**Afterward**

You've now learned many methods of using statistics to analyze data and draw conclusions. However, that's the easy part. The hard part is determining which test to use and for what purposes. For example:

- At-risk youth go through summer chess programs with mentors. How would you test whether or not the summer chess program helped at-risk youth perform better in school?

- You're curious to know how chivalry towards women is changing as more women reach leadership positions in the corporate world. How would you measure "chivalry"? What data would you collect and where would you obtain your sample?

- You want to know how educational attainment differs by the predominant sector (agriculture, services, or industry) of the area in which subjects live. What test(s) would you use?

Hopefully, whenever you read any conclusions based on statistics tests you'll be able to critique the methodologies used. In turn, you'll get better and better at determining robust statistical research methods.

The journey doesn't end here. Feel free to post any questions about statistics at turnthewheel.org/street-smart-stats.

**Further resources:**

- National Center for Education Statistics ELS2002 Dataset of select variables
    - If you are reading Statistics Fundamentals Succinctly, I have re-created this dataset so it will have slightly different data than shown in the book's examples. However, the variables are the same and you can perform the same operations in R.
- http://www.comscore.com/Insights/Data-Mine - lots of fun findings based on data
- http://infosthetics.com/ - awesome data visualizations
- http://www.shodor.org/interactivate/activities/Histogram/ - see how histogram changes as you adjust the bin size
- You can analyze any public microdata (e.g., http://www.bls.gov/data/, http://nces.ed.gov/datatools/index.asp?DataToolSectionID=4) to analyze trends and relationships

**Answers to textbook quizzes**

**Analyze distributions, Lesson 3**
1. Which is true about this distribution?

Since the distribution is skewed to the right (the tail being longer on the right side), there are some large values that will affect the mean more than the median. The mode will still be where the highest frequency occurs. So we would expect that mode < median < mean.

2. Which symbols (<, >, =) should go in the blanks to make this statement true for this distribution?

This is a normal distribution, in which case mean = median = mode.

3. In the table below, the row headers list positive characteristics that we'll ideally have in a measure of center. Which characteristics are true for each measure?

| | Mean | Median | Mode |
|---|---|---|---|
| Has a simple equation | 🟩 | | |
| Will always change if any data value changes | 🟩 | | |
| Not affected by change in bin size | 🟩 | 🟩 | |
| Not affected severely by outliers | | 🟩 | 🟩 |
| Easy to find on a histogram | | | 🟩 |

**Standard deviations, Lesson 4**
1. If the mean is 10 and the standard deviation is 5, which values are more than one standard deviation from the mean?

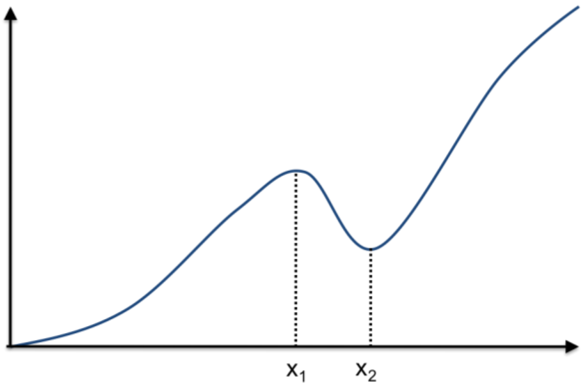
16 (answer choice B), 2 (answer choice C), and -1 (answer choice E) are correct since they are all more than 5 units from 10 (the mean). Since the standard deviation is 5, this means that these values are more than one standard deviation from the mean.

2. The mean is 12 and the standard deviation is 4. Which value is 1.5 standard deviations from the mean?

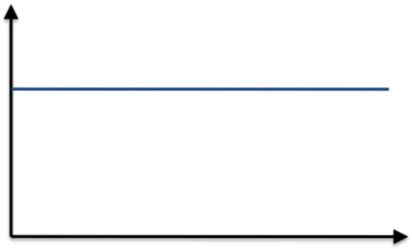
1.5 standard deviations is 1.5*4 = 6. Values 6 units from the mean (12) are 6 and 18.

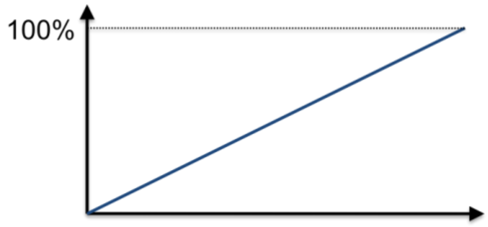
**PDF and CDF, Lesson 6**

1. No, this CDF is not possible. Let’s look at the two points on the x-axis, labeled below. This CDF implies that there are a certain proportion of values less than $x_1$, but even fewer values less than $x_2$, even though $x_2$ is greater than $x_1$. This would mean there are a negative number of values in-between $x_1$ and $x_2$, which doesn’t make sense.

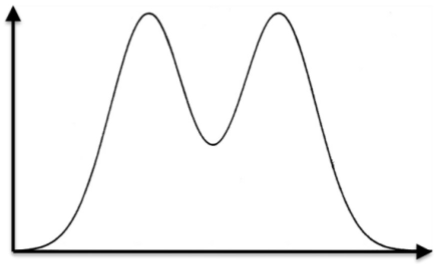


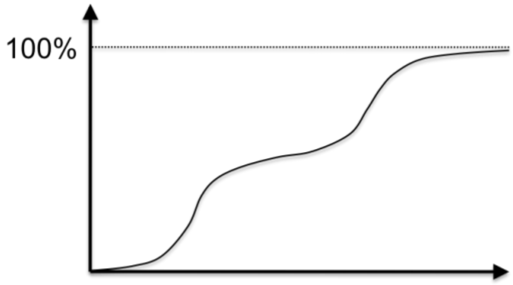
2.





3.





**Sampling Distributions, Lesson 7**

1. Find the z-score for a sample mean of 40 from sample of size 250 from the population of Klout scores. Remember the population mean $\mu$ is 37.719 and the population standard deviation $\sigma$ is 16.037.

   z = $\frac{40 - 37.719}{16.037/\sqrt{250}}$ = 2.25

2. Using the same example from #1, use the z-table to find the proportion of samples of size 250 with means less than 40.
   Will the proportion of samples of size 250 with means less than 40 will be higher or lower for n = 250 than n = 20?

   The proportion of samples of size 250 with means less than 40 is 0.9878. This proportion is much higher than it would have been for n = 20 because samples of size 20 are more prone to error (the sampling distribution will have a wider spread) and therefore it's more likely for a sample of size 20 to have a mean greater than 40 than it would be for a sample of size 250.

3. Approximately ___% of sample means fall within ___ of the population mean?
   d. 95%; $2\sigma/\sqrt{n}$

**Point estimate and CI, Lesson 8**

In a certain city, all coffee shops opened at 7:00 am. After lower-than-desired returns on investment (ROI), 12 coffee shops decided to open at 6:00 am. The following quarter, they saw their ROI rise to an average of 17%. Before, the average return on investment for all coffee shops was 11%, with standard deviation 2.4%.

1. If all coffee shops decided to open at 6:00 am, what would be our best guess for average

ROI for the population of coffee shops?

Our best guess is 17% (our point estimate is the sample mean).

2. Since we can't know for sure what the new ROI average for the population would really be, what is a confidence interval that we can be reasonably sure will contain average ROI?

   Our approximate 95% confidence interval will be $\pm 2$ standard errors from the sample mean:
   $(\bar{x}_I - \frac{2\sigma}{\sqrt{n}}, \bar{x}_I + \frac{2\sigma}{\sqrt{n}})$ = (17% - $\frac{2(2.4\%)}{\sqrt{12}}$, 17% + $\frac{2(2.4\%)}{\sqrt{12}}$) = (15.61%, 18.39%)
   This new confidence interval does not even contain the old population mean (11%), so we can guess that opening at 6:00 am had a treatment effect.

**CI for larger sample, Lesson 8**

A sample of size 24 should have a smaller confidence interval because larger sample sizes result in greater precision. This time the confidence interval would be

(17% - $\frac{2(2.4\%)}{\sqrt{24}}$, 17% + $\frac{2(2.4\%)}{\sqrt{24}}$) = (16.02%, 17.98%)

**Significance Levels, Lesson 9**
Perform a *one-tailed test* for the following z-scores. Determine the alpha level at which the z-score is significant; in other words, fill in the second blank in the following statement:

A z-score of ___ is significant at the ___ level.
*E.g., A z-score of -1.78 is significant at the 0.05 level.*

| | |
|---|---|
| z-score = 3.15 | $\alpha = 0.001$ |
| z-score = -3.01 | $\alpha = 0.01$ |
| z-score = -2.29 | $\alpha = 0.05$ |

**Type I and Type II errors, Lesson 9**
1. You order a hot beverage at a cafe. You don't know how hot it is. If it's too hot and you take a sip, you'll burn your tongue. But maybe the beverage is just right, and if you wait it'll be too cold.

$H_0$: The beverage is fine to drink now.
$H_a$: The beverage is too hot to drink.

| | | **Decision** | |
|---|---|---|---|
| **Truth** | **$H_0$ true** | **Type I error**<br>You think the beverage is too hot so you wait to drink it, but it's actually just right and by the time you drink it, it's too cold. | **Correct**<br>You decide the beverage is fine to drink now, and it is! |
| | **$H_0$ false** | **Correct**<br>You decide the beverage is too hot, so you wait to drink it. Indeed, it was too hot, so when you drink it it's perfect! | **Type II error**<br>You decide the beverage is fine to drink now, but it's too hot and you burn your tongue. |

2. You're about to go out to run a few errands, and you're not sure if it's going to rain in the next hour. If it rains, you want to bring your umbrella. But if it doesn't, you'll end up carrying your umbrella around for nothing, which will be annoying.

$H_0$: It's not going to rain.<br>
$H_a$: It will rain.

| | | **Decision** | |
|---|---|---|---|
| | | **Reject $H_0$**<br>You bring your umbrella | **Retain $H_0$**<br>You don't bring your umbrella |
| **Truth** | **$H_0$ true**<br>It doesn't rain | **Type I error**<br>You think it will rain, but it doesn't. You carry your umbrella around for nothing. | **Correct**<br>You think it's not going to rain, and you're right! |
| | **$H_0$ false**<br>It rains | **Correct**<br>You think it will rain, and it does! Good thing you brought your umbrella. | **Type II error**<br>You think it's not going to rain, but it does and you get drenched. |

**Two-tailed test, Lesson 9**

$H_0$: $\mu_I = \mu$<br>
$H_a$: $\mu_I \neq \mu$

These are the things we know:

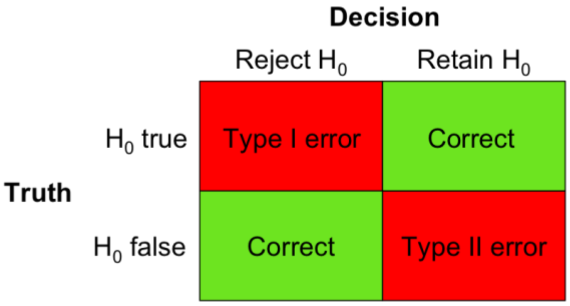
| | |
|---|---|
| Population mean: | $\mu$ = 7.47 |
| Population standard deviation: | $\sigma$ = 2.41 |
| Sample size: | $n$ = 30 |
| Sample mean: | $\bar{x}$ = 8.3 |

1. Based off what we know, what decision would we make: reject the null or fail to reject the null?

   The z-score of our sample mean is (8.3 - 7.47)/(2.41/$\sqrt{30}$) = 1.89. Since we're doing a two-tailed test at $\alpha = 0.05$, the critical values are $\pm 1.96$. 1.89 does not land in this critical region; therefore we fail to reject the null.

2. Let’s say that the true population mean after the intervention is actually 7.8 ($\mu_I = 7.8$). Which quadrant of the decision/truth grid played out?

   The true population mean 7.8 is (7.8 - 7.47)/(2.41/$\sqrt{30}$) = 0.75 standard errors above $\mu$ which is also not in the critical region. (We know this without calculating the z-score simply by observing that 7.8 is less than the sample mean 8.3, and we’ve seen 8.3 falls short of the critical region.) Therefore, in this case we made the correct decision: we failed to reject H0 when in fact $H_0$ was true (top right quadrant).



| | | Decision | |
|---|---|---|---|
| | | Reject $H_0$ | Retain $H_0$ |
| Truth | $H_0$ true | Type I error | Correct |
| | $H_0$ false | Correct | Type II error |

**Practice using the t-table, Lesson 10**

1. Find t-critical value for a one-tailed alpha level of 0.05 with 12 degrees of freedom.
   t* = 1.782 or -1.782

2. n = 30. Find t-critical values for two-tailed test at alpha = 0.05.
   t* = ±2.045

3. Is a t-statistic of 2.641 significant at $\alpha = 0.05$ (n = 5) for a two-tailed test?
   No; the t-critical values are ±2.776, and the t-statistic of 2.641 does not lie outside these critical values. Therefore it’s not significant and we would fail to reject the null.

4. Is a t-statistic of 2.641 significant at $\alpha = 0.05$ (n = 20) for a two-tailed test?
   Yes. Obtaining the same t-statistic for a larger sample size results in more certainty that the sample mean approximates the population mean, and therefore we’re more certain that the population mean is closer to 2.641. This results in t-critical values that are closer to 0: ±2.093.

**Finch Beak Widths, Lesson 10**

Let's tackle this problem using R. First, we'll download the finch beak width data as a csv file and save it to our working directory.

In the Google spreadsheet,
File > Download as > Comma-separated values (.csv)

Change name of file to finches.csv

Drag the file to your working directory. (Type `getwd()` to find out what your working directory is.)

```
finches = read.csv(file = "finches.csv", head = TRUE, sep = ",")
attach(finches)
mean(beak_width)
sd(beak_width)
```

You should get that
$\bar{x}$ = 6.47 and $s$ = 0.40.

Find the t-statistic, $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$:

```
(6.47-6.07)/(0.4/sqrt(500))
```

You should get 22.36. This is huge! Way larger than the two-tailed t-critical value of ~1.984. Therefore, we reject $H_0$ and conclude that the finches on the Galápagos Islands do indeed have different-sized beaks than normal ones.