Mock Final Exam (with some solutions)

Advanced Statistics for Data Science

Spring 2022

Note: This is a mock final exam. The actual exam will be of similar structure and difficulty level, although there might be changes.

Instructions

- You have 3 hours to complete the exam.
- The exam contains two parts. Part I contains 8 problems, each has a maximal credit of 5 points. Part II contains 3 questions, each has a maximal credit of 20 points. The maximal number of points in the exam is 100.
- For maximal grade, you should answer *all* problems correctly.
- You may bring to the exam up to two personal two-sided A4 pages containing relevant material.

Part I

For the following problems, either indicate **True** or **False** or fill above the blank lines to complete a correct statement or answer (whichever applies).

1. (5 points) Random variables X, Y and Z are independent $\mathcal{N}(0,1)$ random variables. The distribution of $X^2/((Y^2+Z^2)/2)$ is called **Answer:** $F_{1,2}$ (F distribution with 1 DoF over 2 DoF).

   Explain: The numerator is χ^2_1. the denominator is $\chi^2_2/2$. The numerator and denominator are independent.

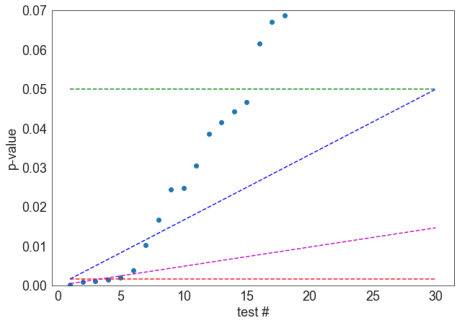
2. (5 points) The treatment group has $\bar{y}_1 = 28.5$ and the control group has $\bar{y}_0 = 31.5$. The standard error $s\sqrt{(1/n_0 + 1/n_1)}$ came out to be 1.5.

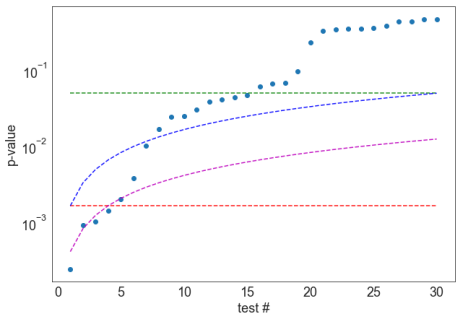
   - The t statistic for equality of the group means is _____ .

- The research claimed a t-test p-value of 0.02559. You redo their computation and get a p-value of 0.05118. What do you think they did wrong? _____

_____

3. (5 points) The figures bellow describe P-values obtained from 30 individual hypothesis tests (the only difference between the figures is the scale of the y-axis, which is logarithmic on the right).





We also have the following legend:

color	curve
green	$y = 0.05$
blue	$y = 0.05 \cdot x/30$
magenta	$y = 0.05 \cdot x/(30 \cdot C_{30})$
red	$y = 0.05/30$

$(C_m = \sum_{i=1}^{m} i^{-1})$

- The tests selected by Binjamin-Hochberg's (BH) procedure for controlling the false discovery rate (FDR) at level $\alpha = 0.05$ are those with P-value rank numbers **Answer:** 1,2,3,4,5,6,7 .

Explain: The blue line is the one that is relevant for BH. The 7-th P-value is the largest one that is below this line.

4. (5 points) Assume that we strongly believe that the response variable y and the covariates vector x are linearly dependent. Which of the following design matrices seems to provide the best least squares coefficients (in the sense of smallest Var $\left[\hat{\beta}\right]$):

$$Z_1 = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad Z_2 = \begin{bmatrix} 1 & -0.2 \\ 1 & -0.1 \\ 1 & 0.1 \\ 1 & 0.2 \end{bmatrix}, \qquad Z_3 = \begin{bmatrix} 1 & -2 \\ 1 & -2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix},$$

Answer: Z_3

Explain: The variance of the predictors in Z_3 is largest, hence the confidence interval for β is smallest.

Additional Explanation: The variance of the first predictor is zero in all cases so we should also consider the second predictor. Among the three cases, the variance of the second predictor is largest in Z_3. A more formal answer involves the S_X term in the confidence interval for $\hat{\beta}$. An heuristic explanation is also acceptable: we are interested in fitting a straight line to 4 noisy measurements in the range $[-2, 2]$. Assuming that the noise variance is constant in all measurements, the best accuracy in of our fit is achieved when we get two measurements at 2 and two at -2.

5. (5 points) We run 10 independent hypotheses tests and obtained P-values $p_1 \leq \ldots \leq p_{10}$. If $p_1 = 0.006$ and $p_{10} = 0.8$, it is possible that we reject 2 or more hypotheses when using Bonferroni correction for significance level 0.05. **(True/False)**

   Answer: False

   Explain: The critical level after a Bonferroni correction is $0.05/10 = 0.005$. However, $0.005 < p_1 \leq p_2$.

6. (5 points) Suppose that we fit a LS model with $p = 5$ predictors and obtain coefficients $\hat{\beta}_j$ for $j = 1, \ldots, 5$. We conduct a t-test for each one of the coefficients to check whether they are different than zero – we obtain that only 2 out of the 5 are significant in the sense that the absolute value of their t statistics exceed the $1 - \alpha/2$ quantile of the t distribution, where $\alpha \in (0, 1)$ is some significant level. Is it possible that all coefficients will turn out to have significant t-test's P-value if we obtain additional measurements? **(True/False)**

   Answer: True

   Explain: The P-value depends on the number of measurements and the size of the effect. Increasing any of these factors can decrease the P-value below the significance level.

   Additional Explanation: Increasing the number of measurements improves the accuracy in estimating the LS coefficients β_j using $\hat{\beta}_j$ in the sense that the confidence interval for β_j (centered at $\hat{\beta}_j$) gets smaller with n (this is the formal way of saying that the accuracy is increased). Connecting this back to test of significance against $\beta_j = 0$: if 0 is not in the $1 - \alpha$ confidence interval of β_j, then the t-test P-value must be smaller than α.

7. (5 points) Suppose that we would like to fit a linear model but our design matrix Z is nearly singular. It makes sense to use ridge regression to regularize our LS estimator. **(True/False)**

   Answer: True

   Explain: Option 1: The ability to handle singular design matrices is one of the motivations of using ridge regression.

   Explain: Option 2: Ridge regression makes a singular design matrix non-singular .

   Additional Explanation: We may also accept False as an answer for a well-explained reason. For example: "False, because ridge regression makes all coefficients non-negative, hence it eliminates any hope for finding a causal relation between the predictors and the response."

8. (5 points) We examine a linear model with 5 predictors. Below are three tables, each potentially describing a path of a model/variable selection procedure for our model. Which of the following paths may correspond to a forward step-wise selection procedure? (left/middle/right)

R^2	variables included
0	$\emptyset$
.3	$\{2\}$
.5	$\{2,3\}$
.6	$\{2,3,5\}$
.62	$\{2,3,5,4\}$

R^2	variables included
.85	$\{1,2,3,4,5\}$
.81	$\{1,2,3,4\}$
.79	$\{2,3,4\}$
.78	$\{2,3\}$
.785	$\{2\}$

R^2	variables included
1	$\emptyset$
.65	$\{2\}$
.6	$\{2,3\}$
.5	$\{2,3,4\}$
.3	$\{2,3,4,5\}$

Explain: _____

_____.

Part II

The questions below may have multiple sections. You should write your response on a separate piece of paper.

1. (20 points) We obtain blood pressure measurements from people in each of three occupations. For simplicity call the occupations A, B and C. Our null hypothesis H_0 is that the average blood pressure in group A equals 2/3 times the average in group B plus 1/3 times the average in group C. The alternative hypothesis is that the A average is not 2/3 times the B average plus 1/3 times the A average.

   Write out the form of the t statistic for testing this hypothesis. State the null distribution of the t statistic and give conditions under which we reject H_0. Introduce and define the notation you need. We can assume that the measurements are independent normally distributed random variables and that they all have the same variance.

   Answer: Set

$$t = \frac{\bar{y}_A - \frac{2}{3}\bar{y}_B - \frac{1}{3}\bar{y}_C}{s\sqrt{\frac{1}{n_A} + \frac{4}{9n_B} + \frac{1}{9n_C}}},$$

   where

$$s^2 = \frac{1}{n-3}\sum_{c\in\{A,B,C\}}\sum_{j=1}^{n_c}(y_{c,j} - \bar{y}_c)^2$$

   and $n = n_A + n_B + n_C$.

   We assume that the variations of all samples around their group mean are normally distributed, independent, and of equal variance. Under the null distribution H_0 we have $t \sim t_{n-3}$. We reject H_0 at significance level $\alpha = 0.05$ if $|t|$ exceeds $t_{n-3}^{0.975}$.

2. (20 points) We have several measurements of some phenomena and we would like to check whether a new measurement originates from the same phenomena (for example, this situation arise to check whether a newly obtained image can be associated with a given class of images, or not). We propose the following model: the observed data $y_1, \ldots, y_n$ obey

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad i = 1, \ldots, n.$$

We would like to check whether some given y_{new} was also sampled from $\mathcal{N}(\mu, \sigma^2)$.

(i) Design a test to make this check. If indeed $y_{new} \sim \mathcal{N}(\mu, \sigma^2)$, then your test can declare the other conclusion with probability at most 0.05. You may express your answer using the quantile function of any of the distribution we have seen in class.

(ii) Suppose that we concluded that $y_{new} \sim \mathcal{N}(\mu, \sigma^2)$. Find a 0.95 confidence interval for y_{new}.

(iii) Suppose that now we have two new measurements y_{new1}, y_{new2}. We assume that both new measurements follow the same normal distribution, but we are unsure whether it is the same distribution as that of $y_1, \ldots, y_n$. Propose a test to check this.

Answer: (i): We set H_0 as the hypothesis that $y_{new} \sim \mathcal{N}(\mu, \sigma^2)$. Under H_0,

$$y_{new} - \bar{y} \sim \mathcal{N}(0, \sigma^2(1 + 1/n)),$$

so

$$t = \frac{y_{new} - \bar{y}}{s\sqrt{1 + 1/n}} \sim t_{n-1}, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Our test rejects H_0 at significance level $\alpha = 0.05$ if $|t| > t_{n-1}^{0.975}$

(another acceptable answer is to incorporate y_{new} into s^2. However, this may inflate s^2 under the alternative hence result in a test of lesser power)

(ii) $y_{new} \in \bar{y} \pm s\sqrt{1 + 1/n} \cdot t_{n-1}^{975}$

(iii) The best option is to use the two-sample t-test with $n_1 = n$ and $n_2 = 2$. In this case,

$$t = \frac{\bar{y}_{new} - \bar{y}}{s\sqrt{\frac{1}{2} + \frac{1}{n}}}, \qquad s^2 = \frac{1}{n + 2 - 2}\left(\sum_{t=1}^{n}(y_i - \bar{y})^2 + (y_{new1} - \bar{y}_{new})^2 + (y_{new2} - \bar{y}_{new})^2\right)$$

We reject the null if $|t|$ exceeds $t_n^{0.975}$.

Another option is to use $\bar{y}_{new} = (y_{new1} + y_{new2})/2$ in a procedure similar to (i). With this option, under the null we have

$$\bar{y}_{new} - \bar{y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{2} + \frac{1}{n}\right)\right),$$

so

$$t = \frac{\bar{y}_{new} - \bar{y}}{s\sqrt{\frac{1}{2} + \frac{1}{n}}}, \qquad s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_i - \bar{y})^2.$$

We reject the null if $|t|$ exceeds $t_{n-1}^{0.975}$.

3. (20 points) We are interested in monitoring air quality in Tel-Aviv area. We propose a linear response model for fine particulate matter levels based on the following predictors that we can measure: time-of-day, temperature, wind intensity.

   - Introduce and define the corresponding linear model notation. Assume that we have measurement of fine particulate matter levels and wind intensity over different days and times of day.
   - Suppose that we suspect that wind direction information can improve our prediction. Given wind direction measurements, propose a method to verify whether to include wind direction in your model.
   - Suppose that in the original (smaller) model, the estimated LS coefficient $\hat{\beta}_{\text{wind}}$ corresponding to wind intensity turned out to be insignificantly different than zero. Is it possible that $\hat{\beta}_{\text{wind}}$ is significantly different than zero in the larger model? explain.