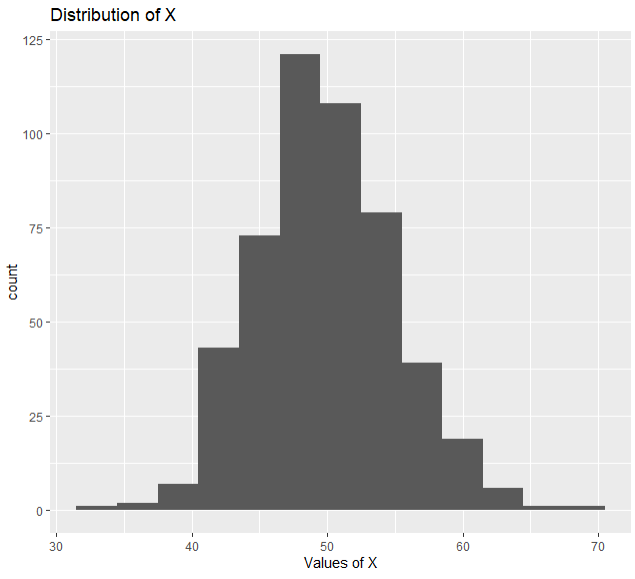
Chapter 8: The Normal Distribution

Overview: In Chapter 7, we explored ways of describing the center and the spread of distributions. We saw that, for symmetric distributions, using the mean and the standard deviation is a way to achieve this and for non-symmetric distributions, it is more appropriate to use the median and quartiles. In this chapter, we discuss a way of smoothing the appearance of our distribution and then we focus on a very special type of distribution called the Normal Distribution.

Motivating Example: Suppose we take a sample of 100 individuals from a population and observe some numerical variable for each individual in the sample. If we then create a histogram to help visualize the distribution of the variable, the histogram will appear as several rectangles. What if we wanted to represent the distribution with a smooth curve instead of rectangles?

Definition: For a given variable X measured on some population, a **density curve** for X is

Sketch of a density curve: Suppose we had the following histogram, a density curve would look like a smooth curve that follows a similar pattern as the histogram:



How to plot a density curve using ggplot2:

It should not be a surprise that we can use the word *geom_density()* in ggplot2.

Note that a density curve has a total area of 1 underneath it.

Example: Consider the $CO2$ dataset built into R and the variable of the carbon dioxide uptake rates from a variety of grass species.

1. The following ggplot command will produce a histogram and a density curve in the same plot.

```
ggplot(CO2, aes(x = uptake)) + geom_histogram(aes(x=uptake ,y= after_stat(density)),
     binwidth = ) + geom_density(aes(x=uptake,y=after_stat(density)), col = "red")
```

2. You can produce the density curve without the histogram with the following command:

```
ggplot(CO2, aes(x = uptake)) + geom_density(col = "blue", fill = "yellow")
```

3. Just like other graphs in ggplot2, you can add colour and title to the density curve.

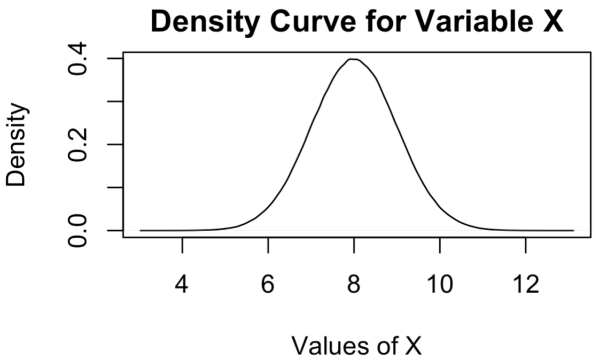
```
ggplot(CO2, aes(x = uptake)) + geom_histogram(aes(x=uptake ,y= after_stat(density)),
     binwidth = 3) + geom_density(aes(x=uptake,y=after_stat(density)), col = "red") +
     ggtitle("Distribution of Carbon Dioxide Uptake") +
     labs(x = expression(paste("carbon dioxide uptake rates (", mu, "mol/", m^2, " sec)")))
```

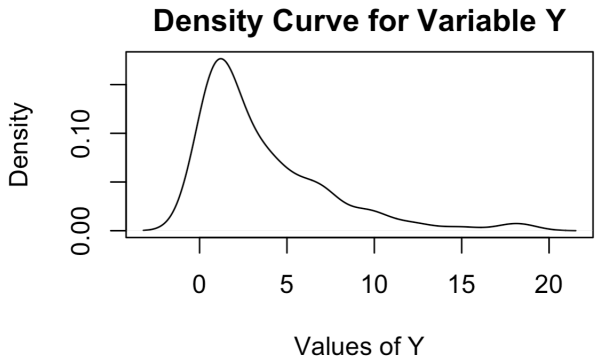
Question: How can we represent the median and mean on density curves?

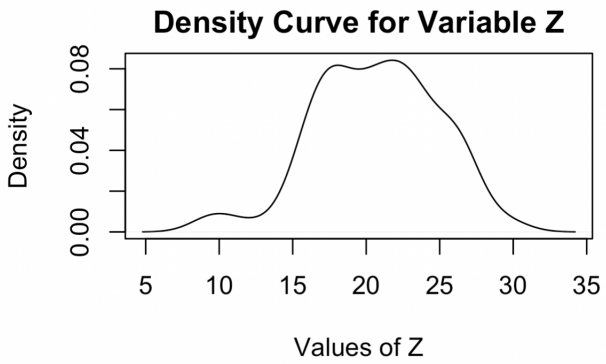
Answer: We know that median represents the value at which 50% of the observations lie below. With a density curve, this means that 50% of the area below the curve will lie to the left of the median and 50% of the area will lie to the right of the median.

For the mean, we think of the point on the density curve at which the curve would balance if it was made of solid material. This can be difficult to eyeball (which is why we normally just compute the mean using software).

Consider the estimated density curves for some variables X, Y and Z:







Summary: When a density curve is symmetric, the mean and the median are both at the same point; the center of the curve. When the curve is not symmetric, the mean gets pulled away from the median, in the direction of the tail of the distribution.

We now introduce one of the most famous distributions in statistics; the Normal Distribution.

A variable X is called **normally distributed** if the density curve is described by a complicated formula (you do not need to know this formula). This distribution is symmetric about its mean and shaped like a bell (that is, has a peak in the center and has tails which fall off quickly).

The following code will produce a normal density curve:

```
x <- c(1:50, by=0.05)
y <- dnorm(x, mean = 20, sd = 5)
normaldata <- as.data.frame(cbind(x, y))

ggplot(normaldata, aes(x = x, y = y)) + geom_line()
```

Properties of the Normal Density Curve:

- The Normal curve is completely described by

- The mean determines

- The standard deviation determines

WARNING: All normal density curves are symmetric and bell-shaped. Not all symmetric and bell-shaped curves are normal.

The 68-95-99.7 Rule: With any Normally distributed variable X, approximately:

Example: Suppose you have a variable X which is normally distributed with mean 12 and standard deviation 3. Determine the range of values that 95% of the observations should fall within.

Practice Question: Approximately what percentage of observations of X should fall between $(3, 21)$?

(A) 68% (B) 95% (C) 99.7%

Question: What if we want to determine the range of values for different percentages?

Answer: In any second year stat course you will learn how to do this by hand (using something called the standard normal distribution). In this course, I will show you two different methods to approximate these values in R.

Recall from Chapter 7, the command:

quantile(X,0.25)

Approximating the Quantiles of a Normally Distributed Variable:

We will illustrate both methods of approximating quantiles using an example.

Example: Download and save the *variable.X.Sample.csv* dataset from Brightspace. Then, read the data into R.

Step 1: Determine if the variable is approximately normal.

We notice:

Step 2 (Method 1): Use the quantile() function to determine the quantiles (often referred to as percentiles) from the sample. This will approximate the true quantiles/percentiles in the population.

Suppose, for example, we wanted to approximate the range of values such that 80% of the observations of X fall between these values:

Sketch:

Practice Question: Approximate the range of values such that 40% of the observations of X fall between these values. Round your answer to 2 decimals as needed.

(A) $(36.49, 37.40)$ (B) $(35.57, 38.26)$ (C) $(35.57, 37.40)$

Practice Question: Approximate the range of values such that 95% of the observations of X fall between these values. Round your answer to 2 decimals as needed.

(A) $(32.54, 40.82)$ (B) $(28.34, 40..82)$ (C) $(31.33, 41.50)$

Note: By the 68-95-99.7 rule, the above range should represent the mean plus or minus 2 standard deviations.

What we have computed above, is a range in which we expect 95% of all of our observations of X to fall between, not just the mean of X.

Step 2 (Method 2): Another way to estimate the quantiles/percentiles of a variable that we think is normally distributed is to use the R function qnorm() (where the q stands for quantile and norm stands for normal).

This function takes in 3 arguments: the quantile/percentile you want, the mean, and the standard deviation. It then computes the exact value of the quantile/percentile for a normal distribution with that mean and standard deviation.

Problem: We only have a sample for the variable X, we do not know its true population mean (μ) and true population standard deviation (σ). What we do know how to find is the sample mean ($\bar{x}$) and the sample standard deviation (s).

- Start by computing the sample mean and standard deviation for the variable X. It might help to name them something so that you can easily refer to them.

- Next, use qnorm() to find the desired quantiles/percentiles. For example, determine the approximate range of values such that 80% of the observations of X fall between these values:

Notice: We get a different range of values than we did using the quantile() function.

Determining the quantile/percentile of an observation: What if we want to go in the other direction? That is, what if we want to know what percentage of observations fall below a given observation?

We use the function pnorm() (where the p stands for percentile and norm still stands for normal). Again, this function takes in 3 arguments: the observation value that you wish to find the percentile for, and the mean and standard deviation of the variable.

Example: Using again the sample mean and sample standard deviation for X:

(a) Determine the percentile of the observation 41.49.

(b) Determine the percentile of the observation 36.41.

Question: What types of variables are normally distributed?

Answer:

- Some real life variables are roughly normal distributed (symmetric and bell-shaped). Examples of types of variables that are often normally distributed are:

- Certain statistics are approximately normally distributed. Examples of such statistics are:

Question: What does it mean for a statistic to be approximately normally distributed?

Answer:

Demo in R Showing the Distribution of different statistics: We will now look at a demonstration in R which shows that the distribution of the sample mean is approximately normal (when the sample is large) and that the distribution of the sample standard deviation is not normal.

The code for this demo will be posted in Brightspace.

Question: Since, for large samples, we know that the sample mean and sample proportion are normally distributed, what are the means and standard deviations for their distributions?

Answer: The mean and standard deviation of the sample mean $\bar{X}$ are:

The mean and standard deviation of the sample proportion $\hat{p}$ are:

Definition: The **standard error** of a statistic is

We often need to estimate the standard error of a statistic. The estimated standard error for the sample mean and the sample proportion are:

Practice Question: Use R, to determine the estimated standard error of the sample mean for the X variable sample from the *variable.X.Sample.csv* dataset that you downloaded earlier in the notes. Round your answer to 3 decimal places.

(A) 0.228 (B) 2.727 (C) 36.759 (D) 143

Question: Why is this useful?

Answer: We can use all these to compute something called **confidence interval**.

Motivating Example: Suppose a lightbulb manufacturer wants to determine the average lifespan of their bulbs. They take a random sample of 80 lightbulbs from their production line and then record the number of hours the lightbulbs stay on before burning out (this test takes a number of days). The resulting data produces a sample mean lifetime of 103.4 hours. How accurate is this estimate? Is the value of the sample mean close to the value of the population mean? How confident can we be that the true average lifetime of their lightbulbs lie within ± 3 hours of this time?

These are some of the questions we seek to discuss further in this part of the chapter.

Definitions: When we collect data from a sample

- An **estimate** of the population parameter

- The **margin of error** of an estimate is

- A **confidence interval**

Question: What does it mean to be $x\%$ confident that the population parameter lies in a confidence interval?

Answer: Suppose you take a sample of size 30 from the population. You can create a confidence interval using that sample. If we say that we are 90% confident that the population parameter lies in that interval, this means:

Note 1: Confidence does not equal Probability.

Note 2: The most common confidence level used in statistics is the 95% confidence level.

Practice Question: Suppose we want to estimate the proportion of vehicles in Victoria that are red. We take a random sample of 320 vehicles and determine that 62 of them are red. The margin of error (at the 95% confidence level) in this sample is approximately 0.044 (or 4.4%).

1. What is the observed value of the sample statistic/ the estimate of the population proportion?

   (A) $\hat{p} = 0.044$ (B) $\hat{p} = 62$ (C) $\hat{p} = 0.19375$ (D) $\hat{p} = 320$

2. Determine a 95% confidence interval for the population proportion p (round to 3 decimal places):

   (A) $(0.15, 0.238)$ (B) $(0.194, 0.238)$ (C) $(0.15, 0.194)$

3. How do we interpret the confidence interval?

Question: How do we compute the margin of error?

Answer: The margin of error is computed by multiplying **the standard error** (standard deviation of the statistic) with something called a **critical value**.

Estimating the Margin of Error for a Proportion:

Example: In the last example, the sample size was 320 and the sample proportion $\hat{p}$ is 0.19375.

So the estimated stardard error is:

Quesiton: How do you determine the critical value or critical number?

Answer: Using the qnorm() function and the confidence level.

Example: Compute the critical number for a 95% confidence interval:

Example: Compute the critical number for a 90% confidence interval:

Practice Question: Compute the critical number for a 84% confidence interval, round your answer to 3 decimal places.

(A) 0.994 (B) -1.405 (C) 1.405

Putting it Together: Now we know how to compute confidence intervals for the population proportion and population mean.

Example: Determine a 90% confidence interval for the population proportion of vehicles in Victoria that are red.

Example: Determine a 95% confidence interval for the population nmean of variable X.

Practice Question (Review of concepts throughout the entire course): Consider the built-in data set UCBAdmissions.

1. If we are interested in the proportion of people that apply to Berkeley University and get accepted, what is the population of interest and what is the parameter of interest?

2. Using the command ?UCBAdmissions, determine the variables in the dataset and describe what kind of variables they are.

For demonstration purpose, the raw data was recreated and saved in the file UCBA.csv

3. Create a variable in R called *totalApplicants* which contains the total number of students who applied to the university in our sample.

4. Create a variable in R called *totalAdmissions* which contains the total number of students who were admitted to the university (across all genders and departments).

5. What is the observed value of the statistic we should use to estimate the population parameter of interest?

6. What is the estimated standard error for $\hat{p}$?

7. What is the critical value for a 92% confidence interval for p?

8. What is the margin of error for our estimate?

9. Determine a 92% confidence interval for the true value of the population proportion.