Review of Quiz 1 & Homework 1

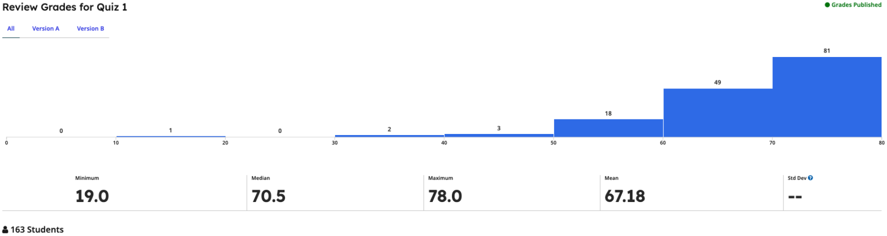
DSC 152 – Week 5 Discussion

Apr. 29th, 2026

Agenda

- Go through quiz (version A) and provide justification for the solutions
- If time permits, go through some reasons why people lost points on some of the homework problems

Some Statistics for Quiz #1



Review Grades for Quiz 1

Minimum	Median	Maximum	Mean	Std Dev
19.0	70.5	78.0	67.18	--

163 Students

Question #1

You want to test whether the true average tip amount differs from \$4.00 using a two-sided one-sample t-test.

(a) Write the null/alternative hypotheses in terms of the appropriate parameter.

We test

$$H_0 : \mu = 4.0 \quad \text{vs.} \quad H_A : \mu \neq 4.0.$$

(b) Write one line of R code to run this t-test on the tip amount in deliveries and output *only* the p-value.

```r
t.test(deliveries$tip_amount, mu=4.0)$p.value
```

We use the `$` operator to get the column name `tip_amount`, and then again to get the p-value using the `p.value` attribute of the test output.

Question #1

(c) Suppose that this test outputs a p-value of approximately 0.14. Which are true? Select all that apply.

■ If H_0 is true, there is a 14% probability of observing a test statistic at least as extreme as the one we computed.
TRUE: This is the definition of the p-value. It measures the probability of obtaining a test statistic as or more extreme than the observed statistic assuming the null hypothesis is true.

☐ There is a 14% probability that the true average tip amount is $4.00.
FALSE: The population mean μ is a fixed value, not a random variable, so it cannot have a probability of being a particular value.

■ At $\alpha = 0.10$, we would fail to reject H_0.
TRUE: By convention, we reject the null hypothesis when our p-value is below α. Here, $p\text{-value} = 0.14 \nleq 0.10 = \alpha$, so we would not have strong enough evidence to reject H_0.

☐ There is a 0.14 probability that H_A is true.
FALSE: The p-value is a measure of compatibility between the data and the null hypothesis, but it does not measure how likely the alternative hypothesis is to being true.

☐ None of the above.
FALSE: Choices 1 and 3 are true, so this cannot be correct.

Question #1

(d) Suppose we wanted one-sided test instead, with the alternative hypothesis being that the true mean tip is less than \$4.00. Based on p-value from (c) and assuming sample mean tip amount is less than \$4.00, can we get one-sided p-value without performing new calculations?

Yes, because (i) the observed sample mean is in the same direction as H_A (i.e., below \$4.00) and (ii) the distribution of the test statistic is symmetric. Then, the one-sided p-value is half of the two-sided, or 0.07 in this case. We would reject the null at the $\alpha = 0.10$ significance level.

(e) Briefly state validity conditions for one-sample t-test.

(i) Random sample (and, thus, independent observations)
(ii) Data are normally distributed (or sample size large enough such that CLT can apply)

Question #1

(f) Suppose we know tip amounts are often highly right-skewed. If we collect a sample of $n = 6$ deliveries from this skewed distribution and run a t-test at $\alpha = 0.1$ given hypotheses in (a), what will likely be true about the Type I error rate?

- ○ It will be approximately 0.10.
  FALSE: The skewness violates the normality of the data condition for validity of the t-test, meaning the Type I error rate is likely to differ from the significance level α.
- ● It is likely to be inflated over 0.10.
  TRUE: The skewness of the data often results in more "extreme" values in one tail than what is expected in the t-distribution, meaning the t-test to reject the null hypothesis more often than the intended $\alpha\%$ rate.
- ○ It is likely to be deflated below 0.10.
  FALSE: The right-skewed distribution has a "heavy" tail that contains more mass than the symmetric t-distribution assumes. This causes the test statistic to fall into the rejection region more often than α would predict, resulting in an inflation of the error rate.
- ○ You will always make a Type I Error.
  FALSE: A Type I error occurs when we reject the null hypothesis when this hypothesis is true. Even with a flawed test, the outcome is still probabilistic; you will still fail to reject H_0 in many instances.

Question #1

(g) A colleague suggests estimating the Type I Error rate in the situation from (f) using simulation. Fill in the code to estimate the Type I error rate when data come from a Gamma distribution with mean 4.

```r
# SOLUTION
count <- 0
for(i in 1:10000){
  skewed_data <- rgamma(n = 6, shape = 1.5, scale = (4 /
    1.5))
  p.val <- t.test(skewed_data, mu=4.0)$p.value
  if(p.val < 0.10){
    count <- count + 1
  }
}
TypeI <- count / 10000
```

We set the null hypothesis value to `mu=4.0`, the p-value, and then we compare the observed p-value with the significance level $\alpha = 0.10$.

Question #2

Because tip amounts are right-skewed, a colleague suggests the use of the sign test instead of the t-test.

(a) Explain what parameter the sign test is testing, and how this is different from the one-sample t-test.

Sign test tests for median $\tilde{\mu}$ of distribution, not mean μ (one-sample t-test). We want to check if 50% of observations fall above/below the hypothesized value.

(b) Using 4 as the null hypothesis value of the parameter of interest, write out the sign sequence (using $+$ and $-$) for the sign test, given 5.5, 4.8, 6.1, 3.2, 4.9, 1.7, 7.2.

5.5 $(+)$, 4.8 $(+)$, 6.1 $(+)$, 3.2 $(-)$, 4.9 $(+)$, 1.7 $(-)$, 7.2 $(+)$; 5 positives, 2 negatives

Question #2

(c) Using this sign sequence, compute the p-value of a two-sided test using the Binomial distribution. [Answer can be in terms of `pbinom()` or `dbinom()` R expressions.]

```
p.val <- pbinom(q = min(5,2), size = 7, prob = 0.5) * 2
```

We need to multiply by two to deal with the fact that we're working with a two-sided test.

(d) Explain how you would determine the significance level that is closest to 0.05. Your answer may include `pbinom` or `dbinom` expressions.

We can compute the binomial tail probabilities under $H_0 : p = 0.5$ for rejection region cutoffs $k \in \{0, 1, \ldots, 7\}$ for

```
pbinom(q = k, size = 7, prob = 0.5)
```

Since the binomial distribution is discrete and only allows for a finite set of possible significance levels α, we choose whichever value for k yields a probability closest to 0.05.
[Note: This question came almost directly from a problem on Homework #1.]

Question #3

(a) A colleague proposes using a bootstrap confidence interval (instead of the t-test) to see if the mean tip amount differs from $4.00. Fill in this R code to construct a 95% bootstrap CI for the mean:

```r
# SOLUTION
set.seed(152)
tip_sample <- c(2.50, 5.80, 3.10, 6.20, 1.90, 4.70, 3.60,
     5.40)
boot_means <- NA
for(i in 1:10000){
    boot_means[i]<-mean(sample(tip_sample, replace=TRUE))
}
ci <- quantile(boot_means, probs=c(0.025, 0.975))
ci
## 2.5% 97.5%
## 3.10 5.19
```

We sample from our `tip_sample` with replacement (that's how bootstrapping works!) and take the mean to get the bootstrapped mean. The `quantile` yields the two quantiles, at the 2.5% and the 97.5% level (for the two-sidedness), that form our CI.
[This question came almost directly from an in-class example and Lab #2.]

Question #3

(b) Based on the 95% bootstrap CI, what is the decision regarding $H_0 : \mu = 4$ at $\alpha = 0.05$?

- ○ Reject H_0, because 4 is inside the interval.
  FALSE: This is the incorrect conclusion for the right justification for HT/CI duality. If the null value is in the interval, it is deemed a plausible value for the population mean, and so we would not have sufficient evidence to reject it.
- ○ Reject H_0, because 4 is outside the interval.
  FALSE: Clearly $4 \in [3.10, 5.19]$, so this statement is incorrect.
- ○ Fail to reject H_0, because 4 is outside the interval.
  FALSE: Clearly $4 \in [3.10, 5.19]$, so this statement is incorrect.
- ● Fail to reject H_0, because 4 is inside the interval.
  TRUE: By the HT/CI duality, we reject the null hypothesis $H_0 : \mu = c$ at the significance level α if and only if $c \notin (1-\alpha)-\text{level CI}$. With $4 \in [3.10, 5.19]$, we fail to reject.

Question #3

(c) A colleague says: "Bootstrap confidence intervals should have a Type I Error rate of approximately 0.05 regardless of the shape of the data distribution, because the bootstrap is a non-parametric procedure." Is this claim correct?

- ☐ Yes. Everything the colleague said is correct.
  FALSE: Non-parametric methods, like bootstrapping, can still rely on assumptions about the data and can fail with highly skewed, small datasets.
- ■ No. The bootstrap can have an inflated Type I error rate.
  TRUE: The bootstrap distribution can underestimate the true variability of the data, especially for small sample sizes, leading to confidence intervals that exclude the true mean too often.
- ☐ No. The colleague is incorrect because the bootstrap is actually a parametric procedure.
  FALSE: The bootstrap is non-parametric because it does not assume any specific functional forms for the population distribution.
- ☐ Yes. However, since the bootstrap is a non-parametric procedure, it only works for testing the median.
  FALSE: Bootstrapping is a tool to estimate the sampling distribution of (almost) any statistic, including the mean, median, etc.
- ■ No. The accuracy of the bootstrap CIs relies on the sample being an accurate representation of the population, which is not always the case.
  TRUE: Resampling works when the sample is representative of the population.; otherwise, the bootstrap will replicate and show the flaws of the sample (small size, unrepresentative sample, etc.)
- ☐ None of the above.
  FALSE: Choices 2 and 5 are correct, so this cannot be true.

Question #4

Estimating statistical power with one-sample t-test for $H_0 : \mu = 4.0$ vs. $H_A : \mu > 4.0$, when $\overline{X} = 6.0$.

(a) Estimate power via simulation with code.

```r
# SOLUTION
count <- 0
for(i in 1:10000){
    samp <- rnorm(n = 30, mean = 6, sd = 3)
    pval <- t.test(samp,mu=4,alternative='greater')$p.value
    if(pval < 0.05){
        count <- count + 1
    }
}
count / 10000
```

Power is probability (proportion) of correctly rejecting H_0. To fill in the code, we just get the p-value of t-test between samp (with mean 6) and with H_0 (with mean 4), where the alternative is greater.

Question #4

(b) Can a theory-based equivalent of the power estimated by this simulation be calculated?

Yes, since the simulation follow the assumptions for a valid t-test. We have normally distributed data, and with the standard deviation being known, we use the `power.t.test` function:

```r
power.t.test(n=30, delta=2, sd=3, sig.level=0.05, type='one.sample', alternative='one.sided')
```

The sample size and standard deviation are the same as our simulated sample, and the alternative is one-sided. We have a one-sample t-test, and our effect size (`delta`) is the difference in the two means, or $6 - 4 = 2$.
This line of code yields the exact probability of rejecting H_0 that the simulation is attempting to approximate.

(c) Can a theory-based equivalent of the power estimated by a simulation whose data followed a GAMMA distribution be calculated?

NO, since a theory-based equivalent estimate requires a closed-form model to describe the test statistic's behavior under the alternative hypothesis. When the data follow $\Gamma(a, b)$ (not a normal or approximate normal), exact distribution for t-statistic is unknown and doesn't fit standard formula.

Simulation is required here since power is estimated empirically, whereas theory-based functions require explicit probability density function to describe statistic.

Question #4

(d) Which of these changes, applied individually, would **increase** the statistical power?

■ Increase the sample size n
TRUE: Increasing n reduces the standard error $(s/\sqrt{n})$, making the distributions more narrow, separating the null from any alternative distribution. This creates a larger rejection region, increasing the power.

■ Increase the effect size δ
TRUE: Power is the probability of detecting a difference; by increasing the distance between the true mean and the null mean, you shift any alternative distribution further away from the null. This results in a larger portion of the alternative distribution falling into the rejection region.

☐ Increasing the population standard deviation σ
FALSE: A higher σ increases the standard error, 'widening' the sampling distributions and creating the possibility for more overlap between the null and alternative distributions. This would cause power to go down.

☐ Decrease the significance level from $\alpha = 0.05$ to $\alpha = 0.01$
FALSE: Reducing α makes the test more conservative (i.e., we have a harsher requirement to meet for 'extremeness') and shrinks the rejection region. With the test not being rejected as often, our power would go down.

☐ None of the above.
FALSE: Choices 1 and 2 are correct, so this cannot be true.

Question #5

We test $H_0 : \mu = 4$ versus $H_A : \mu \neq 4$, where μ is the average tip amount. We know $\overline{X} = 4.03$ and $s = 1.6$, with a resulting p-value of 0.0002. The 95% CI is ($4.01, $4.05).

(a) At $\alpha = 0.05$, select all that are true:

☑ Since the p-value is extremely small, we reject H_0 at $\alpha = 0.05$.
TRUE: The decision rule is to reject H_0 if the p-value is smaller than α. Clearly, $0.0002 < 0.05$, so we have sufficient evidence to reject.

☐ Since the sample mean is only $0.03 away from $\mu = 4$, we actually do not reject H_0 at $\alpha = 0.05$ even though the p-value is small.
FALSE: Statistical significance is determined by the p-value as compared to our significance level α, not by the raw distance between the mean and the null distribution.

☑ Since the 95% CI is very close to the null hypothesis value, the difference from $4 may not be enough to act on even though we should reject H_0.
TRUE: While the result is statistically significant (we reject H_0), a $0.03 difference on a $4.00 base is likely negligible in the real-world, suggesting that "statistically significant" does not always mean "meaningful" or "practically significant".

☐ Since the p-value is extremely small, this indicates there is a meaningful difference from $4 even though the estimated mean is not very different from $4.
FALSE: A small p-value only indicates the observed difference is not likely to be due to chance.

☐ None of the above.
FALSE: Since choices 1 and 3 were correct, this statement cannot be true.

Question #5

(b) Consider a different scenario, where a study with $n = 12$ finds $\overline{X} = \$6.80$ and a p-value of 0.11 for $H_0 : \mu = 4$ vs. $H_A : \mu \neq 4$. Select all that are true.

- ☐ We reject H_0 at $\alpha = 0.05$ since the sample mean is so different from \$4.
  FALSE: Decisions on statistical significance are made on p-values, not on raw difference values. Since the p-value is larger than $\alpha = 0.05$, we fail to reject H_0, despite the larger difference between mean values.
- ☐ If the true tip is \$4.00 and the conditions of the test were met, then the probability of observing a tip that is at least \$6.80 or at most \$1.20 is 11%.
  FALSE: P-values refer to observing a test statistic (sample mean) as extreme as the observed one, not raw values (exact tip amount). This statement incorrectly assigns a probabilistic statement to an individual data point instead of a statistic.
- ☐ We fail to reject H_0 at $\alpha = 0.05$ and should end the investigation here, concluding that there is no meaningful difference from \$4.
  FALSE: Failing to reject H_0 does not prove that the mean is \$4. To truly capture whether this difference (effect size, essentially) is meaningful, we might conduct a power study.
- ■ The result indicates a larger sample size may be needed to determine whether there is a true difference from \$4.00.
  TRUE: This larger observed difference (\$6.80 vs. \$4.00) suggests an effect may exist, but the small sample size meant the test yielded a non-statistically significant p-value. Increasing n reduces standard error and increases power of our study.

Review of Homework #1

In the next few slides I'll provide some notes for where some people happened to lose some points on the assignment.

Notes on Question #1

Some notes on why people lost points for question #1:

- Part (a): Did not include contextual explanation for computed significance level actually means
- Part (b): Did not have a clear interpretation that the test conducted does not have a great ability to detect improvement
- Part (c): Mostly deducted points because people did not correctly interpret their plots within the context of the results from previous parts

Notes on Question #2

Some notes on why people lost points for question #2:

- Remembering the sign test assumes a two-sided alternative hypothesis

Notes on Question #3

Some notes on why people lost people for question #3:

- Part (b): Missing explanations for why the conclusions between the t-test and the sign test results can differ
- Part (b): Invalid explanations that the two tests are testing for the same parameter (related to previous comment)
- Part (c): Did not include justification for why t-test should win (i.e., brief illustration that it is the theoretically most optimal test for the mean in terms of statistical power)