PHP2510 EXAM #2 STUDY GUIDE

Topics Covered: Sampling Distributions & the Central Limit Theorem, Hypothesis Testing, Experimental Design, Confidence Intervals, Non-Parametric Tests, Categorical Comparisons

You may not use a computer, calculator, nor any reference materials. You have the full 80 minutes to complete the exam.

Grading

The exam is graded out of 50 points (although there are 58 points available).

Unless otherwise noted, questions 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

When questions are worth 1-2 points, half-credit may be given for answers that show reasonable levels of statistical comprehension.

When asked for an explanation, you should be clear and concise. 1 point may be subtracted for superfluous information that does not address the question.

Structure

The exam is organized into 5 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
- EXTRA CREDIT: additional challenge questions

Resources

Be sure to:

- Review lecture notes/slides, including the learning objectives for each week; I followed these when designing the exam
- Review labs (although coding in R is not the focus of the test, the labs highlight concepts that will be prominent)
- Review Ed Discussion posts

- Review the readings
- Review the handouts
- Utilize practice problems provided by the book, especially the ones I've suggested on Ed Discussion
- Utilize Week 11 Lab as office hours to ask any additional questions
- Study the concepts highlighted below

Fact Fluency (non-exhaustive)

W1-6

Per the original Ed Discussion Fact Fluence post

- Probability (notation and definition)
- Conditional probability (notation and definition)
- Law of total probability (and moving from marginals to joints)
- Independent events
- Sensitivity, specificity, and positive predictive value of a test
- Random variable (with proper notation)
- Realization of a random variable (with proper notation)
- An iid sample (with proper notation)
- Distribution
- Parameter (of a distribution; of a population)
- A pdf / pmf / cdf
- Statistic (generally, and as an estimator)
- Expectations (and comfort with integrals / summations; linearity of expectations)
- Variance / standard deviation / standard error / standard error of the mean
- Central limit theorem (generally, and applied to different assumptions about the population)
- Standardizing a normal random variable

Hypothesis Testing & Experimental Design

- Overarching logic
- Test statistic
- Null vs alternative hypothesis
- Alpha and beta
- P-value
- Statistical significance
- Rejection region and critical value
- T-tests, z-tests, proportions tests (and how population assumptions create these)
    - Degrees of freedom
    - Pooled variances
    - Use of CLT
    - Core assumptions
    - One sample vs two sample
- Using s^2
- Power

- Effect size
- Relationships between alpha, beta, power, sample size

Confidence Intervals

- Motivation
- Interpretation
- General Form & symmetry
- Assumptions
- Relationship to hypothesis tests and p-values
- Margin of Error

Non-Parametric Tests

- Bootstrapping
- Permutation Tests
- Rank tests
- Reasons for using
- Assumptions (or lack thereof)
- Relationship to hypothesis testing

Analysis of Categorical Variables

- Contingency table
- Chi-squared distribution
- Main tests (independence, homogeneity, model fit)
- General Formula (including interpretation)
- Degrees of freedom

A Few More Practice Problems

Use AI to help you study! Consider using the following prompt in Gemini (or whatever tool you prefer). If you are unsure if the question is relevant or accurately answered by AI, post on Ed Discussion and I will comment.

I am an ivy-league student taking an advanced graduate course in biostatistics. I need a practice exam for my second mid-term. Follow these parameters:
~ Topics Covered: Sampling Distributions & the Central Limit Theorem, Hypothesis Testing, Power Calculations, Confidence Intervals, Non-Parametric Tests, Categorical Comparisons
~ The exam will be split into 5 components:
CONCEPTS: assesses my understanding of statistical concepts via short-form answers
DERIVATIONS & FORMULAE: assesses my ability to use and explain common formulae
END-TO-END ANALYSIS: assesses my ability to solve multi-part word problems
INTERPRETING R CODE: assesses my ability to interpret R code and output
EXTRA CREDIT: additional my questions
~ My professor likes to teach from a theoretical perspective. I won't be allowed a calculator nor cheat sheet. I won't be expected to do distribution lookups. I may be given formulae and asked to use or interpret them. We typically need to use mathematical notation (i.e. probability statements). The professor also likes to give multi-step problems that get progressively harder.

I originally had most of these questions picked out for the labs, but removed them for timing reasons. Consider them as additional practice problems.

CHIHARA 8.16

According to a survey, 26% of residents of Illinois have completed high school. A professor thinks his county might be lower, so creates his own survey; 69 of 310 have completed high school in his survey. Are the data consistent with the professor's hypothesis?

SPEEGLE 8.22

148 subjects' body temps were measured 1-4 times daily for 3 consecutive days, for 700 observations. Does the standardized typical test statistic follow a t-distribution?

SPEEGLE 8.12

You have a random sample of 20 observations from a normal population with unknown mean and sd. Suppose x_bar = 2.3 and s = 1.2
Find the 98% confidence interval for the mean
What confidence interval is (2.0, 2.6)?

Backup Exam #1 Problem (updated)

Suppose a researcher has collected data on alcohol use from 100 random people on a given day. This table summarizes the data collected.

Drinks	Frequency
0	50
1	25
2	10
3	8
4	4
5+	3

1. This researcher decides to use the Poisson(λ) distribution to model her data. See below for Poisson distribution details.
    a. Why is the Poisson distribution reasonable?
    b. What statistic would you suggest using to estimate λ? Why?
    c. Using the Central Limit Theorem, what is the approximate distribution of the sample mean?
2. What is your estimate of λ? Calculate exactly.

3. What numbers would you compare to decide if that fits the data well for the non-drinking population?
4. How would you perform a formal test or whether the model fits all the data? Please show the formula you would use, the numbers that would enter into the formula, and how you'd make a final decision
5. You discover that the researcher's data is from pairs of married couples. Would that concern you? Why or why not?

For X ~ Poisson(λ):

- $f(x) = \lambda^x e^{-\lambda}/x!$ for x = 0, 1, 2, …
- $E[X] = \lambda$; $V[X] = \lambda$

Practice Problems ANSWERS

CHIHARA 8.16

We can form a confidence interval with the standard equation: statistic +/- crit_value * se
statistic = 69/310
crit_value = 1.96
se = sqrt(p*(1-p)/n), where we plug in 69/310 for p and 310 for n

```
p.hat <- 69/310
p.hat + c(-1,1)*1.96*sqrt(p.hat*(1-p.hat)/310)
```

26% is within the CI, so the data is consistent with the original survey. However, how this question is phrased is a bit unique. It asks if the data is consistent with the professor's hypothesis. That is true too, since our estimate is lower than 26%. So the data is consistent with both!

SPEEGLE 8.22

No. There is correlation within subjects, so our "iid" assumption is incorrect and the test statistic does not have a t-distribution.

SPEEGLE 8.12

The CI is in the form: x_bar +/- crit.value * se.
For (a): The standard error is s/sqrt(20); the critical value is the 99th quantile of a t-distribution with 19 df

```
> 2.3 + c(-1,1)*qt(.99, 19)*1.2/sqrt(20)
[1] 1.618585 2.981415
```

For (b), we know the 0.6 = 2*crit_value*1.2/sqrt(20); that means that the critical value is 1.118.

```
> pt(1.118, 19)
[1] 0.8612479
```

So we used the 86th quantile of the t-distribution, meaning alpha/2 = 0.14. So we have a 1-0.28 = 72% CI.

Backup Exam #1 Problem (updated)

1a. The outcome variable is a count of a relatively rare event over a given time period

1b. The sample mean because it is an unbiased estimator of λ (per Poisson distribution facts, the population is centered at λ; the sample mean has the same center as the population)

1c. The sample mean is approximately N(population mean, population variance / n) = N(λ, λ/n)

2. (0*50 + 1*25 + 2*10 + 3*8 + 4*4 + 5*3)/(50+25+10+8+4+5) = (25+20+24+16+15)/(50+25+10+8+4+3) = 100/100 = 1

3. Per the Poisson distribution facts, f(0) = e^(-1). Compare this number against the data collected for the non-drinking population, where p= 50/100.

4. Sum((O-E)^2 / E); we take 6 different sums; O == count from the table; E == 100*Poisson(1) pdf evaluated at the value (0-5). We compare the total to a chi-square distribution with 4 df

5. Yes this is concerning. Our analysis technique assumes iid data, and married couples would have correlated drinking habits.