431 Class 14

Thomas E. Love, Ph.D.

2022-10-18

Today's Agenda

- New Examples: Two Studies from the Cleveland Clinic
- Comparing Two Population Means
  - In a Study using Independent Samples
    - T tests (Pooled and Welch) and Bootstrap and Wilcoxon Rank Sum Approaches
  - In a Study using Matched (Paired) Samples
    - Reviewing what we discussed in Class 13

Version 2022-10-17 00:10:39

Today's Packages

```r
source("c14/data/Love-boost.R") # for bootdif() function

library(broom)
library(glue) # for inserting results into plots
library(Hmisc) # for smean.cl.boot(), mostly
library(infer) # tidy inference methods
library(kableExtra) # for neatening tables
library(janitor); library(naniar)
library(tidyverse)

theme_set(theme_bw())
```

Comparing Means: Two Study Designs

You can afford n = 400 outcome measurements, and want to compare the outcome's mean under exposure A to the outcome's mean under exposure B.

1. Select a random sample of 200 people from the target population, each of whom provide an outcome under exposure A, and then an outcome under exposure B.
2. Select a random sample of 400 people from the target population, then randomly assign 200 to receive exposure A and the remaining 200 to receive exposure B.

- What are the main differences between the studies?
- Study 1 uses **paired samples**, since each result under exposure A is matched to the exposure B result from the same subject. Calculating paired B - A differences for each subject makes sense.
- Study 2 uses **independent samples**, where there is no pairing/matching of individual observations across exposures.

A Study Involving Two Independent Samples

The Supraclavicular Data

These come from the Cleveland Clinic’s Statistical Education Dataset Repository, which is a great source of examples for me, but not for your Project B.

```r
supra_raw <- read_csv("c14/data/Supraclavicular.csv", show_col_types = F) |
  clean_names() |> mutate(subject = as.character(subject))

dim(supra_raw)
```

```
[1] 103  17
```

The Supraclavicular data come from Roberman et al. “Combined Versus Sequential Injection of Mepivacaine and Ropivacaine for Supraclavicular Nerve Blocks”. *Reg Anesth Pain Med* 2011; 36: 145-50.

Supraclavicular Study Objective (in brief)

> This study consisted of 103 patients, aged 18 to 70 years, who were scheduled to undergo an upper extremity procedure suitable for supraclavicular anesthesia. These procedures were expected to be associated with considerable postoperative pain.

> We tested the hypothesis that sequential supraclavicular injection of 1.5% mepivacaine followed 90 seconds later by 0.5% ropivacaine provides a quicker onset and a longer duration of analgesia than an equidose combination of the 2 local anesthetics.

> Patients were randomly assigned to either (1) combined group-ropivacaine and mepivacaine mixture; or (2) sequential group-mepivacaine followed by ropivacaine. The primary outcome was time to 4-nerve sensory block onset.

All quotes here are from the Supraclavicular study description

Study Description (1/2)

- We selected 103 subjects from the population of all people:
  - ages 18-70 years
  - scheduled to undergo an upper extremity procedure suitable for supraclavicular anesthesia
  - who would have been eligible to participate in the study (details are fuzzy)

Study Description (2/2)

- We have randomly allocated subjects to one of two treatments (sequential or mixture.)
- For each subject, we have an outcome (onset time) associated with the treatment they received.
- The subjects were sampled from the population of interest independently of each other, so that the outcomes we see are not matched (or paired) in any way.

Key Question

Does the (true population) mean onset time differ between the two treatments?

Variables of interest to us (n = 103)

Variable	Description
`group`	1 = mixture, 2 = sequential (randomly assigned)
`onset_sensory`	Time to 4 nerve sensory block onset (min.)

Creating the `supra` analytic data

```r
supra <- supra_raw |>
  mutate(trt = fct_recode(factor(group), "mixture" = "1",
                            "sequential" = "2")) |>
  rename(onset = onset_sensory) |>
  select(subject, trt, onset, group)

head(supra)
```

```
# A tibble: 6 × 4
  subject trt        onset group
  <chr>   <fct>      <dbl> <dbl>
1 1       mixture        0     1
2 2       sequential     7     2
3 3       sequential    24     2
4 4       mixture        4     1
5 5       mixture       30     1
6 6       sequential     4     2
```

Summaries: Onset by Treatment

```r
mosaic::favstats(onset ~ trt, data = supra) |>
  kbl(digits = 2) |> kable_classic(font_size = 28)
```

trt	min	Q1	median	Q3	max	mean	sd	n	missing
mixture	0	4	7.5	13.5	50	11.42	11.46	52	0
sequential	1	7	10.0	19.5	50	15.25	12.08	51	0

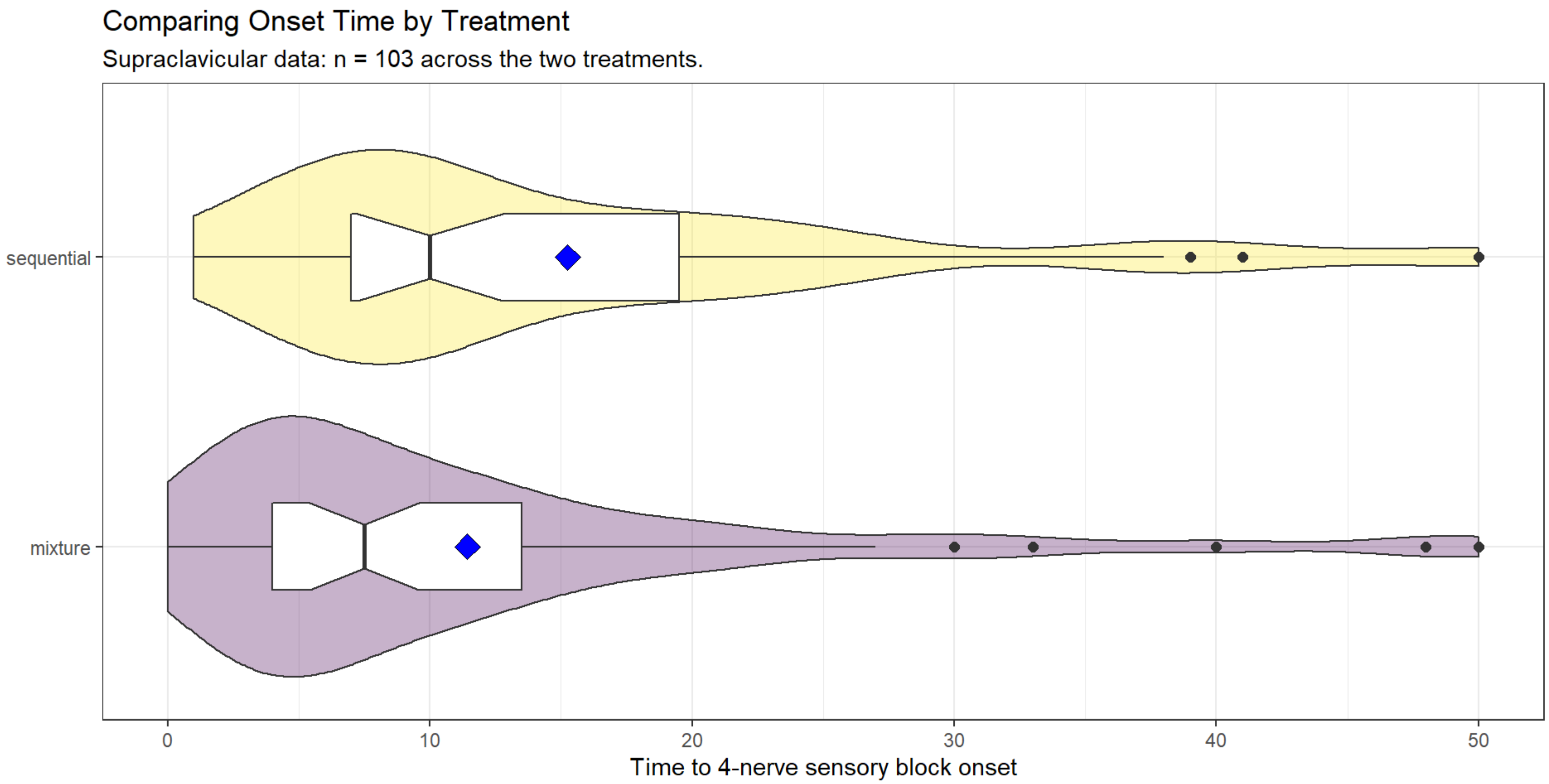
If we’re comparing the difference in means, in which order will we want to see the two `trt`s?

DTDP: Compare onset by treatment

We'll add a blue diamond to indicate the means in each group, too.

```r
ggplot(supra, aes(x = trt, y = onset)) +
  geom_violin(aes(fill = trt)) +
  geom_boxplot(width = 0.3, outlier.size = 2, notch = T) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  guides(fill = "none") +
  scale_fill_viridis_d(alpha = 0.3) +
  coord_flip() +
  labs(y = "Time to 4-nerve sensory block onset",
       x = "",
       title = "Comparing Onset Time by Treatment",
       subtitle = glue("Supraclavicular data: n = ", nrow(supra), " across
```

DTDP: Compare onset by treatment



Formal Language of Hypothesis Testing

- Null hypothesis \(H_0\)
  - \(H_0\): population mean onset time with sequential = population mean onset time with mixture
  - \(H_0\): difference in population means (sequential - mixture) = 0

Formal Language of Hypothesis Testing

- Alternative (research) hypothesis \(H_A\) or \(H_1\)
  - \(H_A\): population mean onset time with sequential \(\neq\) population mean onset time with mixture
  - \(H_A\): difference in population means (sequential - mixture) \(\neq\) 0

Two (related) next steps

1. Given the data, we can then calculate an appropriate test statistic, then compare that test statistic to an appropriate probability distribution to obtain a \(p\) value. Small \(p\) values favor \(H_A\) over \(H_0\).
2. More usefully, we can use an appropriate probability distribution to help use the data to construct an appropriate **confidence interval** for the difference in population means.

Comparing Two Population Means

With **independent samples** (as in this scenario) we have at least four alternatives.

1. Compare population means using a pooled t test or CI.
2. Compare population means using a Welch’s t test/ CI.
3. Compare population means using a bootstrap approach to generate a test or CI.
4. Compare the difference in locations using a Wilcoxon rank sum test or CI.

Option 1: t test

Compare population means using a pooled t test or confidence interval

- This assumes equal population variances of the outcome in the two treatment groups.
- This also assumes Normality of the outcome in each of the two treatment groups.
- This is the result of a linear model of outcome ~ treatment.

Model yielding pooled t-test

- Pooled t test and associated 90% CI for the difference in population means.

```
m1 <- lm(onset ~ trt, data = supra)

tidy(m1, conf.int = TRUE, conf.level = 0.90) |>
  kbl(digits = 3) |> kable_classic_2(font_size = 28)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	11.423	1.632	6.999	0.000	8.714	14.133
trtsequential	3.832	2.319	1.652	0.102	-0.019	7.682

What can we conclude about the difference in means?

Two-Sample `t.test()` approach

We can obtain the same results for the t test comparing two independent samples, and assuming equal variances, with…

```r
t.test(onset ~ trt, data = supra,
       var.equal = TRUE, conf.level = 0.90)
```

```

    Two Sample t-test

data:  onset by trt
t = -1.652, df = 101, p-value = 0.1016
alternative hypothesis: true difference in means between group mixture and
group sequential is not equal to 0
90 percent confidence interval:
 -7.68230689  0.01865682
sample estimates:
   mean in group mixture mean in group sequential 
                11.42308                 15.25490 
```

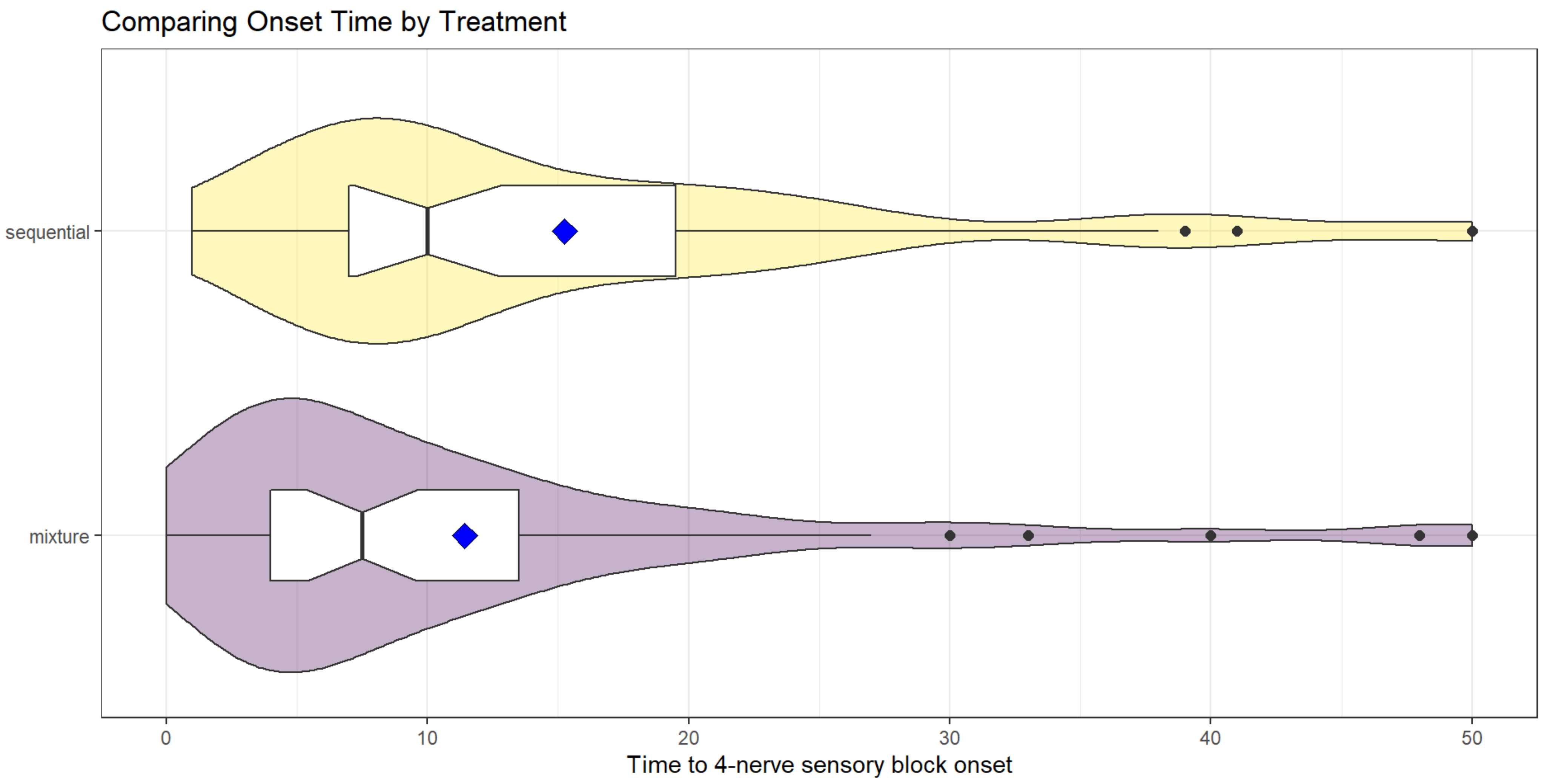
Assessing Pooled T test Assumptions

In preparing a t test with equal variances, we assume that:

- each of the samples (sequential and mixture) are drawn from a Normally distributed population
- each of those populations have the same variance

Do these seem like reasonable assumptions in this case? (See plot on next slide)

Onset Time by Treatment



Option 2: Welch's t test

Let's first consider dropping the "equal variances" assumption. Instead, we'll compare the population means using Welch's t test or confidence interval

- This does not assume equal population variances of the outcome.
- This does assume Normality of the outcome in each of the two treatment groups.

Welch’s t test approach

Here is the Welch’s t test comparing two independent samples, without assuming equal variances…

```r
t.test(onset ~ trt, data = supra, conf.level = 0.90)
```

```

	Welch Two Sample t-test

data:  onset by trt
t = -1.6512, df = 100.47, p-value = 0.1018
alternative hypothesis: true difference in means between group mixture and group sequential is not equal to 0
90 percent confidence interval:
 -7.6845015  0.0208514
sample estimates:
   mean in group mixture mean in group sequential 
                11.42308                 15.25490 
```

Comparing the two "T tests"

```r
t1 <- t.test(onset ~ trt, data = supra, conf.level = 0.90,
             var.equal = TRUE)
w1 <- t.test(onset ~ trt, data = supra, conf.level = 0.90)

bind_rows(tidy(t1), tidy(w1)) |>
  select(method, estimate, conf.low, conf.high, p.value) |>
  kbl(digits = 3) |> kable_classic_2(font_size = 24, full_width = F)
```

method	estimate	conf.low	conf.high	p.value
Two Sample t-test	-3.832	-7.682	0.019	0.102
Welch Two Sample t-test	-3.832	-7.685	0.021	0.102

Balanced Design?

It turns out that if we have a **balanced design** (equal sample sizes in the two groups) then the Pooled t approach and the Welch’s t approach yield essentially the same results.

- So these will be very similar if \(n_1 = n_2\).

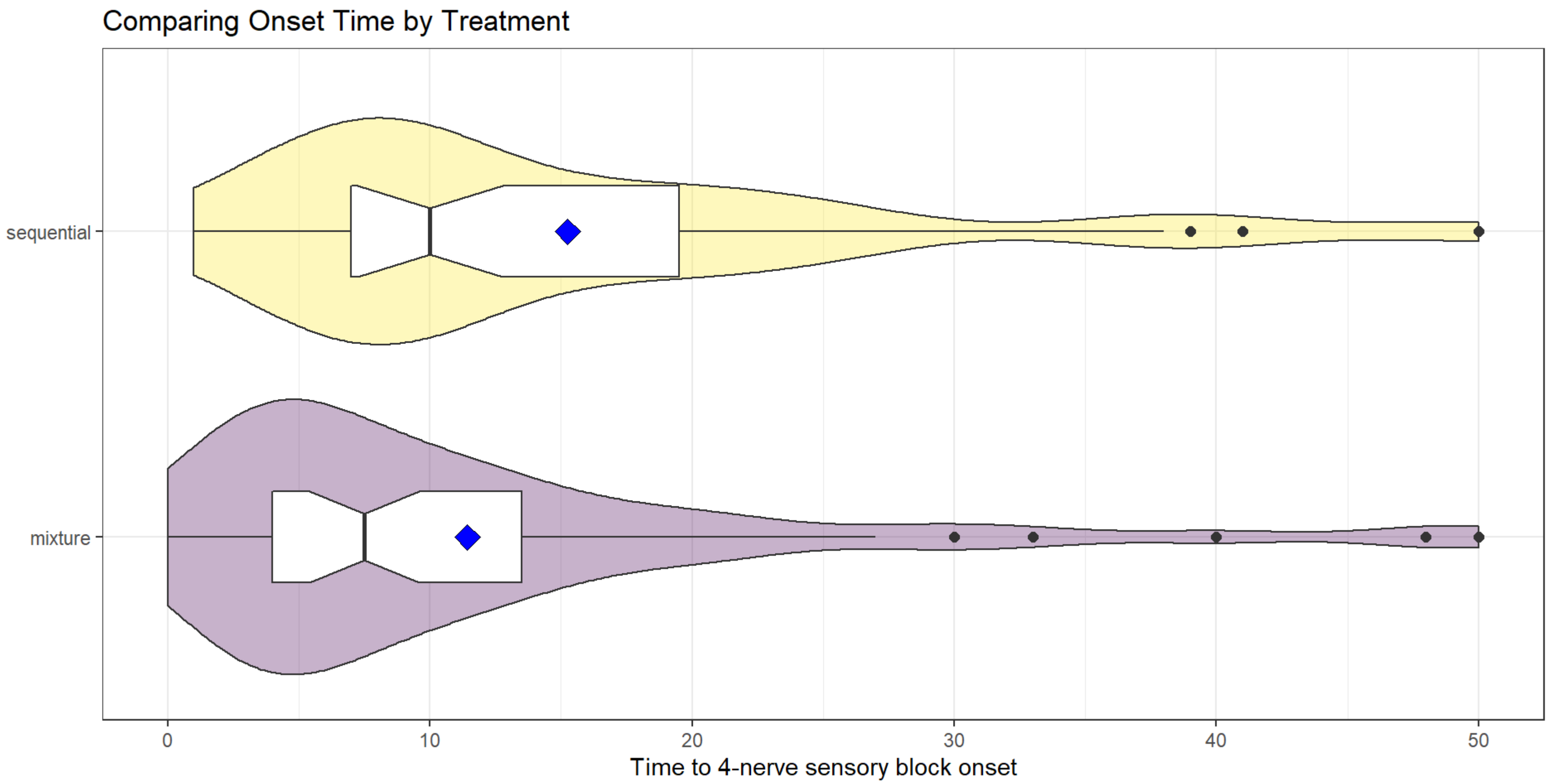
```
# A tibble: 2 × 2
  trt            n
  <fct>      <int>
1 mixture       52
2 sequential    51
```

What about the Normality assumption?

```r
ggplot(supra, aes(x = trt, y = onset)) +
  geom_violin(aes(fill = trt)) +
  geom_boxplot(width = 0.3, outlier.size = 2, notch = T) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  guides(fill = "none") +
  scale_fill_viridis_d(alpha = 0.3) +
  coord_flip() +
  labs(y = "Time to 4-nerve sensory block onset",
       x = "",
       title = "Comparing Onset Time by Treatment")
```

- Does it seem reasonable to assume that the onset times are Normally distributed across the populations of sequential and mixed subjects, based on these samples of data?

What about the Normality assumption?



Option 3: Bootstrap

Compare the population means using a bootstrap approach to generate a confidence interval.

- This does not assume either equal population variances or Normality.

Using `infer`: Obtaining Test Statistic

```r
obs_diff_means <- supra |>
  specify(formula = onset ~ trt) |>
  calculate(stat = "diff in means", order = c("sequential", "mixture"))

obs_diff_means
```

```
Response: onset (numeric)
Explanatory: trt (factor)
# A tibble: 1 × 1
   stat
  <dbl>
1  3.83
```

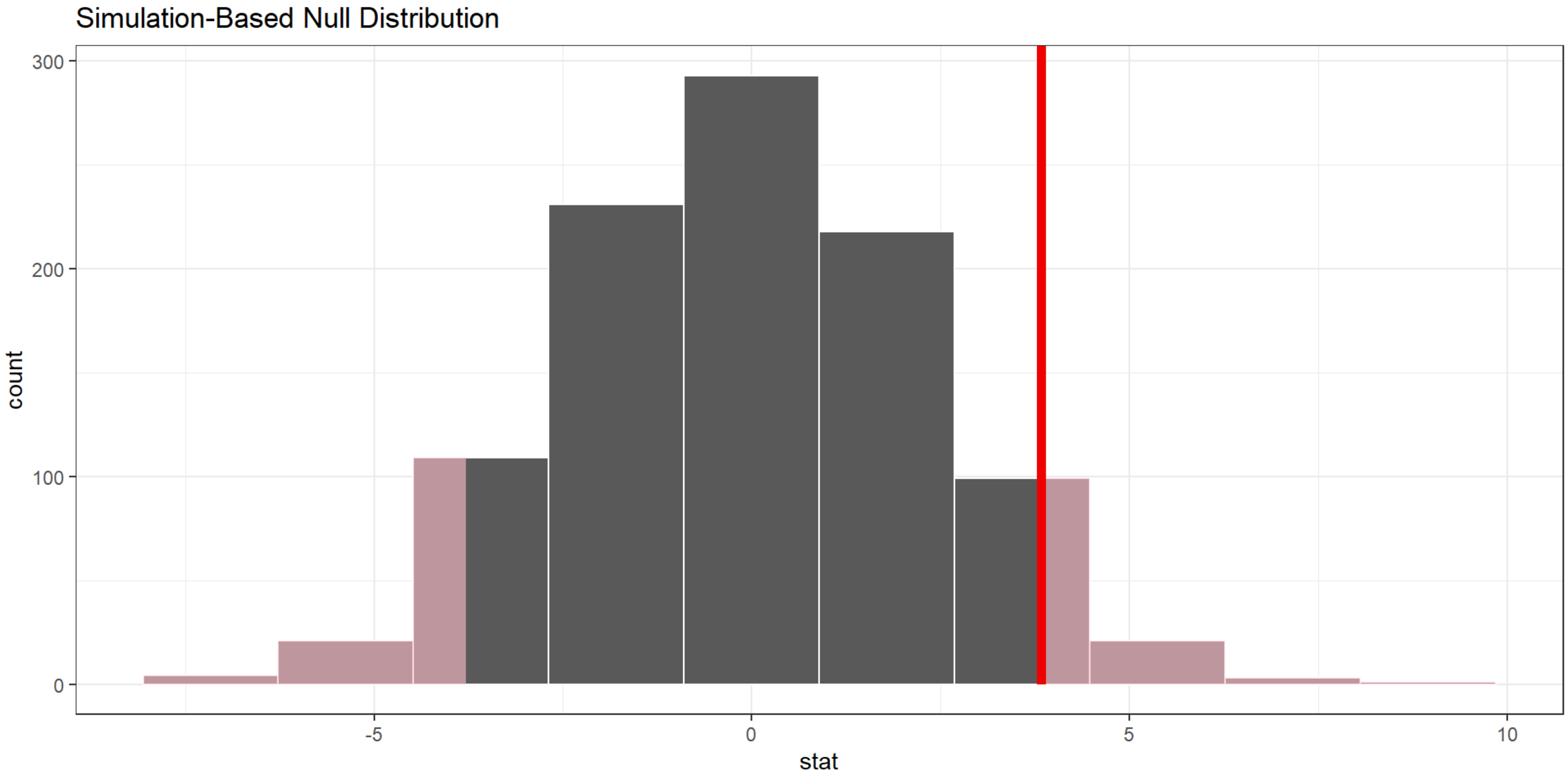
Using infer: Null Distribution

```r
set.seed(432) ## set a seed
null_distribution_supra <- supra |>
  specify(formula = onset ~ trt) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in means", order = c("sequential", "mixture"))

head(null_distribution_supra)
```

```
Response: onset (numeric)
Explanatory: trt (factor)
Null Hypothesis: independence
# A tibble: 6 × 2
  replicate  stat
      <int> <dbl>
1         1 -3.97
2         2  1.04
3         3 -3.94
4         4 -2.15
5         5 -3.90
6         6 -2.62
```

Visualize p value

```r
visualize(null_distribution_supra, bins = 10) +
  shade_p_value(obs_stat = obs_diff_means, direction = "both")
```



Get p-value from permutation test

```r
null_distribution_supra |>
  get_p_value(obs_stat = obs_diff_means, direction = "both")
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1   0.106
```

90% Confidence Interval via Bootstrap

```r
set.seed(432)
bootstrap_distribution <- supra |>
  specify(formula = onset ~ trt) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("sequential", "mixture"))

percentile_ci <- bootstrap_distribution |>
  get_confidence_interval(level = 0.90, type = "percentile")

percentile_ci
```

```
# A tibble: 1 × 2
  lower_ci upper_ci
     <dbl>    <dbl>
1  -0.0470     7.58
```

`bootdif` bootstrap CI approach

Consider the **bootstrap**, without assuming the population distributions are Normal, or have the same variance, at the expense of requiring some random sampling, which can lead to some conflicts.

- We’ll use the `bootdif()` function I’ve provided in the `Love-boost.R` script.

```
set.seed(20221018)
bootdif(y = supra$onset, g = supra$trt, conf.level = 0.90, B.reps = 2000)
```

```
Mean Difference            0.05            0.95 
      3.8318250       0.1230581       7.5400830 
```

Using a bootstrap approach

- If we'd set a different seed or selected a different number of bootstrap replications, we'd get a different result.

```
set.seed(431)
bootdif(y = supra$onset, g = supra$trt, conf.level = 0.90, B.reps = 2000)
```

```
Mean Difference            0.05            0.95 
     3.83182504     -0.08301282      7.55837104 
```

```
bootdif(y = supra$onset, g = supra$trt, conf.level = 0.90, B.reps = 10000)
```

```
Mean Difference            0.05            0.95 
     3.83182504      0.04654977      7.53755656 
```

- This doesn't mean to suggest that we "shop around" until we find an appealing result, of course.

Wilcoxon-Mann-Whitney rank sum

Compare the population locations with a Wilcoxon rank sum test or confidence interval

- This does not assume either equal population variances or Normality, but doesn’t describe the difference in population means or medians.
- The estimator for the rank sum test is a difference in location parameters.
    - This estimates the median of the difference between a sample from x and a sample from y.

Wilcoxon-Mann-Whitney Rank Sum Test

\(H_0\): Difference in Location Parameters is 0, vs. two-tailed \(H_A\): Difference in Location Parameters \(\neq\) 0

```r
wilcox.test(onset ~ trt, data = supra, alt = "two.sided", mu = 0,
            paired = FALSE, conf.int = TRUE, conf.level = 0.90)
```

```
	Wilcoxon rank sum test with continuity correction

data:  onset by trt
W = 974, p-value = 0.02027
alternative hypothesis: true location shift is not equal to 0
90 percent confidence interval:
 -5.999995 -1.000057
sample estimates:
difference in location 
             -3.000052 
```

Our Gathered Estimates

Method	\(\mu_S - \mu_M\)	90% CI	p-value
Pooled t	3.832	(-0.019, 7.682)	0.102
Welch’s t	3.832	(-0.021, 7.685)	0.102
Bootstrap A	3.832	(-0.047, 7.580)	0.106
Bootstrap B	3.832	(+0.123, 7.540)	< 0.10

Thinking about those estimates

All of these results are in minutes (recall 0.08 minutes = 4.8 seconds) so are these **clinically meaningful** differences in this context?

- Do these data involve random sampling?
- What population(s) do these data represent?
- What can we say about the *p* values associated with these approaches?

A Study Involving Two Matched (Paired) Samples

The Hypoxia MAP Data

From Cleveland Clinic’s Statistical Education Dataset Repository.

```r
hypox_raw <- read_csv("c14/data/HypoxiaMAP.csv", show_col_types = F) |>
  clean_names() |>
  mutate(subject = row_number())

dim(hypox_raw)
```

```
[1] 281  37
```

Source: Turan et al. “Relationship between Chronic Intermittent Hypoxia and Intraoperative Mean Arterial Pressure in Obstructive Sleep Apnea Patients Having Laparoscopic Bariatric Surgery” *Anesthesiology* 2015; 122: 64-71.

Background and Study Description

> [The Hypoxia MAP study] retrospectively examined the intraoperative blood pressures in 281 patients who had laparoscopic bariatric surgery between June 2005 and December 2009 and had a diagnosis of OSA within two preoperative years.

> Time-weighted average (TWA) intraoperative MAP was the main outcome in the study. MAP (or mean arterial pressure) is a term used to describe an average blood pressure in a subject.

> MAP is normally between 65 and 110 mmHg, and it is believed that a MAP > 70 mmHg is enough to sustain the organs of the average person. If the MAP falls below this number for an appreciable time, vital organs will not get enough oxygen perfusion, and will become hypoxic, a condition called ischemia.

Our Objective with these Data

We will focus today on two measurements of MAP for each subject (outside of some missing data).

- MAP1 = time-weighted average mean arterial pressure from ET intubation to trocar insertion, in mm Hg.
- MAP2 = time-weighted average mean arterial pressure from trocar insertion to the end of the surgery, in mm Hg.

We are interested in estimating the **difference** between the two MAP levels, across a population of subjects like those enrolled in this study.

Our Key Variables

- For each subject, we have two outcomes to compare: their MAP1 and their MAP2.

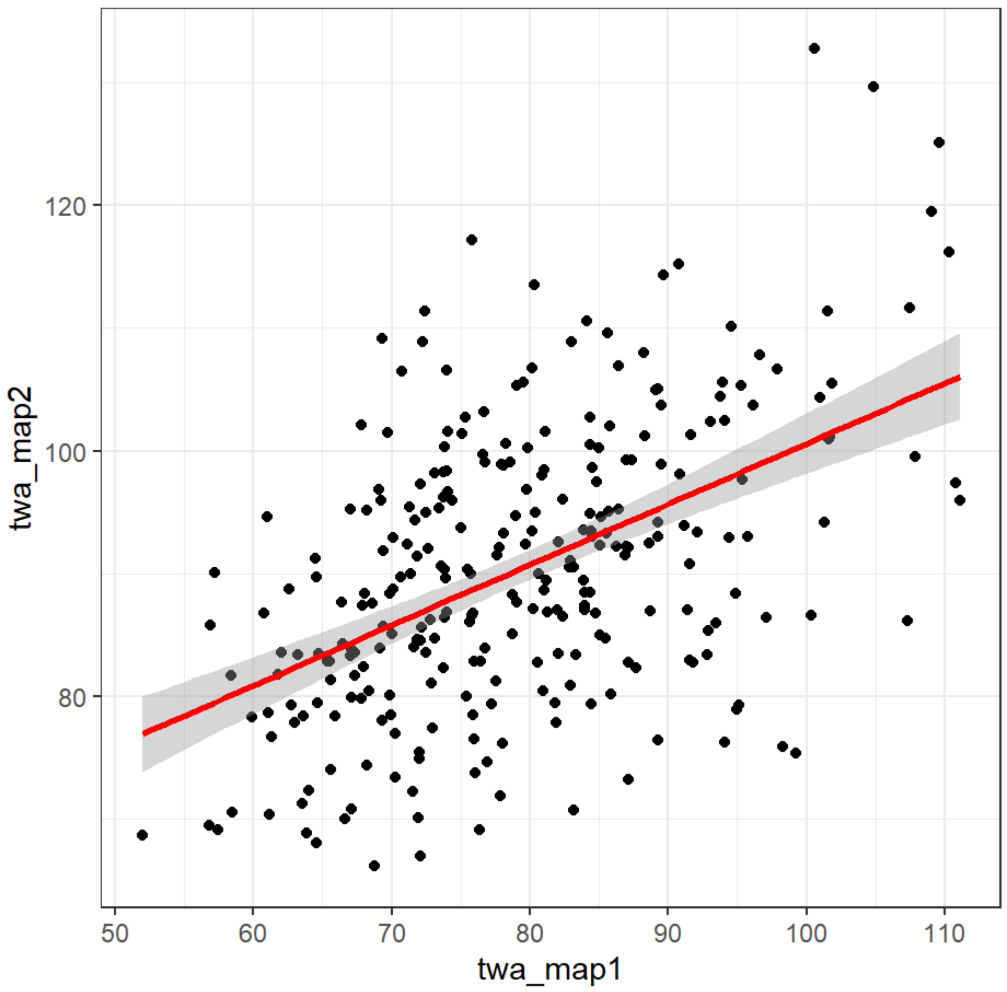
```
hypox <- hypox_raw |>
  select(subject, twa_map1, twa_map2) |>
  mutate(map_diff = twa_map2 - twa_map1)

head(hypox, 4)
```

```
# A tibble: 4 × 4
  subject twa_map1 twa_map2 map_diff
    <int>    <dbl>    <dbl>    <dbl>
1       1     67.9     87.4    19.5 
2       2     67.0     83.3    16.3 
3       3     91.6     83.0    -8.59
4       4     67.1     79.9    12.8 
```

We have Paired Samples in this setting

- Every MAP1 value is connected to the MAP2 value for the same subject. We say that the MAP1 and MAP2 are paired by subject.
- Are the pairings relatively strong?
  - As we'll see, the Pearson correlation of MAP1 and MAP2 across the subjects with complete data is 0.494.
  - Can we draw a plot?
- It makes sense to calculate the (paired) difference in MAP values for each subject, so long as there aren't any missing data.

Scatterplot of MAP 1 vs. MAP 2



Each subject provides a MAP1 and a MAP2

Are there any missing values?

```r
miss_var_summary(hypox)
```

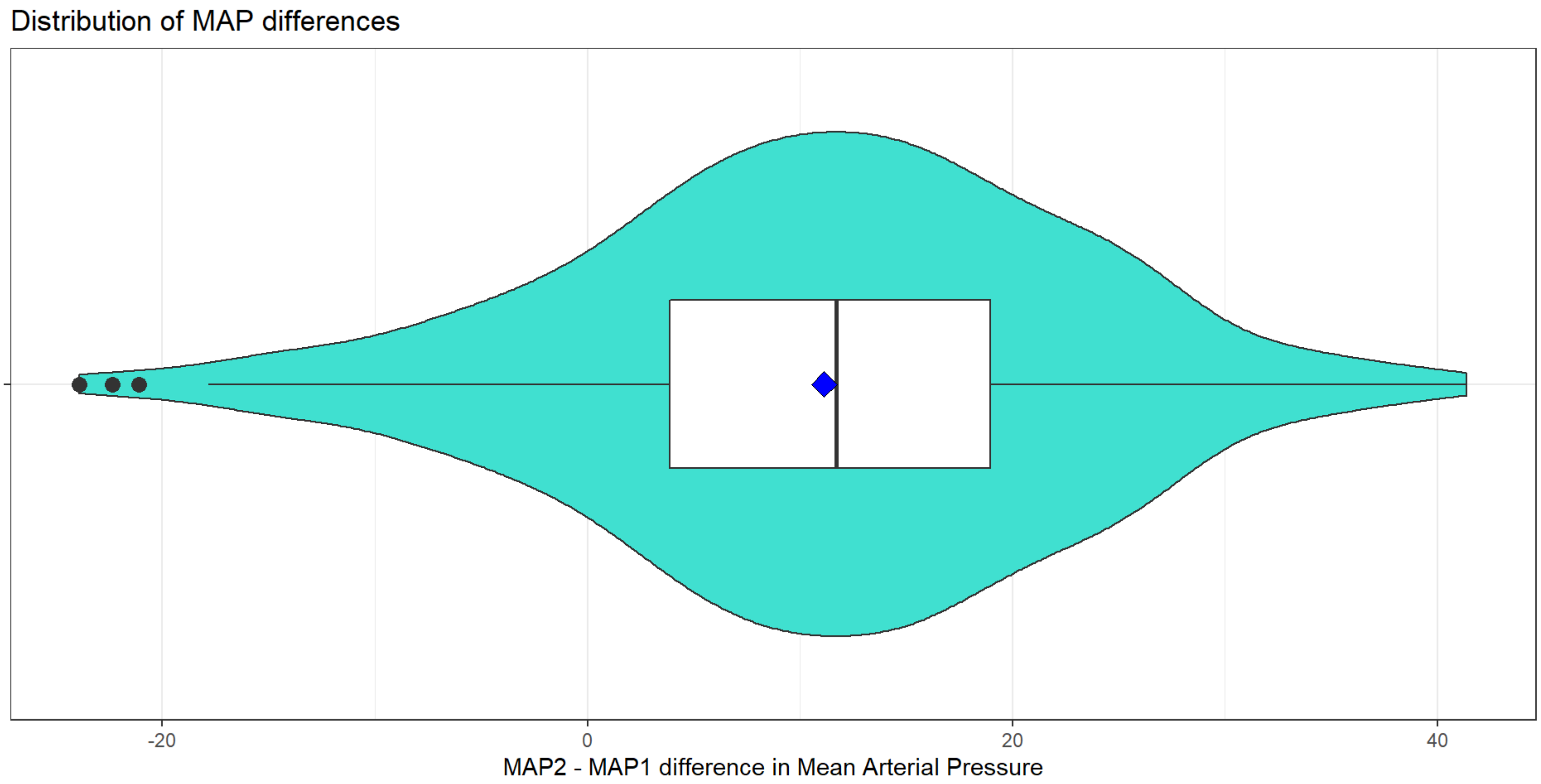
```
# A tibble: 4 × 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 twa_map1      4     1.42
2 map_diff      4     1.42
3 subject       0     0   
4 twa_map2      0     0   
```

```r
hypox <- hypox |> filter(complete.cases(map_diff))
```

Boxplot of the MAP differences

```r
ggplot(data = hypox, aes(x = map_diff, y = "")) +
  geom_violin(fill = "turquoise") +
  geom_boxplot(width = 0.3, outlier.size = 3) +
  stat_summary(fun = "mean", geom = "point",
               shape = 23, size = 4, fill = "blue") +
  labs(x = "MAP2 - MAP1 difference in Mean Arterial Pressure",
       y = "", title = "Distribution of MAP differences")
```

Boxplot of the MAP differences



Numerical Summaries

Is the mean of `map_diff` equal to the difference between the mean of `map2` and the mean of `map1`?

```r
res1 <- as_tibble(bind_rows(
  mosaic::favstats(~ twa_map1, data = hypox),
  mosaic::favstats(~ twa_map2, data = hypox),
  mosaic::favstats(~ map_diff, data = hypox))) |>
  mutate(item = c("map1", "map2", "map_diff")) |>
  select(item, n, mean, sd, min, median, max)

res1 |> kbl(digits = 2) |> kable_classic(font_size = 28, full_width = F)
```

item	n	mean	sd	min	median	max
map1	277	79.24	11.74	51.96	78.02	111.10
map2	277	90.38	11.69	66.17	89.74	132.71
map_diff	277	11.14	11.78	-23.90	11.71	41.37

Comparing Paired Samples

- Null hypothesis \(H_0\)
  - \(H_0\): population mean of paired differences (MAP2 - MAP1) = 0
- Alternative (research) hypothesis \(H_A\) or \(H_1\)
  - \(H_A\): population mean of paired differences (MAP2 - MAP1) \(\neq\) 0

Two (related) next steps

1. Given the data, we can then calculate the paired differences, then an appropriate test statistic based on those differences, which we compare to an appropriate probability distribution to obtain a \(p\) value. Again, small \(p\) values favor \(H_A\) over \(H_0\).
2. More usefully, we can calculate the paired differences, and then use an appropriate probability distribution to help use the data to construct an appropriate **confidence interval** for the population of those differences.

Paired T test via Linear Model

```
m3 <- lm(map_diff ~ 1, data = hypox)

summary(m3)$coef
```

```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 11.13646  0.7078298 15.73325 2.971357e-40
```

```
confint(m3, conf.level = 0.90)
```

```
               2.5 %   97.5 %
(Intercept) 9.743031 12.52989
```

```
summary(m3)$r.squared
```

```
[1] 0
```

Tidied Regression Model

```r
tidy(m3, conf.int = TRUE, conf.level = 0.90) |>
  select(term, estimate, conf.low, conf.high) |>
  kbl(digits = 3) |> kable_minimal(full_width = F)
```

term	estimate	conf.low	conf.high
(Intercept)	11.136	9.968	12.305

```r
tidy(m3, conf.int = TRUE, conf.level = 0.90) |>
  select(term, estimate, std.error, statistic, p.value) |>
  kbl(digits = 3) |> kable_minimal(full_width = F)
```

term	estimate	std.error	statistic	p.value
(Intercept)	11.136	0.708	15.733	0

Paired T test via t.test

```r
t.test(hypox$map_diff, conf.level = 0.90)
```

```
	One Sample t-test

data:  hypox$map_diff
t = 15.733, df = 276, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
  9.968265 12.304660
sample estimates:
mean of x 
 11.13646 
```

Paired T CI yet another way

```
smean.cl.normal(hypox$map_diff, conf = 0.90)
```

```
     Mean     Lower     Upper 
11.136462  9.968265 12.304660 
```

The function `smean.cl.normal` (and that’s an L, not a 1 after C) comes from the `Hmisc` package.

So does the `smean.cl.boot` function we’ll see on the next slide, which will let us avoid the key assumption of Normality for the population of paired differences.

Bootstrap for Comparing Paired Means

```r
set.seed(2022)
Hmisc::smean.cl.boot(hypox$map_diff, conf = 0.90, B = 1000)
```

```
    Mean    Lower    Upper 
11.13646 10.01252 12.30191 
```

Gathered Estimates from our Paired Samples

Method	Estimate and 90% CI	Assumes Normality?
Paired t	11.14 (9.97, 12.30)	Yes
Bootstrap	11.14 (10.01, 12.30)	No

We estimate that the time-weighted average mean arterial pressure is 11.14 mm Hg higher (90% CIs shown above) after trocar insertion than it is during the period from ET intubation to trocar insertion, based on our sample of 277 subjects with complete data in this study.

Evaluating Our Estimates

- Does it matter much whether we assume Normality here?
- What can we say about the *p* values here?
- Is this a random sample of subjects?
- What population do these data represent?

Next Time

What if we want to compare population proportions/rates/percentages rather than means?

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
```