Objectives

- Density curves
- Measuring center and spread for density curves
- Normal distributions
- The 68-95-99.7 (Empirical) rule
- Standardizing observations
- Calculating probabilities using the standard Normal Table (CIS Chapter 8, p 105 – mainly p114)
- Inverse Normal calculations

Books: OS3: Section 3.1 (entire section). IPS: p242, Section 4.3
Topic: Density Curves

- Learning targets:
  - Understand how to interpret the relative frequency plot of data.
  - Understand how to calculate proportions in a sample based on a relative frequency graph.
  - Understand what a density curve is.
  - Understand that the area under a curve corresponds to proportions in a population.
  - Understand that the shapes of density curves can be very different, with many different features.
  - Be able to make a rough sketch of the density curve of a variable and be able to place the mean and standard deviation on it.
Histogram and density curves

- If the data is a continuous numerical variable, it can take any value in a given range of numbers
  - Heights of people can lie anywhere between 0.5 meters to 2 meters.
  - Weights of pigs can lie anywhere between 150 pounds to 600 pounds.
- For any sample from a continuous numerical variable we plot the relative frequency histogram.
- The “histogram” of the population of a continuous random variable is called the density curve. It differs from the histogram of a sample in a few important ways:
  - The height of the relative frequency histogram gives the proportion/chance over any given interval.
  - For sum of all heights is one.
  - The area under the density curve gives the proportion/chance over any given interval.
  - The area under all of the curve is one.
- As we rarely observe the entire population, we do not know the true density curve. Based on the sample we can often obtain a good estimate using statistical software.
Histograms and density curves in statcrunch

- Lab practice: Load data into Statcrunch. Plot a histogram. In Display options on the histogram menu, there is an option called Overlay distrib. In this menu there is a list of density shapes with different shapes. You can overlay your histogram to see which best fits your data.

- Below is the relative frequency histogram of MLB salaries.
Reading the height of the relative frequency histogram, the proportion of those earning between 0-250K dollars is 55%. On the other hand, the proportion given by the density is the area under the red line from 0-250K.

On the relative frequency histogram of MLB salaries, we plot two different densities. Which fits the data better?
Plotting densities and visualization

- Make a sketch of what you think the density curve of human heights are (use mean 67 inches and standard deviation 7 inches as an aid).
- What is the area below the entire curve?
- On the plot show the proportion of human heights less than 60 inches.
- On the plot show the proportion of human heights greater than 75 inches.
- On the plot show the proportion of human heights lying between 60 and 75 inches.
- On the next slide we give an examples of different types of density curves. Match a type of variable to each plot.
Density curves come in any imaginable shape.

The only restrictions is that the $y$-axis cannot take negative values and the area below the curve is one.
Topic: The normal distribution

- Learning targets:
  - Understand that the normal distribution is one particular family of density curves.
  - Understand that the normal distribution is determined by its mean and standard deviation.
  - Understand the main features of the normal distribution.
  - Calculate z-scores and percentiles using the normal distribution (using both tables and statcrunch). Understand that all these calculations are based on the **assumption** the data is normal. If the data is not normal then the calculated percentile can be wrong.
  - Make comparisons using percentiles.

- OS3, page 128 onwards.

- Free normal calculators
  
The normal family of density plots

- We now introduce a family of density functions which are extremely useful in statistics. It is called the normal distribution.

- Here are some reasons that they are important in statistics
  - Some variables (but not all) have a density which is close to a normal distribution. These include biological measurements, some type of exam scores etc.
  - If we can assume a variable is normally distributed, it allows us to calculate probabilities easily (for example, if weights were normally distributed, you can easily calculate the percentile for your weight by knowledge of the population mean and standard deviation).
  - The normal distribution forms the basis of statistical inference. For this reason you should become very familiar with all the normal calculations we do from now on. As we will be using these ideas throughout the course.
Definition: Normal density

Normal distributions are a family of symmetrical, bell-shaped density curves defined by a mean $\mu$ (mu) and a standard deviation $\sigma$ (sigma). We denote a normal distribution by $\text{Normal}(\mu, \sigma)$ or $N(\mu, \sigma)$.

The formula for the density curve is somewhat complicated:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}$$

$e = 2.71828...$ the base of the natural logarithm
$\pi = \pi = 3.14159...$
Examples of normal density curves

Here, means are the same ($\mu = 15$) while standard deviations are different ($\sigma = 2, 4, \text{ and } 6$).

Here, means are different ($\mu = 10, 15, \text{ and } 20$) while standard deviations are the same ($\sigma = 3$).
Load the calf data into Statcrunch. Here the aim is to compare the histogram of calf weights to the normal density curve.

- Graphics -> Histogram
  - Select a weight variable (such as 8 weeks)
  - Select relative frequency (in Type options).
  - In Overlay distrib choose normal density. Statcrunch will calculate the sample mean and standard deviation of the weights and use this to center the normal density curve.

Remember the fit won’t be perfect.
A crude rule for checking Normal Distributions

- About 68% of all observations are within 1 standard deviation (1 × \(\sigma\)) of the mean (\(\mu\)).
- About 95% of all observations are within 2 × \(\sigma\) of the mean \(\mu\).
- Almost all (99.7%) observations are within 3 × \(\sigma\) of the mean.
- Also called the empirical rule because it works approximately for data and many other distributions. E.g., typically 90%-99% of data are within two st. dev.’s of the mean.

Notation: \(\mu\) (mu) is the mean of the idealized curve, while \(\bar{x}\) is the mean of a sample. \(\sigma\) (sigma) is the standard deviation of the idealized curve, while \(s\) is the s.d. of a sample.
One-sided

- Using the symmetry of the normal distribution and the plot on the right as a guide we make the following observations:
  - About 34% of all observations are within 1 standard deviation to the left of the mean.
  - About 47.5% of all observations are within 2 standard deviations to the left of the mean.

\[
\text{mean } \mu = 64.5 \quad \text{standard deviation } \sigma = 2.5 \\
\text{Normal}(\mu, \sigma) = \text{Normal}(64.5, 2.5)
\]
Important observation

- The proportions under the curve do not change linearly with the number of standard deviations from the mean.
- 68% of the observations are within **one standard deviation** of the mean.
- **This does not mean**
  - $2 \times 68 = 136\%$ of the observations lie within **two standard deviations** of the mean. In fact, **95\%** of the observations are within **two standard deviations** of the mean.
  - $0.5 \times 68 = 34\%$ of the observations lie within **half a standard deviation** of the mean. In fact, **38\%** of observations lie within half a standard deviation of the mean.
- Later in this chapter we explain how these proportions are calculated. You will need to use software of statistical tables to calculate probabilities.
Do the weights of 8 week old calves satisfy the empirical rule?

- Load the calf data into Statcrunch, and make a relative frequency histogram of the calf data (use the bin width 5). Overlay a normal distribution.

- Note the sample mean and standard deviation of the calf data and construct the intervals (a) one standard deviation from the mean (b) two standard deviations from the mean (c) three standard deviations from the mean). And count the proportion of calves in these intervals.

- This is just a small sample, but it would appear that calf weights `roughly’ satisfy this rule.
Z-score

- It is important to note that all observations that are within one standard deviation of the mean have z-score less than one.

- This is a plot of a normal distribution with mean 64.5 and sd = 2.5. An interval which is one standard deviation from the mean is [62,67]. Any observation inside this interval will have a z-transform that is less than one.

- All observations in blue in the plot have a z-score less than one.
- Similarly any observation which is outside the interval will have z-score greater than one.

- All observations in blue will have score greater than one.

- This is true for an interval. If an interval is within x-standard deviations of the mean. Then the z-score for any observation in that interval will have a z-score less than x. The z-score for any observation outside the interval will have a z-score more than x.
Do baseball salaries satisfy the empirical rule?

The green line is the mean. And the idea of symmetric intervals about the mean has no meaning, since the proportions on either side are very different.

The answer is no it doesn’t. Using symmetric intervals about the mean is not meaningful for non-symmetric data.

<table>
<thead>
<tr>
<th>Summary statistics:</th>
<th>2015 to 2020 Salary * 10 ** (-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column</td>
<td>n</td>
</tr>
<tr>
<td>SalaryMillion</td>
<td>817</td>
</tr>
</tbody>
</table>
Question about healthy calves and the Empirical Rule

- **Using the empirical rule:** You are presented with an 8 week calf whose weight is 95 pounds. Is he healthy?
- Calculating the z-transform:
  \[ z = \frac{95 - 142.6}{17} = -2.8 \]

- The calf is -2.8 standard deviations below the mean.
- -2.8 is quite far from where most weights lie.
- If healthy calf weights are normally distributed, then by the empirical rule 95% of healthy calves lie within 2 standard deviations of the mean. This means, by symmetry of the normal distribution, 2.5% of healthy calves have a weight more than two standard deviations to the left of the mean.
- We conclude this is quite an unusual/rare event for healthy calves.
Negative and positive z-transforms

- If the data is less than the mean, then the z-transform will be negative.
  - Example: The 95 pound calf, in the previous example, was 2.8 standard deviations \(= 2.8 \times 17\) less than (to the left of) the mean of 142 pounds. Its z-transform is \(-2.8\).

- If the data is greater than the mean, then the z-transform will be positive.
  - Example: A 190 pound calve is 2.8 standard deviations to the right of the mean (142 pounds).
    Its z-transform is \(2.8\).

- In both examples the calves are 2.8 standard deviations from the mean, but in completely different directions.

- Remember the z-transform is always free of units of measurement.
Z-scores and the normal density

- If the data is normal, z-scores (defined at the end of Chapter 3) can be used to calculate percentiles.
- Example: The heights of women are close to normally distributed with mean 64.5 inches and standard deviation 2.5 inches.
- Question: A woman has a height of 71 inches, is she exceptionally tall?
- Answer: The z-transform calculates how close 71 is to the mean but takes into account the spread of heights:

\[ z = \frac{71 - 64.5}{2.5} = 2.6 \]

- This is 2.6 standard deviations to the right of the mean. Since females heights are close to normal we use the 68-95-99.7 rule. 5% of women are more than 2 standard deviations from the mean. Therefore she has to be in the top 2.5% (5% divided by two) of heights.
- She is tall, but what is the exact percentile? To do this we need to calculate the area to the LEFT of 71 on the density curve. If the data is normal, this is area is easy to find (if it is not, then it can be difficult).
The standard Normal distribution

The z-transform transforms Normal(\(\mu, \sigma\)) curve into the standard normal curve: Normal(0,1).

For each \(x\) we calculate a new value, z-score.
Calculation: Women heights

Women’s heights follow the N(64.5", 2.5") distribution. What percent of women are shorter than 71 inches tall?

mean $\mu = 64.5"$

standard deviation $\sigma = 2.5$

$x$ (height) = 71"

We calculate $z$, the standardized value of $x$:

$$z = \frac{x - \mu}{\sigma}, \quad z = \frac{71 - 64.5}{2.5} = 2.6 \quad 2.6 \text{ s.d. above the mean}$$

To find the percent of women are shorter than 71 inches tall, we need to find the area to the left of $z = 2.6$. For this, we must use a special table.
To find the percentile. Go to the normal tables at https://www.stat.tamu.edu/~suhasini/teaching301/z_Table.pdf

Look for 2.6 on the row and 0.00 (2.6+0.00 = 2.6). Search for the intersection. It gives you 0.9953.

0.9953 = 99.53% which corresponds to the percentile for a female of height 71 inches under the assumption the distribution is normal with mean 64.5 and standard deviation 2.5.

Conclusion:
99.53% of women are shorter than 71".
By subtraction, 1 – 0.9953, or 0.46% of women are taller than 71".
She is in the 99.53 percentile.
Equivalent: She is in the top 0.47 percentile.
Using the standard Normal table

Table A gives the area under the standard Normal curve to the left of any z value.

<table>
<thead>
<tr>
<th>z</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.3</td>
<td>.0005</td>
<td>.0005</td>
<td>.0005</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0003</td>
</tr>
<tr>
<td>-3.2</td>
<td>.0007</td>
<td>.0007</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0005</td>
<td>.0005</td>
<td>.0005</td>
</tr>
<tr>
<td>-3.1</td>
<td>.0010</td>
<td>.0009</td>
<td>.0009</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0007</td>
<td>.0007</td>
<td>.0007</td>
</tr>
<tr>
<td>-3.0</td>
<td>.0013</td>
<td>.0013</td>
<td>.0012</td>
<td>.0012</td>
<td>.0011</td>
<td>.0011</td>
<td>.0011</td>
<td>.0010</td>
<td>.0010</td>
<td>.0010</td>
</tr>
<tr>
<td>-2.9</td>
<td>.0019</td>
<td>.0018</td>
<td>.0017</td>
<td>.0016</td>
<td>.0016</td>
<td>.0015</td>
<td>.0015</td>
<td>.0014</td>
<td>.0014</td>
<td>.0014</td>
</tr>
<tr>
<td>-2.8</td>
<td>.0026</td>
<td>.0025</td>
<td>.0024</td>
<td>.0023</td>
<td>.0023</td>
<td>.0022</td>
<td>.0021</td>
<td>.0021</td>
<td>.0020</td>
<td>.0019</td>
</tr>
<tr>
<td>-2.7</td>
<td>.0035</td>
<td>.0034</td>
<td>.0033</td>
<td>.0032</td>
<td>.0031</td>
<td>.0030</td>
<td>.0029</td>
<td>.0028</td>
<td>.0027</td>
<td>.0026</td>
</tr>
<tr>
<td>-2.6</td>
<td>.0047</td>
<td>.0045</td>
<td>.0044</td>
<td>.0043</td>
<td>.0041</td>
<td>.0040</td>
<td>.0039</td>
<td>.0038</td>
<td>.0037</td>
<td>.0036</td>
</tr>
<tr>
<td>-2.5</td>
<td>.0062</td>
<td>.0060</td>
<td>.0059</td>
<td>.0057</td>
<td>.0055</td>
<td>.0054</td>
<td>.0052</td>
<td>.0051</td>
<td>.0049</td>
<td>.0048</td>
</tr>
<tr>
<td>-2.4</td>
<td>0.0082</td>
<td>0.0080</td>
<td>0.0078</td>
<td>0.0075</td>
<td>0.0073</td>
<td>0.0071</td>
<td>0.0069</td>
<td>0.0068</td>
<td>0.0066</td>
<td>0.0064</td>
</tr>
<tr>
<td>-2.3</td>
<td>0.0107</td>
<td>0.0104</td>
<td>0.0102</td>
<td>0.0099</td>
<td>0.0096</td>
<td>0.0094</td>
<td>0.0091</td>
<td>0.0089</td>
<td>0.0087</td>
<td>0.0084</td>
</tr>
<tr>
<td>-2.2</td>
<td>0.0139</td>
<td>0.0136</td>
<td>0.0132</td>
<td>0.0129</td>
<td>0.0125</td>
<td>0.0122</td>
<td>0.0119</td>
<td>0.0116</td>
<td>0.0113</td>
<td>0.0110</td>
</tr>
<tr>
<td>-2.1</td>
<td>0.0179</td>
<td>0.0174</td>
<td>0.0170</td>
<td>0.0166</td>
<td>0.0162</td>
<td>0.0158</td>
<td>0.0154</td>
<td>0.0151</td>
<td>0.0146</td>
<td>0.0143</td>
</tr>
<tr>
<td>-2.0</td>
<td>0.0223</td>
<td>0.0222</td>
<td>0.0217</td>
<td>0.0212</td>
<td>0.0207</td>
<td>0.0202</td>
<td>0.0197</td>
<td>0.0192</td>
<td>0.0188</td>
<td>0.0183</td>
</tr>
<tr>
<td>-1.9</td>
<td>0.0281</td>
<td>0.0274</td>
<td>0.0268</td>
<td>0.0262</td>
<td>0.0256</td>
<td>0.0250</td>
<td>0.0244</td>
<td>0.0238</td>
<td>0.0233</td>
<td>0.0227</td>
</tr>
<tr>
<td>-1.8</td>
<td>0.0351</td>
<td>0.0344</td>
<td>0.0336</td>
<td>0.0329</td>
<td>0.0322</td>
<td>0.0314</td>
<td>0.0307</td>
<td>0.0300</td>
<td>0.0294</td>
<td>0.0287</td>
</tr>
<tr>
<td>-1.7</td>
<td>0.0436</td>
<td>0.0427</td>
<td>0.0418</td>
<td>0.0409</td>
<td>0.0401</td>
<td>0.0392</td>
<td>0.0384</td>
<td>0.0376</td>
<td>0.0368</td>
<td>0.0360</td>
</tr>
<tr>
<td>-1.6</td>
<td>0.0537</td>
<td>0.0526</td>
<td>0.0516</td>
<td>0.0505</td>
<td>0.0495</td>
<td>0.0485</td>
<td>0.0475</td>
<td>0.0465</td>
<td>0.0455</td>
<td>0.0445</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.0655</td>
<td>0.0643</td>
<td>0.0630</td>
<td>0.0618</td>
<td>0.0606</td>
<td>0.0594</td>
<td>0.0582</td>
<td>0.0570</td>
<td>0.0558</td>
<td>0.0546</td>
</tr>
<tr>
<td>-1.4</td>
<td>0.0793</td>
<td>0.0778</td>
<td>0.0764</td>
<td>0.0749</td>
<td>0.0735</td>
<td>0.0721</td>
<td>0.0708</td>
<td>0.0694</td>
<td>0.0680</td>
<td>0.0666</td>
</tr>
</tbody>
</table>

.0082 is the area under N(0,1) left of z = -2.40

.0080 is the area under N(0,1) left of z = -2.41

0.0069 is the area under N(0,1) left of z = -2.46
Probability calculators

- We can calculate probabilities using Statcrunch.
  - Stat -> Calculators -> Select the normal distribution.
  - Here you can choose the mean and standard deviation and calculate the area on the left or right. This area corresponds to the proportion of the population less than or greater than a value.
- [http://onlinestatbook.com/2/calculators/normal_dist.html](http://onlinestatbook.com/2/calculators/normal_dist.html)
Assume that female heights are normally distributed with a mean height of 64.5 inches and standard deviation is 2.5 inches. What proportion of females have a height less than 63 inches (hint: use the first page of the tables and not the second)?

- A. 60%
- B. 76%
- C. 27%
- D. -60%
Tips on using Table A

Because the Normal distribution is symmetric, there are 2 ways that you can calculate the area under the standard Normal curve to the right of a z value.

\[
\text{area to right of } z = \text{area to left of } -z \quad \text{or} \quad A(z) = 1 - A(-z)
\]
Symmetry of the normal

Compare the first and second page of the z-tables

The standard normal is symmetric at zero. The proportion to the left of zero is 50%. The proportion to the right of zero is 50%.

The proportion of the left of Z is the same as the proportion to the right of -Z. Check. The proportion to the left of Z = -1.2 is 0.1151. The proportion to the right of Z = 1.2 is 1 - 0.8849 = 0.1151.
More Calculations: Scores in SATs

One way to get admitted to A&M requires a score of at least 1300 on the combined critical reading and mathematics SAT exams. The SAT scores for 2010 were approximately normal with mean 1016 and standard deviation 212.

What proportion of students taking the SAT in 2010 have this requirement?

\[ x = 1300 \]
\[ \mu = 1016 \]
\[ \sigma = 212 \]
\[ z = \frac{(x - \mu)}{\sigma} = \frac{(1300 - 1016)}{212} \approx 1.34 \]

Table A: area under \( N(0,1) \) to the left of \( z = 1.34 \) is 0.9099.

\[ \text{area right of 1300} = \text{total area} - \text{area left of 1300} = 1 - 0.9099 = 0.0901 \]

Approximately 9.1% of students scored at least 1300.

Side note: The actual data may contain students who scored exactly 1300. However, the proportion of scores exactly equal to 1300 is zero for a normal distribution.
On the standard normal z-tables, what is the proportion to the right of $z = -1.43$?

- (A) 99.57%
- (B) 92.36%
- (C) 7.64%
- (D) 1.43%
- (E) -1.43%

http://www.easypolls.net/poll.html?p=59c1203de4b0e283b4d6a759
Assume that female heights are normally distributed with a mean height of 64.5 inches and standard deviation is 2.5 inches. What proportion of females have a height more than 62 inches?

- A. 15.87%
- B. 99%
- C. 69.15%
- D. 1%
- E. 84.13%
Tips on using Table A

To calculate the area between two \( z \)-values, first get the area under Normal(0,1) to the left of each \( z \)-value from Table A.

Then subtract the smaller area from the larger area.

A common mistake made by students is to subtract the \( z \) values instead of subtracting the areas.

The area between \( z_1 \) and \( z_2 \) is the area left of \( z_1 \) minus the area left of \( z_2 \).
**Calculation Practice:**

- **Question:** What proportion of females have height **between** 60.2 to 70 inches?

- **Answer:** Calculate the z-transform corresponding to 60.2

  \[ z_1 = \frac{60.2 - 64.5}{2.5} = -1.72 \]

  Look look the z-tables. Since -1.72 = -1.7 - 0.02, use the intersection of the column and row inside the table.

  We see from the table it is 0.0427.
Calculate the z-transform corresponding to 70

\[ z_1 = \frac{70 - 64.5}{2.5} = 2.2 \]

Look up 2.2 in the table. Since 2.2 = 2.2 + 0.00 we find the intersection of the row and column inside the table.

It is 0.9861

Since we want the proportion of heights between 60.2 and 70 we take the differences of the areas.

The answer is 0.9861 – 0.0427 = 0.9434.

94.34% of women are between 60.2 and 70 inches.
Question Time

- Assume that male heights are normally distributed with a mean height of 67 inches and standard deviation is 3.5 inches. What proportion of males have a height between 62 and 65.4 inches?
  - A. 26.11%
  - B. 24.7%
  - C. 64%
  - D. 1.36%
  - E. 2.64%

http://www.easypolls.net/poll.html?p=59baca91e4b00b7ab9d84733
There are various ways to gain entrance into Texas A&M. We mentioned on the previous slide SATs, but there are ACTs too.

The ACTs have a different scoring system to the SATs, these range from 1-36.

How to compare students who have taken different exams?

The easiest way is by comparing their percentiles, if one student A is in the top 10% SAT scores whereas student B is in the top 5% ACT scores. It is clear that student B did better in their exams.

This assumes that the same ability group took both set of exams.

**Question**: SATs have mean score 1025 and standard deviation 200, whereas ACT scores have mean 20 and standard deviation 5.

Betty scores 1400 on her SATs,

Joan scores 31 on her ACT.

Which student did `better'?
Answer Both students did better than the mean. But to compare the performance of each student we need know how well they did compared with everyone else who took the same exam.

- This means calculating the proportion of students who got more than 1400 in their SATs
- And more than 31 in their ACTs.

This requires us to know the scores of all students who did both exams.

Often such data is unavailable.

Instead we assume that SAT and ACT scores are close to normally distributed.

- SAT scores are almost normally distributed with mean 1025 and standard deviation 200.
- ACT scores are close normally distributed with mean 20 and standard deviation 5.

Assuming the scores are normally distributed allows us to make comparisons between distribution using only the knowledge of the mean and standard deviation of scores.
Making comparisons

- We have placed both SATs and ACTs on the same “scale”

Assuming the density of both SATs and ACTs allows us to make a proper comparison.

We see that Joan did slightly better.

We will make the calculation by hand on the next page.
The calculation: We first make a z-score for both Betty and Joan:

- Betty’s z-score is $z = (1400 - 1025)/200 = 1.875$
- Joan’s z-score is $z = (31 - 20)/5 = 2.2$.
- Both did better than the mean.

Using the normal tables:
- we see that Betty is in the 96.7 percentile,
- whereas Joan is in the 98.6 percentile.

So Joan did slightly better than Betty, since only 1.4% of students did better than Joan, whereas 3.3% students did better than Betty.

Equivalently, we can just compare z-transforms (since the scores are assumed to be normal). Joan did slightly better as her grade is 2.2 standard deviations right of the mean, whereas Betty is 1.875 standard deviations right of the mean. Since $2.2 > 1.875$, Joan did better.
We can also translate Joan’s grade into a SAT grade using the z-transform.

Since Joan is 2.2 standard deviations from the mean, this means if she took the SAT she would be 2.2 standard deviations from the SAT mean.

Thus Joan’s ACT grade translated into a SAT grade

\[
1025 + 2.2 \times 200 = 1465
\]

In other words, assuming that SAT and ACT grades were normally distribution. Joan’s equivalent SAT grade is 1465.
Assessing the validity of the calculations?

- In all statistical analysis we need to take a step back and ask ourselves whether the calculations were **meaningful**. Let's go through them step by step:
  - Comparing the percentiles for the grades in both exams is a reasonable thing to do. It gives us an idea of where each student stands with respect to the other students who took the same exam.
  - We first calculate the relative distance/z-score/z-transform.
  - To associate a probability to these values, we look up the z-scores in the z-tables.
  - Using the z-tables implies **we have assumed** that the distribution of grades for both SATs and ACTs are normally distributed.
  - It can be argued that SATs are normally distributed; the variable can only take integer scores, but 100-2100 is so fine there is very little difference between integers and non-integers.
How well does the normal density fit ACT grades?

Left, is the true distribution of ACT grades. By counting the height of the blocks which are less than or equal to 31, Joan’s score is in the 89% (percentile). This is the true probability. Comparing this to the normal calculation of 98.6% we see that the normal approximation over estimated Joan’s percentile. The ACT plot is far “thicker” in the tails than the normal distribution.
The impact of assuming normality

- Often data is assumed to come from a normal distribution.
- The main reason is that it makes calculation of percentiles very easy.
  - If a variable is normally distributed you only need the mean and standard deviation to calculate its percentile.
  - If it is not normally distributed, you need to find the density curve which closely approximates the data. This takes more effort or can be impossible to do.
- If you go to the medical facility, they may quote the percentile of your height or weight, which is calculated under the assumption the data is normally distributed.
- In the next slide, we investigate if the normal assumption makes sense for certain examples.
21% of baseball players do not earn negative salaries, as suggested by the normal distribution!

Plots of Baseball salaries and ACT scores. The red curve is the normal distribution. The areas under the red curves do not correspond to the true proportions in the population.

50% of ACT scores are not greater than the mean of 22, as the normal distribution suggests. It is more like 42%. The normal approximation gives about 1.2% scoring between 10-11. Whereas the actual data gives 2.5% (excluding 11 itself).
Aim: calculate proportion of calves whose weight at 0.5 weeks is less than 90 pounds.
Comparing calculations

- If we are selecting the calf from just the sample (the 44 calves that we were observing), then the chance is simply the sum of the heights of the bins less than 90, that is:
  - $0.341 + 0.159 + 0.068 = 0.568$. The proportion of calves less than 90 pounds is 56.8%.

- On the other hand if we want the proportion over the population of calves, we need calculate the area below the density plot of calf weights. The density plots of calf weights is unknown. If we have reason to believe the density plot of calves is close to a normal density plot, then we calculate the proportion using $z$-transforms:
  - The mean and standard deviation of the density plot is 90.11 and 7.7 respectively. Thus to calculate the probability we make a $z$-transform $z = (90 - 90.11)/7.7 = -0.014$. Looking this up on the tables gives a proportion close to 0.49 (49%). Hence, assuming the density plot of calf weights are normal the proportion of calves less than 90 pounds is 49%.

- We see that the normal approximation for this calculation is off by about 7%.

- We need to be careful when using the normal distribution.
Question: How well does the normal density approximate the proportions for this data?

A. The fit seems relatively good. Using the normal will give the correction proportions.
B. The normal will give the completely wrong answer for the proportion between 7.5 and 10.
C. The normal will give the wrong proportion for values less 7.5.
D. (B) and (C)

http://www.easypolls.net/poll.html?p=59c12506e4b0e283b4d6a75f
A farmer wants to enter either his cow or pig for the heaviest animal competition. The winning animal is the heaviest animal in its category (cows or pigs).

- It is known that the weight of cows is approximately normally distributed with mean 280 pounds and standard deviation 20 pounds ($N(280,20)$).
- The weight of pigs is approximately normally distributed with mean 250 pounds and standard deviation 50 pounds ($N(250,50)$).

His prize cow weighs 330 pounds and prize pig weighs 310 pounds. The contest only allows one animal per farmer, which animal should he enter?

A. His prize Cow  
B. His prize Pig

http://www.easypolls.net/poll.html?p=59c143fde4b04d2fa1bd451b
Solution

- It makes sense to enter the heaviest animal is relative to its species.
- The z-score for the cow = \( \frac{330-280}{20} = 2.5 \) standard deviations from the mean. This corresponds to the 99.3% percentile.
- The z-score for the pig is \( \frac{310-250}{50} = 1.2 \) this corresponds to the 88.4 percentile. Despite the pig’s weight lying further from the mean, there is a lot of variation in pig weight.
- The farmer should enter the cow, since only 0.7% of cows are heavier than her.
It is important to monitor a foal in the first few days after birth. In particular it is important that the mare gives the foal essential colostrum. In order to do this, the foal must feed frequently from their mother (on average a newborn foal feeds 4 times in an hour). However, if the foal feeds too frequently, this suggests the mother is not providing enough colostrum and the foal may need veterinary assistance.

The distribution of feed times (time between the feeds) of healthy foals follow a normal distribution. But for each breed of horses the distributions vary slightly:

- Breed 1: $N(15, 3)$ (mean 15 minutes, standard deviation 3)
- Breed 2: $N(15, 2)$ (mean 15 minutes, standard deviation 2)

Suppose a vet uses a blanket threshold and examines any foal whose feed time drops less than 13 minutes. What are the implications of this policy?

A. Breed 1 will be examined more often than Breed 2.
B. Breed 2 will be examined more often than Breed 1.
C. Both breeds will be examined about the same.
Answer

- Each breed has its own distribution for healthy horses. But the same blanket threshold is used:

Because there is more variability (standard deviation) for breed one than breed two (but both have the same mean) we see that breed one will be examined more often than breed two.

This probabilities can be calculated using the z-score and looking but the normal tables.
Question Time

- Same question again but now we compare Breed 2 with Breed 3.
- The distribution of feed times (time between the feeds) of healthy foals follow a normal distribution. But for each breed of horses the distributions vary slightly:
  - Breed 2: \( N(15, 2) \) (mean 15 minutes, standard deviation 2)
  - Breed 3: \( N(16, 4) \) (mean 16 minutes, standard deviation 4)
- Suppose a vet uses a blanket threshold and examines any foal whose feed time drops less than 13 minutes. This means, in general:
  A. Breed 2 will be examined more often than Breed 3.
  B. Breed 3 will be examined more often than Breed 2.
Within z-standard deviation calculations

Question: If a population is normally distributed what proportion of the population will lie within 0.5 standard deviations from the mean.

Answer: This looks like mission impossible!

- You may be asking, what is the mean, what is the standard deviation?? The point is that this information is not required.
- When we say 0.5 standard deviations from the mean this is the same as the z-score being $= -0.5$ and $0.5$.

Here we have two plots. One for the distribution of heights, which are assumed normal with mean $64.5$ and standard deviation $2.5$. The other is a standard normal with mean zero and standard deviation one. The blue area corresponds to 0.5 standard deviations from the mean. The blue area for both plots is the same.
The area between -0.5 and 0.5 is $0.6915 - 0.3085 = 0.383$.

About 38.3% of the population are within 0.5 standard deviations of the mean.

Remember the number of standard deviations from the mean, corresponds to the z-score.

Reason why: if the mean is $\mu$ and standard deviation is $\sigma$, then the point that is 0.75 standard deviations to the right of the mean is

$$\mu + 0.75 \times \sigma$$

The z-score of this is

$$z = \frac{\mu + 0.75 \times \sigma - \mu}{\sigma} = 0.75$$
Question Time

If a distribution is normally distributed what proportion of the population will lie within 1.5 standard deviations of the mean?

- A. 93.32%
- B. 6.68%
- C. 86.64%
- D. 1.5%

http://www.easypolls.net/poll.html?p=59c1256be4b04d2fa1bd44b4
Inverse normal calculations

We may also want to find the observed range of values that correspond to a given proportion/ area under the curve.

For that, we use Table A **backward**:

- We first find the desired area/ proportion in the *body of the table*.
- We then read the corresponding *z*-value from the left column and top row.

For an area of 1.25% (0.0125) to the left of $z$, the *z*-value is $-2.24$. 

**TABLE A Standard normal probabilities**

<table>
<thead>
<tr>
<th>$z$</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.3</td>
<td>.0005</td>
<td>.0005</td>
<td>.0005</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0003</td>
</tr>
<tr>
<td>-3.2</td>
<td>.0007</td>
<td>.0007</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0005</td>
<td>.0005</td>
</tr>
<tr>
<td>-3.1</td>
<td>.0010</td>
<td>.0009</td>
<td>.0009</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0007</td>
<td>.0007</td>
</tr>
<tr>
<td>-3.0</td>
<td>.0013</td>
<td>.0013</td>
<td>.0012</td>
<td>.0012</td>
<td>.0011</td>
<td>.0011</td>
<td>.0011</td>
<td>.0011</td>
<td>.0010</td>
<td>.0010</td>
</tr>
<tr>
<td>-2.9</td>
<td>.0019</td>
<td>.0018</td>
<td>.0018</td>
<td>.0017</td>
<td>.0016</td>
<td>.0016</td>
<td>.0015</td>
<td>.0015</td>
<td>.0014</td>
<td>.0014</td>
</tr>
<tr>
<td>-2.8</td>
<td>.0026</td>
<td>.0025</td>
<td>.0024</td>
<td>.0023</td>
<td>.0023</td>
<td>.0022</td>
<td>.0021</td>
<td>.0021</td>
<td>.0020</td>
<td>.0019</td>
</tr>
<tr>
<td>-2.7</td>
<td>.0035</td>
<td>.0034</td>
<td>.0033</td>
<td>.0032</td>
<td>.0031</td>
<td>.0030</td>
<td>.0029</td>
<td>.0028</td>
<td>.0027</td>
<td>.0026</td>
</tr>
<tr>
<td>-2.6</td>
<td>.0047</td>
<td>.0045</td>
<td>.0044</td>
<td>.0043</td>
<td>.0041</td>
<td>.0040</td>
<td>.0039</td>
<td>.0038</td>
<td>.0037</td>
<td>.0036</td>
</tr>
<tr>
<td>-2.5</td>
<td>.0062</td>
<td>.0060</td>
<td>.0059</td>
<td>.0057</td>
<td>.0055</td>
<td>.0054</td>
<td>.0052</td>
<td>.0051</td>
<td>.0049</td>
<td>.0048</td>
</tr>
<tr>
<td>-2.4</td>
<td>.0082</td>
<td>.0080</td>
<td>.0078</td>
<td>.0075</td>
<td>.0073</td>
<td>.0071</td>
<td>.0069</td>
<td>.0068</td>
<td>.0066</td>
<td>.0064</td>
</tr>
<tr>
<td>-2.3</td>
<td>.0107</td>
<td>.0104</td>
<td>.0102</td>
<td>.0099</td>
<td>.0096</td>
<td>.0094</td>
<td>.0091</td>
<td>.0089</td>
<td>.0087</td>
<td>.0084</td>
</tr>
<tr>
<td>-2.2</td>
<td>.0139</td>
<td>.0136</td>
<td>.0132</td>
<td>.0129</td>
<td>.0126</td>
<td>.0123</td>
<td>.0122</td>
<td>.0120</td>
<td>.0118</td>
<td>.0116</td>
</tr>
<tr>
<td>-2.1</td>
<td>.0173</td>
<td>.0170</td>
<td>.0166</td>
<td>.0163</td>
<td>.0160</td>
<td>.0158</td>
<td>.0156</td>
<td>.0153</td>
<td>.0150</td>
<td>.0148</td>
</tr>
</tbody>
</table>
Example: Female heights

- Female heights tend to be normally distributed with $N(64.5,2.5)$.

Questions:
- (a) How tall is a female in the 75% percentile?

Answers:
- (a) Look up 0.75 inside the z-table, it is 0.674. This means that someone who is in the 75 percentile is 0.674 standard deviations to the right of the mean. That person is $64.5 + 0.674 \times 2.5 = 66.2$ inches tall.
Example: Female heights

- Female heights tend to be normally distributed with $N(64.5, 2.5)$.
- Questions:
  - (b) How tall is a female who is in the top 10% percentile?
- Answers:
  - (b) Top 10% = 90% percentile. Look up 0.9 in the $z$-table, it is 1.28. Using the same argument as above that person is $64.5 + 1.28 \times 2.5 = 67.7$ inches tall.
Female heights tend to be normally distributed with $N(64.5,2.5)$ (mean 64.5 inches and standard deviation 2.5 inches). How tall is a female in the top 20$^{th}$ percentile for heights?

A. 66.63 inches  
B. 66.5 inches  
C. 62.38 inches  
D. 65 inches

http://www.easypolls.net/poll.html?p=59c15479e4b04d2fa1bd4548
Example: Female heights

- Female heights tend to be normally distributed with $N(64.5, 2.5)$.

Questions:
- (c) How tall is a female who is in the bottom 2.5% percentile?

Answers:
- (c) Look up 0.025 in the z-tables – 1.96. Using the same argument as above that person is $64.5 - 1.96 \times 2.5 = 59.6$ inches tall.
Question: Construct an interval centered about the mean, where 95% of female heights lie.

Answer: Look up 2.5% and 97.5% in z-tables; [-1.96, 1.96] (this interval is symmetric about 0).

This interval stays that 95% of heights will lie within 1.96 standard deviations (either way) of the mean.

Translating this into heights we have that 95% of heights lie between [64.5 -1.96 × 2.5, 64.5 + 1.96 × 2.5] = [59.6,69.4] inches.

Observe and compare 1.96 standard deviations from the mean to the 68-95-99.7% rule which corresponds to 1 standard deviation, 2 standard deviation and 3 standard deviations from the mean.

2 standard deviations is in fact an approximation of 1.96 standard deviations from the mean.
Female heights tend to be normally distributed with mean 64 inches and standard deviation 2.5: \( N(64.5,2.5) \). Construct an interval centered about the mean of 64.5 where 80% of heights will lie.

A. \([61.3, 67.7]\]

B. 66.63

C. \([62.38, 66.63]\]

http://www.easypolls.net/poll.html?p=59c156a1e4b04d2fa1bd4551
Topic: QQplots

- Learning Targets:
  - Understand that QQplots are a graphical tool that allows to see if the data has approximately a normal density curve.
  - Understand how deviations from the straight line explain features about the underlying true distribution.
Using QQplots to check normality of data

- By simply superimposing a normal distribution over a histogram it is difficult to check how close the distribution is to normal.
- Typically to check for normality of data we make a QQplot.
- The idea is behind this plot is similar to checking the 68-95-99.7% but extended to all multiples of the standard deviation not just 1, 2, and 3.
- The data is close to normal if the points lie along the x=y line.
- Warning: The QQplot has **nothing** to do with linear regression. The line that you see in the plot is **not** the line of best fit.
- In the following few slides we consider a few examples:
- Making a QQplot in Statcrunch.
Observe that most points lie close to $x=y$ line. Few in the tails lie off the line.
For right skewed data the QQplot has a U-type shape.
QQplot for left skewed data looks like an inverted U
QQplot for uniform and thick tailed data (data whose tails are not much thinner than the center) have an S shape.
In this data set, the response is either 0 or 1. The vertical lines correspond to each of these responses.
The horizontal lines we see are due to several weights having the same value (due to rounding). Eg. The first horizontal line corresponds to 5 calves with the same weight. The weights are not exactly normal, but it does not deviate massively from normality.
Question Time

Based on the plot below which statement is correct

- (A) The horizontal lines show that the ACT scores are integer valued
- (B) The distribution of ACT scores differ from normality especially in the right and left tails.
- (C) The ACT scores lie close to the line and are normal.
- (D) There is a clear linear correlation in the ACT scores.
- (E) A and B.
- (F) C and D.
Accompanying problems associated with this Chapter

- Quiz 4
- Quiz 4 – parts 2
- Homework 2