BMR 617

The central limit theorem and the standard error of the mean

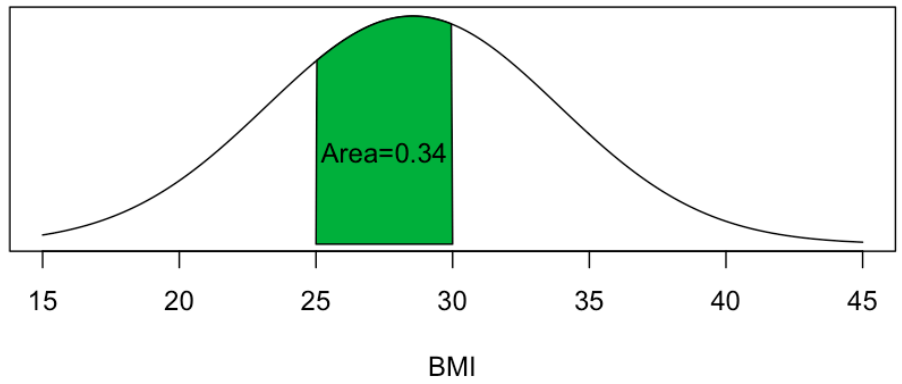
March 2nd 2021

Distributions

- Loosely speaking a *distribution* of a data set describes how likely it is a point in the data set will take on a particular value
- We can talk about distributions of samples, and of populations
- The distribution of a sample can always be completely known, because we know the values of every data point in the data set
- The distribution of a population is generally not known

Distributions of quantitative variables

- When talking about continuous quantitative variables, because the variable can take on any value on a continuum, the probability the variable is *exactly* equal to any given value is zero
- We have to talk about the probability the variable lies within a given range



The normal distribution

- The normal distribution is a particular distribution with certain properties
- It is symmetrical about its mean
- It is entirely determined by its mean and standard deviation
- It has the properties that the probability a value lies within one standard deviation of the mean is approximately 0.68, and within two standard deviations of the mean is approximately 0.95.

Sampling and the central limit theorem

- Suppose we take a sample of *n* values from some population, which has mean μ and standard deviation σ
  - The population can have *any* distribution
- We can compute the mean of the sample, m_1
  - The mean is an approximation to the mean of the population
  - We want to know how good an approximation it is
- Suppose we repeat this, and sample another *n* values
  - We will get a different sample mean, m_2
- If we repeat this over and over, we will have a whole set of different sample means

The central limit theorem

- The central limit theorem is one of the most important theoretical results in statistics. It states that:

  The sample means of a collection of samples of size *n* drawn from a distribution with mean μ and standard deviation σ is approximately normally distributed, with mean μ and standard deviation $\sigma/\sqrt{n}$
- Remember that we can (loosely speaking) interpret the standard deviation as the average distance of a point in a data set from the mean of the data set

Interpreting the central limit theorem

- What the central limit theorem tells us is:
  1. If we take a sample of size *n* and compute the sample mean, *m*, the sample mean approximates the mean of the population from which the sample was drawn
  2. The average error between the sample mean and the population mean is σ/n, where σ is the standard deviation of the population.
  3. Sample means are approximately normally distributed, no matter the distribution of the original population
     - The approximation improves the larger the value of n
     - The approximation also depends on the underlying distribution of the population

The problem with the central limit theorem

- One important thing to note about the central limit theorem is that the standard deviation of sample means, which we loosely interpret as the average distance (error) between a sample mean and the population mean is $\sigma/\sqrt{n}$,
  - where σ is the standard deviation ***of the population***
- The problem here is we never really know the entire population, so we never actually know σ
- We can, of course, calculate the standard deviation ***of the sample***, *s*, which is an approximation to σ
  - But how good an approximation depends on the distribution of the population

The standard error of the mean

- If we take a sample of size *n*, and compute the standard deviation *s*, then the quantity $s/\sqrt{n}$ is called the *standard error of the mean*
- It is an approximation to $\sigma/\sqrt{n}$, which is the average distance from the population mean to the mean of a sample of size *n*
  - I.e. it is an approximation to the average error we make in using the sample mean as an approximation to the population mean

Bar Charts

- We've previously plotted quantitative data using box plots and column scatter plots
  - Probably the best way to show the data
  - All data are shown, not just summary statistics
- Some people still prefer bar charts
- Use "error bars" to summarize spread in the data

Error bars: Standard Deviation or Standard Error of the Mean?

- When showing error bars on bar charts we (currently) have two options: standard deviation, or standard error of the mean
- Which to use?
- Remember our interpretation:
  - The standard deviation is the average distance of a point in the data set from the mean
    - It's a measure of the spread in the sample
  - The standard error of the mean is the average error in using the sample mean as an approximation of the population mean
    - It's a measure of the precision of using this sample for inference about the population

Error bars: standard deviation or SEM?

- Use standard deviation when the primary goal is to describe your data set
  - Typically used to describe potential confounding variables in, for example, clinical trials
  - Show that both control and treatment groups have similar age distributions, for example
- Use standard error of the mean when the primary goal is statistical inference
  - We want to show that our sample is representative of the population
  - At least quantify the extent to which it represents the population

The most important thing is to clearly state what your error bars represent

Aside: bench experiments and populations

- Consider our TH/B6 mouse diet data.
- For example, let's just consider the Cholesterol data for TH mice fed the Chow diet
- We have six values which comprise the sample
  - But what is the population?

Aside: bench experiments and populations

- Consider our TH/B6 mouse diet data.
- For example, let's just consider the Cholesterol data for TH mice fed the Chow diet
- We have six values which comprise the sample
  - But what is the population?
- The population here is somewhat abstract:
  - It's the set of all possible values we could get from repeating this experiment
  - We assume there is some "global" Cholesterol value for TH mice on Chow
  - All individual measures are deviations from this "true" value based on the individual mouse and experimental condition

Bar Charts in R/ggplot

- Start with just TH mice:

```
library(tidyverse)
met <- read_csv("https://denvirlab.marshall.edu/BMR617-2021/data/TH-B6-metabolic.csv")
met <- separate(met, MouseID, sep="-", into=c("Strain", "Diet", "ID"))
th <- filter(met, Strain == "TH")
```

- We have to group and summarize the data:

```
th_grouped <- group_by(th, Diet)
th_summary <- summarise(th_grouped,
    MeanCholesterol=mean(Cholesterol),
    n=n(),
    sd=sd(Cholesterol),
    sem=sd/sqrt(n))
```

- And now we can plot it:

```
ggplot(th_summary, aes(x=Diet, y=MeanCholesterol)) +
  geom_bar(stat="identity", fill="#00B140") +
  geom_errorbar(aes(ymin=MeanCholesterol-sem,
                    ymax=MeanCholesterol+sem),width=0.2)
```

Can you figure out how to plot error bars with the standard deviation instead of the standard error of the mean?

Plotting grouped bar charts

- Plot all the data, change the fill of the bars by strain:

```
met_grouped <- group_by(met, Diet, Strain)
met_summary <- summarise(met_grouped,
    MeanCholesterol=mean(Cholesterol),
    n=n(),
    sd=sd(Cholesterol),
    sem=sd/sqrt(n))
ggplot(met_summary, aes(x=Diet, y=MeanCholesterol, fill=Strain)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_errorbar(aes(ymin=MeanCholesterol-sem
                    ymax=MeanCholesterol+sem),
      position=position_dodge(0.9), width=0.2)
```

More plotting techniques

- The layers in our plot are combined using +
- We can save a layer, and then add to it
  - Makes experimenting and manipulating easier

```
barchart <- ggplot(met_summary, aes(x=Diet, y=MeanCholesterol, fill=Strain)) +
  geom_bar(stat="identity", position=position_dodge())
semErrorBars <- geom_errorbar(aes(ymin=MeanCholesterol-sem,
                                  ymax=MeanCholesterol+sem),
          position=position_dodge(0.9), width=0.2)
sdErrorBars <- geom_errorbar(aes(ymin=MeanCholesterol-sd, ymax=MeanCholesterol+sd),
                             position=position_dodge(0.9), width=0.2)
barchart + semErrorBars
barchart + sdErrorBars
```

Adding labels and titles

- Use `xlab` and `ylab` to modify the labels for the axes
- Use `ggtitle` to add a main title

```
barchart + semErrorBars + ylab("Cholesterol (mg/dl)") +
       ggtitle("Cholesterol by Strain and Diet")
```