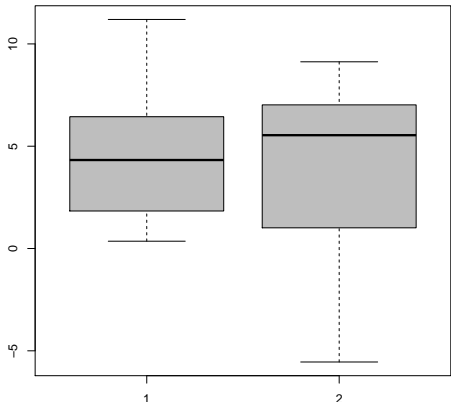


Practice Midterm Exam Solutions

Statistics 139

Problem 1. [3 points each] Parts are unrelated unless otherwise specified.

- (a) One scientist believes a new treatment will improve survival 6 months while another investigator believes it will improve 3 months over standard treatment. A valid t -based 95% confidence interval for a mean difference between treatment and control was calculated to be (2, 8).
- A) Both scientists' claims are equally plausible since they are both inside the confidence interval.
 - B) 3 months is more plausible since it is closer to the null hypothesis of a mean difference of 0.
 - C) **6 months is more plausible since it is closer to the observed sample mean.**
 - D) Cannot be determined from the information given.
- (b) You'd like to perform a hypothesis test to determine whether the the top 10% of earners in Massachusetts is different than it is in New Hampshire (as measured by the 90th percentile of income). A simple random sample of 500 income tax returns is taken within each state. Which test would make the most sense to perform to compare these two groups?
- A) Randomization test
 - B) **Permutation test**
 - C) Bootstrap test
 - D) Proportion z -test
- (c) You'd like to perform a test to determine whether the independent groups in the boxplot to the right come from distributions with similar centers. Which test is most appropriate?
- 
- A) Unpooled t -test
 - B) Permutation test
 - C) **Rank Sum Test**
 - D) ANOVA F -test
- (d) A test of $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$ is conducted on the same population independently by two different researchers. They both use the same sample size and the same value of $\alpha = 0.05$. Which of the following will be the same for both researchers?
- A) The p -value of the test.
 - B) **The power of the test if the true $\mu = 6$.**

- C) The value of the test statistic.
 - D) The decision about whether or not to reject the null hypothesis.
- (e) In a hypothesis test the decision was made to not reject the null hypothesis. Which type of mistake could have been made?
- A) Type 1.
 - B) Type 2.
 - C) Type 1 if it's a one-sided test and Type 2 if it's a two-sided test.

Problem 2. [5 points each unless stated] Parts are unrelated unless otherwise stated.

- (a) A recent study claimed that getting rid of annual medical check-ups would be harmful since patients that attend their annual check-ups have better health outcomes than those patients that skip them. Provide one reason why this claim may be incorrect. Be specific.

Since patients are not randomly assigned to attending their check-ups vs. skipping them, there is potential for confounding factors. It is reasonable to assume that patients that choose to attend their check-ups make other healthy choices as well (take their medicine, better diet, don't smoke, etc...), so the check-ups may not be causing the observed difference in these two groups in health outcomes.

- (b) You perform a hypothesis test of the mean using a sample size of four units, and you do not reject the null hypothesis. Your research colleague says this statistical test provides conclusive evidence that the null hypothesis is reasonable. Do you agree or disagree with his conclusion? Explain your reasoning in three or fewer sentence.

The key is that the conclusion was based on a sample size of just 4, which means there was little to no power to reject a null hypothesis even if the alternative hypothesis was correct. Note: saying the null hypothesis is **reasonable** is a correct conclusion when the null hypothesis is not rejected.

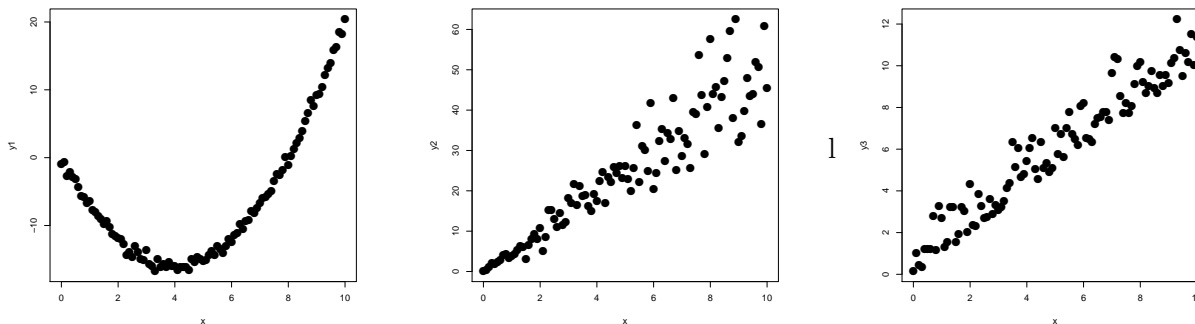
- (c) In comparing 10 groups, you notice that Y_7 is the largest and Y_3 is the smallest, and proceed to test the hypothesis that $H_0 : \mu_3 - \mu_7 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made?

Even though only one formal statistical comparison is being made, this hypothesis was generated after considering informal comparisons of all 10 groups. An analyst should never use the data to generate hypotheses: this is a form of *p-hacking* or *data mining* to find a significant result.

- (d) The median test score on a Stat 139 exam was 85 (out of 100). Would the mean be expected to be above, below, or around the same value of 85? In one or two sentences, explain why.

Since the median is much closer to 100 than 0 (the two boundary conditions), the distribution is likely to be left-skewed, thus the mean is likely to be below the median of 85.

- (e) In the plot provided below, draw a scatter of points that clearly violates one of the assumptions of linear regression, but not the others. Be sure to mention which assumption it violates.



Answers will vary, but the above 3 graphs are the most reasonable graphs to show. The assumption violated in each (going left to right) is linearity, constant variance, and Normality.

- (f) A regression is run in order to determine whether the last $n = 102$ monthly returns of Microsoft Stock prices (`msft`) mimic that of McDonald's Stock prices (`mcd`). Based on the R-output below, determine whether a slope of $\beta_1 = 1$ is reasonable.

```
> summary(lm(msft~mcd))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.000278	0.006720	-0.041	0.967
mcd	0.783439	0.160266	4.888	3.97e-06

This can be determined based on the confidence interval for β_1 or based on a formal hypothesis test (either way, $df = 100 \Rightarrow t^* = 1.984$):

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{s_{\hat{\beta}_1}} = \frac{0.7834 - 1}{0.1603} = -1.35 \quad \text{or} \quad \hat{\beta}_1 \pm t^* s_{\hat{\beta}_1} = 0.7834 \pm 1.984(0.1603) = (0.465, 1.101)$$

Since the magnitude of the calculated statistic is less than the critical value of 1.984 (and the CI does contain the null value of 1), we would conclude that $\beta_1 = 1$ is a reasonable value for the true slope.

- (g) Let $t = \frac{\hat{\beta}_1 - \beta_{H_0}}{\hat{\sigma}_e \sqrt{\frac{1}{(n-1)S_X^2}}}$ be the usual t -statistic for the OLS slope estimate in a simple regression model. Derive the distribution of t^2 through representation. What assumptions do you need for this result to be exact?

$$t_{n-2}^2 \sim \left(\frac{N(0,1)}{\sqrt{\chi_{n-2}^2/(n-2)}} \right)^2 \sim \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2}$$

One of the keys is that the Normal r.v. and the χ^2 r.v. in the t -distribution are independent, and thus the two χ^2 r.v.s in the F -distribution are independent (which is necessary for the result to hold). In order for this to be the exact sampling distribution, we need all 4 assumptions in regression to hold (linearity, constant variance, normality, and independence) plus the null hypothesis to hold: $\beta_1 = 0$.

- (h) In the multiple regression model based on OLS with no missing data (briefly explain your answer):
- **True**** or False: R^2 cannot be negative.
 - **True**** or False: R^2 is the percent of variability in the response variable associated with the predictors.
 - True or ****False****: R^2 equals the sum of the squares of the separate correlation coefficients r of the response with each predictor separately.
 - True or ****False****: R^2 may decrease when an additional explanatory variable is added.
 - For OLS regression, the R^2 cannot be negative since the worst case scenario is using $\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{Y}$, which leads to $R^2 = 0$.
 - The formula is $R^2 = 1 - SSE/SST$ (the reduction in the variability of Y) which leads to this interpretation.

iii. This only holds if the predictors are all centered at zero and uncorrelated to each other. If there is any correlation among them (or with the column of ones for the intercept), then there will be some redundancy in the predictive capability of the set of predictors.

iv. The worst case scenario is that the new introduced variable's estimated slope is $\hat{\beta}_p = 0$, and thus you are predicting the same \hat{y} as without it, leading to the exact same R^2 .

Problem 3. [16 points total]

A friend tells you that she is the master of the game ‘Rock-Paper-Scissors’ and reports that she truly wins each round of the game with 0.8 probability, but you think she is just bragging (or even lying) and really wins with 0.5 probability. You decide to test your friend on her claim by playing 25 independent rounds of the game, and decide to believe her if she wins 18 or more of the rounds.

- (i) Write down the hypotheses, the test statistic, and determine the true reference distribution for this test statistic for this study.

$H_0 : p = 0.5$ vs. $H_A : p = 0.8$, where p is the true proportion of games that she wins. X , the number of games out of 25 that she wins is the test statistic, and it truly has a reference distribution of $X \sim \text{Binom}(n = 25, p = 0.5)$ when the null is true.

- (ii) Calculate an approximate Type 1 error rate for this test.

Under the null, X ’s distribution can be approximated by a Normal: $X \sim N(\mu = np = 12.5, \sigma^2 = np(1 - p) = 6.25 = 2.5^2)$. Thus:

$$P(X \geq 18) = P\left(\frac{X - \mu}{\sigma} \geq \frac{18 - 12.5}{2.5}\right) = P(Z \geq 2.2) = 1 - 0.9861 = 0.0139.$$

Note, this ignores the continuity correction. To correct for using a continuous distribution to approximate a discrete one (a Binomial r.v. can only be positive whole numbers), $P(X > 17.5) = P(Z > 2) = 1 - 0.9772 = 0.0228$ would be a better approximation.

- (iii) Calculate an approximate power for this test.

Under the alternative, X ’s distribution can be approximated by a Normal: $X \sim N(\mu = 20, \sigma^2 = 2^2)$. Thus

$$P(X \geq 18) = P\left(\frac{X - \mu}{\sigma} \geq \frac{18 - 20}{2}\right) = P(Z \geq -1) = 0.8413.$$

If we corrected for continuity, then it would evaluate to $P(X > 17.5) = P(Z > -1.25) \approx 0.89$.

- (iv) You play your 25 rounds of the game and she wins 16 of them. What do you conclude statistically from this test? What do you conclude practically?

Since the observed statistic $X = 16$ is not in the rejection region (18 or more), we do not reject the null hypothesis. Statistically, your friend may not be any better than you at Rock-Paper-Scissors. Practically, there is a good chance that she is actually better since she did win $16/25 = 64\%$ of the rounds (and this test allowed for very little Type I error).

Problem 4. [25 points total] The following is the R-output for a regression to predict the number of bowls of noodle soup sold at a hip new pho restaurant in town based on the high temperature outside the restaurant that day:

```
> mean(temp)
[1] 52.7
> sd(temp)
[1] 17.35
> mean(soup)
[1] 225.12
> sd(soup)
[1] 86.82
> summary(lm(soup~temp))
```

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -----          26.84   <2e-16 ***
temp         -----          -14.51   <2e-16 ***
```

```
Residual standard error: 37.80 on 48 degrees of freedom
Multiple R-squared:  0.8143,    Adjusted R-squared:  0.8105
F-statistic: 210.5 on 1 and 48 DF,  p-value: < 2.2e-16
```

(a) [3 points] What is the estimated correlation between `temp` and `soup`?

Must be negative since the plot shows a negative association.

$$r = \sqrt{R^2} = -\sqrt{0.8143} = -0.9024$$

(b) [4 points] Calculate the estimated simple regression line for these data.

The estimated regression line is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 463.11 - 4.516X$.

$$\hat{\beta}_1 = r \frac{s_y}{s_x} = -0.9024 \left(\frac{86.82}{17.35} \right) = -4.516$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 225.12 - (-4.516)(52.7) = 463.11$$

(c) [5 points] Calculate a 95% confidence interval for estimating the mean number of bowls of soup sold when the high temperature outside is 52.7 degrees.

We need to use the 95% confidence interval to estimate μ at a particular $X_0 = 52.7$:

$$(\hat{\beta}_0 + \hat{\beta}_1 X_0) \pm t_{df=n-2}^* \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}} = (463.11 - 4.516 \cdot 52.7) \pm 2.0106(37.8) \sqrt{\frac{1}{50} + 0} = (214.4, 235.9)$$

We could have saved some time by using the fact that that point estimate at $\bar{X} = 52.7$ will be $\bar{Y} = 225.12$, and the standard error will just be the standard deviation of the residuals divided by square root of n (since the second term under the square root drops out of the calculation).

```
qt(0.975,df=48)
```

```
## [1] 2.010635
```

(d) [4 points] For a new randomly sampled day when the high temperature is known to be $X = 52.7$, what is the approximate probability that the number of bowls sold will be within the interval in part (c), assuming all assumptions are correct?

This is equivalent to determining what level of a prediction interval will be in those bounds (it will also be centered at 163.69). Thus we can just solve for the t_{48}^* from one side of it, and convert that to a probability (or simply use a normal distribution instead of a t):

$$t^* = \frac{1}{2} \left(\frac{\text{upper bound} - \text{lower bound}}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}} \right) = \frac{(235.9 - 214.7)/2}{37.8 \sqrt{1 + \frac{1}{50} + 0}} = 0.2777$$

This leads to a probability of approximately 0.218 of a random day (based on the prediction interval) falling within this range.

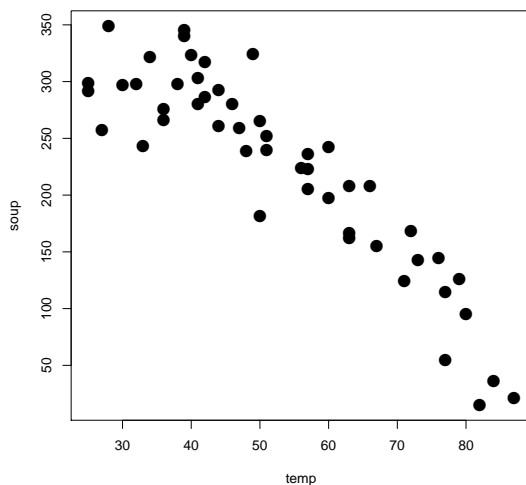
```
1-2*(1-pt(0.2777,df=48))
```

```
## [1] 0.2175647
```

```
1-2*(1-pnorm(0.2777))
```

```
## [1] 0.2187573
```

(e) [3 points] Here's the scatterplot of these data:



Provide the best set of transformations on $Y = \text{soup}$ and $X = \text{temp}$ to make a better regression model. Explain your choice.

Since the spread of the points in the Y direction seems constant, we should look to either transform just X or use a polynomial function of X . It essentially appears to be a rough quadratic function (an “acceleratingly” negative relationship), and thus a quadratic model below seems most appropriate:

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$

where we would expect β_2 to be estimated to be negative and most likely β_1 will be estimated to be positive (since the slope is likely positive when temperature is zero).

(f) [6 points] A second predictor, `heat_index`, is considered to be used to predict soup sales. If `heat_index` is equal to $10 * (\text{temp} - 90)$, then:

- i. What are the estimates of the intercept and slope if `heat_index` is used as the only predictor for `soup`

This is just a linear transformation of the original predictor. The mean and variance of this new X^* will be $\bar{X}^* = 10 * (\bar{X} - 90) = 10 * (52.7 - 90) = -373$ and $S_X^* = 10 * 17.35 = 173.5$. Thus:

$$\hat{\beta}_1^* = r \frac{s_y}{s_X^*} = -0.9085 \left(\frac{86.82}{173.5} \right) = -0.4516$$

$$\hat{\beta}_0^* = \bar{y} - \hat{\beta}_1 \bar{X}^* = 225.12 - (-0.4516)(-373) = 56.67$$

- ii. Give an estimate of the value of R^2 for a multiple regression model to predict `soup` from both `heat_index` and `temp`? Explain how you came to that estimate.

R^2 will remain unchanged from the model with just `temp` as the new introduced variable is redundant (perfectly collinear with what is already there). Thus $R^2 = 0.8143$ still.