c) $\Pr(X < 1)$.

\[ \alpha \] (18) You are working as a quality control supervisor for Carroll Brothers Boots of Wichita Falls, Texas, whose motto is ‘‘We skin ’em, you wear ’em, no sissies welcome’’. You are in charge of the size nine men’s boots. No boot can be made to be exactly size nine, and as part of your job you measure the difference between the actual size and
size nine. This difference is a normally distributed random variable with mean 8 mm and a standard deviation of 3 mm. You decide that each hour you will take a sample of 16 boots and measure their sample mean. Your decision rule is to shut down the operation and call it out of control if the sample mean you observe exceeds 8.75 mm.

a) Assume the process is in control and the true mean is really 8 mm. Having taken your quality assurance sample of size 16, what is the probability that you will shut the operation down?

b) Now suppose that the true mean has shifted to 10 mm. What is the probability that your decision rule will shut the operation down?

髫 (19) Take a sample of size 16 from Normal (μ = 3, σ = 4). Compute

a) Pr(\(\bar{X} \leq 3\))

b) Pr(2 ≤ \(\bar{X} \leq 5\)).

髫 (20) Suppose that I take 3 observations from a normally distributed population with unknown mean and variance. I observe:

5, 7, 9

a) Compute the sample mean of these numbers.

b) Compute the sample standard deviation of these numbers.

c) Now pretend \(\bar{X} = 6\), \(s = 3\). These are not correct, but pretend they are anyway. Find a 90% confidence interval for the true population mean.

髫 (21) Suppose I want to know the average height of undergraduate females, and I know that heights are normally distributed with standard deviation 3”.

a) If I want to take a sample so that I can be 95% confident that the sample mean will be within 1/4” of the true mean, what sample size should I take?

b) Suppose I sample \(N = 25\) females, and find that \(\bar{X} = 66”\). Find a 90% confidence interval for the mean.

髫 (22) In 1960-1970, the statistics department gave an average grade GPA of 2.60. There is some controversy in the department as to whether we have become easier since I arrived in 1974. So, we took a random sample of 2500 grades since 1974 and found that the mean was 2.63. We announced that there was a significant difference. In terms of this example, distinguish between the lay and statistical definition of significance.

髫 (23) Piedmont Airlines announces that it has been its experience that it can expect Flight 233 from Washington to Raleigh to be 80% full. Distinguish between the lay and statistical meaning of the word ‘‘expect’’ here.

髫 (24) I have been called in to consult with ‘‘Good Cash Machines’’ to help them with their auditing. They know that sales records are normally distributed with standard deviation 16. This book or claimed average sale is $83, and they want to know whether this is correct. So, they take a sample of size 64 and reject the claim if the sample mean exceeds $86.
a) What is the Type I error of their decision procedure?

I want to find out by a sample survey what percentage of A&M students favor divestment from South Africa. I have no idea of the correct figure, but want to take a sample size large enough that the sample fraction is within .03 of the true figure with 90% confidence. What sample size should I use?
CHAPTER 14

VARIOUS PROBLEMS

These problems come from my previous classes. Some will be used in class, while others can be used by you in preparing for exams.
2. Spier Industries, a high technology firm, requires constant infusions of new capital for research and development and growth. As the parent company of a wholly owned subsidiary, PPD Drilling, Spier views PPD Drilling as a ‘‘cash cow’’ -- a chief source of cash from operations. PPD Drilling pays back fifty percent of after-tax profit as cash dividend and is designed to serve Spier's cash flow needs. The annual after-tax profit probability distribution for PPD Drilling is assessed to have a mean of $16 million and a standard deviation of $3.4 million. If after-tax profit falls below $10 million, given its burdensome dividend policy, PPD will have to go to the market and borrow to maintain its plant and equipment. What is the probability that additional debt financing would be required?

4. A marketing survey at Apple Computer Corporation reveals that eight percent of a sample of 2000 households plan to consider buying a home computer during the Christmas holidays. Assuming that the sample can be viewed as an unbiased random sample of the 25 million household target market population, calculate a .99 confidence interval estimate of the total unit demand for home computers among the household population.

5. Your employer, a large Los Angeles amusement park, is planning its children's TV programming budget. With the high price TV spots on Saturday morning, you, as marketing manager, want to document your decision if prime Saturday time is the airtime you decide to buy. It is desired to estimate (1) the percentage of viewing children who watch Saturday morning TV within three percentage points at the 95 percent confidence level; (2) the mean viewing time of Saturday morning TV within 15 minutes at the 95 percent confidence level. Previous studies suggest a mean of three hours with an hour standard deviation for Saturday morning viewing.

a) What sample size of children needs to be interviewed to achieve both objectives?

b) If the setup overhead costs of a well designed survey is $25,000, plus $15 per interviewee, what total cost should be allocated for the survey?

6. A shifty-eyed car dealer tells you that the VW Bug he is trying to sell you gets 30 mpg. You don’t trust him, so you get 9 of your friends to come by at various stages, to test-drive the car. They report the following mileage figures back to you:

23, 24, 25, 26, 26, 26, 27, 28, 29

Do you believe the car-dealer's claim, in the light of this evidence?
VARIOUS PROBLEMS

(HINT: find a 95% confidence interval for the true mean gas mileage per gallon.)

8. Floyd Lawson, the Mayberry town barber, has recently hired a masseuse and is interested in determining the effect this has had on business. For 9 days his profits have averaged $\overline{X} = $50 with a sample standard deviation of $s = $20. Assuming one day’s profits are normally distributed with unknown mean $\mu$ and unknown variance,
   a) Construct a 99% confidence interval for $\mu$, based on the above statistic.

9. Suppose I am trying to estimate the proportion of the beer-drinking population who would drink Guinness beer if I were to start exporting it to College Station. I have no idea in advance what this proportion is likely to be. Determine how large a sample I need to take to be 95% sure that my estimate of the proportion is within 2 percentage points of the true proportion.

10. In a survey in which 100 randomly picked middle-class families were interviewed, it was found that their mean medical expenses during a year were $770 with $s = $120. Do the data indicate that the true mean medical expenses during the year are significantly different from $750? Test at the .05 significance level.

11. A competitor of the MacDonalds Fast Food chain wished to show that the Quarter Pounder does not live up to its name: i.e. to show that it weights less than the asserted 4 ounces. A random sample of 16 Quarter Pounds had an average weight of 3.87 ounces with a standard deviation for the sample of .70 ounces
   a) Construct a 95% confidence interval for the mean weight of all MacDonalds Quarter Pounders.
   b) Did the sample present sufficient evidence to conclude at the 10% level of significance that MacDonalds is skimping?

12. a) As sample size increases, does the variance of the sample mean $\overline{X}$
   i) increase,
   ii) decrease, or
   iii) remain the same?

   b As sample size increases, does the length of a 95% confidence interval
   i) increase,
   ii) decrease, or
iii) remain the same?

13. An investigation is undertaken to determine how the administration of a growth hormone affects the weight gain of pregnant rats. Weight gains during gestation are recorded for 6 control rats and for 6 rats receiving the growth hormone. The following summary statistics are obtained:

<table>
<thead>
<tr>
<th></th>
<th>Control Rats</th>
<th>Hormone Rats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean</td>
<td>41.8</td>
<td>60.8</td>
</tr>
<tr>
<td>Sample s.d.</td>
<td>7.6</td>
<td>16.4</td>
</tr>
</tbody>
</table>

a) At the 5% level of significance, is the mean weight gain different for the rats receiving the hormone than for the rats in the control group.

14. a) An automobile company executive claims that the average gas mileage of a new make of car is 33 mpg. In an EPA test, the mileages of 50 cars were recorded: the mean mileage in this sample was 29.5 with $s^2 = 128$. Test at significance level .01 whether or not the executive’s claim seems to be true.

15. A company takes a random sample of salesmen from its western and eastern sales offices and computes the earnings for each salesman. The data are:

<table>
<thead>
<tr>
<th>Sales District</th>
<th>Sample Size</th>
<th>$\bar{X}$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern</td>
<td>150</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>Western</td>
<td>140</td>
<td>11,000</td>
<td>80</td>
</tr>
</tbody>
</table>

Construct a 95% confidence interval for the difference in mean earnings between the eastern and western districts.

16. The brothers of the MM fraternity (the ‘Moo-Moos’) challenge the sisters of the φφ (the ‘Fifi’s’) to a beer-drinking contest. The following figures represent the amount of beer (in pints) consumed in a 2-minute interval (as recorded by a sober bystander) by the competitors:

<table>
<thead>
<tr>
<th>MM’s</th>
<th>φφ’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3, 4.0, 3.8</td>
<td>6.1, 5.8, 6.4, 6.3</td>
</tr>
<tr>
<td>4.1, 3.9, 4.2</td>
<td>6.0, 6.2, 6.1</td>
</tr>
</tbody>
</table>

Does the above provide evidence for a difference in the beer-drinking capacities of both groups? Conduct a test at the 5% significance level.
17. Suppose a drug company has come up with a new serum that could help prevent a disease. The question is whether the serum is effective or not. The courses of action open are to market the serum or not market it.
   a) What are the two types of errors that could be committed?
   b) As far as the drug company is concerned, which error is more serious?
   c) As far as the Federal Drug Administration (which protects consumer interest) is concerned, which error is more serious?

19. The mean and standard deviation for the life of a random sample of 100 light bulbs were calculated to be 1285 and 150 hours, respectively. Estimate the mean life of the population of light bulbs from which the sample was drawn, and place bounds on the error of estimation.

20. A survey is taken to update census data on water usage in Riverside county. From the population of manufacturing firms employing over 25 employees, 58 firms from a simple random sample of 625 drawn use water in the manufacturing process. Estimate at the .95 confidence level the percentage of all manufacturing firms, employing over 25 employees which use water in the manufacturing process.

21. Your Mira Mesa Law Clinic has a word-processing system on a lease with option-to-buy condition and currently is evaluating the performance of the accompanying $10,000 picture-ready printer unit. Printers commonly come with a service contract, but the printers that can minimize shut-down cost will be the most attractive. Shut-down cost includes the opportunity cost of lost time and out-of-pocket expenses of salaried employees left idle. The down time lapse between the point of break-down and the start-up point showed an average of 4 1/2 hours with a standard deviation of 2 hours for sixteen observations. Assume a normal distribution for lapse time is plausible. Construct a 98% confidence interval estimate for the printer's mean 'down' time per break-down. Using the upper confidence limit as the underlying true mean down time, how probable is it that a printer remains down an entire 8 hour day.

22. As part of its job enrichment program for its middle management, a firm has agreed to pay for all the MBA tuition cost up to two courses per semester. Four hundred signed letters of intent to enroll were received from managers taking advantage of the enrichment program. To gain some attitudinal information about the program as well as an estimate of tuition expense for next year's budget, the personnel department picked and interviewed a random sample of
50 from the 400. A survey indicated that next year's average cost per manager will be $2500 with a standard deviation of $625. Calculate the interval estimate of the total cost at a .95 confidence level for all managers.

25. The Santee School district is being investigated for achievement testing irregularities. The average test score in mathematics for last spring was 92. To test the validity of these results the State Attorney General’s office ordered a comparable test to be given to a random sample of 64 of these students. The mean score of this sample was 80 with a standard deviation of 12 points. From this, do you accept that 92 was a valid score at the .05 level of significance?

26. City planners want to test the effectiveness of solar water heaters. To be adequate a water heater needs to be able to maintain a water temperature of at least 120° F. The population distribution of water temperature is assumed to be normal. A test of 16 water heaters is made and the sample standard deviation S is calculated to be 8, and the sample mean $\bar{X}$ is 117° F.
   a) At a .05 level of significance do you accept the null hypothesis that the water temperature is at least 120° F?

27. A Rochester, New York film plant produces ready-to-use film cartridges for instant pictures. The thickness of the coating material on the film is a cause for concern, because too much or not enough coating destroys the quality of the picture. When functioning properly, the manufacturing process places on average a coating of .040 mm maximum thickness of film on the photographic paper used in the cartridge. The standard deviation of coating thickness is .002 mm. For every batch produced, several cartridges are tested to assess the thickness of the coating applied. It is important that for the sample of ten cartridges tested, the mean thickness of coating must be no more than .043 mm. A value over these limits is taken as serious enough to stop the process and investigate the production line.
   d) Calculate the probability that the process will be stopped, even though the process is properly functioning (that is, with maximum coating thickness at .040 mm average).
   b) Calculate the probability that the process will be stopped if it is depositing an average .042 of film coating.

   Assume that the process on average is depositing .043 mm film coating, an amount equal to the upper tolerance limit. What is the chance that the process is allowed to continue without investigation after the sample of 10 cartridges is tested?
33. The Bona Fide Washing Machine Company knows that the standard deviation of the lengths of life of motors received from supplier I is 400 hours. Calculate the 95 percent confidence interval for the mean length of life of the motors received from this supplier, based on a sample of 40 motors where the mean length of life is 4,500 hours.

34. A US steel firm wants to determine the proportion of its customers who bought Japanese steel during the past year. It selects a random sample of 100 of its customers, and finds that 36 percent of them bought Japanese steel then.
   a) Calculate a 90 percent confidence interval for this production.

36. A state university wants to estimate the average amount that its students paid in state income taxes in 1983. The standard deviation of state income tax paid by the students is estimated to be about $60. The university wants the probability to be 0.95 that the sample mean will differ by no more than $10 from the true mean. If a random sample of the students is selected, how big should it be?

38. The number of cough drops included in a box varies from box to box. A manufacturer of cough drops wants to test whether the mean number in a box of its cough drops equals 12. It picks a random sample of 60 boxes, and finds that the mean number in a box equals 11.95, the sample standard deviation being 0.09. It sets $\alpha$ equal to 0.05. What should the firm’s decision be?

39. A firm produces metal wheels. The mean diameter of the wheels should be 4 inches. Because of chance variation and other factors, the diameters of the wheels vary, the standard deviation being 0.05 inches. To test whether the mean is really 4 inches, the firm selects a random sample of 50 wheels (and finds that the sample mean diameter equals 3.97 inches). What should the firm’s decision be?

40. A class contains 60 students. The teacher wants to test whether the mean IQ in the class equals 120. He chooses a random sample of 36 students (without replacement), and find that the mean and standard deviation of the IQ’s in the sample are 122.8 and 10.9, respectively. What should the teacher’s decision be? Take $\alpha = 0.05$. 
41. The president of the Crooked Arrow National banks wants to test whether 60 percent of the bank’s loans are made to persons who reside in the city where the bank is located. The bank’s statistician chooses a random sample of 200 of the people to whom the bank has made loans and finds that 52 percent reside in this city.
   a) If a 5% significance level is used, should this hypothesis be accepted or rejected? (Use a two-tailed test.)
   b) If a 1% significance level is used, should this hypothesis be accepted or rejected? (Use a two-tailed test.)

42. An Atlantic City gambling casino wants to make sure that the probability of winning on a certain slot machine is 0.4 (no more, no less). The machine is operated 200 times, and the operator wins 89 times.
   a) If $\alpha$ is set at 0.01, is the difference between the sample proportion and 0.4 statistically significant?

43. The Internal Revenue Service wants to determine whether the percentage of personal income tax returns filed on time by taxpayers is less than last year’s percentage, 81 percent. It selects a random sample of 400 of this year’s tax returns and finds that 317 of them were filed on time.
   a) Can the Internal Revenue Service be reasonably sure that the percentage filed on time has declined?
   b) If the significance level (that is, $\alpha$) is set at 0.05, is the apparent decline statistically significant?

44. It is claimed that men are better than women at a certain clerical task. To see whether this is the case, a firm chooses a random sample of 100 men and a random sample of 100 women. Each person is given this task to do, and his or her performance is graded from zero to 100. For the men, the sample mean is 60.8 and the sample standard deviation is 9.9. For the women, the sample mean is 58.4 and the sample standard deviation is 8.7.
   a) If the significance level is set at 0.05, does the evidence indicate that men are better than women at this task?

45. A chemical firm has two very large research laboratories, one in the United States and one in Europe. It wants to determine whether the chemists at its US laboratory tend to be older or younger than those at its European laboratory. It selects a random sample of 80 chemists from each laboratory, the results
being as follows:

<table>
<thead>
<tr>
<th></th>
<th>Mean age (years)</th>
<th>Standard deviation (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>50.6</td>
<td>8.7</td>
</tr>
<tr>
<td>Europe</td>
<td>47.3</td>
<td>7.6</td>
</tr>
</tbody>
</table>

a) If the significance level is set at 0.01, does the evidence indicate that the men age is the same at the two laboratories?

b) If the significance level is set at 0.05 instead, what does the evidence indicate?

46. A pharmaceutical firm wants to determine whether a weight-reducing drug has a different effect on adults over 40 years old than on adults that are no more than 40 years old. Twelve people over 40 are given the drug; their mean weight loss is 8.9 pounds, and the standard deviation is 4.1 pounds. Twelve people no more than 40 are given the drug; their mean weight loss is 11.3 pounds, and the standard deviation is 3.8 pounds.

a) Based on a two-tailed test with $\alpha$ equal to 0.05, are these data consistent with the hypothesis that the average effect of this drug is the same in both age groups?

b) Would your answer to (a) change if the significance level were 0.01 rather than 0.05?

50. A herd of 26 milking cows is divided at random into two groups of size 13, which are fed two different diets. The first diet has the economic advantage of producing a useful byproduct in its manufacture, making it cheaper, so it will be used unless there is evidence that the resulting milk yields are lower than those of the second diet.

The average daily milk yields over a three-week period are recorded as follows: two cows from the second group became ill and had to be removed from the experiment.

1st diet: 39, 43, 47, 38, 41, 37, 114, 42, 45, 36, 41, 39, 72.
2nd diet: 35, 46, 48, 40, 51, 46, 40, 49, 40, 50, 46.

Can the question of which diet to use be settled from these figures?

53. Leukemia remission times (months)

$X$-sample (exposed to drug 6-MP): 6, 6, 6, 7, 10, 13, 16, 22, 23, 32, 32

$Y$-sample (control): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17,
Carry out a suitable test to investigate whether the drug 6-MP helps to delay the recurrence of leukemia.

55. Thirty-five patients with moderately severe hypertension were given 2 drugs on different days to establish the relative values of the drugs in reducing systolic blood pressure. For each drug the measurement of blood pressure was taken two hours after the administration of the drug. The blood pressure was also recorded 15 minutes before the administration of the drug.

Do the drugs differ in their abilities to lower systolic blood pressure?

<table>
<thead>
<tr>
<th>Patient</th>
<th>CAPT Before</th>
<th>CAPT After</th>
<th>DIST Before</th>
<th>DIST After</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>210</td>
<td>201</td>
<td>195</td>
<td>190</td>
</tr>
<tr>
<td>2</td>
<td>169</td>
<td>165</td>
<td>187</td>
<td>186</td>
</tr>
<tr>
<td>3</td>
<td>187</td>
<td>166</td>
<td>189</td>
<td>186</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
<td>157</td>
<td>169</td>
<td>171</td>
</tr>
<tr>
<td>5</td>
<td>167</td>
<td>147</td>
<td>177</td>
<td>166</td>
</tr>
<tr>
<td>6</td>
<td>176</td>
<td>145</td>
<td>166</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>185</td>
<td>168</td>
<td>188</td>
<td>163</td>
</tr>
<tr>
<td>8</td>
<td>206</td>
<td>180</td>
<td>189</td>
<td>170</td>
</tr>
<tr>
<td>9</td>
<td>173</td>
<td>147</td>
<td>180</td>
<td>168</td>
</tr>
<tr>
<td>10</td>
<td>146</td>
<td>136</td>
<td>167</td>
<td>163</td>
</tr>
<tr>
<td>11</td>
<td>174</td>
<td>151</td>
<td>163</td>
<td>155</td>
</tr>
<tr>
<td>12</td>
<td>201</td>
<td>168</td>
<td>184</td>
<td>163</td>
</tr>
<tr>
<td>13</td>
<td>198</td>
<td>179</td>
<td>171</td>
<td>175</td>
</tr>
<tr>
<td>14</td>
<td>148</td>
<td>129</td>
<td>172</td>
<td>168</td>
</tr>
<tr>
<td>15</td>
<td>154</td>
<td>131</td>
<td>165</td>
<td>147</td>
</tr>
</tbody>
</table>
CHAPTER 15

COURSE DATASETS

SKEENA RIVER SOCKEYE SALMON

This is the Skeena River Sockeye Salmon data set for a 28 year period. The dataset name is SKEENA. The main variables are the number of spawners, i.e., mature fish who are about to spawn and die, and the number of recruits, i.e., new fish who are born and then recruited into the fishery. The important problem here is to predict the number of recruits from a given number of spawners.

AORTIC STENOSIS

This example consists of two sets of kids. There is a group of normal, healthy kids (dataset name is NORMVALV), and a group of kids with aortic stenosis (dataset name is ABNOVALV). The two variables are body surface area (BSA) and aortic valve area (AVA); I also use log(1+BSA) and log(1+AVA). The goal of the project is to set normal ranges of aortic valve areas for normal kids who have a given body surface area, and then to use this in a prediction experiment. Specifically, a kid will have his/her BSA and AVA measured, and if he/she has an AVA not in the normal range, then he/she is a candidate for aortic stenosis, and further tests will be performed.

REACTION TIMES

Auditory and visual reaction times were observed in a 1992 Texas A&M student experiment for males and females. The data are in the dataset AUDVREG. Gender=0 is the males, while gender=1 is the females.

HORMONE ASSAY

This is an industrial experiment comparing two methods for performing a hormone
assay. There is a standard or reference method, and a new or test method. The goal is to see how good a job the new, test method does of predicting the standard, reference method. The dataset is called HORMONE.

**ESTERASE ASSAY**

This is a typical immunoassay, the dataset name being ESTERASE. The amount of esterase is being used to predict radioactive counts.

**SUICIDE CLINICAL TRIAL**

This data, called the SUICIDE dataset, is a typical clinical trial. Those randomized to the control or placebo group are also in the dataset names SUICONT, while those randomized to the treatment group are also in the data set SUICASE. This is baseline data, i.e., data observed at the start of the experiment before therapy began. The key feature of interest here is whether the randomization worked, i.e., whether the treatment and control groups are essentially the same. If they are, then any differences we observe by the end of the study can be ascribed to the therapy.

**STUDY OF BREAST CANCER IN YOUNGER WOMEN**

The main dataset is called HANES, and contains part of a large study of breast cancer and nutrition in younger women. Women had their age, body mass index, saturated fat, alcohol usage, etc. assessed at the beginning of the study, and were followed for 10 years after that. I have included all 59 women who developed breast cancer; they are also all by themselves in the dataset HANECASE. Of the 3086 women who remained cancer--free, I selected 60 of them at random. They are alone in the dataset HANECONT.

**COUNTING CHOCOLATE CHIPS**

In the dataset CHOCCHIP, I have the results of an experiment done in the 1991--92 academic year. I taught two sections of STAT--302 that semester, and asked the class on two occasions to count chocolate chips. The two largest means occurred when I did the experiment at the end of class, the two smallest during class.