Week 9 Live Session

w203 Instructional Team

Common hypothesis testing errors

For each of the following scenarios, explain the key error in the statistical procedure.

a) Bill hypothesizes that the average student drinks between 100 and 200 grams of caffeine during a take-home lab. He measures mean caffeine intake for a random sample of 50 lab-takers, then computes the p-value associated with his hypothesis.

In the frequentist framework, the null hypothesis can't be a range like 100-200. The null has to be specific enough to predict the sampling distribution of our test statistic. In this case, an average of 100 grams would imply one distribution, an average of 200 would imply another distribution, and all numbers in between would imply other distributions. There is no way to arrive at a sinble p-value in this situation.

b) Mike likes peanuts. Mike likes peanuts so much that he conducts a study to show how peanut allergies are an NIH sponsored hoax. He recruits 20 toddlers and randomly assigns each into two groups: peanut butter and brown sugar paste. To Mike's delight, he fails to find evidence for a difference between the groups ($p = .34$). Mike concludes by accepting the null hypothesis (that peanut allergies do not exist).

Never accept the null hypothesis! The only valid decisions are reject and fail-to-reject. The type 2 error rate is not being controlled here, so it is incorrect to interpret a high p-value as support for the null.

c) Anne replicates Mike's study and estimates a p-value of .03, she concludes that the alternative hypothesis has a 97% chance of being true.

The p-value is NOT the chance that the null hypothesis is false. This is a common misconception, but the p-value is only the chance of getting a statistic as extreme as that observed assuming the null is true. A statement about the altenate hypothesis being true would be a Bayesian statement, and the probability would be a subjective probability. Anne is therefore incorrectly placing a Bayesian meaning on a Frequentist framework.

d) Tim asks 50 passengers on the 8am Staten Island Ferry to complete his survey about attitudes toward atheists. He finds a statistically significant difference between attitudes toward atheists and attitudes toward scientologists ($p = .04$). Huzzah! Tim concludes that the US public is more fearful of atheists than scientologists.

Tim doesn't have a random sample from the US public. The staten island ferry probably contains a very distinctive subset of the US population.

Comparing Means

The file united_states_senate_2014_v2.csv contains data on the 100 members of the US senate that served in 2014. We will consider this group to be a sample (for example, from some generative process that creates senators).

```r
S = read.csv("united_states_senate_2014_v2.csv")
summary(S)
```

```
##           Senator.Names    Gender          State             Party
##  Alan Franken    : 1    Female:20   Alabama   : 2   Democrat   :53
##  Amy Klobuchar   : 1    Male  :80   Alaska    : 2   Independent: 2
##  Angus King      : 1                 Arizona   : 2   Republican :45
##  Barbara Boxer   : 1                 Arkansas  : 2
##  Barbara Mikulski: 1                 California: 2
##  Benjamin Cardin : 1                 Colorado  : 2
##  (Other)         :94                 (Other)   :88
##             Religion  Campaign.Money.Raised..millions.of...
##  Protestant     :49   Min.   : 0.100
##  Catholic       :27   1st Qu.: 4.575
##  Jewish         :10   Median : 7.550
##  Other Christian: 7   Mean   : 9.645
##  Mormon         : 2   3rd Qu.:13.800
##  Unaffiliated   : 2   Max.   :44.200
##  (Other)        : 3
##  Campaign.Money.Spent..millions.of...   NRA.Rating
##  Min.   : 0.200                       A      :34
##  1st Qu.: 2.975                       F      :34
##  Median : 6.000                       A+     : 9
##  Mean   : 8.227                              : 5
##  3rd Qu.:12.225                       AQ     : 5
##  Max.   :43.400                       C      : 3
##                                       (Other):10
```

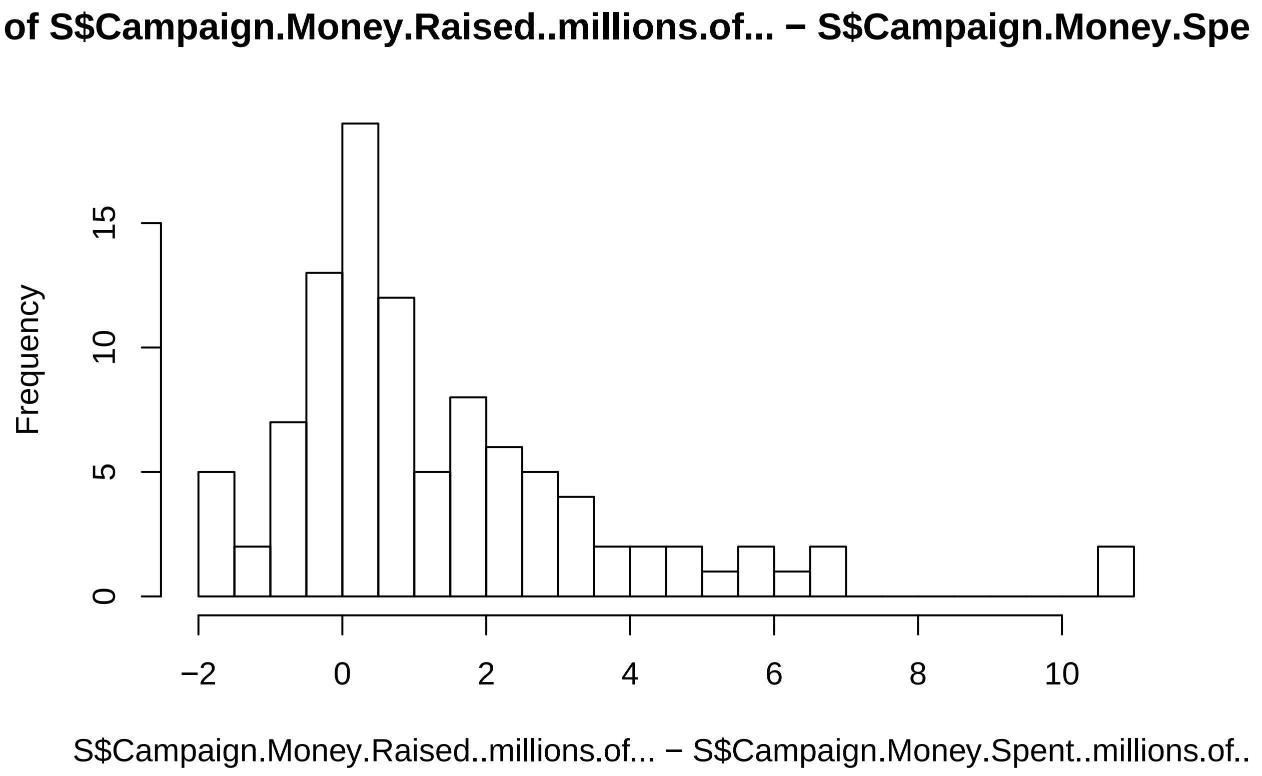
You will be placed in a breakout room and assigned a single question to investigate using this dataset. Each question involves a comparison of means.

In your breakout rooms, examine the data and decide what type of test is most appropriate. You may select a paired or an unpaired test. You may also select a parametric or a nonparametric test. Conduct your test and interpret your results.

Room 1: Is there a difference between the amount of money a senator raises and the amount spent?

There is a clear pairing in this case - each senator corresponds to one measurment of money raised and one of money spent. We have metric variables, and a large sample, so we can use the CLT to justify a parametric test. Just in case, we examine a histogram (of the difference variable) to see if there's an unusual departure from normality, especially an unusual skew. There seems to be no cause for alarm here, so we proceed with the parametric test.

```
hist(S$Campaign.Money.Raised..millions.of... - S$Campaign.Money.Spent..millions.of..., breaks=20)
```



```r
t.test(S$Campaign.Money.Raised..millions.of..., S$Campaign.Money.Spent..millions.of..., paired=T)
```

```
##
##  Paired t-test
##
## data:  S$Campaign.Money.Raised..millions.of... and S$Campaign.Money.Spent..millions.of...
## t = 5.9944, df = 99, p-value = 3.329e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9486232 1.8873768
## sample estimates:
## mean of the differences
##                   1.418
```

note that we would get the same result by running a one-sample t-test on the difference:

```r
t.test(S$Campaign.Money.Raised..millions.of... - S$Campaign.Money.Spent..millions.of...)
```

```
##
##  One Sample t-test
##
## data:  S$Campaign.Money.Raised..millions.of... - S$Campaign.Money.Spent..millions.of...
## t = 5.9944, df = 99, p-value = 3.329e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.9486232 1.8873768
## sample estimates:
## mean of x
##     1.418
```

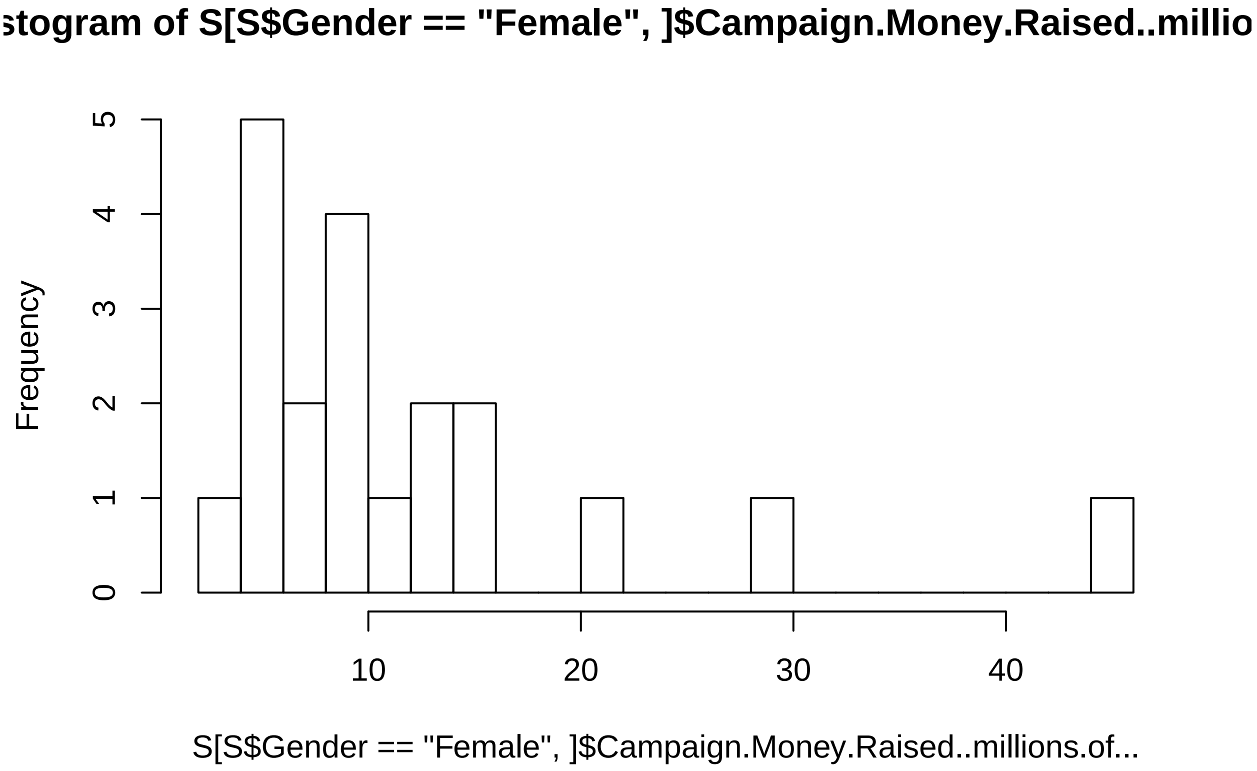
Room 2: Do female Democratic senators raise more or less money than female Republican senators?

There are not many observations here (20), so the CLT does not apply. We check the histogram of the variable and find a moderate skew. The t-test is robust to moderate deviations from normality, so you could argue that it is still valid here. However, since this is a borderline case, we will be more conservative and run a rank test.

```
summary(S$Gender)
```

```
## Female   Male
##     20     80
```

```
hist(S[S$Gender=="Female",]$Campaign.Money.Raised..millions.of..., breaks=20)
```



```
wilcox.test(Campaign.Money.Raised..millions.of... ~ Party, data = S[S$Gender=="Female",])
```

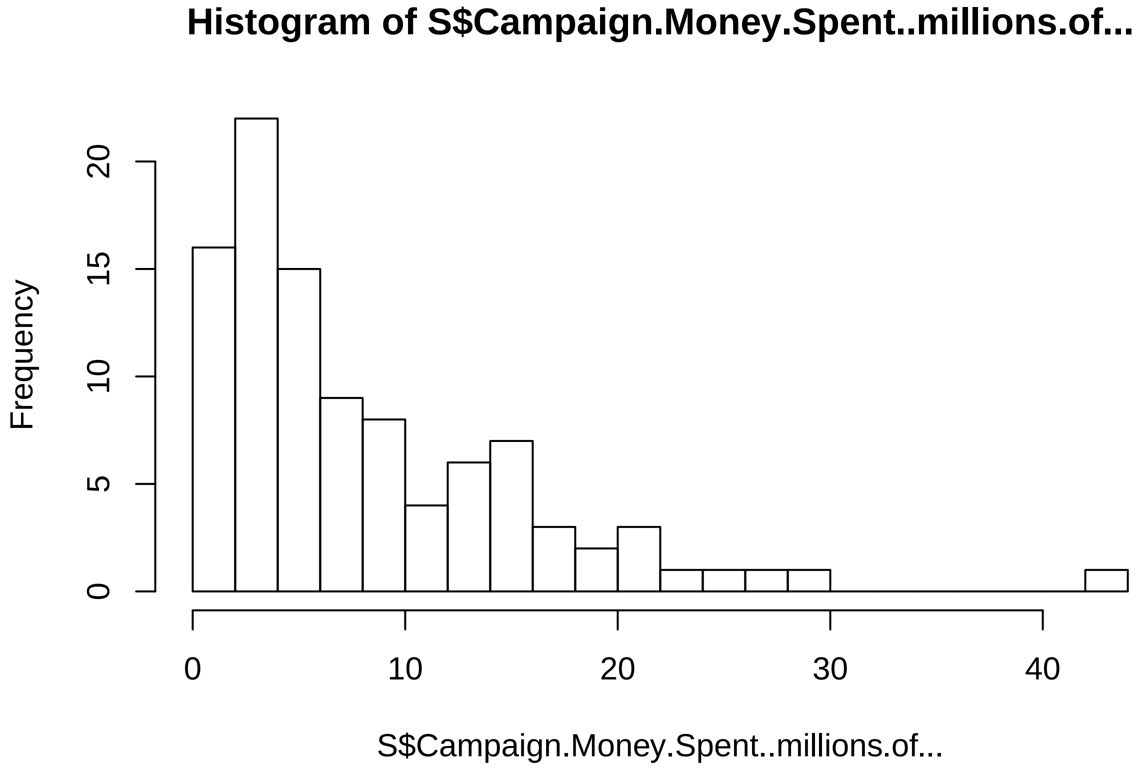
```
## Warning in wilcox.test.default(x = c(15.3, 13.8, 11.7, 9.7, 29.7, 9.9,
## 6.2, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Campaign.Money.Raised..millions.of... by Party
## W = 58, p-value = 0.01593
## alternative hypothesis: true location shift is not equal to 0
```

Room 3: Do protestant Senators spend more or less money than non-protestant senators?

This is a pretty straightforward problem. We have a metric variable and a large sample size, so we can invoke the CLT. There is no pairing we can use - in fact, the two groups have different sizes, so there would be no wait to pair. Just in case, we check the histogram to make sure there isn't a severe skew that would make us worry whether the CLT is working. Everything looks ok, so we proceed with an unpaired parametric test.

```
hist(S$Campaign.Money.Spent..millions.of..., breaks=20)
```



```
table(S$Religion)
```

```
##
##       Buddhism        Catholic       Christian          Jewish
##              1              27               1              10
##         Mormon Other Christian      Protestant    Unaffiliated
##              2               7              49               2
##    Unspecified
##              1
```

```
t.test(S$Campaign.Money.Spent..millions.of... ~ S$Religion == "Protestant")
```

```
##
##  Welch Two Sample t-test
##
## data:  S$Campaign.Money.Spent..millions.of... by S$Religion == "Protestant"
## t = 1.2856, df = 97.177, p-value = 0.2016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.038329  4.857697
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            9.162745            7.253061
```

Room 4: Does the NRA prefer male senators or female senators?

The NRA grade is an ordinal variable. The difference between A and B may not be the same as the difference between B and C. This means, for example, that it wouldn't be valid to average an A and a C, and call the result equivalent to a B. For this reason, we will run a non-parametric test. Note that a large sample does not fix the problem with the non-metric scale. The CLT addresses issues of non-normality, but it doesn't fix the fact that the intervals may not be equal.

Room 5: Choose your own question to investigate.

Demonstration of Confidence Intervals

The following exercise is meant to demonstrate what the confidence level in a confidence interval represents. For this exercise, we will assume a standard normal population distribution and simulate what happens when we draw a sample and compute a confidence interval.

Your task is to complete the following function so that it,

1) simulates and stores n draws from a standard normal distribution
2) based on those draws, computes a valid confidence interval with confidence level α.

Your function should return a vector of length 2, containing the lower bound and upper bound of the confidence interval.

```
sim_conf_int = function(n, alpha) {
  # Your code to
  # 1. simulate n draws from a standard normal dist.
  # 2. compute a confidence interval with confidence level alpha
 # return(c(-1,1))  # replace with the interval you compute.

  v = rnorm(n)
  low = mean(v) - qt(1-alpha/2, n-1)* sd(v)/sqrt(n)
  high = mean(v) + qt(1-alpha/2, n-1)* sd(v)/sqrt(n)
  return(c(low,high))
}
```

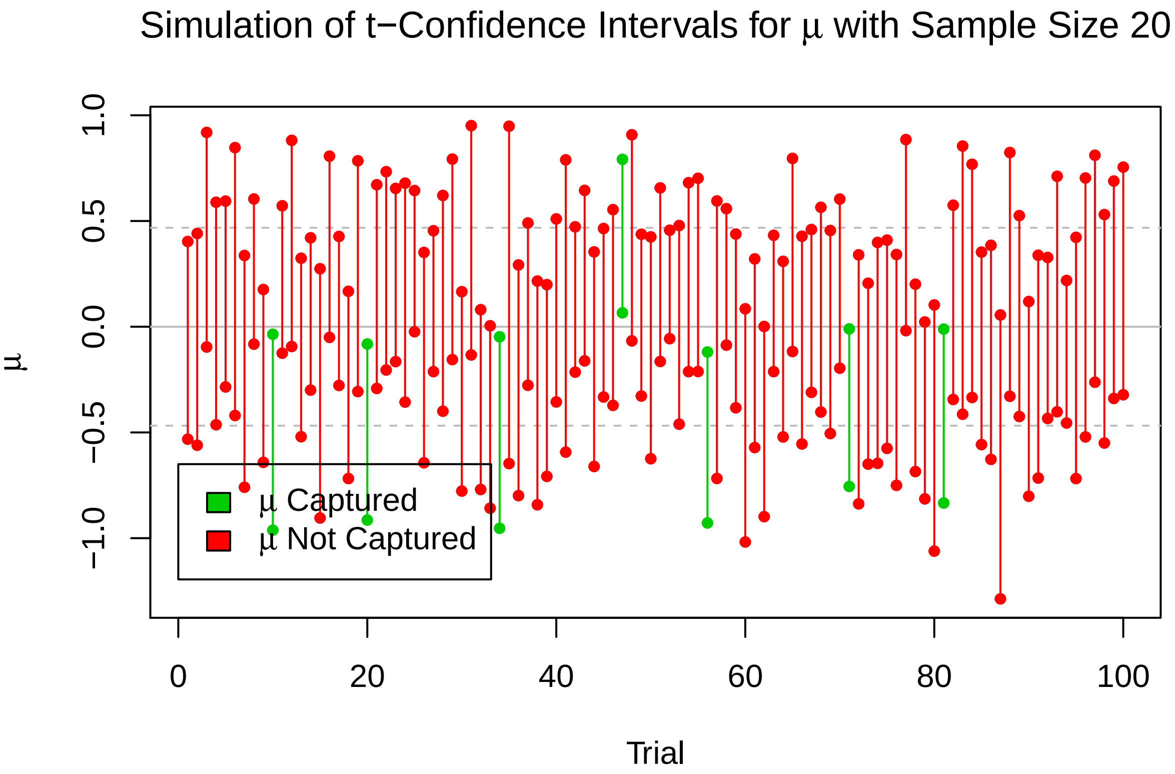
When your function is complete, you can use the following code to run your function 100 times and plot the results.

```
many_conf_int = function(m, n, alpha) {
  results = NULL
  for(i in 1:m) {
    interval = sim_conf_int(n, alpha)
    results = rbind(results, c(interval[1], interval[2], interval[1]<0 & interval[2]>0))
  }
  resultsdf = data.frame(results)
  names(resultsdf) = c("low", "high", "captured")
  return(resultsdf)
}

n = 20
cints = many_conf_int(100, n, .05)

plot(NULL, type="n",xlim=c(1,100),ylim=c(min(cints$low), max(cints$high)), xlab="Trial",ylab=expression
abline(h = c(0, qt(0.975, n-1)/sqrt(n), qt(0.025, n-1)/sqrt(n)), lty = c(1,2,2), col = "gray")
points(cints$high, col = 3-cints$captured, pch = 20)
points(cints$low, col = 3-cints$captured, pch = 20)
for(i in 1:100)
   {
     lines(c(i,i), c(cints$low[i],cints$high[i]), col = 3-cints$captured[i], pch = 19)
     }
title(expression(paste("Simulation of t-Confidence Intervals for ", mu,
                          " with Sample Size 20")))
```

```
legend(0,-.65, legend = c(expression(paste(mu," Captured")),
                          expression(paste(mu," Not Captured"))), fill = c(3,2))
```


Simulation of t−Confidence Intervals for μ with Sample Size 20
μ
Trial
🟩 μ Captured
🟥 μ Not Captured