MAD – Data Analysis & Biostatistics in R

Inference & Hypothesis Tests

James R. Hunter, Ph.D.

DIPA, EPM, UNIFESP

25 September 2020

MAD

Section 1

But, First, Association and Correlation

Two-Way Views of Variables

- Relationship between 2 variables
- Variables stratified by different levels of a 2nd variable
  - Age (numeric) by Gender (categorical)
  - Viral Load (numeric) by Age_Group (categorical)
  - Death from SARS-CoV-2 (Logical) by Gender (categorical)
  - CD4+ T Cell Count by HIV-1 Viral Load (both numeric)

Portray Relationships in Table or Graph Format

- When the variables are both categorical
  - we use counts and proportions
- When at least one is numeric, we generally use means for that category

Example #1 with summarytools::ctable()

- With SEADE comorbidity data set, show the number of deaths by gender
- Frequency Table

```
library(summarytools)
ctable(sp_comorb$death, sp_comorb$sex)
```

```
## Cross-Tabulation, Row Proportions
## death * sex
## Data Frame: sp_comorb
##
## ------- ----- ------------- ------------- --------------
##           sex        female          male          Total
##   death
##   FALSE             93 (47.0%)   105 (53.0%)   198 (100.0%)
##    TRUE             42 (41.2%)    60 (58.8%)   102 (100.0%)
##   Total            135 (45.0%)   165 (55.0%)   300 (100.0%)
## ------- ----- ------------- ------------- --------------
```

Example #2 with dplyr

- With fm, show mean age per sex

```r
fm %>%
  group_by(sex) %>%
  summarise(mean_age = mean(age))
```

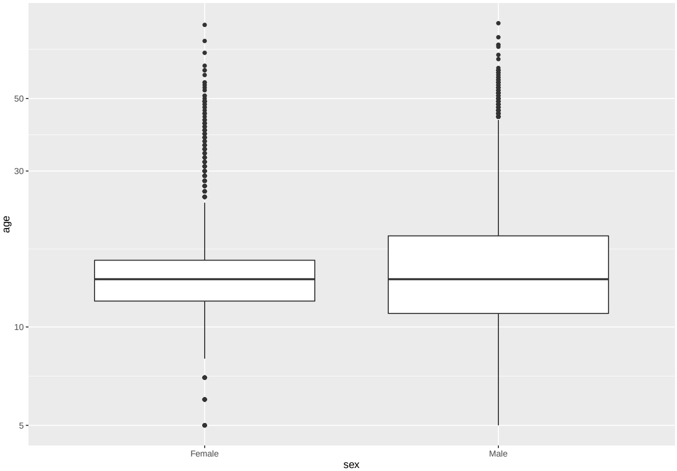
```
## # A tibble: 2 x 2
##   sex    mean_age
##   <fct>     <dbl>
## 1 Female     15.4
## 2 Male       17.1
```

Get More Information on `age` and `sex`

- Violin plot - advanced form of boxplot
    - Useful when too many points to use `geom_jitter()`
    - Shows density of points along y-axis

Boxplot of age and sex

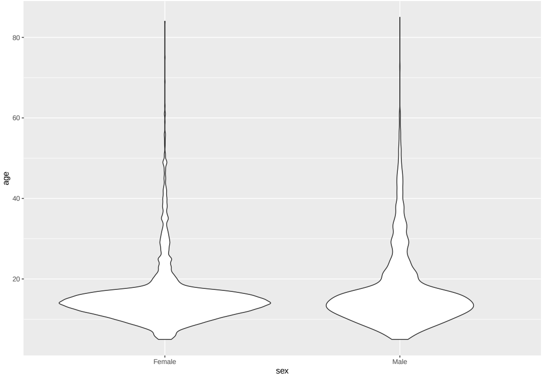
```
fm %>%
  ggplot(aes(x = sex, y = age)) +
  geom_boxplot() +
  labs(x = "sex", y = "age") +
  scale_y_log10()
```



Violin Plot of Same Data

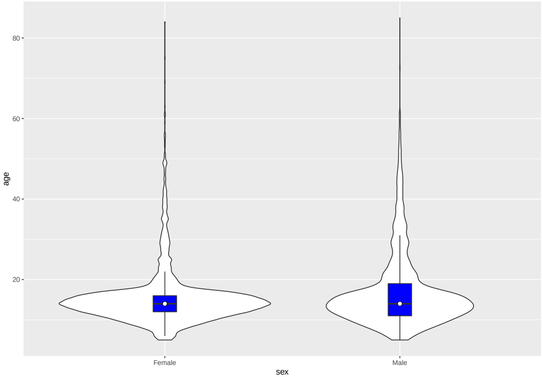
- Clearer view of where most of the values lie

```
gg_viol <- fm %>%
  ggplot(aes(x = sex, y = age)) +
  geom_violin() +
  labs(x = "sex", y = "age")
gg_viol
```



Put Boxplot Information into Violin Plot

```
gg_viol +
geom_boxplot(width = .1, fill = "blue", outlier.colour = NA) +
stat_summary(fun = median, geom = "point", fill = "white", shape = 21, size = 2.5)
```



Example #3 with cor()

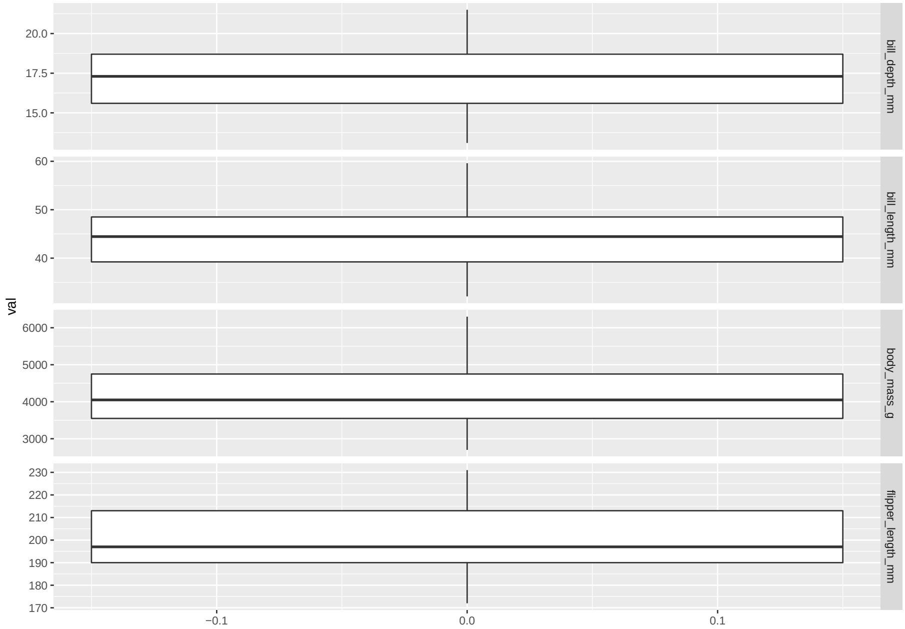
- Use Penguin example data set to determine correlation among
  - lengths of bills and flippers
  - mass
- First, look at data

```r
library(summarytools)
peng %>%
  select(bill_length_mm:body_mass_g) %>%
  descr(stats = "common") %>%
  knitr::kable()
```

	bill_depth_mm	bill_length_mm	body_mass_g	flipper_length_mm
Mean	17.151170	43.921930	4201.7544	200.91520
Std.Dev	1.974793	5.459584	801.9545	14.06171
Min	13.100000	32.100000	2700.0000	172.00000
Median	17.300000	44.450000	4050.0000	197.00000
Max	21.500000	59.600000	6300.0000	231.00000
N.Valid	342.000000	342.000000	342.0000	342.00000
Pct.Valid	99.418605	99.418605	99.4186	99.41860

How Are These Distributed?

```
peng_long <- peng %>%
  mutate(peng_num = 1:nrow(peng)) %>%
  select(peng_num, bill_length_mm:body_mass_g) %>%
  pivot_longer(cols = bill_length_mm:body_mass_g, names_to = "v", values_to = "val")

peng_long %>% group_by(peng_num) %>%
  ggplot(aes(y = val)) +
  geom_boxplot(width = .3) +
  facet_grid(rows = "v", scales = "free")
```

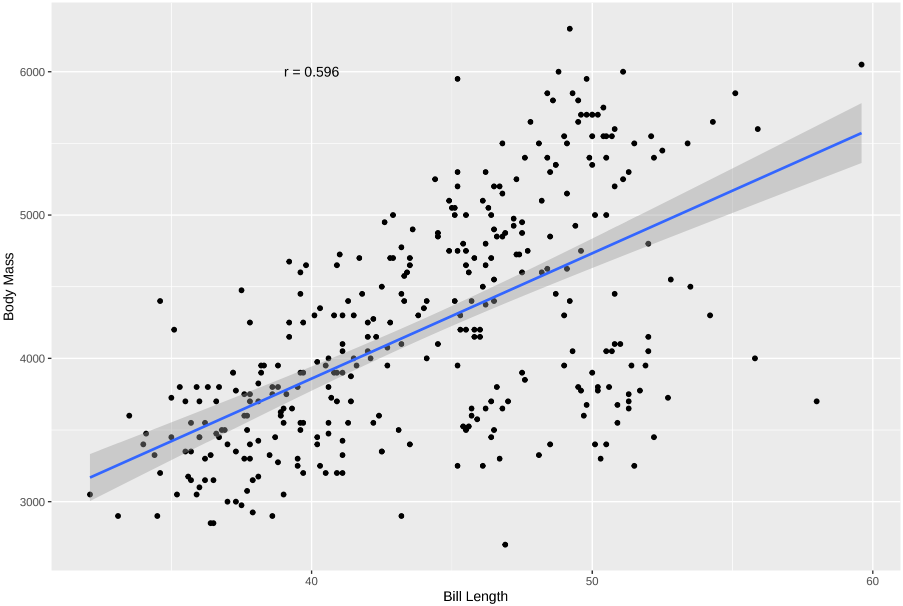

20.0
17.5
15.0
bill_depth_mm
60
50
40
bill_length_mm
val
6000
5000
4000
3000
body_mass_g
230
220
210
200
190
180
170
flipper_length_mm
−0.1
0.0
0.1

How Do These Variables Interact?

- Body Mass x Bill Length

```r
peng %>%
  ggplot(aes(x = bill_length_mm, y = body_mass_g)) +
  geom_point() +
  labs(x = "Bill Length", y = "Body Mass", title = "Penguin Body Measurements") +
  geom_smooth(method = "lm")
```

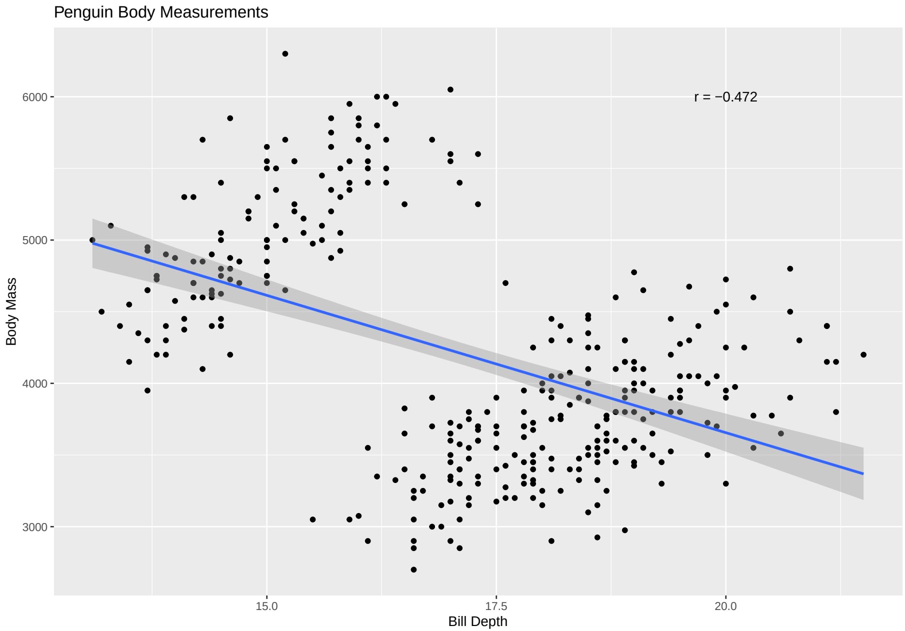
Penguin Body Measurements



Another Pair of Measurements

- Body Mass x Bill Depth

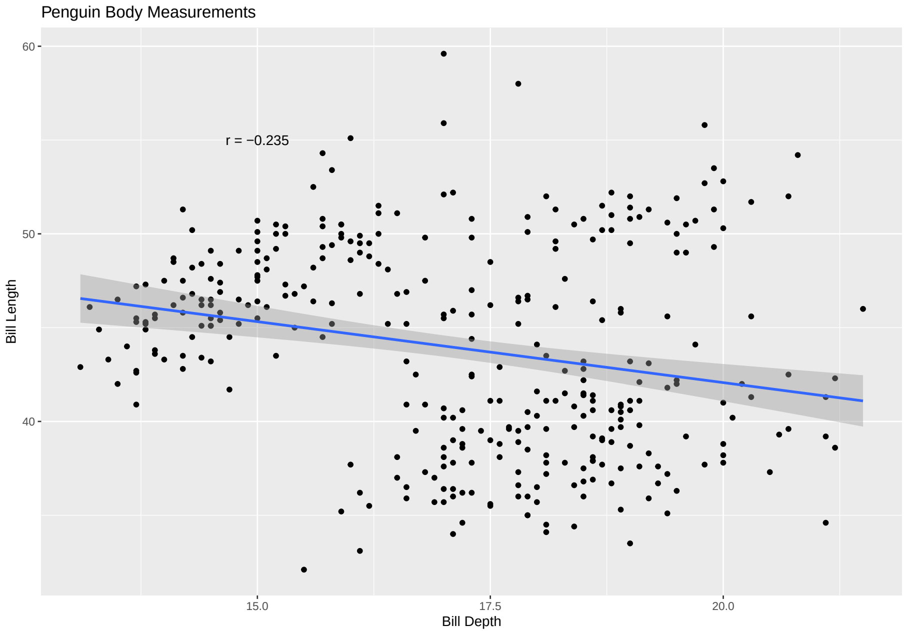
```r
peng %>%
  ggplot(aes(x = bill_depth_mm, y = body_mass_g)) +
  geom_point() +
  labs(x = "Bill Depth", y = "Body Mass", title = "Penguin Body Measurements") +
  geom_smooth(method = "lm")
```


Penguin Body Measurements
r = −0.472
Body Mass
6000
5000
4000
3000
15.0
17.5
20.0
Bill Depth

Another Pair of Measurements

- Bill Length x Bill Depth

```r
peng %>%
  ggplot(aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point() +
  labs(x = "Bill Depth", y = "Bill Length", title = "Penguin Body Measurements") +
  geom_smooth(method = "lm")
```


Penguin Body Measurements
r = −0.235
60
50
40
15.0
17.5
20.0
Bill Depth
Bill Length

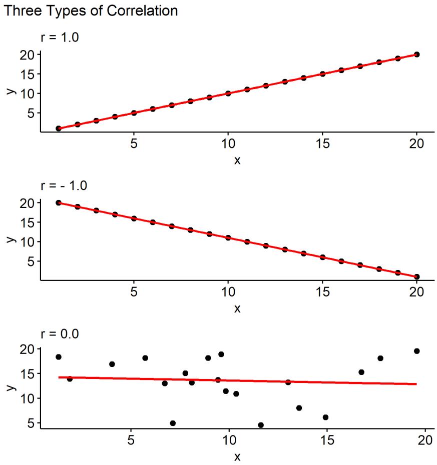
Three Distinct Type of Relationships

- Body Mass x Bill Length
  - As bill length gets longer, body mass increases
  - Positive trend
- Body Mass x Bill Depth
  - As bill depth gets larger, body mass decreases
  - Negative trend
- Bill Length x Bill Depth
  - As bill depth gets larger, no clear trend in bill length
  - Still a bit negative, but largely flat
  - Are they associated? correlated?

Correlation of These Variables

- Concept of Pearson's correlation coefficient
  - Population: "ρ" (Greek letter rho)
  - Sample: "r"
- Measures how one variable *varies* against another
- Varies from -1 to 1
  - r = -1: perfect negative relationship between variables
  - r = 0: no relationship between variables
  - r = 1: perfect positive relationship between variables

Three “Pure” Correlations



But, What is Correlation?

- We looked at variance for 1 variable
- When we have 2 variables, we look at covariance
  - How much variance exists when considering both variables

$$Cov_{x,y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}$$

- Looks just like variance formula
- Denominator: $N - 1$ same as variance
- Instead of squaring deviations, we have 2 deviations multiplied

OK, Correlation is a Form of Covariance?

- Problem with covariance: How to interpret units
- Correlation *standardizes* covariance by dividing out estimated standard deviation of each
- Turns it into a pure number: no units

$$r_{x,y} = \frac{Cov(x,y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

- $\hat{\sigma}_x$ = estimated population standard deviation of x

Calculating Correlation (r)

- Base R function `cor()`
- If we give it two specific variables to correlate, it will return correlation coefficient
- If we give it vector of variables, it will return matrix of correlations among all
- Argument: `use =` controls missing data (`NA`)
  - If data has missing values, `use = "complete.obs"`

cor() Case 1: two variables

```r
cor(peng$bill_length_mm, peng$body_mass_g, use = "complete.obs")
```

```
## [1] 0.5951098
```

cor() Case 2: Matrix of Variables

```r
peng %>%
  select(bill_length_mm:body_mass_g) %>%  # choose numeric vars
  cor(., use = "complete.obs") %>%
  knitr::kable()
```

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
bill_length_mm	1.0000000	-0.2350529	0.6561813	0.5951098
bill_depth_mm	-0.2350529	1.0000000	-0.5838512	-0.4719156
flipper_length_mm	0.6561813	-0.5838512	1.0000000	0.8712018
body_mass_g	0.5951098	-0.4719156	0.8712018	1.0000000

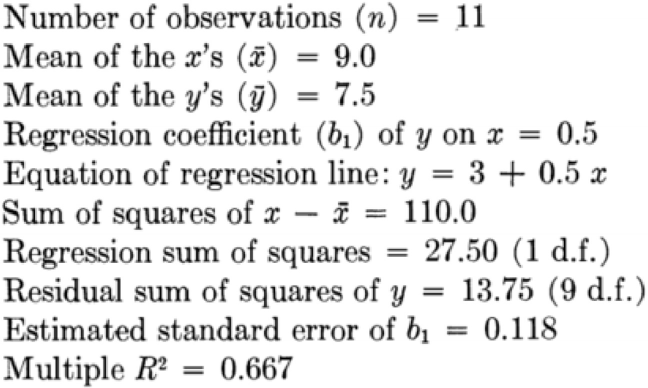
Correlation – Interpretation

- **Always** look at a scatterplot before interpreting correlation
- Same correlation coefficient could mean different things
- Check what you are correlating: spurious correlations

Anscombe's Quartet

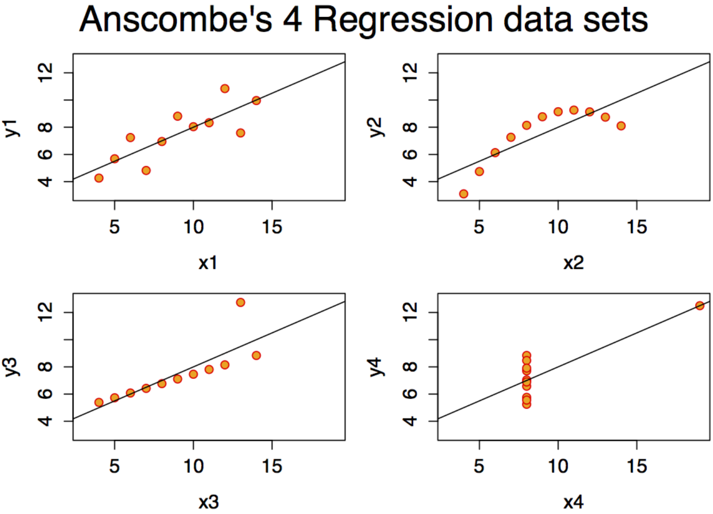
- Physicist Anscombe in 1973 showed same summary measures
- All with r = 0.8167
- Could come from very different data sets

Anscombe's EDA



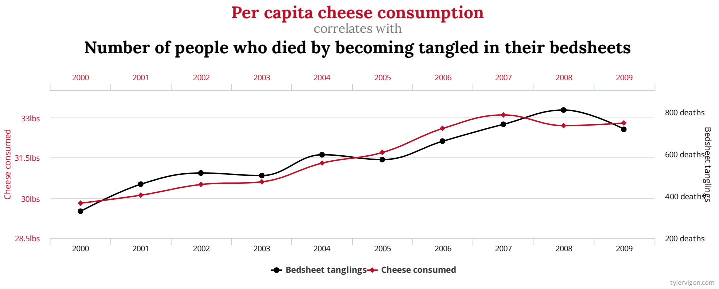
Number of observations (n) = 11
Mean of the x's ($\bar{x}$) = 9.0
Mean of the y's ($\bar{y}$) = 7.5
Regression coefficient (b_1) of y on x = 0.5
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of y = 13.75 (9 d.f.)
Estimated standard error of b_1 = 0.118
Multiple R^2 = 0.667

What They Look Like



Spurious Correlations

- The following correlation is above 0.95



Vigen, Tyler, https://www.tylervigen.com/spurious-correlations

Pearson Correlation vs. Others

- r actually measures degree of linear relationship between two variables
  - How tightly they line up together
- What if their relationship is not linear?
- Use *Spearman* correlation
  - Based on ranks
  - Instead of working with variances of numbers
  - Uses relative ranks of variables
  - Non-parametric

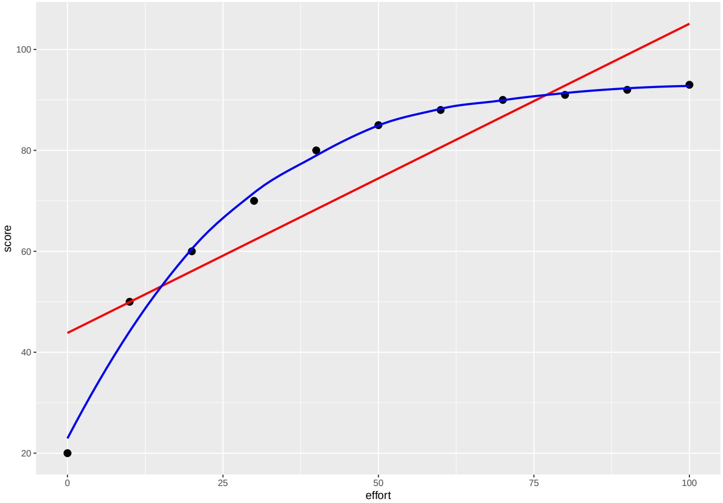
Toy Example: Effort vs. Test Score

- Work that goes into studying for a test
  - Not linear
  - 0 effort can even produce a score of 20 (of 100)
  - 20% effort can get you a 60
  - But, to get a top score requires much more effort
- Is effort correlated to test score?
- *Thanks to D. Navarro, Learning Statistics with R, for this example*

```
test <- tibble(effort = seq(from = 0, to = 100, by = 10),
               score = c(20, 50, 60, 70, 80, 85, 88, 90, 91, 92, 93))
test %>% knitr::kable()
```

effort	score
0	20
10	50
20	60
30	70
40	80
50	85
60	88
70	90
80	91
90	92
100	93

```
ggplot(test, aes(x = effort, y = score)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  geom_smooth(method = "loess", color = "blue", se = FALSE)
```



Correlation Using Pearson and Spearman

```r
cor(test$effort, test$score, method = "pearson")
```

```
## [1] 0.884
```

```r
cor(test$effort, test$score, method = "spearman")
```

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

- Spearman coefficient $= 1$
  - Rank of each point same for both variables

```r
rank(test$effort)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11
```

```r
rank(test$score)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11
```

Section 2

Inference

Populations x Samples

- Population: People enrolled in course
  - Can select various samples of 10 of people in course
- Population: Brazilians

Do the Measurements Make a Difference?

- We can measure a number of characteristics of variables
  - Point estimates
    - Sample mean ($\bar{x}$)
    - Sample standard deviation (s)
- So what?
  - One sample could have a mean of 10.5, another of 17
  - Do they represent the same population?
  - Sampling error: how an estimate varies from 1 sample to another
  - Can select many samples of 2,000 people to stand in for all Brazilians
  - How many people in sample infected with COVID-19?
  - Does that number **accurately** represent # of Brazilians infected?

Section 3

Statistical Inference Allows Us to Work with Samples

Example: Support for Compulsory COVID Vaccination in Terra de Ninguem

- Small country with population of 212,910 people
- 80% support compulsory vaccination
- Will a sample of 500 Terrans show same proportion?

Terminology of Proportions

- Working with proportions
- Proportions are decimal representation of percentages
- 80% of population = proportion of 0.8
- Proportion = p
- Estimate of a proportion = $\hat{p}$

Set up Terran Population

- 80% of population will support vaccination, 20% not
- Use function `rep()` from base R
  - `rep` stands for “repeat” or “replicate”

```
tdn_pop <- 212910
support_not <- fct_relevel(factor(c(rep("support", 0.8 * tdn_pop), rep("not", 0.2 *
summarytools::freq(support_not, cumul = F)
```

```
## Frequencies
## support_not
## Type: Factor
##
##                   Freq   % Valid   % Total
## ------------- -------- --------- ---------
##       support   170328     80.00     80.00
##           not    42582     20.00     20.00
##          <NA>        0                0.00
##         Total   212910    100.00    100.00
```

Take Sample

- Pull a random 1,000 Terrans from population
- See what their proportion of support is
  - Sum of Terrans who support divided by sample size

```r
N <- 1000
sample_1 <- sample(support_not, size = N)
# Calculate p-hat
samp1_prop <- sum(sample_1 == "support") / N
samp1_prop
```

```
## [1] 0.801
```

```r
samp1_error <- samp1_prop - .8
samp1_error
```

```
## [1] 0.001
```

Try It Again – New Sample

```r
sample_2 <- sample(support_not, size = N)
# Calculate p-hat
samp2_prop <- sum(sample_2 == "support") / N
samp2_prop
```

```
## [1] 0.838
```

```r
samp2_error <- samp2_prop - .8
samp2_error
```

```
## [1] 0.038
```

Single Simulations – Boring

- Let’s do 10,000 simulations and see if we get close to the population value of 0.8

```
set.seed = 42
sims <- 10000
sim_results <- numeric(length = sims) # holds sims # of results

for (i in 1:sims) {
  samp <- sample(support_not, size = N)
  res <- sum(samp == "support") / N
  sim_results[i] <- res
}
```

Summary Statistics for sim_results

```r
summarytools::descr(sim_results, stats = "common", round.digits = 4)
```

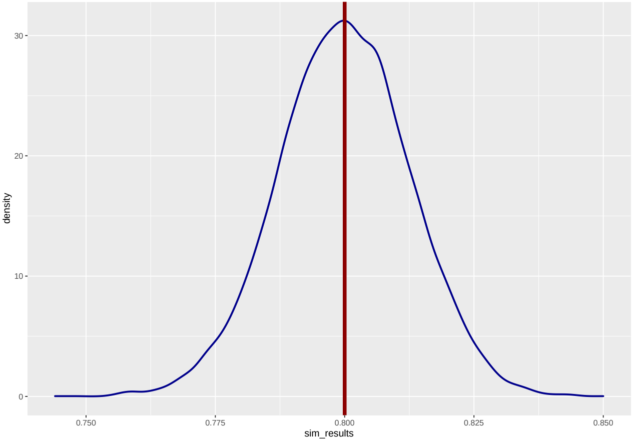
```
## Descriptive Statistics
## sim_results
## N: 10000
##
##                      sim_results
## --------------- -------------
##            Mean        0.7998
##         Std.Dev        0.0178
##             Min        0.7260
##          Median        0.8000
##             Max        0.8660
##         N.Valid    10000.0000
##       Pct.Valid      100.0000
```

- Here, the `Std.Dev` is of the distribution of samples and measures the errors across all 10,000 samples
- In this case, it is called standard error
- Relates only to deviations of sampling distributions

Density Curve of sim_results

- Alternative to a histogram

```
tibble(sim_results) %>%
  ggplot(aes(x = sim_results)) +
  geom_density(color = "darkblue", size = 1) +
  geom_vline(xintercept = 0.8, colour = "darkred", size = 2)
```



Section 4

Central Limit Theorem

Our Graph and Normal Curves

- Our graph looks like a normal curve
- Thanks to Central Limit Theorem
- **If** observations are independent (come from a random sample)
- **If** observations meet the **success-failure** condition
- **Then** estimated proportion ($\hat{p}$) will follow a normal distribution

Success-Failure Condition

- Size of sample needs to be large enough that
  - $np \geq 10$
  - $n(1-p) \geq 10$
- In our case, wildly exceed these tests

Parameters of Normal Proportions

- Normal distribution has 2 parameters: mean (μ) and standard deviation (s)
    - Remember: In dealing with trials, we substitute standard error (SE)
- Mean: $\mu_{\hat{p}} = p$
- Standard Error:

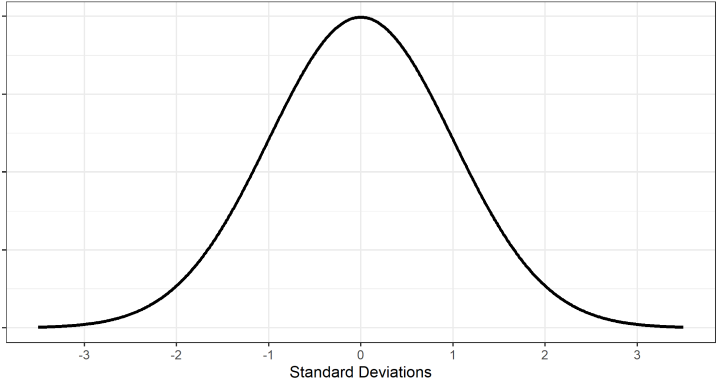
$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Section 5

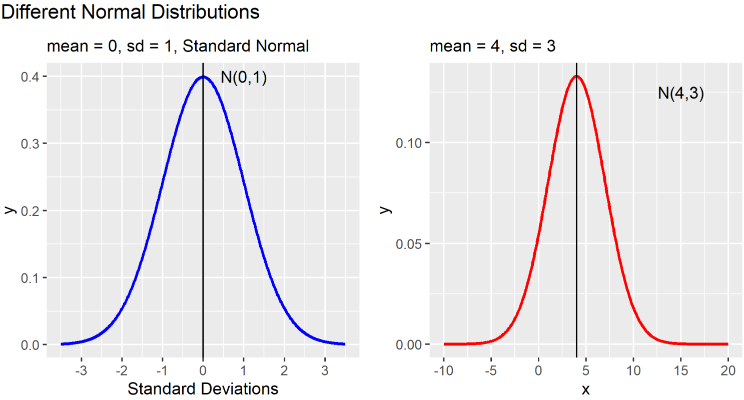
Normal Distribution

Familiar bell-shaped curve

- Symmetrical around mean



Two Different Normal Curves



Standardizing (Normalizing) a Distribution

- Take any distribution of numbers
- *Centralize* it by subtracting out the mean (μ)
- *Standardize* it by dividing by the standard deviation (σ)
- Called the Z-Score
    - Z commonly used to indicate a normalized variable

$$Z_{x_i} = \frac{(x_i - \mu)}{\sigma}$$

Standarize Penguin Weights

- Palmer Penguins data set (peng)
  - Variable body_mass_g

```
(mean_peng_mass <- mean(peng$body_mass_g, na.rm = TRUE))
```

```
## [1] 4201.754
```

```
(sd_peng_mass <- sd(peng$body_mass_g, na.rm = TRUE))
```

```
## [1] 801.9545
```

```
peng <- peng %>%
  mutate(body_mass_z = (body_mass_g - mean_peng_mass)/sd_peng_mass)

(mean(peng$body_mass_z, na.rm = TRUE))
```

```
## [1] 8.237527e-17
```

```
(sd(peng$body_mass_z, na.rm = TRUE))
```

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

Summary of Penguin Body Mass

```r
peng2 <- peng %>%
  mutate(peng_num = 1:344)
summarytools::descr(peng2$body_mass_g, stats = "common")
```

```
## Descriptive Statistics
## peng2$body_mass_g
## N: 344
##
##                    body_mass_g
## --------------- -------------
##            Mean       4201.75
##         Std.Dev        801.95
##             Min       2700.00
##          Median       4050.00
##             Max       6300.00
##         N.Valid        342.00
##       Pct.Valid         99.42
```

Where Does a Single Penguin Fit in the Mass?

- Work with penguin # 197

```
p181 <- peng2 %>%
  filter(peng_num == 197) %>%
  select(species, body_mass_g, sex, year, peng_num)
p181
```

```
## # A tibble: 1 x 5
##   species body_mass_g sex    year peng_num
##   <chr>         <dbl> <chr> <dbl>    <int>
## 1 Gentoo         5550 male   2008      197
```

In Distribution of Gentoo Penguins, How Big Is He?

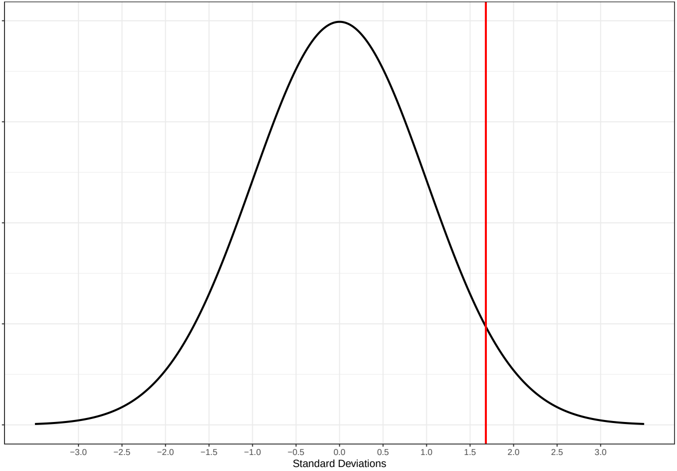
- Mean of penguins: 4201.75g
- Difference: 5550 - 4201.75 = 1348.25
- Calculate Z score: divide by s = 801.9545

```
(z_score <- (5550 - mean_bm) / std_dev_bm)
```

```
## [1] 1.6812
```

What Does That Mean?

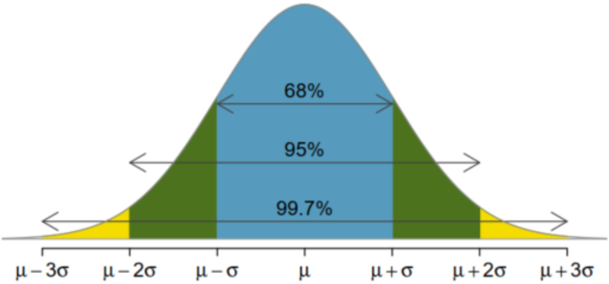
- Normalized value of 1.6811996 means he is 1.6811996 standard deviations away from mean



Calculate Exactly How Big He Is Relatively

- Function `pnorm()` of a Z-score tells you the percentile he falls in
- pnorm(z_score) = pnorm(1.68) = 0.9536379
- Out of all the penguins, he falls in the 95th percentile
  - Bigger than all but 5% of the penguins

Refresh Standard Deviations



Section 6

Normal Distribution & Decisions about Proportions

Return to the Terrans

- Support for compulsory COVID-19 vaccination
- We know population value is 80% support
- We take a sample of 300 from city X, result is 75% support
- Is this figure different than the 80% or is it within a sampling error margin?
- Is this deviation of 5% from 80% simply due to chance
  - Does the data provide strong evidence that the population proportion is different from 80%?[1]
- Test idea that a real difference exists or not
  - Hypothesis test
  - Confidence interval

[1] Diez, et.al., OpenIntro Stats, 4th ed.

Hypothesis Test

- 2 competing notions
  - There is **no difference** between the view of people in city X and all Terrans (skeptical)
  - There is a **real difference**
- These are hypotheses
  - Ideas we can test with data
- We will test the hypothesis of no real difference (H_0) or the null hypothesis
- Other hypothesis: H_1 or alternate hypothesis

Result of Hypothesis Test

- By comparing the difference in proportions against a normal distribution
- If the difference is larger than we could expect by chance
    - We **reject** the null hypothesis
    - Note: **not** accept the alternative hypothesis
- If not larger, need to accept hypothesis that no real difference exists

Hypothesis Test

- H_0: $p = 0.8$
  - (proportion equal to population proportion)
- H_1: $p \neq 0.8$
  - (proportion not equal to population proportion)

First Issue: Have We Met Success-Failure Condition?

- n = 300, p = 0.8
- np = 300 * 0.8 = 240
- n(1 - p) = 300 * 0.2 = 60
- Both are greater than 10

Compute Sample Error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

```r
p <- 0.8; n <- 300
se <- sqrt(p * (1 - p) / n)
se
```

```
## [1] 0.02309401
```

Use SE to Compute Z-Score

```
cityx_support <- 0.75
z_support <- (cityx_support - p) / se
z_support
```

```
## [1] -2.165064
```

Determine Where on Normal Curve This Value Lies

```r
(p_val_cityx <- pnorm(z_support))
```

```
## [1] 0.01519141
```

- This is the famous p-value
- Compare p-value to a standard, which we normally call α (Greek alpha)
- If value is more extreme than our standard value (α), we say the difference is statistically significant
- We need to find out why city X harbors so many non-supporters

Definition of p-value

The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.[2]

[2] Diez, et.al., OpenIntro Stats, 4th ed.

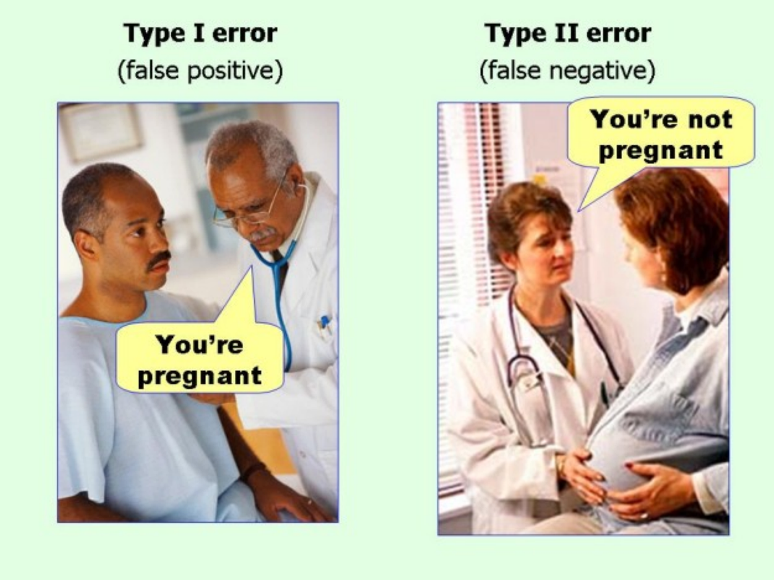
Types of Error

Table 4: Truth Table

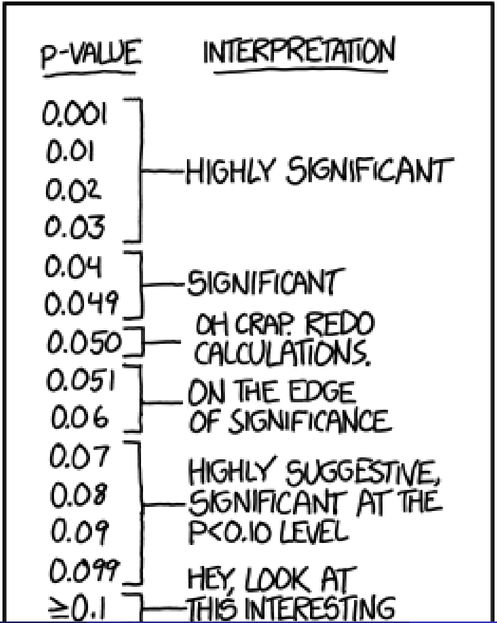
truth	h0_accept	h0_reject
H0 True	ok	Type 1 Error
H1 True	Type 2 Error	ok

- p(Type 1 Error) $= \alpha$

Types of Error - 2



Interpreting the p-Value


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING

Confidence Interval Method

- If Central Limit Theorem conditions satisfied
  - They are
  - $\therefore$ data follows a normal distribution
- 95% of all values fall within 1.96 standard deviations of mean
  - 2.5% on each side will be outside this limit
- Z value for 95%: `qnorm(.975)` = 1.959964
  - `qnorm()` stands for *quantile function*
- We can double check this with `pnorm()`
  - `pnorm(1.9601)` = 0.9750079
  - Returns the percentage we gave it

Constructing Confidence Interval

- We want an interval for a 95% confidence interval that extends
    - 1.96 standard deviations from point estimate ($\hat{p}$)
- This says we are 95% confident that population proportion will fall in this interval
    - point estimate ± 1.96 x standard error
    - Remember that standard error stands in for standard deviation
- The 1.96 x SE is called the margin of error

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

95% Confident – What Does This Mean?

- Refers to all the different possible samples of size 300 we could have taken of the 212,910 Terrans
- 95% of all the confidence intervals related to these samples will include the true p of 0.8
- We are therefore 95% confident that **our** sample has the population p in its confidence interval

Our Sample Confidence Interval

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

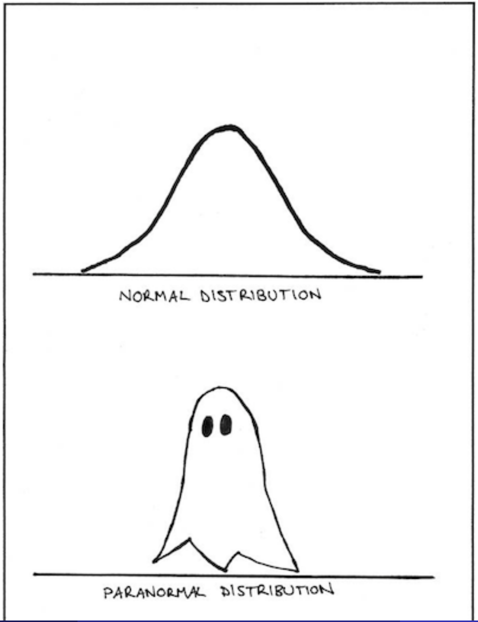
```r
p <- 0.8
p_hat <- 0.75
n <- 300
ci_high <- p_hat + (qnorm(0.975) * sqrt(p * (1 - p)/n))
ci_lo <- p_hat - (qnorm(0.975) * sqrt(p * (1 - p)/n))

paste("95% Confidence Interval:", round(ci_lo, 4), "to", round(ci_high, 4))
```

```
## [1] "95% Confidence Interval: 0.7047 to 0.7953"
```

Conclusion Based on Confidence Interval

- Interval does not include the true proportion, 0.80
- We need to find out why city X harbors so many non-supporters
    - *just like hypothesis test*
- You will get same results with confidence intervals and hypothesis tests

Before We Leave Normal Distributions



Section 7

Comparing Means - t-Tests

Compare “Adelie” and “Gentoo” Penguins’ Body Mass

- Two of the penguin species
- How much bigger is one species than the other?
- Is that difference important/significant?

```
peng3 <- peng %>%
  filter(!is.na(body_mass_g),
         species != "Chinstrap") %>%
  mutate(peng_num = 1:274)
```

New Question: Are Gentoos Bigger than Adelies?

- Summary stats by species

```
peng3  %>%
  group_by(species) %>%
  summarytools::descr(body_mass_g, stats = "common")
```

```
## Descriptive Statistics
## body_mass_g by species
## Data Frame: peng3
## N: 151
##
##                      Adelie    Gentoo
## ---------------- --------- ---------
##             Mean   3700.66   5076.02
##          Std.Dev    458.57    504.12
##              Min   2850.00   3950.00
##           Median   3700.00   5000.00
##              Max   4775.00   6300.00
##          N.Valid    151.00    123.00
##        Pct.Valid    100.00    100.00
```

Mean of Gentoos Bigger

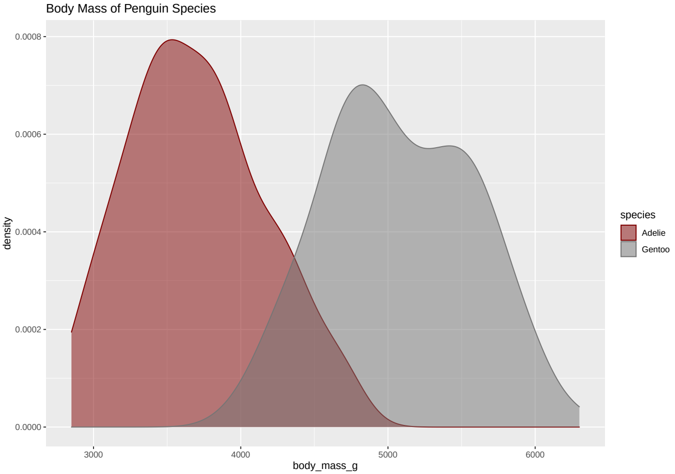
- Is that because Gentoos are really bigger?
- Is it a result of a sampling error?
    - The penguins that were chosen for this sample
- Would a different sample produce a different result?
- If this is a sample, do we know the population standard dev (σ)?
    - No, because this is not a full census
        - Only 274 penguins of thousands
    - We only know sample standard dev (s)

What Our Question Really Is

Is the population mean of Gentoos greater than that of Adelies?

- Is the difference in means **real** or result of **chance**
- Based on the *sample* mean and standard deviation of the penguins measured

Comparative Density Graph of Species



Code for Comparative Density Graph

```r
uchic <- ggsci::pal_uchicago("default")(9)[1:3]
peng3 %>%
  mutate(species = factor(species)) %>%
  ggplot(aes(x = body_mass_g, color = species, fill = species)) +
  geom_density(alpha = .5) +
  labs(title = "Body Mass of Penguin Species")+
  scale_fill_manual(values = uchic) +
  scale_color_manual(values = uchic)
```

How Do We Measure “Larger”?

- Locate the value of the difference on the curve of a distribution in terms of standard deviations
    - How many standard deviations away from a mean difference of 0 is the *measured* difference
    - Just as we did with locating a single penguin on a normal curve
- Translate that number of standard deviations into location on the curve
- Compare that location against the standard we want to use to measure it

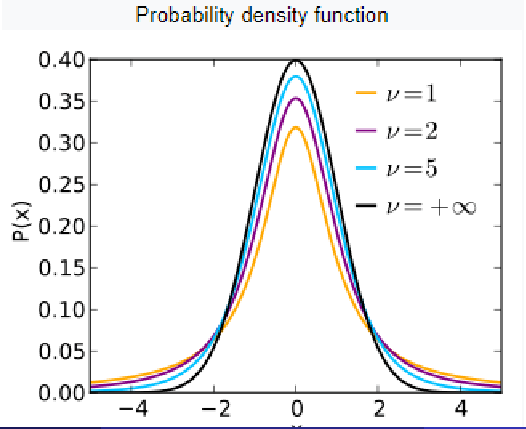
t-Distribution Rather than Normal Distribution

- In this case, we are working with a sample
  - We do **not** know the population standard deviation that normal needs
- We use "Student's t-Distribution"
  - Family of distributions
  - Single parameter of t-distribution: degrees of freedom(df)
  - For t, df is sample size - 1

$$df = n - 1$$

Shape of Student's t-Distribution

- Comparison to normal
  - Bell-shaped
  - Symmetric
  - Fatter tails, lower center
- Larger df, closer to normal



What is Difference That Interests Us

- Mean of Gentoo - Mean of Adelie
- R can handle most of calculations
- Because done in base R, t-tests work better with vectors
- t-test function (`t.test()`) will report most results you need

Code for t-Test

```r
gentoo <- peng3 %>%
  filter(species == "Gentoo") %>%
  select(body_mass_g)
adelie <- peng3 %>%
  filter(species == "Adelie") %>%
  select(body_mass_g)

# conduct test
 tt <- t.test(x = gentoo$body_mass_g, y = adelie$body_mass_g,
       alternative = "two.sided")
 tt
```

```
##
##  Welch Two Sample t-test
##
## data:  gentoo$body_mass_g and adelie$body_mass_g
## t = 23.386, df = 249.64, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1259.525 1491.183
## sample estimates:
## mean of x mean of y
##  5076.016  3700.662
```

What Do the Test Results Mean?

- Null hypothesis: Difference between species is 0
  - Results report H_1
- Sample estimates
  - mean of x: 5076.0162602 (Gentoo)
  - mean of y: 3700.6622517 (Adelie)
  - Difference: 1375.3540085
- 95% Confidence Interval
  - 1259.525172, 1491.182845
  - Does not include 0 - value we wanted to test
- t-statistic: 23.3860277
  - Calculated as t_{df}=(sample mean - null value)/SE
- SE and degrees of freedom are complicated to calculate
  - Because sample sizes are different for the two species
- p-value
  - Well below α of 0.05

Conclusion

- p-value more extreme than 0.05
- We **reject** the null hypothesis of no difference
- Practical conclusion: difference in species exists
- What obstacles to understanding source of differences
  - Confounding variables
  - Bias in sampling

Other Types of t-Tests

- Paired samples
- Single sample
- Non-parametric comparison of means
- We shall return!

Section 8

Probability

History

- Starting point for statistics
- Subject of Interest since Twelfth Century
- Originally derived from games of chance
- Great thinkers
  - Fibonacci, 12th Century
  - Girolamo Cardano (*Liber de Ludo Alae*), 15th cent.
  - Chevalier de Méré, Blaise Pascal, Pierre de Fermat, 16th cent.
  - Abraham de Moivre, Gauss, Bernoulli(s), LaPlace, 17th - 18th cents.

Probability Scale

- Probability is a number between 0 and 1
- A pure number; no units
- Probability 0: the event is impossible
- Probability 0.5: the event is equally possible and impossible
- Probability 1.0: the event is certain

Simple Example

- What is the probability that a 1 will appear on a fair die that you throw on the table?
- Die has 6 numbers
- Only will will turn up
- 1 chance in 6 or p = 1/6 or 16.67%

$$p[1] = p[2] = p[3] = p[4] = p[5] = p[6] = 0.1667$$

What is the Probability of Winning the Mega-Sena?

- Select 6 numbers between 1 and 60
- Can we select the same number twice
  - NO – without replacement
  - How do you calculate it
- Only 1 combination of 6 numbers can win

How Do We Count in Probability?

- Frequentist view of probability
- Count all the possible solutions with the desired result (**WIN**)
- Count all the possible solutions
- Compare the two

$$P_{win} = \frac{\text{all the desired solutions}}{\text{all possible solutions}}$$

Mega-sena Probability

- Numerator: Only 1 combination wins
- Denominator: All the possible combinations
- **How many combinations of 6 numbers from 1 to 60 can be chosen**

Combinations

- If we take r objects from a set of n objects without replacement and without reference to order, how many different combinations are possible?

$$\begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{r!(n-r)!}$$

- Spoken: “n choose r”
- Alternative written form: ${}_nC_r$

Factorials

$$5! = 5 * 4 * 3 * 2 * 1 = 120$$

- Factorials grow very quickly

Number	Factorial
1	1
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	362880
10	3628800

$$\frac{1}{50063860} = 0.000000020 = 2.0^{-08}$$

Odds of Winning Mega-Sena

- Odds used frequently in medical research
- Definition: ratio of chances of winning to chances of losing

$$Odds = \frac{\text{chances of winning}}{\text{chances of losing}}$$

$$\text{Odds Mega-Sena} = \frac{1}{50,063,380 - 1} = \frac{1}{50,063,379}$$

Combinations and Permutations

- How many different codons are possible using the 4 DNA bases?
- $4^3 = 4 * 4 * 4 = 64$
- Permutation: order counts
    - GGC - glycine
    - CGC - arginine
- $(\text{results per event})^{\text{events}} = n^r$
- When either you have “with replacement” or “order counts”
    - Use powers