Homework Zero

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

Complete before the first class, do not *turn in*

This "homework" is intended as a self test of your knowledge of the basic statistical concepts and calculations that are required for 41100 - Applied Regression Analysis. The prerequisite for my class is a thorough understanding of the material covered in 41000 or its equivalent. The problems below will help you gauge your readiness for 41100. If this "homework" is at all challenging or the concepts unfamiliar, aside from problems marked with a $\star$, you should carefully consider your options.

The problems marked with a $\star$ are more difficult conceptually and are beyond the scope of a basic statistics course, but connect the more basic ideas to concepts covered in 41100.

The prerequisite is not strictly enforced by the registration system, but be warned that lecture material, assignments, and exams will assume you have this knowledge and additional help/tutoring on these topics will not be provided. This course moves quickly to more advanced material.

This "homework" is not to be turned in and will not graded. You should complete it before the first class and decide if this course is right for you. Solutions are not available. What ends up happening is that people look at the answers and think, "oh yeah, I knew that" without really trying and get a false sense confidence. Either you feel confident in your understanding of the material or you don't; a solution set won't change that.

The problems are grouped into five loose sections, which overlap in material and are not in any particular order.

1 Summation Notation

(a) Refer to this table to answer the questions below.

i	1	2	3	4
Z_i	2.0	-2.0	3.0	-3.0

(i) Compute $\sum_{i=1}^{4} z_i$

(ii) Compute $\sum_{i=1}^{4} (z_i - \bar{z})^2$

(iii) What is the sample variance? Assume that the z_i are i.i.d.. *Note that i.i.d. stands for "independent and identically distributed".*

(b) Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample (independent and identically distributed) from some distribution with mean μ and variance σ^2. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Carefully prove that:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 + (\overline{X} - \mu)^2.$$

(c) For a general set of n numbers, $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_n\}$ show that

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i - \bar{X})Y_i$$

2 Normal Distribution

(a) Suppose that $X \sim \mathcal{N}(-10, 25)$, i.e., X has a Normal distribution with a mean of -10 and variance of 25.

(i) Compute: $\mathbb{P}[X > -10]$, $\mathbb{P}[X < -20]$, and $\mathbb{P}[X = 0]$.

(ii) Find $\mathbb{P}[-22 \leq X \leq -12]$ by expressing it in terms of Z, the standard Normal random variable.

(iii) Find two numbers z_1 and z_2 such that $\mathbb{P}[|X| \leq z_1] = 0.95$ and $\mathbb{P}[|X| \leq z_2] = 0.90$, respectively.

(b) A random sample of size 100 is taken from a Normal population with mean $\mu = 50$ and standard deviation $\sigma = 5$. What is the probability that the *mean* of the sample will:

(i) exceed 51?

(ii) fall between 49.5 and 50.5?

(iii) be less than 48?

3 Sampling Distributions

(a) Suppose we have a random sample $\{Y_i, i = 1, \ldots, n\}$, where $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 4)$ for $i = 1, \ldots, n$.

(i) What is the variance of the sample mean?

(ii) What is the expectation of the sample mean?

(iii) What is the sampling distribution of the sample mean?

(iv) What is the variance for another i.i.d. realization Y_{n+1}?

(v) What is the standard error of $\bar{Y}$?

(b) ($\star$) Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample (i.i.d.) from a Normal distribution with mean 5 and variance 1. Define the set of random variables $\{Y_1, Y_2, \ldots, Y_n\}$ by

$$Y_i = \begin{cases} 1 & \text{if } X_i > 6.96 \\ 0 & \text{if } X_i \leq 6.96. \end{cases}$$

Find the sampling distribution of $W = \sum_{i=1}^{n} Y_i$.

4 Marginal and Conditional Moments/Distributions

(a) Suppose that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, $\text{var}(X) = \text{var}(Y) = 1$, and $\text{corr}(X, Y) = 0.5$. Compute:

(i) $\mathbb{E}[3X - 2Y]$

(ii) $\text{var}(3X - 2Y)$

(iii) $\mathbb{E}[X^2]$

(b) Suppose $X \sim \mathcal{N}(2, 3)$, $U \sim \mathcal{N}(0, 1)$, and $\text{corr}(X, U) = \rho$. Define $Y = 3X + U$. Separately for both $\rho = 0$ and $\rho = 0.3$, compute the following and comment on how your answer changes with ρ and why.

(i) $\mathbb{E}[Y]$

(ii) $\mathbb{V}[Y]$

(iii) $\mathbb{E}[Y|X]$

(iv) $\mathbb{V}[Y|X]$

(c) Below are the midterm and final exam scores of 20 students in 41100. Scores are out of a total of 100. Use this data to answer the questions below.

Student:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Midterm	66	90	85	65	84	83	47	84	80	92	76	62	87	64	74	57	72	84	87	49
Final	65	87	70	74	74	82	65	84	76	79	66	69	93	56	84	69	70	97	81	66

(i) What is the average final exam score for students scoring at least 80% on the midterm?

(ii) Scores above 85% were assigned an A. What was the probability of getting an A on both exams?

(iii) For students that got an A on the midterm, what was the probability of getting an A on the final?

(iv) Using M to denote midterm score and F to denote final score, write what the above parts of this question are asking for in conditional probability/expectation notation. For example, " the average final exam score for students scoring at least 80% on the midterm" would be $\mathbb{E}[F \mid M \geq 80]$.

(d) You are the proud owner of eight McDonald's franchises in the suburbs of Chicago. You decide to do a little experiment by setting the price of a Happy Meal across the restaurants. (Assume that demographics and clientele are similar for each franchise.) You gather data on the number of Happy Meals sold at each franchise during a week of the pricing experiment.

franchise (i)	1	2	3	4	5	6	7	8
`price` (\$)	1.5	1.5	1.75	2.0	2.0	2.25	2.5	2.5
`sales`	420	450	420	380	440	380	360	360

(i) Ignore `price`. If the `sales` are i.i.d. with mean 390 and variance σ^2, what would you estimate for σ^2?

(ii) Now, assume that you model `sales` as independently distributed with variance σ^2 and mean $\mathbb{E}(\texttt{sales}_i) = 500 - 60 \cdot \texttt{price}_i$. What would you estimate for σ^2? By comparison to your estimate in (i), what does this say about this model?

(iii) Find the correlation between `price` and `sales`.

5 Hypothesis Testing and Confidence Intervals

(a) Suppose we sample $\{Y_i, i = 1, \ldots, n\}$, where $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$, and that you want to test the null hypothesis $H_0 : \mu = 10$ *vs* the alternative $H_a : \mu \neq 10$, at the 5% level.

(i) What test statistic would you use? How do you estimate σ?

(ii) What is the distribution for this test statistic *if* the null is true?

(iii) What is the distribution for the test statistic if the null is true and $n \to \infty$?

(iv) Define the test rejection region. (I.e., for what values of the test statistic would you reject the null?)

(v) How would compute the p-value associated with a particular sample?

(vi) What is the 95% confidence interval for μ? How should one interpret this interval?

(vii) If $\bar{Y} = 11$, $s_y = 1$, and $n = 9$, what is the test result? What is the 95% CI for μ?

(b) Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of coin flips, where θ is the probability of getting heads.

(i) Appeal to the Central Limit Theorem to construct a level α test of $H_0 : \theta = 1/4$ versus $H_1 : \theta \neq 1/4$. (Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject).

(ii) Appeal to the Central Limit Theorem to construct a $1 - \alpha$ confidence interval for θ.

(iii) Referring to the part above, how many coins must be flipped so that the interval has width less than 0.2, regardless of the true value of θ? What happens your answer as $\alpha \to 0$?

(iv) ($\star$) Suppose that $n = 5$ and all 5 coins have come up tails. For this data, conduct the test from part (i) and compute the confidence interval of part (ii). What do these results tell you?

(c) Suppose $X \sim \mathcal{N}(2, 3)$, $U \sim \mathcal{N}(0, 1)$, and $\text{corr}(X, U) = 0$. Define $Y = 3X + U$.

(i) What is the distribution of Y given X?

(ii) ($\star$) Find a 95% prediction interval for Y given X (a "confidence" interval that contains $Y|X$ 95% of the time).

(d) The table below shows 41100 midterm and final grades for five students from the full-time program and five from the evening/weekend program. Scores are out of a total of 100. Use this data to answer the hypothesis testing questions below. When answering all questions below, explicitly write out the null and alternative hypotheses, give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject. Carefully show how you compute each required piece of the test statistic.

Full-time		Evening/weekend	
Midterm	Final	Midterm	Final
61	91	91	85
71	87	65	53
56	82	84	74
47	75	84	84
83	85	91	77

(i) Appealing to the Central Limit Theorem, conduct a test of the hypothesis that the average midterm score of full-time students is different from the average midterm score of evening/weekend students.

(ii) Repeat the above for the final exam.

(iii) Appealing to the Central Limit Theorem, conduct a test of the hypothesis on average full-time students performed better on the final than the midterm.

(iv) Now assume that test scores are Normally distributed. Re-do parts (i) and (iii) without using the central limit theorem, but instead constructing exact finite sample tests. What about your results change? What does this tell you?

(e) ($\star$) Suppose you observe a random sample $X_i, i = 1, \ldots, n$ from a Normal population with unknown mean μ_X and known variance equal to 1. Consider testing $H_0 : \mu_X = 0$ versus $H_1 : \mu_X \neq 0$. Use $\bar{X}$ as an estimator for μ_X.

(i) Construct a 5% level test based on the t statistic. (Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject).

(ii) Suppose now you observe *another* sample $Y_i, i = 1, \ldots, n$ from a Normal population with unknown mean μ_Y and known variance equal to 1, independent of the first (note the sample sizes are the same, but the means may be different). Let $\bar{Y}$ be an estimator for μ_Y. Construct a 5% level test based on the T statistic for $H_0 : \mu_Y = 0$ versus $H_1 : \mu_Y \neq 0$. (Give the test statistic, its distribution, the critical region, and describe when you reject the null and when you fail to reject).

(iii) Each test in parts (i) and (ii) are 5% level tests. Find the probability that at least one of the tests gives a false positive.

(iv) I would like to test the null hypothesis

$$H_0 : \mu_X = 0 \text{ AND } \mu_Y = 0$$

against the alternative

$$H_0 : \mu_Y \neq 0 \text{ OR } \mu_Y \neq 0.$$

I will reject the null if $\bar{X} > c$ OR $\bar{Y} > c$. Find the value of c so that this is a 5% level test.

(v) Part (iv) refers to testing two restrictions. What happens to the value of c as the number of restrictions grows, but the level stays fixed at 5%?