431 Class 15

Thomas E. Love

2017-10-17

Today's R Setup

```
devtools::install_github('jtleek/slipper')

library(slipper); library(tidyverse)

dm192 <- read.csv("data/dm192.csv") %>% tbl_df

source("Love-boost.R")
```

Statistical Inference and the dm192 data

1. Hypothesis Testing and P Values
2. Comparing Two Population Means using Paired Samples
3. Comparing Two Population Means using Independent Samples
4. Some Early Thoughts on Power and Sample Size

- Assignment 4's deadline.

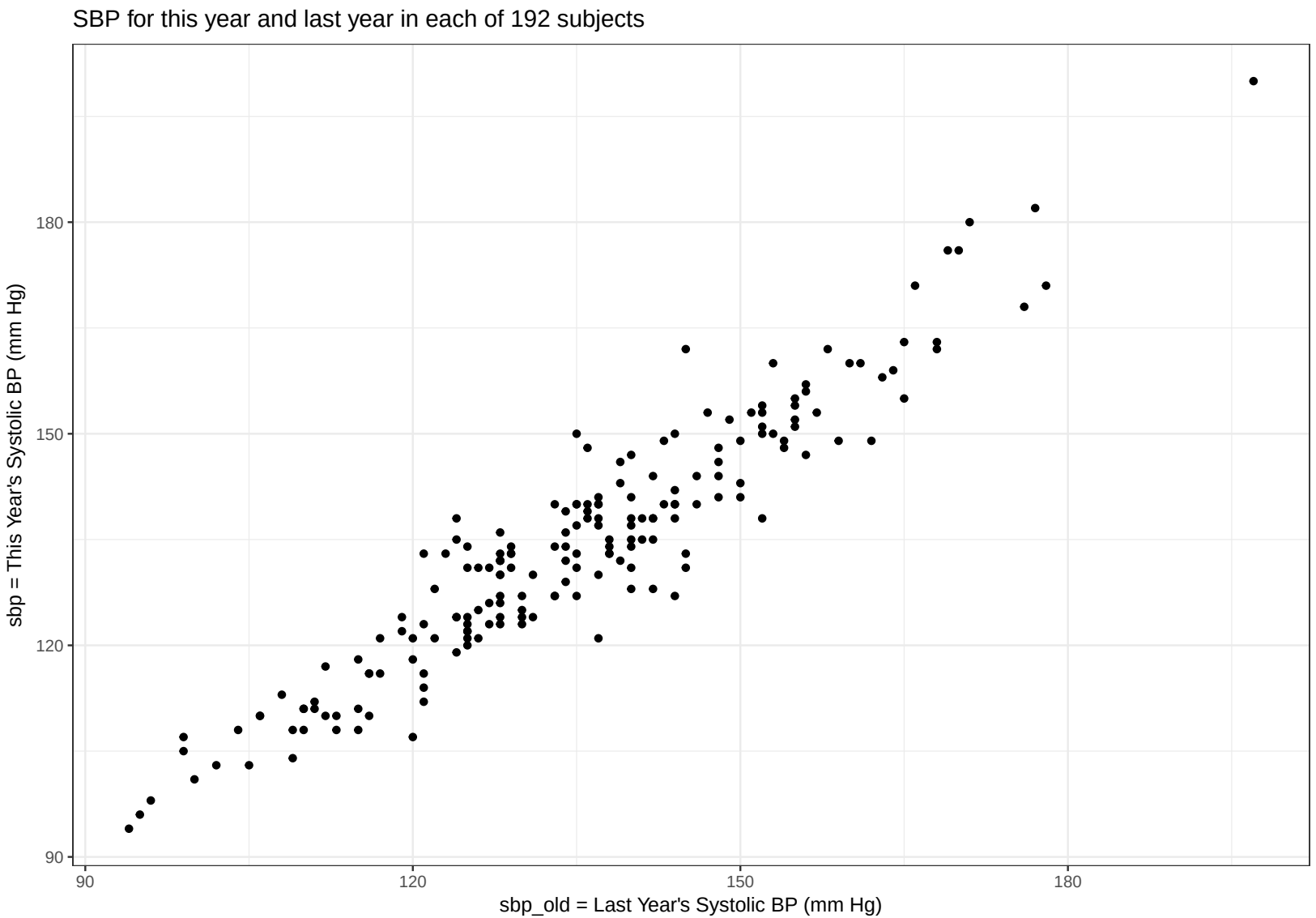
Comparing Population Means via Paired Samples

The dm192 data has current systolic blood pressure (sbp), and systolic blood pressure from last year (sbp_old). Suppose we want to describe the mean SBP change in not just our sample, but instead the entire **population** (adults who live in NE Ohio with diabetes) over the past year.

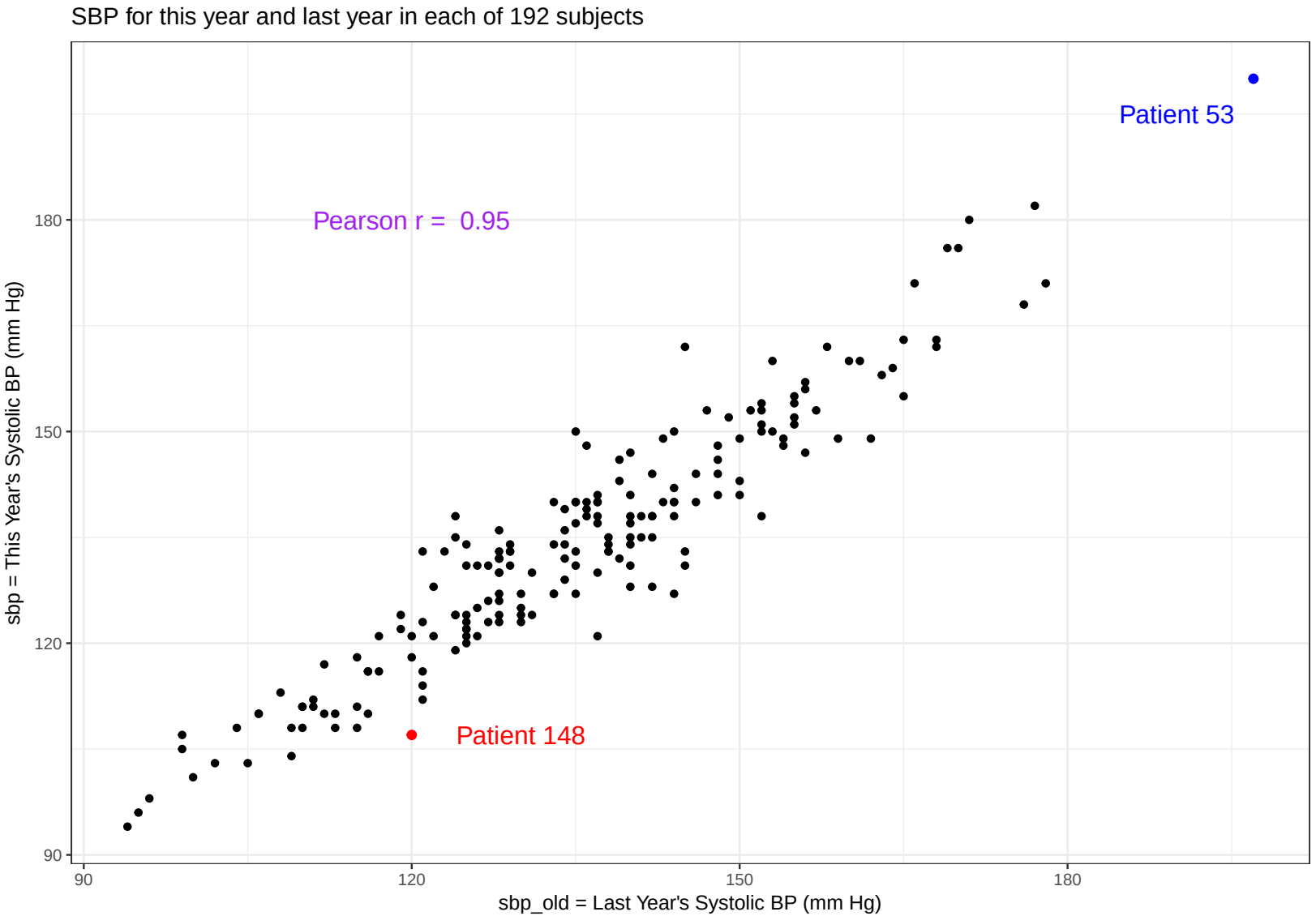
```
dm_first <- select(dm192, pt.id, sbp, sbp_old)
summary(dm_first)
```

```
     pt.id             sbp           sbp_old
 Min.   :  1.00   Min.   : 94.0   Min.   : 94.0
 1st Qu.: 48.75   1st Qu.:123.0   1st Qu.:124.0
 Median : 96.50   Median :133.0   Median :135.0
 Mean   : 96.50   Mean   :134.2   Mean   :135.0
 3rd Qu.:144.25   3rd Qu.:144.5   3rd Qu.:145.2
 Max.   :192.00   Max.   :200.0   Max.   :197.0
```

Each subject provides both a `sbp_old` and `sbp`

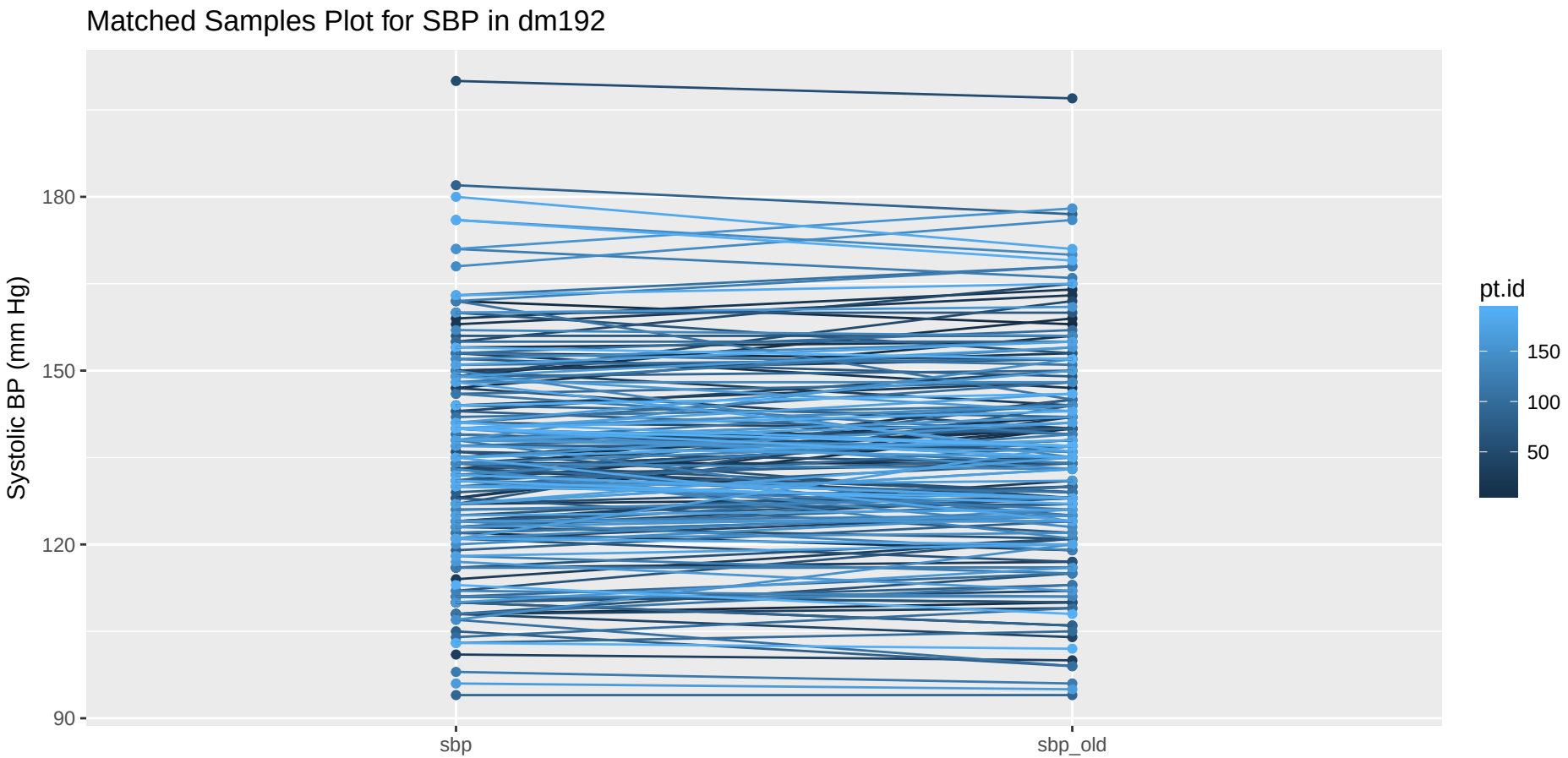


The Impact of Pairing



A Matched Samples Plot ("After - Before" Plot)

Each subject provides both a value for `sbp` and one for `sbp_old`:



Patient 53 is the patient on top, with `sbp` = 200, and `sbp_old` = 197.

Paired Samples? Calculate Paired Differences

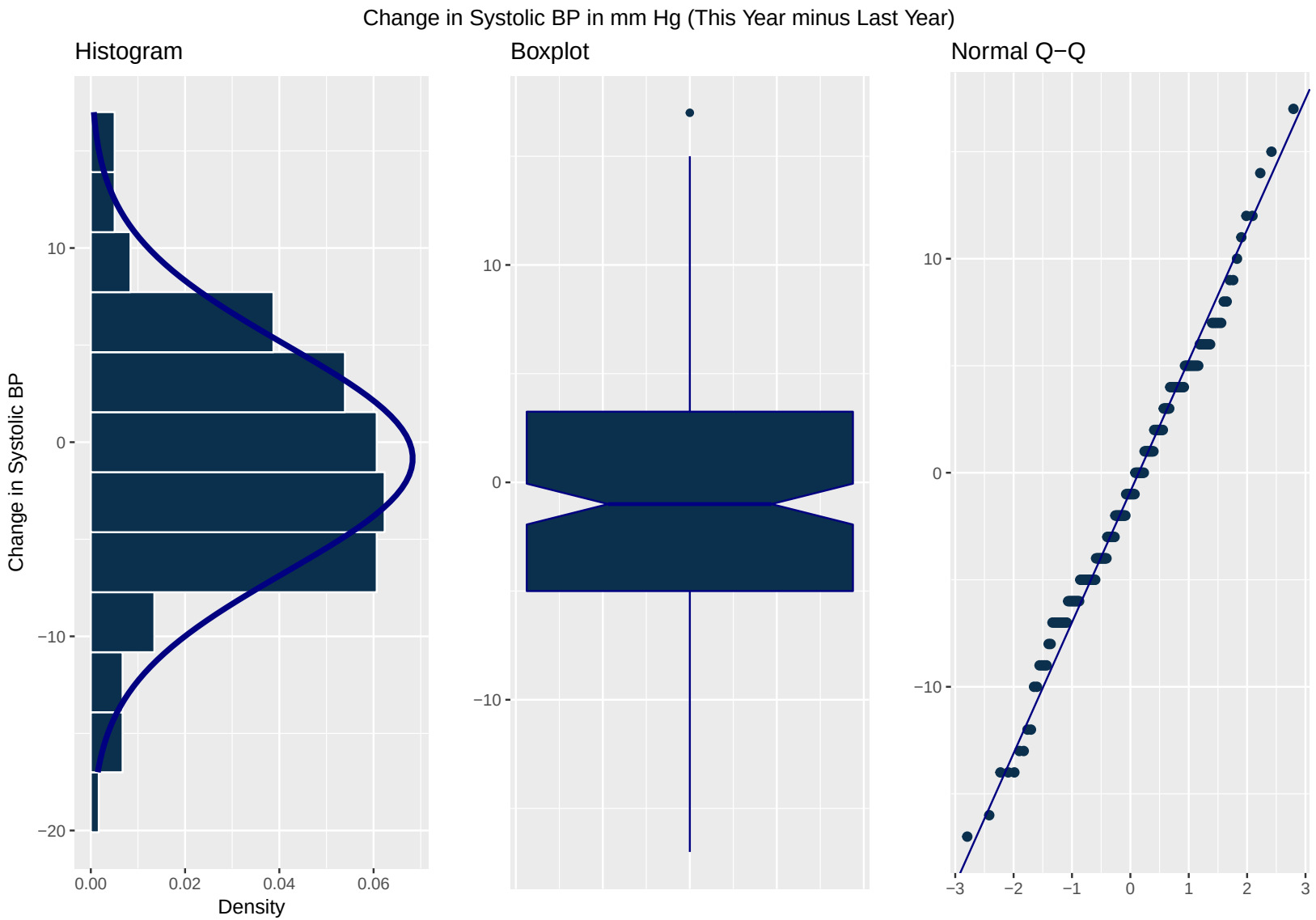
```
dm_first$diffs <- dm_first$sbp - dm_first$sbp_old;
dm_first[1:3,]
```

```
# A tibble: 3 x 4
  pt.id   sbp sbp_old diffs
  <int> <int>   <int> <int>
1     1   108     110    -2
2     2   162     158     4
3     3   135     142    -7
```

```
mosaic::favstats(dm_first$diffs)
```

```
 min Q1 median   Q3 max        mean        sd   n
 -17 -5     -1 3.25  17 -0.8385417 5.840818 192
 missing
       0
```

EDA for the Paired Differences



t test for the Paired Differences

```
t.test(dm_first$sbp, dm_first$sbp_old, paired = TRUE)
```

```
    Paired t-test

data:  dm_first$sbp and dm_first$sbp_old
t = -1.9893, df = 191, p-value = 0.04809
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -1.669983188 -0.007100145
sample estimates:
mean of the differences
             -0.8385417
```

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 SBP 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 SBP This Year = Population Mean SBP 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 SBP in Diabetes Study

Since our null hypothesis is

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

our alternative hypothesis will therefore cover all other possibilities:

H_A: Population Mean SBP This Year – Population Mean SBP 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

1. 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.
2. 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.
3. 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 the Part B 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 SBP in Diabetes Study

```
t.test(dm_first$sbp-dm_first$sbp_old)
```

```
    One Sample t-test

data:  dm_first$sbp - dm_first$sbp_old
t = -1.9893, df = 191, p-value = 0.04809
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.669983188 -0.007100145
sample estimates:
 mean of x
-0.8385417
```

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 SBP in Diabetes data

```
wilcox.test(dm_first$sbp - dm_first$sbp_old, conf.int=TRUE)
```

```
    Wilcoxon signed rank test with continuity
    correction

data:  dm_first$sbp - dm_first$sbp_old
V = 6714, p-value = 0.04065
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 -1.999947e+00 -4.688972e-05
sample estimates:
(pseudo)median
    -0.9999959
```

Should we reject or retain $H_0 : \mu_d = 0$ based on this test?

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 Twins data

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

```
set.seed(4311); Hmisc::smean.cl.boot(dm_first$diffs)
```

```
      Mean        Lower        Upper
-0.83854167 -1.66666667 -0.05195313
```

So, according to this confidence interval, a reasonable range (with 95% confidence) for μ, the population mean of the unadjusted – adjusted differences is (-1.67, -0.052). Should we reject or retain $H_0 : \mu = 0$?

What does this confidence interval suggest about the p value?

Using `slipper` to run a bootstrap CI

For paired differences, we can bootstrap the mean or any other summary. . .

```
# requires library(slipper)
# to install slipper: devtools::install_github('jtleek/slipper
set.seed(4312)
dm_first %>% slipper(mean(diffs), B = 500) %>%
  filter(type == "bootstrap") %>%
  summarize(ci_low = quantile(value, 0.025),
            ci_high = quantile(value, 0.975))
```

```
    ci_low     ci_high
1 -1.67487 -0.03619792
```

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

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

Approach	p value	95% CI for μ_d	Conclusion re: H_0: $\mu_d = 0$
t Test	0.048	(-1.67, -0.007)	$p < 0.05$, so reject H_0
Wilcoxon	0.041	(-2.0, -0.0004)	$p < 0.05$, so reject H_0
Bootstrap	< 0.05	(-1.67, -0.052)	CI for μ excludes 0 so reject H_0

Our Conclusions for the SBP in Diabetes Study

So, in this case, using any of these methods, we draw the same conclusion – to reject H_0 at the 5% significance level and conclude as a result that:

1. there is a statistically significant difference between the population mean SBP of patients this year as compared to last year.
2. the population mean this year – last year difference in SBP, which we have called μ_d, is statistically significantly different from zero.
3. In fact, the confidence intervals universally tell us that this population mean is negative – SBP was (slightly) smaller this year than last year at the 95% confidence level.

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.

Independent Samples Study Designs

- Independent samples designs do not impose such a matching, but instead sample two unrelated sets of subjects, where each group receives one of the two exposures.
- The two groups of subjects are drawn independently from their separate populations of interest.
- One obvious way to tell if we have an independent samples design is that this design does not require the sizes of the two exposure groups to be equal.

The best way to establish whether a study uses paired or independent samples is to look for the **link** between the two measurements that creates paired differences.

- Deciding whether or not the samples are paired (matched) is something we do before we analyze the data.

How Big A Sample Size Do I need?

1. What is the budget?
2. What are you trying to compare?
3. What is the study design?
4. How big an effect size do you expect (hope) to see?
5. What was that budget again?
6. OK, tell me the maximum allowable rates of Type I and Type II error that you want to control for. Or, if you like, tell me the confidence level and power you want to have.
7. And what sort of statistical inference do you want to plan for?

Type I and Type II Error

Once we know how unlikely the results would have been if the null hypothesis were true, we must make one of two choices:

1. The p value is not small enough to convincingly rule out chance. Therefore, we cannot reject the null hypothesis as an explanation for the results.
2. The p value was small enough to convincingly rule out chance. We reject the null hypothesis and accept the alternative hypothesis.

How small must the p value be in order to rule out the null hypothesis?

- The standard choice is 5%.
- This standardization has advantages and disadvantages, and there are many situations for which a 5% cutoff may be unwise.

Disease Screening Analogy

Consider being tested for disease. Most tests for diseases are not 100% accurate. The lab technician or physician must make a choice:

1. In the opinion of the medical practitioner, you are healthy. The test result was weak enough to be called **negative** for the disease.
    - Potential Error: You actually have the disease but have been told you do not. This is a **false negative**. This is what we refer to as a **Type II** error, and the maximum allowable rate of Type II error is called β.
2. In the opinion of the medical practitioner, you have the disease. The test results were strong enough to be called **positive** for the disease.
    - Potential Error: You are actually healthy but have been told you have the disease. This is a **false positive**. This is what we refer to as a **Type I** error, and the maximum allowable rate of Type I error is α.

Relation of α and β to Error Types

- α is the probability of rejecting H_0 when H_0 is true.
  - So $1 - \alpha$, the confidence level, is the probability of retaining H_0 when that's the right thing to do.
- β is the probability of retaining H_0 when H_A is true.
  - So $1 - \beta$, the power, is the probability of rejecting H_0 when that's the right thing to do.

–	H_A is True	H_0 is True
Test Rejects H_0	Correct Decision (1 - β)	Type I Error (α)
Test Retains H_0	Type II Error (β)	Correct Decision (1 - α)

Sample Size & Power of a Paired Sample t Test

For a paired-samples t test, R can estimate any one of the following elements, given the other four, using the `power.t.test` command, for either a one-sided or two-sided paired t test...

- n = the sample size (# of pairs) being compared
- δ = `delta` = the true difference in means between the two groups
- s = `sd` = the true standard deviation of the paired differences
- α = `sig.level` = the significance level for the comparison (maximum acceptable risk of Type I error)
- $1 - \beta$ = power = the power of the paired t test to detect the effect of size δ

A Small Example: Studying Systolic BP

- I want to do a study to compare population mean SBP before and after I give people this intervention. I'll measure everyone twice - once before and once after the intervention.

So, tell me, how big a sample do I need?

A Small Example: Studying Systolic BP

- I want to do a study to compare population mean SBP before and after I give people this intervention. I'll measure everyone twice - once before and once after the intervention.
- I have some old data, and in those people, I saw that the standard deviation of changes in SBP readings was about 20 mm Hg.

So, tell me, how big a sample do I need?

A Small Example: Studying Systolic BP

- I want to do a study to compare population mean SBP before and after I give people this intervention. I'll measure everyone twice - once before and once after the intervention.
- I have some old data, and in those people, I saw that the standard deviation of changes in SBP readings was about 20 mm Hg.
- I want to do a t test, because I expect the changes to be Normally distributed.

So, tell me, how big a sample do I need?

A Small Example: Studying Systolic BP

- I want to do a study to compare population mean SBP before and after I give people this intervention. I'll measure everyone twice - once before and once after the intervention.
- I have some old data, and in those people, I saw that the standard deviation of changes in SBP readings was about 20 mm Hg.
- I want to do a t test, because I expect the changes to be Normally distributed.
- I want to make sure I can detect a change in mean SBP associated with my intervention of as little as 10 mm Hg.

So, tell me, how big a sample do I need?

A Small Example: Studying Systolic BP

- I want to do a study to compare population mean SBP before and after I give people this intervention. I'll measure everyone twice - once before and once after the intervention.
- I have some old data, and in those people, I saw that the standard deviation of changes in SBP readings was about 20 mm Hg.
- I want to do a t test, because I expect the changes to be Normally distributed.
- I want to make sure I can detect a change in mean SBP associated with my intervention of as little as 10 mm Hg.
- I don't know what power I should use, or what confidence level I should use, but usually in my field we use 80% power and 95% confidence, so I guess that's what I should do.

So, tell me, how big a sample do I need?

Power of a Paired T test

So, do we know four out of five?

- n = the sample size (# of pairs) being compared
- δ = `delta` = the true difference in means between the two groups
- s = `sd` = the true standard deviation of the paired differences
- α = `sig.level` = the significance level for the comparison (maximum acceptable risk of Type I error)
- $1 - \beta$ = power = the power of the paired t test to detect the effect of size δ

power.t.test for a Paired Sample Study

```
power.t.test(delta = 10, sd = 20, type = "paired",
     alt = "two.sided", sig.level = 0.05, power = 0.80)
```

```
     Paired t test power calculation

              n = 33.3672
          delta = 10
             sd = 20
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

What if we wanted 90% power instead?

```
power.t.test(delta = 10, sd = 20, type = "paired",
     alt = "two.sided", sig.level = 0.05, power = 0.90)
```

```
     Paired t test power calculation

              n = 43.99552
          delta = 10
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

What if the effect was half as large?

```
power.t.test(delta = 5, sd = 20, type = "paired",
     alt = "two.sided", sig.level = 0.05, power = 0.90)
```

```
     Paired t test power calculation

              n = 170.0511
          delta = 5
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences*
```

What if we didn't pair the samples?

```
power.t.test(delta = 5, sd = 20, type = "two.sample",
     alt = "two.sided", sig.level = 0.05, power = 0.90)
```

```
     Two-sample t test power calculation

              n = 337.2008
          delta = 5
             sd = 20
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Sample Size & Power: Pooled t Test

For an independent-samples t test, with a balanced design (so that $n_1 = n_2$), R can estimate any one of the following elements, given the other four, using the `power.t.test` function, for a one-sided or two-sided t test.

- n = the sample size in each of the two groups being compared
- δ = `delta` = the true difference in means between the two groups
- s = sd = the true standard deviation of the individual values in each group (assumed to be constant, since we assume equal population variances)
- α = `sig.level` = the significance level for the comparison (maximum acceptable risk of Type I error)
- $1 - \beta$ = power = the power of the t test to detect the effect of size δ

If you want a two-sample power calculation for an unbalanced design, you will need to use a different library and function in R.

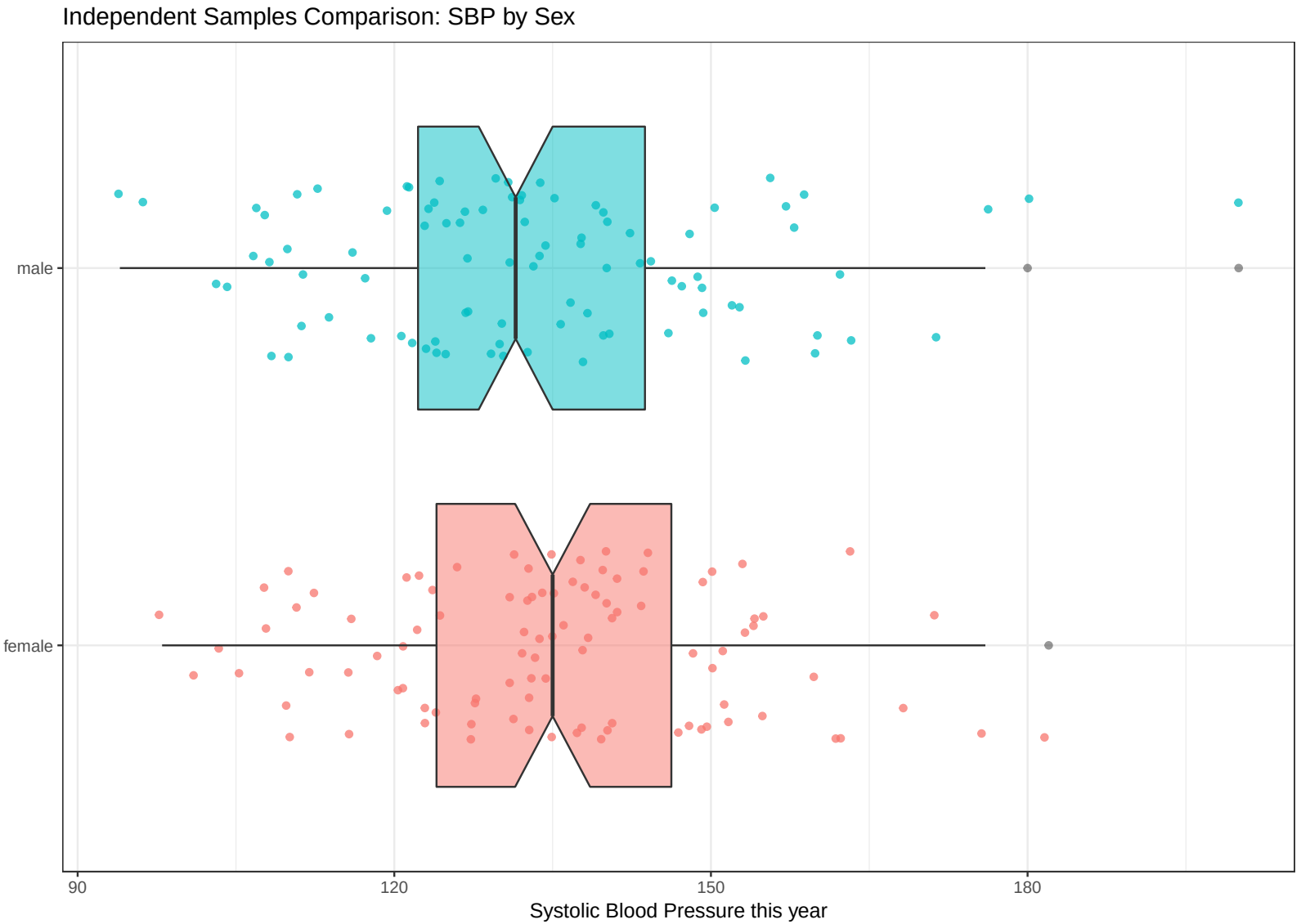
What if the Samples Aren't Paired?

In the dm192 frame, we might also consider looking at a different kind of comparison, perhaps whether the average systolic blood pressure is larger in male or in female adults in NE Ohio living with diabetes.

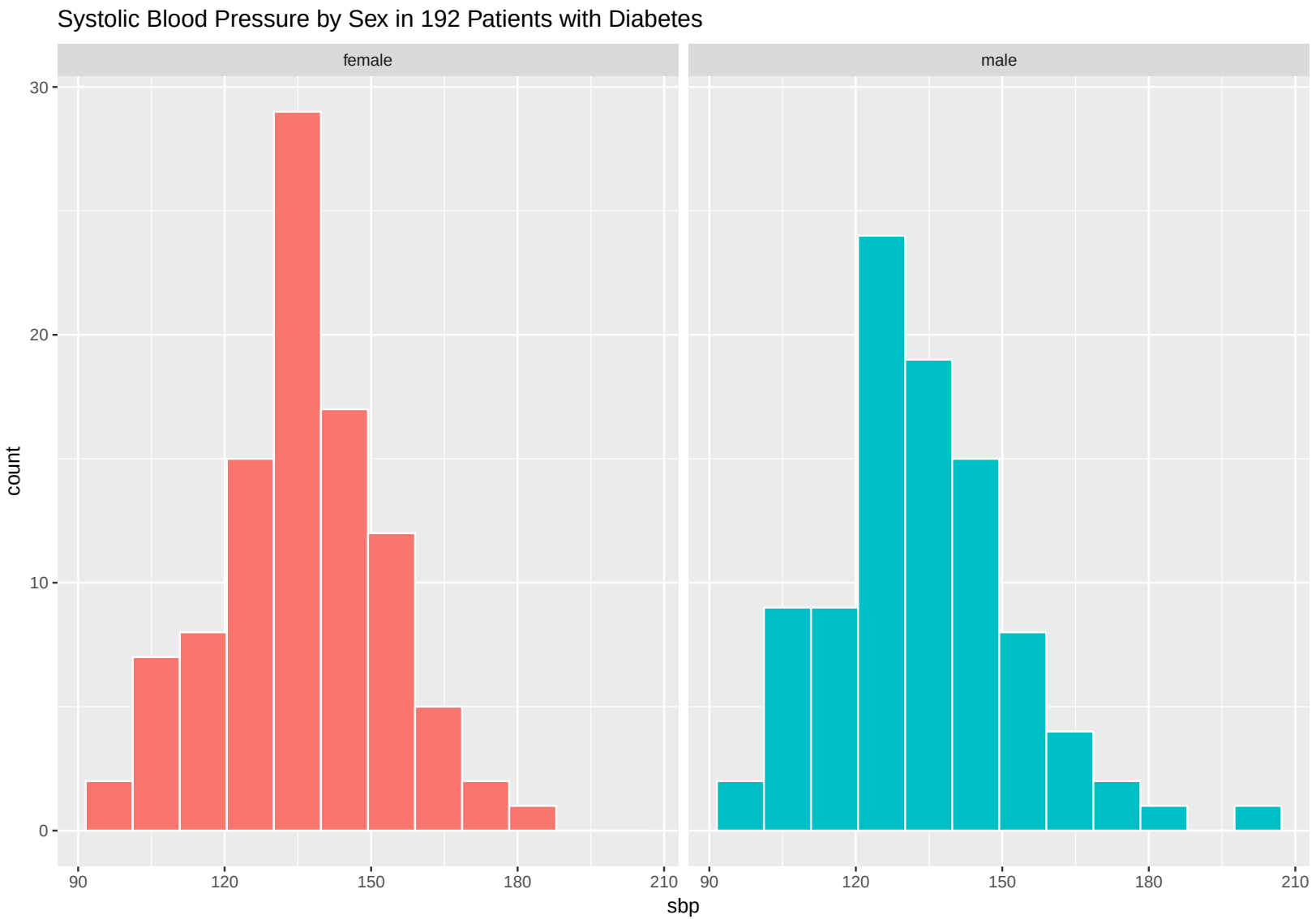
```
dm_second <- select(dm192, pt.id, sex, sbp)
summary(dm_second)
```

```
     pt.id            sex          sbp       
 Min.   :  1.00   female:98   Min.   : 94.0  
 1st Qu.: 48.75   male  :94   1st Qu.:123.0  
 Median : 96.50               Median :133.0  
 Mean   : 96.50               Mean   :134.2  
 3rd Qu.:144.25               3rd Qu.:144.5  
 Max.   :192.00               Max.   :200.0  
```

Our comparison now is between females and males



Another Way to Picture Two Independent Samples



Numerical Summary for Two Independent Samples

```
by(dm_second$sbp, dm_second$sex, mosaic::favstats)
```

```
dm_second$sex: female
 min  Q1 median     Q3 max     mean       sd  n
  98 124    135 146.25 182 135.1327 16.75637 98
 missing
       0
------------------------------------------------
dm_second$sex: male
 min     Q1 median     Q3 max     mean       sd  n
  94 122.25  131.5 143.75 200 133.2447 18.82785 94
 missing
       0
```

Hypotheses Under Consideration

The hypotheses we are testing are:

- H_0: mean in population 1 = mean in population 2 + hypothesized difference Δ_0 vs.
- H_A: mean in population 1 $\neq$ mean in population 2 + hypothesized difference Δ_0,

where Δ_0 is almost always zero. An equivalent way to write this is:

- $H_0 : \mu_1 = \mu_2 + \Delta_0$ vs.
- $H_A : \mu_1 \neq \mu_2 + \Delta_0$

Yet another equally valid way to write this is:

- $H_0 : \mu_1 - \mu_2 = \Delta_0$ vs.
- $H_A : \mu_1 - \mu_2 \neq \Delta_0$,

where, again Δ_0 is almost always zero.

Testing Options for Independent Samples

1. Pooled t test or Indicator Variable Regression Model (t test assuming equal population variances)
2. Welch t test (t test without assuming equal population variances)
3. Wilcoxon-Mann-Whitney Rank Sum Test (non-parametric test not assuming populations are Normal)
4. Bootstrap confidence interval for the difference in population means

Assumptions of the Pooled T test

The standard method for comparing population means based on two independent samples is based on the t distribution, and requires the following assumptions:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.
3. [Normal Population] The two populations are each Normally distributed
4. [Equal Variances] The population variances in the two groups being compared are the same, so we can obtain a pooled estimate of their joint variance.

The Pooled Variances t test in R

Also referred to as the t test assuming equal population variances:

```
t.test(dm_second$sbp ~ dm_second$sex, var.equal=TRUE)
```

```
    Two Sample t-test

data:  dm_second$sbp by dm_second$sex
t = 0.73467, df = 190, p-value = 0.4634
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -3.181093  6.957037
sample estimates:
mean in group female   mean in group male
            135.1327             133.2447
```

Assumptions of the Welch t test

The Welch test still requires:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.
3. [Normal Population] The two populations are each Normally distributed

But it doesn't require:

4. [Equal Variances] The population variances in the two groups being compared are the same.

Welch's t test is the default choice in R.

Welch t test without assuming equal population variances

```
t.test(dm_second$sbp ~ dm_second$sex)
```

```
    Welch Two Sample t-test

data:  dm_second$sbp by dm_second$sex
t = 0.73288, df = 185.39, p-value = 0.4646
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -3.194236  6.970180
sample estimates:
mean in group female   mean in group male
            135.1327             133.2447
```

Assumptions of the Wilcoxon-Mann-Whitney Rank Sum Test

The Wilcoxon-Mann-Whitney Rank Sum test still requires:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.

But it doesn't require:

3. [Normal Population] The two populations are each Normally distributed
4. [Equal Variances] The population variances in the two groups being compared are the same.

It also doesn't really compare population means.

Wilcoxon-Mann-Whitney Rank Sum Test

```
wilcox.test(dm_second$sbp ~ dm_second$sex, conf.int = TRUE)
```

```
    Wilcoxon rank sum test with continuity
    correction

data:  dm_second$sbp by dm_second$sex
W = 5035.5, p-value = 0.2649
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -2.000061  7.999993
sample estimates:
difference in location
              2.999918
```

The Bootstrap

This bootstrap approach to comparing population means using two independent samples still requires:

1. [Independence] The samples for the two groups are drawn independently.
2. [Random Samples] The samples for each of the groups are drawn at random from the populations of interest.

but does not require either of the other two assumptions:

3. [Normal Population] The two populations are each Normally distributed
4. [Equal Variances] The population variances in the two groups being compared are the same.

The bootstrap procedure I use in R was adapted from Frank Harrell and colleagues. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/BootstrapMeansSoftware

The `bootdif` function

The procedure requires the definition of a function, which I have adapted a bit, called `bootdif`, which is part of the `Love-boost.R` script on the web site, and is also part of this Markdown file.

As in our previous bootstrap procedures, we are sampling (with replacement) a series of many data sets (default: 2000).

- Here, we are building bootstrap samples based on the SBP levels in the two independent samples (M vs. F).
- For each bootstrap sample, we are calculating a mean difference between the two groups (M vs. F).
- We then determine the 2.5^{th} and 97.5^{th} percentile of the resulting distribution of mean differences (for a 95% confidence interval).

Using the `bootdif` function to compare means based on independent samples

So, to compare systolic BP (our outcome) across the two levels of sex (our grouping factor) for the adult patients with diabetes in NE Ohio, run the following. . .

```
set.seed(4314); bootdif(dm_second$sbp, dm_second$sex)
```

```
Mean Difference           0.025           0.975
      -1.887972      -6.977860        2.917249
```

Note that the two columns must be separated here with a comma rather than a tilde (~).

This CI describes the male - female difference (i.e. the negative of the F-M difference used earlier) – we can tell this by the listed sample mean difference.

Can we use slipper instead?

For differences in means between independent samples, we can use the tidy function in broom to obtain the point estimate, and then slipper on that result.

```
broom::tidy(t.test(dm_second$sbp ~ dm_second$sex))
```

```
  estimate estimate1 estimate2 statistic   p.value
1 1.887972  135.1327  133.2447 0.7328845 0.4645545
  parameter  conf.low conf.high
1  185.3938 -3.194236   6.97018
                     method alternative
1 Welch Two Sample t-test   two.sided
```

Using `slipper` to run a bootstrap CI

For comparing the means of independent samples:

```
# requires library(slipper)
set.seed(4313)
dm_second %>%
  slipper((broom::tidy(t.test(sbp ~ sex))$estimate),
          B = 500) %>%
  summarise(bootci_low = quantile(value, 0.025),
            bootci_high = quantile(value, 0.975))
```

```
  bootci_low bootci_high
1  -2.687109    6.970497
```

Results for the SBP and Sex Study

Procedure	2-sided p value for $H_0 : \mu_F = \mu_M$	95% CI for $\mu_F - \mu_M$
Pooled t test	0.463	(-3.2, 7.0)
Welch t test	0.465	(-3.2, 7.0)
Rank Sum test	0.265	(-2.0, 8.0)
Bootstrap CI	$p > 0.05$	(-2.9, 7.0) via `bootdif`
Bootstrap CI	$p > 0.05$	(-2.7, 7.0) via `slipper`

What conclusions should we draw, at $\alpha = 0.05$?

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.

On Reporting p Values

When reporting a p value and no rounding rules are in place from the lead author/journal/source for publication, follow these conventions. . .

1. Use an italicized, lower-case p to specify the p value. Don't use p for anything else.
2. For p values above 0.10, round to two decimal places, at most.
3. For p values near α, include only enough decimal places to clarify the reject/retain decision.
4. For very small p values, always report either $p < 0.0001$ or even just $p < 0.001$, rather than specifying the result in scientific notation, or, worse, as $p = 0$ which is glaringly inappropriate.
5. Report p values above 0.99 as $p > 0.99$, rather than $p = 1$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if p $<$ .05,

which morphed into a

- **rule** for editors: reject the submitted article if p $>$ .05.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$,

which morphed into a

- **rule** for editors: reject the submitted article if $p > .05$,

which morphed into a

- **rule** for journals: reject all articles that report p-values[1]

[1] http://www.nature.com/news/psychology-journal-bans-p-values-1.17001 describes the recent banning of null hypothesis significance testing by *Basic and Applied Psychology*.

From George Cobb - on why *p* values deserve to be re-evaluated

The **idea** of a p-value as one possible summary of evidence

morphed into a

- **rule** for authors: reject the null hypothesis if $p < .05$, which morphed into a
- **rule** for editors: reject the submitted article if $p > .05$, which morphed into a
- **rule** for journals: reject all articles that report p-values.

Bottom line: **Reject rules. Ideas matter.**

Posted to an American Statistical Association message board Oct 14 2015