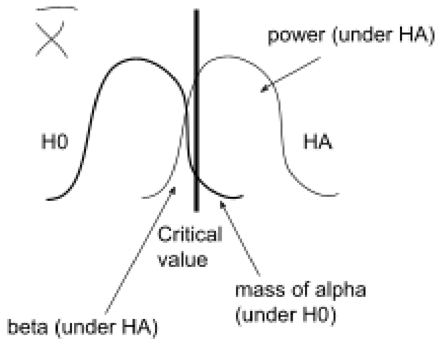
PHP2510

EXAM #2 ANSWER KEY

Section 1: Concepts (15 pts)

1.1. What is the core difference in underlying assumptions between prop.test and t.test?
Prop.test assumes the outcome of interest, for the population, follows a Bern(p) distribution, while t.test makes no distributional assumptions on the outcome, apart that it has some true mean and variance.

1.2. Draw a picture that shows how relaxing your alpha-level (from 0.05 to 0.10, perhaps) increases power. Briefly explain your picture.



“Relaxing” the alpha level moves the critical value to the left (calculated under H0). That increases power and decreases beta (the probability of a type 2 error), which are calculated under HA

1.3. How do you make a rejection decision for a particular null hypothesis if only the confidence interval is reported?
If the null hypothesis’ parameter value is within the confidence interval, we do NOT reject that null hypothesis. If the null hypothesis’ parameter value of interest is NOT within the confidence interval, we DO reject that null hypothesis.

1.4. Explain why non-parametric approaches are useful for highly skewed data. Describe a situation with, or what is meant by, 'highly skewed' data.
They avoid making distributional assumptions, or invoking the CLT, which happens more slowly for skewed data. “Highly skewed” means the data is non-symmetric. One common example of this is with count data, as some subjects may have particularly large counts, but counts under 0 are impossible.

1.5. A 95% confidence interval should have 95% coverage. What does this mean?
If we think about repeating the process (sampling and CI construction) many times, the 95% confidence interval should contain the true parameter value (CI “covers” it) 95% of the time.

Section 2: Derivations & Formulae (15 points)

2.1. A Poisson(λ) random variable has mean = variance = λ. If you are willing to assume you have an iid sample from this distribution (as the population outcome), what would be the form of the confidence interval? *Hint: make sure your final answer does not require knowing λ, since that is an unknown value we are trying to estimate. You may use a z statistic for your critical value.*

$$\mathbf{\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}}$$

2.2. You typically need to 4x your sample size to cut your confidence interval width in half. Identify a situation when this holds true, and show how we get this exact result.

The standard confidence interval for a one-sample group means t-test has the form:

$$\mathbf{\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)}$$

That means the total width is:

$$\mathbf{2 * t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)}$$

Because this has sqrt(n) in the denominator, the result holds.

2.3. The following formula is used for a test of independence (or homogeneity) between two categorical variables. Assume you need to explain this formula to a colleague that has little background in calculus or statistics. Define each term ($\sum$, "Observed", "Expected", "~", χ^2_{df}) and how it works to provide evidence about relevant research question.

$$\sum \frac{(Observed\ -\ Expected)^2}{Expected} \sim \chi^2_{df}$$

We create a table of counts for the two categorical variables we are interested in. For each cell in the table, we take:

- **Observed: The actual count (for that cell) in our dataset**
- **Expected: The count (for that cell) that we would expect if we did have independence between the two categorical variables**

The difference between these two numbers, squared and standardized by the "Expected" count, gives us evidence: the larger this number is, the more unlikely that independence actually does hold in the population that we sampled from. We add up those values across all the table cells (that is the meaning of the summation), and see if that is "too large", using a specific reference distribution (that is the meaning of ~ chi^2_df), because statistical theory says this is how the value should behave if the two categorical variables really were independent.

2.4. Review the two formulae below. Describe the setting when these arise. When would you use the latter instead of the former?

$$\frac{\overline{X_1} - \overline{X_2} \ - \ (\mu_1 - \ \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim \ t$$

$$\frac{\overline{X_1} - \overline{X_2} \ - \ (\mu_1 - \ \mu_2)}{s_p \ \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \ t$$

These are the standardized test statistics for a two sample t-test that is used for determining whether two group means are different. The second formula is used when you are willing to make an assumption of equal variances between the two groups, so you use the pooled standard error.

2.5. When performing multiple hypothesis tests, the "global type 1 error rate" is defined as the probability of making any errors under the global null hypothesis (that all individual null hypotheses are true). Write a formula that expresses the global type 1 error rate (α_G) in terms of the individual tests' type 1 error rates (α_i). Assume you have k independent tests.

$$\begin{aligned}
\alpha_G &= P(\text{any errors} \mid \text{global null}) \\
&= 1 - P(\text{no errors} \mid \text{global null}) \\
&= 1 - P(\text{no error on test1, test2, } \ldots, \text{ testk} \mid \text{global null}) \\
&= 1 - P(\text{no error on test1} \mid \text{global null}) \times P(\text{no error on test2} \mid \text{global null}) \times \cdots \times P(\text{no error on testk} \mid \text{global null}) \\
&= 1 - P(\text{no error on test1} \mid H_{0,1}) \times P(\text{no error on test2} \mid H_{0,2}) \times \cdots \times P(\text{no error on testk} \mid H_{0,k}) \\
&= 1 - (1-\alpha_1)(1-\alpha_2)\ldots(1-\alpha_k) \\
&= 1 - \prod_{i=1}^{k}(1-\alpha_i)
\end{aligned}$$

Section 3: End-to-End Analysis (12 points)

3.1. [6 points; 1 per subquestion] 16 subjects had their bone density measured before and after a treatment. Your dataset has a mean difference (after - before) of 0.02 g/cm^2 with a sample standard deviation (often called "s") of the differences of 0.04 g/cm^2.

A. Calculate the standard error of the mean difference.
   s/sqrt(n) = 0.04 / sqrt(16) = 0.01
B. Calculate the standardized test statistic for testing the null hypothesis of no change.
   (0.02 - 0) / (0.04 / sqrt(16)) = 2
C. The 99.5th percentile of the t-distribution with 15 df is 2.95. Construct the 99% CI for the true mean difference.
   0.02 +/- 2.95 * (0.04 / sqrt(16)) = 0.02 +/- 0.0295 = [-0.0095, 0.0495]
D. Does the result show evidence of a significant change at α = 0.01? Just state yes or no.
   No [0 is within the CI; 2 is not in the rejection region]
E. Given only 16 subjects, is it true that the data should be normally distributed, or somewhat close to it, for the results to be trustworthy? State yes or no.
   Yes [we are relying on CLT, which is otherwise suspect with small n]
F. This is paired data. Describe in what way that influenced our analysis procedure.

We took the paired differences (within a given subject) at the beginning, turning a two-sample problem (with correlation between groups) into a one-sample problem with independence.

3.2. [6 points, 1 per subquestion] You are studying whether a new sleep medication is effective. You recruit patients to participate in an experiment: **For one week**, they take the medication at 10pm each night then go to bed with a monitoring device. Via the device, you record the total number of days each patient entered REM sleep before 2am. The outcome may be a number 0-7. You want to estimate, at the population level, the effectiveness of the drug in terms of how many days per week it enables people to enter REM sleep within 4 hours. But, of course, you only have your sample (of n patients in the experiment), so you need to perform statistical inference.

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

   Outcome is X ~ Bin(7,p)
   The population parameter of interest is 7p
   X_i ~ iid i=1, …, n draws from X was collected as data

B. Why would the sample mean be a reasonable statistic to use in this setting? State in one sentence.

   It is an unbiased and consistent estimator of 7p

C. Using the central limit theorem, what is the approximate distribution of your sample mean? *Hint: use the fact that a Bern(p) has expectation p and variance p(1-p) – which you proved in homework #1! – or the fact that a Bin(n,p) has expectation np and variance np(1-p). Be careful with your 'n's (there are two <u>different</u> ones floating around in this problem).*

   CLT says that the sample mean is approximately N(population mean, population variance / n) = N(7p, 7p*(1-p)/n)

D. Construct a 95% confidence interval for the true average number of days per week using 1.96 as your critical value. Make sure to express it without any unknown parameters in the equation.

$$\mathbf{\bar{X} \pm 1.96\sqrt{\frac{\bar{X}(1-\bar{X}/7)}{n}}}$$

E. Your boss asks you whether it's likely that the underlying *probability* of the drug working on a given day is 80%. How could you use the existing confidence interval from (D) to answer her question?

   Check if 7*.8 = 5.6 is within the CI. If yes, our data aligns with that underlying probability and we do NOT reject the statement.

F. You are given the data in 0s/1s, with each patient taking up 7 rows. Why would it be inappropriate to perform a proportions test on the data? State in one sentence.

   Sleep data from the same patient is likely correlated, so that analysis would violate the independence assumption.

Section 4: Interpreting R Code (10 points)

4.1. [3 points, 1 per subquestion] Here we will use resampledata3::Bangladesh, a dataset with levels of arsenic, chlorine and cobalt in a sample of 271 wells in Bangladesh. Answer the following questions, using the output from the R code below. You may assume the data is appropriate for the test performed.

A. What is the average chlorine level in the wells?
78.08

B. Identify 1 value of average chlorine levels for which you would reject the null hypothesis (at alpha = 0.05).
Any number less than 52.87263 OR greater than 103.29539

C. Is there statistically significant evidence that wells with Arsenic levels > 22 have different average Cobalt levels than those with Arsenic levels <= 22 (at alpha = 0.05)? State Yes or No and the directionality of the difference if Yes.
Yes, wells with arsenic levels > 22 have higher cobalt levels

```
> d <- resampledata3::Bangladesh
> t.test(d$Chlorine)

        One Sample t-test

data:  d$Chlorine
t = 6.0979, df = 268, p-value = 3.736e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  52.87263 103.29539
sample estimates:
mean of x
 78.08401

> t.test(d$Cobalt ~ d$Arsenic > 22)

        Welch Two Sample t-test

data:  d$Cobalt by d$Arsenic > 22
t = -2.3224, df = 267.94, p-value = 0.02096
alternative hypothesis: true difference in means between group FALSE and group TRUE
is not equal to 0
95 percent confidence interval:
 -0.19542973 -0.01609982
sample estimates:
mean in group FALSE  mean in group TRUE
          0.4516788           0.5574436
```

4.2. [1 point] Mice weigh 340g with sd 30g in one town. Another town will check if their mice weigh more. They plan to sample mice and perform a test at the 0.1 alpha-level. They want to know what n must be chosen to detect an effect with 85% probability, assuming the town has mice that weigh 354g on average. Review the following R code. Is it correct? State Yes or No.

```
> power.t.test(delta = 14, sd = 30, power = 0.85, sig.level = 0.1, alternative = "one.sided",
type = "one.sample")
```

Yes

4.3. [4 points] A researcher is studying mercury contamination in fish. They collect fish from two bodies of water: Area A (n=40) and Area B (n=14). The data is highly right-skewed. The researcher is interested in the ratio of the medians (Median_A / Median_B) as an effect size.

A. [2 points] Give three distinct reasons why a non-parametric analysis would be appropriate here.

   (1) small n; (2) skew; (3) non-typical parameter of interest

B. [2 points] Provide a clear, one-sentence interpretation of the confidence interval from the R code below. What is your conclusion about the comparison between the areas? Be specific about the interpretation (make sure to state more than just the rejection decision).

   With 95% confidence, we believe the true ratio of the medians (area A / area B) of mercury contamination is contained in the interval [0.245, 0.693]. Because only numbers less than 1 are in the interval, we conclude there is statistically significant evidence that Area A has a <u>lower</u> median level of mercury contamination.

```
> # 'mercury_data' is a data.frame with 'mercury' (numeric) and 'area' (factor: "A", "B")
> set.seed(123)
> R <- 2000  # Number of bootstrap replicates
> boot_ratios <- numeric(R) # Vector to store results
>
> # Separate the data
> data_A <- mercury_data$mercury[mercury_data$area == "A"]
> data_B <- mercury_data$mercury[mercury_data$area == "B"]
>
> # Get the sample sizes
> n_A <- length(data_A)
> n_B <- length(data_B)
>
> # Run the bootstrap
> for (i in 1:R) {
+   # 1. Resample with replacement
+   resample_A <- sample(data_A, size = n_A, replace = TRUE)
+   resample_B <- sample(data_B, size = n_B, replace = TRUE)
+
+   # 2. Calculate the statistic for this replicate and store
+   median_A <- median(resample_A)
+   median_B <- median(resample_B)
+   boot_ratios[i] <- median_A / median_B
+ }
>
> # 3. Calculate the 95% percentile interval
> ci_95 <- quantile(boot_ratios, probs = c(0.025, 0.975))
>
> # Calculate and print the observed statistic
> obs_median_A <- median(data_A)
> obs_median_B <- median(data_B)
> print(paste("Observed Ratio:", round(obs_median_A / obs_median_B, 2)))
[1] "Observed Ratio: 0.34"
>
> # Print the final CI
> print("95% Percentile CI:")
[1] "95% Percentile CI:"
> print(ci_95)
    2.5%     97.5%
0.2454593 0.6928071
```

Section 5: Extra Credit (6 points)

5.1. [1 point] In what situation might you have (even extreme) statistical significance, but clinically irrelevant findings?
When n (the sample size) is very large, we can detect very small effect sizes.

5.2. [1 point] A rank test is robust. What does this mean? Please answer in 1-2 sentences.
The test still works appropriately (i.e. the reference distribution is correct) even in the presence of skewed data or outliers. That is, the results don't depend on certain distributional assumptions (like normality) on the outcome variable.

5.3. [4 points; 1 per subquestion] If X ~ exp(λ), then E[X] = 1/λ; V[X] = $(1/λ)^2$. Assume you have a large iid sample from X. Construct a 90% Confidence Interval for λ. *Note: the typical/general confidence interval form is not appropriate because the sample mean is not centered at the population parameter of interest. Follow these steps:*

A. Determine the central limit result applied to $\bar{X}$
B. Update (A) by appropriate re-using $\bar{X}$ in the variance term
C. Using (B), write a relevant probability statement about (standardized) $\bar{X}$; use the fact that the 95th quantile of a N(0,1) is 1.645.
D. Rearrange for λ [*this is somewhat tedious algebra; consider perfecting the rest of the exam before spending lots of time here!*] then report the upper and lower bounds of your probability statement.

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

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

$$\mathbf{P\left(-1.645 \leq \frac{\bar{X} - 1/\lambda}{\bar{X}/\sqrt{n}} \leq 1.645\right) \approx 0.90}$$

$$\mathbf{\left[\frac{1}{\bar{X}\left(1 + \frac{1.645}{\sqrt{n}}\right)}, \frac{1}{\bar{X}\left(1 - \frac{1.645}{\sqrt{n}}\right)}\right]}$$