PSTAT 5LS Lab 6

YOUR NAME HERE

Week of November 11, 2024

Section 1

Announcements & Recap

Announcements

- Homework 6 due by 11:59pm on F 11/15
- ECoach journal entry 8 due by 11:59pm on M 11/18
- Exam 2 during lecture on R 11/21
    - Slide Set 5: Inference for One Proportion (Theory)
    - Slide Set 6: Inference for One Mean
    - Slide Set 7: Inference for the Difference in Two Means
    - HW 5-6
    - Name That Scenario: determine if the parameter of interest is p, μ, $\mu_1 - \mu_2$, μ_d
    - Don't forget about the ideas in Slide Sets 1 and 2 (e.g., observational study or experiment, explanatory and response variables, describing a distribution using a histogram)

Section 2

Learning Objectives

R Learning Objectives

1. Create a graphical display of the t distribution.
2. Interpret R output providing confidence intervals and hypothesis tests for inference on population means.

Statistical Learning Objectives

1. Learn to use the t distribution
2. Get experience making confidence intervals for population means
3. Understand hypothesis tests for population means

Functions Covered in this Lab

1. `plot_t()`
2. `pt()`
3. `qt()`
4. `t.test()`
5. `subset()`

Section 3

Lab Tutorial

One Population Mean

We are shifting our focus from categorical data (proportions) to *numeric* data (means). Inference for one mean is a new tool in our statistical toolkit that will let us answer different questions about data. Keep in mind that the parameter we're interested in now is μ (mu), the population mean.

The t Distribution

Recall from lecture that the t **distribution**, like the standard normal distribution ($N(0,1)$) that we've seen before, is symmetric about zero and bell-shaped.

The t distribution, however, has "thicker" ("heavier") tails than the normal distribution. This is so that we can account for the increased uncertainty introduced when we use the sample standard deviation s to estimate the population standard deviation σ.

The t Distribution

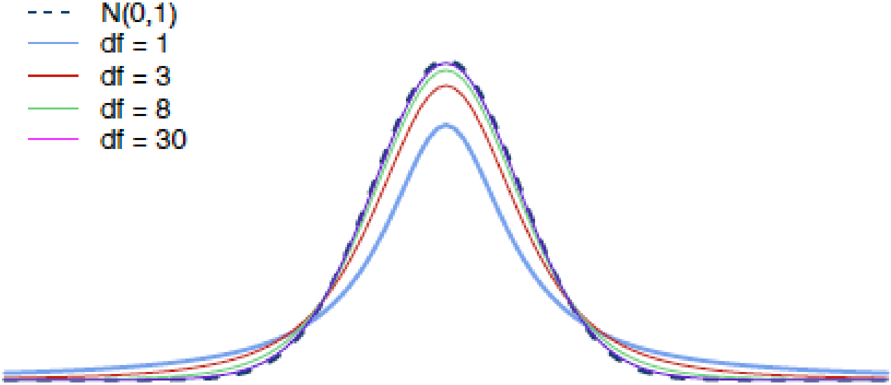
The t distribution is a “family” of distributions, meaning there are an infinite number of t distributions. The same is true of the normal distribution – there are infinite different normal distributions.

We describe which normal distribution we need to specify two parameters, the mean μ and the standard deviation σ.

We identify which t distribution we're talking about using just one parameter, the **degrees of freedom (df)**.

t Distribution Versus the Standard Normal Distribution

As the degrees of freedom increase, the t distribution gets closer and closer to a standard normal distribution, $N(0,1)$:



Graphing the t Distribution

We can use the `plot_t()` function in the `stats250sbi` package to make a graphical display of the t distribution. As you might expect, `plot_t()` is similar to `plot_norm()`.

Arguments to send to `plot_t()`:

- `df`: This is the degrees of freedom for the sample; this week, we can find `df` by computing $n - 1$, where n is the sample size.
- `shadeValues`: This is the value(s) that we wish to identify, and shade either to the left or to the right of. To include two values we will combine them by using the `c()` function.
- `direction`: This is the direction to shade. The choices are: “less”, “greater”, “beyond”, “between”. The text must be written in double quotes.
- `col.shade`: Optional color choice for the graph.

Recall an Example from Lecture

In lecture, you worked with data from a random sample of 53 female ring-tailed lemurs at the Duke Lemur Center in North Carolina. The question was whether the average weight for female ring-tailed lemurs differs from the 2213 gram average weight for female ring-tailed lemurs in the wild.

Our hypotheses were

$$H_0 : \mu = 2213 \text{ and } H_A : \mu \neq 2213$$

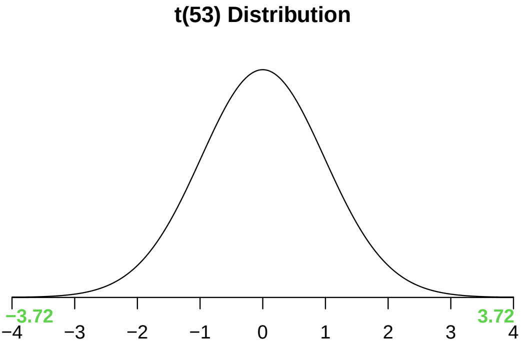
The test statistic for that example was

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2406.6 - 2213}{\frac{378.7}{\sqrt{53}}} = 3.722$$

Graphing the *t* Distribution

The p-value for this hypothesis test is the area in both tails of the $t(52)$ distribution. Run the `tryIt1` chunk in your notes to plot the p-value for this example.

```
plot_t(3.722, df = 53,
       shadeValues = c(-3.722, 3.722),
       direction = "beyond")
```



Finding Probabilites Under a t Distribution

We can find probabilities related to the t distribution using the `pt()` function. The `pt()` function is analogous to the `pnorm()` function we used with normal distributions.

Arguments to send to `pt()`:

- `q`: The observation is sometimes called a **quantile**, thus the **q** notation.
- `df`: This is the degrees of freedom for the sample; this week, we can find `df` by computing $n - 1$, where n is the sample size.
- `lower.tail`: By default, this argument is set to `TRUE`, meaning that we *want* the lower tail (i.e., to shade to the left). If we *don't want* the lower tail, and we actually want the *upper* tail (i.e., to shade to the right), we should set this to `FALSE`.

The mean of a t distribution is 0, and the standard deviation of a t distribution is a function of the the degrees of freedom, so we *don't* need to specify the mean and the standard deviation to use `pt()`.

Using pt() to Find a p-value

The p-value for the lemur example from lecture was approximately 0.0005. How would we have calculated that if it had not been in the R output for the hypothesis test?

Our test statistic is $t = 3.722$, the degrees of freedom are $df = n - 1 = 52$, and we had a two-sided alternative. Let's calculate the p-value in the `tryIt2` code chunk in your notes. Don't forget to specify the value of the test statistic and the degrees of freedom.

```
2*pt(q = 3.722, df = 52, lower.tail = FALSE)
```

```
## [1] 0.0004865675
```

Finding the Quantile on a *t* Distribution

Also, we can use `qt()` to get quantiles of the t distribution. The `qt()` function will be helpful to find t^* critical values needed for confidence intervals for means.

Arguments to send to `qt()`:

- `p`: The probability to the **left by default** of the quantile we wish to find. If we want the probability to the *right*, we should tinker with `lower.tail` as specified below.
- `df`: This is the degrees of freedom for the sample; this week, we can find `df` by computing $n - 1$, where n is the sample size.
- `lower.tail`: By default, this argument is set to `TRUE`, meaning that we *want* the lower tail (i.e., to shade to the left). If we *don't want* the lower tail, and we actually want the *upper* tail (i.e., to shade to the right), we should set this to `FALSE`.

Back to the Penguins!

Run the `loadPenguins` code chunk in your notes so that we can use the penguins data set.

Let's say we want to construct a confidence interval for the population mean flipper length of penguins living in the Palmer Archipelago, or conduct a hypothesis test for that mean.

In order to use our technology for constructing confidence intervals and performing hypothesis tests for means, we need two conditions to hold. **What are they?**

Conditions for a CI or HT for the Population Mean

The conditions are

1. The observations must be *independent* of one another.
2. When the sample is small, we require that the *sample observations come from a normally distributed population*. We can relax this condition more and more for larger and larger sample sizes.

How can we check that we meet both of these conditions?

Checking the Conditions for a CI or HT for the Population Mean

To check the conditions:

1. Verify that we have taken a random sample from the population. If we don't have a random sample, we should consider whether it's reasonable or not to believe that the observations are independent of one another.
2. Examine a histogram of the sample data. When the sample is small, we require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger and larger sample sizes.

Saying "small *n*" and "large *n*" is still vague, so here are a few guidelines:

- slight skew is okay for sample sizes up to about 15
- moderate skew is okay for sample sizes up to about 30
- strong skew is okay for sample sizes of about 60 or more

Are the Observations for the Penguins Data Set Independent of One Another?

Do you think that the penguins data is from a random sample? If not, do you think that the observations of penguins are independent of one another?

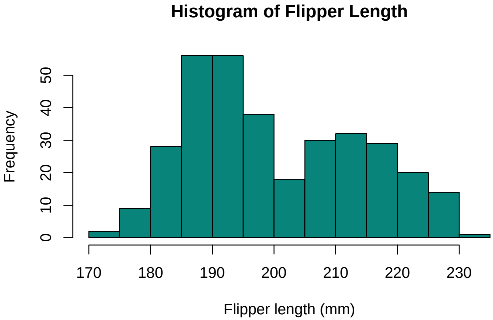
Are Penguin Flipper Lengths Normally Distributed?

Let's make a histogram in the `tryIt3` code chunk of your notes document of the penguin flipper length. (Don't forget to first run the chunk to read in the data!)

```
hist(penguins$flipper_length_mm,
     main = "Histogram of Flipper Length",
     xlab = "Flipper length (mm)",
     col = "#09847A")
```

Do you think that the distribution is approximately bell-shaped?

Histogram of Penguin Flipper Lengths



Dealing with a Bimodal Distribution

In lecture, Professor Miller has talked about bimodal distributions being indicative of subgroups. Flipper length differs for the three species of penguins, so we need to look at them individually.

For the purpose of our work in this lab, we will work the flipper lengths for the Gentoo penguins.

Subsetting the Data

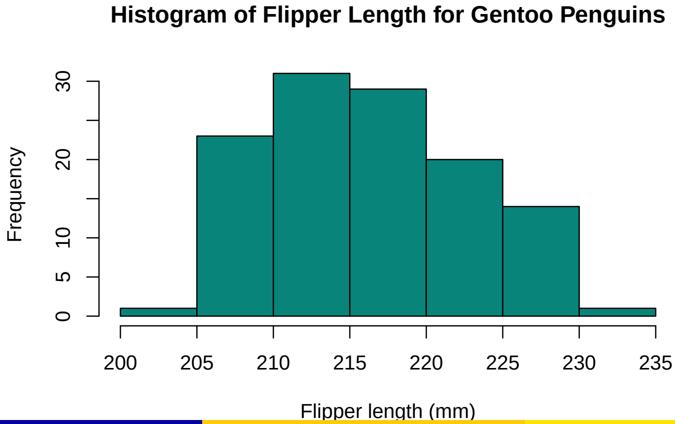
Run the gentoo_subset code chunk in your notes document to create a data set with only the 119 Gentoo penguins.

```
gentoo <- subset(penguins, penguins$species == "Gentoo")
```

Note that the Environment pane now has gentoo listed.

Create a Histogram to Examine the Distribution

Let's examine the flipper lengths for the Gentoo penguins in the `tryIt4` code chunk. Be sure to specify your variable and a label for the x-axis (we've taken care of the title for you).



Checking Normality

Use the histogram you created to check the normality condition. Does it appear that the flipper lengths for the Gentoo penguins came from a normally distributed population. If not, what can we do?

Hypothesis Test for a Mean

When we have the data, we can use a new function called `t.test()` to have R compute the confidence interval for the mean flipper length.

Note: We must have the data to run `t.test()`. If we only have summary statistics, we need to calculate the test statistic by hand and then use the function `pt()` in R to calculate the p-value.

Hypothesis Test for a Mean

According to Australia's Department of Climate Change, Energy, the Environment and Water, Gentoo penguins have an average flipper length of 230 mm. Is the average flipper length for the Gentoo penguins living on Palmer Archipelago different? Use $\alpha = 0.05$.

Thus we are testing:

$H_0 : \mu = 230$

$H_a : \mu \neq 230$

where μ is the mean flipper length in the population of Gentoo penguins living in the Palmer Archipelago.

Using t.test() to Run this Hypothesis Test

Since we have the data in R, let's use `t.test()` to run the hypothesis test.

Try this code out in the `tryIt5` code chunk in your notes document. Make sure to specify the hypothesized mean in the `mu` argument and the alternative in the `alternative` argument ("two.sided", "less", "greater").

```r
t.test(gentoo$flipper_length_mm,
       mu =  230,
       alternative =  "two.sided")
```

Using t.test() to Run this Hypothesis Test

```
##
##  One Sample t-test
##
## data:  gentoo$flipper_length_mm
## t = -21.145, df = 118, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 230
## 95 percent confidence interval:
##  216.0398 218.4308
## sample estimates:
## mean of x
##  217.2353
```

To Reject or Not to Reject?

When R reports a p-value of $<$ 2.2e-16, it is telling us that the p-value is less than 2.2×10^{-16}. In other words, the p-value is close to zero. Because our p-value is less than the significance level of 0.05, we reject the null hypothesis.

Conclusion for the Hypothesis Test

We've decided to reject the null hypothesis, so what's the conclusion?

Creating a Confidence Interval for the Mean Flipper Length

Since we rejected the null hypothesis that the mean flipper length of Gentoo penguins is 230 mm, it makes sense to come up with an estimate for this parameter.

Let's create a 95% confidence interval for the mean flipper length of Gentoo penguins.

First, let's do this by hand.

Creating a Confidence Interval for the Mean Flipper Length

As you learned in lecture, when we have a sample of n independent observations from a nearly normal distribution, a confidence interval for the population mean μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

Creating a Confidence Interval for the Mean Flipper Length

Here are the summary statistics for the mean flipper length of Gentoo penguins:

```
summary(gentoo$flipper_length_mm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   203.0   212.0   216.0   217.2   221.5   231.0 
```

```
sd(gentoo$flipper_length_mm)
```

```
## [1] 6.585431
```

Run the `tryIt6` chunk in your notes to have R calculate these summary statistics for you.

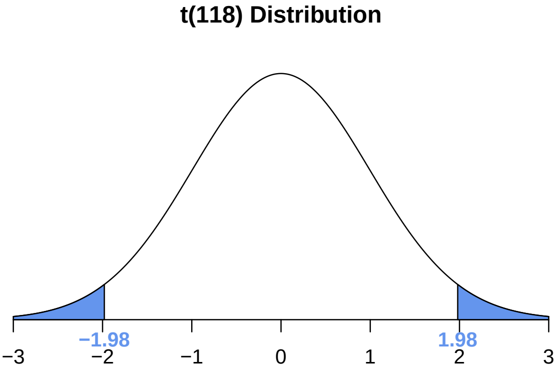
Creating a Confidence Interval for the Mean Flipper Length

We have some parts of the confidence interval calculation, but we still need the critical value t^*:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 217.2353 \pm t^* \frac{6.585431}{\sqrt{119}}$$

Finding the Critical Value t^* with R

We are using the t distribution with $df = n - 1 = 119 - 1 = 118$. For a 95% confidence interval, we want an area of 0.95 between $-t^*$ and t^*, leaving 1 – 0.95 = 0.05 to be split between the two tails (0.025 in each tail).



Finding the Critical Value t^* with R

We will use the qt() function get t^*. There is 0.025 area in each tail, so we can find t^* as, leaving $1 - 0.95 = 0.05$ to be split between the two tails.

```
qt(p = 0.025, df = 118, lower.tail = FALSE)
```

```
## [1] 1.980272
```

```
qt(p = 0.975, df = 118, lower.tail = TRUE)
```

```
## [1] 1.980272
```

Note that both of these get us the same t^* value–it's up to you which code you prefer. Enter the values for p, df, and lower.tail in the tryIt7 code chunk in your notes.

Creating a Confidence Interval for the Mean Flipper Length

Now we can finish calculating the 95% confidence interval for the population mean flipper length of Gentoo penguins living in the Palmer Archipelago.

$$217.2353 \pm 1.980272 \cdot \frac{6.585431}{\sqrt{119}} = 217.2353 \pm 1.195 = 216.040 \text{ to } 218.430$$

Interpreting a Confidence Interval for One Mean

We estimate with 95% confidence that the mean flipper length for all Gentoo penguins in the Palmer Archipelago is between 216.04 mm and 218.43 mm.

(This interval gives a range of reasonable values for the mean flipper length for all Gentoo penguins in the Palmer Archipelago.)

Creating a Confidence Interval for the Mean Flipper Length Using R

When we have the data, it's much easier to use `t.test()` to construct a confidence interval. For confidence intervals, we do not specify the `alternative` (we aren't testing hypotheses), but we need to specify the confidence level as a decimal in `conf.level`. As with hypothesis tests, we can only use `t.test()` if we have the data in R.

Creating a Confidence Interval for the Mean Flipper Length

Since we only want a confidence interval (we are not doing a hypothesis test), we do not need to specify a hypothesized mean (we don't have one!) nor do we have to specify the alternative. We *do* have to specify the confidence level. Try this code out in the `tryIt8` code chunk in your notes document (be sure to specify `conf.level`).

```r
t.test(gentoo$flipper_length_mm, conf.level = 0.95)
```

Creating a Confidence Interval for the Mean Flipper Length

```
##
##  One Sample t-test
##
## data:  gentoo$flipper_length_mm
## t = 359.85, df = 118, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  216.0398 218.4308
## sample estimates:
## mean of x
##  217.2353
```

As expected, R produce the same confidence interval that we found by hand.