431 Class 14

Thomas E. Love

2017-10-12

Today's R Setup

```
library(boot); library(broom); library(magrittr)
library(tidyverse)

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

Today's Agenda

1. Project Task B Group Meetings
2. Statistical Inference and the `dm192` data
    - Hypothesis Testing and P Values
    - Comparing Two Population Means using Paired Samples
    - Comparing Two Population Means using Indepdendent Samples

Project Task A is due on Friday 2017-10-13 at noon.

Project Task B Group Meetings

Go.

Description of the dm192 data

I stored the dm192.csv data in a subdirectory of my class 14 project directory called data.

```
dm192 <- read.csv("data/dm192.csv") %>% tbl_df
head(dm192,5) # show just the first 5 rows
```

```
# A tibble: 5 x 14
  pt.id practice   sbp   dbp   a1c   ldl   age    sex
  <int>   <fctr> <int> <int> <dbl> <int> <int> <fctr>
1     1        A   108    71   5.8    58    44   male
2     2        A   162    92  11.6    54    28 female
3     3        B   135    84    NA    NA    58 female
4     4        C   133    87  12.7   112    56   male
5     5        D   128    72   6.8   105    54 female
# ... with 6 more variables: race <fctr>, hisp <fctr>,
#   insurance <fctr>, statin <int>, sbp_old <int>,
#   a1c_old <fctr>
```

The Large Sample Formula for the CI around μ

The two-tailed 100(1-α)% confidence interval for a population mean μ (based on the Normal distribution) is:

- The Lower Bound is $\bar{x} - Z_{\alpha/2}(\sigma/\sqrt{n})$ and the Upper Bound is $\bar{x} + Z_{\alpha/2}(\sigma/\sqrt{n})$

where $Z_{\alpha/2}$ is the value that cuts off the top $\alpha/2$ percent of the standard Normal distribution (the Normal distribution with mean 0 and standard deviation 1).

Obtaining the $Z_{\alpha/2}$ value using `qnorm`

We can obtain this cutoff value from R by substituting in the desired proportion for alphaover2 into the `qnorm` function as follows:

```
qnorm(alphaover2, lower.tail=FALSE)
```

For example, if we are building a 95% confidence interval, we have 100(1-α) = 95, so that α is 0.05, or 5%. This means that the cutoff value we need to find is $Z_{0.05/2} = Z_{.025}$, and this turns out to be 1.96.

```
qnorm(0.025, lower.tail=FALSE)
```

```
[1] 1.959964
```

Commonly Used Cutoffs based on the Normal Distribution

- If we're building a two-tailed 95% confidence interval, we'll use $Z_{.025} = 1.96$
- For a two-tailed 90% confidence interval, we use $Z_{.05} = 1.645$
- For a two-tailed 99% confidence interval, we use $Z_{.005} = 2.576$
- For a two-tailed 50% confidence interval, we use $Z_{.25} = 0.67$
- For a two-tailed 68% confidence interval, we use $Z_{.16} = 0.99$

Lots of CIs use the Normal distribution

- The usual 95% confidence interval for large samples is an estimate $\pm$ 2 standard errors[1].
- Also, from the Normal distribution, an estimate $\pm$ 1 standard error is a 68% confidence interval, and an estimate $\pm$ 2/3 of a standard error is a 50% confidence interval.
- A 50% interval is particularly easy to interpret because the true value should be inside the interval about as often as it is not.
- A 95% interval is thus about three times as wide as a 50% interval.
- In general, the larger the confidence required, the wider the interval will need to be.

[1] The use of 2 standard errors for a confidence interval for a population mean is certainly reasonable whenever n is 60 or more. This is because the t distribution with 59 degrees of freedom has a 0.025 cutoff of 2.0, anyway.

Large-Sample CI for Systolic BP Mean, μ

Since we have a fairly large sample ($n = 192$), we could consider using a large-sample approach (assuming the sample standard deviation is equal to the population standard deviation, and then using the Normal distribution) to estimate a confidence interval for the mean systolic blood pressure in the population of all adults with diabetes who live in Northeast Ohio. The 95% confidence interval is calculated as $\bar{x} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$, and here we will assume that $s = \sigma$ which may be reasonable with a fairly large sample size.

- We have $n = 192$ observations, and since we want a 95% confidence interval, $\alpha = 0.05$
- Our sample mean $\bar{x} = 134.21$ and standard deviation $s = 17.78$
- So the standard error is 1.28

The 95% CI is thus $134.21 \pm 1.96(1.28)$, or (131.7, 136.72) using the Normal distribution.

- Our 95% CI based on the t distribution was (131.7, 136.7).

Assumptions of a t-based Confidence Interval

"Begin challenging your assumptions. Your assumptions are your windows on the world. Scrub them off every once in awhile or the light won't come in." (Alan Alda)

1. Sample is drawn at random from the population or process.
2. Samples are drawn independently from each other from a population or process whose distribution is unchanged during the sampling process.
3. Population or process follows a Normal distribution.

Can we drop any of these assumptions?

Only if we're willing to consider alternative inference methods.

What is a Bootstrap and Why Should I Care?

The bootstrap (and in particular, what's known as bootstrap resampling) is a really good idea that you should know a little bit about[2].

If we want to know how accurately a sample mean estimates the population mean, we would ideally like to take a very, very large sample, because if we did so, we could conclude with something that would eventually approach mathematical certainty that the sample mean would be very close to the population mean.

But we can rarely draw enormous samples. So what can we do?

[2] See Good PI Hardin JW Common Errors in Statistics – a very helpful book.

Resampling is A Big Idea

If we want our sample mean to accurately estimate the population mean, we would ideally like to take a very, very large sample, so as to get very precise estimates. But we can rarely draw enormous samples. So what can we do?

Oversimplifying, the idea is that if we sample (with replacement) from our current data, we can draw a new sample of the same size as our original.

- And if we repeat this many times, we can generate as many samples of, say, 192 systolic blood pressures, as we like.
- Then we take these thousands of samples and calculate (for instance) the sample mean for each, and plot a histogram of those means.
- If we then cut off the top and bottom 5% of these sample means, we obtain a reasonable 90% confidence interval for the population mean.

Bootstrap: Estimating a confidence interval for μ

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 (the default is 1,000 using our approach) to obtain a set of 1,000 sample means.
4. Sort those 1,000 sample means in order, and estimate the 90% confidence interval for the population mean based on the middle 90% of the 1,000 bootstrap samples.
5. Send us a result, containing the sample mean, and a 90% confidence interval for the population mean

When is a Bootstrap Confidence Interval for μ 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
- and that the samples are identically distributed (even though that distribution may not be Normal.)

A downside is that you and I will get (somewhat) different answers if we resample from the same data.

90% CI for population mean μ using bootstrap

The command that we use to obtain a CI for μ using the basic nonparametric bootstrap and without assuming a Normally distributed population, is `smean.cl.boot`, a part of the `Hmisc` package in R.

```
set.seed(43101)
Hmisc::smean.cl.boot(dm192$sbp, conf = 0.90)
```

```
    Mean     Lower     Upper
134.2083 131.9633 136.1424
```

Comparing Bootstrap and T-Based Confidence Intervals

- The `smean.cl.boot` function (unlike most R functions) deletes missing data automatically, as does the `smean.cl.normal` function, which produces the t-based confidence interval.

```
Hmisc::smean.cl.boot(dm192$sbp, conf = 0.90)
```

```
    Mean     Lower     Upper
134.2083 132.2234 136.4904
```

```
Hmisc::smean.cl.normal(dm192$sbp, conf = 0.90)
```

```
    Mean     Lower     Upper
134.2083 132.0876 136.3291
```

Rerunning 90% CI for μ via Bootstrap

```
set.seed(43102); Hmisc::smean.cl.boot(dm192$sbp, conf = 0.9)
```

```
    Mean    Lower    Upper
134.2083 132.1195 136.3187
```

```
set.seed(43103); Hmisc::smean.cl.boot(dm192$sbp, conf = 0.9)
```

```
    Mean    Lower    Upper
134.2083 132.0880 136.3180
```

```
set.seed(43104)
  Hmisc::smean.cl.boot(dm192$sbp, conf = 0.9, B = 2000)
```

```
    Mean    Lower    Upper
134.2083 132.1404 136.4534
```

Bootstrap: Changing the Confidence Level

```
set.seed(43105); Hmisc::smean.cl.boot(dm192$sbp, conf = 0.90)
```

```
    Mean    Lower    Upper
134.2083 132.0823 136.3029
```

```
set.seed(43106); Hmisc::smean.cl.boot(dm192$sbp, conf = 0.95)
```

```
    Mean    Lower    Upper
134.2083 131.7492 136.8180
```

```
set.seed(43107); Hmisc::smean.cl.boot(dm192$sbp, conf = 0.99)
```

```
    Mean    Lower    Upper
134.2083 130.7445 137.4845
```

Bootstrap for a One-Sided Confidence Interval

If you want to estimate a one-sided confidence interval for the population mean using the bootstrap, then the procedure is as follows:

1. Determine α, the significance level you want to use in your one-sided confidence interval. Remember that α is 1 minus the confidence level. Let's assume we want a 90% one-sided interval, so $\alpha = 0.10$.
2. Double α to determine the significance level we will use in the next step to fit a two-sided confidence interval.
3. Fit a two-sided confidence interval with confidence level $100(1 - 2\alpha)$. Let the bounds of this interval be (a, b).
4. The one-sided (greater than) confidence interval will have a as its lower bound.
5. The one-sided (less than) confidence interval will have b as its upper bound.

One-sided CI for μ via the Bootstrap

Suppose that we want to find a 90% one-sided upper bound for the population mean systolic blood pressure among Northeast Ohio adults with diabetes, μ, using the bootstrap.

Since we want a 90% confidence interval, we have $\alpha = 0.10$. We double that to get $\alpha = 0.20$, which implies we need to instead fit a two-sided 80% confidence interval.

```
set.seed(43108); Hmisc::smean.cl.boot(dm192$sbp, conf = 0.80)
```

```
    Mean     Lower     Upper
134.2083 132.7234 135.7714
```

Since the upper bound of this two-sided 80% CI is 135.77, that will also be the upper bound for a 90% one-sided CI.

Additional Notes on the Bootstrap

Bootstrap resampling confidence intervals do not follow the general confidence interval strategy using a point estimate $\pm$ a margin for error.

- A bootstrap interval is often asymmetric, and while it will generally have the point estimate (the sample mean) near its center, for highly skewed data, this will not necessarily be the case.
- I usually use either 1,000 (the default) or 10,000 bootstrap replications for building confidence intervals - practically, it makes little difference.

The bootstrap may seem like the solution to all problems in theory, we could use the same approach to find a confidence interval for any other statistic – it's not perfect, but it is very useful.

- It does eliminate the need to worry about the Normality assumption in small sample size settings, but it still requires independent and identically distributed samples.

Bootstrap Resampling: Advantages and Caveats

Bootstrap procedures exist for virtually any statistical comparison - the t-test analog above is just one many possibilities, and bootstrap methods are rapidly gaining on more traditional approaches in the literature thanks mostly to faster computers.

The bootstrap produces clean and robust inferences (such as confidence intervals) in many tricky situations.

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).

Bootstrap CI for the Population Median, Step 1

If we are willing to do a small amount of programming work in R, we can obtain bootstrap confidence intervals for other population parameters besides the mean. One statistic of common interest is the median. How do we find a confidence interval for the population median using a bootstrap approach? Use the `boot` package, as follows.

In step 1, we specify a new function to capture the medians from our sample.

```
f.median <- function(y, id)
{    median ( y[id])  }
```

Bootstrap CI for the Population Median, Step 2

In step 2, we summon the boot package and call the boot.ci function:

```
set.seed(431787)
boot.ci(boot (dm192$sbp, f.median, 1000),
        conf=0.90, type="basic")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot(dm192$sbp, f.median, 1000), conf = 0.9
    type = "basic")

Intervals :
Level      Basic
90%   (130.5, 134.0 )
Calculations and Intervals on Original Scale
```

Bootstrap CI for the Population Median vs. Mean

- Note that the sample **median** of the SBP data is 133 mm Hg.
- Our 90% confidence interval for the population **median** SBP among NE Ohio adults with diabetes is (130.5, 134) according to the bootstrap, using the random seed 431787.
- The sample **mean** of the SBP data is 134.2 mm Hg.
- The 90% bootstrap CI for the population **mean** SBP, μ, is (132.1, 136.5) if we use the random seed 43121.

The Wilcoxon Signed Rank Procedure for CIs

It turns out to be difficult to estimate an appropriate confidence interval for the median of a population, which might be an appealing thing to do, particularly if the sample data are clearly not Normally distributed, so that a median seems like a better summary of the center of the data. Bootstrap procedures are available to perform the task.

The Wilcoxon signed rank approach can be used as an alternative to t-based procedures to build interval estimates for the population *pseudo-median* when the population cannot be assumed to follow a Normal distribution.

As it turns out, if you're willing to assume the population is **symmetric** (but not necessarily Normally distributed) then the pseudo-median is actually equal to the population median.

What is a Pseudo-Median?

The pseudo-median of a particular distribution G is the median of the distribution of (u + v)/2, where both u and v have the same distribution (G).

- If the distribution G is symmetric, then the pseudomedian is equal to the median.
- If the distribution is skewed, then the pseudomedian is not the same as the median.
- For any sample, the pseudomedian is defined as the median of all of the midpoints of pairs of observations in the sample.

Getting the Wilcoxon Signed Rank-based CI in R

```
wilcox.test(dm192$sbp, conf.int=TRUE, conf.level=0.95)
```

```
    Wilcoxon signed rank test with continuity
    correction

data:  dm192$sbp
V = 18528, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 131.4999 136.0000
sample estimates:
(pseudo)median
         133.5
```

Interpreting the Wilcoxon CI for the Population Median

If we're willing to believe the `sbp` values come from a population with a symmetric distribution, the 95% Confidence Interval for the population median would be (131.5, 136)

For a non-symmetric population, this only applies to the *pseudo-median*.

Note that the pseudo-median (133.5) is actually fairly close in this situation to the sample mean (134.2) as well as to the sample median (133), as it usually will be if the population actually follows a symmetric distribution, as the Wilcoxon approach assumes.

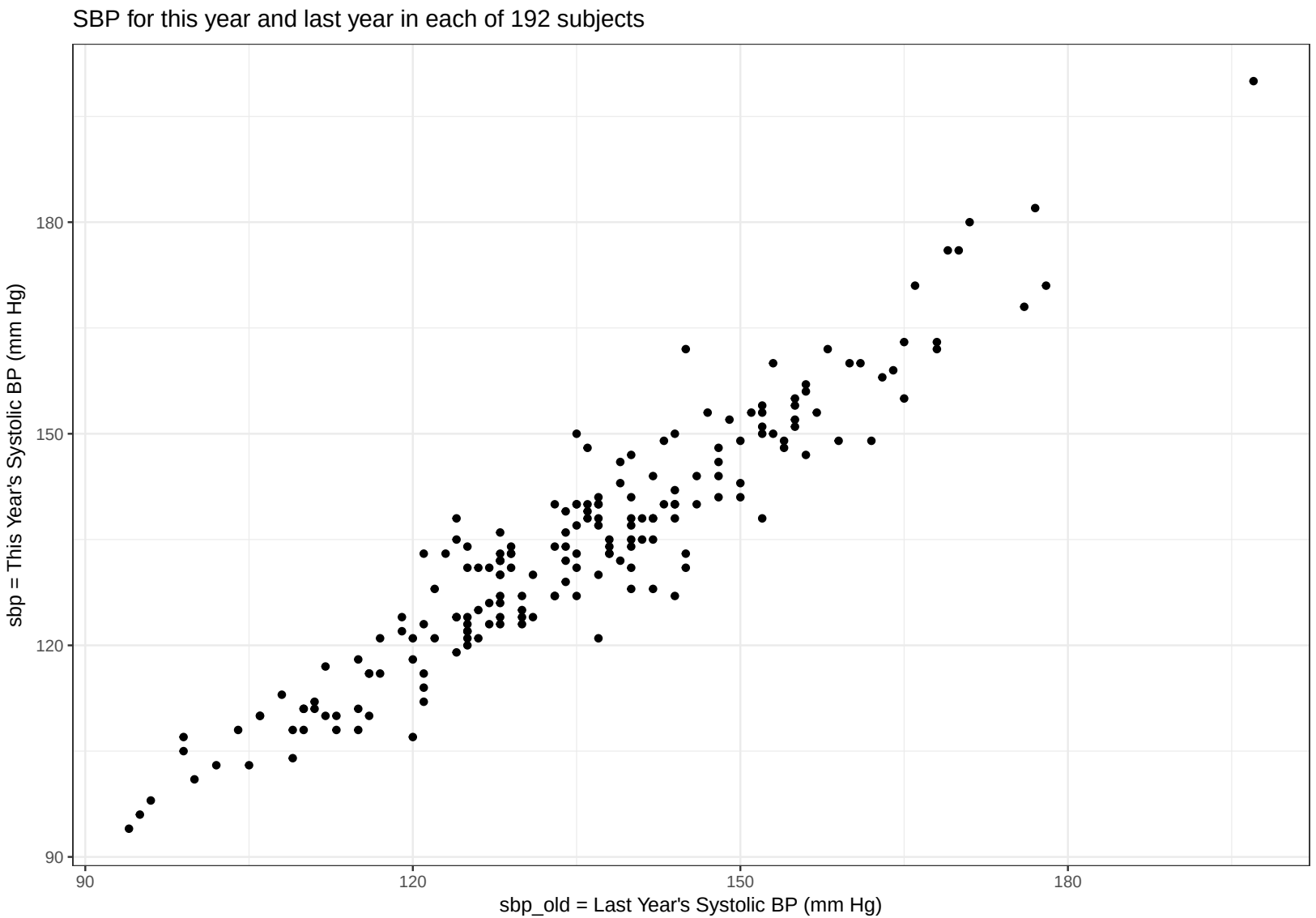
Comparing Population Means via Paired Samples

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.

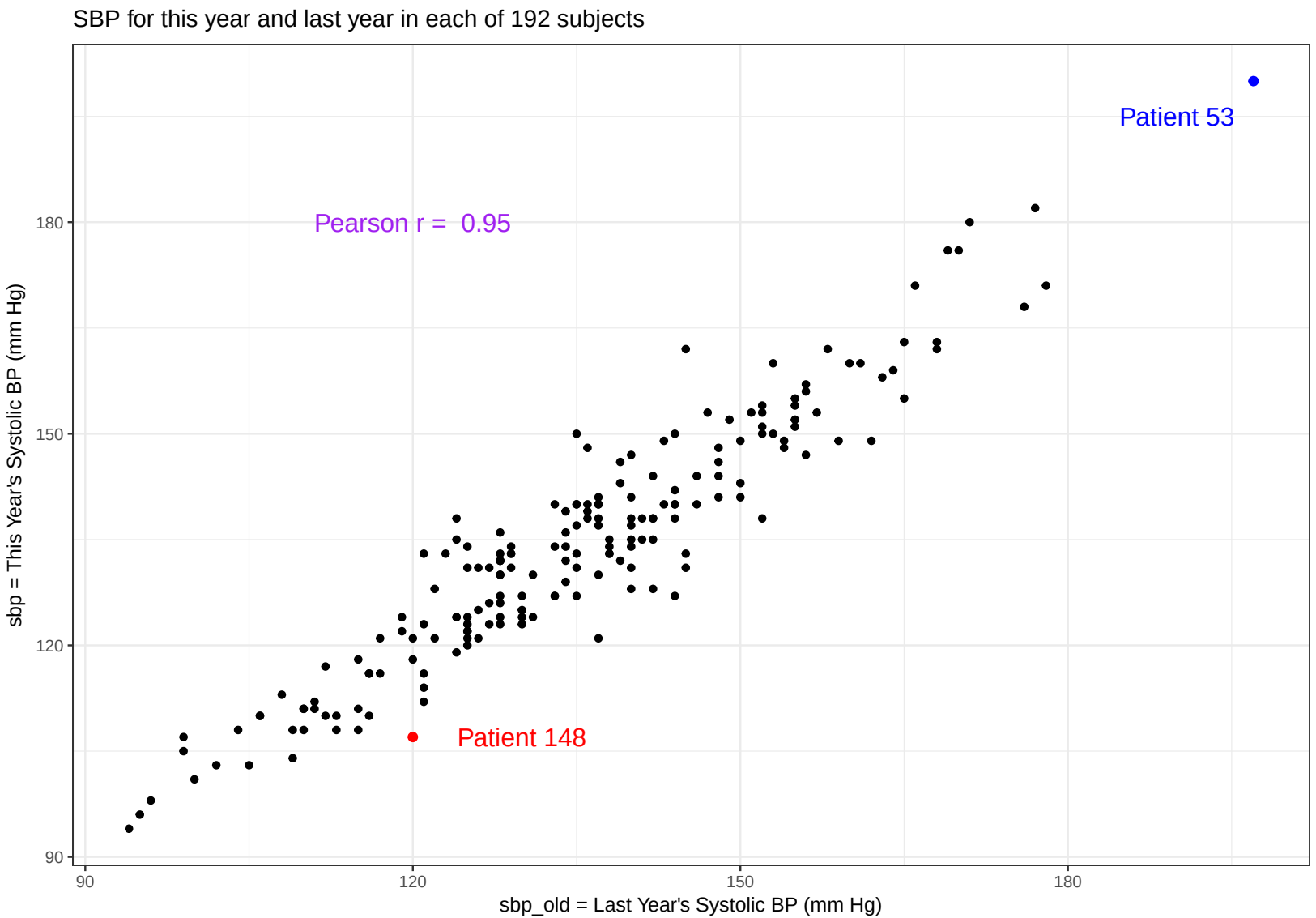
```
dm_first <- select(dm192, pt.id, sbp, sbp_old)
summary(dm_first)
```

```
     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`

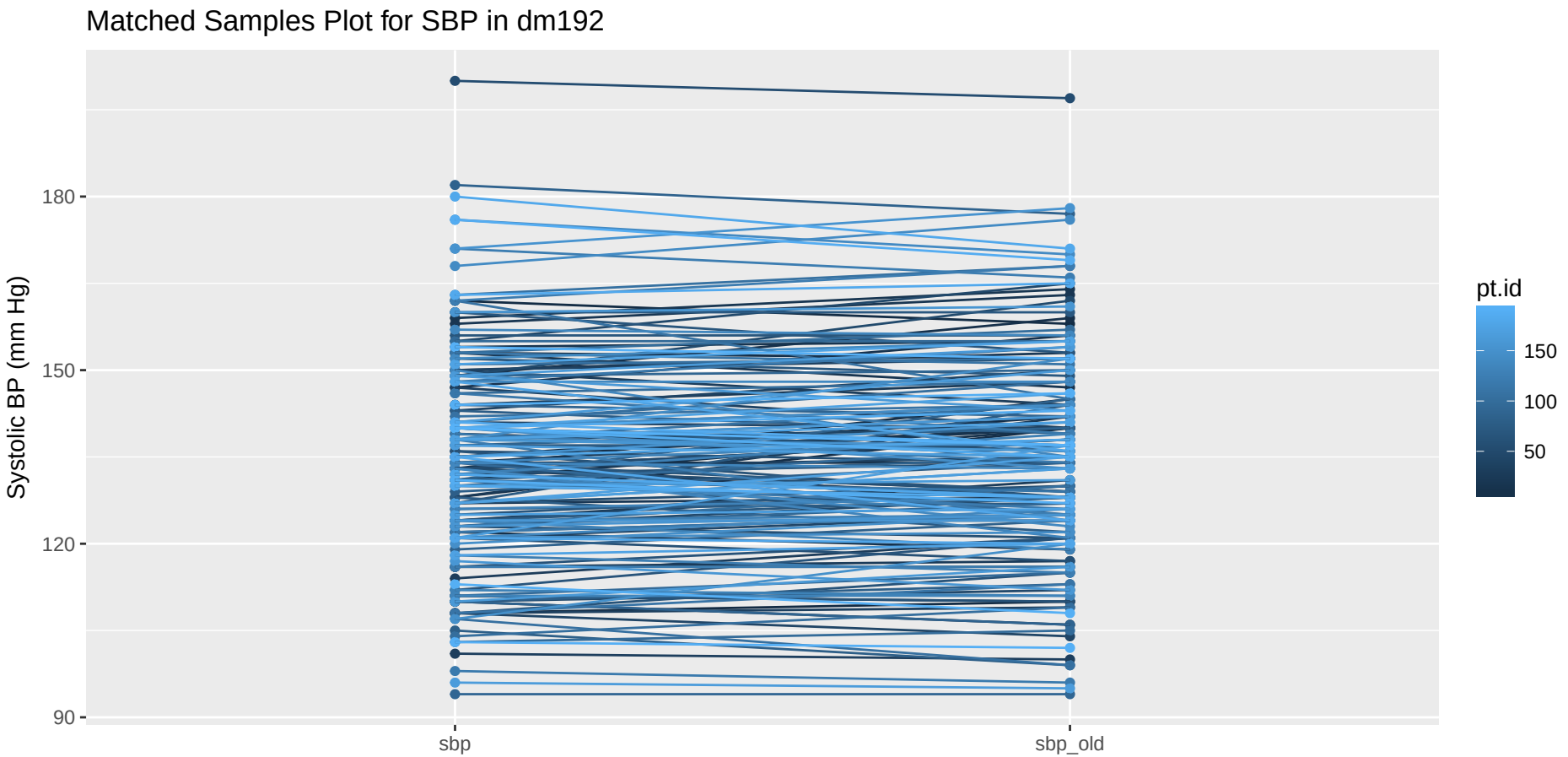


The Impact of Pairing



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

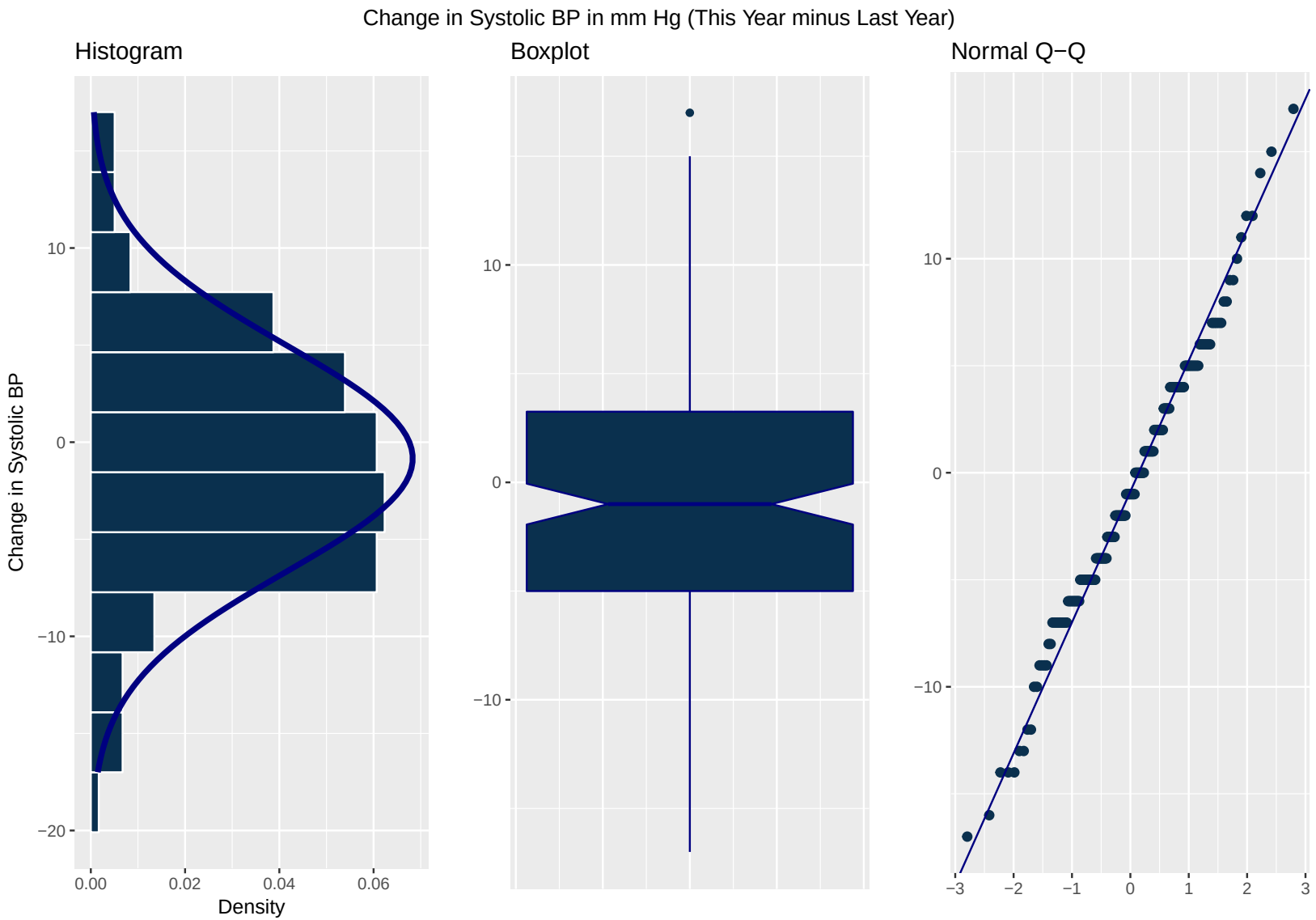
```
dm_first$diffs <- dm_first$sbp - dm_first$sbp_old;
dm_first[1:3,]
```

```
# 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?

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.

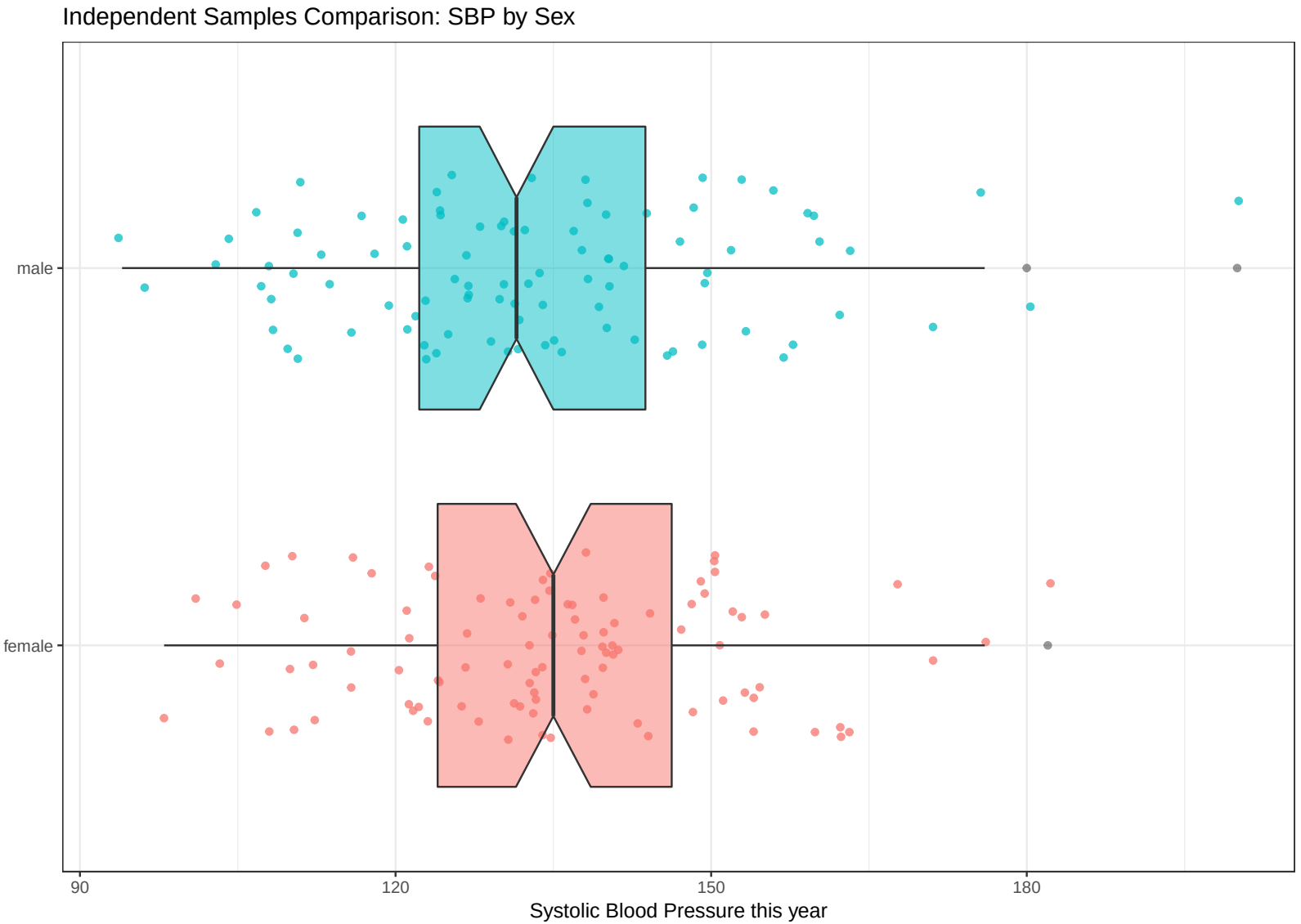
What if the Samples Aren't Paired?

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.

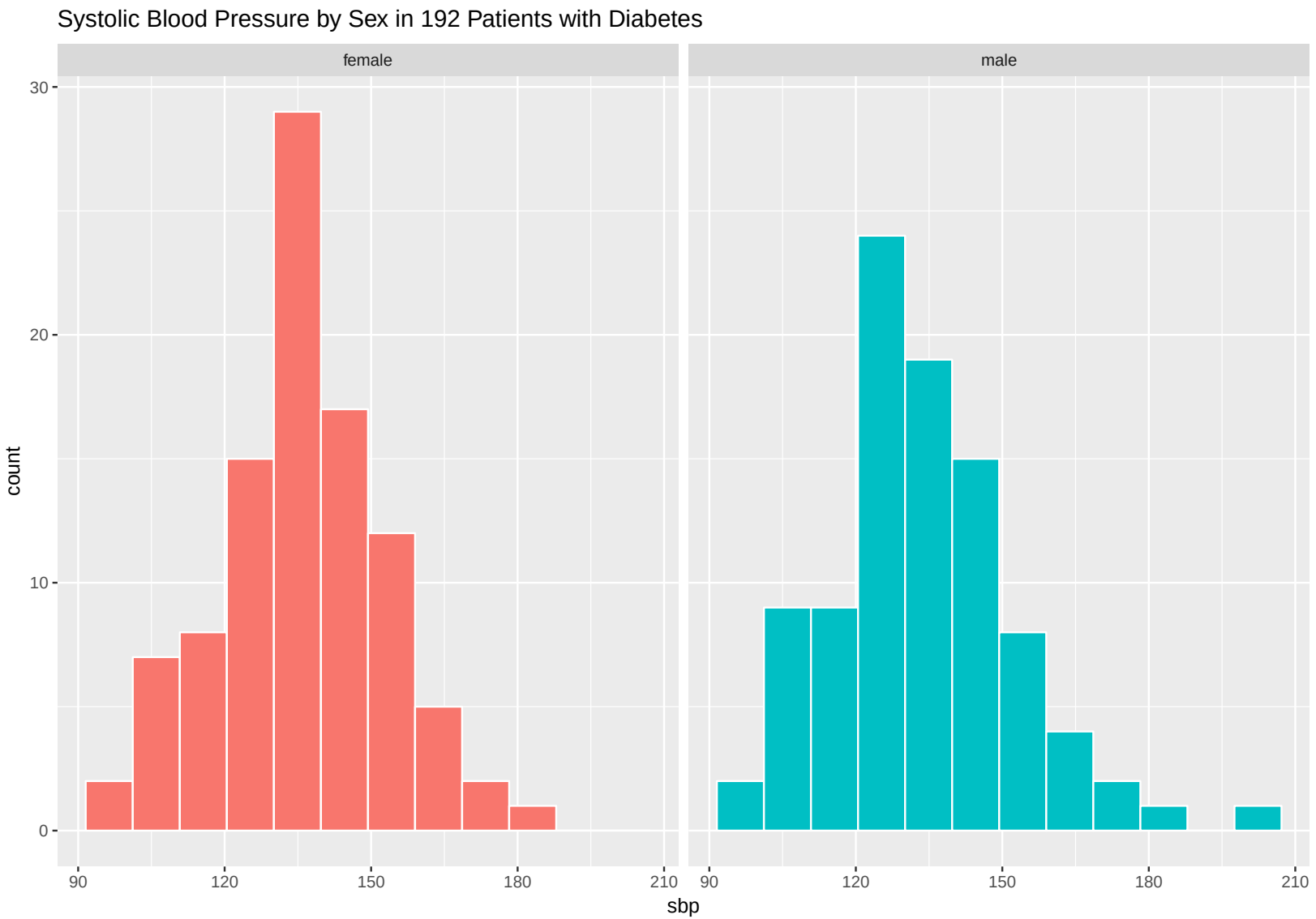
```
dm_second <- select(dm192, pt.id, sex, sbp)
summary(dm_second)
```

```
     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 2.5th and 97.5th 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.

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)

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[3]

[3] 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