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/

Leek, Chapter 6

- 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.

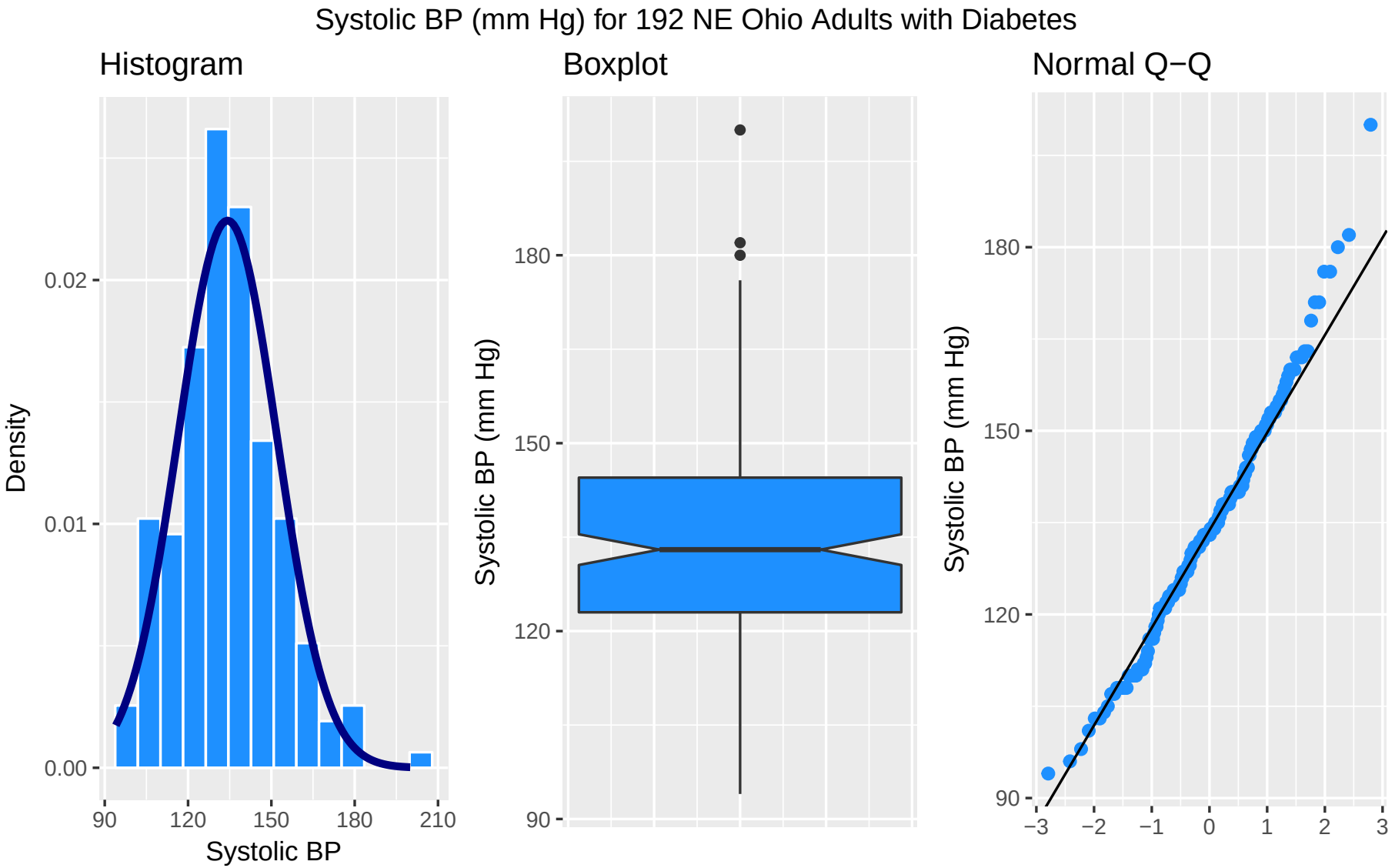
How Did I Build These Graphical Summaries

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



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 summaring 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, μ, 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 μ, the mean across all NE Ohio adults with diabetes. So we need to estimate it.

A 90% Confidence Interval for μ

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 μ turns out to be (132.1, 136.3) mm Hg. How would you interpret this result?

A 90% Confidence Interval for μ

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 μ 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, μ, falls between 132.1 and 136.3 mm Hg.

A 90% Confidence Interval for μ

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 μ 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, μ, 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 μ 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-α)%, 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 μ.

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 σ 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 μ, 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-α)% 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-α)% 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:

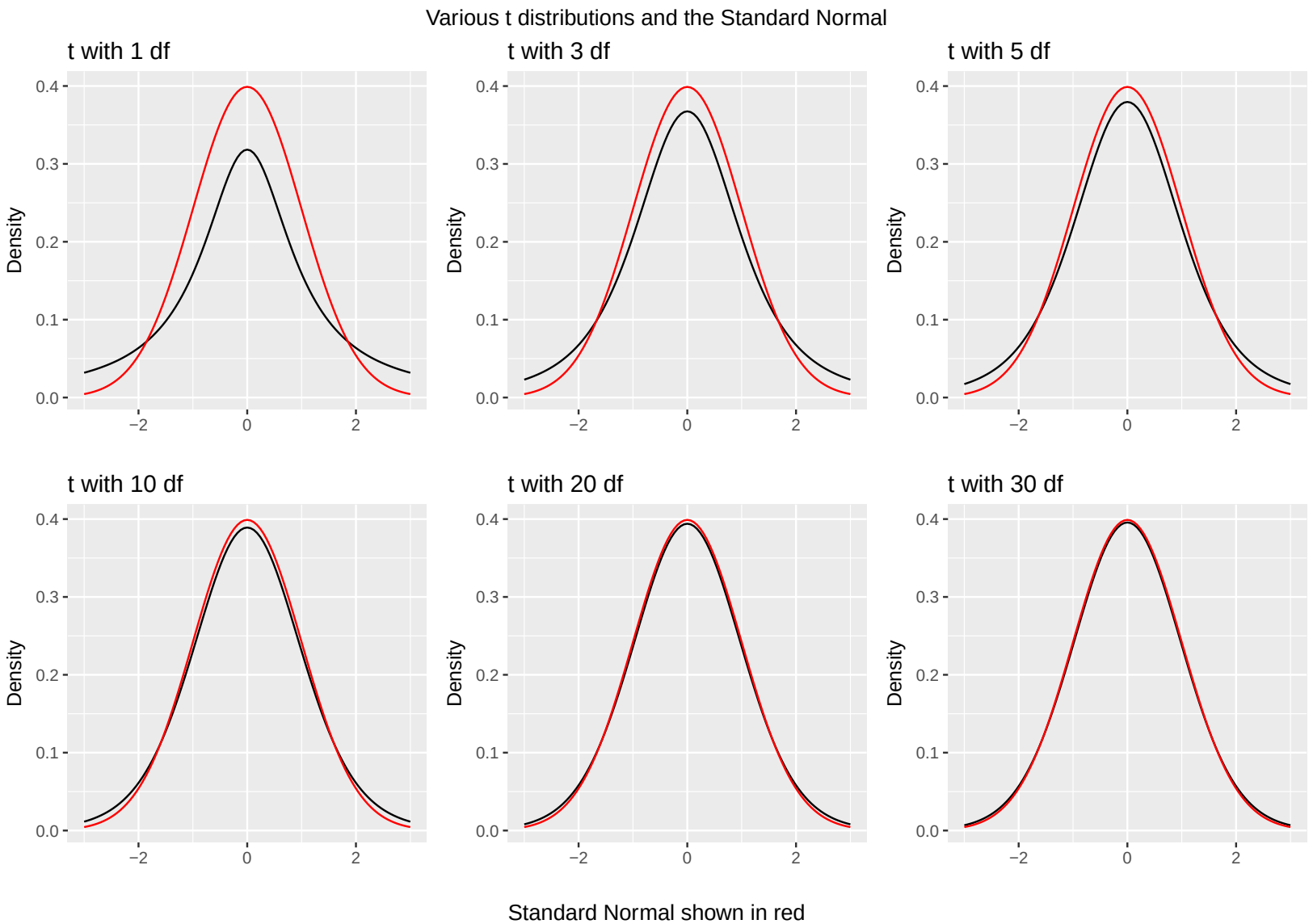
```
qt(alphaover2, df = n-1, lower.tail=FALSE)
```

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



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

- The 90% CI for μ 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

`qt(0.10/2, df = 192-1, lower.tail=FALSE)` = 1.6528705

- So the 90% CI for μ is 134.21 $\pm$ 1.653 x 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 μ

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[1].

[1] 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 α.

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

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

One-sided vs. Two-sided Confidence Intervals

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

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

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

One-Sided CI for μ

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

```
    One Sample t-test

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

Another One-Sided CI for μ

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

```
    One Sample t-test

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

Relationship between One-Sided and Two-Sided CIs

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

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

Why does this happen?

Why, indeed?

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

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

Interpreting the Result

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

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

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

Table of t-based estimates follows. . .

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

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

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

Table of t-based estimates follows. . .

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

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

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-α)% confidence interval using the Normal distribution, based on the sample mean $\bar{x}$, the sample size n, and the (known) population standard deviation σ.

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 betterm if you answer my questions.
4. Don't make me search for things - label what you're doing using my labels from the instructions.