431 Class 14

Thomas E. Love

2018-10-11

Today's R Setup

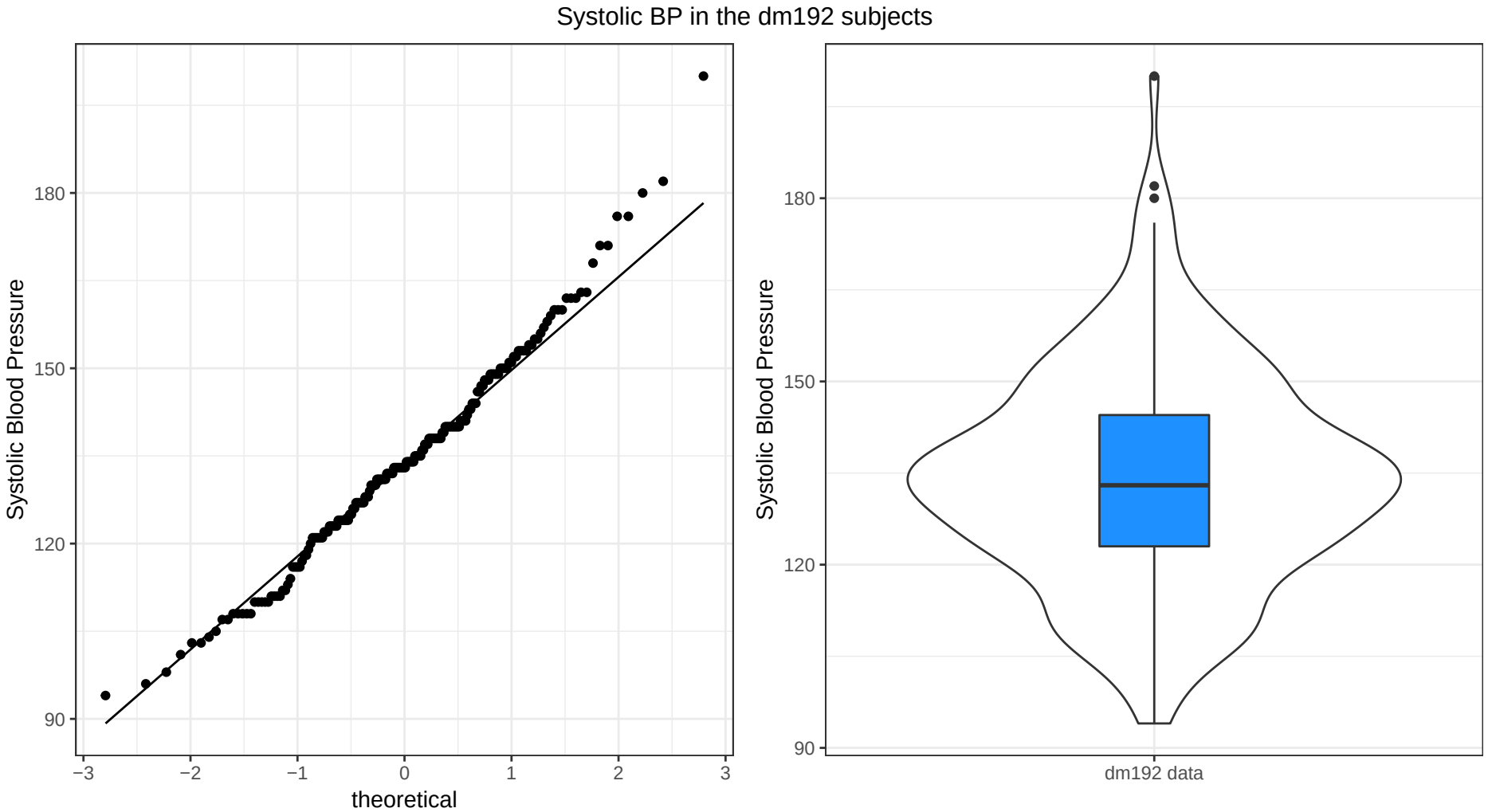
```
library(boot); library(Hmisc); library(broom)
library(tidyverse) # always load tidyverse last

source("Love-boost.R") # script from our Data page
```

```
dm192 <- read.csv("data/dm192.csv") %>% tbl_df
mosaic::favstats(~ sbp, data = dm192)
```

```
 min  Q1 median    Q3 max     mean       sd   n
  94 123    133 144.5 200 134.2083 17.77899 192
 missing
       0
```

sbp in dm192 is "Normalish" but not clearly Normal



Building Confidence Intervals for the Population Mean

Confidence Intervals for Population Mean μ

There are four options that we'll see today:

1. Use the t distribution
   - Assumes the population has a Normal distribution
   - Estimates unknown σ with sample SD
2. Use a standard Normal (Z) distribution
   - Assumes the population has a Normal distribution
   - Assumes σ is known, or large sample (n $\geq$ 60)
3. Use Wilcoxon signed rank procedure
   - Assumes the population has a symmetric distribution
   - Pseudo-median must be of interest and similar to μ
4. Use bootstrap procedure
   - No distributional assumption, μ is of interest
   - Can also be used for other parameters besides μ.

Getting R to build a CI for μ with `t.test`

Happily, R does all of the work.

```
t.test(dm192$sbp, conf.level = 0.90,
       alternative = "two.sided")
```

```
    One Sample t-test

data:  dm192$sbp
t = 104.6, df = 191, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 132.0876 136.3291
sample estimates:
mean of x
 134.2083
```

Summarizing/Tidying the Confidence Interval

```
tt <- t.test(dm192$sbp, conf.level = 0.90,
       alternative = "two.sided")
broom::tidy(tt) # from broom package
```

estimate	statistic	p.value	parameter	conf.low	conf.high
134.2083	104.5979	1.43e-170	191	132.0876	136.3291

method	alternative
One Sample t-test	two.sided

Our 90% confidence interval for the true population mean SBP in NE Ohio adults with diabetes, based on our sample of 192 patients, is (132.1, 136.3) mm Hg[1].

[1] Since the actual SBP values are integers, we should at most include one decimal place in our confidence interval.

What if we want a two-sided 95% CI instead?

The `t.test` function in R has an argument to specify the desired confidence level.

```
t.test(dm192$sbp, conf.level = 0.95, alt = "two.sided")
```

```
    One Sample t-test

data:  dm192$sbp
t = 104.6, df = 191, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 131.6775 136.7392
sample estimates:
mean of x
 134.2083
```

CIs using different Confidence Levels

Below, we see two-sided confidence intervals for various levels of α.

Confidence Level	α	Two-Sided Interval Estimate for SBP Population Mean, μ	Point Estimate for SBP Population Mean, μ
80% or 0.80	0.20	(132.6, 135.9)	134.2
90% or 0.90	0.10	(132.1, 136.3)	134.2
95% or 0.95	0.05	(131.7, 136.7)	134.2
99% or 0.99	0.01	(130.9, 137.5)	134.2

What is the relationship between the confidence level and the width of the confidence interval in the table?

One-sided vs. Two-sided Confidence Intervals

In some situations, we are concerned with either an upper limit for the population mean μ or a lower limit for μ, but not both.

If we, as before, have a sample of size n, with sample mean $\bar{x}$ and sample standard deviation s, then:

- The upper bound for a one-sided 100(1-α)% confidence interval for the population mean is $\mu \leq \bar{x} + t_{\alpha,n-1}(\frac{s}{\sqrt{n}})$, with lower "bound" $-\infty$.
- The corresponding lower bound for a one-sided 100(1 - α) CI for μ would be $\mu \geq \bar{x} - t_{\alpha,n-1}(\frac{s}{\sqrt{n}})$, with upper "bound" ∞.

One-Sided CI for μ (Lower Bound only)

```
t.test(dm192$sbp, conf.level = 0.90, alt = "greater")
```

```
    One Sample t-test

data:  dm192$sbp
t = 104.6, df = 191, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
90 percent confidence interval:
 132.5583      Inf
sample estimates:
mean of x
 134.2083
```

Another One-Sided CI for μ (Upper bound only)

```
t.test(dm192$sbp, conf.level = 0.90, alt = "less")
```

```
    One Sample t-test

data:  dm192$sbp
t = 104.6, df = 191, p-value = 1
alternative hypothesis: true mean is less than 0
90 percent confidence interval:
     -Inf 135.8584
sample estimates:
mean of x
 134.2083
```

Relationship between One-Sided and Two-Sided CIs

Note the relationship between the *two-sided* 80% confidence interval, and the *one-sided* 90% confidence interval.

Confidence Level	α	Type of Interval	Interval Estimate for Population Mean SBP, μ
80% or 0.80	0.20	Two-Sided	(132.56, 135.86)
90% or 0.90	0.10	One Sided ($>$)	$\mu > 132.56$

Why does this happen?

Why, indeed?

- The 90% two-sided interval is placed so as to cut off the top 5% of the distribution with its upper bound, and the bottom 5% of the distribution with its lower bound.
- The 95% "less than" one-sided interval is placed so as to have its lower bound cut off the top 5% of the distribution.

Confidence Level	α	Type of Interval	Interval Estimate for Population Mean SBP, μ
90% or 0.90	0.10	Two-Sided	(132.09, 136.33)
95% or 0.95	0.05	One Sided ($>$)	$\mu > 132.09$

Interpreting the Result

(132.1, 136.3) mm Hg is a 90% two-sided confidence interval for the population mean SBP among NE Ohio adults with diabetes.

- Our point estimate for the true population mean SBP among NE Ohio adults with diabetes is 134.2 mm Hg. The values in the interval (132.1, 136.3) represent a reasonable range of estimates for the true population mean SBP among NE Ohio adults with diabetes, and we are 90% confident that this method of creating a confidence interval will produce a result containing the true population mean SBP among NE Ohio adults with diabetes.
- Were we to draw 100 samples of size 192 from the population described by this sample, and use each such sample to produce a confidence interval in this manner, approximately 90 of those confidence intervals would cover the true population mean SBP among NE Ohio adults with diabetes.

Changing α and One-Sided vs. Two-Sided CIs for μ

Table of t-based estimates follows. . .

Confidence Level	α	2-Sided Interval Estimate for μ, Population Mean SBP	1-Sided Lower Bound for μ
80%	0.20	(132.6, 135.9)	$\mu > 133.1$
90%	0.10	(132.1, 136.3)	$\mu > 132.6$
95%	0.05	(131.7, 136.7)	$\mu > 132.1$
99%	0.01	(130.9, 137.5)	$\mu > 131.2$

- Point Estimate is 134.2 for each of these interval estimates.

Changing α and One-Sided vs. Two-Sided CIs for μ

Table of t-based estimates follows. . .

Confidence Level	α	2-Sided Interval Estimate for μ, Population Mean SBP	1-Sided Lower Bound for μ
80%	0.20	(132.6, 135.9)	$\mu > 133.1$
90%	0.10	(132.1, 136.3)	$\mu > 132.6$
95%	0.05	(131.7, 136.7)	$\mu > 132.1$
99%	0.01	(130.9, 137.5)	$\mu > 131.2$

- Point Estimate is 134.2 for each of these interval estimates.
- Leek: Confirm that estimates have reasonable signs and magnitudes. Do they?

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.

Large Sample Approaches (in Brief)

When you have a large sample size, say, more than 60 observations, the difference between a confidence interval for a population mean based on the t distribution and a confidence interval based on the Normal distribution are usually trivial.

If we were in the position of knowing the standard deviation of the population of interest precisely, we could use that information to build a 100(1-α)% confidence interval using the Normal distribution, based on the sample mean $\bar{x}$, the sample size n, and the (known) population standard deviation σ.

The Large Sample Formula for the CI around μ

If we have a very large sample, we might:

1. Assume the sample standard deviation is an excellent estimate of the population standard deviation σ, and
2. Use the Standard Normal (mean = 0, sd = 1) distribution to build our two-tailed 100(1-α)% confidence interval for a population mean μ:
   - The Lower Bound is $\bar{x} - Z_{\alpha/2}(\sigma/\sqrt{n})$
   - 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 Normal distribution with mean 0 and standard deviation 1.

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

Specify the desired alphaover2 proportion in the `qnorm` function:

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

Building a 95% CI for the population mean SBP

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

For a 95% confidence interval, we have 100(1-α) = 95, so that α is 0.05, or 5%. The cutoff value we need is $Z_{0.05/2} = Z_{.025}$, and this is 1.96.

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

```
[1] 1.959964
```

Common Cutoffs from the Normal Distribution

The usual 95% confidence interval for large samples is an estimate $\pm$ 2 standard errors[2], which is of course an approximation to what the Normal distribution suggests (1.96).

Type of Interval	Confidence	Cutoff Value
Two-tailed	95% CI	$Z_{.025} = 1.96$
Two-tailed	90% CI	$Z_{.05} = 1.645$
Two-tailed	99% CI	$Z_{.005} = 2.576$
Two-tailed	50% CI	$Z_{.25} = 0.67$
Two-tailed	68% CI	$Z_{.16} = 0.99$

[2]The use of 2 standard errors for a confidence interval for a population mean is reasonable if the sample data are approximately Normal and n $\geq$ 60, since the t distribution with 59 degrees of freedom has a 0.025 cutoff of 2.0, anyway.

Lots of CIs use the Normal distribution

- A point estimate $\pm$ 1 standard error is a 68% confidence interval.
- A point 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 (point estimate $\pm$ 2 standard errors) 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.

Large-Sample CI for Systolic BP Mean, μ

The 95% CI using the Normal distribution is $\bar{x} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$

- $n = 192 \geq 60$, so we can consider a large-sample approach.
- We will assume $s = \sigma$.
- 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 of the mean 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 (131.68, 136.74).

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 the bootstrap 90% confidence interval estimate for the population mean.

See Good PI Hardin JW *Common Errors in Statistics* for some theory.

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 (selecting one subject doesn't change the probability that another subject will also be selected)
- and that the samples are identically distributed (even though that distribution may not be Normal.)

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

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(20181011)
Hmisc::smean.cl.boot(dm192$sbp, conf = 0.90)
```

```
    Mean     Lower     Upper
134.2083 132.2078 136.2708
```

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 132.0206 136.2036
```

```
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.

Next Time

Comparing Two Population Means

- Using Paired (Matched) Samples
- Using Independent Samples

Hypothesis Testing and p Values