

# 431 Class 13

[thomaseLove.github.io/431](https://thomaseLove.github.io/431)

2020-10-06

# Today's Agenda

- Confidence Intervals for a Mean
  - with indicator variable regression or with a  $t$  distribution
  - with the bootstrap
  - with the Wilcoxon signed-rank procedure
  - Interpreting the Results
- $p$  values and statistical significance
- The `dm431` data

# Today's R Packages and Data

```
library(broom)
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)

theme_set(theme_bw())

dm431 <- readRDS("data/dm431_2020.Rds")
```

# Fundamentals of Statistical Inference

# 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 Questions: Making Inferences From A Sample

- ① What is the population about which 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 study population not represent the target population?

For more, see Spiegelhalter, Chapter 3.

# Systolic Blood Pressure in the dm431 data

Here, I will look at systolic blood pressure values from a sample of 431 adult patients living in Northeast Ohio between the ages of 31 and 70, who have a diagnosis of diabetes, as gathered in the dm431 data.

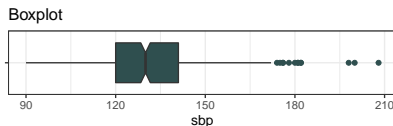
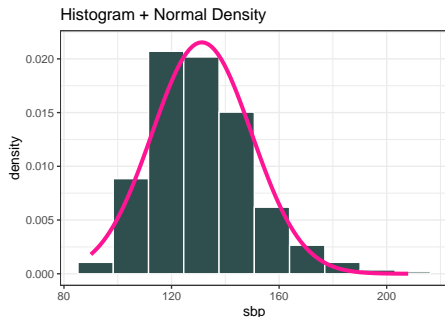
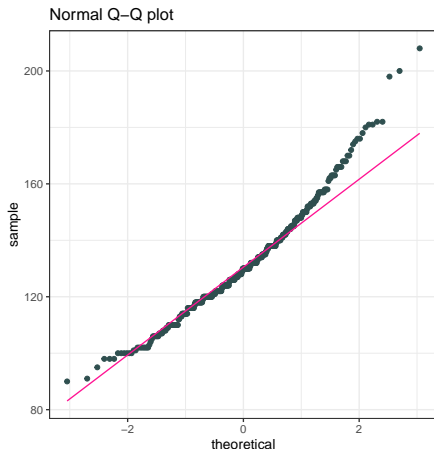
- These data are simulated to mirror some details from real data gathered by *Better Health Partnership*.
- The dm431 data contains multitudes, but for now, we're just looking at 431 systolic blood pressure values, gathered in the sbp variable.

```
mosaic::favstats(~ sbp, data = dm431)
```

min	Q1	median	Q3	max	mean	sd	n	missing
90	120	130	141	208	131.2645	18.52038	431	0

- See next slide. How reasonable is a Normal model for sbp?

# Graphical Summaries: sbp in dm431



Does a Normal model seem *very*, *somewhat* or *not* reasonable?



# 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.

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.

The key tradeoffs are

- cost vs. precision (larger samples produce narrower intervals), and
- precision vs. confidence in the correctness of the statement.

# Our Assumptions

Suppose that

- systolic BPs across the population of NE Ohio adults ages 31-70 living with diabetes follows a Normal distribution (with mean  $\mu$  and standard deviation  $\sigma$ .)
- the 431 adults in our `dm431` tibble are a random sample from that population.

We know the sample mean (131.26 of our 431 adults, but we don't know  $\mu$ , the mean across **all** NE Ohio adults ages 31-70 living with diabetes.

So we need to estimate it, by producing a **confidence interval for the true (population) mean**  $\mu$  of all adults with diabetes ages 31-70 living in NE Ohio based on this sample.

# Available Methods

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

- ① A **t-based** estimate and confidence interval, available from an intercept-only linear model, or (equivalently) from a t test.
  - This approach will require an assumption that the population comes from a Normal distribution.
- ② A **bootstrap** confidence interval, which uses resampling to estimate the population mean.
  - This approach won't require the Normality assumption, but has some other constraints.
- ③ A **Wilcoxon signed rank** approach, but that won't describe the mean, only a pseudo-median.
  - This also doesn't require the Normality assumption, but no longer describes the population mean (or median) unless the population can be assumed symmetric. Instead it describes the *pseudo-median*.

# Starting with A Good Answer

Use indicator variable regression to produce a t-based interval.

```
model1 <- lm(sbp ~ 1, data = dm431)

tidy(model1, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error,
         conf.low, conf.high, p.value) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high	p.value
(Intercept)	131.26	0.89	129.79	132.74	0

- Our point estimate for the population mean SBP ( $\mu$ ) is 131.26 mm Hg.
- Our 90% confidence interval is (129.79, 132.74) mm Hg for  $\mu$ .

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

- Our 90% confidence interval estimate for  $\mu$  turns out to be (129.79, 132.74) mm Hg. How do we interpret this result?
- Some people think this means that there is a 90% chance that the true mean of the population,  $\mu$ , falls between 129.79 and 132.74 mm Hg.

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

- Our 90% confidence interval estimate for  $\mu$  turns out to be (129.79, 132.74) mm Hg. How do we interpret this result?
- Some people think this means that there is a 90% chance that the true mean of the population,  $\mu$ , falls between 129.79 and 132.74 mm Hg.
- That's not correct. Why not?

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

- Our 90% confidence interval estimate for  $\mu$  turns out to be (129.79, 132.74) mm Hg. How do we interpret this result?
- Some people think this means that there is a 90% chance that the true mean of the population,  $\mu$ , falls between 129.79 and 132.74 mm Hg.
- That's not correct. Why not?
- The population mean  $\mu$  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.

# So what do we have confidence in?

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 of our 90% confidence interval for  $\mu$ ?



# A somewhat better interpretation

- Our 90% confidence interval for  $\mu$  is (129.79, 132.74) mm Hg.

If we used this same method to sample data from the true population of adults ages 31-70 with diabetes in NE Ohio and built 100 such 90% confidence intervals, then 90 of them would contain the true population mean. We don't know whether this one interval we built contains  $\mu$ , though.

- We call  $100(1 - \alpha)\%$ , here, 90%, or 0.90, the *confidence* level, and
- $\alpha = 10\%$ , or 0.10 is called the *significance* level.

The indicator variable approach we've used is identical to a t test.

# Using t test to find the CI for $\mu$

```
t1 <- dm431 %$%  
  t.test(sbp, conf.level = 0.90, alt = "two.sided")  
  
t1
```

## One Sample t-test

```
data: sbp  
t = 147.14, df = 430, p-value < 2.2e-16  
alternative hypothesis: true mean is not equal to 0  
90 percent confidence interval:  
 129.794 132.735  
sample estimates:  
mean of x  
 131.2645
```

# Tidying the t test results

```
tidy(t1) %>% select(estimate, conf.low, conf.high) %>%  
  kable(digits = 2)
```

estimate	conf.low	conf.high
131.26	129.79	132.74

```
tidy(t1) %>% select(method, alternative, statistic,  
                    parameter, p.value) %>%  
  kable()
```

method	alternative	statistic	parameter	p.value
One Sample t-test	two.sided	147.1418	430	0

- The `statistic` is the t statistic (estimate / standard error)
- The `parameter` describes the degrees of freedom (here,  $n - 1$ )

# 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.

- 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 upper 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	(129.79, 132.74)
95% or 0.95	0.05	One Sided ( $<$ )	$\mu < 132.74$
95% or 0.95	0.05	One Sided ( $>$ )	$\mu > 129.79$

Want to calculate the t-based CI by hand?

# What is the formula for the t-based CI?

Many confidence intervals follow a general strategy using a point estimate  $\pm$  a margin for error.

We 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  $s_x$ . The two-sided  $100(1-\alpha)\%$  confidence interval is:

$$\bar{x} \pm t_{\alpha/2, n-1} \left( \frac{s_x}{\sqrt{n}} \right)$$

- $SE(\bar{x}) = \frac{s_x}{\sqrt{n}}$  is the standard error of the sample mean
- The margin of error for this CI is  $t_{\alpha/2, n-1} \left( \frac{s_x}{\sqrt{n}} \right)$ .
- $t_{\alpha/2, n-1}$  is the value that cuts off the top  $\alpha/2$  percent of the  $t$  distribution, with  $n - 1$  degrees of freedom. Obtain in R with:

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

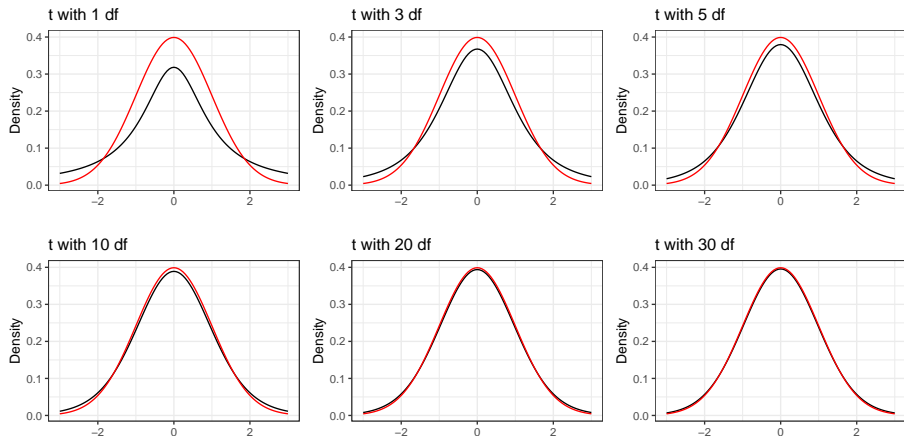
# Student's t distribution

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



# “Hand-Crafting” the 90% confidence interval for $\mu$

Let's build a 90% confidence interval for the true mean SBP across the entire population of NE Ohio adults ages 31-70 with diabetes.

$\alpha$	$n$	$\bar{x}$	$s$	$SE(\bar{x}) = s/\sqrt{n}$
0.10	431	131.26	18.52	0.89

The two-sided  $100(1-\alpha)\%$  confidence interval (based on a  $t$  test) is:  
 $\bar{x} \pm t_{\alpha/2, n-1}(s/\sqrt{n})$ , or

- The 90% CI for  $\mu$  is  $131.26 \pm t_{0.10/2, 431-1} (0.89)$ 
  - To calculate the  $t$  cutoff value for  $\alpha = 0.10$  and  $n = 431$ , we use

`qt(0.10/2, df = 431-1, lower.tail=FALSE) = 1.648405`

- So the 90% CI for  $\mu$  is  $131.26 \pm 1.6484 \times 0.89$ , or
- $131.26 \pm 1.47$ , or  $(129.79, 132.73)$

# A Few Thoughts on Hypothesis Testing about a Population Mean

# Four Steps to Complete a Hypothesis Test

- 1 Specify the null hypothesis,  $H_0$
- 2 Specify the research or alternative hypothesis,  $H_1$ , sometimes called  $H_A$
- 3 Specify the approach to be used to make inferences to the population based on sample data.
  - We must specify  $\alpha$ , 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 an appropriate point estimate and confidence interval (and maybe a  $p$  value.)

# In our Setting

- ❶ Null Hypothesis:  $H_0 : \mu = 0$
- ❷ Alternative Hypothesis:  $H_A : \mu \neq 0$
- ❸ Test statistic:  $t = \text{estimate} / \text{standard error}$  with  $\alpha = 0.10$  since we're using a 90% confidence interval
- ❹ Our 90% confidence interval for  $\mu$  is (129.79, 132.74) mm Hg.
  - Does  $H_0 : \mu = 0$  seem consistent with the data, or do we find a detectable difference between our data's estimates of  $\mu$  and the assumption that  $\mu = 0$ ?
  - Be careful. Data lie all the time.

# Defining a $p$ Value (but not very well)

The  $p$  value estimates the probability that we would obtain a result as much in favor or more in favor of the alternative hypothesis  $H_A$  as we did, assuming that  $H_0$  is true.

- 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.

## How people use the $p$ Value

- If the  $p$  value is less than  $\alpha$ , this suggests we might reject  $H_0$  in favor of  $H_A$ , and declare the result statistically significant.

But we won't be comfortable with doing that, at least in time.

# 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 closer to** 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.

*P-values have taken quite a beating lately. These widely used and commonly misapplied statistics have been blamed for giving a veneer of legitimacy to dodgy study results, encouraging bad research practices and promoting false-positive study results.*

*Last week, I attended the inaugural METRICS conference at Stanford, which brought together some of the world's leading experts on meta-science, or the study of studies. I figured that if anyone could explain p-values in plain English, these folks could.*

(Christie Aschwanden, FiveThirtyEight, 2015-11-24)

## FiveThirtyEight

---

Politics

Sports

**Science**

Podcasts

Video

NOV. 24, 2015, AT 12:12 PM

# Not Even Scientists Can Easily Explain P-values

By Christie Aschwanden

Filed under Scientific Method



Link: <https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/>



# Do Scientists Get This Right?

*Scientists regularly get it wrong, and so do most textbooks, said Steven Goodman (co-director of the METRICS conference.) Even after spending his “entire career” thinking about p-values, he said he could tell me the definition, “but I cannot tell you what it means, and almost nobody can.”*

*When Goodman speaks to large audiences of scientists, he often presents correct and incorrect definitions of the p-value, and they “very confidently” raise their hand for the wrong answer. “Almost all of them think it gives some direct information about how likely they are to be wrong, and that’s definitely not what a p-value does,” Goodman said.*

(Christie Aschwanden, FiveThirtyEight, 2015-11-24)

# Can we define a p value better?

*I've come to think that the most fundamental problem with p-values is that no one can really say what they are.*

*What I learned by asking all these very smart people to explain p-values is that I was on a fool's errand. Try to distill the p-value down to an intuitive concept and it loses all its nuances and complexity, said science journalist Regina Nuzzo, a statistics professor at Gallaudet University. "Then people get it wrong, and this is why statisticians are upset and scientists are confused."*

*You can get it right, or you can make it intuitive, but it's all but impossible to do both.*

(Christie Aschwanden, FiveThirtyEight, 2015-11-24)

# 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.

# Confidence Intervals using Bootstrap Resampling

# 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 “431 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 90% confidence interval

The bootstrap can be used to build a confidence interval for  $\mu$  without the assumption that the population follows a Normal distribution.

```
set.seed(2020)
```

```
Hmisc::smean.cl.boot(dm431$sbp, conf.int = .90, B = 1000)
```

Mean	Lower	Upper
131.2645	129.8165	132.7594

- I often use  $B = 1,000$  (the default) or 10,000 bootstrap replications for building CIs, but it isn't usually critical.
- A bootstrap interval is often asymmetric, and for highly skewed data, the point estimate might not be near the center of the interval.
- A “downside” is that you and I will get (somewhat) different answers if we resample from the same data with different seeds.

# Bootstrap vs. t-Based Confidence Intervals

- Hmisc's `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.

```
set.seed(431)
```

```
dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.90)
```

	Mean	Lower	Upper
131.2645	129.8046	132.7290	

```
dm431 %$% Hmisc::smean.cl.normal(sbp, conf = 0.90)
```

	Mean	Lower	Upper
131.2645	129.7940	132.7350	

# Bootstrap: Estimating a confidence interval for $\mu$

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

The bootstrap idea can be used to produce interval estimates for almost any population parameter, not just the mean.



# When is a Bootstrap CI 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 (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.)

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).

## The Wilcoxon Signed Rank Approach (if the data come from a symmetric population)

# The Wilcoxon Signed Rank Procedure for CIs

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

## What is a Pseudo-Median?

- For any sample, the pseudomedian is defined as the median of all of the midpoints of pairs of observations in the sample.
- As it turns out, if you're willing to assume the population is **symmetric** (but not necessarily Normally distributed) then the pseudo-median is equal to the population median.

# Wilcoxon rank sum based 90% confidence interval

```
wilcox.test(dm431$sbp, conf.int = TRUE, conf.level = 0.90)
```

Wilcoxon signed rank test with continuity  
correction

data: dm431\$sbp

V = 93096, p-value < 2.2e-16

alternative hypothesis: true location is not equal to 0

90 percent confidence interval:

129.0 131.5

sample estimates:

(pseudo)median

130

# Interpreting the Wilcoxon Signed Rank CI

If we're willing to believe the sbp values come from a population with a symmetric distribution, the 90% Confidence Interval for the population median would be (129, 131.5)

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

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

```
mosaic::favstats(~ sbp, data = dm431)
```

min	Q1	median	Q3	max	mean	sd	n	missing
90	120	130	141	208	131.2645	18.52038	431	0

# Tidying the Wilcoxon Results

```
w1 <- dm431 %$%  
  wilcox.test(sbp, conf.int=TRUE, conf.level=0.90)  
  
tidy(w1) %>% select(method, alternative) %>% kable()
```

method	alternative
Wilcoxon signed rank test with continuity correction	two.sided

```
tidy(w1) %>% select(estimate, conf.low, conf.high) %>% kable()
```

estimate	conf.low	conf.high
130	129	131.5

# Reviewing Our Three Methods

For estimating the population mean...

- ❶ A **t-based** estimate and confidence interval, available from an intercept-only linear model, or (equivalently) from a t test.
  - This approach will require an assumption that the population comes from a Normal distribution.
- ❷ A **bootstrap** confidence interval, which uses resampling to estimate the population mean.
  - This approach won't require the Normality assumption, but has some other constraints.
- ❸ A **Wilcoxon signed rank** approach, but that won't describe the mean, only a pseudo-median.
  - This also doesn't require the Normality assumption, but no longer describes the population mean (or median) unless the population can be assumed symmetric. Instead it describes the *pseudo-median*.