STA130 W5

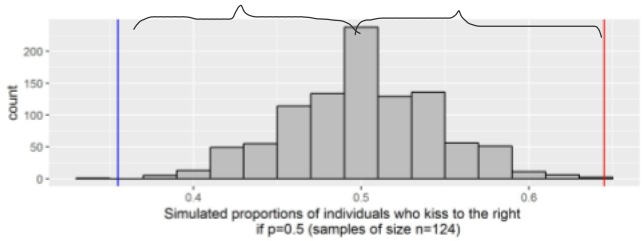
Wednesday, October 2, 2019 12:29 PM

The logic of hypothesis testing(kissing example)
1. State hypotheses
2. Calculate test statistic(from sample data)
3. Simulate values of the statistic under the null hypothesis (Ho)
4. Evaluate the evidence against Ho
5.Make a conclusion
   - We have strong evidence against the null hypothesis that people have no preference of .side when they kiss.

p = 0.5 vs p ≠ 0.5
P-hat = 80/24 = 0.645
*p-hat: test statistic

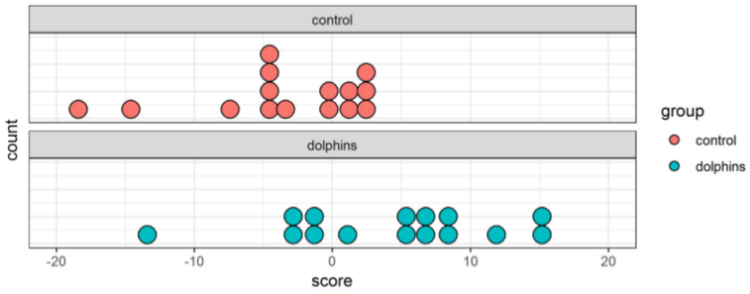
Distance btw p-hat and 0.5
0.145 both side

Distance between p-hat and 0.5




```
sim %>%  filter(p_right>=0.645 | p_right<=0.355)%>%
summarise(p_value = n() / repetitions)
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.001
```
*bounded btw 0 - 1 exclusive
*P value cannot be 0, but it can be very close to 0

Swimming with dolphins: A treatment for mild depression?
Research question: Does the presence of dolphins help some depression patients improve at a higher rate than other individuals in similar circumstances?
Sample: 30 adults (18-65 years old) with mild or moderate depression, randomized to one of two groups
Control group
1 hour/day of swimming and snorkeling
Treatment group
1 hour/day of swimming and snorkeling... with dolphins
Outcome: After two weeks record the change in depression symptoms for each person



1.Is there a difference in the scores for the two groups?
Yes, there may be a difference
2.What test statistic would you calculate to compare the two groups?
Means, median, proportion of people who improve in each group, range(max-min), sd/variance

1. State hypotheses

Translate research question into statements involving parameters
μ = "mu" represents the mean
Ho: μd - μc = 0 or μd = μc
Ha: μd -μc ≠ 0 or μd ≠ μc
The mean score is **not** the same for the 2 groups
Where μd = mean score for people who swim w. dolphins
μc = without dolphins

2. Calculate the test statistic

```
mean_data <- data %>%
group_by(group) %>% dolphin and control groups
summarise(means = mean(score))
mean_data
## # A tibble: 2 x 2
##   group    means
##   <fct>    <dbl>
## 1 control  -3.46
## 2 dolphins  4.18
```

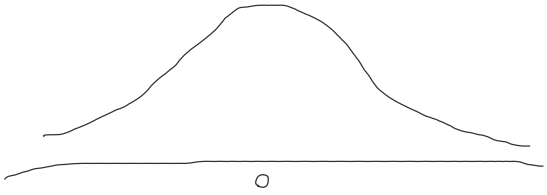
Sample mean in the dolphin group

$$\text{Test statistic} = \hat{\mu}_{dolphin} - \hat{\mu}_{control} = 4.18 - (-3.46) = 7.64$$

```
# The diff function calculates the difference between values in a vector
test_stat <- mean_data %>%
    summarise(test_stat = diff(means)) diff: 4.18 - (-3.46) (value in row 2 minus val. In row
    1)
test_stat
## # A tibble: 1 x 1
 ##   test_stat
##       <dbl>
## 1     7.64
```

3. Model behaviour of the statistic under Ho

What is the null hypothesis Ho?
No difference in the group means
If there was no real difference between the two treatments (dolphins vs control), what would you expect the distribution of **μ-hat dolphins - μ-hat control** to look like?



Under null hypothesis, shuffle the group wont change the mean(15.8)

μ-hat d = 0.15	**μ-hat c = 5.27 -- 3.81**
μ-hat d = 6.175	**μ-hat c = 7.03 -- (-1.76)**

What does it mean to simulate data under Ho ?
If Ho what is the relationship between the grouping (i.e. dolphins vs control) and the outcome (i.e. the improvement score)?
No connection btw group label (colour) and the response (score) under Ho
We want to simulate many possible values of what the statistic could have been if Ho was true to estimate the distribution of its possible values under Ho

Simulating values of the test statistic under **without** a computer
- Get n cards (n1 of one colour, and n2 of the second colour)
- Shuffle the cards and distribute one to each of the n = n1 + n2 observational units
- Calculate the statistic for each group defined by cards of each colour, and then take the difference
- Repeat the above many times

We can simulate shuffling the individuals into groups using R
The **sample()** function can be used to shuffle (i.e. reorder) the labels
By default, sample() returns a random sample of the same length as the original vector, **without replacement**

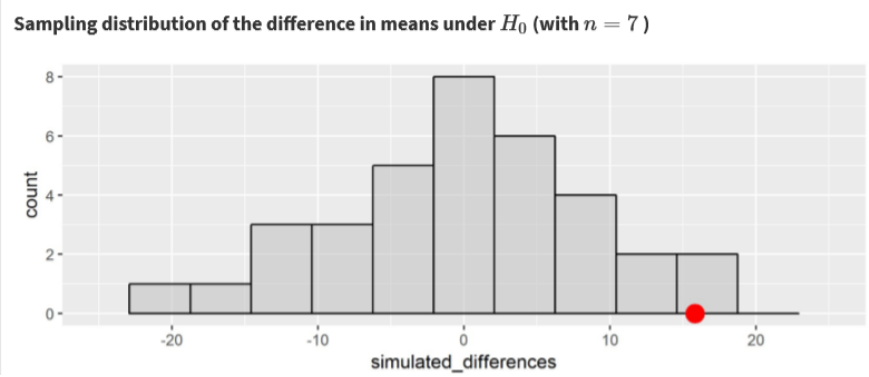
```
a_vector <- c(1,1,1,2,2)
a_vector
```

```
## [1] 1 1 1 2 2
sample(a_vector)
## [1] 2 1 1 2 1
sample(a_vector)
## [1] 2 1 1 1 2
```

*Replacing original group labels with shuffled version of group labels

Original data for 7 people	After shuffling the group labels
data_n7 %>% select(group, score) Real data ## group score ## 4 dolphins 15.5 ## 6 dolphins -2.8 ## 8 dolphins 9.0 ## 11 dolphins 14.9 ## 41 control 2.0 ## 13 control -14.6 ## 14 control -7.4	# One possible grouping under H_0 data_n7 %>% select(group, score) %>% **mutate**(group=**sample**(group)) One simulated(shuffled) dataset ## group score ## 1 dolphins 15.5 ## 2 control -2.8 ## 3 control 9.0 ## 4 dolphins 14.9 ## 5 control 2.0 ## 6 dolphins -14.6 ## 7 dolphins -7.4 # Another possible grouping under H_0 data_n7 %>% select(group, score) %>% mutate(group=sample(group)) ## group score ## 1 control 15.5 ## 2 dolphins -2.8 ## 3 dolphins 9.0 ## 4 control 14.9 ## 5 dolphins 2.0 ## 6 control -14.6 ## 7 dolphins -7.4

How many possible groupings are there?
There are 7! / (4! * 3!) = 35 possible groupings ("7 choose 4")
For each possible grouping, we can calculate the value that the statistic would take. This gives an estimate of the sampling distribution of the test statistic **under the assumption that there is no difference between the groups (i.e. the treatment does not affect**



In our real sample of 7 people, the difference of means was 15.8. If the null hypothesis of no difference between treatment groups was true, is it likely that we observe a difference of this size?
We would expect to see a diff at least as extreme approx. 4/35 times(~0.1) Estimated p-value

Exact permutation approach
- Consider all possible groupings
- Calculate the test statistic for each grouping
- Plot a histogram of the sampling distribution of the statistic under Ho
- Calculate the p-value and evaluate the evidence against Ho

As the sample size increases, the number of possible groupings increases very quickly - too many

Random permutation test
Instead of looking at all possible groupings, we can estimate the sampling distribution by looking at only a random sample of the possible groupings
Ideally, it is **good to get 10,000 groupings** (or shuffles), but often 1,000 is enough.
Steps:
1. Shuffle (or re-order) the group variable
2. Calculate the statistic based on the new groupings
3. Repeat many times to estimate the sampling distribution of the statistic under H0
4. Estimating the sampling distribution of the test statistic under Ho for the full study data (n=30)

Set simulation values

Setup (last week)
```
set.seed(101)
```

```
repetitions <- 1000
simulated_differences <- rep(NA, times=repetitions)
Calculate the test statistic
test_stat <- data %>% data: original data
group_by(group) %>% One row for each group, mean for each group
summarise(means = mean(score)) %>%
summarise(diff(means)) %>%
as.numeric() Turn the data frame into one number
test_stat
## [1] 7.64

Shuffle the group labels many times and calculate the new test statistic each time
for(i in 1:repetitions)
{
 value <- data %>%
mutate(group=sample(group)) %>% Shuffle the group
group_by(group) %>% Based on shuffled group
summarise(means = mean(score)) %>% Mean based on  new group
summarise(diff(means)) %>%
 as.numeric()
  simulated_differences[i] <- value Save value into ith slot of results vector
}
                                    digits
round(simulated_differences, 1)[1:80] Print values from 1 to 80
```

Estimating the sampling distribution of the **test statistic** under Ho for the full study data (n=30)

Diff in mean scores for dolphin/control groups

```
Plot the histogram
data_frame(simulated_differences) %>%
ggplot(aes(x=simulated_differences)) +
geom_histogram(bins = 15, color="black", fill="gray")
```

Evaluate the evidence against Ho - p-value

The p-value is the proportion of observations in the estimated sampling distribution of the statistic under Ho that are more extreme than our observed test statistic, **7.6** Based on full sample (n=30)

```
data_frame(simulated_differences) %>%
ggplot(aes(x=simulated_differences)) +
geom_histogram(bins = 15, color="black", fill="gray") +

data_frame(simulated_differences) %>%  simulated_differences: the vector
filter(simulated_differences >= abs(test_stat)  |
        simulated_differences <= -abs(test_stat)) %>% Keeps only simulated values a t least as extreme as obs. test stat(7.6)
summarise(pvalue = n() / repetitions) Calculate proportion of simulated values that are kept
*p-value = 0.083
```

5. Make a conclusion

Recall:

- A large P-value means the data are consistent with the null hypothesis.
- A small P-value means the data are inconsistent with the null hypothesis.

If treatment (dolphins vs control/snorkeling) has no effect on improving depression scores, the chance of seeing a difference at least as large as what we observed is only 0.003. In other words, these data provide strong evidence that the mean change in depression scores is different for individuals in the "dolphin" and "control" groups.

5. Make a conclusion based on a significance level α

Sometimes, you'll see some conclusions talking about **statistical significance** (or **a statistically significance difference**)

- A significance level (α) set in advance determines the cut-off for how unusual/extreme the test statistic has to be (assuming is true) in order to reject the assumption that is true (i.e. to conclude statistical significance)
- can be chosen to be any number between **(0, 1)**, but typically **α = 0.05**
- RULE: Reject H_0 if p-value<= α
- It is better to report the **p-value** and **comment** on the strength of evidence against instead of only reporting whether the result is/isn't statistically significant

Suppose we had decided to use a 1% significance level to conduct this test. What conclusion would we make (recall that our p-value was 0.003)?

Reject Ho because the pvalue (0.003) is smaller than 0.01

	One-sample test of a proportion	Test to compare the value of a parameter across two groups (could be mean, median, proportion, sd, ...)
1. State hypotheses	$H_0: p = p_0$ vs $H_A: p \neq p_0$	$H_0: \theta_1 = \theta_2$ vs $H_A: \theta_1 \neq \theta_2$ where θ_k is the value of the parameter (mean/median/proportion/sd/...) in group k
2. Compute test statistic	$\hat{p} = \frac{\#\ with\ characteristic}{n}$	$\hat{\theta}_1 - \hat{\theta}_2$ where $\hat{\theta}_k$ is the value of the parameter (mean / median / proportion / sd...) in the sample from group k
	Estimate the sampling distribution of $\hat{p}$: • Flip a coin n times repeatedly and compute $\hat{p}$ each time (if $p_0 = 0.5$)	*Estimate the sampling distribution of $\hat{\theta}_1 - \hat{\theta}_2$* • Consider n playing cards that match the number of observations in each group (say n_1 red cards and n_2 black

3. Simulate test statistic under H_0	• Simulate taking many samples from the population where $p = p_0$ by repeatedly using sample() to select n elements with replacement from the appropriate vector, based on appropriate probabilities (from the null hypothesis), and computing $\hat{p}$ each time	cards). Randomly distribute one card to each observation and compute $\hat{\theta}_1 - \hat{\theta}_2$ based on the new groups (red and black). Re-shuffle and redistribute the cards many times. • Simulate many random re-arrangements of the group labels across observations by repeatedly using sample() to shuffle group labels and define new groups, and then compute $\hat{\theta}_1 - \hat{\theta}_2$ each time
4. Assess evidence against H_0	Estimate p-value as the proportion of the statistic values simulated under H_0 which are at least as far away from p_0 as $\hat{p}$ is.	Estimate p-value as the proportion of the statistic values simulated under H_0 which are at least as far away from 0 as $\hat{\theta}_1 - \hat{\theta}_2$ is.
5. Make a conclusion	Reject H_0 if $pvalue \leq \alpha$	

P-value	Evidence
p-value > 0.10	no evidence against H_0
0.05 < p-value < 0.10	weak evidence against H_0
0.01 < p-value < 0.05	moderate evidence against H_0
0.001 < p-value < 0.01	strong evidence against H_0
p-value < 0.001	very strong evidence against H_0

Other statistics we could use: Mean, medians, proportion improved

Recall: Positive scores indicate that depression scores improved during the study
Research question: Is the proportion of improvement different for patients exposed to dolphins and patients who are not exposed to dolphins?
Ho: Pd - Pc = 0
Ha: Pd - Pc ≠ 0
Where Pd is the **proportion** who improve after swimming w. dolphins
Pc after snorkeling

```
data <- data %>%
mutate(status = ifelse(score > 0,
        yes="Improved",
         no="Didn't improve"))
data %>% group_by(group) %>%
summarise(n=n(),
     n_improve = sum(status=="Improved"),
      prop_improve = n_improve / n)
```

```
## # A tibble: 2 x 4
##   group       n n_improve prop_improve
##   <fct>   <int>    <int>       <dbl>
## 1 control    15        6       0.4
## 2 dolphins   15       10       0.667
```

What would be a useful test statistic for this hypothesis test?
Pd - Pc = 0.667 - 0.4 = 0.267

```
prop_data <- data %>%
group_by(group) %>%
summarise(n_improve = sum(status=="Improved"),
         n=n(),
        prop_improve = n_improve / n) The only diff from previous slides
test_stat <- prop_data %>%
 summarise(test_stat = diff(prop_improve)) %>%  to calculate diff in proportion across 2 rows
as.numeric()
We calculate proportion of improved instead of mean score for each group
test_stat
## [1] 0.2666667

set.seed(151) repetitions <- 1000; Set up simulation
simulated_values <- rep(NA, repetitions);

Simulate values under the null hypothesis by shuffling group labels
for(i in 1:repetitions){
simdata <- data %>%
   mutate(group = sample(group)) %>%
   group_by(group) %>%
   summarise(n=n(),
         n_improve = sum(status=="Improved"),
         prop_improve = n_improve / n)
sim_prop_diff <- simdata %>% summarise(value = diff(prop_improve))
simulated_values[i] <- as.numeric(sim_prop_diff)
}
sim <- data_frame(prop_diff = simulated_values)

Looks like p-value will be pretty large
ggplot(sim, aes(x=prop_diff)) +
geom_histogram(binwidth=0.15, fill="gray", color="black") +
labs(x = "Difference in proportion of improved patients for dolphin
     and control groups, assuming no difference between groups")
```

```
ggplot(sim, aes(x=prop_diff)) +
geom_histogram(binwidth=0.15, fill="gray", color="black") +
geom_vline(xintercept=c(test_stat, -test_stat), color="red") +
labs(x = "Difference in proportion of improved patients for dolphin
     and control groups, assuming no difference between groups")
```

Calculate the p-value

```
sim %>% filter(prop_diff >= abs(test_stat) | prop_diff <= -abs(test_stat)) %>%
 summarise(pvalue = n() / repetitions)
```

Based on this p-value is there evidence against the null hypothesis?
No coz the pvalue is larger than 0.1
Can we conclude that the null hypothesis is true or false?
No, we can never be certain

5.
If the treatment (dolphin vs control/snorkeling) has no effect on the proportion of patients who improve, the chance of observing a difference in proportions as large or larger than 26.6% is 0.261. In other words, these data provide **no** evidence that the proportion of individuals who improve is different in the dolphin and snorkeling groups.

The logic of hypothesis testing: Courtroom analogy

- Assumption of innocence

Type 1 and Type 2 errors

We use hypothesis tests to make conclusions about reality based on data.
Since we are making conclusions based on things that vary (i.e. are uncertain), there is potential for our conclusions to be incorrect.

Type I error: Reject H_0 when is H_0 true

Even when we set to be very small (i.e. need a very extreme/unusual observed test statistic to reject H_0, we could still observe a very unusual outcome and end up rejecting $H_0$when we should not)

Type II error: Do not reject H0 when H0 is false (and should be rejected!)
When we don't reject a null hypothesis (i.e. the results don't look unusual compared to the sampling distribution assuming H_0 is true), it is still possible that H_0 may not be true