431 Class 13

github.com/THOMASELOVE/2019-431

2019-10-08

Our Agenda (Notes Chapters 16-18)

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

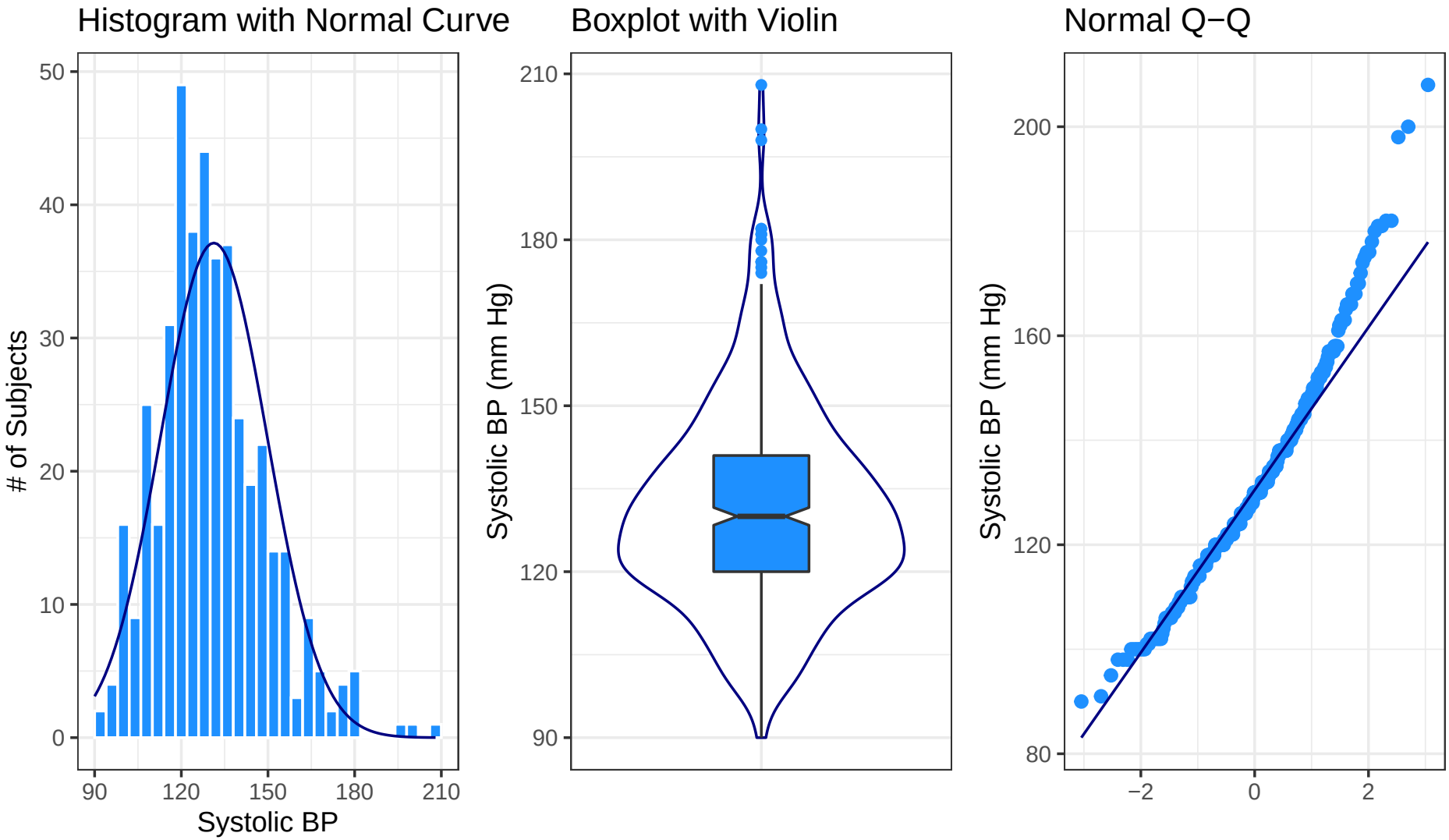
Today's Setup and Data

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

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

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

The boot package will be introduced today.

Graphical Summaries: `sbp` in `dm431`

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



Confidence Intervals for a Population Mean

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*.

Our Goal

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 our sample of 431 such adults.

Results so far (from Class 12)

90% Confidence Intervals for μ

Basis	90% CI
t-distribution	(129.79, 132.74) mm Hg
bootstrap	(129.85, 132.79) mm Hg

(used `set.seed(4312019)` in bootstrap)

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(2019431)
boot.ci(boot (dm431$sbp, f.median, 1000),
        conf=0.90, type="basic")
```

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

CALL :
boot.ci(boot.out = boot(dm431$sbp, f.median, 1000), conf = 0.9
    type = "basic")

Intervals :
Level      Basic
90%   (130, 133 )
Calculations and Intervals on Original Scale
```

Bootstrap CI for the Population Median vs. Mean

- Note that the sample **median** of the SBP data is 130 mm Hg.
- Our 90% confidence interval for the population **median** SBP among NE Ohio adults with diabetes is (130, 133) according to the bootstrap, using the random seed 2019431.
- The sample **mean** of the SBP data is 131.3 mm Hg.
- The 90% bootstrap CI for the population **mean** SBP, μ, is (129.8, 132.8) if we use the random seed 2019431.

The Wilcoxon Signed Rank Procedure for CIs

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

tidy(w1) %>%
  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     130.     129.      132. Wilcoxon sig~ two.sided
```

estimate	conf.low	conf.high	method	alternative
<dbl>	*<dbl>*	*<dbl>*	*<chr>*	*<chr>*
130.	129.	132.	Wilcoxon signed rank test with continuity correction	two.sided

Confidence Intervals for a Population Proportion

Moving on from Means to Proportions

We've focused on creating statistical inferences about a population mean when we have a quantitative outcome. Now, we'll tackle a **categorical** outcome.

We'll estimate a confidence interval around an unknown population proportion, or rate, symbolized with π, on the basis of a random sample of n observations from the population of interest.

The sample proportion is called $\hat{p}$, which is sometimes, unfortunately, symbolized as p.

- This $\hat{p}$ is the sample proportion - not a p value.

Hemoglobin A1c $<$ 8 rate?

The dm431 data yields these results on whether each subject's Hemoglobin A1c level (a measure of blood sugar control) is below 8%[1].

```
dm431 %$%
  tabyl(a1c < 8)
```

```
 a1c < 8   n     percent valid_percent
   FALSE 147 0.341067285     0.3434579
    TRUE 281 0.651972158     0.6565421
      NA   3 0.006960557            NA
```

What can we conclude about the true proportion of Northeast Ohio adults ages 31-70 who live with diabetes whose A1c is below 8%?

[1] Having an A1c $<$ 8 is a good thing, generally, if you have diabetes.

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.

- 281 of our 431 adult patients, or 65.2% have A1c < 8.

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) proportion** with A1c < 8, across all adults with diabetes ages 31-70 living in NE Ohio, based on this sample.

A Confidence Interval for a Proportion

A 100(1-α)% confidence interval for the population proportion π can be created by using the standard normal distribution, the sample proportion, $\hat{p}$, and the standard error of a sample proportion, which is defined as the square root of $\hat{p}$ multiplied by $(1 - \hat{p})$ divided by the sample size, n.

Specifically, that confidence interval estimate is $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

where $Z_{\alpha/2}$ = the value from a standard Normal distribution cutting off the top $\alpha/2$ of the distribution, obtained in R by substituting the desired $\alpha/2$ value into: `qnorm(alpha/2, lower.tail=FALSE)`.

- *Note*: This interval is reasonably accurate so long as $n\hat{p}$ and $n(1 - \hat{p})$ are each at least 5.

Estimating π in the A1c $<$ 8 data

- We'll build a 95% confidence interval for the true population proportion, so $\alpha = 0.05$
- We have n $= 431$ subjects
- Sample proportion is $\hat{p} = .652$, since $281/431 = 0.652$.

The standard error of that sample proportion will be

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.652(1-0.652)}{431}} = 0.023$$

Confidence Interval for π = Pr(A1c < 8)

Our 95% confidence interval for the true population proportion, π, of people whose A1c is below 8 is:

$$\hat{p} \pm Z_{.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.652 \pm 1.96(0.023) = 0.652 \pm 0.045$$

or (0.607, 0.697).

To verify that $Z_{0.025} = 1.96...$

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

```
[1] 1.959964
```

Likely Accuracy of this Confidence Interval?

Since $n\hat{p} = (431)(0.652) = 281$ and $n(1 - \hat{p}) = (431)(1 - 0.652) = 150$ are substantially greater than 5, the CI should be reasonably accurate.

What can we conclude from this analysis?

- Point estimate of the population proportion with A1c < 8 is 0.652
- 95% confidence interval for the population proportion is (0.607, 0.697)

What is the "margin of error" in this confidence interval?

- The entire confidence interval has width 0.09 (or 9 percentage points.)
- The margin of error (or half-width) is 0.045, or 4.5 percentage points.

Happily, that's our last "by hand" calculation.

R Methods to get a CI for a Population Proportion

I am aware of at least three different procedures for estimating a confidence interval for a population proportion using R. All have minor weaknesses: none is importantly different from the others in many practical situations.

1. The `prop.test` approach (also called the Wald test)

```
prop.test(x = 281, n = 431)
```

2. The `binom.test` approach (Clopper and Pearson "exact" test)

```
binom.test(x = 281, n = 431)
```

3. Building a confidence interval via a `SAIFS` procedure

```
saifs.ci(x = 281, n = 431)
```

The prop.test approach (Wald test)

The prop.test function estimates a confidence interval for π:

```
prop.test(x = 281, n = 431)
```

```
    1-sample proportions test with continuity
    correction

data:  281 out of 431, null probability 0.5
X-squared = 39.211, df = 1, p-value = 3.804e-10
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.6046536 0.6965367
sample estimates:
        p
0.6519722
```

binom.test (Clopper-Pearson "exact" test)

```
binom.test(x = 281, n = 431)
```

```
    Exact binomial test

data:  281 and 431
number of successes = 281, number of trials =
431, p-value = 2.795e-10
alternative hypothesis: true probability of success is not equ
95 percent confidence interval:
 0.6049168 0.6969247
sample estimates:
probability of success
             0.6519722
```

Estimating a Rate More Accurately

Suppose you have some data involving n independent tries, with x successes. The most natural estimate of the "success rate" in the data is x / n.

But, strangely enough, it turns out this isn't an entirely satisfying estimator. Alan Agresti provides substantial motivation for the $(x + 1)/(n + 2)$ estimate as an alternative[2]. This is sometimes called a *Bayesian augmentation*.

[2] This note comes largely from a May 15 2007 entry in Andrew Gelman's blog at http://andrewgelman.com/2007/05/15

Use (x + 1)/(n + 2) rather than x/n

- The big problem with x / n is that it estimates p = 0 or p = 1 when x = 0 or x = n.
- It's also tricky to compute confidence intervals at these extremes, since the usual standard error for a proportion, $\sqrt{np(1-p)}$, gives zero, which isn't quite right.
- (x + 1)/(n + 2) is much cleaner, especially when you build a confidence interval for the rate.
- The only place where (x + 1)/(n + 2) will go wrong (as in the SAIFS approach) is if n is small and the true probability is very close to 0 or 1.
  - For example, if n = 10, and p is 1 in a million, then x will almost certainly be zero, and an estimate of 1/12 is much worse than the simple 0/10.
  - However, how big a deal is this? If p might be 1 in a million, are you going to estimate it with an experiment using n = 10?

Practical Impact of Bayesian Augmentation

It is likely that the augmented `(x + 1) / (n + 2)` version yields more accurate estimates for the odds ratio or relative risk or probability difference, but the two sets of estimates (with and without the augmentation) will be generally comparable, so long as. . .

a. the sample size in each exposure group is more than, say, 30 subjects, and/or

b. the sample probability of the outcome is between 0.1 and 0.9 in each exposure group.

Bayesian Augmentation: Add a Success and a Failure

You'll get slightly better results if you use $\frac{x+1}{n+2}$ rather than $\frac{x}{n}$ as your point estimate, and to fuel your confidence interval using either the `binom.test` or `prop.test` approach.

- The results will be better in the sense that they'll be slightly more likely to meet the nominal coverage probability of the confidence intervals.
- This won't make a meaningful difference if $\frac{x}{n}$ is near 0.5, or if the sample size n is large. Why?

Suppose you want to find a confidence interval when you have 2 successes in 10 trials. I'm suggesting that instead of `binom.test(x = 2, n = 10)` you might want to try `binom.test(x = 3, n = 12)`

SAIFS confidence interval procedure

SAIFS = single augmentation with an imaginary failure or success[3]

- Uses a function I built in R for you (Part of `Love-boost.R`)

```
saifs.ci(x = 281, n = 431)
```

```
Sample Proportion             0.025             0.975
            0.652             0.605             0.698
```

`saifs.ci` already builds in a Bayesian augmentation, so we don't need to do that here.

[3] see Notes Part B for more details.

Results for "A1c < 8" Rate (x = 281, n = 431)

Method	95% CI for π
`prop.test`	0.605, 0.697
`binom.test`	0.605, 0.697
`saifs.ci`	0.605, 0.698

Our "by hand" result, based on the Normal distribution, with no continuity correction, was (0.607, 0.697).

So in this case, it really doesn't matter which one you choose. With a smaller sample, we may not come to the same conclusion about the relative merits of these different approaches.

Assumptions behind Inferences about π

We are making the following assumptions, when using these inferential approaches:

1. There are n identical trials.
2. There are exactly two possible outcomes (which may be designated as success and failure) for each trial.
3. The true probability of success, π, remains constant across trials.
4. Each trial is independent of all of the other trials.

Accuracy of these Inferences about a Proportion

We'd like to see that both $n\hat{p}$ = observed successes and $n(1 - \hat{p})$ = observed failures exceed 5.

- If not, then the intervals may be both incorrect (in the sense of being shifted away from the true value of π), and also less efficient (wider) than necessary.

None of these approaches is always best

When we have a sample size below 100, or the sample proportion of success is either below 0.10 or above 0.90, caution is warranted[4], although the various methods often yield similar responses.

95% CI Approach	Wald	Clopper-Pearson	SAIFS
X = 10, n = 30	0.179, 0.529	0.173, 0.528	0.148, 0.534
X = 10, n = 50	0.105, 0.341	0.1, 0.337	0.083, 0.333
X = 90, n = 100	0.82, 0.948	0.824, 0.951	0.829, 0.96
X = 95, n = 100	0.882, 0.981	0.887, 0.984	0.894, 0.994

[4] We might consider using the Bayesian augmentation, especially for 'prop.test' or 'binom.test'

Next Up: Comparing Two Populations (Chapters 19-24 in the Notes)