# Homework 1

Due Thursday April 16th at midnight

This homework covers Lectures 1-4. Please write your solutions in an R Markdown file and submit your pdf output to Gradescope. All code for your solutions should be shown in your pdf file (which it will do by default; just don't turn this off anywhere).

## Introduction

Suppose that you are a data scientist for the UCSD women's basketball team, and Head Coach Heidi VanDerveer wants you to investigate a few statistical questions regarding the performance of one of her players, Redshirt Junior Guard Rosa Smith.

### Question 1

Smith has not historically been a great 3-point shooter (29.6% in 2023-2024, 26.9% in 2024-2025), but this season showed early signs of greater success. You would like to test the following hypotheses:

$$H_0 : p = 0.30$$
$$H_A : p > 0.30$$

where $p$ is Smith's true 3-point shooting percentage. Coach VanDerveer has approached you with this question after 4 games, at which point Smith has taken 9 3-point attempts.

**(a)** You would like to use the standard $\alpha = 0.05$, but you realize that due to the discrete nature of the Binomial distribution, this is not exactly possible. Write code to show that the rejection region corresponding to a significance level that is as close to 0.05 as possible (either above or below it) is: $\{6, 7, 8, 9\}$. Also state what that significance level actually is for this rejection region, and explain to Coach VanDerveer how you found it and what it means.

*Hint*: the `pbinom` function may be useful. Note *very* carefully what happens with the `lower.tail` argument and strictness of inequalities, and the fact that a p-value is defined as the probability of observing a result at least as extreme as the data result.

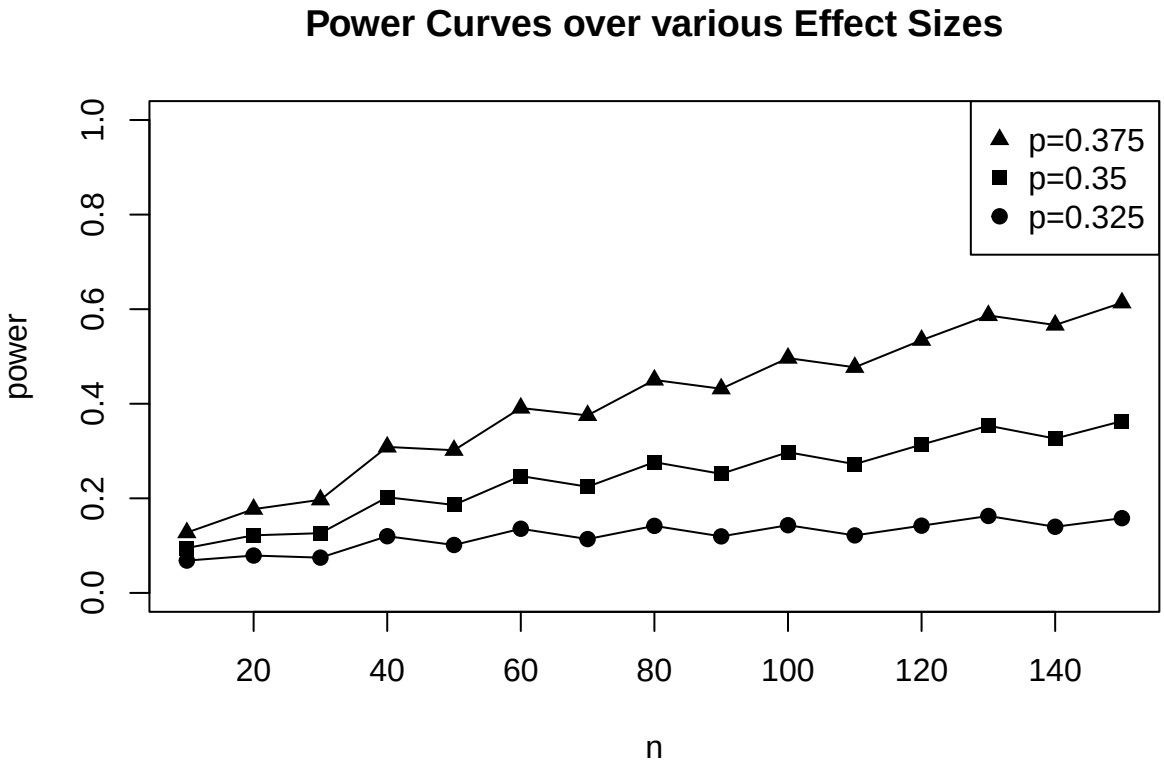
**(b)** With the rejection region from the previous question, suppose that Smith's true 3-point shooting percentage this season is 37.5%. If this were truly the case, find the power to detect evidence of the $H_A$ stated above. Explain to Coach VanDerveer what the result of your power calculation means, whether this seems good or bad, and why.

**(c)** As mentioned above, Coach VanDerveer has approached you after 4 games, at which point Smith has taken 9 3-point attempts. She made 5 of them. Based on this, perform the hypothesis test stated at the beginning of the question with the rejection region stated in part (a). Report the p-value and your conclusions in light of the result from the power calculation in the previous question.

**(d)** Write code to recreate these power curves over a range of effect sizes and sample sizes ranging from 10 to 150. Note that for each sample size, the nominal $\alpha$ level cannot be exactly equal to 0.05 due to the same reason as in Part (a); below, the rejection regions resulting in the significance level that is as close to 0.05 as possible for sample sizes from 10 to 30 has already be calculated; you may copy this into your code (without the quotation marks or the `## [1]`).

```
## [1] "rej_regions_cutoff <- c(6,10,14,17,21,24,28,31,35,38,42,45,48,52,55)"
```

- *Note 1*: The `pbinom` function may again be useful here, but again be very careful about how it handles strictness of inequalities.
- *Note 2*: There was code shown in Lab 2 that you can use as a starting point, and modify as you need to for this situation.




Also in your answer, summarize what we observe in these power curves.

**(e)** Now, a dataset of Rosa Smith's full 2025-2026 seasons statistics are provided as a tab-separated file, `smith.txt`. Load this into your R Markdown file. Then, examine the `3FG/A` column (a snapshot of the first several rows is shown below).

| | Date | Opponent | GS | MIN | FGM/A | %...6 | 3FG/A | %...8 | FTM/A | %...10 | OFF | DEF | TOT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11/07/25 | Denver | NA | 16 | 2-3 | 0.667 | 1-2 | 0.500 | 0-0 | 0.000 | 1 | 3 | 4 |
| 2 | 11/12/25 | Sacramento St. | NA | 10 | 1-1 | 1.000 | 1-1 | 1.000 | 0-0 | 0.000 | 0 | 1 | 1 |
| 3 | 11/16/25 | at San Francisco | NA | 30 | 4-7 | 0.571 | 1-2 | 0.500 | 4-4 | 1.000 | 1 | 4 | 5 |
| 4 | 11/22/25 | Air Force | * | 30 | 2-5 | 0.400 | 2-4 | 0.500 | 3-4 | 0.750 | 0 | 5 | 5 |
| 5 | 11/24/25 | Occidental | * | 23 | 3-7 | 0.429 | 1-4 | 0.250 | 4-4 | 1.000 | 0 | 3 | 3 |
| 6 | 11/28/25 | at Washington | * | 40 | 8-14 | 0.571 | 2-6 | 0.333 | 1-1 | 1.000 | 0 | 1 | 1 |
| 7 | 11/30/25 | at Portland St. | * | 34 | 7-15 | 0.467 | 2-5 | 0.400 | 1-1 | 1.000 | 0 | 5 | 5 |
| 8 | 12/06/25 | Long Beach St. | * | 32 | 11-18 | 0.611 | 4-10 | 0.400 | 1-2 | 0.500 | 2 | 2 | 4 |
| 9 | 12/15/25 | at California Baptist | * | 29 | 5-12 | 0.417 | 2-5 | 0.400 | 0-0 | 0.000 | 0 | 1 | 1 |
| 10 | 12/20/25 | Northern Ariz. | * | 20 | 3-9 | 0.333 | 0-5 | 0.000 | 0-0 | 0.000 | 0 | 3 | 3 |
| 11 | 01/01/26 | Cal Poly | * | 34 | 7-11 | 0.636 | 4-8 | 0.500 | 3-4 | 0.750 | 1 | 4 | 5 |

Notice that the formatting of this column is not quite optimal for doing any analyses with it; for example:

- The value of `1-2` in Row 1 indicates that she made 1 out of 2 attempts in that game
- The value of `1-1` in Row 2 indicates that she made 1 out of 1 attempt in that game

Write code to take the `3FG/A` column and extract the two values separately and store them as numeric values. Then, perform any necessary calculations to do the hypothesis test shown at the top of this question, now with these values from the entire season. Report:

- The sample proportion of 3-point attempts that Smith made in the 2025-2026 season (including the post-season)
- The p-value from the hypothesis test
- A brief summary of your conclusions. Use the standard p-value cutoff of 0.05.

Reminder: all code must be shown for full credit.

### Question 2

With the full dataset loaded in Question 1, we will now investigate the number of assists that Smith had in each game. This is stored in the `AST` column of the dataframe.

**(a)** Plot a histogram of the number of assists that Smith had in each game. You may create it using either base R graphics or `ggplot`, but either way, be sure to adjust axes labels and the title to things that are meaningful and easy to read, and adjust font sizes to be readable if needed as well.

**(b)** We would like to perform a hypothesis test with these data. Noting that the distribution looks fairly skewed, we realize that a t-Test may not be a valid test here, so we will perform a Sign Test with the following hypotheses:

$$H_0 : \tilde{\mu} = 1.5$$
$$H_A : \tilde{\mu} > 1.5$$

where $\tilde{\mu}$ is Smith's true median number of assists. Find the rejection region corresponding to the significance level that is as close to $\alpha = 0.05$ as you can, and state what this rejection region is and the significance level as part of your answer. Report your p-value and conclusions.

**(c)** As noted in class, the only condition required for the Sign Test is that the observations are independent. Comment on whether that condition is likely satisfied here. *Note*: there are many possible correct answers in which you could argue this in either direction. All reasonable attempts will receive full credit for this question.

**Question 3**

**(a)** Plot a histogram of the number of minutes that Smith played in each game. Again you may create it using either base R graphics or `ggplot`, but either way, be sure to adjust axes labels and the title to things that are meaningful and easy to read, and adjust font sizes to be readable if needed as

**(b)** Noting that this distribution looks fairly symmetric and bell-shaped, a t-Test is probably valid here. However, suppose we still wanted to consider a Sign Test. Suppose that Coach VanDerveer aimed to play Smith for 33.5 minutes in each game, with symmetric spread around that. She wants to know if she succeeded in this. Perform a t-Test for:

$$H_0 : \mu = 33.5$$
$$H_A : \mu \neq 33.5$$

and also a Sign Test for:

$$H_0 : \tilde{\mu} = 33.5$$
$$H_A : \tilde{\mu} \neq 33.5$$

where $\mu$ and $\tilde{\mu}$ are the population mean and median, respectively, Report your p-value and conclusions from each and comment briefly. For each test, use the standard p-value cutoff of 0.05.

**(c)** Apart from testing slightly different things (the Sign Test is a test of the median and a t-Test is a test of the mean), they also may have substantially different power. In particular, when the distributional conditions of the t-Test are met, the t-Test is theoretically the most optimal test you can do for a test of the mean in terms of statistical power.

Code a simulation to demonstrate at least one scenario in which a t-Test has higher power than a Sign Test to detect a deviation from a mean or median respectively (use the same null hypothesis value either way). You do not need to show a variety of situations; just one is sufficient for this question. However, the situation you demonstrate should show the t-Test having at least 20% higher power (e.g. 80% power for the t-Test vs. 60% power for the Sign Test) – thus, it may require you to try several situations before you get one that satisfies this.

You may use a p-value cutoff of 0.05 in either case, even though the Sign Test may not have a significance level that is exactly 0.05. Use at least 10,000 replicates in your simulations. Note that even with this many replicates, results will vary from run to run unless you set a seed – so you may choose to do this particularly if your result is fairly close to the 20% requirement to make sure that any given knit doesn't put you under it.

Briefly summarize your findings, including:

- Details of your simulation: what distribution did you use, with what parameter values, and what sample size
- The estimates of power from the t-Test and the Sign Test in that situation