Extra practice problems

Study design, data types, and descriptive statistics [L1, L2, L3]

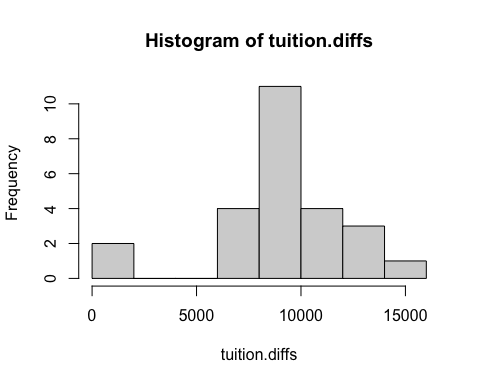
STAT218

### Test information

### Practice problems

1. The tuition dataset contains in-state and out-of-state tuition at a random sample of 25 public universities from 2011-2012.
   1. [L3] Visualize the distribution of differences between in-state and out-of-state tuition. Comment on whether the assumptions for inference using the model seem appropriate.
   2. [L4] Calculate and interpret a 95% confidence interval for the mean difference between in-state and out-of-state tuition.
   3. [L4] Interpret your interval in context following the style introduced in class.

# load data  
load('data/tuition.RData')  
  
# part a: visualize distribution of differences; are assumptions for use of t model met?  
tuition.diffs <- tuition$out.of.state - tuition$in.state  
hist(tuition.diffs, breaks = 10)



# part b: 95% interval estimate for differences  
t.test(tuition.diffs)$conf.int

[1] 7687.079 10389.721  
attr(,"conf.level")  
[1] 0.95

1. [challenge problem] Cancer rates and sunspot activity. The dataset cancer contains skin cancer rates per 100,000 people in Connecticut each year from 1938 to 1972. Each year is also classified as following a period of higher than average or lower than average sunspot activity. The delta variable is the change in cancer rate relative to the previous year. In this problem, you’ll perform inference on the mean delta by sunspot activity level to determine whether higher than average sunspot activity is associated with an increase in mean skin cancer rates from the prior year.
   1. [L4] Estimate the mean delta (irrespective of sunspot activity level). Provide both a point estimate and standard error, and interpret the estimate in context. Does the estimate suggest that the cancer rate is increasing or decreasing? Explain.
   2. [L4, L5] Perform a test for mean delta to determine whether the mean cancer rate is increasing. Use a 5% significance level, and report your test result together with an interval estimate following the narrative style introduced in class.
   3. [L3] Plot the ‘raw’ cancer rate (*i.e.*, not the delta) against year. (Add the argument type = 'b' to draw a path connecting the observations.) Is your answer in (b) consistent with any trend(s) you see?
   4. [L3] Make a side-by-side boxplot of the delta variable for each level of sunspot activity. Comment on the plot: does there seem to be a difference?
   5. [L5] Test whether the mean change in cancer rate is higher in years with higher than average sunspot activity. Carry out inference at the 5% significance level.

# load data  
load('data/cancer.RData')  
  
# part a: point estimate of mean delta and standard error  
mean(cancer$delta)

[1] 0.1176471

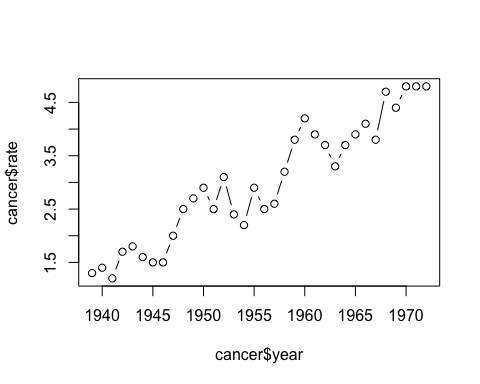
sd(cancer$delta)/sqrt(length(cancer))

[1] 0.1944186

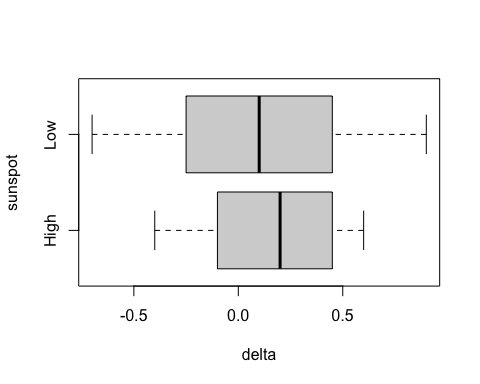
# part b: is delta increasing? test at 5% level  
t.test(cancer$delta, mu = 0, alternative = 'greater', conf.level = 0.95)

One Sample t-test  
  
data: cancer$delta  
t = 1.7642, df = 33, p-value = 0.04348  
alternative hypothesis: true mean is greater than 0  
95 percent confidence interval:  
 0.004791988 Inf  
sample estimates:  
mean of x   
0.1176471

# part c: plot   
plot(cancer$year, cancer$rate, type = 'b')



# part d: boxplots by activity level; different?  
boxplot(delta ~ sunspot, data = cancer, horizontal = T)



# part e: test for a difference in mean delta by sunspot activity at the 5% level  
t.test(delta ~ sunspot, alternative = 'greater', data = cancer, conf.level = 0.95)

Welch Two Sample t-test  
  
data: delta by sunspot  
t = 0.57547, df = 31.997, p-value = 0.2845  
alternative hypothesis: true difference in means between group High and group Low is greater than 0  
95 percent confidence interval:  
 -0.1472952 Inf  
sample estimates:  
mean in group High mean in group Low   
 0.16000000 0.08421053

1. [L5] Physiological indicators of schizophrenia. The dataset Sleuth3::case0202 contains data on volumes of the left hippocampus in cubic centimeters for 15 pairs of monozygotic twins; one twin in each pair was schizophrenic and the other was not.

Tuberculuosis and survival in guinea pigs. The dataset tubercle contains survival times in days of guinea pigs that were randomly assigned to receive a dose of tubercle bacilli, the bacterial pathogen that causes tuberculosis, or to a control group.