Week 9 Live Session

w203 Instructional Team

Announcements

Feedback about Quiz 1.

Remember that Lab 3 is being released this week.

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.

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).

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.

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.

Comparing Means

The file united_states_senate_2014.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.csv")
summary(S)
```

```
##             Senator.Names    Gender          State             Party
##  Alan "Al" 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...
##  Min.   : 0.200
##  1st Qu.: 2.975
##  Median : 6.000
##  Mean   : 8.227
##  3rd Qu.:12.225
##  Max.   :43.400
##
```

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?

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

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

Room 4: 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.
}
```

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)
```

```r
  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$high, col = 2+cints$captured, pch = 20)
points(cints$low, col = 2+cints$captured, pch = 20)
for(i in 1:100)
   {
     lines(c(i,i), c(cints$low[i],cints$high[i]), col = 2+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))
```