431 Class 12

github.com/THOMASELOVE/2019-431

2019-10-03

Today's Agenda (Notes Chapters 16-17)

1. Statistical Inference and the `dm431` data
   - Point Estimates and Confidence Intervals for a Population Mean (quantitative data)
2. Group Work on Project Study A Proposal

Today's Setup and Data

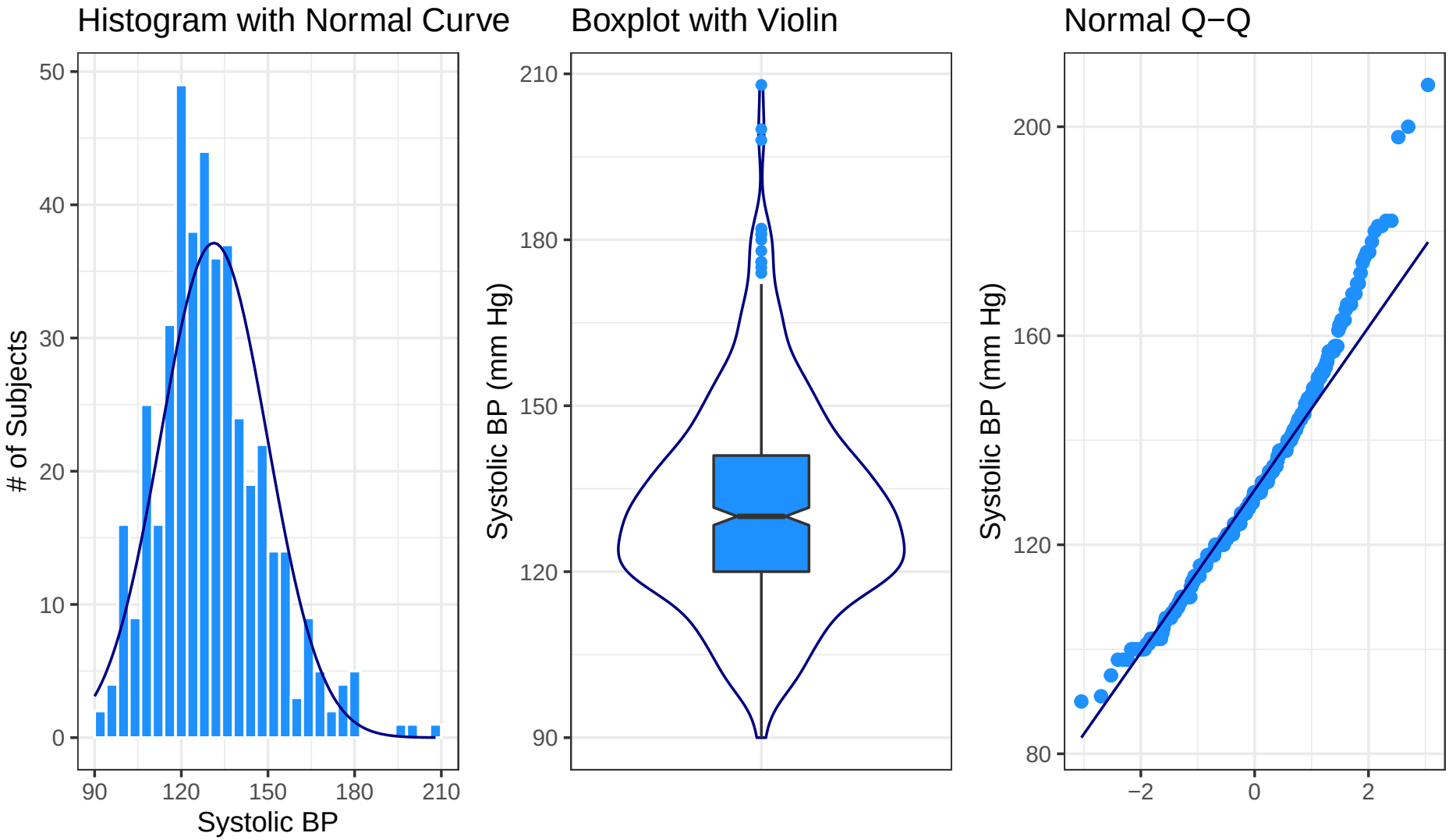
```
library(magrittr); library(janitor)
library(patchwork); library(here);
library(broom)
library(tidyverse)

source(here("R", "Love-boost.R"))

dm431 <- readRDS(here("data", "dm431.Rds"))
```

Graphical Summaries: `sbp` in `dm431`

Systolic BP (mm Hg) for 431 NE Ohio Adults with Diabetes



Confidence Intervals for a Population Mean

Our Sample and Our Population

Sample: 431 adult patients living in Northeast Ohio between the ages of 31 and 70, who have a diagnosis of diabetes.

- Sample Mean Systolic Blood Pressure = 131.3

Our population: **All** adult patients living in Northeast Ohio between the ages of 31 and 70, who have a diagnosis of diabetes.

Our first inferential goal will be to produce a **confidence interval for the true (population) mean** systolic blood pressure 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 for the mean (or median).
3. A **Wilcoxon signed rank** confidence interval (for the pseudo-median).

Population Mean Estimation using the t distribution

What do we need? (Besides a computer running R.)

1. An assumption that the data in our sample come from a population that follows a Normal distribution.
2. An assumption that random sampling from the population is a good model for how the data were collected.
    - We assume samples were taken from the population independently, and they have identical distributions.
3. A pre-specified confidence level 100*(1 - α) for our confidence interval.
4. The sample itself, to determine the sample size n (of non-missing values), the sample mean $\bar{x}$ and the sample standard deviation s_x.
    - These will let us calculate:
        - our point estimate of the population mean μ
        - the standard error of the sample mean
        - the margin of error (half-width) of our confidence interval

Building a 90% Confidence Interval for μ

```
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
```

- The sample mean $\bar{x} = 131.26$, and this is also our point estimate of the population mean μ.
- The sample standard deviation is $s_x = 18.52$.
- We have $n = 431$ observations
- If we want a 90% confidence interval, then $\alpha = 0.10$.

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 σ is the standard deviation of the measurements in the population.
- We often estimate this particular standard error with its sample analogue, $s_x/\sqrt{n}$, where s_x is the sample standard deviation.
- Other statistics have different standard errors.
  - For p, the sample proportion, $\sqrt{p(1-p)/n}$ is the standard error using a sample of size n.
  - For r, the sample Pearson correlation, $\sqrt{\frac{1-r^2}{n-2}}$ is the standard error using n pairs of observations.

Standard Error of the Mean for the SBP data

The standard deviation of the SBP data turns out to be 18.52, with $n = 431$ observations, so we estimate the standard error of the mean is

$$SE_{mean}(\text{SBP}) = \frac{SD(\text{SBP})}{\sqrt{n}} = \frac{18.52}{\sqrt{431}} = 0.89$$

This standard error will play an important role in the development of our confidence interval using the t distribution.

```
dm431 %$% psych::describe(sbp) %>%
  select(n, mean, sd, se)
```

```
     n   mean    sd   se
X1 431 131.26 18.52 0.89
```

Confidence Interval for a population mean

We can build a 100(1-α)% 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-α)% confidence interval (based on a t test) is:

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

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:

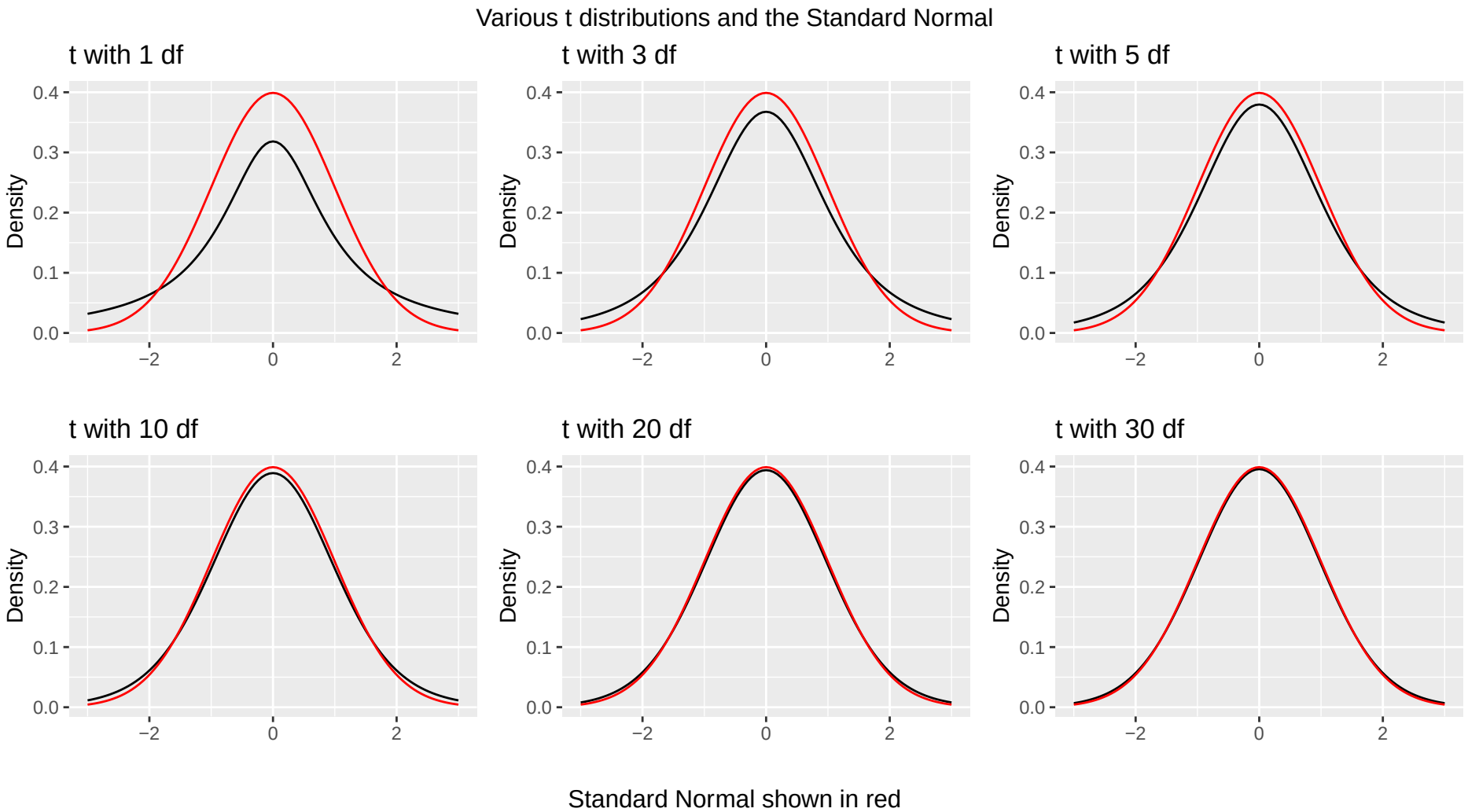
```
qt(alphaover2, df = n-1, lower.tail=FALSE)
```

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



"Hand-Crafting" the 90% confidence interval for μ

α	n	$\bar{x}$	s_x	$\text{SE}(\bar{x})$
0.10	431	131.26	18.52	0.89

Our two-sided $100(1-\alpha)\%$ confidence interval is:

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

We need the t cutoff value for $\alpha = 0.10$ and $n = 431$.

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

So our 90% confidence interval is:

$$131.26 \pm 1.648(0.89), \text{ or } 131.26 \pm 1.47, \text{ or } (129.79, 132.73)$$

What is the margin of error in this confidence interval?

Getting R to build a CI for μ

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

```
t1 <- dm431 %$% t.test(sbp, conf.level = 0.90,
                       alternative = "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
```

Summarizing the Confidence Interval

```
tidy(t1) %>% # from broom package
  select(estimate, conf.low, conf.high, method, alternative)
```

```
# A tibble: 1 x 5
  estimate conf.low conf.high method          alternative
     <dbl>    <dbl>     <dbl> <chr>           <chr>
1     131.     130.      133. One Sample t~ two.sided
```

estimate <dbl>	conf.low <dbl>	conf.high <dbl>	method <chr>	alternative <chr>
131.2645	129.794	132.735	One Sample t-test	two.sided

Since the actual SBP values are integers, we should probably include no more than one additional significant figure in our confidence interval.

We've Seen This Result Before

This intercept-only linear regression model yields the same estimates.

```
model1 <- lm(sbp ~ 1, data = dm431)
tidy(model1, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

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

- Our point estimate for the population mean SBP (μ) will be 131.26 mm Hg based on the dm431 sample.
- Our 90% confidence interval estimate for μ turns out to be (129.79, 132.74) mm Hg.

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

```
model1 <- lm(sbp ~ 1, data = dm431)
tidy(model1, conf.int = TRUE, conf.level = 0.95) %>%
  select(estimate, std.error, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

estimate	std.error	conf.low	conf.high
131.26	0.89	129.51	133.02

```
dm431 %$% t.test(sbp, conf.level = 0.95) %>%
  tidy() %>%
  select(estimate, conf.low, conf.high, alternative) %>%
  knitr::kable(digits = 2)
```

estimate	conf.low	conf.high	alternative
131.26	129.51	133.02	two.sided

Using Different Levels of Confidence

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

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	(130.1, 132.4)	131.3
90% or 0.90	0.10	(129.8, 132.7)	131.3
95% or 0.95	0.05	(129.5, 133)	131.3
99% or 0.99	0.01	(129, 133.6)	131.3

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

```
dm431 %$%
  t.test(sbp, conf.level = 0.90, alt = "greater") %>%
  tidy() %>% select(estimate, conf.low, conf.high)
```

```
# A tibble: 1 x 3
  estimate conf.low conf.high
     <dbl>    <dbl>     <dbl>
1     131.     130.       Inf
```

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	(130.12, 132.41)
90% or 0.90	0.10	One Sided ($>$)	$\mu > 130.12$

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

Interpreting the Result

(129.79, 132.74) mm Hg. is a 90% two-sided confidence interval for the population mean SBP among NE Ohio adults with diabetes. How can we interpret that?

- Our point estimate for the true population mean SBP among NE Ohio adults with diabetes is 131.26 mm Hg. The values in the interval (129.79, 132.74) 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 ages 31-70 with diabetes.
- Were we to draw 100 samples of size 431 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 ages 31-70 with diabetes.

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.

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*.

Confidence Intervals using Bootstrap Resampling

Bootstrap 90% confidence interval

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

```
set.seed(431)
Hmisc::smean.cl.boot(dm431$sbp, conf.int = .90, B = 1000)
```

```
    Mean     Lower     Upper
131.2645 129.8046 132.7290
```

The bootstrap will be less effective (in some ways) than the t-distribution approach when the data really do follow a Normal distribution.

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, 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: 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(20191003)
dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.90)
```

```
    Mean     Lower     Upper
131.2645 129.8724 132.7802
```

Bootstrap vs. 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.

```
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
```

Rerunning 90% CI for μ via Bootstrap

```
set.seed(2019431)
  dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.9)
```

```
    Mean    Lower    Upper
131.2645 129.7817 132.8150
```

```
set.seed(4312019)
  dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.9)
```

```
    Mean    Lower    Upper
131.2645 129.8552 132.7912
```

Bootstrap: Changing the Confidence Level

```
set.seed(43105); Hmisc::smean.cl.boot(dm431$sbp, conf = 0.90)
```

```
    Mean    Lower    Upper
131.2645 129.7326 132.7942
```

```
set.seed(43106); Hmisc::smean.cl.boot(dm431$sbp, conf = 0.95)
```

```
    Mean    Lower    Upper
131.2645 129.6032 133.0082
```

```
set.seed(43107); Hmisc::smean.cl.boot(dm431$sbp, conf = 0.99)
```

```
    Mean    Lower    Upper
131.2645 129.2106 133.6638
```

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)
dm431 %$% Hmisc::smean.cl.boot(sbp, conf = 0.80)
```

```
    Mean     Lower     Upper
131.2645 130.0694 132.3237
```

The upper bound of this two-sided 80% CI 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).