431 Class 13

Thomas E. Love, Ph.D.

2022-10-11

Today's Agenda

- A New Example from a 2020 letter to NEJM
- Hypothesis Testing and Interval Estimation
  - Quantitative Outcome: Paired Samples
  - Quantitative Outcome: Single Sample
  - t, Bootstrap and Wilcoxon methods
- Slides 1-70 in-class, the rest are for at-home study

Today's Packages

```
library(infer) ## new today, part of tidymodels
library(readxl) ## to read in Excel sheet
library(broom) ## also part of tidymodels
library(Hmisc) ## for help with bootstrapping
library(kableExtra) ## for table tidying

library(janitor); library(naniar); library(patchwork)
library(tidyverse)

theme_set(theme_bw())
```

- Visit https://infer.tidymodels.org/ for more on infer.
- Visit https://moderndive.com/ especially Section III for more of a textbook-style presentation.

Something Happened? Signal or Noise?

Very often, sample data indicate that something has happened…

- the proportion of people who respond to this treatment has changed
- the mean value of this measure appears to have changed

Before we get too excited, it's worth checking whether the apparent result might possibly be the result of random sampling error. Statistics provides multiple ways to do this.

Making Inferences From A Sample

1. What is the population about which we aim to make an inference?
2. What is the sample available to us to make that inference?

- Who are the individuals fueling our inference?
- What data are available from those individuals?

3. Why might the study population not represent the target population?

For more, see Spiegelhalter, Chapter 3

Point Estimation and Confidence Intervals

- A **point estimate** provides a single best guess as to the value of a population or process parameter.
- A **confidence interval** can convey how much error one must allow for in a given estimate. It includes an interval estimate and a probability statement (confidence level.)

The key tradeoffs in estimation are

- cost vs. precision (larger samples produce narrower intervals), and
- precision vs. confidence in the accuracy of the statement.

Today’s Example

We’ll look at (part of) a 2020 NEJM letter which reports on a study of 70 inpatients with Covid-19.

> … [w]e tested saliva specimens collected by the patients themselves and nasopharyngeal swabs collected from the patients at the same time point by health care workers.

Wyllie et al. Saliva or Nasopharyngeal Swab Specimens for Detection of SARS-CoV-2.

N Engl J Med 2020; 383:1283-1286. DOI: 10.1056/NEJMc2016359 (2020-09-24)

Today's Data

```r
sal <- read_excel("c13/data/nejm_saliva.xlsx") |>
  clean_names() |> mutate(subject = as.character(subject))

dim(sal); pct_complete_case(sal)
```

```
[1] 70  5
```

```
[1] 100
```

```r
head(sal)
```

```
# A tibble: 6 × 5
  subject np_n1  s_n1 np_titre    s_titre
  <chr>   <dbl> <dbl>    <dbl>      <dbl>
1 1        30.0  28.1   805000   2612380.
2 3        45    38.0     2400      2400
3 4        29.5  26.9  1110000   5391649.
4 6        45    31.8     2400    264166.
5 7        29.9  25.9   841000  10200000
6 10       25.9  19.3  9970000 595000000
```

A Codebook (ignoring `subject`)

Variable	Description
`np_n1`	cycle threshold for PCR assay targeting SARS-CoV-2 N1 sequence via Nasopharyngeal Swab Sample
`s_n1`	cycle threshold via Saliva Sample
`np_titre`	Detected copies/ml of SARS-CoV-2 RNA via Nasopharyngeal Swab
`s_titre`	Detected copies/ml of SARS-CoV-2 RNA via Saliva

- More details available in Appendix 1 and 2 of Wyllie et al. (2020)
- There is an equation to take `n1` values to `titre` values.

Key Question A

Do the two sampling approaches (nasopharyngeal and saliva) provide meaningfully different results?

- Does the population of NP minus Saliva paired differences follow a distribution centered around zero?
- If we'd obtained data from the entire target population, is it reasonable that the mean difference between Saliva and NP would be zero?

This is a **paired samples** question, comparing NP to Saliva, where each subject provided both an NP and a Saliva result.

Question A: Paired Differences

Each subject provides a Nasopharyngeal response (`np_n1`) and a paired (by subject) Saliva response (`s_n1`).

- We'll analyze the paired differences (`np_n1` - `s_n1`).

```r
sal <- sal |> mutate(diff = np_n1 - s_n1)
## last three subjects shown below
tail(sal, 3) |> select(subject, np_n1, s_n1, diff) |>
  kbl() |> kable_minimal(full_width = FALSE, font_size = 24)
```

subject	np_n1	s_n1	diff
294	38.59	28.31	10.28
306	45.00	45.00	0.00
366	35.87	45.00	-9.13

Data Setup for Question A

We have a sample of 70 observations of the saliva - NP paired differences. We want to know if they might plausibly come from a distribution with mean zero.

```
stem(sal$diff)
```

```

  The decimal point is 1 digit(s) to the right of the |

  -2 | 1
  -1 | 776
  -1 | 4
  -0 | 9987
  -0 | 4444433332211
   0 | 00000000122233333444
   0 | 5555777777888899
   1 | 00111333
   1 | 7
   2 | 023
```

Key Question B

Could the sample of N1 saliva data have plausibly come from a population with mean 35?

- If we could have sampled the entire target population, is it reasonable that the mean of that population would be 35?
- Is the true mean of N1 (saliva) = 35?

This is a **one-sample** question, comparing N1 in Saliva to 35, and ignoring the Nasopharyngeal data.

Data Setup for Question B

We have a sample of 70 observations of the N1 values for Saliva. We want to know if they might plausibly come from a distribution with mean 35.

```r
stem(sal$s_n1)
```

```
  The decimal point is 1 digit(s) to the right of the |

  1 | 4
  1 | 5799
  2 | 000111233444
  2 | 55667778888
  3 | 011122234444
  3 | 5556667899
  4 | 004
  4 | 55555555555555555
```

Key Insight

A paired samples analysis and a one-sample analysis are done in exactly the same way.

- If you have paired data, take paired differences, and treat them as you would a single sample.
- So, I can walk through Question A's analysis, and Question B will take the same approach.

The comparison of Saliva vs. Nasopharyngeal specimens seems of greater interest, so we'll start with that, and Question A.

Question A Analyses

DTDP for Paired Differences

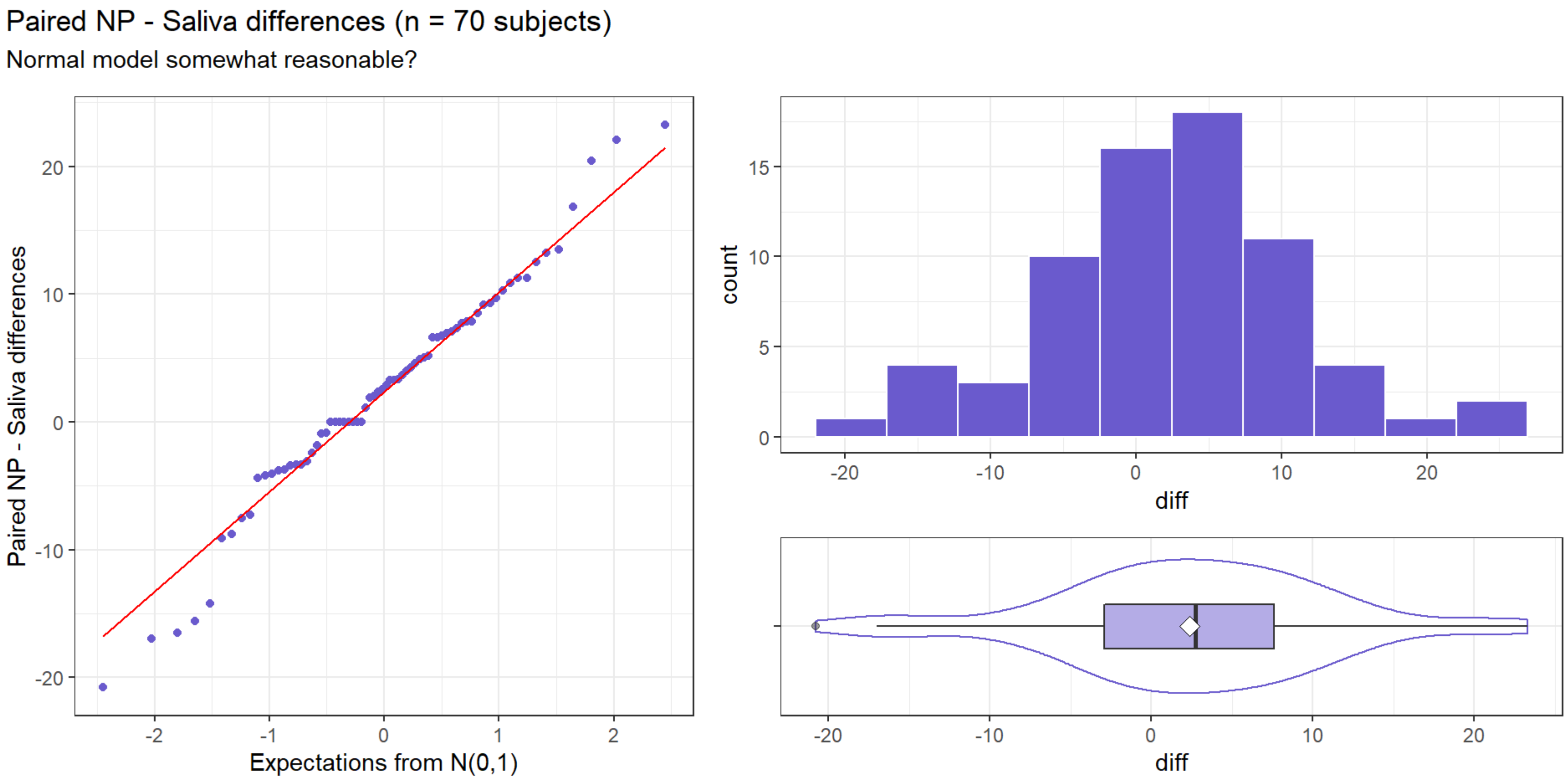
```r
p1 <- ggplot(sal, aes(sample = diff)) +
  geom_qq(col = "slateblue") + geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(y = "Paired NP - Saliva differences",
       x = "Expectations from N(0,1)")

p2 <- ggplot(sal, aes(x = diff)) +
  geom_histogram(bins = 10, col = "white", fill = "slateblue")

p3 <- ggplot(sal, aes(x = diff, y = "")) +
  geom_violin(col = "slateblue") +
  geom_boxplot(fill = "slateblue", alpha = 0.5, width = 0.3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 3, fill = "white") +
  labs(y = "")

p1 + (p2/p3 + plot_layout(heights = c(2,1))) +
  plot_annotation(title = "Paired NP - Saliva differences (n = 70 subjects)
                  subtitle = "Normal model somewhat reasonable?")
```

DTDP for Paired Differences



Paired Differences by the Numbers

```r
mosaic::favstats(~ diff, data = sal) |>
  kbl(digits = 3) |> kable_minimal(font_size = 24)
```

min	Q1	median	Q3	max	mean	sd	n	missing
-20.77	-2.923	2.725	7.615	23.28	2.36	8.672	70	0

```r
Hmisc::describe(sal$diff)
```

```
sal$diff 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
      70        0       62    0.998     2.36    9.601  -14.984   -7.666 
     .25      .50      .75      .90      .95 
  -2.923    2.725    7.615   11.375   15.328 

lowest : -20.77 -16.99 -16.53 -15.61 -14.22, highest:  13.48  16.84  20.45 
22.06  23.28
```

Hypothesis Testing / Confidence Intervals for a Mean of Paired Differences

1. Using the **t distribution** to produce a test and confidence interval about the population mean of the paired differences, \(\mu\).

- Assumes that the paired differences are drawn from a Normal distribution.

2. Using the **bootstrap** to produce a test or confidence interval about \(\mu\)

- Doesn’t assume Normality of the paired differences.

3. Using the **Wilcoxon signed rank** procedure to produce a test/CI about the population *pseudo-median* of the paired differences.

- Pseudo-median is close to the mean/median only if differences are *symmetric*.

Hypothesis Testing Elements

1. Specify the null hypothesis, \(H_0\).
2. Specify the alternative hypothesis, \(H_A\).
3. Specify \(\alpha\), tolerable Pr(incorrectly rejecting \(H_0\)).
    - Confidence level is 100(1-\(\alpha\))%, so 95% confidence means we will tolerate \(\alpha = 0.05\).
4. Specify the approach to be used to make inferences based on the sample.
5. Obtain the data and summarize it to obtain appropriate p-value, and/or CI.

For our Question A

\(H_0: \mu = 0\) vs. \(H_A: \mu \neq 0\).

1. Our null hypothesis is that the Saliva and NP approaches yield the same results. This would lead to their paired differences having mean zero.
2. If the null hypothesis is not true, it means that the alternative (that the true mean of the paired differences is either greater than or less than 0) must be true.
3. We'll use a 95% confidence level, corresponding to \(\alpha\) = 0.05.

Available Approaches for Inference

- t-based approach (either via indicator variable regression or direct t test)
- bootstrap approach via `infer` package
- bootstrap confidence interval via `smean.cl.boot()`
- Wilcoxon signed rank

T-based Approach

Indicator Variable Regression

This produces the t test result for the mean of `diff`.

```r
m1 <- lm(diff ~ 1, data = sal)

tidy(m1, conf.int = TRUE, conf.level = 0.95) |>
  kbl(digits = 3) |> kable_classic(font_size = 24)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.36	1.037	2.277	0.026	0.292	4.428

Another way to get the t results

This also produces the t test result for the mean of `diff`.

```r
t.test(sal$diff, mu = 0, conf.level = 0.95)
```

```

	One Sample t-test

data:  sal$diff
t = 2.2765, df = 69, p-value = 0.02592
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2918899 4.4275387
sample estimates:
mean of x 
 2.359714 
```

Hand-calculation of the t statistic

The one-sample t test uses as its test statistic:

\[t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.36-0}{8.672/\sqrt{70}} = \frac{2.36}{1.0365} = 2.277 \]

- Sample mean \(\bar{x}\), standard deviation \(s\), null hypothesized mean = \(\mu_0\).

Obtaining the p value

The t distribution is indexed by its degrees of freedom. - In this case, we have \(n = 70\) and df = \(n-1\) = 69.

In R, we can obtain a two-tailed p value for our test statistic of 2.277 using 69 degrees of freedom with:

```r
pt(2.277, df = 69, lower.tail = FALSE)*2
```

```
[1] 0.02589075
```

Defining a *p* Value (but not very well)

The *p* value estimates the probability that we would obtain a result as much in favor or more in favor of the alternative hypothesis \(H_A\) as we did, assuming that \(H_0\) is true.

- The *p* value is a conditional probability of seeing evidence as strong or stronger in favor of \(H_A\) calculated **assuming** that \(H_0\) is true.

How people use the *p* Value

- If the *p* value is less than \(\alpha\), this suggests we might reject \(H_0\) in favor of \(H_A\), and declare the result statistically significant.

But we won’t be comfortable with doing that, at least in due time.

What the *p* Value isn’t

The *p* value is not a lot of things. It’s **NOT**

- The probability that the alternative hypothesis is true
- The probability that the null hypothesis is false
- Or anything like that.

The *p* value **is closer to** a statement about the amount of statistical evidence contained in the data that favors the alternative hypothesis \(H_A\). It’s a measure of the evidence’s credibility.

Confidence Interval via t distribution

The two-sided 100(1-\(\alpha\))% confidence interval for \(\mu\) is:

\[\bar{x} \pm t_{\alpha/2, n-1} (\frac{s_x}{\sqrt{n}}) ==> 2.36 \pm t_{0.025, 69} (\frac{8.672}{\sqrt{70}})\]

and we can obtain \(t_{0.025, 69}\) from R with:

```r
qt(0.025, 69, lower.tail = FALSE)
```

```
[1] 1.994945
```

\[2.36 \pm 1.9949 \times 1.0365 => 2.36 \pm 2.07, \mbox{ or } (0.29, 4.43) \]

Interpreting the Confidence Interval

Some people think this means that there is a 95% chance that the true mean of the population, μ, falls between 0.29 and 4.43. Not true.

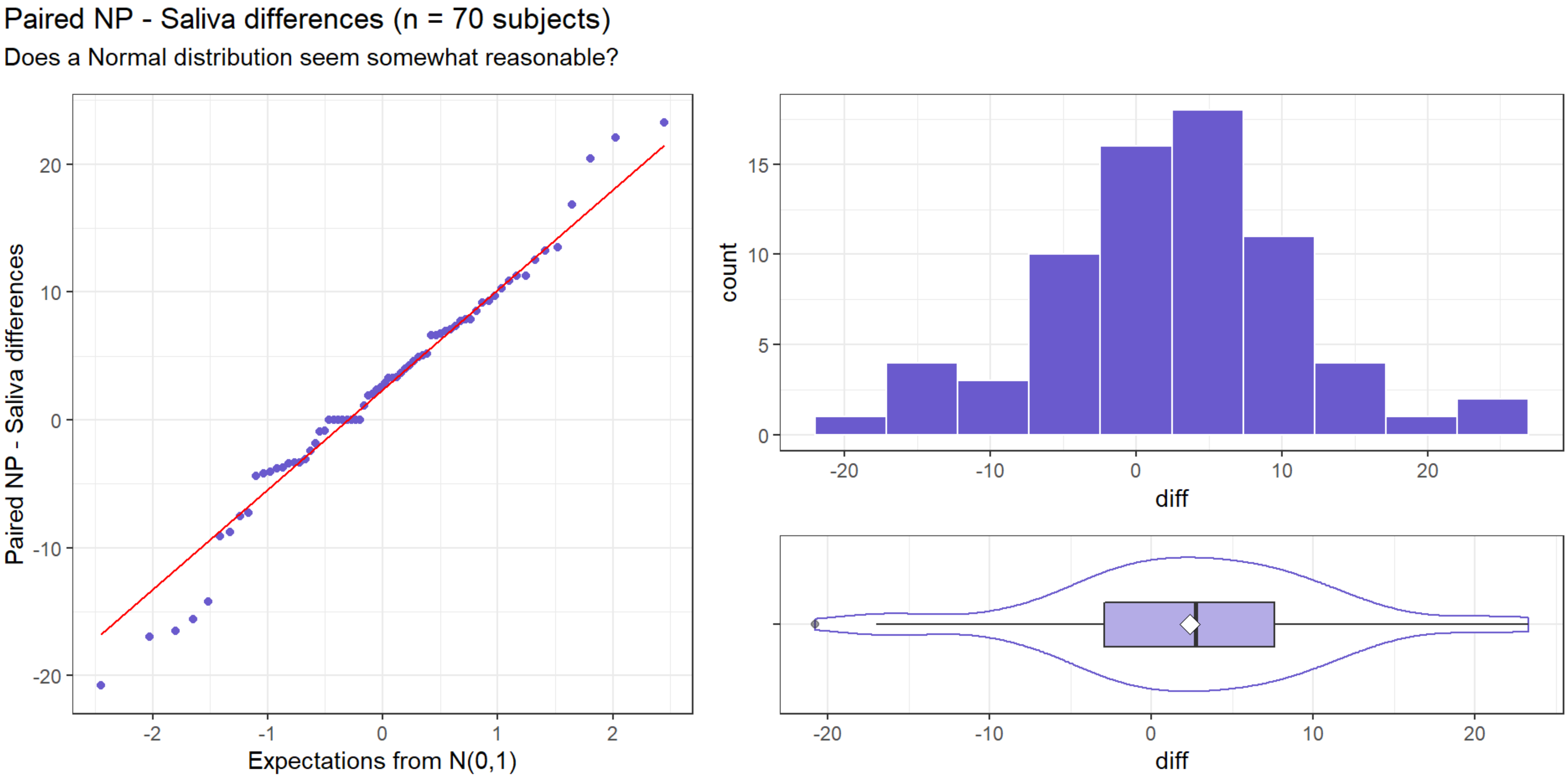
Our confidence is in the process. If we built 100 confidence intervals this way, 95 would be expected to contain the true value of the parameter \(\mu\).

- We are accounting for one particular type of error (called sampling error) in developing our interval estimate, while assuming all other potential sources of error are negligible.

Assumptions of the t-based approaches

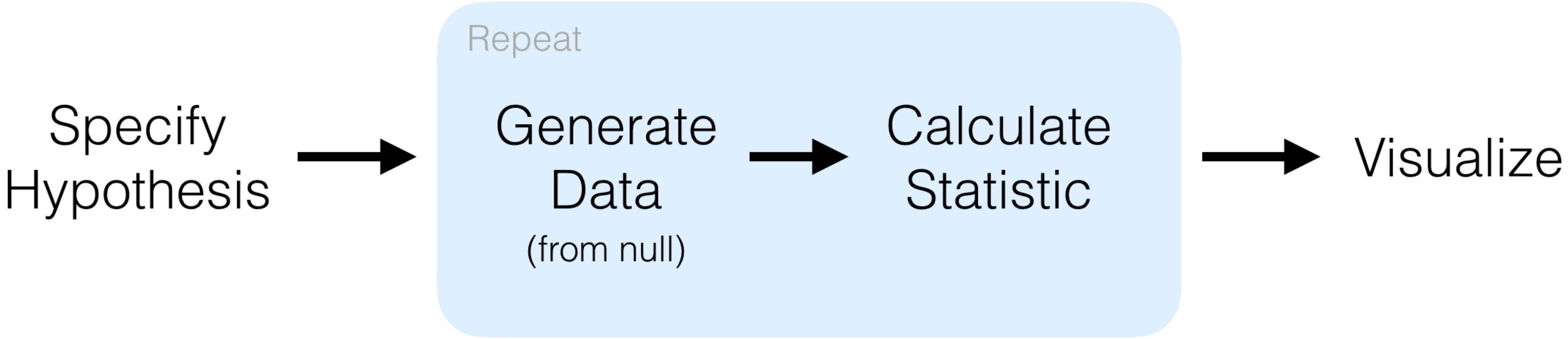
1. The data (specifically the paired differences) are a random sample from what we would observe if we could obtain the entire target population.
2. The subjects who provide the responses we observe are selected independently from the target population. In other words, if I am selected, it does not change the probability that you will be selected if we are each in the target population.
3. The target population's paired differences follow a Normal distribution.

Our Paired Differences, Again



Bootstrap Approach

What the infer package does



A randomization test using infer tools

This is a randomization-based analog to the 1-sample t test.

First, we calculate the observed statistic:

```r
observed_statistic <- sal |>
  specify(response = diff) |>
  calculate(stat = "mean")

observed_statistic
```

```
Response: diff (numeric)
# A tibble: 1 × 1
   stat
  <dbl>
1  2.36
```

Compare to a Null Distribution?

Our next goal is to compare this observed statistic to a null distribution, generated under the assumption that the mean was actually 0, to get a sense of how likely it would be for us to see this observed mean difference in the population.

Our null hypothesis is still \(H_0: \mu = 0\) vs. the two-tailed alternative \(H_A: \mu \neq 0\).

Using the bootstrap to generate the null distribution

We can generate the null distribution using the bootstrap.

- In the bootstrap, for each replicate, a sample of size equal to the input sample size is drawn (with replacement) from the input sample data.
- This allows us to get a sense of how much variability we'd expect to see in the entire population so that we can then understand how unlikely our sample mean would be.

`infer` package has 4 main verbs

Verb	Activity
`specify()`	specify variable, or relationship between variables, that interests us
`hypothesize()`	declare the null hypothesis
`generate()`	generate data reflecting the null hypothesis
`calculate()`	obtain a distribution of statistics from the generated data to form the null distribution

Generate the null distribution

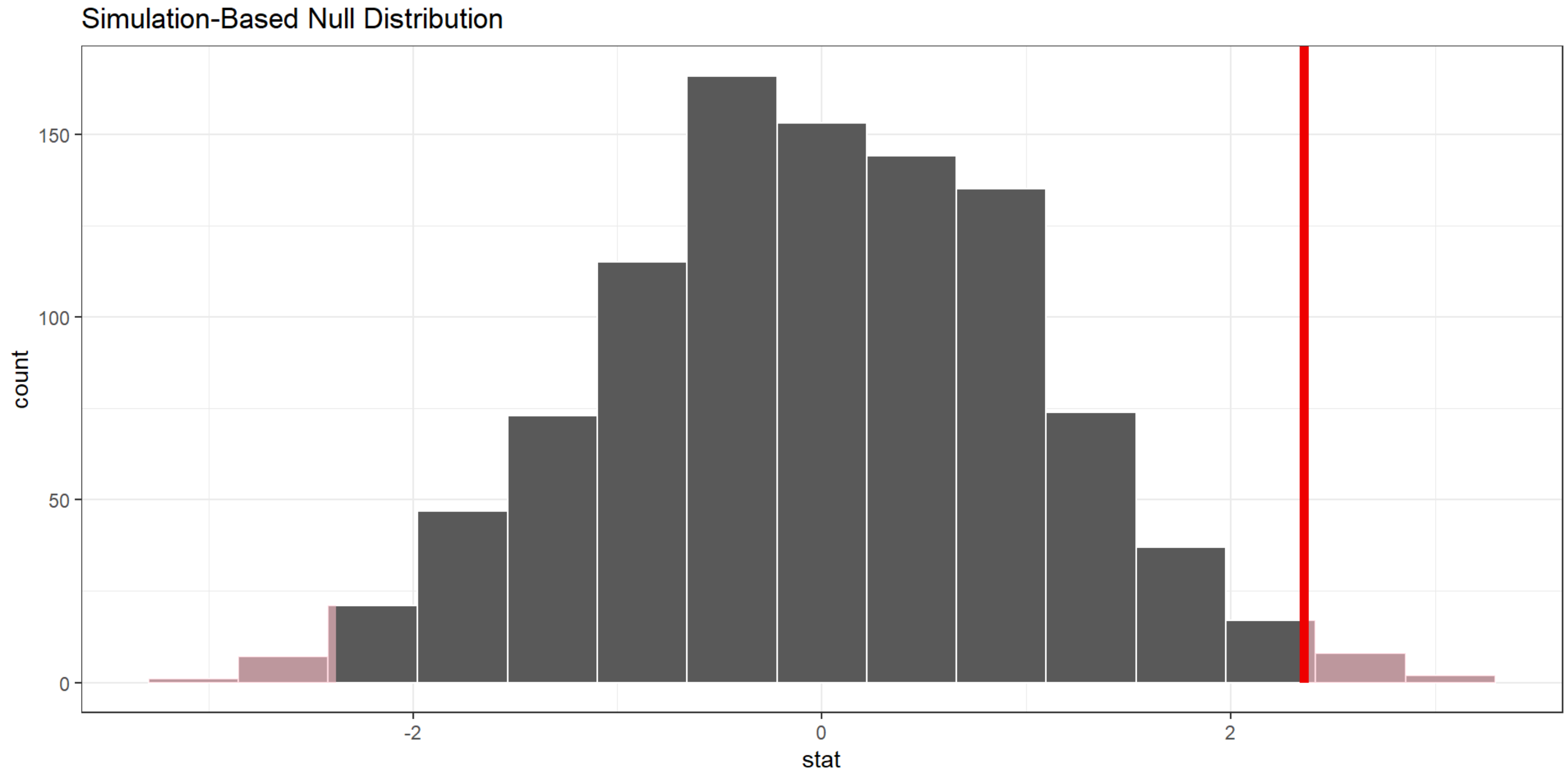
Using the bootstrap, we need to set a seed so we can replicate our work later:

```r
set.seed(20221011)
null_dist_diffs <- sal |>
  specify(response = diff) |>
  hypothesize(null = "point", mu = 0) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean")
```

Resulting Null Distribution

Get a sense of where our observed statistic falls.

```
null_dist_diffs |>
  visualize() +
  shade_p_value(observed_statistic, direction = "two-sided")
```



Calculating the bootstrap *p* value

```r
p_value_1_sample <- null_dist_diffs |>
  get_p_value(obs_stat = observed_statistic,
              direction = "two-sided")

p_value_1_sample
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1   0.022
```

Thus, if the true mean of the paired differences was really 0, our approximation of the probability that we would see a test statistic as or more extreme than is approximately 0.022

Bootstrap CI from `infer`

Using a reasonable approach based on the bootstrap sample generated by `infer`:

```r
ci_diffs <- null_dist_diffs |>
  get_confidence_interval(point_estimate = observed_statistic,
                          level = 0.95, type = "se")

ci_diffs
```

```
# A tibble: 1 × 2
  lower_ci upper_ci
     <dbl>    <dbl>
1    0.343     4.38
```

Bootstrap CI from `smean.cl.boot()`

\(H_0: \mu = 0\) vs. \(H_A: \mu \neq 0\).

We've previously seen a quick way to get a 95% bootstrap confidence interval for a mean via `smean.cl.boot()` from `Hmisc`:

```
set.seed(431)
smean.cl.boot(sal$diff, conf.int = 0.95, B = 2000)
```

```
     Mean     Lower     Upper 
2.3597143 0.3687571 4.4138643 
```

What can we conclude about our hypotheses in light of this interval?

When is a Bootstrap CI for \(\mu\) Reasonable?

The interval will be reasonable as long as we are willing to believe that:

- the original sample was a random sample (or at least a completely representative sample) from a population,
- and that the samples are independent of each other (selecting one subject doesn't change the probability that another subject will also be selected)
- and that the samples are identically distributed (even though that distribution may not be Normal.)

It is still possible that the results can be both:

- **inaccurate** (i.e. they can, include the true value of the unknown population mean less often than the stated confidence probability) and
- **imprecise** (i.e., they can include more extraneous values of the unknown population mean than is desirable).

Assumptions of the bootstrap

1. The data (specifically the paired differences) are a random sample from what we would observe if we could obtain the entire target population.
2. The subjects who provide the responses we observe are selected independently from the target population. In other words, if I am selected, it does not change the probability that you will be selected if we are each in the target population.

Wilcoxon Signed Rank Procedure

The Wilcoxon Signed Rank Procedure

The Wilcoxon signed rank approach builds interval estimates for the population *pseudo-median* when the population can only be assumed to be symmetric.

- For any sample, the pseudo-median is defined as the median of all of the midpoints of pairs of observations in the sample.
- As it turns out, if you're willing to assume the population is **symmetric** (but not necessarily Normally distributed) then the pseudo-median is equal to the population median.

Wilcoxon based 95% confidence interval

```r
wilcox.test(sal$diff, mu = 0, conf.int = TRUE, conf.level = 0.95)
```

```

	Wilcoxon signed rank test with continuity correction

data:  sal$diff
V = 1330, p-value = 0.01333
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 0.6150213 5.1700248
sample estimates:
(pseudo)median 
      3.015033 
```

Interpreting the Wilcoxon Signed Rank CI

If we're willing to believe the `diff` values come from a population with a symmetric distribution, the 95% Confidence Interval for the population median would be (0.6, 5.2). For a non-symmetric population, this only applies to the *pseudo-median*. The pseudo-median will be fairly close to the sample mean and median if the population actually follows a symmetric distribution. Here, the estimated pseudo-median was 3.015.

```
mosaic::favstats(~ diff, data = sal)
```

```
   min      Q1 median    Q3   max     mean       sd  n missing
 -20.77 -2.9225  2.725 7.615 23.28 2.359714 8.672247 70       0
```

Wilcoxon Assumptions

1. The data (specifically the paired differences) are a random sample from what we would observe if we could obtain the entire target population.
2. The subjects who provide the responses we observe are selected independently from the target population. In other words, if I am selected, it does not change the probability that you will be selected if we are each in the target population.
3. The data are reasonably assumed to come from a symmetric population, so that the pseudo-median is of interest.

Question A Conclusions

Does the population of NP minus Saliva paired differences follow a distribution centered around zero?

- Sample of paired differences reasonably modeled by Normal distribution.
- \(H_0: \mu = 0\) vs. \(H_A: \mu \neq 0\).
- Results from our four approaches on the next slide.

Question A Results

Procedure	*p*	Estimate	95% CI
t	0.026	\(\hat{\mu} = \bar{x}\) = 2.36	(0.29, 4.43)
Bootstrap via `infer`	0.022	\(\hat{\mu} = \bar{x}\) = 2.36	(0.34, 4.38)
Bootstrap via `Hmisc`	< 0.05	\(\hat{\mu} = \bar{x}\) = 2.36	(0.37, 4.41)
Wilcoxon	0.013	ps.-med. = 3.02	(0.62, 5.17)

What can we conclude?

Question B Analyses

Question B (Single Sample)

\(H_0: \mu = 35\) vs. \(H_A: \mu \neq 35\) where \(\mu\) is now the population mean of the N1 values for Saliva Samples.

Again, we’ll use a 95% confidence level, or \(\alpha = 0.05\).

- DTDP
- t approaches
- bootstrap with `infer`, bootstrap with `smean.cl.boot()`,
- Wilcoxon signed rank
- Conclusions

DTDP for Saliva N1 values

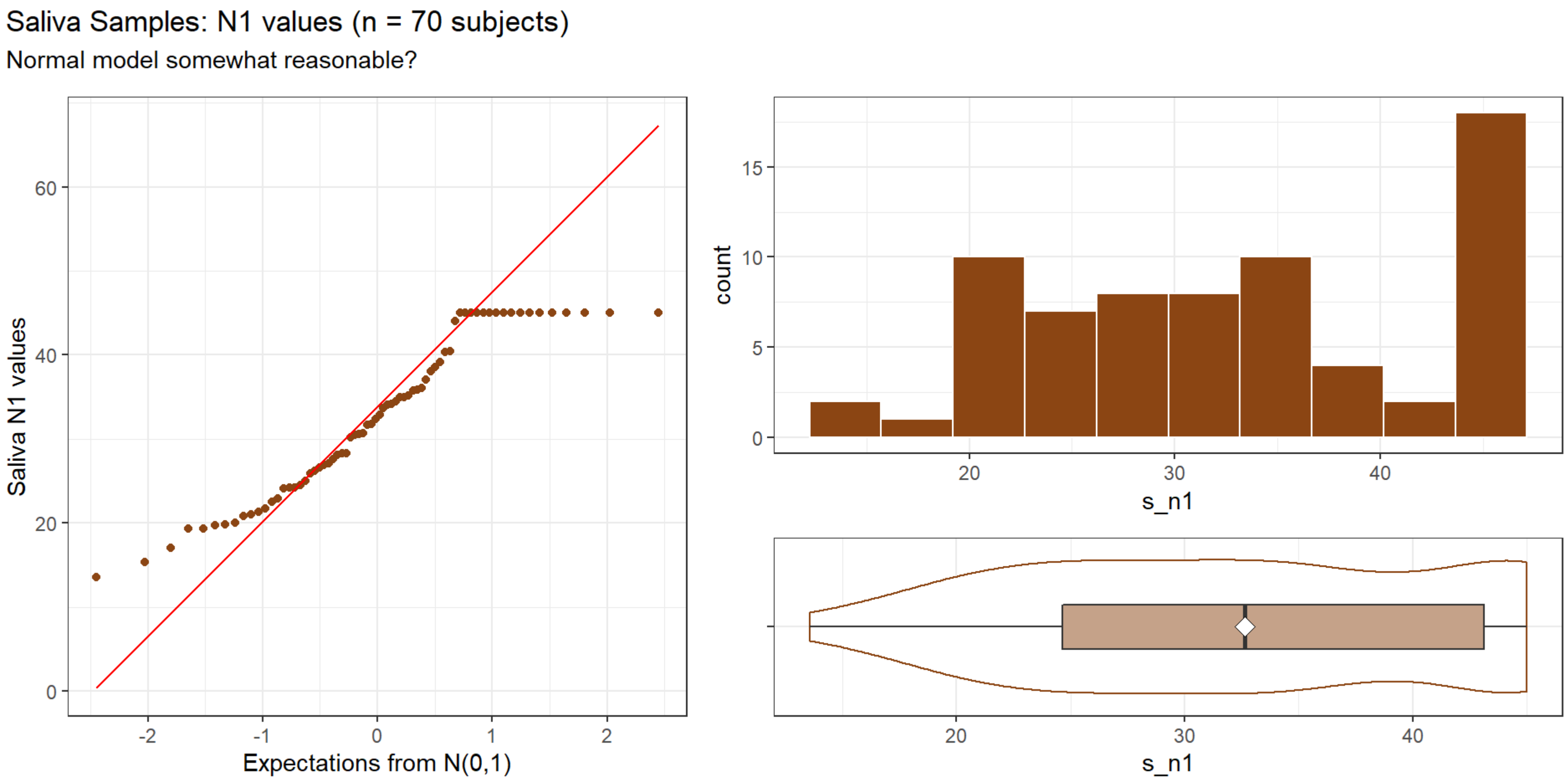
```r
p1 <- ggplot(sal, aes(sample = s_n1)) +
  geom_qq(col = "saddlebrown") + geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(y = "Saliva N1 values",
       x = "Expectations from N(0,1)")

p2 <- ggplot(sal, aes(x = s_n1)) +
  geom_histogram(bins = 10, col = "white", fill = "saddlebrown")

p3 <- ggplot(sal, aes(x = s_n1, y = "")) +
  geom_violin(col = "saddlebrown") +
  geom_boxplot(fill = "saddlebrown", alpha = 0.5, width = 0.3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 3, fill = "white") +
  labs(y = "")

p1 + (p2/p3 + plot_layout(heights = c(2,1))) +
  plot_annotation(title = "Saliva Samples: N1 values (n = 70 subjects)",
                  subtitle = "Normal model somewhat reasonable?")
```

DTDP for Saliva N1 values



Numerical Summaries

```
mosaic::favstats(~ s_n1, data = sal) |>
  kbl(digits = 2) |> kable_minimal(font_size = 24)
```

min	Q1	median	Q3	max	mean	sd	n	missing
13.57	24.67	32.66	43.11	45	32.63	9.43	70	0

T approach

Even though the Normal assumption for the N1 values is hard to believe.

```r
tt <- t.test(sal$s_n1, mu = 35, conf.level = 0.95)
tidy(tt) |> kbl(digits = 3) |> kable_classic_2()
```

estimate	statistic	p.value	parameter	conf.low	conf.high
32.631	-2.103	0.039	69	30.383	34.878

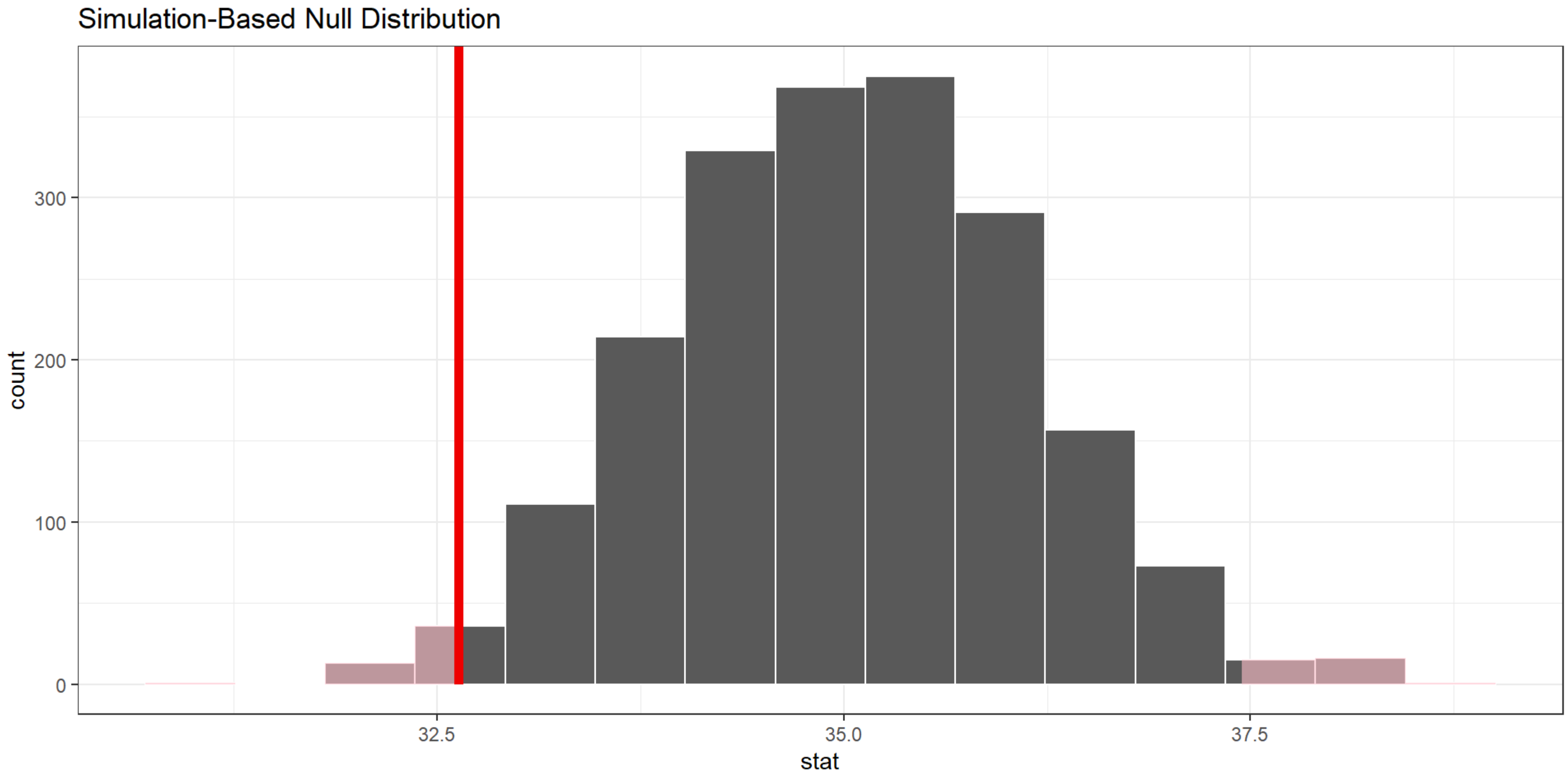
Bootstrap via `infer`

```r
observed_statistic <- sal |>
  specify(response = s_n1) |>
  calculate(stat = "mean")

set.seed(2022)
null_dist_1_sample <- sal |>
  specify(response = s_n1) |>
  hypothesize(null = "point", mu = 35) |>
  generate(reps = 2000, type = "bootstrap") |>
  calculate(stat = "mean")

null_dist_1_sample |>
  visualize() +
  shade_p_value(observed_statistic, direction = "two-sided")
```

Bootstrap via `infer`



Bootstrap p value via infer

```r
p_value_1_sample <- null_dist_1_sample |>
  get_p_value(obs_stat = observed_statistic,
              direction = "two-sided")

p_value_1_sample
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1   0.029
```

Bootstrap 95% CI via `infer`

```
ci_1_sample <- null_dist_1_sample |>
  get_confidence_interval(point_estimate = observed_statistic,
                          level = 0.95, type = "se")

ci_1_sample
```

```
# A tibble: 1 × 2
  lower_ci upper_ci
     <dbl>    <dbl>
1     30.5     34.8
```

Bootstrap 95% CI via Hmisc

```r
set.seed(43134)
smean.cl.boot(sal$s_n1, conf.int = 0.95, B = 2000)
```

```
    Mean    Lower    Upper 
32.63086 30.34701 34.84223 
```

Wilcoxon Signed Rank

```r
wilcox.test(sal$s_n1, mu = 35, conf.int = TRUE, conf.level = 0.95)
```

```
	Wilcoxon signed rank test with continuity correction

data:  sal$s_n1
V = 912, p-value = 0.05304
alternative hypothesis: true location is not equal to 35
95 percent confidence interval:
 30.17004 35.00498
sample estimates:
(pseudo)median 
      32.70503 
```

Question B Results

Procedure	*p*	Estimate	95% CI
t	0.039	\(\hat{\mu} = \bar{x}\) = 32.63	(30.38, 34.88)
Bootstrap via `infer`	0.029	\(\hat{\mu} = \bar{x}\) = 32.63	(30.5, 34.8)
Bootstrap via `Hmisc`	< 0.05	\(\hat{\mu} = \bar{x}\) = 32.63	(30.35, 34.84)
Wilcoxon	0.053	ps.-med. = 32.71	(30.17, 35.00)

What can we conclude?

What’s in the rest of these slides

Another one-sample example, with some slight variations in approach, and more details in some spots.

Session Information

```r
sessionInfo()
```

```
R version 4.2.1 (2022-06-23 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22000)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base
```

For Home Study

Key Question for Home Study

Can we test to see whether the mean of the base-10 logged (RNA copies per milliliter) in the NP samples is detectably different from 4.5?

- Our null hypothesis is \(H_0: \mu = 4.5\) vs. the two-tailed alternative \(H_A: \mu \neq 4.5\).

Create a new variable

```r
sal <- sal |>
  mutate(np_log = log10(np_titre))
```

DTDP: Base-10 Logarithm of np_titre

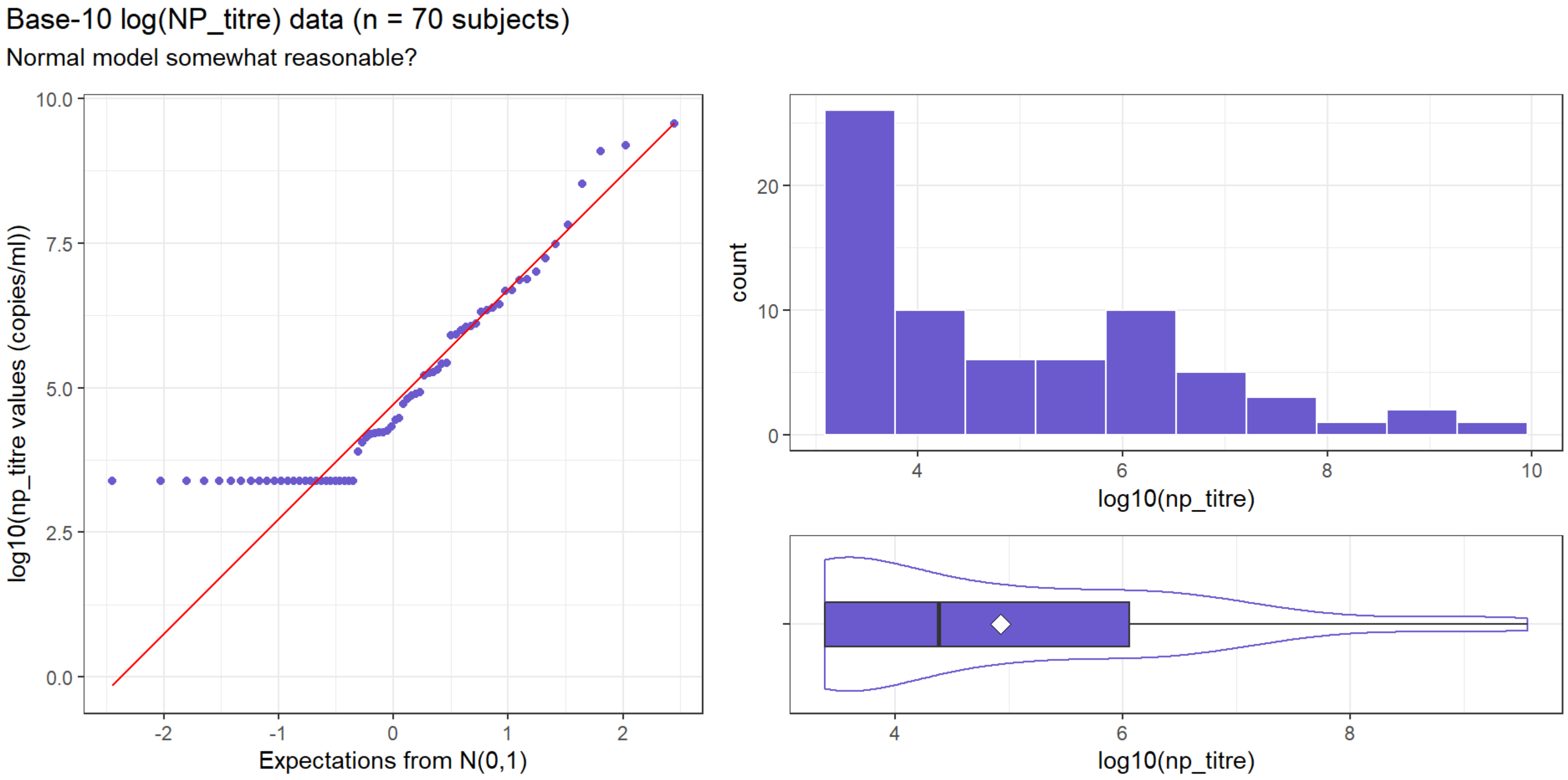
```r
p1 <- ggplot(sal, aes(sample = log10(np_titre))) +
  geom_qq(col = "slateblue") + geom_qq_line(col = "red") +
  theme(aspect.ratio = 1) +
  labs(y = "log10(np_titre values (copies/ml))",
       x = "Expectations from N(0,1)")

p2 <- ggplot(sal, aes(x = log10(np_titre))) +
  geom_histogram(bins = 10, col = "white", fill = "slateblue")

p3 <- ggplot(sal, aes(x = log10(np_titre), y = "")) +
  geom_violin(col = "slateblue") +
  geom_boxplot(fill = "slateblue", width = 0.3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 3, fill = "white") +
  labs(y = "")

p1 + (p2/p3 + plot_layout(heights = c(2,1))) +
  plot_annotation(title = "Base-10 log(NP_titre) data (n = 70 subjects)",
                  subtitle = "Normal model somewhat reasonable?")
```

DTDP: Base-10 Logarithm of `np_titre`



Numerical Summaries

Raw `np_titre` data:

```r
mosaic::favstats(~ np_titre, data = sal) |> kbl() |> kable_minimal(font_size= 24)
```

min	Q1	median	Q3	max	mean	sd	n	missing
2400	2400	24350	1147500	3.64e+09	98527107	489700726	70	0

Base-10 logarithm of `np_titre`:

```r
mosaic::favstats(~ log10(np_titre), data = sal) |> kbl(digits = 3) |> kable_minimal()
```

min	Q1	median	Q3	max	mean	sd	n	missing
3.38	3.38	4.382	6.06	9.561	4.927	1.669	70	0

Note the log of the mean (\(log_{10}(98527107)\) = 7.994) isn't the mean of the logs (4.927).

Our Assumptions

Suppose that

- logged cells/ml results across the population of all inpatients with a SARS-CoV-2 diagnosis follow a Normal distribution (with mean \(\mu\) and standard deviation \(\sigma\).)
- the 70 adults in our `sal` tibble are a random sample from that population.

What else do we know?

We know the sample mean (4.93) of our outcome, but we don’t know \(\mu\), the mean across **all** inpatients with a SARS-CoV-2 diagnosis.

So we need to estimate it, by producing a **confidence interval for the true (population) mean** \(\mu\).

Available Methods

To build a point estimate and confidence interval for the population mean, we could use

1. A **t-based** estimate and confidence interval, available from an intercept-only linear model, or (equivalently) a t test.
    - This approach will require an assumption that the population comes from a Normal distribution.

Available Methods

2. A **bootstrap** confidence interval, which uses resampling to estimate the population mean.
    - This approach won't require the Normality assumption, but has other constraints.
3. A **Wilcoxon signed rank** approach, but that won't describe the mean, only a pseudo-median.
    - This also doesn't require the Normality assumption, but no longer describes the population mean (or median) unless the population can be assumed symmetric. Instead it describes the *pseudo-median*.

Starting with A Good Answer

Indicator variable regression to produce a t-interval.

```r
model1 <- lm(np_log ~ 1, data = sal)
tidy(model1, conf.int = TRUE, conf.level = 0.95) |>
  select(term, estimate, std.error, conf.low, conf.high, p.value) |>
  kbl(digits = 2) |> kable_minimal(font_size = 24)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	4.93	0.2	4.53	5.33	0

- Point estimate of population mean (\(\mu\)) is 4.93 mm Hg.
- 95% confidence interval is (4.53, 5.33) for \(\mu\).

Interpreting the 95% CI for \(\mu\)

- Some people think this means that there is a 95% chance that the true mean of the population, \(\mu\), falls between 4.53 and 5.33. Not true.
- The population mean \(\mu\) is a constant **parameter** of the population of interest. That constant is not a random variable, and does not change.
- So the actual probability of the population mean falling inside that range is either 0 or 1.

So what do we have confidence in?

Our confidence is in our process.

- It’s in the sampling method (random sampling) used to generate the data, and in the assumption that the population follows a Normal distribution.
- It’s captured in our accounting for one particular type of error (called *sampling error*) in developing our interval estimate, while assuming all other potential sources of error are negligible.

Interpreting the CI

Our 95% confidence interval for \(\mu\) is (4.53, 5.33).

If we used this method to sample data from the target population of inpatients with SARS-CoV-2 and build 100 such intervals, then 95 of them would contain the true population mean. We don’t know whether this particular interval contains \(\mu\), though.

- 100(1 - \(\alpha\))%, here 95%, or 0.95 is the *confidence* level.
- \(\alpha\) = 5%, or 0.05 is called the *significance* level.

This approach is identical to a t test.

Formula for the t-based CI?

Many confidence intervals follow a general strategy using a point estimate \(\pm\) a margin for error.

We build a 100(1-\(\alpha\))% confidence interval using the \(t\) distribution, using the sample mean \(\bar{x}\), the sample size \(n\), and the sample standard deviation \(s_x\). The two-sided 100(1-\(\alpha\))% confidence interval is:

\[\bar{x} \pm t_{\alpha/2, n-1} (\frac{s_x}{\sqrt{n}})\]

Ancillary Elements of the CI

- \(SE(\bar{x}) = \frac{s_x}{\sqrt{n}}\) is the standard error of the sample mean
- The margin of error for this CI is \(t_{\alpha/2, n-1} (\frac{s_x}{\sqrt{n}})\).
- \(t_{\alpha/2, n-1}\) is the value that cuts off the top \(\alpha/2\) percent of the \(t\) distribution, with \(n - 1\) degrees of freedom. Obtain in R with:

```
qt(alphaover2, df = n-1, lower.tail=FALSE)
```

Five Steps to Complete a Hypothesis Test

1. Specify the null hypothesis, \(H_0\)
2. Specify the research or alternative hypothesis, \(H_1\), sometimes called \(H_A\)
3. Specify the approach to be used to make inferences to the population based on sample data.
   - We must specify \(\alpha\), the probability of incorrectly rejecting \(H_0\) that we are willing to accept. Often, we use \(\alpha = 0.05\)
4. Obtain the data, and summarize it to obtain an appropriate point estimate and confidence interval (and maybe a \(p\) value.)
5. Draw a conclusion

Five Steps of a Hypothesis Test

1. Specify the null hypothesis.

Here, we have \(H_0: \mu = 4.5\), or in general \(H_0: \mu = \mu_0\).

2. Specify the research (alternative) hypothesis.

Here, we have \(H_A: \mu \neq 4.5\)

Five Steps of a Hypothesis Test

3. Calculate a test statistic based on the data and null hypothesis value.

The one-sample t test uses as its test statistic:

\[t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.927-4.5}{1.669/\sqrt{70}} = 2.14 \]

where \(\bar{x}\) is the sample mean and \(s\) is the sample standard deviation.

Five Steps of a Hypothesis Test

4. Obtain an appropriate p value by comparing the test statistic to the reference distribution identified by the null hypothesis, and sample size.

- Here, we have n = 70, so we have n - 1 = 69 degrees of freedom for our estimate.
- In R, we can obtain a two-tailed p value for our test statistic of 2.14 using 69 degrees of freedom with:

```r
pt(2.14, df = 69, lower.tail = FALSE)*2
```

```
[1] 0.03589283
```

Step 5: Make a decision (based on the p value)

This is the part I don’t like. Everything up to here is fine.

If we establish a tolerable Type I error rate, \(\alpha\), then

- if \(p < \alpha\), we can reject our null hypothesis in favor of the alternative.
- if \(p \geq \alpha\), we must fail to reject our null hypothesis.

Comparing the p-value to \(\alpha\)

So, if \(\alpha = 0.05\), and we have

- \(H_0: \mu = 4.5\) vs. \(H_A: \mu \neq 4.5\)
- and obtain a two-tailed \(p\) value = 0.036

what should we conclude?

One-Sample t test

- \(H_0: \mu = 4.5\) vs. the two-tailed alternative \(H_A: \mu \neq 4.5\).

```r
t.test(sal$np_log, mu = 4.5)
```

```
	One Sample t-test

data:  sal$np_log
t = 2.1409, df = 69, p-value = 0.03582
alternative hypothesis: true mean is not equal to 4.5
95 percent confidence interval:
 4.529121 5.325034
sample estimates:
mean of x 
 4.927078 
```

Tidied One-Sample t test

- \(H_0: \mu = 4.5\) vs. the two-tailed alternative \(H_A: \mu \neq 4.5\).

```r
tt <- t.test(sal$np_log, mu = 4.5)
tidy(tt) |> kbl(digits = 2) |> kable_paper(font_size = 24)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
4.93	2.14	0.04	69	4.53	5.33	One Sample t-test	two.sided

Approaches that Don’t Assume Normality

Hypothesis Testing about a Population Mean (or Median) that don’t require the assumption of Normality:

1. with infer() tools, a randomization test for the mean relying on the bootstrap
2. via a bootstrap confidence interval for the mean (or the median)
3. with the Wilcoxon signed-rank test (tests population pseudo-median)

A randomization test using infer tools

This is a randomization-based analog to the 1-sample t test.

First, we calculate the observed statistic:

```
observed_statistic <- sal |>
  specify(response = np_log) |>
  calculate(stat = "mean")

observed_statistic
```

```
Response: np_log (numeric)
# A tibble: 1 × 1
   stat
  <dbl>
1  4.93
```

Next Goal

Our next goal is to compare this observed statistic to a null distribution, generated under the assumption that the mean was actually 4.5, to get a sense of how likely it would be for us to see this observed mean if the true logged counts/ml in the population was really 4.5.

Again, our null hypothesis is \(H_0: \mu = 4.5\) vs. the two-tailed alternative \(H_A: \mu \neq 4.5\).

Using the bootstrap to generate the null distribution

We can generate the null distribution using the bootstrap.

- In the bootstrap, for each replicate, a sample of size equal to the input sample size is drawn (with replacement) from the input sample data.
- This allows us to get a sense of how much variability we'd expect to see in the entire population so that we can then understand how unlikely our sample mean would be.

Generate the null distribution

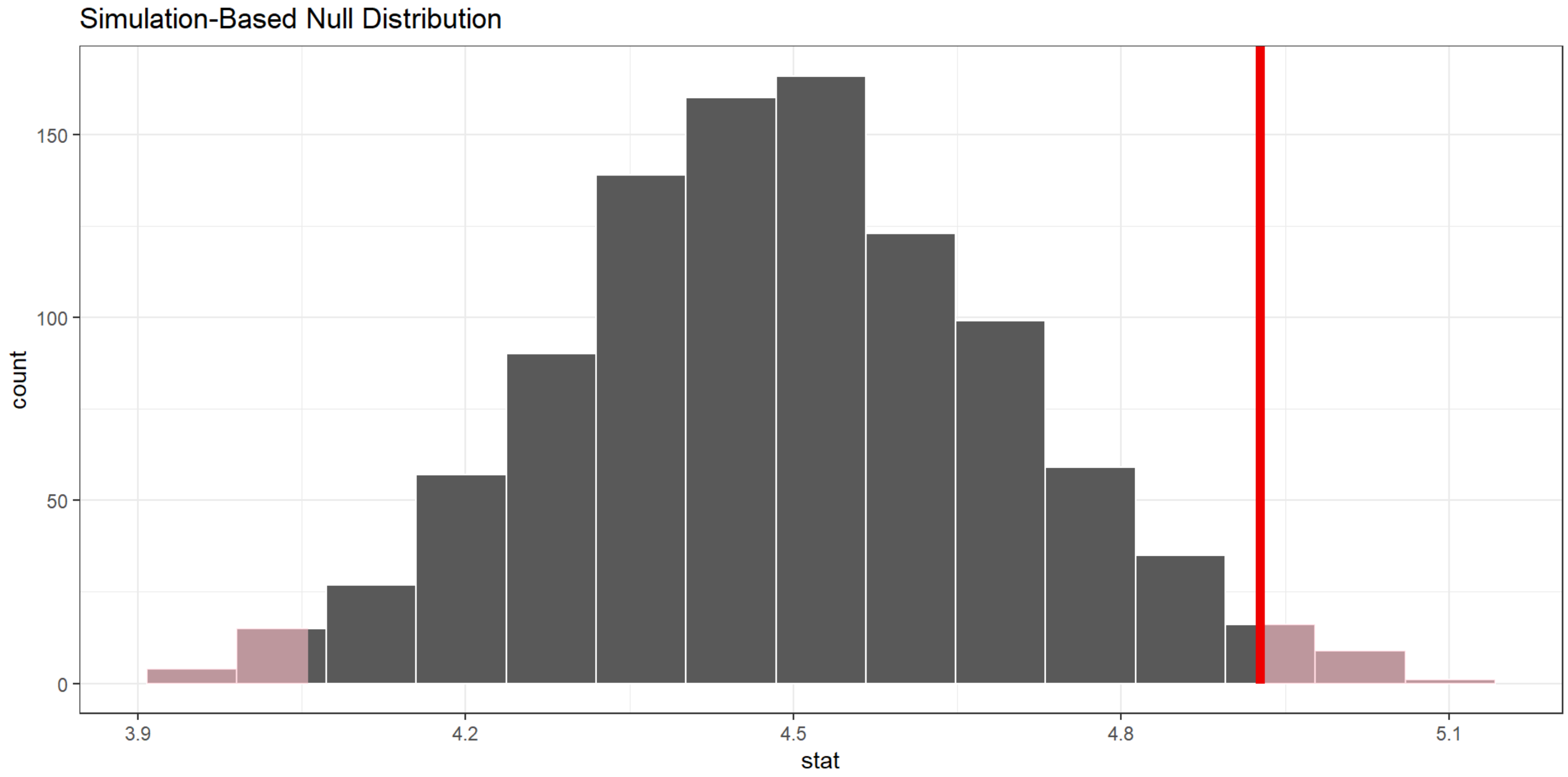
Using the bootstrap, we need to set a seed so we can replicate our work later:

```r
set.seed(431)
null_dist_1_sample <- sal |>
  specify(response = np_log) |>
  hypothesize(null = "point", mu = 4.5) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean")
```

Resulting Null Distribution

Get a sense of where our observed statistic falls.

```
null_dist_1_sample |>
  visualize() +
  shade_p_value(observed_statistic, direction = "two-sided")
```



Calculating the *p* value

```r
p_value_1_sample <- null_dist_1_sample |>
  get_p_value(obs_stat = observed_statistic,
              direction = "two-sided")

p_value_1_sample
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1    0.03
```

Thus, if the true mean logged counts/ml was really 4.5, our approximation of the probability that we would see a test statistic as or more extreme than is approximately 0.04.

What if our null hypothesis changed?

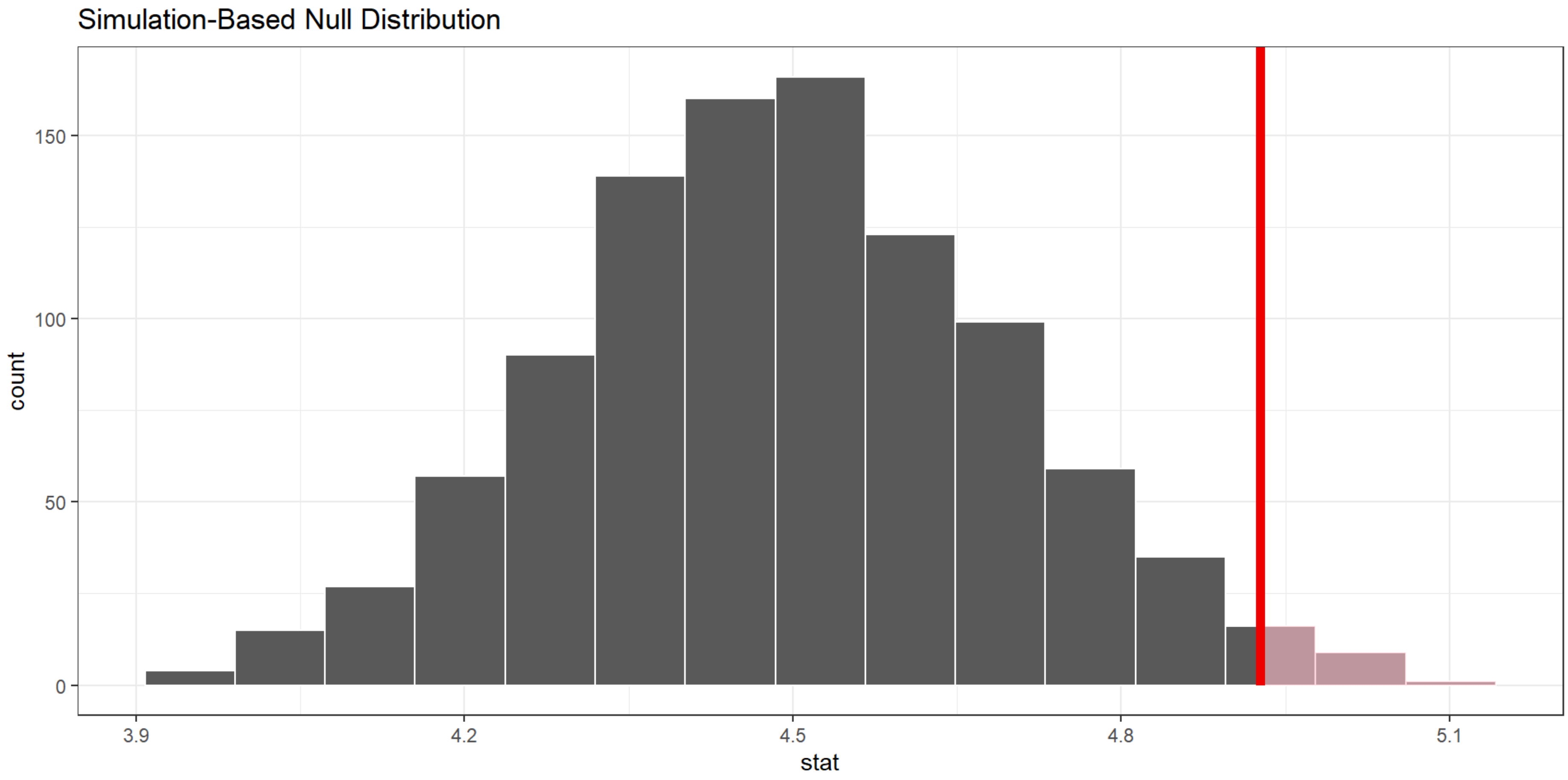
We currently have \(H_0: \mu = 4.5\) vs. the two-tailed alternative \(H_A: \mu \neq 4.5\), and we obtain a p value of about 0.04.

Consider \(H_0: \mu <= 4.5\) vs. the one-tailed alternative \(H_A: \mu > 4.5\).

- If we used the same null distribution we created previously, then we should have a p value of about 0.04/2 = 0.02

Visualize 1-tailed p value

```r
null_dist_1_sample |>
  visualize() +
  shade_p_value(observed_statistic, direction = "greater")
```



Calculating the *p* value

Again, we’re now looking at \(H_0: \mu <= 4.5\) vs. the one-tailed alternative \(H_A: \mu > 4.5\).

```
p_value_1_sample <- null_dist_1_sample |>
  get_p_value(obs_stat = observed_statistic,
              direction = "greater")

p_value_1_sample
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1   0.015
```

One-Sided t test and CI?

```r
t.test(sal$np_log, mu = 4.5, alternative = "greater")
```

```

    One Sample t-test

data:  sal$np_log
t = 2.1409, df = 69, p-value = 0.01791
alternative hypothesis: true mean is greater than 4.5
95 percent confidence interval:
 4.594493      Inf
sample estimates:
mean of x 
 4.927078 
```

Bootstrap Mean via Confidence Interval and `smean.cl.boot()`

- \(H_0: \mu = 4.5\) vs. the two-tailed alternative \(H_A: \mu \neq 4.5\).

95% confidence interval via bootstrap…

```r
set.seed(43102)
smean.cl.boot(sal$np_log, conf.int = 0.95, B = 2000)
```

```
    Mean    Lower    Upper 
4.927078 4.567397 5.332890 
```

What can we conclude from this interval about our hypotheses?

Bootstrap CI for \(\mu\)

What the computer does:

1. Resample the data with replacement, until it obtains a new sample that is equal in size to the original data set.
2. Calculates the statistic of interest (here, a sample mean.)
3. Repeat the steps above many times (default is 1,000 with our approach) to obtain a set of 1,000 results (here: 1,000 sample means.)
4. Sort those 1,000 results in order, and estimate the 90% confidence interval for the population value based on the middle 90% of the 1,000 bootstrap samples.
5. Send us a result, containing the sample estimate, and the bootstrap 90% confidence interval estimate for the population value.

The bootstrap idea can be used to produce interval estimates for almost any population parameter, not just the mean.

What about p values?

```r
set.seed(431)
smean.cl.boot(sal$np_log, conf = 0.9)
```

```
    Mean    Lower    Upper 
4.927078 4.587132 5.259062 
```

1. What can we say about the *p* value for \(H_0: \mu = 4.5\) vs. \(H_A: \mu \neq 4.5\) based on this bootstrap?
2. What can we say about the *p* value for \(H_0: \mu = 5\) vs. \(H_A: \mu \neq 5\) based on the bootstrap?

When is a Bootstrap CI for \(\mu\) Reasonable?

The interval will be reasonable as long as we are willing to believe that:

- the original sample was a random sample (or at least a completely representative sample) from a population,
- and that the samples are independent of each other (selecting one subject doesn't change the probability that another subject will also be selected)
- and that the samples are identically distributed (even though that distribution may not be Normal.)

It is still possible that the results can be both:

- **inaccurate** (i.e. they can, include the true value of the unknown population mean less often than the stated confidence probability) and
- **imprecise** (i.e., they can include more extraneous values of the unknown population mean than is desirable).

The Wilcoxon Signed Rank Procedure

The Wilcoxon signed rank approach builds interval estimates for the population *pseudo-median* when the population can only be assumed to be symmetric.

- For any sample, the pseudo-median is defined as the median of all of the midpoints of pairs of observations in the sample.
- As it turns out, if you're willing to assume the population is **symmetric** (but not necessarily Normally distributed) then the pseudo-median is equal to the population median.

Wilcoxon based 95% confidence interval

```r
wilcox.test(sal$np_log, mu = 4.5, conf.int = TRUE, conf.level = 0.95)
```

```
	Wilcoxon signed rank test with continuity correction

data:  sal$np_log
V = 1478, p-value = 0.1664
alternative hypothesis: true location is not equal to 4.5
95 percent confidence interval:
 4.342776 5.182472
sample estimates:
(pseudo)median 
      4.789161 
```

Interpreting the Wilcoxon Signed Rank CI

Again, the pseudo-median would be close to the sample mean and median if the population actually follows a symmetric distribution.

```
mosaic::favstats(~ np_log, data = sal)
```

```
      min       Q1  median       Q3      max     mean       sd  n missing
 3.380211 3.380211 4.38235 6.059674 9.561101 4.927078 1.668991 70       0
```