PHP2510 FINAL STUDY GUIDE

Topics Covered: Basic Probability, Random Variables, Diagnostic Testing, Exploratory Data Analysis w/ R, Sampling Distributions, Hypothesis Testing, Confidence Intervals, Non-Parametric Testing, Categorical Data Comparisons, Linear Regression

You may not use a computer, calculator, nor any reference materials (apart from what is provided in this exam). The exam is designed to take 2 hrs, but up to 3 hours will be given.

Grading

This exam is graded out of 150 points (although there are 159 points available).

All subquestions are worth 3 points, awarded as follows:

- 3 points if fully correct, or with very minor flaws in arithmetic
- 2 points for an attempt that contains some flaws but shows comprehension of most core statistical concepts
- 1 point for an attempt that contains major flaws or misunderstanding of most core statistical concepts
- 0 points for no attempt whatsoever

Please be clear and concise; any explanation can be done in 3 sentences or fewer. 1 point may be subtracted for excessive information that does not address the question. Please use accurate notation and answer in mathematical terms when possible.

Structure

The exam is organized into 4 parts:

CONCEPTS: assesses your understanding of statistical concepts via short-form answers
DERIVATIONS & FORMULAE: assesses your ability to use and explain common formulae
END-TO-END ANALYSIS: assesses your ability to solve multi-part word problems
INTERPRETING R CODE: assesses your ability to interpret R code and output

Note: the +9 points of "extra credit" problems are sprinkled throughout and not explicitly labelled as such.

A table of common distributions with their facts will be provided.

Resources

Sick of SPEEGLE and CHIHARA? Consider reading:

- https://mixtape.scunning.com/02-probability_and_regression: 2.1-2.10 (2.11-2.16 is optional and relevant, but also has technical derivations that you are NOT expected to

master), is a nice recap of some of the foundational probability concepts we did early in the class
- https://dev.peopleanalytics-regression-book.org/linear_regression.html is a great OLS regression reference
- Review this post for all "fact fluency" terms

Helpful Hints

- There are no regression formulae you are expected to memorize or utilize (in terms of how to estimate parameters), but you need to know: how to write out the model being fit (see slides) and convert that to R code (or vice-versa per this post); how to use output to create confidence intervals or perform formal tests, etc.
- The test has a few problems about rank statistics (which were in scope for Exam #2, but minimally prevalent in the actual exam). Make sure you review that content.

Additional Practice Problems

Q1-Q8 are generally more focused on recent material and word-to-notation long-form problems. Q9+ are generally more focused on previous material, especially distributions and foundational probability.

Q1. You are studying whether a new sleep medication is effective. You recruit 100 patients for an experiment: For one week, they take medication at 10pm each night and go to bed with a monitoring device. Via the device, you record the total number of days each patient entered REM sleep before 2am. There are two medications (treatment vs control) being studied, and patients are randomized to one of them (for all 7 days). The outcome for each patient may be a number 0-7. You assume the data comes from a Binomial(7,p_T) distribution for each patient in the treatment group; Binomial(7,p_C) for each patient in the control group.

Your colleague loves regression; he borrows your data and runs the following code:

Call:
lm(formula = REM_days ~ as.factor(treatment), data = data)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.4000	0.1400	31.43	< 2e-16
as.factor(treatment)TRUE	0.9300	0.6200	1.500	0.1368

Residual standard error: 1.4 on 98 degrees of freedom

1. Write out the model being fit
2. Given the problem set-up, which of the following classic OLS regression assumptions are violated? Select all that apply and explain why you selected them.
    a. Normality

    b. Linearity
    c. Homoscedasticity
    d. Independence
3. What does the statistical significance of the intercept tell you? Interpret it within the context of the problem.
4. What does the lack of statistical significance of the as.factor(treatment)TRUE coefficient tell you? Interpret it within the context of the problem.

Q2. [CHALLENGE] You are studying whether a new sleep medication is effective. You recruit 100 patients for an experiment: they take medication at 10pm each night and go to bed with a monitoring device. Via the device, you record whether each patient entered REM sleep before 2am (a yes/no variable, coded 1/0). You follow up a month later to collect the devices, and realize you forgot to tell the patients to wear the device EVERY night. So, of course, each patient has a different number of days (which is known via the device settings) of actual data collection. In this problem (unlike in others, but now more realistic), you allow each patient to have their own underlying probability of REM sleep given the medication. You still want to estimate, at the population level, the average effectiveness of the drug (in terms of how likely it enables the average people to enter REM sleep within 4 hours).

1. What statistic would you use, despite this data collection issue? Hint:
    a. Define the data generation process for each patient
    b. Find an unbiased statistic for the relevant parameter for each patient.
    c. Combine (b) across 100 patients.
2. A different researcher analyzes your data in a simpler way: They take the total number of "yes"s (across all 100 patients) and divide by the total number of days of data observation (across all 100 patients). Explain when and why this would be biased. Hint: think about the data collection issue, and what patient behavior needs to happen to systematically taint our sample.

NOTE: In my opinion, this problem is more challenging than anything on the final. But it is a good example of making sure you are comfortable with word problems, distributions, parameters, estimators, and notation.

Q3. You are studying the recovery time for a specific severe bacterial infection. You collect the length of hospital stay (in hours) from a sample of n patients who were admitted with the infection and treated with a new protocol. Because hospital stay is a continuous, non-negative outcome that often has a right-skewed distribution, it can be modeled by an Exponential distribution (assume this fits the data well). You want to perform inference on the true average length of hospital stay in hours for the population of patients with this infection under the new protocol.

Use the fact that Exp(λ) has expected value $1/\lambda$ and variance $(1/\lambda)^2$

1. Create notation for this problem. Make sure you clearly define:
    Your outcome variable and its theoretical distribution.
    The population parameter of interest (that you want to estimate).
    The data you actually collected.

2. Why would the sample mean be a reasonable statistic to use in this setting? State in one sentence.
3. Using the central limit theorem, what is the approximate distribution of your sample mean?
4. Construct a 95% confidence interval for the true average length of hospital stay (in hours), using 1.96 as your critical value. Make sure to express it without any unknown parameters in the equation.
5. You report a confidence interval of [100, 150]. Your colleague collects data from a different hospital, for the same research question. They report their CI for the average length of hospital stay, but in days (not hours), as [4.1, 6.3]. What do you conclude, if anything?
6. The other hospital gives you the data, but it's corrupted - every 10th row is missing! Describe a scenario where this corruption is <u>not</u> problematic for the inference methods that we've learned.

Q4. In the context of regression, explain the following in non-technical language:

1. An interaction term
2. Heteroscedasticity
3. Residuals
4. Extrapolation

Q5. Do diagnostic checks on an OLS regression prove that your model assumptions are correct? Why or why not?

Q6. Consider the following R code:

```
m1 <- lm(Y ~ X1 + X2*X3, data = data)
m2 <- lm(Y ~ X1 + X2, data = data)
```

1. Are the two models nested?
2. True or False: you use the anova() command to formally test the fit of two nested models, but AIC to compare two non-nested model.

Q7. A researcher has two continuous variables and decides to 'bucket' (categorize) each of them and perform a chi-squared test on the resulting contingency table. Why is this not recommended?

Q8. Consider the following two statements:

- About half of all published research is wrong (it claims HA, but H0 is true)
- Results with p-values < 0.05 always get published

Do these statements contradict each other (even without "cheating" or "p-hacking" – assume the reported p-values are "correct" when answering the question)? Why or why not?

Q9. Your town requests, starting Jan 1, for each urgent care to record the number of patients treated until a specific rare disease is encountered.

1. What is a reasonable distribution to use to model this?

2. How would you use the data to estimate the prevalence of the rare disease among urgent care patients?
3. How would you decide if your choice of (1) is appropriate for the data collected?

[Q10 - Q12] Each sample of water has a 15% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Let X = the number of samples that contain the pollutant in the next 20 samples analyzed.

- Question 10. Specify the distribution of X (state the distribution name and parameter values)
- Question 11. Find the probability that in the next 20 samples, exactly 3 contain the Pollutant.
- Question 12. Find the probability that $X \geq 2$.

Question 13. How many ways are there to fill out a 20 question True/False Exam, assuming you do not skip any questions?

Question 14. Label each statement True or False

- If two events A and B are independent, then $P(A \mid B) = P(A)/P(B)$.
- If two events A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$
- If two events A and B are mutually exclusive, they must also be independent.

Question 15. What is the definition of p-value?

Question 16. Describe a variable X that would follow a Bernoulli distribution with parameter p=.3. This can be from any field of study.

Question 17. Consider the data set flights which is part of the nycflights13 package in R. This data set contains information on flight departures, delays, arrivals of flights leaving NYC to travel domestically in 2013. Suppose we are interested in the flight patterns of flights headed for Boston Logan Airport. The following tibble describes key descriptive statistics (mean, std deviation, min, max, n) of the variable dep_delay which indicates number of minutes a flight is delayed (values<0 indicate flights that departed early):

tibble: 1 x 5

mean_dd	sd_dd	min_dd	max_dd	n
<dbl>	<dbl>	<dbl>	<dbl>	<int>
8.73	34.8	-23	437	15508

Briefly explain what this tells us about the variable and the shape of the distribution.

Additional Practice Problems (Solved)

Q1

1.1.

$$\text{REM\_days}_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

where T_i equals 1 if observation i is treated; 0 for not treated; epsilon_i is N(0, sigma^2)

1.2.
Normality - our outcome is binomial, not normal
Homoscedasticity - because in a binomial, the variance is a function of the mean, and we have two different means (treatment and control), this is not true
There are no explicit violations for linearity (E[Y | X] is linear in X trivially) or independence.

1.3. We reject the null hypothesis that the control group has a 0 probability of underlying REM sleep success on average

1.4. We fail to reject the null hypothesis that the treatment group has no difference in underlying REM sleep success probability compared to the control group (i.e. we conclude the treatment doesn't work any more than the control)

Q2
2.1.

$$X_i \sim \text{Binomial}(n_i, p_i) \text{ for patient } i$$
$$\frac{X_i}{n_i} \text{ is an unbiased estimator for } p_i$$
$$\frac{1}{100}\sum_{i=1}^{100}\left(\frac{X_i}{n_i}\right) \text{ is a good estimator for the average } p_i$$

2.2. If patients with different underlying p_i wear the device in different amounts (for example, high p_i patients tend to have high n_i), our sample is tainted (the sample of patient-days is not iid). The procedure in (1) deals with it by looking within each patient first. But procedure (2) would "count" patients with high n_i more (and low n_i less), thereby biasing our estimate of the average p.

Q3
3.1.

$$X \sim \text{Exp}(\lambda)$$
$$\mu = E[X] = 1/\lambda$$
$$X_i \text{ iid draws from } X$$

3.2. It is an unbiased and consistent estimator of $1/\lambda$

3.3.

$$\bar{X} \sim N\left(\frac{1}{\lambda}, \frac{1/\lambda^2}{n}\right)$$

3.4.

$$\left[\bar{X} - 1.96\left(\frac{\bar{X}}{\sqrt{n}}\right), \quad \bar{X} + 1.96\left(\frac{\bar{X}}{\sqrt{n}}\right)\right]$$

3.5 You can convert their CI to hours by multiplying by 24, therefore getting [94, 144]. Due to the extreme overlap with ours, we can infer that their data aligns with ours.

3.6 If the ordering of the data is random, this is not a problem, because we can claim we have a (smaller) iid sample so no assumptions are violated.

Note: It is important to see that in this problem we are interested in estimating 1/lambda, not lambda itself (hence we use X_bar, not 1/X_bar).

Q4.
4.1. An interaction term: This is a predictor variable, defined as the product of two other predictor variables. It allows for the effect (on the expected value of the outcome) of one predictor variable to depend on the level of the other predictor variable.
4.2. Heteroscedasticity: The opposite of homoscedasticity. This occurs when the variability of the outcome (for a given X) is not constant (across all X).
4.3. Residuals: The difference between the actual observed value of the outcome (for a given X) and the value predicted by the regression model (for that given X). A residual is the estimate of the true error term.
4.4. Extrapolation: If we use the regression model to make predictions for data points that lie outside the range of the original data used to fit the model, we are extrapolating. This often happens for the intercept.

Q5. No. The diagnostic checks allow us to look for "red flags" - that is, evidence that our assumptions are wrong. But "passing" all diagnostic checks doesn't prove our model is right; it just shows it is plausible.

Q6. Yes; True

Q7. We throw away information and therefore lose statistical power to detect an association

Q8. No - the second statement means that studies that get a low p-value "by chance" make it through the publication filter; what gets published is then a non-representative sample of all research. Therefore, statement 1 can be a natural consequence of statement 2. This is analogous to having low PPV for a rare disease.

Q9. Geometric(p) because each patient (entering the urgent care) can be viewed as a "trial" that is successful if the patient has the rare disease. Because the mean of a Geometric(p) is 1/p (this would be given to you on the exam), I would use 1 / X_bar (that is, the reciprocal of the average

patients counted across the urgent cares) to estimate the underlying prevalence. I would use a Chi-squared goodness-of-fit test to determine if the Geo(1/X_bar) fit the data well.

Q10: X~Binomial(20, 0.15)

Q11 [the binomial pdf will be given to you, but you need to know how to use it]
P(X=3) = (20!/(3!*17!))*(0.15)^3*(0.85)^17 (≈ 0.24)

Q12 [the binomial pdf will be given to you, but you need to know how to use it]
P(k>=2) = 1 - P(k<2) =1−[P(X=0)+P(X=1)]
= 1 - [(20!/(0!*20!))*(0.15)^0*(0.85)^20 + (20!/(1!*19!))*(0.15)^1*(0.85)^19]
(≈ 1- (0.04 +0.14) ≈ 0.82)

Q13. 2^20

Q14: F, T, F

Q15: The p-value is the probability under the null hypothesis of obtaining a test statistic at least as extreme as the one obtained.

Question 16: Let X be an indicator for whether a randomly selected patient misses their scheduled medical appointment. Then X=1 if the patient misses the appointment and X=0 otherwise, and suppose the probability of missing the appointment is 30%. Since this describes a single trial with two outcomes, X follows a Bernoulli(0.3) distribution.

Q17: The mean delay is about 8.7 minutes, but the standard deviation is large (34.8 minutes), indicating substantial variability in delay times. The minimum value is –23 minutes, meaning at least one flight left early, while the maximum delay of 437 minutes shows that at least one flight was extremely late. Together, these values suggest that the distribution of dep_delay is right-skewed.