Outline of Lecture 12, STAT 303, Sections 503–504

H.J. Newton

February 12, 1996

1 The Normal (bell-shaped) distribution

- Basic properties of the normal curve:
  - Notation: $X \sim \mathrm{N}(\mu, \sigma^2)$ means a population having the normal distribution with mean μ and variance σ^2.
  - Special case: $Z \sim \mathrm{N}(0, 1)$.
- Calculating probabilities and quantiles for a single observation:
  - For $Z \sim \mathrm{N}(0, 1)$, using `Calculator -> Statistical tables -> Normal -> z` will give $\Pr(Z < z)$, $\Pr(Z > z)$, and $\Pr(Z < -|z| \text{ or } Z > |z|)$. To find probabilities for $X \sim \mathrm{N}(\mu, \sigma^2)$, first calculate the Z score, $Z = (x - \mu)/\sigma$, and then use the same procedure as for Z.
  - For $Z \sim \mathrm{N}(0, 1)$, using `Calculator -> Inverse statistical tables -> Normal -> p` will give the pth percentile (quantile) of Z. The pth percentile of $X \sim \mathrm{N}(\mu, \sigma^2)$ is $\mu + \sigma \times$ pth percentile of Z.
- **The central limit theorem**. Let X represent a population having mean μ and variance σ^2 and *any distribution*. If we collected $\bar{X}$'s for millions of samples of size n from X, then this population of $\bar{X}$'s would be approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.
- Calculating probabilites and quantiles for a sample mean. These are the same as for X except we use $\sigma/\sqrt{n}$ instead of σ.

2 Questions being asked in Chapters 6, 7, 8, and 10 (parametric statistical inference)

- How do I estimate the parameter (or parameters) of a population (or populations) given a sample (or samples) from the population (or populations):
  - With a single guess? (**point estimation**)
  - With a range of guesses I can be 'confident' includes the true value? (**confidence interval**)
- How do I decide if a specified value (or values) is reasonable value for a parameter (or parameters)? (**tests of hypotheses**)
- What parameters are we interested in?
  - The mean of a continuous population
    - When population has normal distribution
    - When population might not have normal distribution but sample size is 'large'
  - The variance (or standard deviation) of a normal population
  - The proportion of 1's in a 0-1 population
  - The difference of means in two normal populations
    - When the samples are independent

        - When the variances of the populations are equal
        - When the variances of the populations are not equal
    - When the samples are not independent (when they come in pairs such as a 'before and after' experiment)
  - The ratio of variances (or standard deviations) in two normal populations
  - The difference in proportions of 1's in two 0-1 populations
  - The means of more than 2 normal populations
  - The proportion of 1's in more than two 0-1 populations

3 How do we answer these questions?

- By using the **sampling distribution of a transformed statistic**

- For example, if we have a random sample of n elements from a $\mathrm{N}(\mu, \sigma^2)$ population, then even though we don't know μ, statistical theory allows us to find two numbers t_1 and t_2 (from a table or computer program) so that

$$\Pr\left(t_1 \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_2\right) = .95,$$

  for example (there are a t_1 and t_2 for any specified probability like the .95 in this example). Note that s is the standard deviation of the sample, and thus we know all the numbers in this equation except μ.

- So what? If we solve this inequality to get μ by itself in the middle, we find

$$\Pr\left(\bar{X} + t_1\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_2\frac{s}{\sqrt{n}}\right) = .95,$$

  which gives us a range of possible values for μ that has a 95% chance of actually containing the its true value. This is called a 95% *confidence interval for* μ.

- For example, if a company is manufacturing ball bearings and randomly selects $n = 16$ of its ball bearings and finds their average diameter is $\bar{X} = 10$mm and the standard deviation of the diameters is $s = 2$, then the 95% confidence interval for the average diameter of *all* of its ball bearings is (we find $t_1 = -2.1314$ and $t_2 = 2.1314$ from Stataquest) $(10 - 2.1314(2/\sqrt{16}), 10 + 2.1314(2/\sqrt{16}))$, which is (9.9343,11.0657) mm.

- In this example, μ is the **parameter** of interest, $\bar{X}$ is a **statistic** estimating μ,

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

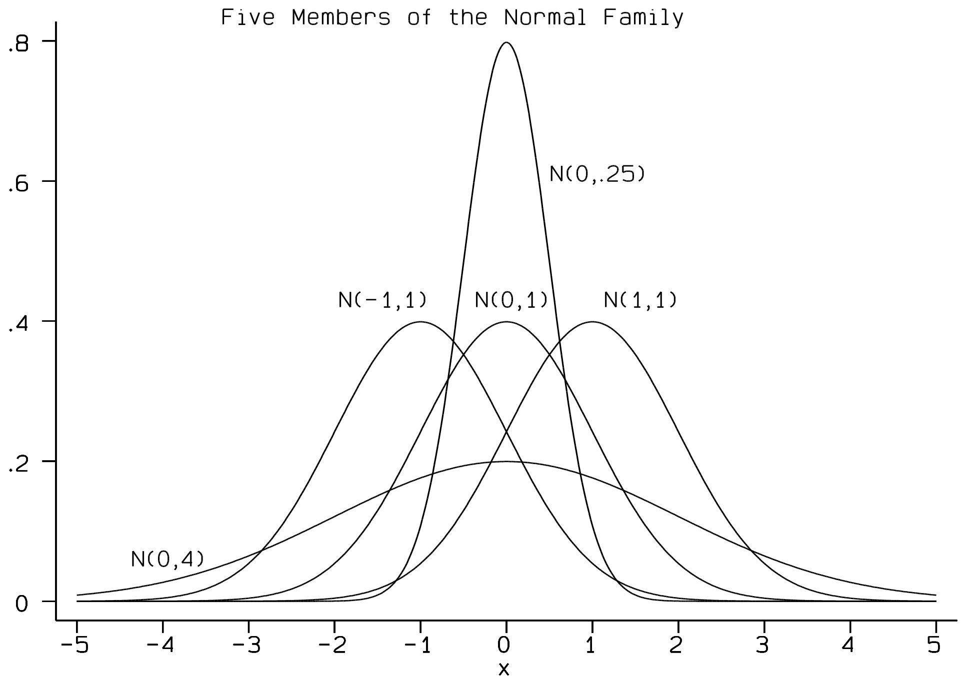
  is called a **transformed statistic**, and t_1 and t_2 are called **critical values**.

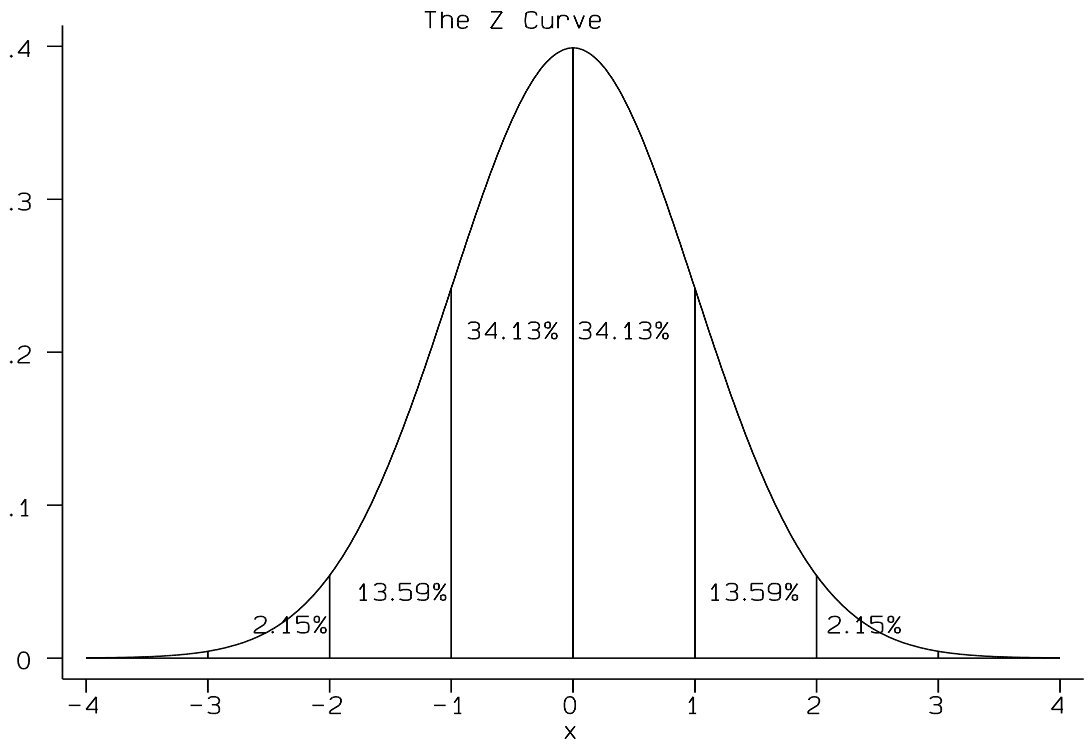
- What are the key elements of this process that we have to learn?

  - In each of the many situations discussed above, what is the transformed statistic (see the table later in this handout) and how do we find critical values? (also in the table)
  - Why does this work? (See the **sampling distribution concept lab**).

Transformed Statistics for One and Two Sample Inferences

Situation			
Transformed Statistic	Should be near	If	df
For μ, Normal sample, σ^2 known (`1 sample normal`)			
$Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	0	$\bar{X} \doteq \mu$	
For μ, Normal sample, σ^2 unknown (`1 sample t`)			
$t = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$	0	$\bar{X} \doteq \mu$	$n-1$
For $\mu_d =$ mean difference in paired data, Normal pairs, σ_d^2 unknown (`paired t`)			
$t = \dfrac{\bar{d} - \mu}{s_d/\sqrt{n}}$	0	$\bar{d} \doteq \mu_d$	$n-1$
For σ^2, Normal sample (`1 sample variance`)			
$\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$	$n-1$	$s^2 \doteq \sigma^2$	$n-1$
For $\mu_1 - \mu_2$, σ_1^2, σ_2^2 known, independent normal samples (`2 sample normal`)			
$Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$	0	$\bar{X}_1 - \bar{X}_2 \doteq \mu_1 - \mu_2$	
For $\mu_1 - \mu_2$, σ_1^2, σ_2^2 unknown but assumed equal, independent normal samples (`2 sample t`)			
$t = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_P^2(1/n_1 + 1/n_2)}}$,	0	$\bar{X}_1 - \bar{X}_2 \doteq \mu_1 - \mu_2$	$n_1 + n_2 - 2$
For π, sample from 0–1, $\min(n\pi, n(1-\pi) \geq 5$, (`1 sample proportion`)			
$Z = \dfrac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$	0	$p \doteq \pi$	
For $\pi_1 - \pi_2$, independent samples from 0–1 populations, $\min(n_1\pi_1, n_1(1-\pi_1) \geq 5$, $\min(n_2\pi_2, n_2(1-\pi_2) \geq 5$, (`2 sample proportions`)			
$Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2}}$	0	$p_1 \doteq \pi_1$, $p_2 \doteq \pi_2$,	
For ratio of variances, independent normal samples (`2 sample variance`)			
$F = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \dfrac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$	1	$s_1^2 \doteq \sigma_1^2$, $s_2^2 \doteq \sigma_2^2$	$df_1 = n_1 - 1$, $df_2 = n_2 - 1$


Five Members of the Normal Family
N(0,.25)
N(-1,1)
N(0,1)
N(1,1)
N(0,4)
.8
.6
.4
.2
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x


The Z Curve
34.13%
34.13%
13.59%
13.59%
2.15%
2.15%
.4
.3
.2
.1
0
-4
-3
-2
-1
0
1
2
3
4
x