DSCC/CSC 462: Computational Introduction to Statistics

Midterm

PRACTICE

Name: Solutions

"I affirm that I will not give or receive any unauthorized help on this exam, and that all work will be my own."

_____

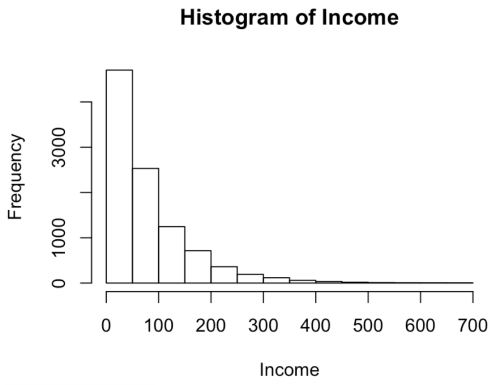
_____

_____

Signature _____

Show your work for all questions in order to receive partial credit for incorrect answers. You may use R, class notes, and a calculator, but no internet-enabled devices. This practice test is roughly twice as long as the actual midterm.

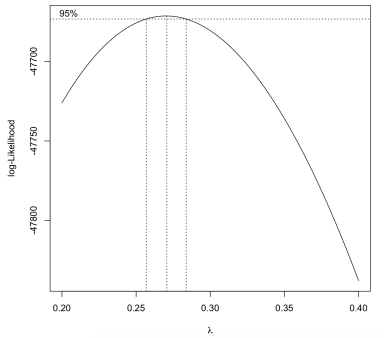
1. A CEO examines the incomes of her employees. A histogram of the incomes is given here:



(a) (2 points) This distribution can best be described as having:

A. Strong positive skew
B. Strong left skew
C. Weak negative skew
D. Weak right skew
E. Symmetry and uniform spread
F. Symmetry and bell-shaped spread

(b) (3 points) The CEO wants to transform the incomes to make their distribution more normal. Consider the following output:




```
> bc1$y[bc1$y==max(bc1$y)]
[1] -47671.4
>
> bc1$y[bc1$x==max(bc1$x)]
[1] -47837.78
>
> bc1$x[bc1$x==max(bc1$x)]
[1] 0.4
>
> bc1$x[bc1$y==max(bc1$y)]
[1] 0.2707071
```

What is the **most** appropriate Box-Cox transformation to make to the data? Write out the functional form of the transformation.

$$y = \frac{x^{0.2707071} - 1}{0.2707071}$$, where x is the original data.

2. (12 points) The density function for random variable X is given as

$$f_X(x) = \begin{cases} ax^3 + bx & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $E(X) = \frac{1}{3}$, find a and b.

① integral of PDF = 1

$$\int_0^1 ax^3 + bx \, dx = 1$$

$$\frac{ax^4}{4} + \frac{bx^2}{2} \Big|_0^1 = 1$$

$$\frac{a}{4} + \frac{b}{2} = 1$$

$$a + 2b = 4$$

$$a = 4 - 2b$$

$$a = 4 - 2(7)$$

$$\boxed{a = -10}$$

② $E[X] = \frac{1}{3}$

$$\int_0^1 x(ax^3 + bx) \, dx = \frac{1}{3}$$

$$\frac{ax^5}{5} + \frac{bx^3}{3} \Big|_0^1 = \frac{1}{3}$$

$$\frac{a}{5} + \frac{b}{3} = \frac{1}{3}$$

$$3a + 5b = 5$$

$$3(4 - 2b) + 5b = 5$$

$$12 - b = 5$$

$$\boxed{b = 7}$$

3. (6 points) An accountant advertises that if he makes any mistakes on your taxes, he will pay you \$50. On average, the accountant makes 2.4 mistakes during a tax season, and the distribution of the number of mistakes follows a Poisson distribution. What is the probability that the accountant will have to pay, in total, more than \$50 this tax season due to mistakes? You can assume that mistakes are independently made.

Let X = # mistakes.

$$Pr(\text{Pay} > \$50) = Pr(>1 \text{ mistake}) = Pr(X > 1) = 1 - Pr(X \le 1)$$

$$= 1 - Pr(X = 0) - Pr(X = 1)$$

$$= 1 - e^{-2.4} - 2.4\, e^{-2.4}$$

$$= 1 - 3.4\, e^{-2.4}$$

$$= \boxed{0.692}$$

4. (8 points) Three treasure chests are presented to a pirate. In one of the chests, there are two gold coins. In a second chest, there are two silver coins. In a third chest, there is one gold coin and one silver coin. The pirate chooses a chest and pulls out one coin. It is gold. Given that the coin is gold, what is the probability that the other coin in the chest is also gold? Use Bayes' rule to answer this question.

GG, SS, GG, SG

$$\Pr(\text{GG} \mid \text{see G}) = \frac{\Pr(\text{see G} \mid \text{GG}) \Pr(\text{GG})}{\Pr(\text{see G})}$$

$$= \frac{\Pr(\text{see G} \mid \text{GG}) \Pr(\text{GG})}{\Pr(\text{see G} \mid \text{GG}) \Pr(\text{GG}) + \Pr(\text{see G} \mid \text{SG}) \Pr(\text{SG}) + 0}$$

$$= \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/2 \cdot 1/3} = \boxed{\frac{2}{3}}$$

5. Three missiles are independently fired at a target, and each has probability 0.4 of hitting the target.

(a) (4 points) Find the probability that the target is hit by exactly two missiles.

Let X = # missiles that hit target. $X \sim \text{Binom}(3, 0.4)$ (n, p)

$$\Pr(X = 2) = \binom{3}{2} (0.4)^2 (0.6)$$

$$= \boxed{0.288}$$

(b) (2 points) What is the mean number of missiles that the target is expected to be hit by?

$$E(X) = np = \boxed{1.2 \text{ missiles}}$$

(c) (2 points) What is the variance of the number of missiles that the target is expected to be hit by?

$$\text{Var}(X) = np(1-p) = \boxed{0.72 \text{ missiles}^2}$$

6. (3 points) The mean age of flight attendants is 40 years old, with a standard deviation of 8 years. What percent of flight attendants are between 20 years old and 60 years old?

A. 68%
B. 72%
C. 75%
D. 84%
E. 95%
F. 98.8%
G. 99.7%

$\frac{20-40}{8} = -2.5$

$\frac{60-40}{8} = 2.5$

Use Chebyshev's Inequality because shape of distribution is unknown.

$1 - \left(\frac{1}{2.5}\right)^2 = 0.84$.

7. (5 points) In poker, a full house occurs when you have three cards of one rank and two cards of another rank. For example if you have three queens and two aces, this would be full house. Three queens, one king, and one ace would NOT be full house. How many ways can you be dealt a poker hand that is a full house?

$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2} = 13 \cdot 4 \cdot 12 \cdot 6 = 3744$

choose value of triple — choose cards for triple — choose value of double — choose cards for double

8. Consider random variables X and Y. X has mean $\mu_X = 25$ and standard deviation $s_X = 3$. Y has mean $\mu_Y = 10$ and standard deviation $s_Y = 2$. The covariance between X and Y is $cov(X, Y) = 1$. The random variable Z is defined as $Z = X + aY$. Suppose we also know that $E(Z) = \mu_Z = 5$.

(a) (3 points) Determine the value of a.

$5 = E(Z) = E(X + aY) = E(X) + a\,E(Y) = \mu_X + a\mu_Y = 25 + a(10)$

$5 = 25 + 10a \rightarrow a = -2$

(b) (5 points) Calculate $var(Z)$.

$Var(Z) = Var(X + aY) = Var(X) + a^2 Var(Y) + 2a \cdot Cov(X, Y)$

$= s_X^2 + (-2)^2 s_Y^2 + 2(-2)\,cov(X, Y)$

$= 9 + 4 \cdot 4 - 4 \cdot 1$

$= 21$

9. (2 points) A doctor sees patients throughout a typical day. Times spent with patients (in minutes) follow an exponential distribution with parameter $\lambda = 0.0625$. On average, what is the expected amount of time that a doctor spends with a given patient?

Let X: amount of time spent with patient. $X \sim \text{Exp}(0.0625)$

$E(X) = \frac{1}{\lambda} = \frac{1}{0.0625} =$ **16 mins**

10. (3 points) Suppose a designer has a palette of 10 colors to work with and wants to design a flag with 3 vertical stripes, all of different colors. How many possible flags can be created?

$10 \cdot 9 \cdot 8 =$ **720**

(10: first stripe; 9: second stripe; 8: third stripe)

11. Of all the clients that come to a particular lawyer's office, 80% are guilty (G). Additionally, 40% of the clients are charged with assault crimes. The lawyer observes that 30% of his clients are both charged with assault crimes and guilty (A).

(a) (4 points) What is the probability that a client is guilty or charged with assault?

Make table:

	A	A^c	
G	0.3	0.5	0.8
G^c	0.1	0.1	0.2
	0.4	0.6	

$\Pr(G \cup A) = \Pr(G) + \Pr(A) - \Pr(G \cap A)$

$= 0.8 + 0.4 - 0.3 =$ **0.9**

(b) (4 points) What is the probability that a client is charged with assault or is not guilty?

$\Pr(A \cup G^c) = \Pr(A) + \Pr(G^c) - \Pr(A \cap G^c)$

$= 0.4 + 0.2 - 0.1 =$ **0.5**

(c) (4 points) Given that a client is not charged with assault, what is the probability that he is not guilty?

$\Pr(G^c \mid A^c) = \frac{\Pr(G^c \cap A^c)}{\Pr(A^c)} = \frac{0.1}{0.6} =$ **0.167**

(d) (3 points) Are guilt and charge independent? Justify your answer.

$\Pr(G \cap A) \stackrel{?}{=} \Pr(G) \cdot \Pr(A)$

$0.3 \stackrel{?}{=} (0.8)(0.4) \Rightarrow$ Guilt and charge are not independent.

$0.3 \neq 0.32$

12. (3 points) You have a dataset with 100 observations. According to Sturges' rule, how many bins should be used when making a histogram of these observations?

$\lceil \log_2(100) + 1 \rceil = \boxed{8}$

13. Consider the data $1, 5, 6, 7, 11$.

(a) (2 points) Calculate the mean of the data.

$\bar{x} = \frac{1+5+6+7+11}{5} = \boxed{6}$

(b) (2 points) Calculate the median of the data.

1 5 **6** 7 11

(c) (2 points) Calculate the 20% trimmed mean of the data.

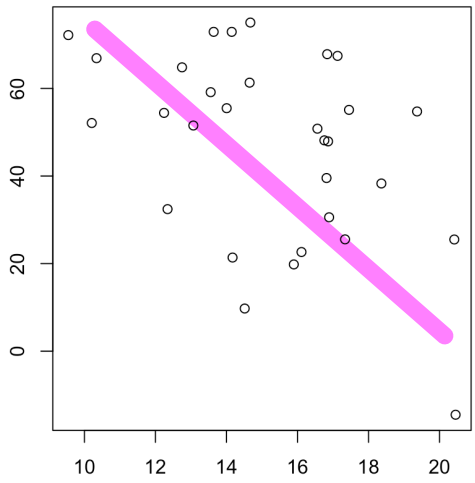
$\bar{x}_{\text{tr}} = \frac{5+6+7}{3} = \boxed{6}$

14. (2 points) Suppose $\bar{x} = 20$ and $s^2 = 100$. What is the value of the coefficient of variation?

$$CV = \frac{s}{\bar{x}} = \frac{10}{20} = \boxed{0.5}$$

15. (2 points) How is the association in the following scatterplot best described?

A. Strong, positive, and linear
B. Strong, negative, and linear
C. Weak, positive, and linear
D. Weak, negative, and linear
E. Strong and non-linear
F. Weak and non-linear
G. No association



16. Suppose that 72% of adults over 30 are married (M) and have a probability of 0.91 of having a child (C). In contrast, the remaining 28% of single adults over 30 have a probability of 0.24 of having a child.

(a) (5 points) If we pick an adult over thirty at random, what is the probability that they have a child?

$$Pr(C) = Pr(C|M)Pr(M) + Pr(C|M^c)Pr(M^c)$$
$$= (0.91)(0.72) + (0.24)(0.28) = \boxed{0.7224}$$

(b) (5 points) What is the probability that an adult over 30 will be married given that they do not have a child?

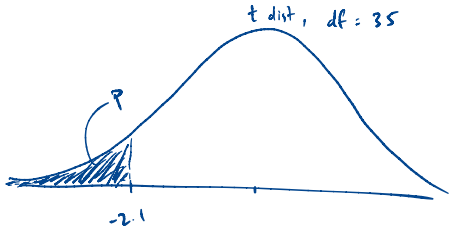
$$Pr(M|C^c) = \frac{Pr(C^c|M)Pr(M)}{Pr(C^c)} = \frac{(0.09)(0.72)}{1 - 0.7224} = \boxed{0.2334}$$

A department store has hired you to analyze customer trends. All subsequent questions included in this exam are based on this setup.

17. The store takes a random sample of $n = 36$ customers. The mean age of the sample of customers is $\bar{x} = 43.7$ years with a standard deviation of $s = 18.0$ years.

    (a) (4 points) Let μ = the average age of customers at the store. Calculate a test statistic for testing the hypothesis $H_0 : \mu \geq 50$ years against $H_1 : \mu < 50$ years.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{43.7 - 50}{18/\sqrt{36}} = \boxed{-2.1}$$



    (b) (2 points) What distribution does this test statistic follow?

t distribution with df = 35.

    (c) (2 points) Calculate the p-value of this test. Write down the R code you used to calculate it.

pt(-2.1, 35) = 0.0215

    (d) (3 points) At the $\alpha = 0.05$ significance level, what conclusion do you reach regarding the null hypothesis?

Reject H_0. Conclude that the mean age of customers is greater than 50.

18. (6 points) The store wants to determine the mean amount of time customers spend on the store's website with a margin of error of 3.5 minutes. It is known from previous analyses that the standard deviation of time spent on the website is $\sigma = 10.0$ minutes. If we want to create a two-sided 95% confidence interval, at least how many customers must be included in the study? The critical value to use in your calculation is $z_{\alpha/2} = 1.96$.

m

$$n = \left\lceil \left(\frac{z_{\alpha/2} \cdot \sigma}{m} \right)^2 \right\rceil = \left\lceil \left(\frac{1.96 \cdot 10}{3.5} \right)^2 \right\rceil = \lceil 31.36 \rceil = \boxed{32}$$

19. Are there differences in spendings in the men's department and the women's department? To examine this question, a sample of $n_1 = 64$ purchases from the men's department and $n_2 = 64$ purchases from the women's department were examined. The average amount spent in the men's department was $\bar{x}_1 = \$98.40$ with a standard deviation of $s_1 = \$48.00$. The average amount spent in the women's department was $\bar{x}_2 = \$125.20$ with a standard deviation of $s_2 = \$64.00$.

(a) Test at the $\alpha = 0.05$ significance level whether the variability in spendings in the men's department is different than the variability in spendings in the women's department. In particular, test the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$.

i. (4 points) Calculate the appropriate test statistic.

$$F = \frac{s_1^2}{s_2^2} = \frac{48^2}{64^2} = \boxed{0.5625}$$

ii. (3 points) What distribution does the test statistic follow? Make sure to specify the appropriate degrees of freedom associated with the distribution.

numerator ↙ ↘ denominator

F distribution with 63 and 63 degrees of freedom

iii. (3 points) Calculate the p-value associated with this test statistic. Write down the R code used to calculate it.

pf(0.5625, 63, 63) = 0.0239

iv. (3 points) At the $\alpha = 0.05$ significance level, what conclusion do you reach regarding the null hypothesis?

Reject H_0. Conclude different variabilities.

(b) Assuming unequal variances, test at the $\alpha = 0.05$ significance level whether the average amount spent in the men's department is less than the average amount spent in the women's department, or $H_0 : \mu_1 - \mu_2 \geq 0$ vs. $H_1 : \mu_1 - \mu_2 < 0$.

i. (6 points) Calculate the appropriate test statistic.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{98.4 - 125.2}{\sqrt{\frac{48^2}{64} + \frac{64^2}{64}}} = \boxed{-2.68}$$

ii. (4 points) Calculate the critical value for this test. Write down the R code used to calculate it.

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = 116.84$$

qt(0.05, 116.84) = -1.658

iii. (3 points) Compare your test statistic to the critical value. What conclusion do you reach regarding the null hypothesis?

$t = -2.68 < -1.658 \rightarrow$ Reject H_0. Conclude that mean amount spent in men's department is less than the mean amount spent in the women's department.

20. (8 points) Do customers spend, on average, more money in the evenings as compared to in the mornings? To investigate this question, a sample of $n_1 = 9$ evening purchases and a sample of $n_2 = 16$ morning purchases were examined. The mean amount spent in the evening was $\bar{x}_1 = \$34.60$ with a standard deviation of $s_1 = \$3.00$. The mean amount spent in the morning was $\bar{x}_2 = \$32.10$ with a standard deviation of $s_2 = \$3.00$. Assume that the amount spent follows a normal distribution and that the variances of the two populations are equal. Create an appropriate one-sided (lower-bound) 95% confidence interval for the $\mu_1 - \mu_2$ difference.

$CI: \left((\bar{x}_1 - \bar{x}_2) + t_\alpha \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \infty \right)$

$s_p = 3$

$t_\alpha = qt(0.05, 9+16-2)$

$= -1.714$

$\rightarrow \left((34.6 - 32.1) - 1.714 \sqrt{9\left(\frac{1}{9} + \frac{1}{16}\right)}, \infty \right)$

$\rightarrow \boxed{(0.3575, \infty)}$

21. Do coupons incentivize customers to buy more products? To test this, a small experiment was conducted. During week 1, a group of $n = 16$ customers were allowed to shop in the store without using coupons. The number of items purchased by each customer during week 1 was recorded. During week 2, the same sample of $n = 16$ customers were allowed to shop in the store and could use coupons. The number of items purchased by each customer was again recorded. For each customer, the week 2 - week 1 difference in the number of items purchased was calculated. The average difference was $\bar{d} = 1.5$ items with a standard deviation of $s_d = 2.5$ items. Assume the number of items bought follows a normal distribution.

(a) Let δ = the mean in week 2 - week 1 differences. Test the hypothesis $H_0 : \delta \leq 0$ against $H_1 : \delta > 0$ at the $\alpha = 0.05$ significance level.

i. (4 points) Calculate the appropriate test statistic.

$t = \frac{\bar{d} - \delta}{s_d/\sqrt{n}} = \frac{1.5 - 0}{2.5/\sqrt{16}} = \boxed{2.4}$

ii. (2 points) What distribution does this test statistic follow?

t distribution with 15 df

iii. (2 points) Calculate the p-value for this test. Write down the R code used to calculate it.

`1 - pt(2.4, 15)` = 0.015

iv. (3 points) At the $\alpha = 0.05$ significance level, what conclusion do you reach regarding the null hypothesis?

Reject H_0 and conclude that coupons increase the number of products purchased.

22. The store sells four designer brands (we will call them Designer A, Designer B, Designer C, and Designer D). It is assumed that all four brands are equally preferred by customers. A sample of $n = 200$ customers is taken, and they are asked their favorite of the four designers. Of the 200 customers, 20 preferred Designer A, 100 preferred Designer B, 50 preferred Designer C, and 30 preferred Designer D. Use an appropriate chi-square test at the $\alpha = 0.05$ significance level to determine whether the four designers are equally preferred.

(a) (2 points) State your null and alternative hypotheses.

H_0 : $P_A = P_B = P_C = P_D = 0.25$

H_1 : At least one of these equalities does not hold.

(b) (4 points) Calculate the expected number of customers to prefer each brand if the null hypothesis is true (four cells total):

Designer A	Designer B	Designer C	Designer D
50	50	50	50

$E = np$

(c) (6 points) Calculate the appropriate test statistic.

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(20-50)^2}{50} + \frac{(100-50)^2}{50} + \frac{(50-50)^2}{50} + \frac{(30-50)^2}{50}$$

$$= 18 + 50 + 0 + 8 = \boxed{76}$$

(d) (2 points) How many degrees of freedom are associated with the distribution of this test statistic?

$df = k - 1 = 4 - 1 = \boxed{3}$

(e) (2 points) Calculate the p-value, and write the R code for calculating the p-value.

$p = 1 - \text{pchisq}(76, 3) = 2.2 \times 10^{-16} \approx 0.$

(f) (3 points) At the $\alpha = 0.05$ significance level, what conclusion do you reach regarding the null hypothesis?

Reject H_0. Not all brands are preferred equally.

23. Let p be the proportion of customers who are women. It is hypothesized that $H_0 : p \leq 0.5$ vs. $H_1 : p > 0.5$. A sample of $n = 100$ customers is taken.

(a) (4 points) At the $\alpha = 0.05$ significance level, calculate the minimum value of $\hat{p}$ that will lead to the rejection of the null hypothesis.

$$\hat{p} = p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} = 0.5 + 1.645\sqrt{\frac{(0.5)(0.5)}{100}}$$

$$= \boxed{0.58225}$$

(b) Assume that the true value of p is $p = 0.6$. Using this information and your result in part a, calculate the power of the test.

i. (4 points) Calculate the z-score for the estimate from part a assuming a true value of $p = 0.6$.

$$z = \frac{0.58225 - 0.6}{\sqrt{\frac{(0.6)(0.4)}{100}}} = \boxed{-0.3623}$$

ii. (3 points) What is the power of the test? Write down the R code you used to calculate it.

```
1 - pnorm(-0.3623)
```

$$= \boxed{0.641}$$

24. Do men and women equally shop online? To investigate this question, let p_1 be the proportion of men who shop online and let p_2 be the proportion of women who shop online. Out of a sample of $n_1 = 24$ men, $x_1 = 16$ reported shopping online in the past month. Out of a sample of $n_2 = 40$ women, $x_2 = 32$ reported shopping online in the past month.

(a) Test whether shopping online is independent of the customer's sex. In particular, consider H_0 : online shopping and sex are independent against H_1 : online shopping and sex are associated.

i. (3 points) Fill in the two-way table with the observed counts and totals (9 cells total):

	Shop online	Do not shop online	Total
Men	16	8	24
Women	32	8	40
Total	48	16	64

ii. (4 points) Calculate the table of expected counts (9 cells total):

	Shop online	Do not shop online	Total
Men	18	6	24
Women	30	10	40
Total	48	16	64

iii. (6 points) Calculate the appropriate test statistic.

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(16-18)^2}{18} + \frac{(8-6)^2}{6} + \frac{(32-30)^2}{30} + \frac{(8-10)^2}{10}$$

$$= \boxed{1.422}$$

iv. (2 points) How many degrees of freedom are associated with the distribution of this test statistic?

$$df = (r-1)(c-1) = (2-1)(2-1) = \boxed{1}$$

v. (4 points) Calculate the critical value, and show your R code.

```
qchisq(0.95, 1) = 3.84
```

vi. (3 points) What conclusion do you reach?

Fail to reject H_0. There is insufficient evidence of a relationship between sex and online shopping.