Assignment 3 – Intro Inference

2181MATH4990 Mateen Shaikh

Due: Tues Feb 25 by 3:30pm

Instructions:

The assignment should be formatted as explained on the course outline. Only include graphs when asked or as an integral part of your answer. Code must be separated and submitted as a plain text file (.txt) containing no more than 100 functional lines of code used to answer questions. Answers should contain only relevant, concise information and will be penalized otherwise.

Versicolor & Virginica Irises

Tests

1. Report the usual 95% confidence interval from `t-test` for true average iris sepal width of Veriscolor. Write one or two sentences to explain this in the context of flowers to someone who doesn't understand what a confidence interval is.
2. The output gives the p-value result of a hypothesis test as well. If it's sensible, explain and interpret the result of the test. If it's not sensible, explain why.
3. Conduct the usual 2-sided t-test that the average iris sepal width of Veriscolor is the same as that of Virginica from the iris data set. Describe what the null hypothesis/null distribution is, and the the corresponding p-value. Also provide one or two sentences to explain this to someone who doesn't understand hypothesis testing.
4. Repeat the above through a permutation test and report on how different your 2-sided p-value is. Show the distribution test statistics under the null and comment on whether it is approximately normal, and hence, whether the t-test would have given a good approximation.
5. For all measurements (Sepal Width, Sepal Length, Petal Width, Petal Length) conduct the one-way ANOVA and conduct the t-test that the measurements between versicolor and virginica are the same. Provide one or two sentences to to explain the result of the ANOVA to someone who does not understand it.

CI

The parametric approach to CI can be shown to be problematic for a variety of reasons. Consider, for example, consider the data

```
x=rep(0,10); x[1]=1
```

6. Use the `t.test` function to estimate the confidence interval for the true mean (a proportion, since the data is binary). Report on this confidence interval and why this is problematic for estimation of the true mean.
7. Bootstrap 1000 times and report the 95% confidence interval. Comment on whether this range for a confidence interval is more appropriate.

Estimating type I error

8. Simulate samples of size 10 from the exponential distribution with mean 1 using `rexp(5,rate=1)`. Extract the p-value from the standard t-test using `t.test(x,mu=1)$p.value` where `x` is your sample. Replicate this at least 10000 times to illustrate the distribution of p-values under the null. Comment on whether this looks uniform for particularly small p-values (say, less than 0.1).

Theoretical

Consistency & Bias

9. The following estimator for sample variance from a sample of size n

   $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{\sum\limits_{i=1}^{n}\left(X_i - \frac{1}{n}\sum\limits_{j=1}^{n} X_j\right)^2}{n};$$

   is biased, but consistent. Show that the bias of this estimator is $\text{bias}(S^2) = \dfrac{-\sigma^2}{n-1}$ where σ^2 is the real variance of the population and S^2 is the estimator (how we estimate it). You may complete this theoretically or illustrate empirically (but not both):

   If you choose to Prove Theoretically
   You may show this theoretically with formulas (should be several lines of math). It will help to remember that $E(X_iX_j) = Cov(X_i, X_j) + E(X_i)E(X_j)$ and that X_i and X_j are statistically independent of one another for $i \neq j$. State your appropriate justification in the margin, as with any proof.

If you choose to illustrate empirically
You may empirically illustrate the bias with a small simulation with varying values of sample size, n. You should visually convey some sense of centre (average/median) and some sense of variability of your estimates (e.g., quartiles, confidence intervals, mean ± standard error, etc.).

Also make sure to overlay the theoretical bias for every value of n in your visualization. An example of pseudocode for such a simulation is shown below. However, you may simulate differently as long as you convincingly convey the bias.

```
numSimulationsPern = 1000 # or any reasonably large number
allValuesOfn = 3:30 #good enough range
theoreticalVariance = 25 #or any consistent number you want
#######################################################
# function to simulate data and calculate the bias 'numSimulations' times
# @param n the sample size to repeatedly simulate
# @return the biases as a vector
s.squared.for.n=function(n){
 # simulate a sample of size n with 'theoreticalVariance', 'numSimulationsPern' times
 # each time, estimate the variance as shown on the assignment
 # store the difference estimated - theoreticalVariance in vector to return
}
#######################################################
allBiases = lapply(valuesOfn,s.squared.for.n);
#returns the results as list of vectors
#

#calculate a 3-number summary for allValuesOfn:
plot the median biases for allValuesOfn, connected by a line
add points for the first quartile for allValuesOfn, connected by a line
add points for the third quartile for allValuesOfn, connected by a line
add points that correspond to the theoretical bias, connected by a line
```

10. Assuming the bias provided above is correct, explain why this estimator is consistent. This should be a one-sentence explanation.

Bootstrap

11. When constructing a bootstrapped sample of size n, from an original sample also of size n, determine the probability that a particular datum will not be included in the bootstrapped sample as n becomes large. The probability may be calculated exactly with calculus or estimated with a small simulation. If you simulate, remember that you do not need an actual data set.