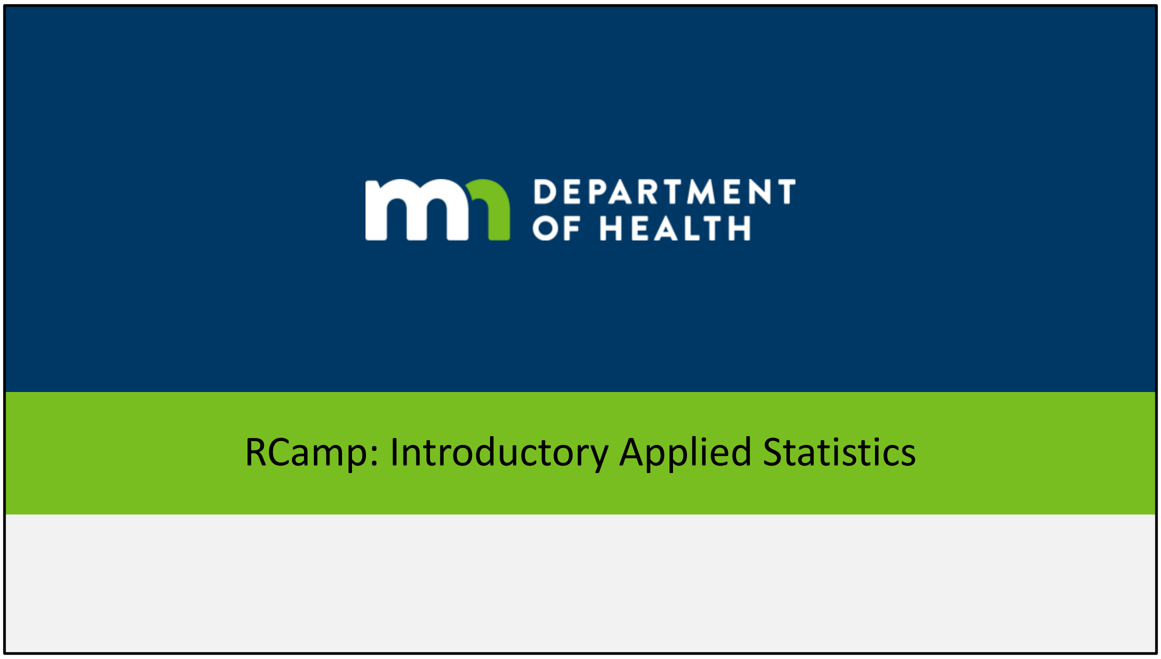


DEPARTMENT OF HEALTH

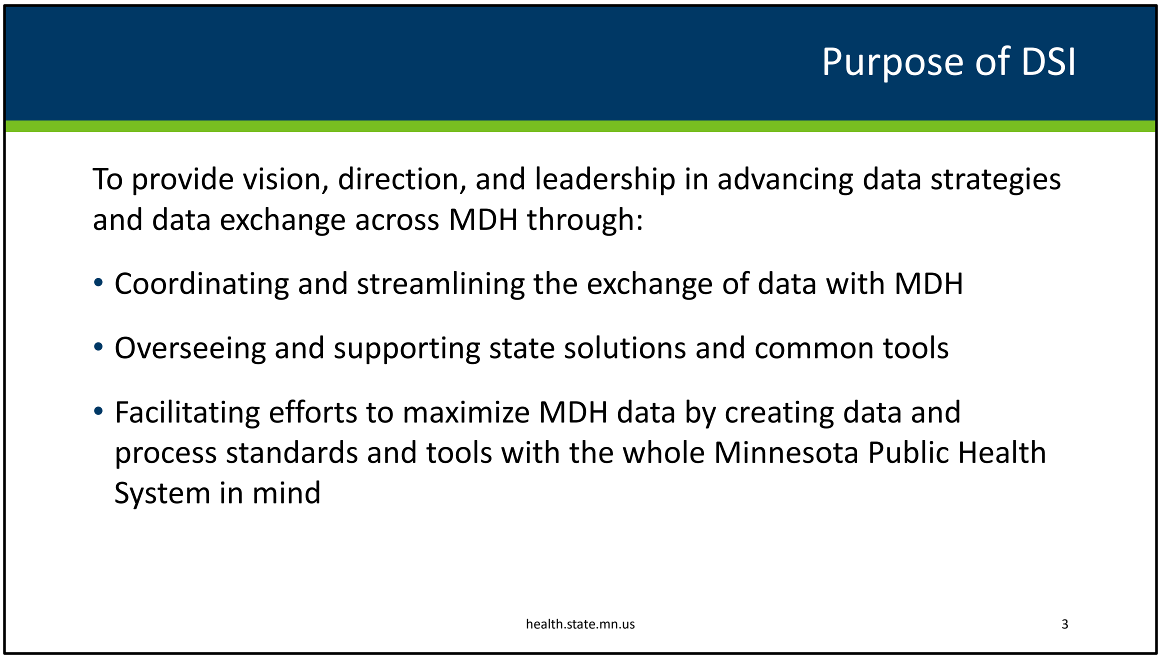
# RCamp: Introductory Applied Statistics

# Tribal-State Relations Acknowledgment Statement

The State of Minnesota is home to 11 federally recognized Indian Tribes with elected Tribal government officials. The State of Minnesota acknowledges and supports the unique political status of Tribal Nations across Minnesota and their absolute right to existence, self-governance, and self-determination. This unique relationship with federally recognized Indian Tribes is cemented by the Constitution of the United States, treaties, statutes, case law, and agreements. The State of Minnesota and Tribal governments across Minnesota significantly benefit from working together, learning from one another, and partnering where possible.

Minnesota Department of Health recognizes, values, and celebrates the vibrant and unique relationships between the 11 Tribal Nations and the State of Minnesota. Partnerships formed through government-to-government relationships with these Tribes will effectively address health disparities and lead to better health outcomes for all of Minnesota.

In our work at the Office of Data Strategy and Interoperability, we demonstrate our commitment to Tribal-State relations by providing free assistance upon request and promoting health equity in data collection and use.

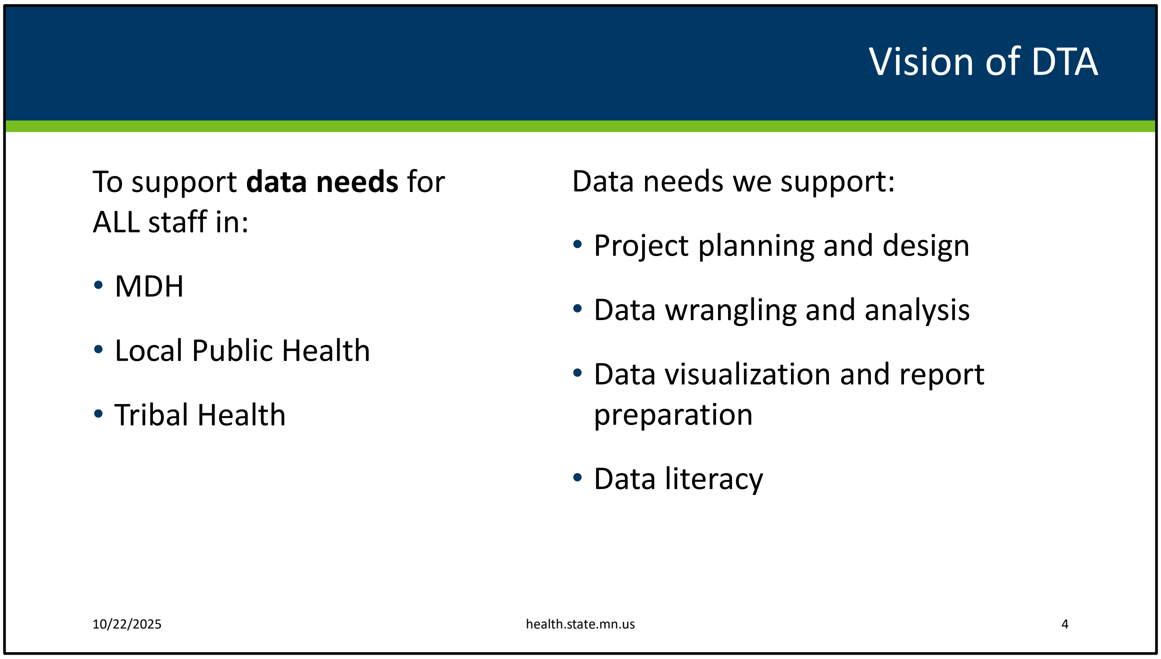


## Purpose of DSI

To provide vision, direction, and leadership in advancing data strategies and data exchange across MDH through:

- Coordinating and streamlining the exchange of data with MDH
- Overseeing and supporting state solutions and common tools
- Facilitating efforts to maximize MDH data by creating data and process standards and tools with the whole Minnesota Public Health System in mind

The purpose of DSI is to provide vision, direction, and leadership across MDH. We accomplish this by streamlining data exchange between MDH and outside partners, including LPH, advocating for common tools and standards within and with MDH, and working toward solutions that help LPH access and use MDH data for their work.



# Vision of DTA

To support **data needs** for ALL staff in:

- MDH
- Local Public Health
- Tribal Health

Data needs we support:

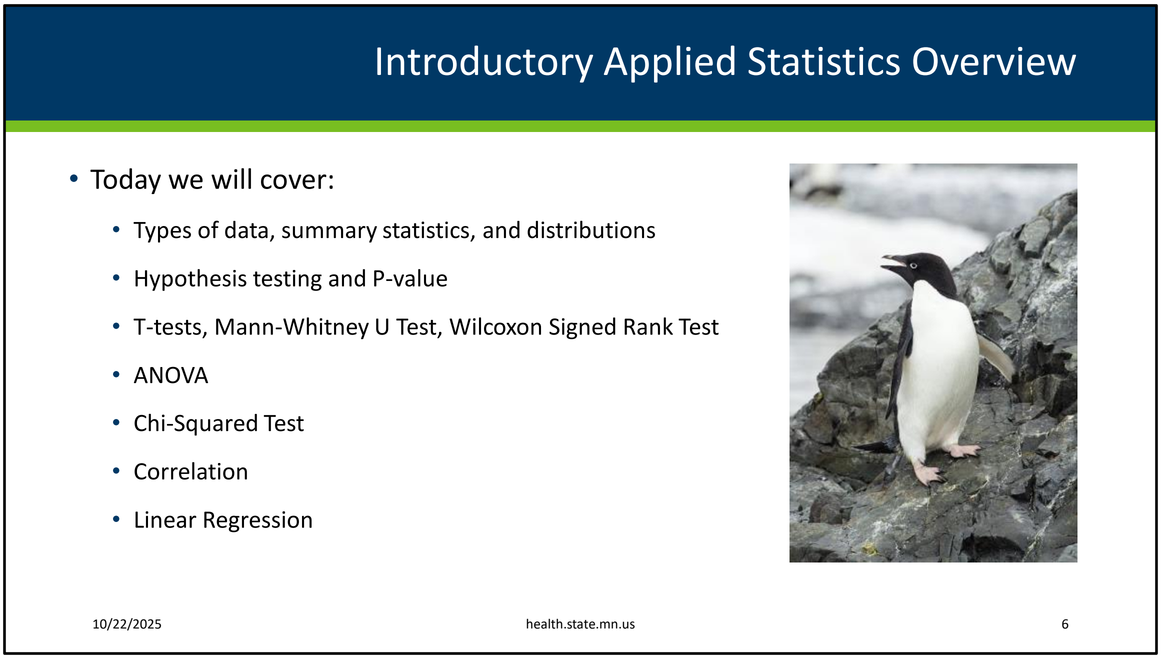
- Project planning and design
- Data wrangling and analysis
- Data visualization and report preparation
- Data literacy

The vision of our unit is to support a range of data needs for ALL government staff working in public health in Minnesota. If you are familiar with the data lifecycle, we referenced that when determining what we could support. We support project design and conceptualization, data cleaning and validation, data analysis, and data visualization. We also provide general support through venues like trainings and communities of practice.

## Acknowledgments

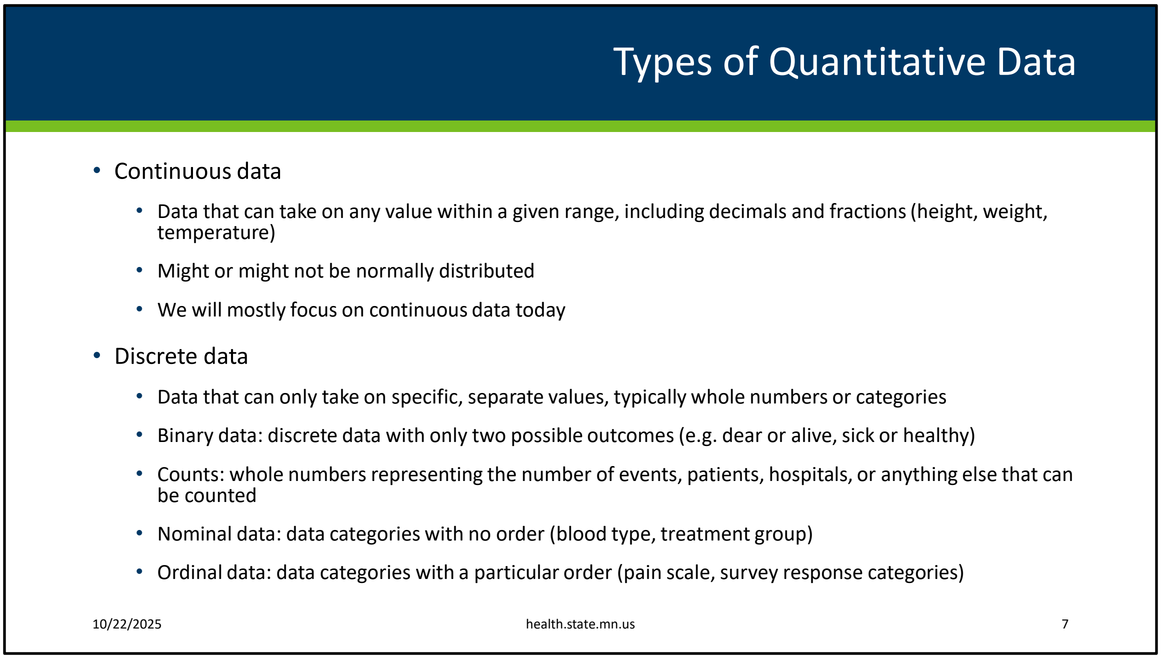
*Thank you to the State of Minnesota R user group!*

★Special thanks to our founders Barbara Monaco, Dorian Kvale, Derek Nagel, and Kristi Ellickson.



# Introductory Applied Statistics Overview

- Today we will cover:
  - Types of data, summary statistics, and distributions
  - Hypothesis testing and P-value
  - T-tests, Mann-Whitney U Test, Wilcoxon Signed Rank Test
  - ANOVA
  - Chi-Squared Test
  - Correlation
  - Linear Regression



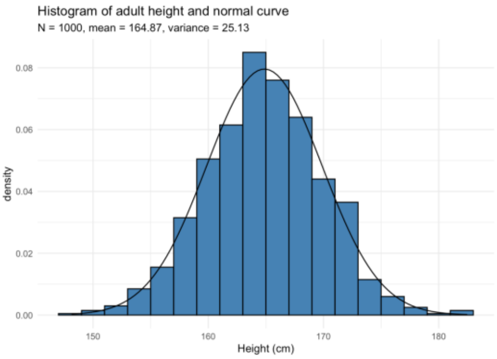
## Types of Quantitative Data

- Continuous data
  - Data that can take on any value within a given range, including decimals and fractions (height, weight, temperature)
  - Might or might not be normally distributed
  - We will mostly focus on continuous data today
- Discrete data
  - Data that can only take on specific, separate values, typically whole numbers or categories
  - Binary data: discrete data with only two possible outcomes (e.g. dear or alive, sick or healthy)
  - Counts: whole numbers representing the number of events, patients, hospitals, or anything else that can be counted
  - Nominal data: data categories with no order (blood type, treatment group)
  - Ordinal data: data categories with a particular order (pain scale, survey response categories)

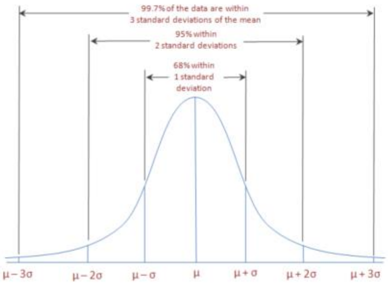
# Summary Statistics

- Summary statistics summarize and provide information about your sample data.
- Summary statistics fall into three main categories:
  - Measures of **location**: where your data is centered at, or where a trend lies
    - Mean, median, mode
  - Measures of **spread** or distribution shape:  how spread out or varied your data set is
    - range, standard deviation, skew and kurtosis
  - Graphs/charts: ways to display summary data
    - histogram, frequency distribution table, box plot, bar chart, scatter plot and pie chart

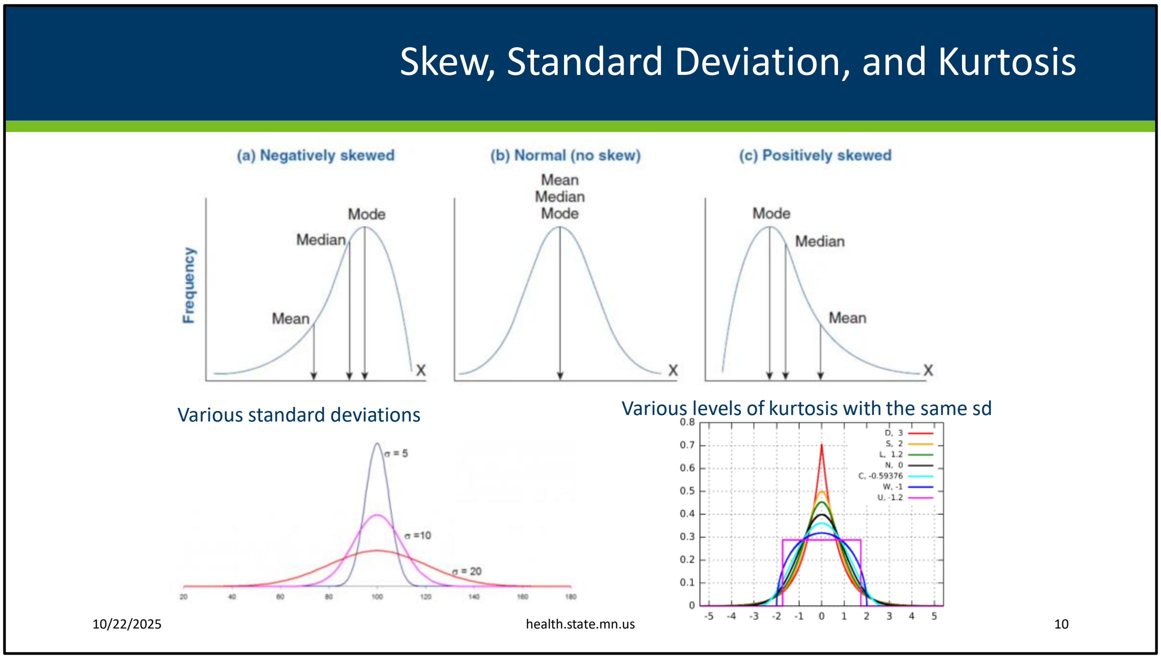
# Normal Distribution



- $\mu \pm \sigma$ includes approximately 68% of the observations
- $\mu \pm 2 \cdot \sigma$ includes approximately 95% of the observations
- $\mu \pm 3 \cdot \sigma$ includes almost all of the observations (99.7% to be more precise)



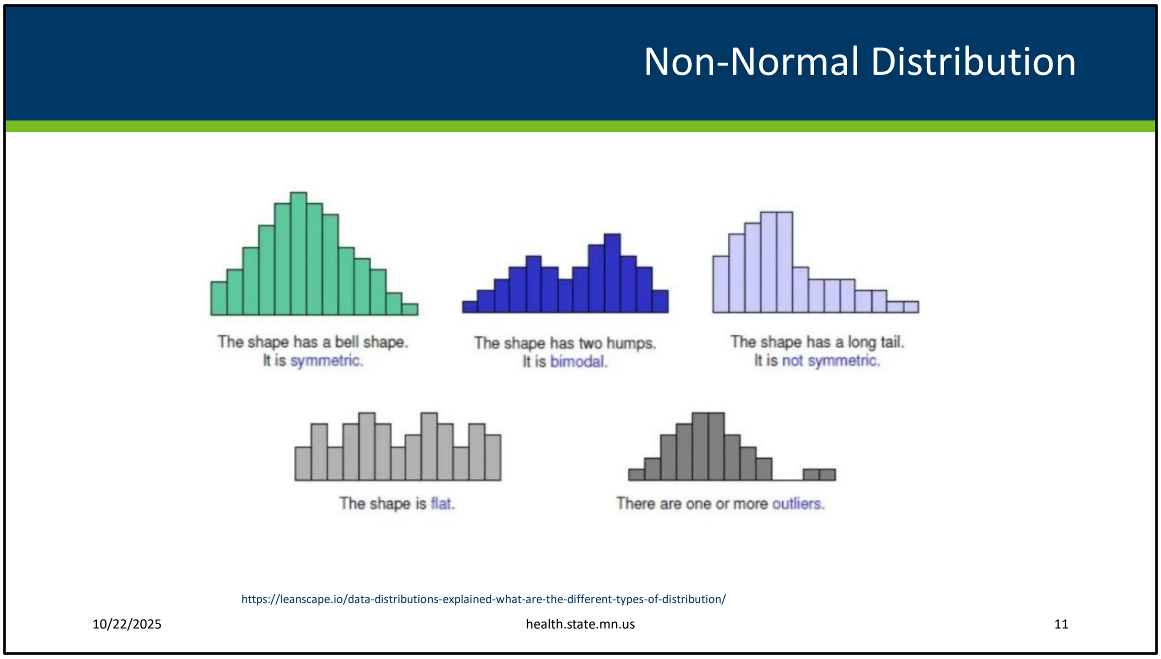
https://statsandr.com/blog/do-my-data-follow-a-normal-distribution-a-note-on-the-most-widely-used-distribution-and-how-to-test-for-normality-in-r/



# Skew, Standard Deviation, and Kurtosis

Various standard deviations

Various levels of kurtosis with the same sd



## Non-Normal Distribution

The shape has a bell shape.
It is symmetric.

The shape has two humps.
It is bimodal.

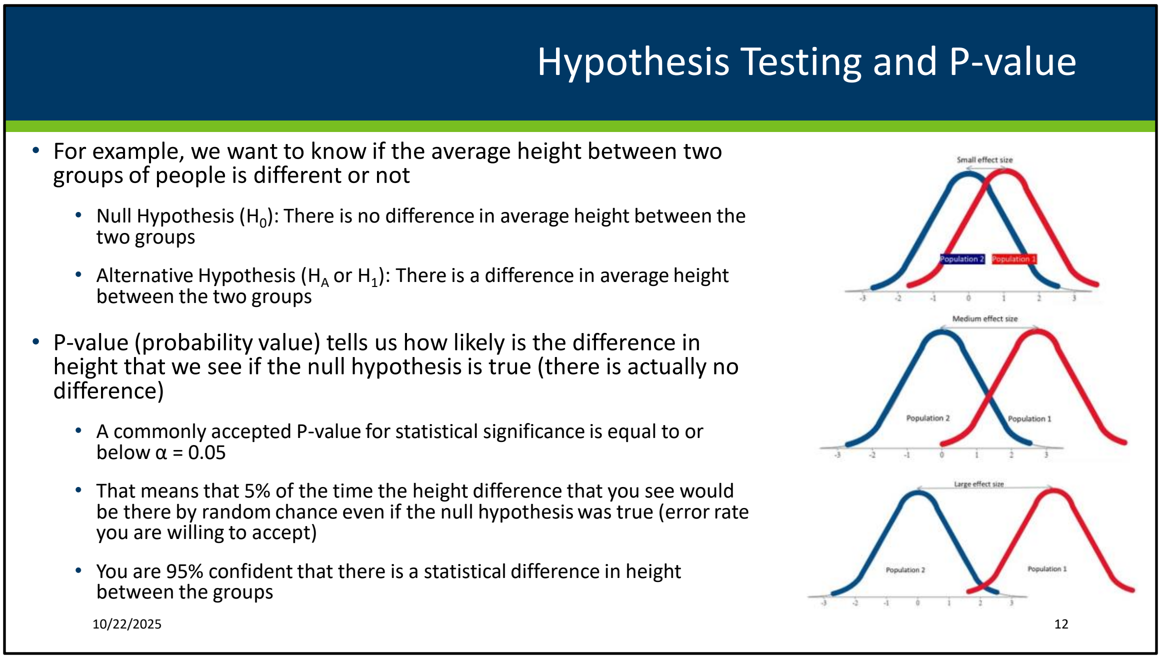
The shape has a long tail.
It is not symmetric.

The shape is flat.

There are one or more outliers.

https://leanscape.io/data-distributions-explained-what-are-the-different-types-of-distribution/

For many statistical tests, normal distribution is required in order to get accurate results!



# Hypothesis Testing and P-value

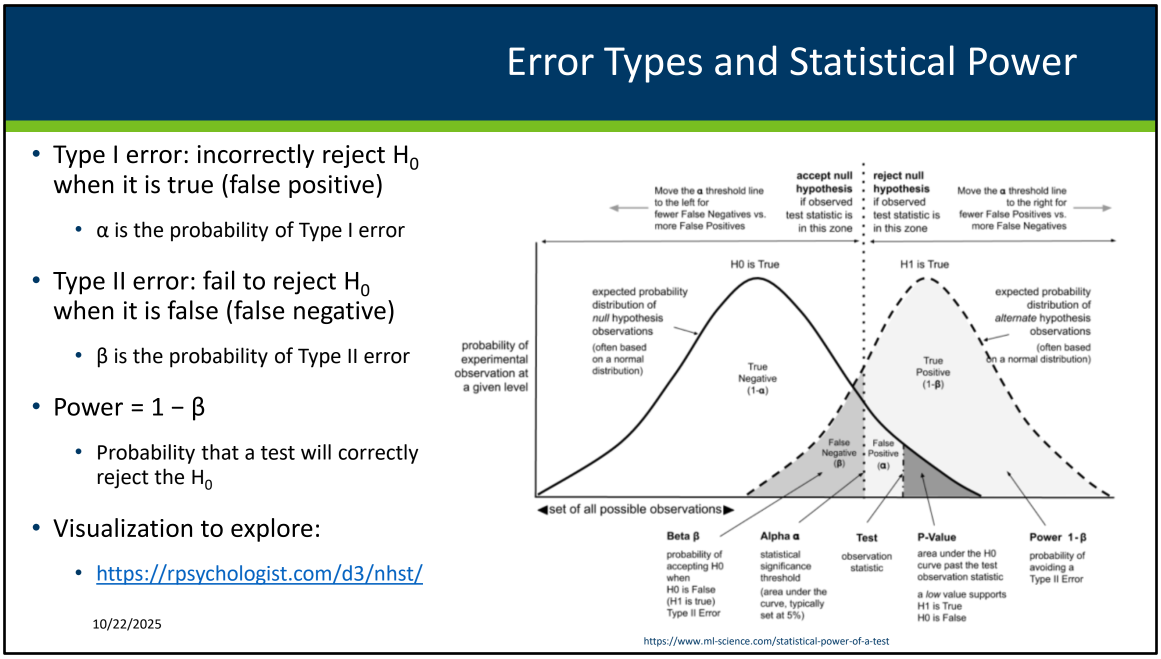
- For example, we want to know if the average height between two groups of people is different or not
  - Null Hypothesis ($H_0$): There is no difference in average height between the two groups
  - Alternative Hypothesis ($H_A$ or $H_1$): There is a difference in average height between the two groups
- P-value (probability value) tells us how likely is the difference in height that we see if the null hypothesis is true (there is actually no difference)
  - A commonly accepted P-value for statistical significance is equal to or below $\alpha = 0.05$
  - That means that 5% of the time the height difference that you see would be there by random chance even if the null hypothesis was true (error rate you are willing to accept)
  - You are 95% confident that there is a statistical difference in height between the groups

10/22/2025

12

P-value: probability of observing data as extreme (or more) than observed if H0 were true.

Significance level (α): a pre-set threshold (commonly 0.05) for declaring statistical significance.



## Error Types and Statistical Power

- Type I error: incorrectly reject $H_0$ when it is true (false positive)
  - α is the probability of Type I error
- Type II error: fail to reject $H_0$ when it is false (false negative)
  - β is the probability of Type II error
- Power = 1 − β
  - Probability that a test will correctly reject the $H_0$
- Visualization to explore:
  - [https://rpsychologist.com/d3/nhst/](https://rpsychologist.com/d3/nhst/)

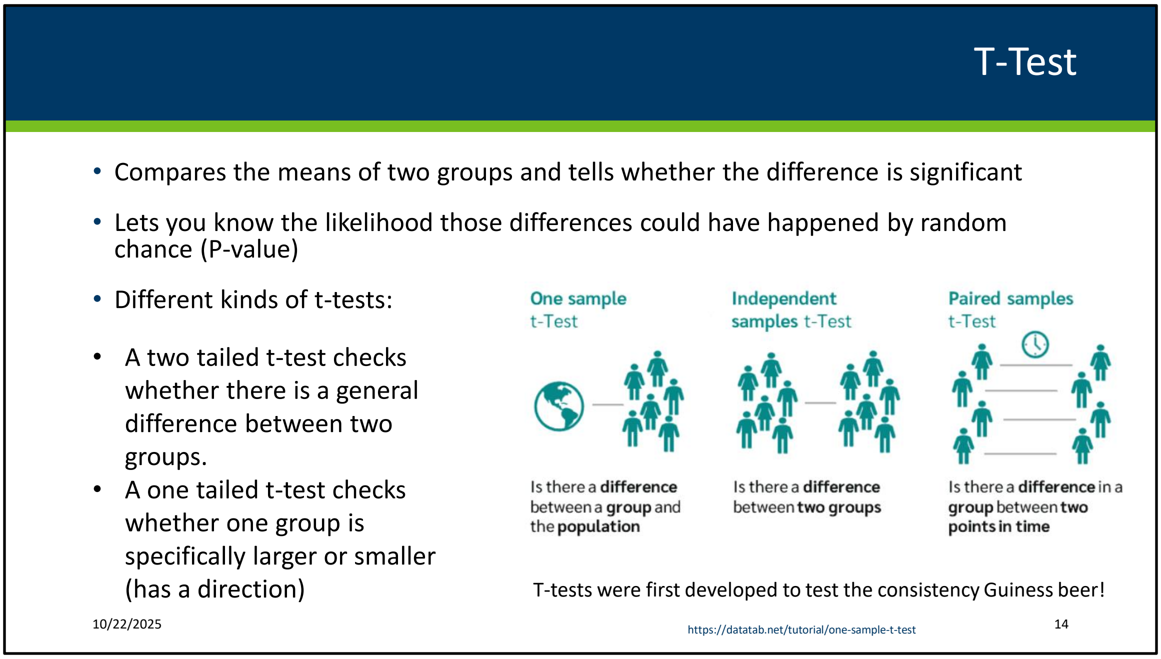
10/22/2025

Move the α threshold line to the left for fewer False Negatives vs. more False Positives
accept null hypothesis if observed test statistic is in this zone
reject null hypothesis if observed test statistic is in this zone
Move the α threshold line to the right for fewer False Positives vs. more False Negatives
H0 is True
H1 is True
expected probability distribution of *null* hypothesis observations (often based on a normal distribution)
expected probability distribution of *alternate* hypothesis observations (often based on a normal distribution)
probability of experimental observation at a given level
True Negative (1-α)
True Positive (1-β)
False Negative (β)
False Positive (α)
◀set of all possible observations▶
Beta β probability of accepting H0 when H0 is False (H1 is true) Type II Error
Alpha α statistical significance threshold (area under the curve, typically set at 5%)
Test observation statistic
P-Value area under the H0 curve past the test observation statistic a *low* value supports H1 is True H0 is False
Power 1-β probability of avoiding a Type II Error

https://www.ml-science.com/statistical-power-of-a-test

P-value: probability of observing data as extreme (or more) than observed if H0 were true.

Significance level (α): a pre-set threshold (commonly 0.05) for declaring statistical significance.



# T-Test

- Compares the means of two groups and tells whether the difference is significant
- Lets you know the likelihood those differences could have happened by random chance (P-value)
- Different kinds of t-tests:

- A two tailed t-test checks whether there is a general difference between two groups.
- A one tailed t-test checks whether one group is specifically larger or smaller (has a direction)

**One sample** t-Test

Is there a **difference** between a **group** and the **population**

**Independent samples** t-Test

Is there a **difference** between **two groups**

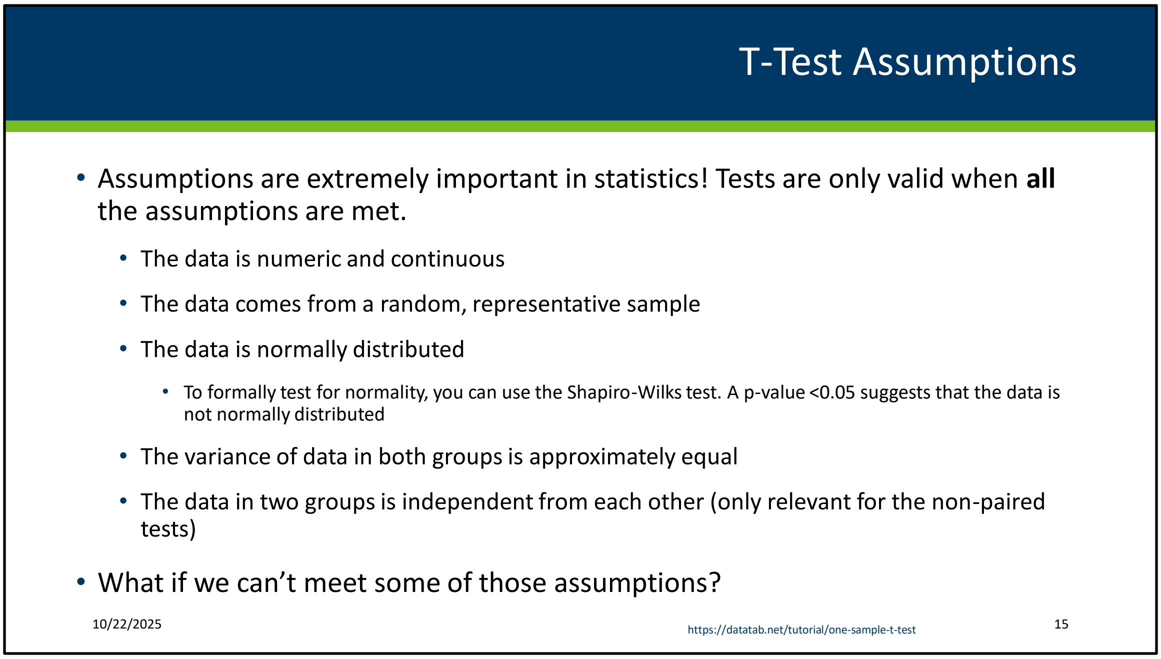
**Paired samples** t-Test

Is there a **difference** in a **group** between **two points in time**

T-tests were first developed to test the consistency Guiness beer!

https://datatab.net/tutorial/one-sample-t-test

A paired t-test assumes that there is less variability between the two measured groups, since it assumes that we are measuring the same subjects twice. By removing this variability that would usually exist if the groups were independent, we make the test more sensitive, and it often results in a lower p-value.



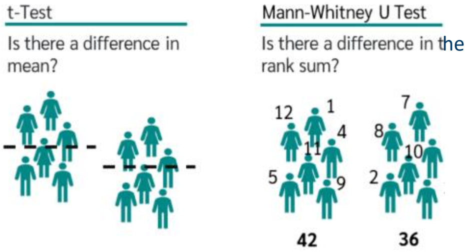
# T-Test Assumptions

- Assumptions are extremely important in statistics! Tests are only valid when **all** the assumptions are met.
  - The data is numeric and continuous
  - The data comes from a random, representative sample
  - The data is normally distributed
    - To formally test for normality, you can use the Shapiro-Wilks test. A p-value <0.05 suggests that the data is not normally distributed
  - The variance of data in both groups is approximately equal
  - The data in two groups is independent from each other (only relevant for the non-paired tests)
- What if we can’t meet some of those assumptions?

https://datatab.net/tutorial/one-sample-t-test

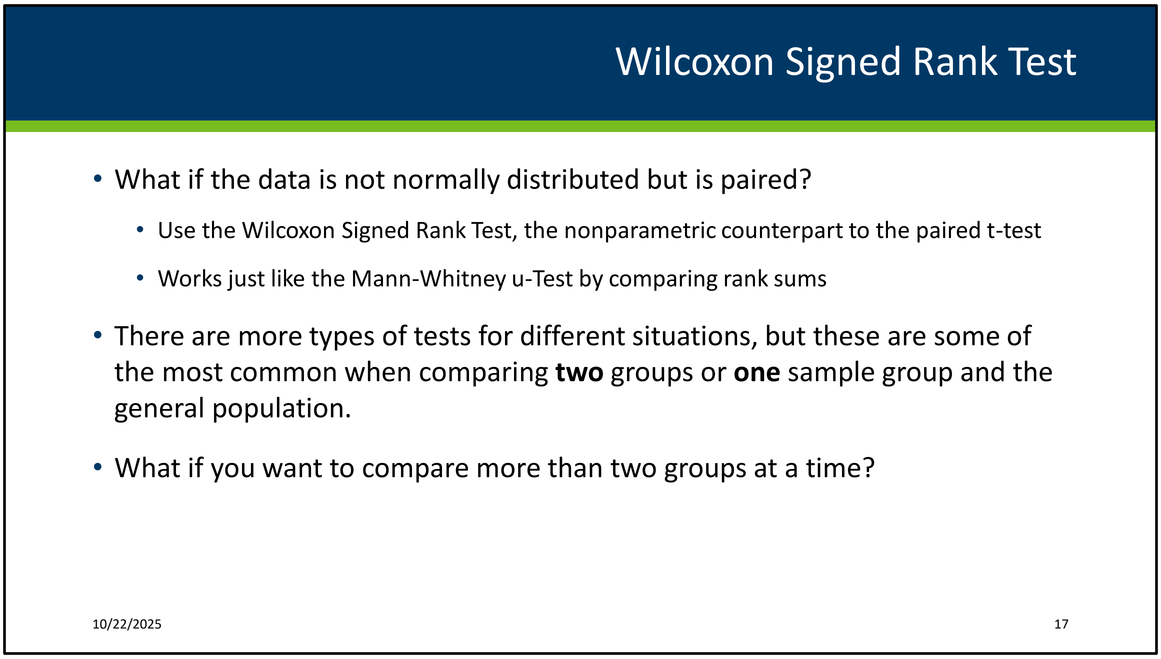
# Mann-Whitney U-Test

- If the data isn't normally distributed, you can use the Mann-Whitney U-Test
- The independent t-test counterpart for data that is not normally distributed.
- Tests that require assumptions about data distributions are called **parametric**. Tests that do not are called **nonparametric**.
- Instead of comparing groups directly, the test ranks everyone in both groups from shortest to tallest, sums the ranks in both groups, and compares this rank sum between the groups.



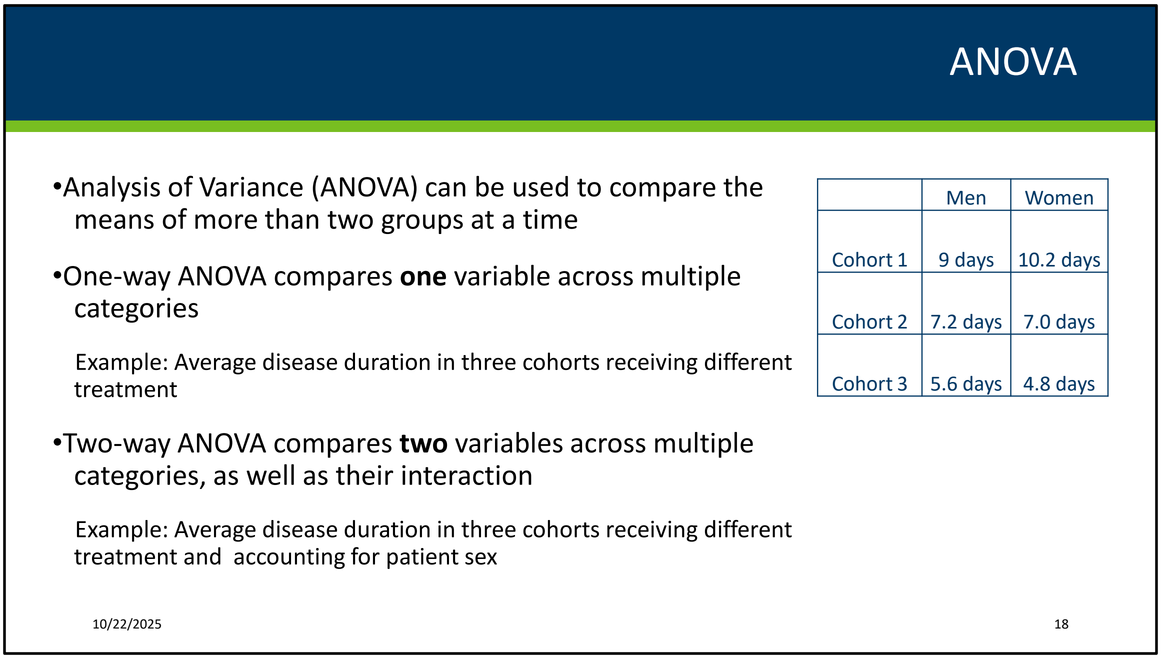
t-Test
Is there a difference in mean?
Mann-Whitney U Test
Is there a difference in the rank sum?
12
1
4
11
5
9
7
8
10
2
42
36

https://datatab.net/tutorial/mann-whitney-u-test



## Wilcoxon Signed Rank Test

- What if the data is not normally distributed but is paired?
  - Use the Wilcoxon Signed Rank Test, the nonparametric counterpart to the paired t-test
  - Works just like the Mann-Whitney u-Test by comparing rank sums
- There are more types of tests for different situations, but these are some of the most common when comparing **two** groups or **one** sample group and the general population.
- What if you want to compare more than two groups at a time?



# ANOVA

- Analysis of Variance (ANOVA) can be used to compare the means of more than two groups at a time

- One-way ANOVA compares **one** variable across multiple categories

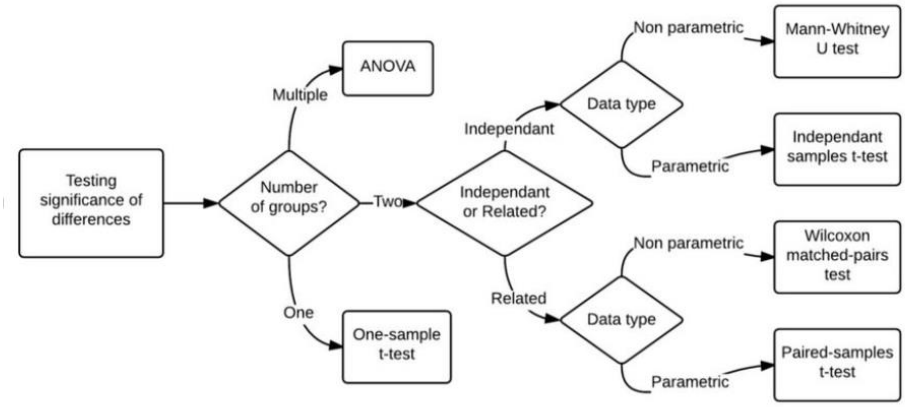
  Example: Average disease duration in three cohorts receiving different treatment

- Two-way ANOVA compares **two** variables across multiple categories, as well as their interaction

  Example: Average disease duration in three cohorts receiving different treatment and  accounting for patient sex

| | Men | Women |
|---|---|---|
| Cohort 1 | 9 days | 10.2 days |
| Cohort 2 | 7.2 days | 7.0 days |
| Cohort 3 | 5.6 days | 4.8 days |

# Statistical Test Decision Tree

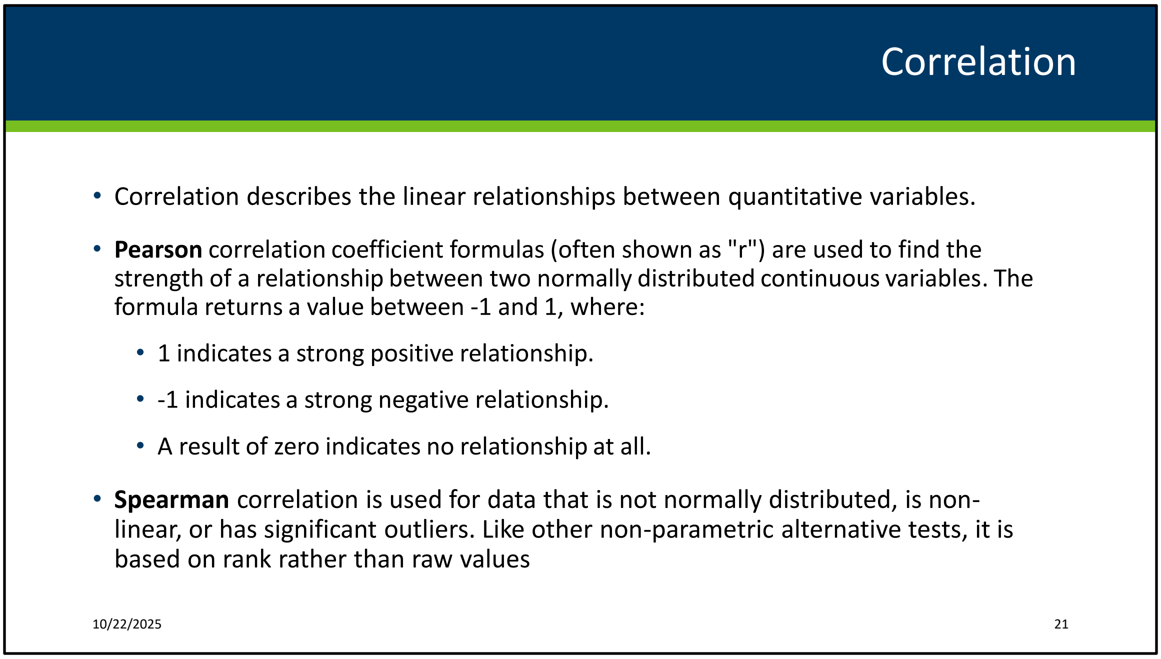


Chaudy, Yaelle. (2015). An Assessment and Learning Analytics Engine for Games-based Learning. 10.13140/RG.2.1.4932.5040.

# Chi-Square Test

- If your data isn’t continuous but is instead categorical, you can use the Chi-Square test
  - Example: we want to tell whether out of a sample of 300 people, diabetes rates are different between men and women
  - $H_0$: Diabetes rate are the same
  - $H_A$: Diabetes rate are different
- Assumptions:
  - Both datasets are categorial
  - Observations are independent
  - Each category (cell in the table) is mutually exclusive
  - Each value in the cell should be 5 or greater

| | Diabetes | No Diabetes | Total |
|---|---|---|---|
| Men | 43 | 114 | 157 |
| Women | 34 | 109 | 143 |
| Total | 77 | 223 | 300 |

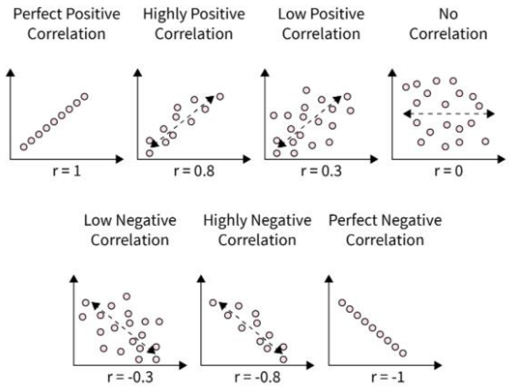


# Correlation

- Correlation describes the linear relationships between quantitative variables.
- **Pearson** correlation coefficient formulas (often shown as "r") are used to find the strength of a relationship between two normally distributed continuous variables. The formula returns a value between -1 and 1, where:
  - 1 indicates a strong positive relationship.
  - -1 indicates a strong negative relationship.
  - A result of zero indicates no relationship at all.
- **Spearman** correlation is used for data that is not normally distributed, is non-linear, or has significant outliers. Like other non-parametric alternative tests, it is based on rank rather than raw values

# Correlation Visual Examples

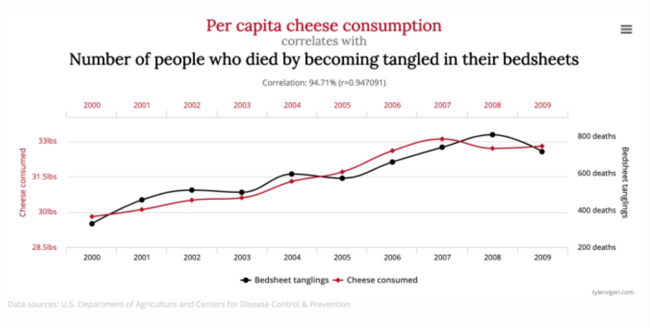
Scatter Plots & Correlation Examples



https://www.scaler.com/topics/data-science/covariance-and-correlation/

# Correlation is not Causation

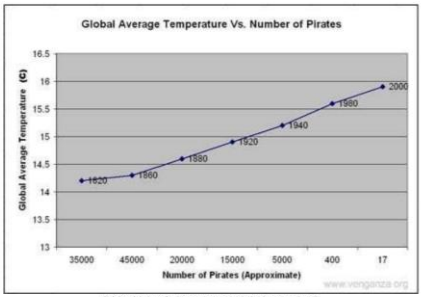
- Correlation is not causation! just because two things correlate does not necessarily mean that one causes the other.



Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% (r=0.947091)

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

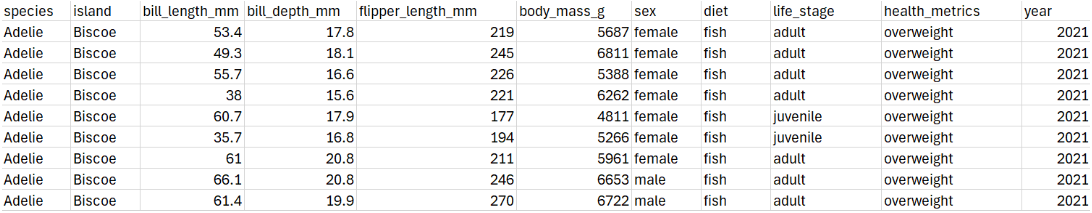


Global Average Temperature Vs. Number of Pirates

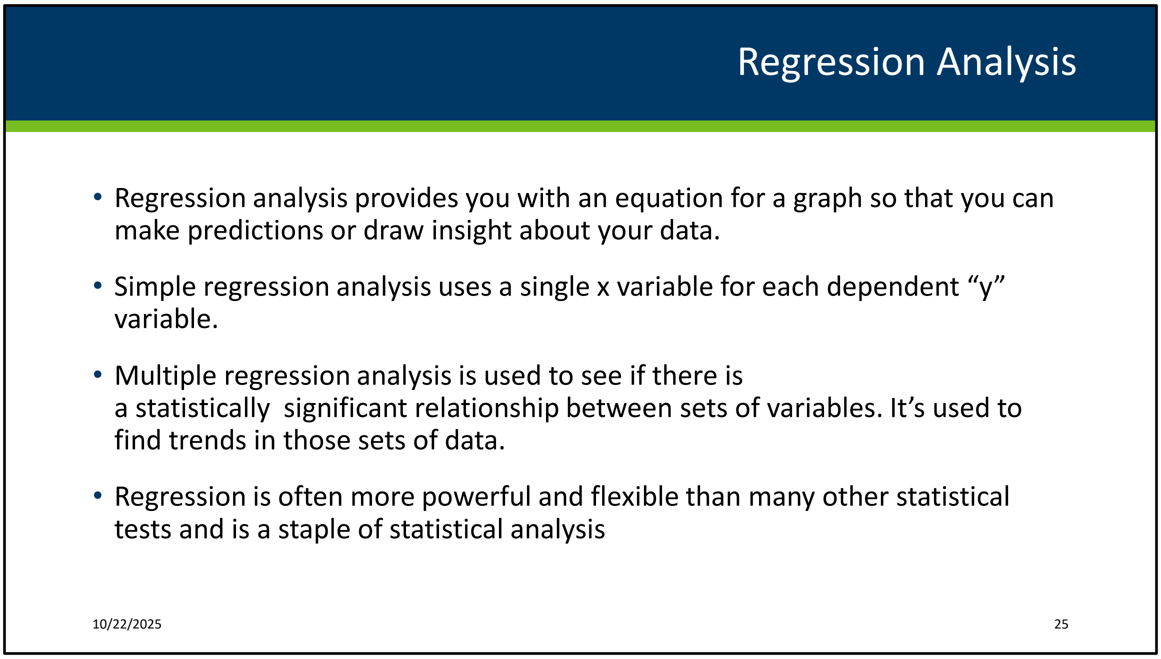
Did a lack of piracy cause global warming?

# Palmer Penguin Dataset

- Now let’s look how these kinds of tests can be run in R
- This data was collected between 2007 and 2009 on the Palmer archipelago in the Antarctic
- https://journal.r-project.org/articles/RJ-2022-020/



| species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex | diet | life_stage | health_metrics | year |
|---|---|---|---|---|---|---|---|---|---|---|
| Adelie | Biscoe | 53.4 | 17.8 | 219 | 5687 | female | fish | adult | overweight | 2021 |
| Adelie | Biscoe | 49.3 | 18.1 | 245 | 6811 | female | fish | adult | overweight | 2021 |
| Adelie | Biscoe | 55.7 | 16.6 | 226 | 5388 | female | fish | adult | overweight | 2021 |
| Adelie | Biscoe | 38 | 15.6 | 221 | 6262 | female | fish | adult | overweight | 2021 |
| Adelie | Biscoe | 60.7 | 17.9 | 177 | 4811 | female | fish | juvenile | overweight | 2021 |
| Adelie | Biscoe | 35.7 | 16.8 | 194 | 5266 | female | fish | juvenile | overweight | 2021 |
| Adelie | Biscoe | 61 | 20.8 | 211 | 5961 | female | fish | adult | overweight | 2021 |
| Adelie | Biscoe | 66.1 | 20.8 | 246 | 6653 | male | fish | adult | overweight | 2021 |
| Adelie | Biscoe | 61.4 | 19.9 | 270 | 6722 | male | fish | adult | overweight | 2021 |



## Regression Analysis

- Regression analysis provides you with an equation for a graph so that you can make predictions or draw insight about your data.
- Simple regression analysis uses a single x variable for each dependent “y” variable.
- Multiple regression analysis is used to see if there is a statistically  significant relationship between sets of variables. It’s used to find trends in those sets of data.
- Regression is often more powerful and flexible than many other statistical tests and is a staple of statistical analysis

# Linear Regression

- Linear regression equation:

$$y = b_0 + b_1 * x + \epsilon, \epsilon \sim \text{Norm}(0, \sigma^2)$$

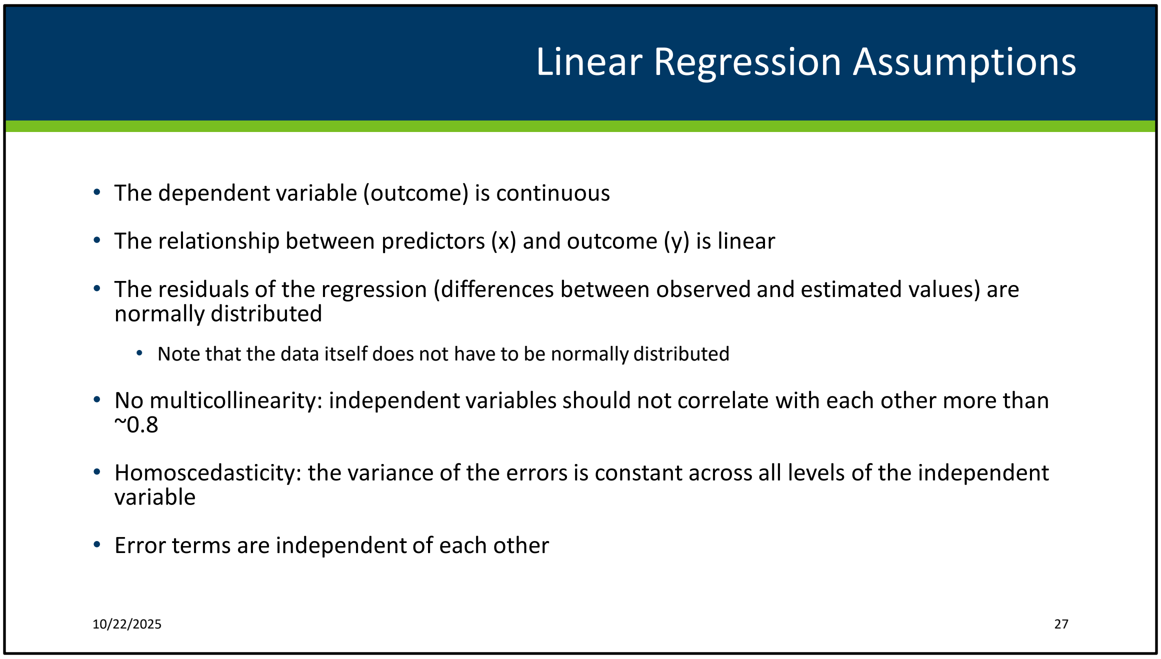
  - y – dependent variable
  - x – independent variable
  - b1 – slope
  - b0 – intercept
  - ε **–** the error term that represents all the variation in y that the model cannot explain

- Linear regression error term is normally distributed

$$\epsilon \sim \text{Norm}(0, \sigma^2)$$

  - 0 – the mean of the normal distribution
  - $\sigma_2$ – standard deviation squared, also known as variance

- Linear regression visualization: https://ryansafner.shinyapps.io/ols_estimation_by_min_sse/

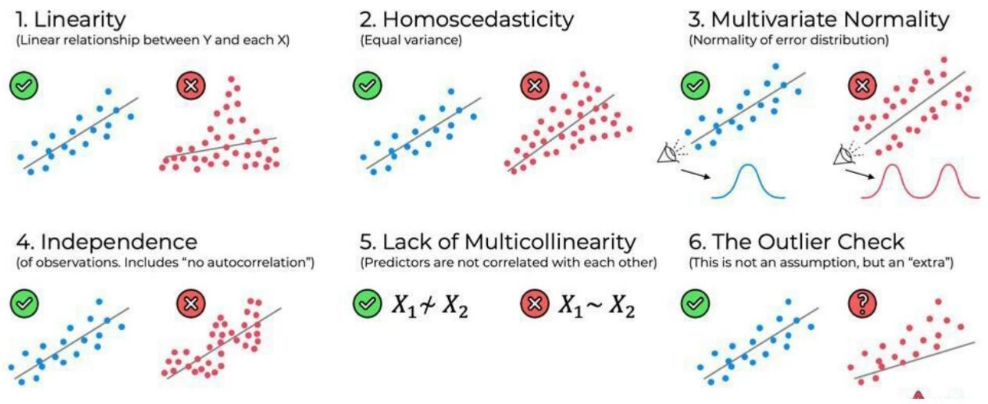


## Linear Regression Assumptions

- The dependent variable (outcome) is continuous
- The relationship between predictors (x) and outcome (y) is linear
- The residuals of the regression (differences between observed and estimated values) are normally distributed
  - Note that the data itself does not have to be normally distributed
- No multicollinearity: independent variables should not correlate with each other more than ~0.8
- Homoscedasticity: the variance of the errors is constant across all levels of the independent variable
- Error terms are independent of each other

*Error dependence might be especially prevalent in time series data where earlier values influence future values

# Linear Regression Assumptions



1. Linearity
(Linear relationship between Y and each X)

2. Homoscedasticity
(Equal variance)

3. Multivariate Normality
(Normality of error distribution)

4. Independence
(of observations. Includes "no autocorrelation")

5. Lack of Multicollinearity
(Predictors are not correlated with each other)

✓ $X_1 \nsim X_2$ ☒ $X_1 \sim X_2$

6. The Outlier Check
(This is not an assumption, but an "extra")

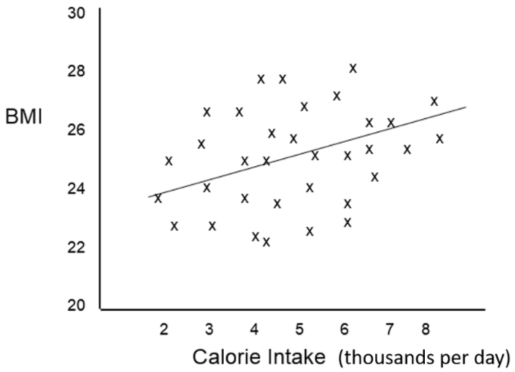
https://www.geeksforgeeks.org/machine-learning/assumptions-of-linear-regression/

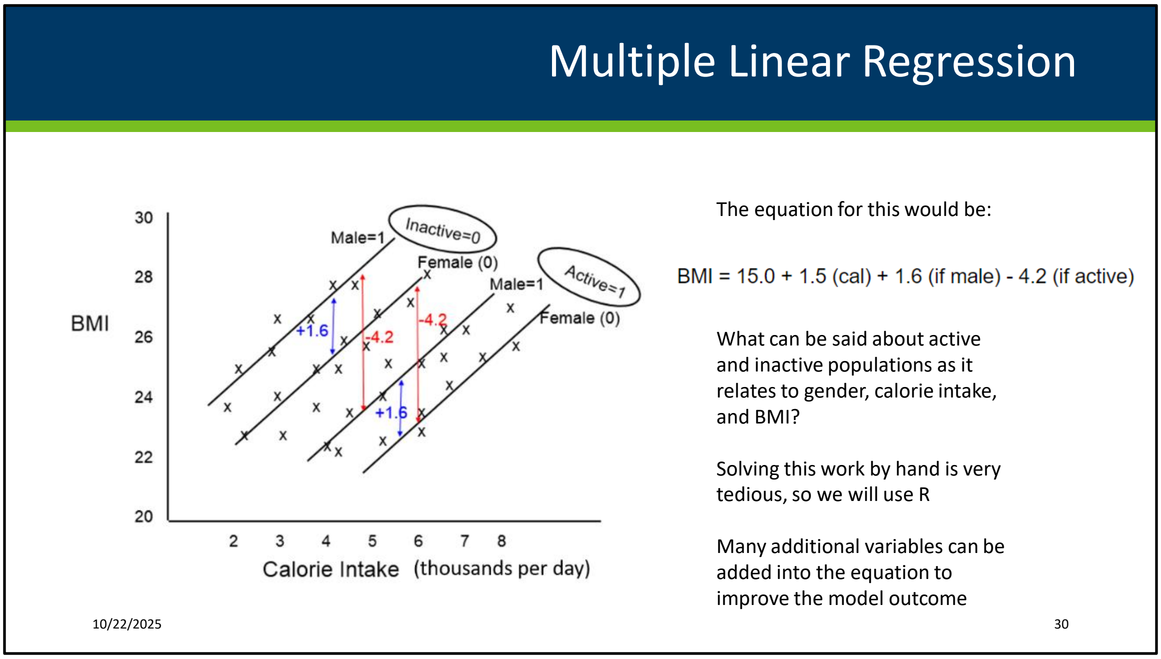
# Multiple Linear Regression

Consider a hypothetical situation evaluating BMI and Calorie Intake.

Multiple R squared value (a statistic to help assess the wellness of the regression fit) in this case was 0.12 , so caloric intake alone does not predict BMI well.

What other variables could influence BMI? Perhaps activity level and gender?





# Multiple Linear Regression

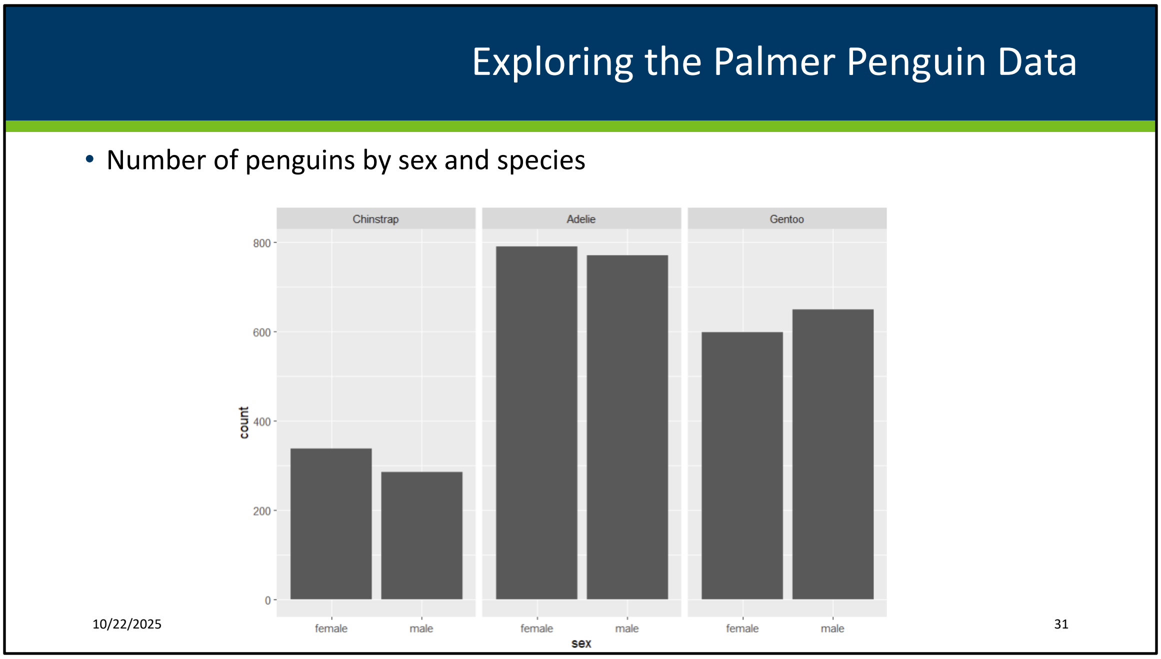
The equation for this would be:

BMI = 15.0 + 1.5 (cal) + 1.6 (if male) - 4.2 (if active)

What can be said about active and inactive populations as it relates to gender, calorie intake, and BMI?

Solving this work by hand is very tedious, so we will use R

Many additional variables can be added into the equation to improve the model outcome



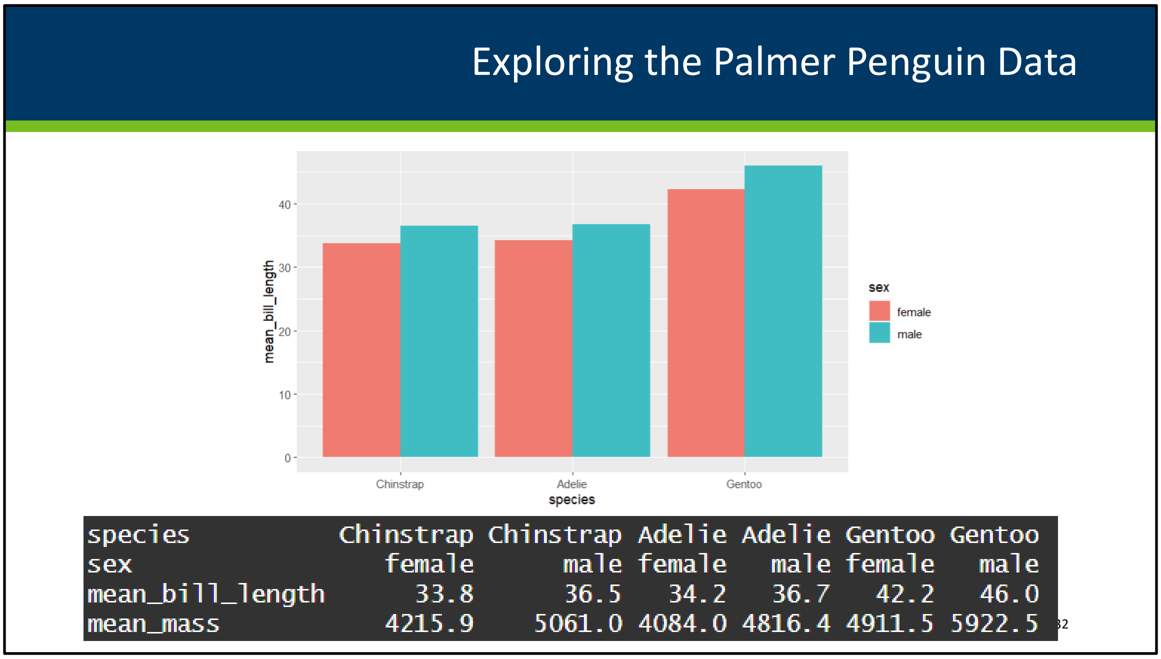
## Exploring the Palmer Penguin Data

- Number of penguins by sex and species

10/22/2025

31

It's usually the best practice to start exploring your data via figures, tables, and graphs and only move on to statistics once you understand it well enough

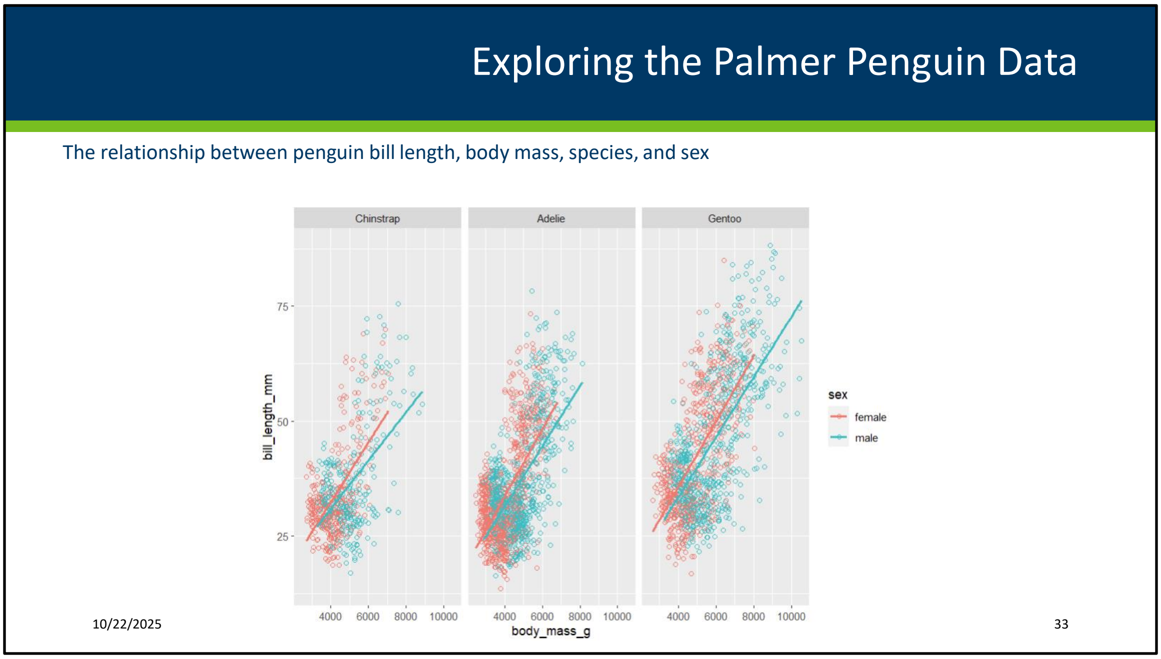


## Exploring the Palmer Penguin Data

```
species          Chinstrap Chinstrap Adelie Adelie Gentoo Gentoo
sex                 female      male female   male female   male
mean_bill_length      33.8      36.5   34.2   36.7   42.2   46.0
mean_mass           4215.9    5061.0 4084.0 4816.4 4911.5 5922.5
```

32

A table can be more useful than a graph if you want to see specific numbers



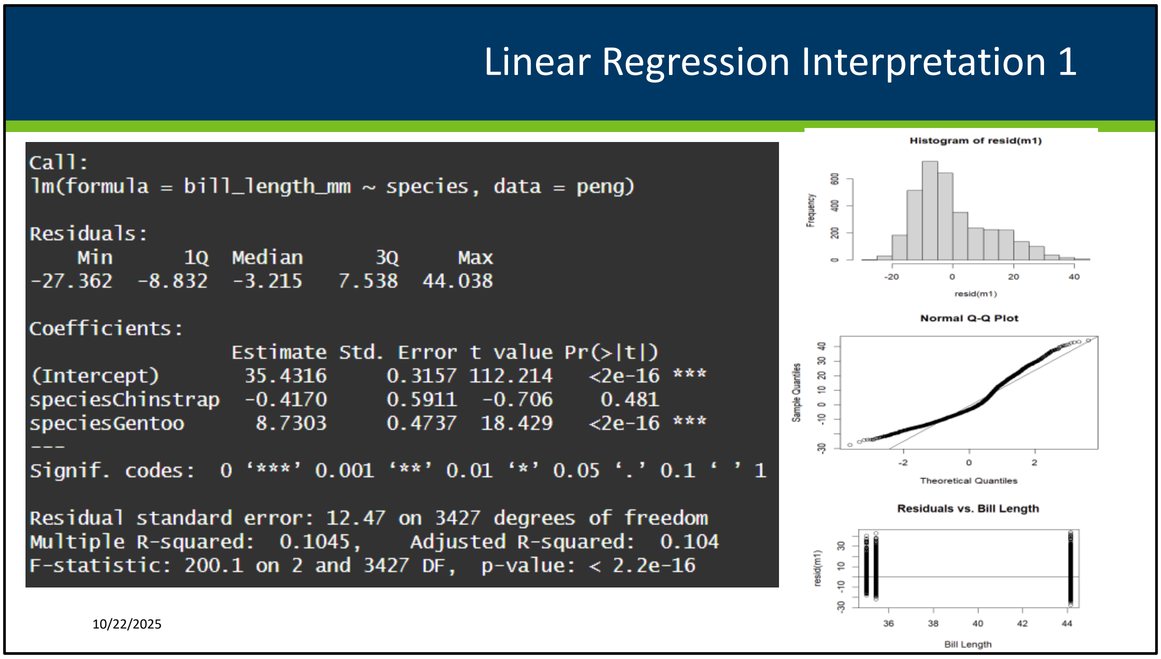
# Exploring the Palmer Penguin Data

The relationship between penguin bill length, body mass, species, and sex

10/22/2025

33

Graphs like these can help visualize complex relationships and build hypotheses. Interpretations? It seems like the relationship between body mass and bill length varies by sex and species



## Linear Regression Interpretation 1

```
Call:
lm(formula = bill_length_mm ~ species, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.362  -8.832  -3.215   7.538  44.038 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       35.4316     0.3157 112.214   <2e-16 ***
speciesChinstrap  -0.4170     0.5911  -0.706    0.481    
speciesGentoo      8.7303     0.4737  18.429   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.47 on 3427 degrees of freedom
Multiple R-squared:  0.1045,	Adjusted R-squared:  0.104 
F-statistic: 200.1 on 2 and 3427 DF,  p-value: < 2.2e-16
```

10/22/2025

Interpretation: mean bill length of Adeline is 35.4mm. Gentoo bill length is 8.73mm longer than Adeline, which is statistically significant at alpha = 0.05. The mean Chinstrap bill length is shorter than Adeline, but this difference isn’t statistically significant.

R-squared: the proportion of the variance of the dependent variable that can be explained by the independent variable. It is 0.104, so only about 10% of the bill length is explained by the species of penguin. Multiple R-squared always increases with additional variables, so Adjusted R-squared is generally a better statistic to consider. Regression models should aim to be parsimonious.

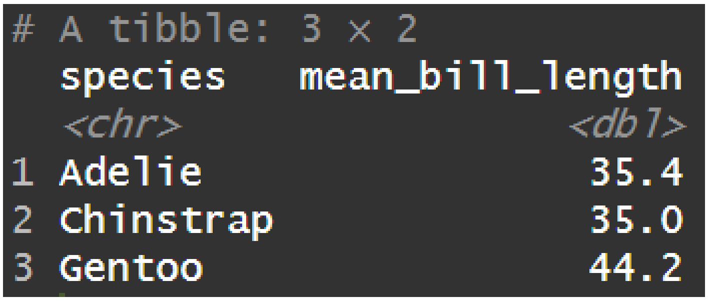
QQ-plot and histogram both show that the residuals aren’t quite normally distributed. QQ-plot plots the quantiles of the residuals against a theoretical distribution. It is best if it is a straight line.

Shapiro-Wilk test might be useful to formally test the normality of these residuals.

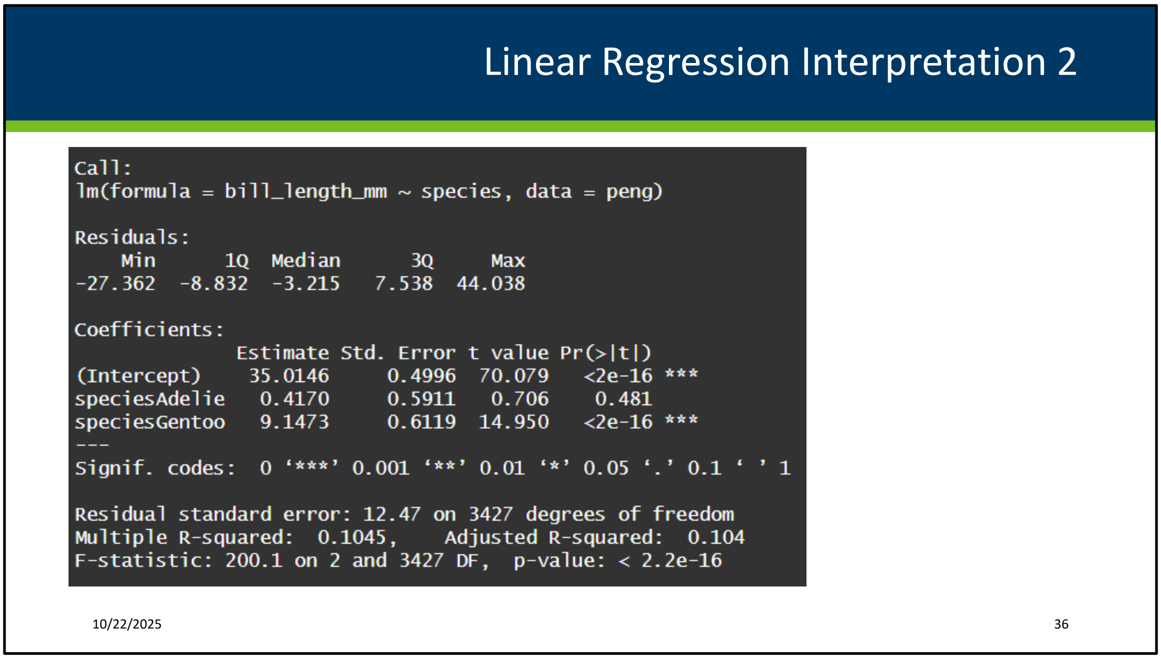
Residuals vs Bill Length is showing the bill length on the x-axis, and how residuals are dispersed for the tree species (three vertical lines). The variance of the residuals is fairly constant, which is good.

# Exploring the Palmer Penguin Data

Which species should serve as the reference category for the regressions?



```
# A tibble: 3 × 2
  species   mean_bill_length
  <chr>                <dbl>
1 Adelie                35.4
2 Chinstrap             35.0
3 Gentoo                44.2
```

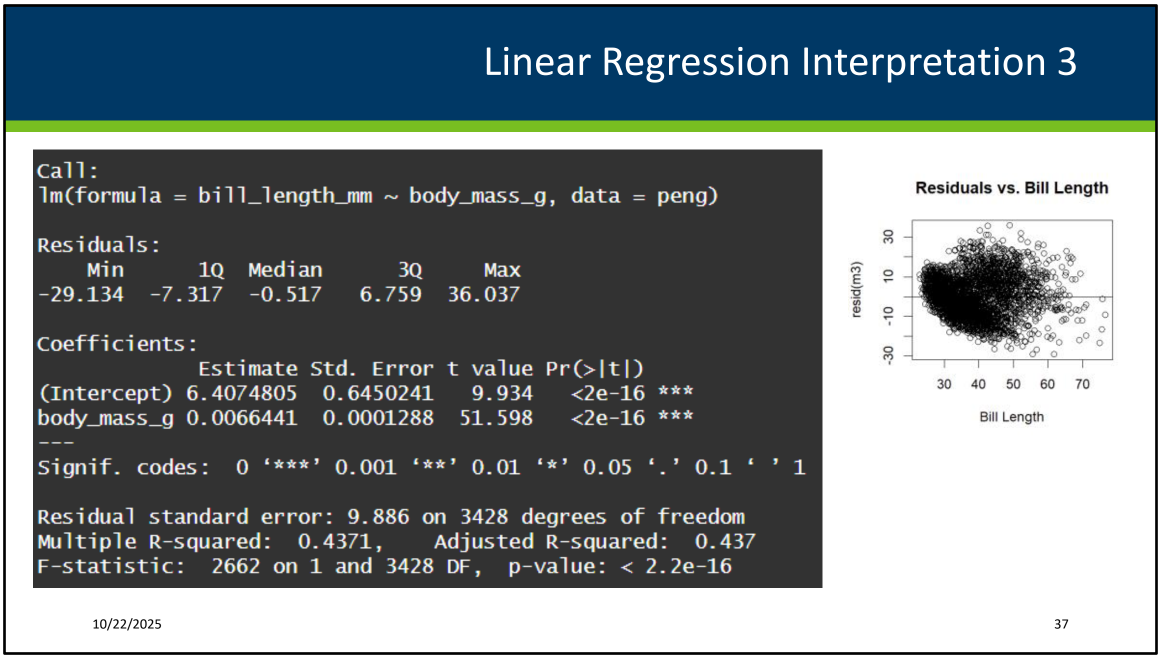


## Linear Regression Interpretation 2

```
Call:
lm(formula = bill_length_mm ~ species, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.362  -8.832  -3.215   7.538  44.038 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    35.0146     0.4996  70.079   <2e-16 ***
speciesAdelie   0.4170     0.5911   0.706    0.481    
speciesGentoo   9.1473     0.6119  14.950   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.47 on 3427 degrees of freedom
Multiple R-squared:  0.1045,	Adjusted R-squared:  0.104 
F-statistic: 200.1 on 2 and 3427 DF,  p-value: < 2.2e-16
```

10/22/2025 36

Average Chinstrap bill length is 35.01mm. The length of Gentoo beak is 9.15mm longer than Chinstrap, which is statistically significant at alpha = 0.05. Adelie bill length isn't statistically significantly different from Chinstrap.



## Linear Regression Interpretation 3

```
Call:
lm(formula = bill_length_mm ~ body_mass_g, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.134  -7.317  -0.517   6.759  36.037 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.4074805  0.6450241   9.934   <2e-16 ***
body_mass_g 0.0066441  0.0001288  51.598   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.886 on 3428 degrees of freedom
Multiple R-squared:  0.4371,	Adjusted R-squared:  0.437 
F-statistic:  2662 on 1 and 3428 DF,  p-value: < 2.2e-16
```

10/22/2025

37

What does the intercept 6.407 mean? Does this have scientific significance?

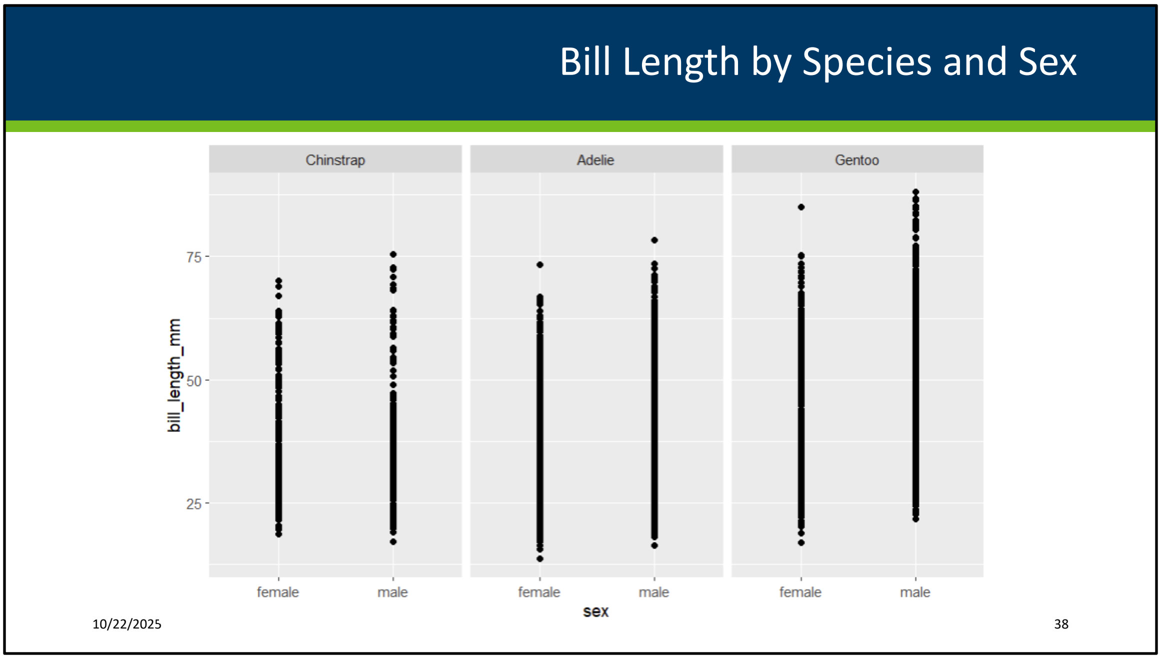
- It is the average length of the bill of all the species in mm if the penguin weighed 0 grams. It’s not a scientifically meaningful value.

For every increase of a penguin’s mass by one gram, their beak length increases by 0.0066mm.

It is better to multiply the coefficients and say that for every increase of a penguin’s mass by one kilogram, it’s bill length increases by 6.64mm

The residuals look fairly normally distributed, but what might be an issue? Could heteroscedasticity be a problem?

Could try transforming the data, or consider non-parametric regressions that we will not cover here.

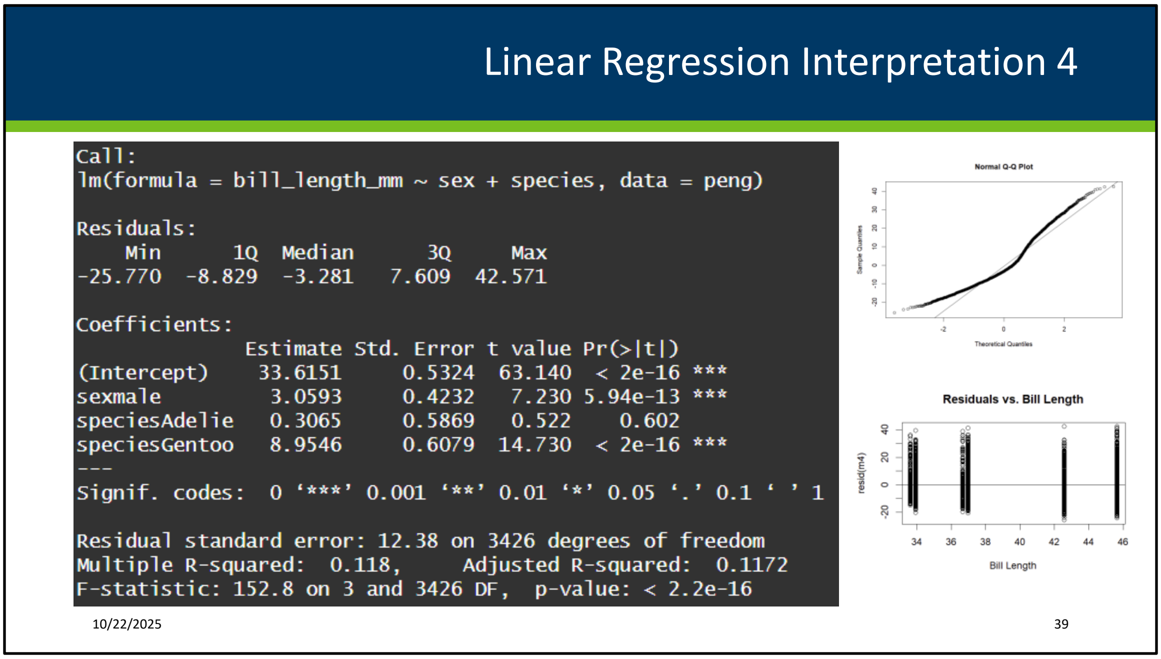


## Bill Length by Species and Sex

10/22/2025

38

Seems like males might have longer bills in Adelie and Gentoo, and not for Chinstrap, but it is difficult to tell



## Linear Regression Interpretation 4

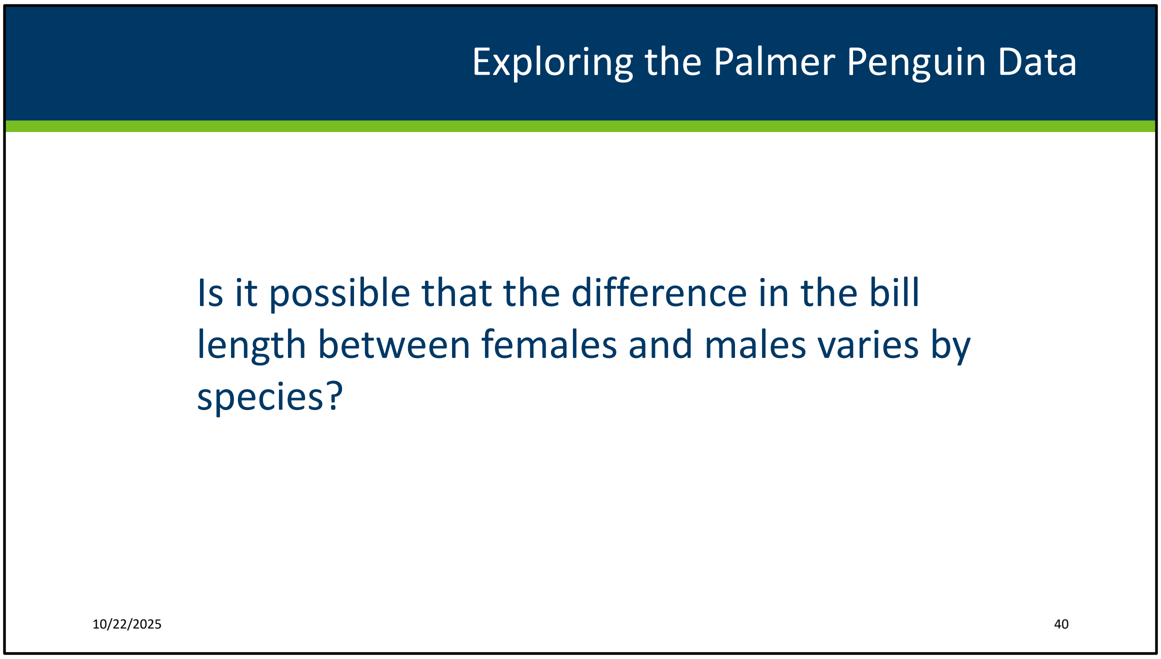
```
Call:
lm(formula = bill_length_mm ~ sex + species, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.770  -8.829  -3.281   7.609  42.571 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    33.6151     0.5324  63.140  < 2e-16 ***
sexmale         3.0593     0.4232   7.230 5.94e-13 ***
speciesAdelie   0.3065     0.5869   0.522    0.602    
speciesGentoo   8.9546     0.6079  14.730  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.38 on 3426 degrees of freedom
Multiple R-squared:  0.118,	Adjusted R-squared:  0.1172 
F-statistic: 152.8 on 3 and 3426 DF,  p-value: < 2.2e-16
```

10/22/2025

39

33.61mm is the bill length of Chinstrap females. The bill length of males of all species is on average 3.06 mm longer than the bills of females (this difference is statistically significant). Bill lengths of all Adelie are 0.307mm longer than bill lengths of Chinstrap females, but this difference isn’t significant. Bill lengths of all Gentoo is 8.95mm longer than female Chinstrap bill lengths, which is statistically significant.

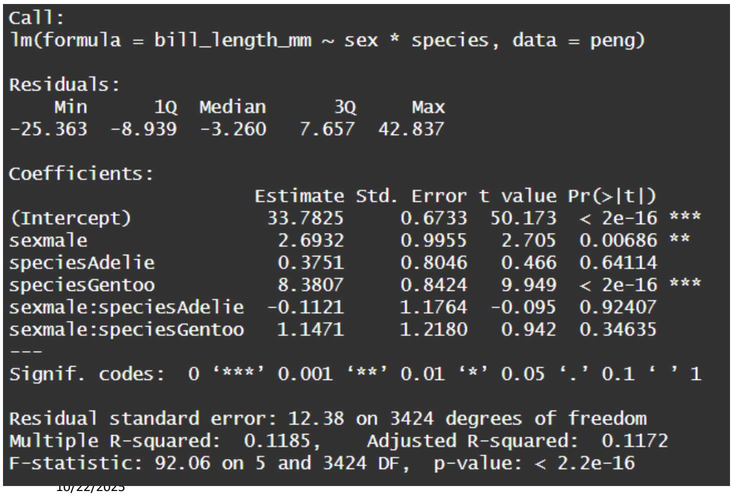
QQplot is not very normal..



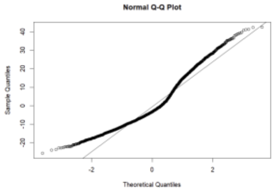
## Exploring the Palmer Penguin Data

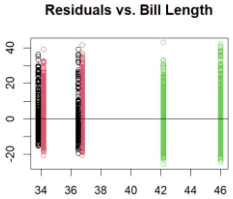
Is it possible that the difference in the bill length between females and males varies by species?

# Linear Regression Interpretation 5



```
Call:
lm(formula = bill_length_mm ~ sex * species, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.363  -8.939  -3.260   7.657  42.837 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            33.7825     0.6733  50.173  < 2e-16 ***
sexmale                 2.6932     0.9955   2.705  0.00686 ** 
speciesAdelie           0.3751     0.8046   0.466  0.64114    
speciesGentoo           8.3807     0.8424   9.949  < 2e-16 ***
sexmale:speciesAdelie  -0.1121     1.1764  -0.095  0.92407    
sexmale:speciesGentoo   1.1471     1.2180   0.942  0.34635    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.38 on 3424 degrees of freedom
Multiple R-squared:  0.1185,	Adjusted R-squared:  0.1172 
F-statistic: 92.06 on 5 and 3424 DF,  p-value: < 2.2e-16
```



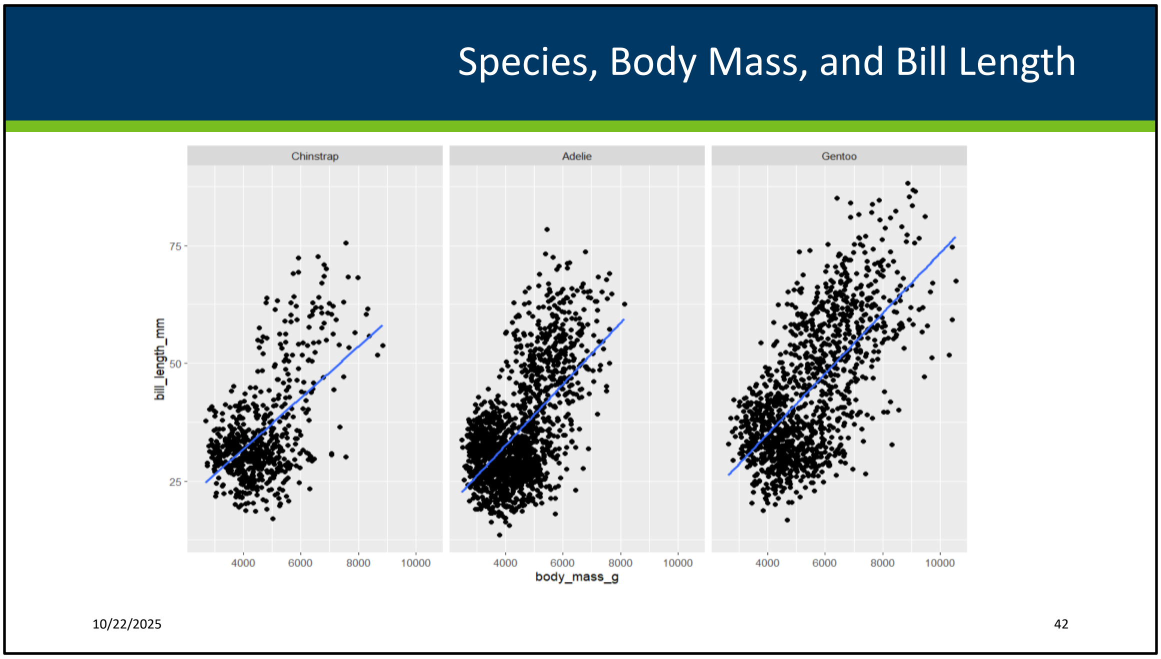


33.78mm – bill length of Chinstrap females. Average length of bills of male Chinstrap is 2.69mm longer, than in Chinstrap females. Bill lengths of female Adelie are 0.375mm longer than bill lengths of female Chinstraps, but this difference is not significant. Bill lengths of female Gentoo are 8.38mm longer than bill lengths of female Chinstrap, which is statistically significant.

-0.1121: difference in bill length between male Adelie and female Adelie is 0.1121mm less than difference in bill length between male Chinstrap and female Chinstrap. But this difference in difference is not significant.

1.147: difference in bill length of male and female Gentoo is 1.147mm bigger, than difference in bill length between male and female Chinstrap, though it’s also not significant.

Average bill length of female Adelie: 33.783+0.3751 = 34.16
Average bill length of male Adelie: 33.783+2.693 + 0.3751 – 0.1121 = 36.739
Average bill length of female Gentoo: 33.78+8.3807 = 42.161
Average bill length of male Gentoo: 33.78+2.69 + 8.3807 + 1.1471 = 45.99

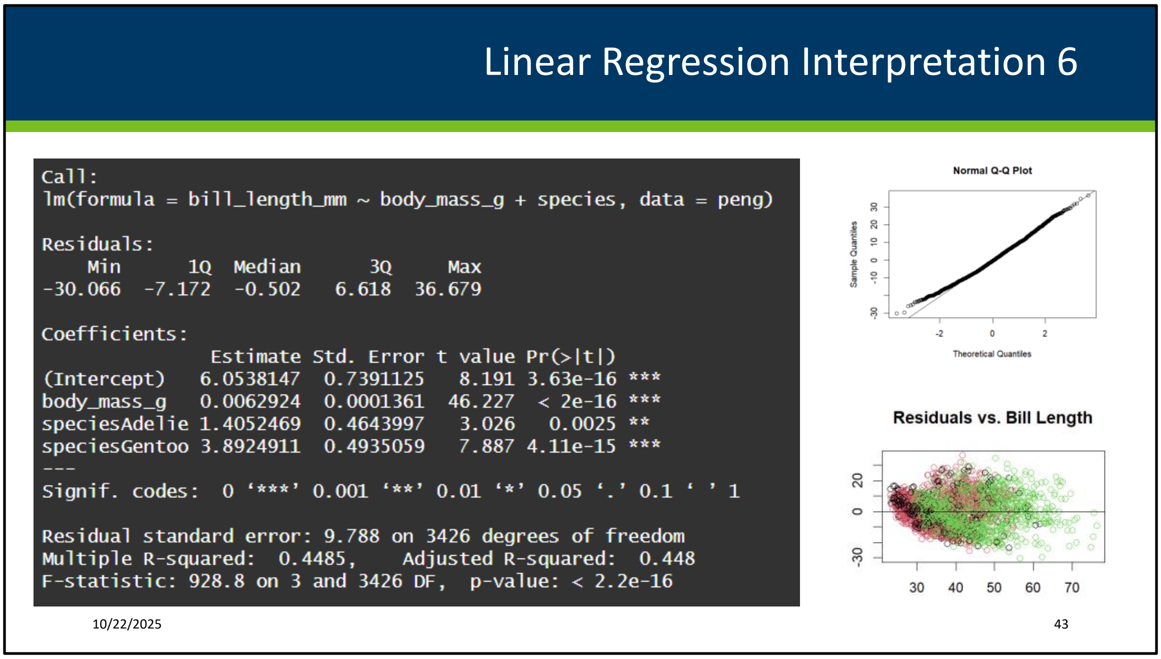


## Species, Body Mass, and Bill Length

10/22/2025

42

If we look at species, mass, and beak length together, we can see that Gentoo generally might be larger and have a longer beak.



## Linear Regression Interpretation 6

```
Call:
lm(formula = bill_length_mm ~ body_mass_g + species, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.066  -7.172  -0.502   6.618  36.679 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.0538147  0.7391125   8.191 3.63e-16 ***
body_mass_g   0.0062924  0.0001361  46.227  < 2e-16 ***
speciesAdelie 1.4052469  0.4643997   3.026   0.0025 ** 
speciesGentoo 3.8924911  0.4935059   7.887 4.11e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.788 on 3426 degrees of freedom
Multiple R-squared:  0.4485,	Adjusted R-squared:  0.448 
F-statistic: 928.8 on 3 and 3426 DF,  p-value: < 2.2e-16
```

10/22/2025

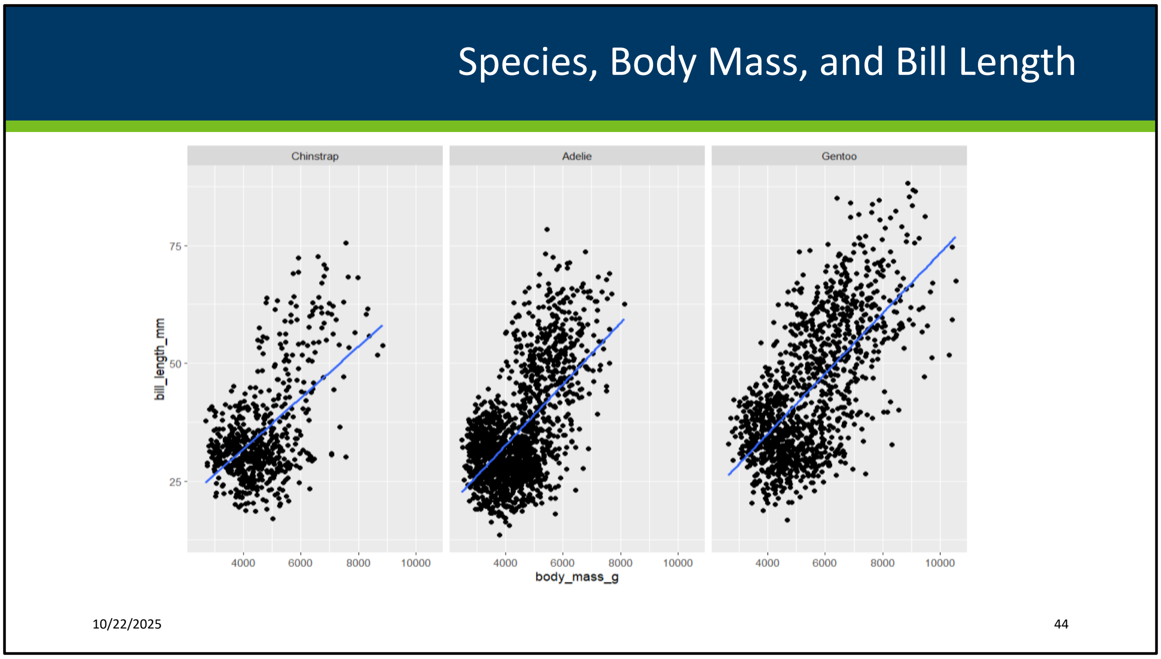
43

6.054 – bill length of Chinstrap at 0 grams of mass
Bill length of all species of penguins increases by 6.29mm for every kilogram of gained weight, which is statistically significant.

Adelie bill length is 1.405mm longer than Chinstrap bill length when both are at 0 grams of mass, while Gentoo bill length is 3.89mm longer at 0 grams of mass.
Does it make sense to discuss these specific differences? No, since 0 mass penguins do not exist.

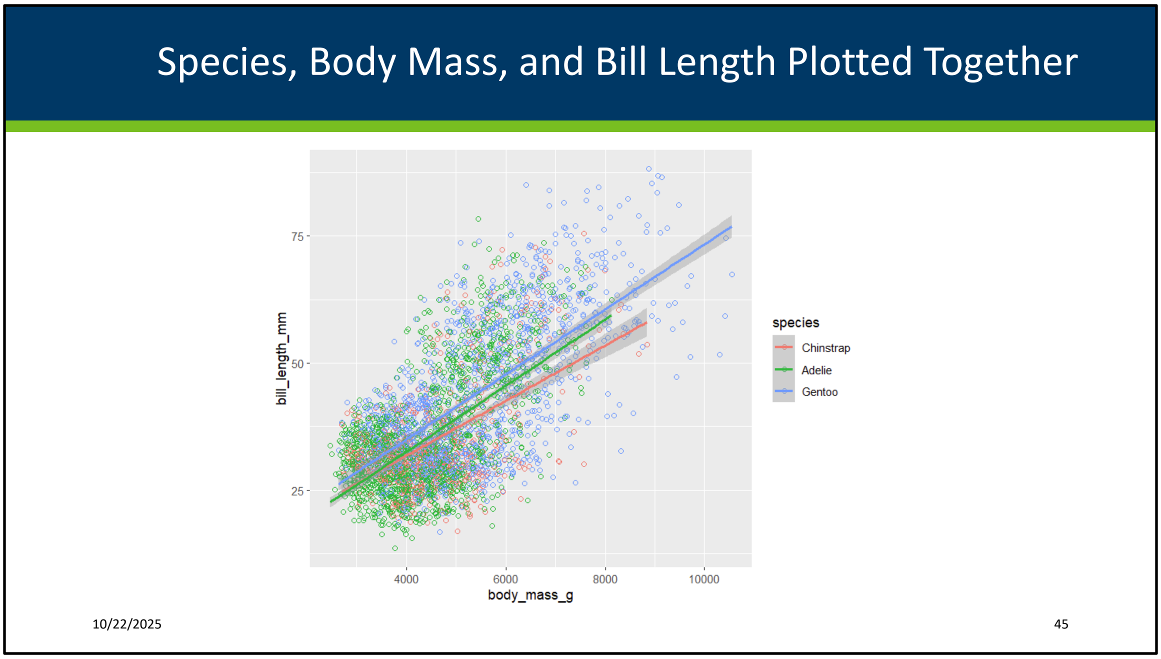
This model assumes that the change in bill length relative to body mass for all species is the same. However, this might not be the case!

Residuals QQplot looks normal. The species-specific vertical lines in the residual vs bill length graph are gone since they are masked by the continuous measures of the body mass



## Species, Body Mass, and Bill Length

Is there any way we could make this graph a bit easier to read?

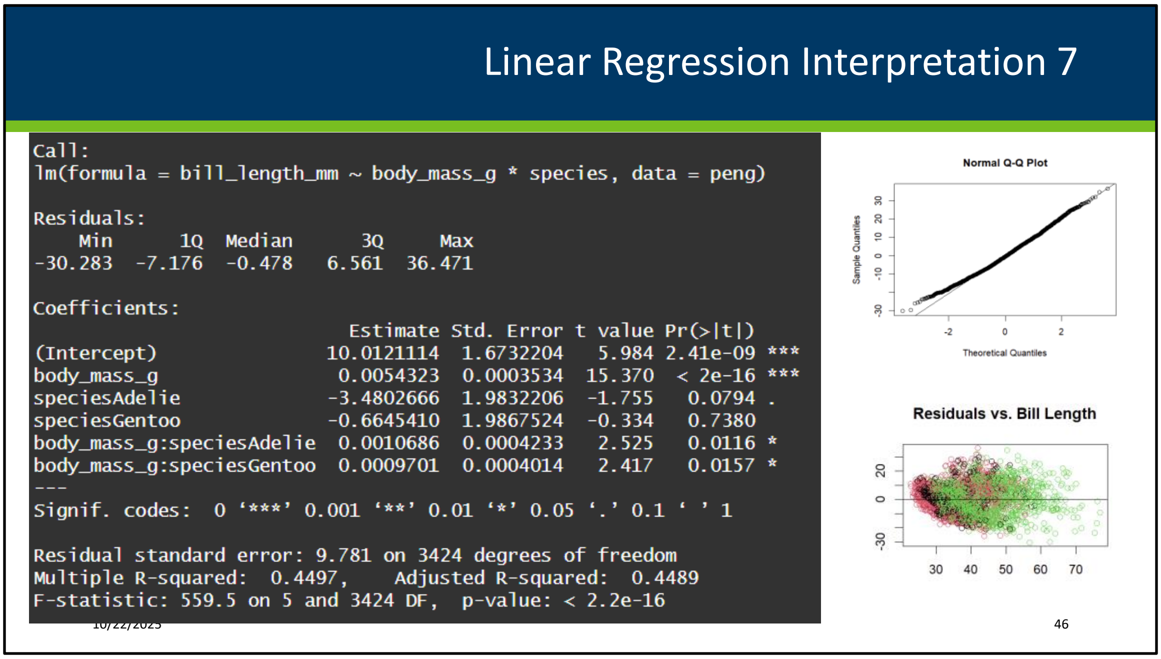


# Species, Body Mass, and Bill Length Plotted Together

10/22/2025

45

By plotting the graphs over each other, we can see that different species have slightly different slopes in the relationship between body mass and bill length



## Linear Regression Interpretation 7

```
Call:
lm(formula = bill_length_mm ~ body_mass_g * species, data = peng)

Residuals:
    Min      1Q  Median      3Q     Max
-30.283  -7.176  -0.478   6.561  36.471

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               10.0121114  1.6732204   5.984 2.41e-09 ***
body_mass_g                0.0054323  0.0003534  15.370  < 2e-16 ***
speciesAdelie             -3.4802666  1.9832206  -1.755   0.0794 .
speciesGentoo             -0.6645410  1.9867524  -0.334   0.7380
body_mass_g:speciesAdelie  0.0010686  0.0004233   2.525   0.0116 *
body_mass_g:speciesGentoo  0.0009701  0.0004014   2.417   0.0157 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.781 on 3424 degrees of freedom
Multiple R-squared:  0.4497,    Adjusted R-squared:  0.4489
F-statistic: 559.5 on 5 and 3424 DF,  p-value: < 2.2e-16
```

10/22/2025

46

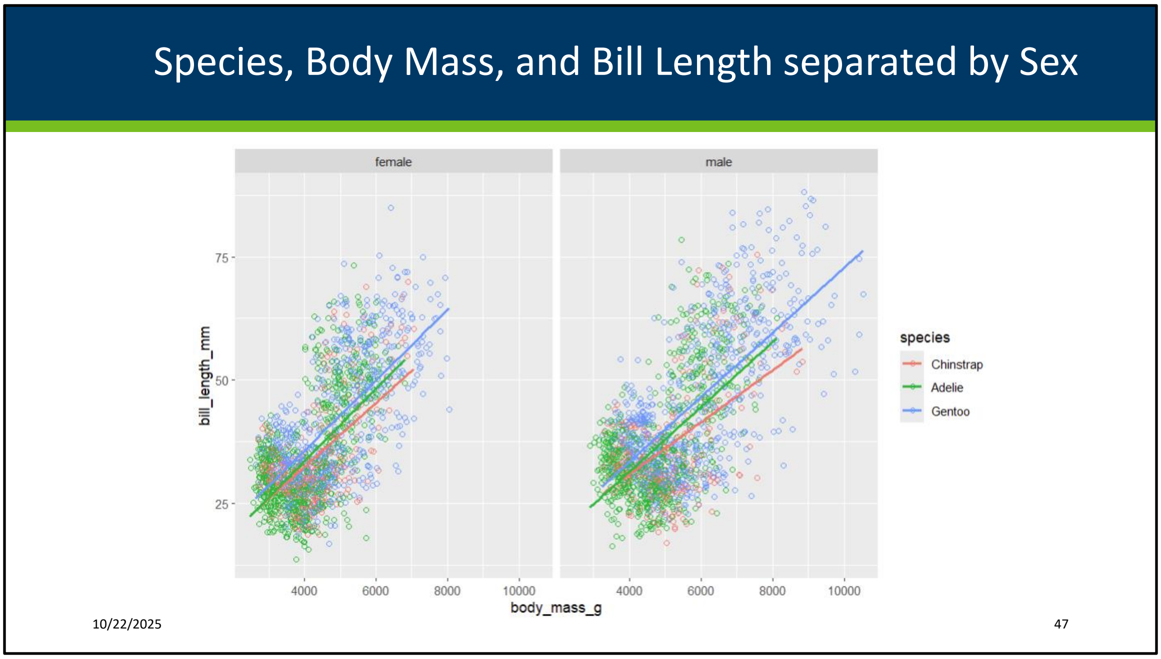
10.01 – bill lengths of all Chinstrap at 0g mass
Chinstrap bill length increases by 5.4mm for every kilogram it gains, which is statistically significant

Adelie bill length is 3.48mm shorter than Chinstrap bill length at 0g, while Gentoo bill length is 0.66mm shorter, though neither are significant

For every gained kilogram of mass Adelie bill length increases by 1.07mm more than the bill of Chinstrap for every kilogram it gains (statistically significant)

For every gained kilogram Gentoo bill length increases by 0.970mm more than the bill of Chinstrap for every kilogram it gains (statistically significant)

What will be the beak length of a 6.5kg Gentoo? 10.01– 0.664 + (0.0054+0.0009701)*6500 = 50.7

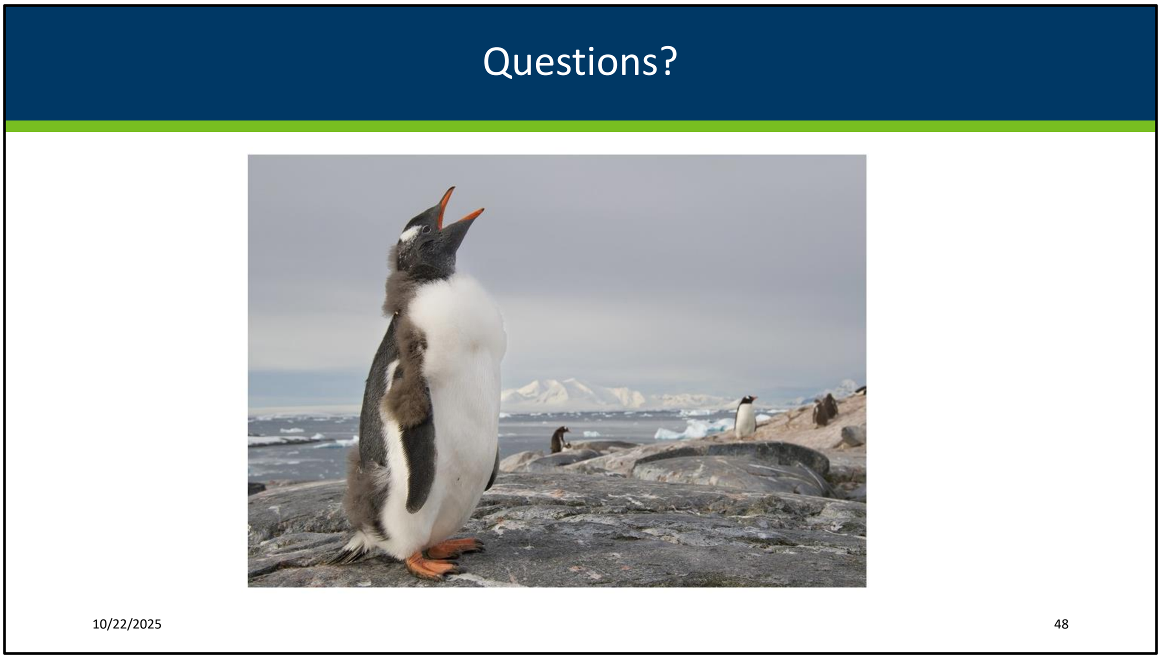


## Species, Body Mass, and Bill Length separated by Sex

10/22/2025

47

Interpretations?



## Questions?

Interpretations?