431 Class 16

github.com/THOMASELOVE/2019-431

2019-10-24

Today's Agenda (Notes Chapters 22-23)

- Statistical Inference and the `dm431` data: Comparing Population Means using Paired Samples
- The Five Steps of a Hypothesis Test

Today's Setup and Data

```
library(magrittr); library(janitor)
library(patchwork); library(here);
library(boot); library(broom)
library(tidyverse)

source(here("R", "Love-boost.R"))

dm431 <- readRDS(here("data", "dm431.Rds"))
```

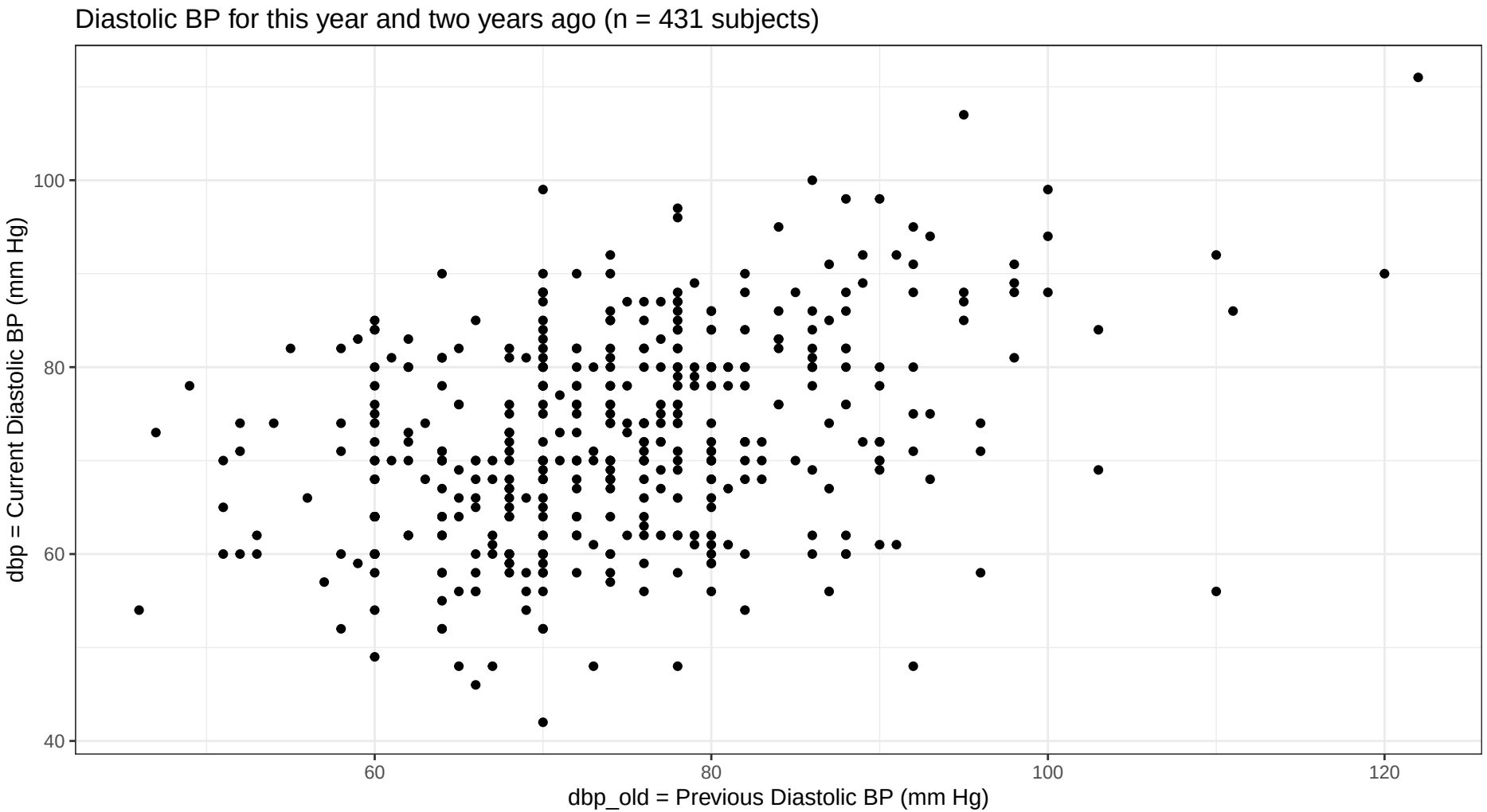
Comparing Population Means via Paired Samples

The dm431 data has current diastolic blood pressure (dbp), and diastolic blood pressure from two years ago (dbp_old) for each subject. Suppose we want to describe the mean DBP change in not just our sample, but instead the entire **population** (adults who live in NE Ohio with diabetes) over the past year.

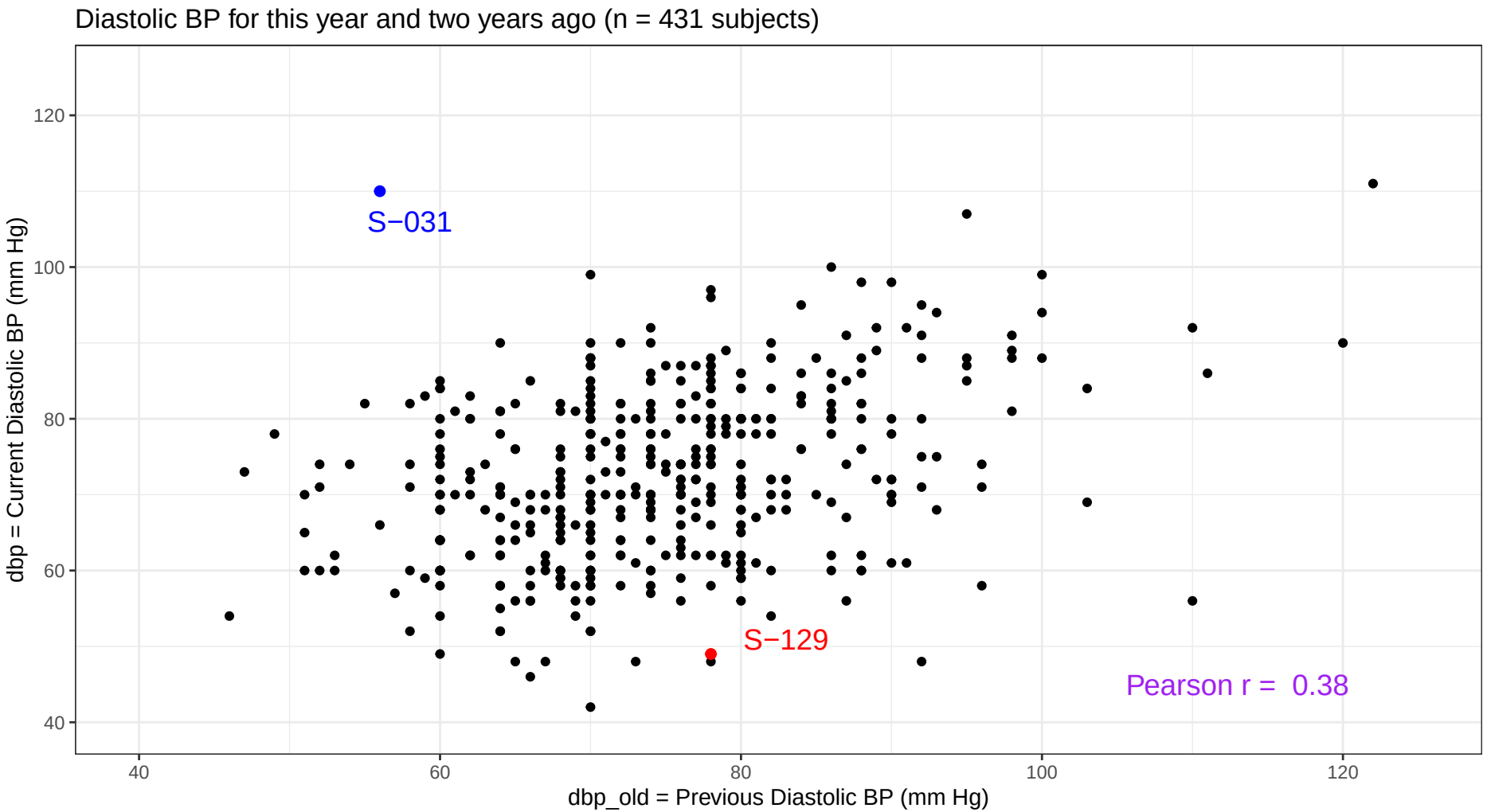
```
dm431 %>% select(dbp, dbp_old) %>% summary()
```

```
      dbp             dbp_old      
 Min.   : 42.00   Min.   : 46.00  
 1st Qu.: 64.00   1st Qu.: 68.00  
 Median : 72.00   Median : 74.00  
 Mean   : 72.62   Mean   : 74.65  
 3rd Qu.: 80.00   3rd Qu.: 80.00  
 Max.   :111.00   Max.   :122.00  
```

Each subject provides both a dbp_old and dbp



The Impact of Pairing



Creating a Before-After Plot

Each subject provides both a value for `dbp` and one for `dbp_old`. To build the plot we want here, we'll need to pivot the data to make it longer, as if we were working with independent samples.

```
## first re-express the data
dm_dbp_longer <-
  dm431 %>% select(subject, dbp, dbp_old) %>%
  pivot_longer(
    cols = starts_with("dbp"),
    names_to = "time", values_to = "DBP")
```

The data, in longer form

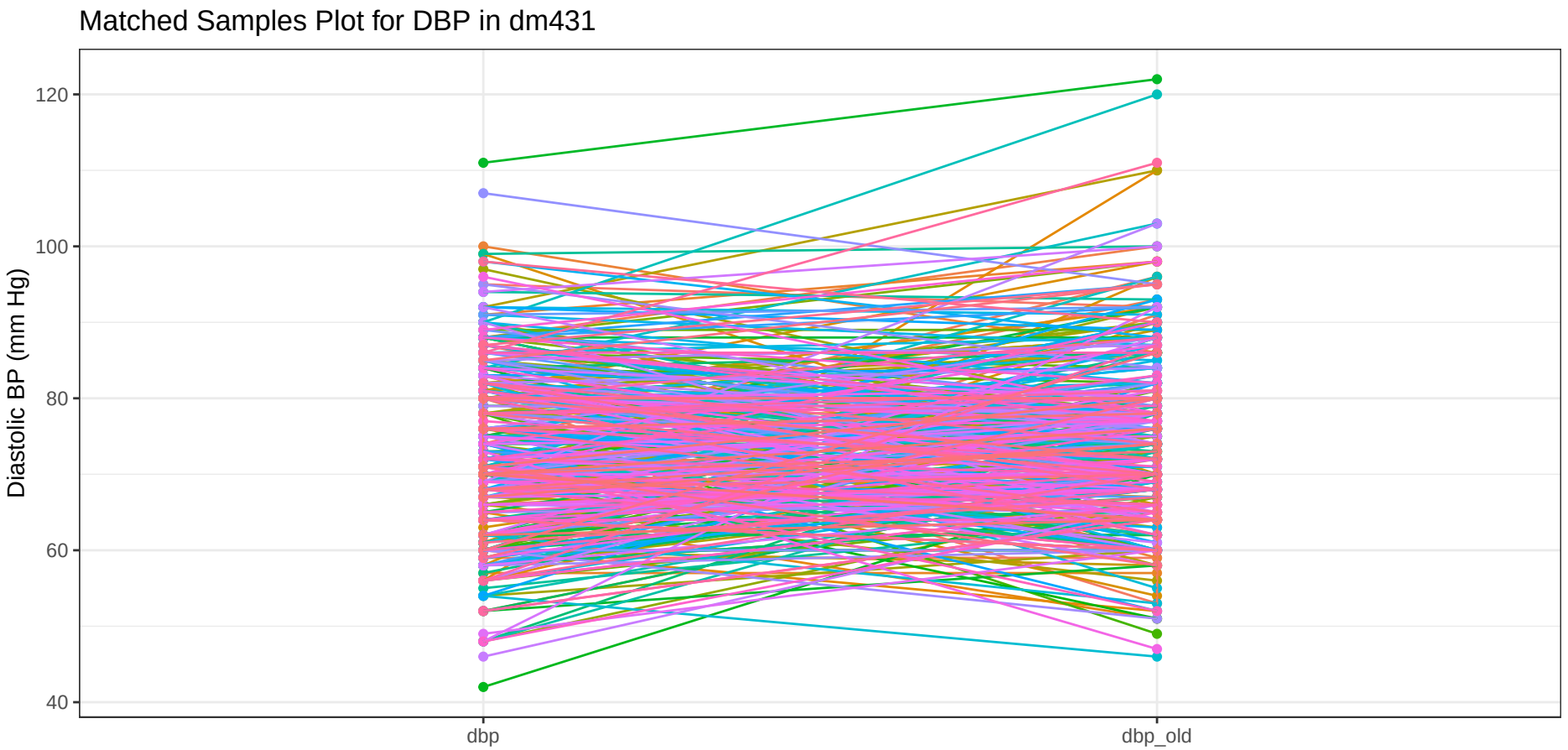
```
dm_dbp_longer %>% filter(subject %in% c("S-029", "S-131"))
```

```
# A tibble: 4 x 3
  subject time       DBP
  <chr>   <chr>    <dbl>
1 S-029   dbp         65
2 S-029   dbp_old     51
3 S-131   dbp         60
4 S-131   dbp_old     74
```

Code: Matched Samples ("After - Before") Plot

```
ggplot(dm_dbp_longer, aes(x = time, y = DBP,
                          group = subject)) +
    geom_point(aes(col = subject)) +
    geom_line(aes(col = subject)) +
    guides(col = FALSE) +
    theme_bw() +
    labs(title = "Matched Samples Plot for DBP in dm431",
         y = "Diastolic BP (mm Hg)", x = "")
```

Matched Samples Plot ("After - Before" Plot)



Patient S-141 is the patient on top, with `dbp` = 111, and `dbp_old` = 122.

Paired Samples? Calculate Paired Differences

```
dm431 <- dm431 %>%
    mutate(dbp_chg = dbp - dbp_old)

mosaic::favstats(~ dbp_chg, data = dm431) %>% round(., 2)
```

```
 min  Q1 median Q3 max  mean    sd   n missing
 -54 -10     -2  6  29 -2.03 12.42 431       0
```

Building Confidence Intervals using Paired Samples

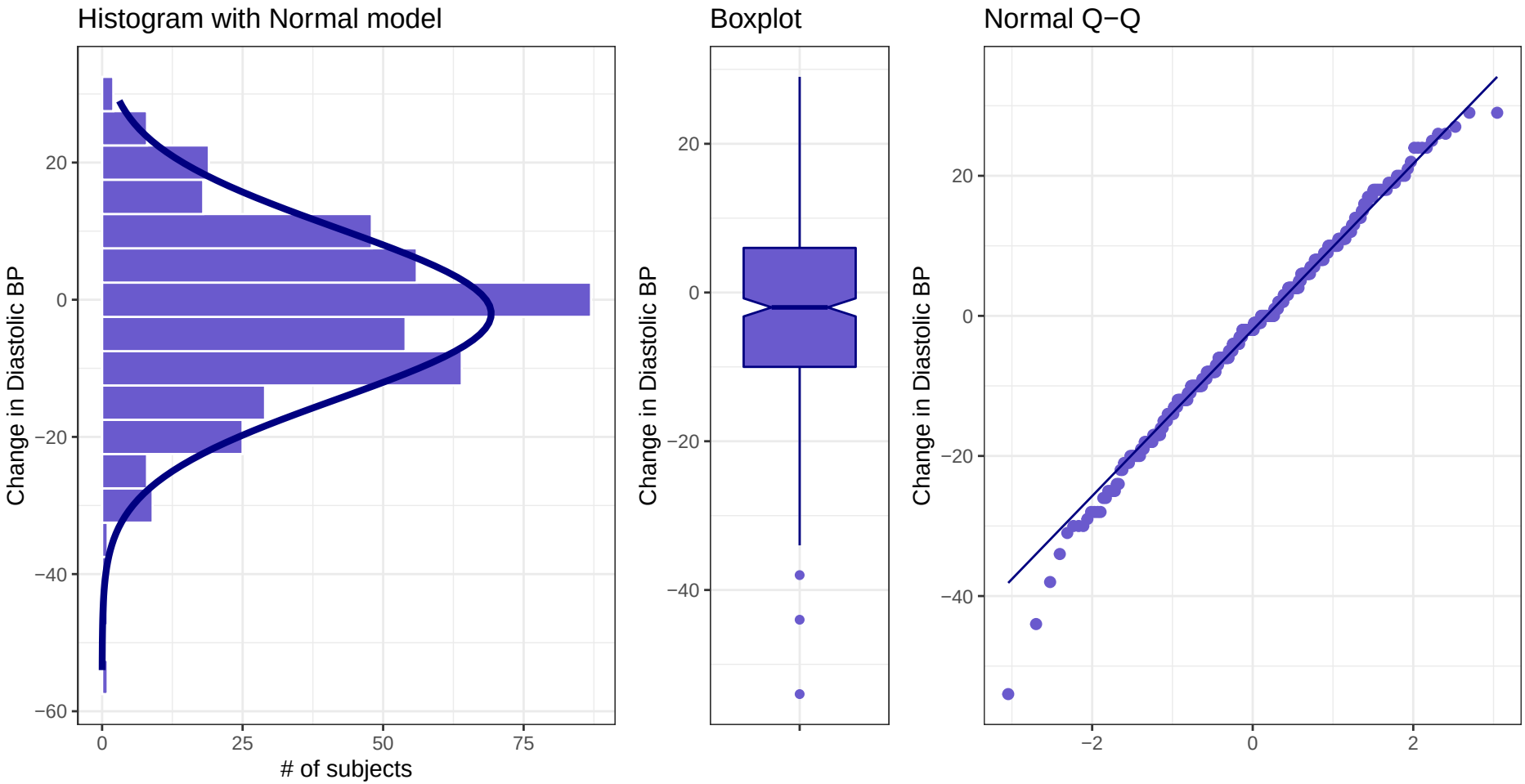
To build a point estimate and confidence interval for the population mean difference, we could use

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.
2. A **bootstrap** confidence interval, which uses resampling to estimate the population mean.
    - This approach won't require the Normality assumption, but has some 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*.

It's just the one-sample situation again, but with paired differences.

EDA for the Paired Differences

Change in Diastolic BP in mm Hg (Current minus Previous)



Intercept-only Regression for the Paired Differences

```
m1 <- lm(dbp_chg ~ 1, data = dm431)
tidy(m1, conf.int = TRUE, conf.level = 0.95) %>%
    select(estimate, conf.low, conf.high)
```

```
# A tibble: 1 x 3
  estimate conf.low conf.high
     <dbl>    <dbl>     <dbl>
1    -2.03    -3.20    -0.850
```

t test for the Paired Differences

```
dm431 %$%
  t.test(dbp, dbp_old, paired = TRUE, conf.level = 0.95)
```

```
    Paired t-test

data:  dbp and dbp_old
t = -3.3869, df = 430, p-value = 0.000772
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -3.2009949 -0.8500492
sample estimates:
mean of the differences
              -2.025522
```

Alternate Specifications

We can obtain the same result with

```
dm431 %$% t.test(dbp - dbp_old, conf.level = 0.95)
```

```
    One Sample t-test

data:  dbp - dbp_old
t = -3.3869, df = 430, p-value = 0.000772
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -3.2009949 -0.8500492
sample estimates:
mean of x
-2.025522
```

Five Steps to Complete a Hypothesis Test

1. Specify the null hypothesis, H_0 (which usually indicates that there is no difference between various groups of subjects)
2. Specify the research or alternative hypothesis, H_1, sometimes called H_A (which usually indicates that there is some difference or some association between the results in those same groups of subjects).
3. Specify the test procedure or test statistic to be used to make inferences to the population based on sample data.
    - Here we specify α, 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 a relevant test statistic, and a resulting p value.
5. Use the p value to either
    - **reject** H_0 in favor of the alternative H_A (concluding that there is a statistically significant difference/association at the α significance level)
    - or **retain** H_0 (and conclude that there is no statistically significant difference/association at the α significance level)

Step 1. The Null Hypothesis

- A null hypothesis is a statement about a population parameter, and it describes the current state of knowledge – the status quo – or our model for the world before the research is undertaken and data are collected.
- It often specifies an idea like "no difference" or "no association" in testable statistical terms.

The Null Hypothesis in the DBP in Diabetes Study

- Here, our null hypothesis will refer to the population mean of the paired differences in systolic blood pressure (in mm Hg) comparing the same subjects last year vs. this year.
- H_0: Population Mean DBP This Year = Population Mean DBP Last Year
  - If there is in fact no difference between the years, then the this year – last year difference will be zero.
- Symbolically, H_0: $\mu_d = 0$, where μ_d is the population mean (this year – last year) difference in systolic BP.
  - Of course, we've built confidence intervals for means like this already.

Step 2. The Alternative Hypothesis

- The alternative or research hypothesis, H_A, is in some sense the opposite of the null hypothesis.
- It specifies the values of the population parameter that are not part of H_0.
- If H_0 implies "no difference", then H_A implies that "there is a difference".

The Alternative Hypothesis in the DBP in Diabetes Study

Since our null hypothesis is

H_0: Population Mean DBP This Year – Population Mean DBP Last Year = 0, or $H_0 : \mu_d = 0$,

our alternative hypothesis will therefore cover all other possibilities:

H_A: Population Mean DBP This Year – Population Mean DBP Last Year $\neq$ 0, or $H_A : \mu_d \neq 0$.

Occasionally, we'll use a one-sided alternative, like $H_A : \mu_d < 0$, in which case, $H_0 : \mu_d \geq 0$.

Step 3: The Test Procedure and Assumptions

We want to compare the population mean of the paired differences, μ_d, to a fixed value, 0.

We must be willing to believe that the paired differences data are a random (or failing that, representative) sample from the population of interest, and that the samples were drawn independently, from an identical population distribution.

Given those assumptions, we have four possible strategies to complete our paired samples comparison:

The Four Strategies for Testing Paired Differences

a. Assume the paired differences come from a Normally distributed population, and perform a **one-sample t test** on the paired differences, and use the resulting p value to draw a conclusion about the relative merits of H_0 and H_A.

b. Or perform a **Wilcoxon signed-rank test** on the paired differences, which would be more appropriate than the t test if the population of paired differences was not Normally distributed, but was reasonably symmetric, and use the resulting p value.

c. Or develop a **bootstrap confidence interval** for the population mean of the paired differences, as we've done in the past. This wouldn't require an assumption about Normality. We'd then use that confidence interval to assess the relative merits of H_0 and H_A.

I'm skipping the **sign test**. See Section 23.7 in the Course Notes.

Step 4: Collect and summarize the data, usually with a *p* value

Of course, in this case, we've already gathered the data. The task now is to obtain and interpret the tests using each of the four procedures listed previously. The main task we will leave to the computer is the calculation of a **p value**.

Defining a *p* Value

The *p* value assumes that the null hypothesis is true, and estimates the probability, under those conditions (i.e. H_0 is true), that we would obtain a result as much in favor or more in favor of the alternative hypothesis H_A as we did.

- 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.

Using the p Value

The way we use the p value is to compare it to α, our pre-specified tolerance level for a certain type of error (Type I error, specifically – rejecting H_0 when it is in fact true.)

- If the p value is less than α, we will reject H_0 in favor of H_A
- If the p value is greater than or equal to α, we will retain H_0.

t Test for the DBP in Diabetes Study

```
dm431 %$% t.test(dbp - dbp_old)
```

```
    One Sample t-test

data:  dbp - dbp_old
t = -3.3869, df = 430, p-value = 0.000772
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -3.2009949 -0.8500492
sample estimates:
mean of x
-2.025522
```

- The alternative hypothesis is `true difference in means is not equal to 0`. Should we retain or reject H_0 at $\alpha = 0.05$?

Wilcoxon Signed Rank for the DBP in Diabetes data

```
dm431 %$% wilcox.test(dbp - dbp_old, conf.int=TRUE)
```

```
    Wilcoxon signed rank test with continuity
    correction

data:  dbp - dbp_old
V = 33462, p-value = 0.001962
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 -3.4999027 -0.9999552
sample estimates:
(pseudo)median
     -2.000001
```

- Should we reject or retain $H_0 : \mu_d = 0$ based on this test at a 5% significance level?

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** 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.

Bootstrap CI for the DBP Changes

Using a significance level of $\alpha = 0.05$ is equivalent to using a confidence level of 100(1-α)% = 95%:

```
set.seed(20191024)
Hmisc::smean.cl.boot(dm431$dbp_chg, conf.int = 0.95)
```

```
     Mean      Lower      Upper
-2.025522 -3.120650 -0.837529
```

Should we reject or retain $H_0 : \mu_d = 0$ at a 5% significance level?

What does this confidence interval suggest about the p value?

Step 5. Draw a conclusion, based on the confidence interval

We have the following results at the 5% significance level (equivalently, at the 95% confidence level, or with $\alpha = 0.05$):

- Sample Mean for DBP now is 72.62.
- Sample Mean for DBP two years ago is 74.65.
- Sample Mean Difference is thus -2.03

Approach	p value	95% CI for μ_d	Conclusion re: H_0: $\mu_d = 0$
t Test	0.0007	(-3.3, -0.9)	$p < 0.05$, so reject H_0
Wilcoxon	0.0020	(-3.5, -1.0)	$p < 0.05$, so reject H_0
Bootstrap	< 0.05	(-3.1, -0.8)	CI for μ excludes 0 so reject H_0

Our Conclusions for the DBP in Diabetes Study

So, in this case, using any of these methods, we conclude that:

a. there is a statistically detectable difference in DBP comparing the current data to two years ago at the 5% significance level, specifically it appears that the results were a little lower this year (our point estimate is 2.03 mm Hg lower.)

b. more importantly, our 95% confidence interval for the size of that difference appears to be on the order of 1-3 mm Hg, which isn't likely to be a medically important improvement.

Did pairing help in this situation to reduce noise?

- Was there a positive correlation of `dbp` and `dbp_old`? Yes, it was 0.38, so there was some reduction in nuisance variation at the subject level.
- What if we did this (incorrectly) assuming independent samples?

Using dm_dbp_longer: independent samples

```
dm_dbp_longer %$% t.test(DBP ~ time, var.equal = TRUE)
```

```
    Two Sample t-test

data:  DBP by time
t = -2.6586, df = 860, p-value = 0.007993
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -3.5209002 -0.5301439
sample estimates:
    mean in group dbp mean in group dbp_old
              72.62413                74.64965
```

Results from independent samples approaches

Comparing the Diastolic BP for the current data (now) to the previous data (old) without accounting for the fact that the same people provided the data in each sample.

Procedure	p for $H_0 : \mu_{now} = \mu_{old}$	95% CI for $\mu_{now} - \mu_{old}$
Pooled t test	0.008	(-3.5, -0.5)
Welch t test	0.008	(-3.5, -0.5)
Rank Sum test	0.025	(-3.0, -0.5)
Bootstrap CI	$p > 0.05$	(-3.5, -0.5)

- What changes here when we (incorrectly) ignore the pairing?

Note I used the seed 20191024 to obtain the bootstrap result.

Paired Samples Study Designs

- Using a paired samples design means we carefully sample matched sets of subjects in pairs, so that the sampled subjects in each pair are as similar as possible, except for the exposure of interest.
- Each observation in one exposure group is matched to a single observation in the other exposure group, so that taking paired differences is a rational thing to do.
- Since every subject must be matched to exactly one subject in the other group, the sizes of the groups must be equal.

A Few Comments on Significance

- **A significant effect is not necessarily the same thing as an interesting effect.** For example, results calculated from large samples are nearly always "significant" even when the effects are quite small in magnitude. Before doing a test, always ask if the effect is large enough to be of any practical interest. If not, why do the test?
- **A non-significant effect is not necessarily the same thing as no difference.** A large effect of real practical interest may still produce a non-significant result simply because the sample is too small.
- **There are assumptions behind all statistical inferences.** Checking assumptions is crucial to validating the inference made by any test or confidence interval.