

# 431 Class 13

Thomas E. Love

2017-10-10

# Today's R Setup

```
library(boot); library(broom); library(magrittr)  
library(tidyverse)  
  
source("Love-boost.R")
```

# Today's Agenda

- 1 Discussing Quiz 1
- 2 The Signal and the Noise, Chapters 4-5
- 3 Leek, Chapter 6
- 4 Statistical Inference and the dm192 data

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

# Quiz 1

The Quiz went reasonably well. One person scored 100/100.

n	Mean	SD	Q1	Median	Q3	Max
51	81.8	10.3	74	82.5	89.5	100

Range	"Grade"	n
89.5 - 100	A	14
84 - 89	A-/B+	11
73 - 83	B	19
below 73	-	7

If you have questions, pose them via email to Dr. Love soon.

# The Signal and the Noise

- Weather forecasters have rapid feedback loops that let them to repeatedly test their models.
- An exciting notion: Predicting results under certain initial conditions for many different examples of initial conditions and then averaging over the results.
- Can we reduce uncertainty in situations that seem hopelessly complicated to analyze, by averaging over the predictions made under different assumptions?
- Computers + Humans, at least in some endeavors, do better than either alone.

# The Signal and the Noise

- Weather forecasters have rapid feedback loops that let them to repeatedly test their models.
- An exciting notion: Predicting results under certain initial conditions for many different examples of initial conditions and then averaging over the results.
- Can we reduce uncertainty in situations that seem hopelessly complicated to analyze, by averaging over the predictions made under different assumptions?
- Computers + Humans, at least in some endeavors, do better than either alone.
- <http://projects.fivethirtyeight.com/2016-swing-the-election/>

- Statistical inference starts with a best estimate of what is happening in the population.
- Define the population, sample, individuals, and data
- Why might your sample not represent the population?
- Importance of Exploration before we attempt Inference
- Key role of assumptions
- Confirm that your estimates have reasonable signs and magnitudes.

# Point Estimation and Confidence Intervals

The basic theory of estimation can be used to indicate the probable accuracy and potential for bias in estimating based on limited samples.

A **point estimate** provides a single best guess as to the value of a population or process parameter.

A **confidence interval** can convey how much error one must allow for in a given estimate.

The key tradeoffs are

- cost vs. precision, and
- precision vs. confidence in the correctness of the statement.

Often, if we are dissatisfied with the width of the confidence interval and want to make it smaller, we have to reconsider the sample – larger samples produce shorter intervals.



# Something Happened! Is this Signal or Noise?

Very often, sample data indicate that something has happened. . .

- the proportion of people who respond to this treatment has changed
- the mean value of this measure appears to have changed

Before we get too excited, it's worth checking whether the apparent result might possibly be the result of random sampling error.

Statistics provides a number of tools for reaching an informed choice (informed by sample information, of course) including confidence intervals and hypothesis tests ( $p$  values), in particular.

# Key Example

Here, I will look at systolic blood pressure values from a sample of 192 adult patients living in Northeast Ohio between the ages of 24 and 74, who have a diagnosis of diabetes, as gathered in the `dm192.csv` data file.

- These data are simulated to mirror some details from real data gathered by the *Better Health Partnership*.
- The `dm192` data has a lot to it, but today, we're just looking at 192 systolic blood pressure values, gathered in the `sbp` variable.

## In the Course Notes

I don't use the `dm192` data in the Part B notes. Instead, I begin with a detailed look at a sample of serum zinc levels in 462 teenage males, as contained in the `serzinc` data frame.

# Description of the dm192 data

I stored the dm192.csv data in a subdirectory of my class 13 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>
```

# A Confidence Interval for the Population Mean

Today, we're focused on our sample of 192 systolic blood pressure values captured in the current time period. The sample summary statistics are:

```
dm192 %$% mosaic::favstats(sbp)
```

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

Our first inferential goal will be to produce a **confidence interval for the true (population) mean** of all adults with diabetes living in NE Ohio based on this sample. We'll assume that

- these 192 adults are a random sample from the population of interest (all adults with diabetes living in NE Ohio), and
- that each sbp value is drawn independently from an identical distribution describing that population.

# Procedures for Building a Confidence Interval

To do this, we will have several different procedures available, including:

- 1 A confidence interval for the population mean based on a **t distribution**, if we assume that the data are drawn from an approximately Normal distribution, using the sample standard deviation. (A wise choice when the data are well described by the Normal.)
- 2 A resampling approach to generate a **bootstrap** confidence interval for the population mean, which does not require that we assume either that the population standard deviation is known, nor that the data are drawn from an approximately Normal distribution, but which has some other weaknesses. (A better choice especially when the data aren't well fit by a Normal model.)
- 3 The **Wilcoxon signed rank** test can also be used to yield a confidence interval statement about the population pseudo-median, a measure of the population distribution's center (but not the population's mean).

# Exploratory Data Analysis for the SBP values

I'll begin by briefly summarizing the `dm192` systolic blood pressure data, using some functions I've built for you. These results include some of the more useful plots and numerical summaries when assessing shape, center and spread.

The `sbp` data in the `dm192` data frame appear to be very well described by a Normal model, as it turns out, with one fairly substantial outlier on the high end of the scale, in particular.

# How Did I Build These Graphical Summaries

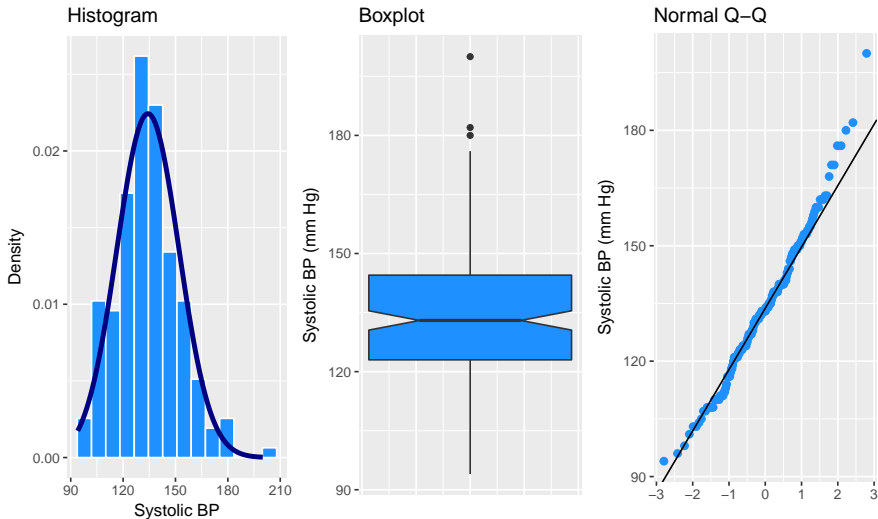
The code presented in the Markdown file builds:

- a histogram (with Normal model superimposed),
- a boxplot (with median notch) and
- a Normal Q-Q plot (with guiding straight line through the quartiles)

for the `sbp` results from the `dm192` tibble. It does this while making use of several functions built specifically for this course, and contained in the script `Love-R-functions.R` which is also included in the Markdown document.

# Graphical Summary of the dm192 systolic BP data

Systolic BP (mm Hg) for 192 NE Ohio Adults with Diabetes





# Key Functions Used in this Work

These functions include:

- `fd_bins` to estimate the Freedman-Diaconis bins setting for the histogram
- `qq_int` and `qq_slope` to facilitate the drawing of a line on the Normal Q-Q plot

You could potentially add `coord_flip()` + to the histogram, and this would have the advantage of getting all three plots oriented in the same direction, but then we (or at least I) lose the ability to tell the direction of skew at a glance from the direction of the histogram.

# How I Built this Graphical Summary, part 1

```
p1 <- ggplot(dm192, aes(x = sbp)) +  
  geom_histogram(aes(y = ..density..),  
                 bins = fd_bins(dm192$sbp),  
                 fill = "dodgerblue", col = "white") +  
  stat_function(fun = dnorm,  
               args = list(mean = mean(dm192$sbp), sd = sd(dm192$sbp)),  
               lwd = 1.5, col = "navy") +  
  labs(title = "Histogram",  
       x = "Systolic BP", y = "Density")  
  
p2 <- ggplot(dm192, aes(x = 1, y = sbp)) +  
  geom_boxplot(fill = "dodgerblue", notch = TRUE) +  
  theme(axis.text.x = element_blank(),  
        axis.ticks.x = element_blank()) +  
  labs(title = "Boxplot",  
       y = "Systolic BP (mm Hg)", x = "")
```

## How I Built this Graphical Summary, part 2

```
p3 <- ggplot(dm192, aes(sample = sbp)) +  
  geom_qq(col = "dodgerblue", size = 2) +  
  geom_abline(intercept = qq_int(dm192$sbp),  
              slope = qq_slope(dm192$sbp)) +  
  labs(title = "Normal Q-Q",  
        y = "Systolic BP (mm Hg)", x = "")  
  
gridExtra::grid.arrange(p1, p2, p3, nrow=1,  
  top = "Systolic BP (mm Hg) for 192 NE Ohio  
        Adults with Diabetes")
```

# Additional Numerical Summaries

Here are some numerical summaries to augment the plots in summarizing the center, spread and shape of the distribution of SBP across these 192 adults.

```
psych::describe(dm192$sbp)
```

	vars	n	mean	sd	median	trimmed	mad	min	max
X1	1	192	134.21	17.78	133	133.64	16.31	94	200
	range	skew	kurtosis	se					
X1	106	0.4	0.49	1.28					

The standard deviation of the SBP data turns out to be 17.78, with  $n = 192$  observations, so the standard error of the mean is

$$se(SBP) = \frac{17.78}{\sqrt{192}} = 1.28$$

This standard error is about to become quite important to us in building statistical inferences about the mean of the entire population of NE Ohio

# Key Questions for Making Inferences from One Sample

- ① What is the population about whom we aim to make an inference?
- ② What is the sample available to us to make that inference?
  - Who are the individuals fueling our inference?
  - What data are available to make an inference?
- ③ Why might this sample not represent the population?

# Defining a Confidence Interval

A confidence interval for a population or process mean uses data from a sample (and perhaps some additional information) to identify a range of potential values for the population mean, which, if certain assumptions hold, can be assumed to provide a reasonable estimate for the true population mean. A confidence interval consists of:

- 1 An interval estimate describing the population parameter of interest (here the population mean), and
- 2 A probability statement, expressed in terms of a confidence level.

# An Example

Suppose that we are willing to assume that the systolic blood pressures across the entire population of NE Ohio adults with diabetes,  $\mu$ , follows a Normal distribution (and so, summarizing it with a mean is a rational thing to do.)

Suppose that we are also willing to assume that the 192 adults contained in the `dm192` tibble are a random sample from that complete population. While we know the mean of the sample of 192 adults, we don't know  $\mu$ , the mean across all NE Ohio adults with diabetes. So we need to estimate it.

## A 90% Confidence Interval for $\mu$

Later, we will find that, with these assumptions in place, we can find a 90% confidence interval for the mean systolic blood pressure across the entire population of NE Ohio adults with diabetes.

- This 90% confidence interval for  $\mu$  turns out to be (132.1, 136.3) mm Hg. How would you interpret this result?



# A 90% Confidence Interval for $\mu$

Later, we will find that, with these assumptions in place, we can find a 90% confidence interval for the mean systolic blood pressure across the entire population of NE Ohio adults with diabetes.

- This 90% confidence interval for  $\mu$  turns out to be (132.1, 136.3) mm Hg. How would you interpret this result?
- Some people think this means that there is a 90% chance that the true mean of the population,  $\mu$ , falls between 132.1 and 136.3 mm Hg.

# A 90% Confidence Interval for $\mu$

Later, we will find that, with these assumptions in place, we can find a 90% confidence interval for the mean systolic blood pressure across the entire population of NE Ohio adults with diabetes.

- This 90% confidence interval for  $\mu$  turns out to be (132.1, 136.3) mm Hg. How would you interpret this result?
- Some people think this means that there is a 90% chance that the true mean of the population,  $\mu$ , falls between 132.1 and 136.3 mm Hg.
- That's not correct. Why not?

# So what do we have confidence in?

- The population mean is a constant **parameter** of the population of interest. That constant is not a random variable, and does not change. So the actual probability of the population mean falling inside that range is either 0 or 1.

Our confidence is in our process.

- It's in the sampling method (random sampling) used to generate the data, and in the assumption that the population follows a Normal distribution.
- It's captured in our accounting for one particular type of error (called *sampling error*) in developing our interval estimate, while assuming all other potential sources of error are negligible.

So what is a more appropriate interpretation?

## 90% CI for $\mu$ is (132.1, 136.3) mm Hg.

What's closer to the truth is:

- If we used this same method to sample data from the true population of adults with diabetes in NE Ohio and built 100 such 90% confidence intervals, then 90 of them would contain the true population mean.
- We call  $100(1-\alpha)\%$ , here, 90%, or 0.90, the *confidence level*, and
- $\alpha = 10\%$ , or 0.10 is called the *significance level*.

If we had instead built a series of 100 different 95% confidence intervals, then about 95 of them would contain the true value of  $\mu$ .

# Estimating a Population Mean

Let's look more closely at the issue of estimating a population mean based on a sample of observations.

We will need three critical pieces - the sample, the confidence level, and the margin of error, which is based on the standard error of a sample mean, when we are estimating a population mean.

In developing a confidence interval for a population mean, we may be willing to assume that the data in our sample are drawn from a Normally distributed population. If so, the most common and useful means of building a confidence interval makes use of the t distribution (sometimes called Student's t) and the notion of a *standard error*.

# The Standard Error of a Sample Mean

The standard error, generally, is the name we give to the standard deviation associated with any particular parameter estimate.

- If we are using a sample mean based on a sample of size  $n$  to estimate a population mean, the **standard error of that sample mean** is  $\sigma/\sqrt{n}$ , where  $\sigma$  is the standard deviation of the measurements in the population.
- We often estimate this particular standard error with its sample analogue,  $s/\sqrt{n}$ , where  $s$  is the sample standard deviation.
- Other statistics have different standard errors.
  - $\sqrt{p(1-p)/n}$  is the standard error of the sample proportion  $p$  estimated using a sample of size  $n$ .
  - $\sqrt{\frac{1-r^2}{n-2}}$  is the standard error of the sample Pearson correlation  $r$  estimated using  $n$  pairs of observations.

# Confidence Intervals for $\mu$ , via the t distribution

In practical settings, we will use the t distribution to estimate a confidence interval from a population mean whenever we:

- are willing to assume that the sample is drawn at random from a population or process with a Normal distribution,
- are using our sample to estimate both the mean and standard deviation, and
- have a small sample size.

# The Formula

We can build a  $100(1-\alpha)\%$  confidence interval using the  $t$  distribution, using the sample mean  $\bar{x}$ , the sample size  $n$ , and the sample standard deviation  $s$ . The two-sided  $100(1-\alpha)\%$  confidence interval (based on a  $t$  test) is:

$$\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$$

where  $t_{\alpha/2, n-1}$  is the value that cuts off the top  $\alpha/2$  percent of the  $t$  distribution, with  $n - 1$  degrees of freedom.

We obtain the relevant cutoff value in R by substituting in values for `alphaover2` and `n-1` into the following line of R code:

```
qt(alphaover2, df = n-1, lower.tail=FALSE)
```



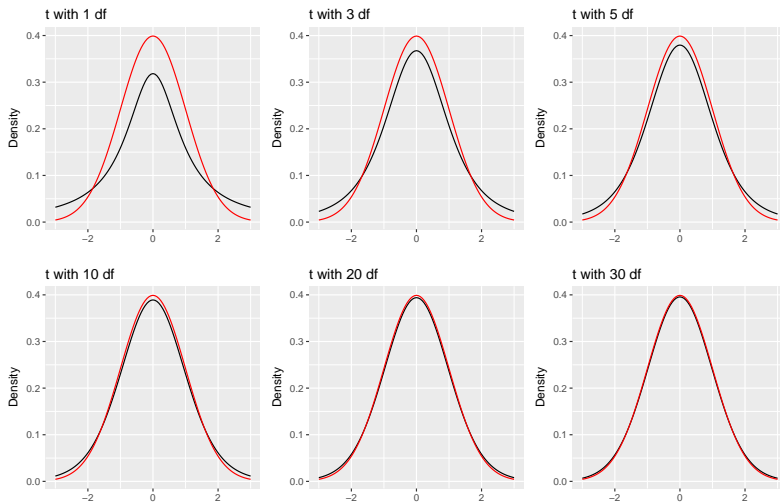
# Student's t distribution

Student's t distribution looks a lot like a Normal distribution, when the sample size is large. Unlike the normal distribution, which is specified by two parameters, the mean and the standard deviation, the t distribution is specified by one parameter, the degrees of freedom.

- t distributions with large numbers of degrees of freedom are more or less indistinguishable from the standard Normal distribution.
- t distributions with smaller degrees of freedom (say, with  $df < 30$ , in particular) are still symmetric, but are more outlier-prone than a Normal distribution

# Six t Distributions and a Standard Normal

Various t distributions and the Standard Normal



Standard Normal shown in red

# Building the CI by hand for the Systolic BP data

In the SBP data, we observe the following results:

```
dm192 %>%  
  summarize(n = length(sbp), sample_mean = mean(sbp),  
            sample_sd = sd(sbp),  
            std.error = sd(sbp)/sqrt(n)) %>%  
  round(digits = 2) %>%  
  knitr::kable()
```

n	sample_mean	sample_sd	std.error
192	134.21	17.78	1.28

# Building the CI by Hand, 2

Let's build a 90% confidence interval for the true mean SBP across the entire population of NE Ohio adults with diabetes.

- The confidence level will be 90%, or 0.90
- The  $\alpha$  value, which is  $1 - \text{confidence} = 0.10$ .
- From the summaries above, we know that
  - $n = 192$ ,
  - $\bar{x} = 134.21$  and
  - $s = 17.78$ ,
  - and that our standard error of the sample mean is 1.28.

# Calculating the CI

The two-sided  $100(1-\alpha)\%$  confidence interval (based on a  $t$  test) is:

$$\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n}), \text{ or}$$

- The 90% CI for  $\mu$  is  $134.21 \pm t_{0.10/2, 192-1} (1.28)$ 
  - To calculate the  $t$  cutoff value for  $\alpha = 0.10$  and  $n = 192$ , we use

$$\text{qt}(0.10/2, \text{df} = 192-1, \text{lower.tail}=\text{FALSE}) = 1.6528705$$

- So the 90% CI for  $\mu$  is  $134.21 \pm 1.653 \times 1.28$ , or
- $134.21 \pm 2.12$ , or  $(132.09, 136.33)$

So, our 90% confidence interval for the true population mean SBP level across NE Ohio adults with diabetes, based on our sample of 192 such adults, is  $(132.1, 136.3)$  mm Hg.

# Getting R to build a CI for $\mu$

Happily, R does all of this work, and with less inappropriate rounding.

```
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 the Confidence Interval

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

	estimate	statistic	p.value	parameter	conf.low
1	134.2083	104.5979	1.434014e-170	191	132.0876
	conf.high	method	alternative		
1	136.3291	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<sup>1</sup>.

---

<sup>1</sup>Since the actual SBP values are integers, we should include no more than one additional significant figure 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
```



# Using Different Levels of Confidence

Below, we see two-sided confidence intervals for various levels of  $\alpha$ .

Confidence Level	$\alpha$	Two-Sided Interval Estimate for SBP Population Mean, $\mu$	Point Estimate for SBP Population Mean, $\mu$
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  $\mu$  or a lower limit for  $\mu$ , 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  $100(1-\alpha)\%$  confidence interval for the population mean is  $\mu \leq \bar{x} + t_{\alpha/2, n-1}(\frac{s}{\sqrt{n}})$ , with lower “bound”  $-\infty$ .
- The corresponding lower bound for a one-sided  $100(1 - \alpha)$  CI for  $\mu$  would be  $\mu \geq \bar{x} - t_{\alpha/2, n-1}(\frac{s}{\sqrt{n}})$ , with upper “bound”  $\infty$ .

# One-Sided CI for $\mu$

```
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 $\mu$

```
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 intervals.

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

90% or 0.90 | 0.10 | One Sided ( $<$ ) |  $\mu < 135.86$

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

95% or 0.95 | 0.05 | One Sided ( $<$ ) |  $\mu < 136.33$

# 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 $\alpha$ and One-Sided vs. Two-Sided CIs for $\mu$

Table of t-based estimates follows. . .

Confidence Level	$\alpha$	2-Sided Interval Estimate for $\mu$ , Population Mean SBP	1-Sided Lower Bound for $\mu$
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 $\alpha$ and One-Sided vs. Two-Sided CIs for $\mu$

Table of t-based estimates follows. . .

Confidence Level	$\alpha$	2-Sided Interval Estimate for $\mu$ , Population Mean SBP	1-Sided Lower Bound for $\mu$
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?

# Large Sample Approaches (in Brief)

When you have a large sample size, say, more than 60 observations, the difference between a confidence interval 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  $100(1-\alpha)\%$  confidence interval using the Normal distribution, based on the sample mean  $\bar{x}$ , the sample size  $n$ , and the (known) population standard deviation  $\sigma$ .

# The Large Sample Formula for the CI around $\mu$

The two-tailed  $100(1-\alpha)\%$  confidence interval for a population mean  $\mu$  (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  $100(1-\alpha) = 95$ , so that  $\alpha$  is 0.05, or 5%. This means that the cutoff value we need to find is  $Z_{0.05/2} = Z_{0.025}$ , 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<sup>2</sup>.
- 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.

---

<sup>2</sup>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, $\mu$

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<sup>3</sup>.

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?

---

<sup>3</sup>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 $\mu$

What the computer does:

- ➊ Resample the data with replacement, until it obtains a new sample that is equal in size to the original data set.
- ➋ Calculates the statistic of interest (here, a sample mean.)
- ➌ Repeat the steps above many times (the default is 1,000 using our approach) to obtain a set of 1,000 sample means.
- ➍ 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.
- ➎ Send us a result, containing the sample mean, and a 90% confidence interval for the population mean

# When is a Bootstrap Confidence Interval for $\mu$ 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 $\mu$ using bootstrap

The command that we use to obtain a CI for  $\mu$  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 $\mu$ 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  $\alpha$ , the significance level you want to use in your one-sided confidence interval. Remember that  $\alpha$  is 1 minus the confidence level. Let's assume we want a 90% one-sided interval, so  $\alpha = 0.10$ .
- 2 Double  $\alpha$  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 $\mu$ 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,  $\mu$ , 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,  $\mu$ , 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.

# Project Task A, due 2017-10-13 at Noon

I have to read lots of these + REDOs - reduce my pain!

- 1 Make one submission, via Canvas, as requested. I'll review it then, and get back to you. I cannot review your materials in advance, or I'd go insane. I process this work in batches.
- 2 The only sorts of questions I'm going to answer in advance are highly specific ones about the instructions for the Task. If you want clarification of what I'm looking for, no problem. If you want us to evaluate your work, you'll have to wait.
- 3 If your response to Task A is more than three pages long, cut it to three pages. For the proposal summary, I specified a word limit. Obey this in your first submission. One page is more than enough for the proposal in *every* case. Some data descriptions can easily be done in a page, as well, but some can't. Try to keep that to two pages. One is better if you answer my questions.
- 4 Don't make me search for things - label what you're doing using my labels from the instructions.