Test 2 practice problems

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

STAT218

### Test 2 information

The test will comprise four problems focused on point estimation, interval estimation, and hypothesis tests for means (learning outcomes L4-L5). Each problem will have multiple parts, some of which may require skills from earlier (especially summary statistics and statistical graphics).

You will have 48 hours to complete the test; a Posit cloud project will be provided with comment outlines to help you organize your calculations. You’ll submit your work via an online form, and will be expected to also upload your R script from your Posit cloud project.

The problems below are intended to help you practice the skills and concepts that will be assessed in the test. An expandable “solution” is provided below each prompt that shows the calculations needed to answer the prompts; of course, resolving the problems satisfactorily also requires interpreting results accurately. You’re encouraged to ask about interpretations in class.

### Practice problems

1. [L3, L4] 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. [L3, L4, L5] 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  
cancer <- read\_csv('data/cancer.csv')  
  
# part a: point estimate of mean delta and standard error  
mean(cancer$delta)  
sd(cancer$delta)/sqrt(length(cancer))  
  
# part b: is delta increasing? test at 5% level  
t.test(cancer$delta, mu = 0, alternative = 'greater', conf.level = 0.95)  
  
# 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)

1. [L4, L5] Studies have provided evidence that the hippocampus is smaller in schizophrenic patients on average. The dataset hippocampus contains data on volumes of the left hippocampus in cubic centimeters for 15 pairs of monozygotic twins; one twin in each pair was affected by schizophrenia and the other was not.
   1. [L5] Compute the pairwise differences in hippocampal volume by twin pair and inspect the distribution. Do assumptions for inference using the model seem plausible?
   2. [L5] Formulate a hypothesis to test whether hippocampal volume is smaller among the affected twin on average. Write the hypotheses in notation.
   3. [L4, L5] Carry out the test in (b) at the 1% significance level. Report your test result along with a corresponding interval estimate following the narrative style introduced in class.

# load and inspect data  
twins <- read\_csv('data/hippocampus.csv')  
head(twins)  
  
# part a: compute differences; check distribution for t inference assumptions  
hvolume.diff <- twins$affected - twins$unaffected  
hist(hvolume.diff, breaks = 5)  
  
# part c: carry out test at 1% significance level  
t.test(hvolume.diff, mu = 0, alternative = 'less', conf.level = 0.99)