STAT 135, Final exam, Spring 2016, H. Pitters

NAME (IN CAPS):_____
SID number and SECTION:_____

There are 6 questions in the exam. Show your work or provide a brief explanation for all answers. This exam is closed books and closed notes. You are allowed to use a calculator and a cheat sheet. Answers should be simplified as much as possible.

Good luck!

1	2	3	4	5	6	Σ

Question 1. Bootstrap.[8]

(1) Briefly describe both the parametric bootstrap and the nonparametric bootstrap. What is the rationale behind the bootstrap and what is it potentially useful for? [3]

(2) There are two types of errors made when computing a sampling distribution via the bootstrap method. Explain these two errors. [3] Consider an estimator, $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ (were $X_1, \ldots, X_n$ denote the observations) for a parameter θ. Apart from computing an estimate for θ, one is usually also interested in computing the sampling error, i.e. the standard deviation of $\hat{\theta}$, or even the entire sampling distribution of $\hat{\theta}$. An analytic expression for the standard error (or sampling distribution) is only feasable in very specific settings. However, the bootstrap method allows to simulate the sampling distribution of $\hat{\theta}$ with the help of a computer.
Parametric bootstrap. Here the common distribution F_θ of $X_1, \ldots, X_n$ is assumed to be from a parametric family $\{F_\theta, \theta \in \Theta\}$. One has to estimate θ, say by the MLE $\hat{\theta}_{MLE}$. Then $F_{\theta_{MLE}}$ is taken to approximate F_θ, and this approximation introduces the first kind of error.
With this approximation, one
   (a) draws with replacement a sample $x_1^*, \ldots, x_n^*$ of size n from $F_{\hat{\theta}_{MLE}}$,
   (b) evaluates the estimator $\hat{\theta}_1^* := \hat{\theta}(x_1^*, \ldots, x_n^*)$,
   (c) and repeats this procedure B times to obtain $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. The empirical distribution of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ is used as an approximation for the distribution of $\hat{\theta}$. This is the second kind of error, which can be controlled by increasing the number B of bootstrap samples.

Nonparametric bootstrap. Here the true distribution F underlying $X_1, \ldots, X_n$ is approximated by the empirical cdf

$$F_n(x) \coloneqq \frac{1}{n}\#\{i \colon x_i \le x\}.$$

Again, this approximation introduces an error, the first kind of error.
With this approximation, one

(a) draws with replacement a sample $x_1^*, \ldots, x_n^*$ of size n from F_n,
(b) evaluates the estimator $\hat{\theta}_1^* \coloneqq \hat{\theta}(x_1^*, \ldots, x_n^*)$,
(c) and repeats this procedure B times to obtain $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. The empirical distribution of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ is used as an approximation for the distribution of $\hat{\theta}$. This is the second kind of error, which can be controlled by increasing the number B of bootstrap samples.

(3) Which of the two errors can be controlled by the number B of bootstrap samples? [2]

A: The second kind of error (as described above) can be controlled by increasing the number B of bootstrap samples.

Question 2. Most powerful test. [8] Suppose that under the null hypothesis a random variable X has a uniform distribution on $[0, 1]$, and under the alternative it has a distribution with density

$$f(y) \coloneqq \begin{cases} 2y & \text{if } y \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Consider a single observation x from X.

(1) What is the most powerful test at level $\alpha = 0.1$? [3]

A: Since both hypotheses are simple, the Likelihood-ratio test is the most powerful test by the Neyman-Pearson Lemma. LRT rejects H_0 for small values of

$$\Lambda(x) \coloneqq \frac{\mathrm{lik}_0(x)}{\mathrm{lik}_A(x)} = \frac{1}{2x}.$$

Equivalently, reject H_0 for large values of X.

(2) Work out the critical value of the most powerful test based on X as the test statistic. [1]

A: For any level $\alpha \in (0, 1)$ we find the critical value, c say, of the LRT by solving

$$\alpha = \mathbb{P}\{\text{type I error}\} = \mathbb{P}\{X \ge c | H_0\} = 1 - c,$$

hence $c = 1 - \alpha = 0.9$.

(3) What is the power of this test? [2]

A: The power is given by

$$1-\beta = 1-\mathbb{P}\{\text{type II error}\} = 1-\mathbb{P}\{\text{accept } H_0|H_A\}$$
$$= 1-\mathbb{P}\{X \le c|H_A\} = 1-(2\int_0^c ydy) = 1-c^2 = 1-0.81 = 0.19.$$

(4) Is this test uniformly most powerful for testing against the (composite) alternative

$$H_A\colon f(y) \coloneqq \begin{cases} \lambda y^{\lambda-1} & \text{if } y \in [0,1], \\ 0 & \text{otherwise,} \end{cases} \qquad \lambda > 1?[2]$$

A: For any $\lambda > 1$ the likelihood ratio is

$$\Lambda(x) = \frac{1}{\lambda y^{\lambda-1}},$$

hence LRT rejects for large values of X. The critical value is the same for all λ, as it only depends on the null distribution of X, which is uniform. Consequently, this test is most powerful for any $\lambda > 1$ and therefore uniformly most powerful.

Question 3. Freshmen statistics. [10] Entering freshmen at a University historically selected one of the five colleges as shown in table **??**. In the most recent

	college	percentages
(1)	Business	15%
(2)	Education	20%
(3)	Engineering	30%
(4)	Liberal Arts	25%
(5)	Science	10%

TABLE 1. Historical data on freshmen's college choices.

class 73 students selected business, 105 selected education, 150 chose engineering, 124 selected liberal arts, and 47 selected science.

(1) From the perspective of the administration each entering freshman randomly chooses a college independently of the other freshmen according to the historical frequencies. Let

$N_i \coloneqq$ # freshmen in most recent class that chose college i.

Specify the distribution of the random vector $(N_1, \ldots, N_5)$. [1]

A: The distribution of $(N_1, \ldots, N_5)$ is multinomial with parameters $73 + 105 + 150 + 124 + 47 = 499$ and $p_1 = 3/20$, $p_2 = 1/5$, $p_3 = 3/10$, $p_4 = 1/4$, $p_5 = 1/10$.

(2) The administration is interested in knowing whether or not the historical percentages have changed, in which case expenses have to be made to adapt to the new situation. Give a conservative null hypothesis H_0 and an alternative H_A. [1]

A:

$$H_0 \colon (p_1, \ldots, p_5) = (\frac{3}{20}, \frac{1}{5}, \frac{3}{10}, \frac{1}{4}, \frac{1}{10})$$
$$H_A \colon (p_1, \ldots, p_5) \neq (\frac{3}{20}, \frac{1}{5}, \frac{3}{10}, \frac{1}{4}, \frac{1}{10}),$$

where p_i denotes the historic frequency of freshmen choosing college i.

(3) Propose a test statistic that discriminates between these two hypotheses and explain it. [2]

Pearson's chi squared statistic

$$X^2 \coloneqq \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i^2},$$

where $c = 5$ denotes the number of categories (here: colleges), E_i is the expected number of counts (here: students) in category i (here: college i), and O_i is the observed number of counts in category i.

(4) What is, approximately, the null distribution of your test statistic? [2]

A: X^2 has approximately a chi squared distribution with $c - 1 = 4$ degrees of freedom.

(5) Does the test reject H_0 for large values or for small values of the test statistic? Explain your answer. [2]

A: Large values of X^2 correspond to large values of differences $(O_i - E_i)^2$ between observed and expected numbers of counts (for at least one category i), which contradicts H_0. Thus, the test rejects H_0 for large values of X^2.

(6) Compute the p-value of your test. [2]

A: Let us first evaluate the X^2 statistic. Given H_0, we have $E_i = 499p_i$, hence

$$E_1 = 74.85, E_2 = 99.9, E_3 = 149.7, E_4 = 124.75, E_5 = 49.9.$$

This yields $X^2 = 0.49$, and the p-value is

$$\mathbb{P}\{Z \geq 0.49\} \approx 0.975,$$

where Z denotes a random variable with χ^2_4 distribution.

Question 4. Diagnosing cancer. [5] Investigators are interested to know whether counting blood platelets might be useful in diagnosing cancer. In a medical study blood platelets were counted for a group of $n_1 = 153$ male cancer patients and for a control group of $n_2 = 35$ healthy males. The data are summarized in table **??**.

Let

$X_i \coloneqq$ number of platelets counted for individual i in group of cancer patients

$Y_i \coloneqq$ number of platelets counted for individual i in control group.

	sample size	sample mean	sample standard deviation
cancer patients	$n_1 = 153$	$\bar{X} = 395$	$S_1 = 170$
healthy males	$n_2 = 35$	$\bar{Y} = 235$	$S_2 = 45.3$

TABLE 2. Platelet counts for cancer patients and controls.

(1) Propose a level $\alpha = 0.05$ test of the null hypothesis $H_0 \colon \mu_X = \mu_Y$ that platelet counts cannot be used to discriminate between cancer patients and healthy males versus the alternative $H_A \colon \mu_X \neq \mu_Y$. [3]

Hint: Sample sizes are large enough to approximate $\bar{X}$ and $\bar{Y}$ by normals.

A: Notice that since the sample variances differ considerably, we do not assume the population variances to be equal and hence do not use the pooled sample variance. Consider then the test statistic

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}$$

which has standard normal distribution. However, since we don't know σ_X^2 and σ_Y^2, we estimate them by the sample variances s_X^2, s_Y^2. The test rejects H_0 for large values of $|Z|$, where Z is defined by

$$Z := \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}$$

In other words, the rejection region is of the form $[c, \infty)$ for some critical value $c > 0$. To find c, we recall that under the null Z is approximately normal (due to large sample size), and solve

$$\alpha = \mathbb{P}\{|Z| > c | H_0\} = 2\mathbb{P}\{Z \leq -c | H_0\}.$$

This is equivalent to

$$\mathbb{P}\{Z \leq c | H_0\} = 1 - \frac{\alpha}{2} = 0.975.$$

From the distribution table we find $c = 1.96$.

(2) What is the p-value of your test? [2]

Evaluating the test statistic yields

$$Z = \frac{|395 - 235|}{\sqrt{\frac{170^2}{153} + \frac{(45.3)^2}{35}}} = 10.17.$$

The p-value is the smallest significance level α at which the test rejects H_0. In this example, it is the probability that $|Z|$ has a value as large as 10.17, or even larger. We compute

$$\mathbb{P}\{|Z| > 10.17 | H_0\} = \Phi(-10.17) + 1 - \Phi(10.17) = 2\Phi(-10.17) \approx 0,$$

since $\Phi(-10.17)$ is very close to 0. This means that one can confidently reject H_0 based on the data. Platelet counts seem to discriminate well between healthy males and males with cancer.

Question 5. Crime reports. [7] The data in table **??** from police records show the number of daily crime reports from a sample $X_1, \ldots, X_{11}$ of days during the winter months and a sample $Y_1, \ldots, Y_{11}$ of days during the summer months. We want

X_i winter	Y_i summer
18	28
20	18
15	24
16	32
21	18
20	29
12	23
16	38
19	28
20	18
39	40

TABLE 3. Police records on crime reports.

to test the null hypothesis

H_0: distribution of crime reports is the same in summer and winter

versus the alternative

H_A: distributions of crime reports in summer and winter differ.

Consider Wilcoxon's sum-of-ranks test for the comparison of unmatched samples from two populations based on the sum-of-ranks $R_2 \coloneqq \sum_{i=1}^{n_2} \operatorname{rank}(Y_i)$ as a test statistic.

(1) Under H_0 what can you say about the value of R_2? [1]
Hint: Recall that

$$\mathbb{E}R_2 = \frac{1}{2}n_2(n_1+n_2+1), \qquad \operatorname{Var}(R_2) = \frac{1}{12}n_1n_2(n_1+n_2+1).$$

(Notice that there were two typos in the formulas given on the exam as $\mathbb{E}R_2 = \frac{1}{2}n_1(n_1+n_2+1)$ and $\operatorname{Var}(R_2) = \frac{1}{2}n_1n_2(n_1+n_2+1)$. If you used these wrong formulas but carried out the rest of the calculations correctly, you still earn full credit.)

A: If the value of R_2 is close to its minimal value $n_2(n_2+1)/2$, respectively its maximal value $n_1n_2 + n_2(n_2+1)/2$, then the data contradict H_0. The data support H_0 if the value of R_2 is close to its mean $\mathbb{E}R_2$.

(2) Compute the sum-of-ranks R_2 for crime records during summer. [2]
A: Recall that for ties the ranks are averaged, i.e. all 4 observations of 18 are assigned the rank 6.5. We obtain $R_2 = 160.5$.

(3) Since sample sizes are large enough (both sample sizes exceed 11), we approximate the sum-of-ranks R_2 by a normal distribution. Using this approximation, compute the p-value of Wilcoxon's rank-sum-test. [2]

A:

$$\mathbb{P}\left\{|\frac{R_2-\mathbb{E}R_2}{\sqrt{\mathrm{Var}(R_2)}}| \geq \frac{160.5-\mathbb{E}R_2}{\sqrt{\mathrm{Var}(R_2)}}\right\} = 2(1-\Phi(\frac{160.5-\mathbb{E}R_2}{\sqrt{\mathrm{Var}(R_2)}}))$$
$$= 2(1-\Phi(\frac{34}{15.23})) = 2(1-2\Phi(2.23)) = 0.026,$$

since $\mathbb{E}R_2 = 126.5$, $\mathrm{Var}(R_2) = 231.92$.

(4) What is the advantage of a nonparametric over a parametric test? [1]

A: For a nonparametric test there are no requirements on the distribution of the populations.

(5) Describe a setting where you would expect a parametric test to outperform a nonparametric test. [1]

A: Consider the setting were both hypotheses H_0 and H_A are simple. From the Neyman-Pearson Lemma we know that the most powerful test is the LRT. A nonparametric test in this setting is probably not as powerful.

Question 6. Linear regression. [10] Consider a sample of n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$.

(1) State the assumptions of a simple linear regression model with response variable y and explanatory variable x. [4]

A: In the simple linear regression model the relationship between explanatory variable x and response y is assumed to be given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for each $i \in \{1, \ldots, n\}$. Moreover, β_0, β_1 are deterministic parameters, the x_i are considered to be fixed, and (ϵ_i) is assumed to be an i.i.d. sequence of random variables such that $\mathbb{E}\epsilon_i = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$.

(2) Give an unbiased estimator for the slope and an unbiased estimator for the intercept of the regression line. [2]

A:

$\hat{\beta}_0 = -\hat{\beta}_1\bar{x} + \bar{y}$, $\hat{\beta}_1 = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2}$.

(3) Since we do not have access to the population, how can we check whether there is a linear relationship in the population before performing a linear regression? Name a graphical and a quantitative method. [2]

A: An informal graphical method is the scatterplot. A quantitative method would be to compute the correlation coefficient r, as a value of r close to zero indicates that there is no linear relationship in the sample.

(4) One assumption of the linear regression model is that the standard deviation of responses about the population line is the same for all values of the explanatory variable. After carrying out the linear regression, how can we check this assumption graphically? [1]

A: We can check this assumption graphically by plotting the residuals $\hat{\epsilon}_i$ versus the predictors x_i. The distribution should be approximately uniform; hence, if there is an obvious pattern in this plot, this indicates that the assumption of equal standard deviations of responses is violated.

(5) Suppose the population meets the regression assumptions. What does a test of

$$H_0 \colon \beta_1 = 0 \qquad \text{versus} \qquad H_A \colon \beta_1 \neq 0$$

tell us about the relationship of x and y? [1]

A: If this test rejects H_0, this indicates that there is indeed a linear relationship between x and y. If H_0 cannot be rejected, the data do not support the assumption of a linear relationship.