Final Exam

Paul Gribble
Winter, 2019

due: April 25, 2019

LEGAL EAGLES

The file `lsat.csv` contains average law school admission test (LSAT) scores and grade point average (GPA) from 82 American law schools who participated in a large study of admission practices (Efron B, Tibshirani RJ (1994). An Introduction to the Bootstrap. CRC Press). You can load the data into R using:

```
fname <- "https://www.gribblelab.org/stats2019/data/lsat.csv"
lsat <- read.table(fname, sep=",", header=TRUE)
```

Question 1 (1 point)

Report the Pearson product-moment correlation coefficient r between LSAT and GPA.

Question 2 (2 points)

Use bootstrapping (n=10,000) to estimate the sampling distribution of r. Plot a histogram of the sampling distribution.

To perform the bootstrap, on each iteration, resample *pairs* of (LSAT,GPA) scores from the original set of 82 observations, with replacement, to generate a new set of 82 observations.

Question 3 (1 point)

Based on the bootstrap results, report the 95% confidence intervals on r. You can use the R function `quantile()` to obtain the lower and upper bounds of the confidence interval.

Question 4 (1 point)

Estimate a linear regression model that predicts GPA based on LSAT:

$$GPA_i = \beta_0 + \beta_1 LSAT_i + \epsilon_i \tag{1}$$

Report the estimated parameters β_0 and β_1. Report the residual standard error.

Question 5 (1 point)

Plot GPA (vertical axis) as a function of LSAT (horizontal axis). Overlay the model prediction as a solid red line.

Question 6 (2 points)

Use bootstrapping (n=10,000) to estimate the sampling distribution of β_0 and β_1. Plot a histogram of the sampling distribution of β_0. Do the same for β_1.

To perform the bootstrap, on each iteration, resample *pairs* of (LSAT,GPA) scores from the original set of 82 observations, with replacement, to generate a new set of 82 observations.

Question 7 (1 point)

Based on the bootstrap results, report the 95% confidence interval for β_0. Do the same for β_1.

DEBUGGING

Question 8 (1 point)

Load in the `InsectSpray` dataset into R like this:

```
attach(InsectSprays)
data(InsectSprays)
summary(InsectSprays)
with(InsectSprays, boxplot(count ~ spray))
```

The data gives a count of the number of insects found in fields after one of six different insecticides (labelled A,B,C,D,E,F) were used. Twelve observations are reported for each of 6 insecticides, for a total of 72 observations.

Filter the data so that only sprays "C", "D" and "E" are considered. You should end up with 36 observations.

Using the dataset of 36 observations, perform a one-way ANOVA. Report the F ratio for the main effect of insecticide.

Question 9 (2 points)

Perform a permutation test (n=10,000) to test the null hypothesis that all sprays result in the same number of insects found.

The observed F ratio F_{obs} will be the statistic of interest that you will use to perform the permutation test.

On each iteration of the permutation test, generate a new sample of 36 observations in which the "C", "D" and "E" labels are randomly shuffled. Perform a one-way ANOVA on the new sample. Compute the F ratio.

Plot a histogram showing the sampling distribution of F generated from the n=10,000 permutation tests.

Compute and report the probability p of obtaining a value of F as large or larger than F_{obs}.

Question 10 (3 points)

Use permutation tests (n=10,000) to perform follow-up tests comparing (1) spray C vs D, (2) spray C vs E and (3) spray D vs E. Don't forget to implement some kind of correction for type-I error accumulation due to multiple comparisons. I suggest bonferroni-holm since it's relatively simple. For each permutation test show a histogram of the sampling distribution of the difference-between-means, and report the p-value for the permutation test.