Math-146 - Exam #1 - Practice (S22)

General Information

- You will have 50 minutes to complete Exam #1
- You are not allowed to use anything on the exam besides a pencil/pen and paper
- The first page of the exam will be a formula sheet. It will be the exact same set of formulas given on the practice exam.

Exam topics

1) **Data basics** - cases, variables, variable types
2) **Univariate graphs and summaries** - frequencies/proportions, barcharts, dotplots/histograms, shape, center (mean/median), spread (standard deviation/range/IQR), and boxplots/five number summaries
3) **Bivariate graphs and summaries** - contingency tables, conditional proportions, stacked barcharts, side-by-side graphs, comparative summaries, scatterplots, correlation, regression, ecological fallacy, and extrapolation
4) **Sampling and confidence intervals** - sampling bias and sampling variability, standard deviation vs. standard error, sampling distributions, bootstrapping and bootstrap distributions, confidence intervals via the 2-SE and percentile methods, confidence intervals using SE formulas
5) **Study design and confounding variables** - observational studies vs. experiments, confounding variables, benefits of random assignment

Types of content to expect

- 1 question consisting of several true/false statements
- 1 question involving matching of definitions or methods
- 1 question resembling those on the homework (ie: a prompt with several parts)
- 1 question resembling those on in-class labs (ie: screenshots of StatKey output involving real data)

Formulas

Table of Statistical Notation:

	Population Parameter	Estimate (from sample)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Regression	β_0, β_1	b_0, b_1

Table of Standard Error Formulas:

Estimate	Standard Error	CLT Conditions
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	normal population or $n \geq 30$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1, 2\}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	normal populations or $n_1 \geq 30$ and $n_2 \geq 30$
r	$\sqrt{\frac{1-\rho^2}{n-2}}$	normal population (both vars) or $n > 30$

Question #1 (true/false)

For each statement (A - F), indicate whether the statement is true (T) or false (F):

A) The standard error of an estimate decreases when a larger sample size is used.
B) Sampling bias will decrease if a larger sample size is used.
C) For the same data, a 99% confidence interval estimate reflects a wider range of plausible values than a 90% confidence interval estimate.
D) The choice of explanatory and response variables matters for regression, but does not matter for correlation.
E) The purpose of random assignment in a randomized experiment is to eliminate the possibility of confounding variables influencing the experiment's outcome.
F) Convenience samples are always biased, so they can never be treated as representative of a population.

Question #2 (matching)

For each scenario (A - F), select the appropriate method from those described in 1 - 9. Not all of the options in 1 - 9 will be used.

A) Finding an association between two categorical variables
B) Describing the spread of a quantitative variable
C) Describing the strength of association between two quantitative variables
D) Finding an association between one categorical variable and one quantitative variable
E) Describing the central tendency of a quantitative variable that is skewed to the right
F) Expressing the variability in a point estimate

1) Median
2) 75th Percentile
3) Correlation coefficient
4) Contingency table
5) Standard error
6) Side-by-side boxplots
7) Standard deviation
8) Stratification
9) Simple random sampling

Question #3 (homework style)

A 2015 study published in the journal *SLEEP*, found that just one session of cognitive behavioral therapy can help people with insomnia. In the study, 40 people who suffered from insomnia were randomly divided into two groups of 20 each. One group participated in a one-hour therapy session, while the other did not receive any treatment. Three months later, 14 people in the therapy group reported improved sleep, but only 3 in the other group reported an improvement.

Part A: What are the cases in this study? Explain in 1-2 sentences.

Part B: What are the variables in this study? Name them and identify each as either categorical or quantitative.

Part C: If these data were laid out in spreadsheet form, with the cases as rows and the variables as columns, how many rows and columns would this spreadsheet contain?

Part D: Of the variables you described in Part B, identify the explanatory and response variable. Is there an association between these two variables? Justify your answer.

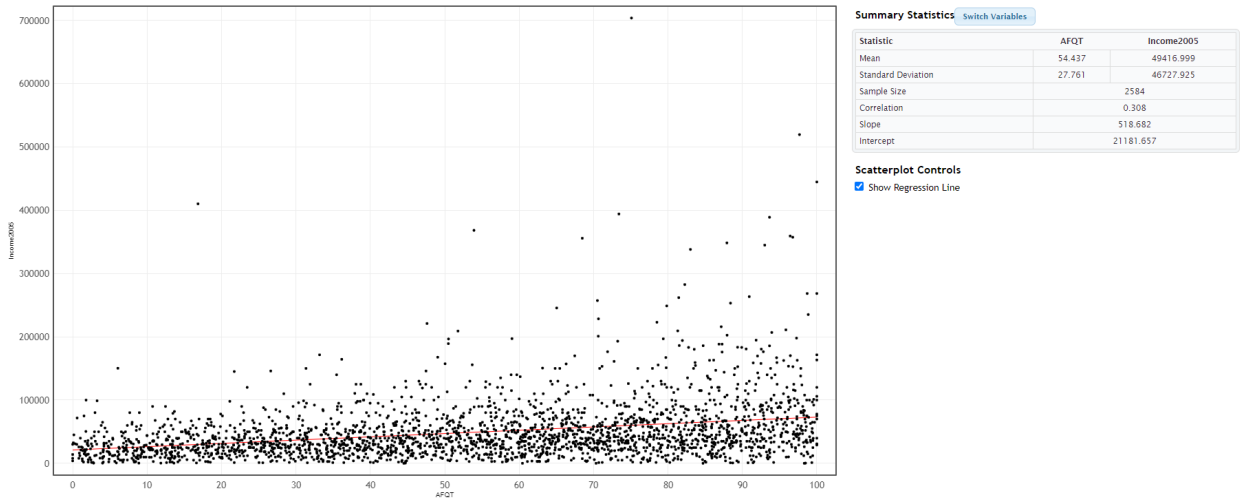
Part E: Briefly explain, describe, or sketch an appropriate graph that could be used to depict the relationship between the explanatory and response variable in this study.

Question #4 (lab style)

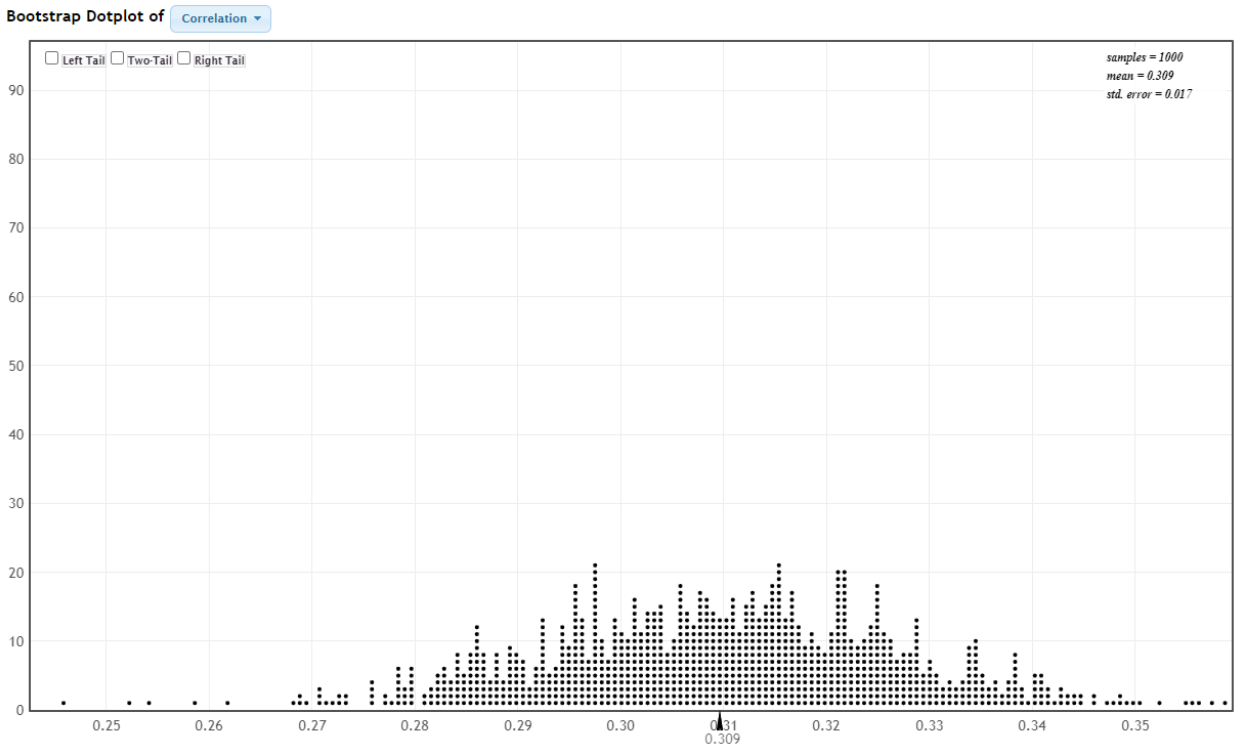
The Armed Forces Qualification Test (AFQT) is a multiple choice aptitude test designed to gauge whether an individual's reasoning and verbal skills are sufficient for military service. Because the test has been so broadly applied, its scores are reported as percentiles and used as a measure of a person's general intelligence.

The data summarized below are a randomly chosen sample of $n = 2584$ Americans who were first tested 1979 (when they were approximately 18 years old) and then re-interviewed in 2006 to obtain data on their educational attainment and annual income. You may assume that everyone who was randomly selected agreed to participate in the follow-up interview.

The graph below displays StatKey output relating AFQT percentile scores and income:



The graph below displays correlation coefficients relating AFQT score and income across 1,000 different bootstrapped samples:



Part A: Describe the relationship between AFQT score and income in these data. Be sure to comment on the form, direction, and strength of the relationship.

Part B: Use linear regression to predict the 2006 income of an individual whose AFQT percentile score is 50 (the median). Show your work.

Part C: Use the 2-SE method (ie: 95% rule) to find a 95% confidence interval estimate for the population-level correlation between AFQT score and income. Show your work.

Part D: Suppose your friend argues that because of sampling bias and sampling variability the correlation seen in the AFQT dataset is not an accurate reflection of the true relationship between scores on the Armed Forces Qualification Test and future income. Briefly explain why neither of these explanations are convincing for this dataset.