Chapter 13: The Normal Distribution

Overview: In Chapter 12, 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.

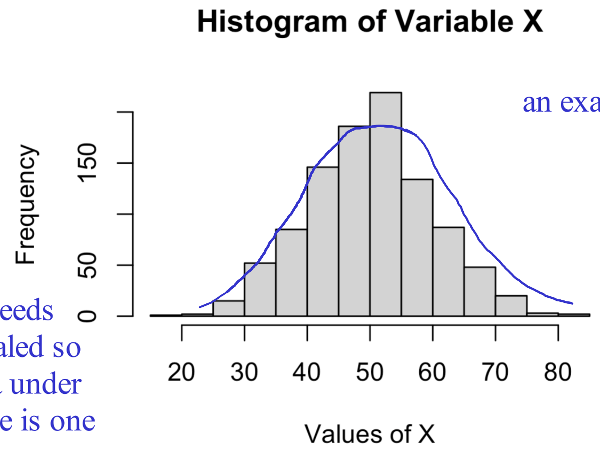
Motivating Example: Suppose we take a sample of 100 individuals from a population and observe some numerical variable for each individual in the sample. If we then create a histogram to help visualize the distribution of the variable, the histogram will appear as several rectangles. What if we wanted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

a theoretical curve which describes the distribution of variable, that is, the possible values of x and how often they are expected to occur in the population.

The curve extends across all possible values of the variable and the area under the curve equals ONE.

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:



an example of what a density curve looks like

y-axis needs to be scaled so that area under the curve is one

How to plot a Density Curve in R: In R, there is a function called density() which will try and estimate the density curve for some variable X.

Since a density curve has a total area of 1 underneath it, we need to first convert our histogram so that the area of all of the rectangles also adds up to 1. The way to do this is to type in the following code when plotting your histogram:

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

1. Create a vector X which contains the values of the uptake variable from the dataset.

```
y = co2$uptake
```

2. Plot a histogram of this variable, make sure that the sum of the areas of the rectangles add to 1.

```
hist (y, main = "Co2 uptake", xlab= "uptake Rate", prob= TRUE, col = "yellow")
```

3. Now, to add an estimate of the density curve to the histogram, use the function lines() and the function density():

```
lines (density(y))
```

4. Just like with the hist() function, you can also add colour to your density curve using the col parameter with the lines() function:

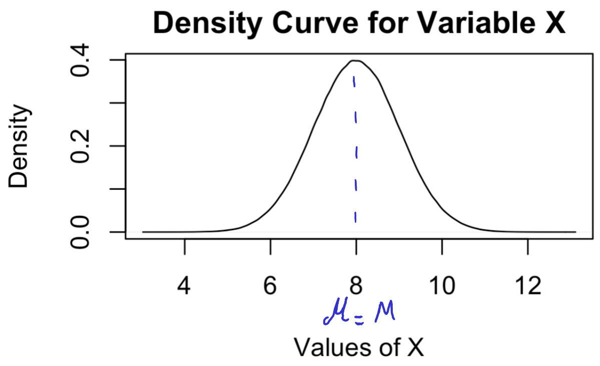
```
lines (density(y), col = "red")
```

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

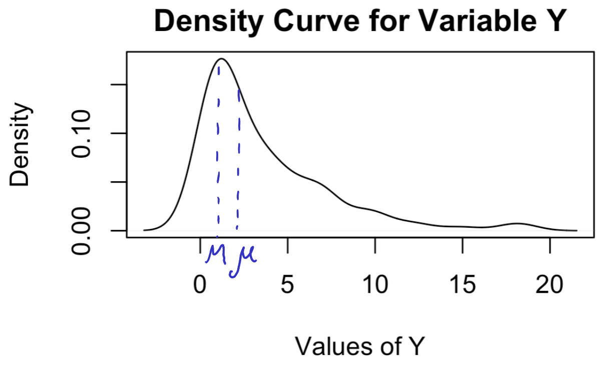
Answer: We know that median represents the value at which 50% of the observations lie below. With a density curve, this means that 50% of the area below the curve will lie to the left of the median and 50% of the area will lie to the right of the median.

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).

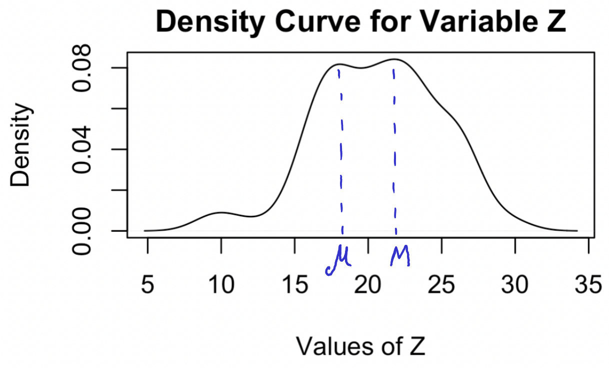
Consider the estimated density curves for some variables X, Y and Z:



Mean ≈ Median



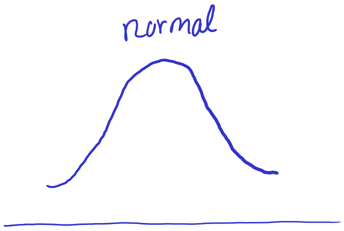
when the distribution is not systemmic, mean and median are often different

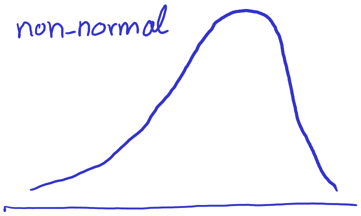


Summary: When a density curve is symmetric, the mean and the median are both at the same point; the center of the curve. When the curve is not symmetric, the mean gets pulled away from the median, in the direction of the tail of the distribution.

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 associated with the variable is symmetric and shaped like a bell (that is, has a peak in the center and has tails which fall off quickly). Normal distribution has bell-shaped density. Not all bell-shaped density is normal.





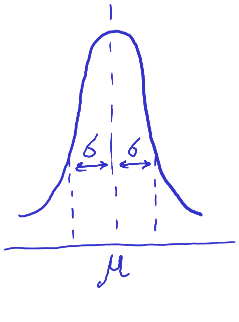
Properties of the Normal Density Curve:

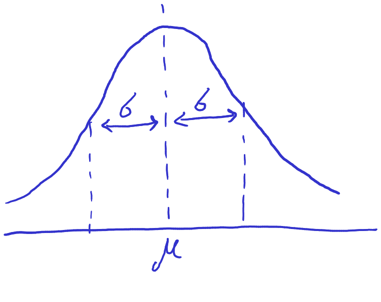
- The Normal curve is completely described by itsmean and standard deviation

  If x is normal with mean μ and standard deviation σ, then we wtite x~ $N(\mu, \sigma)$

- The mean determines the centre of distribution

- The standard deviation determines the shape of the curve (how tall or flat the curve is. The standard deviation is the distance from the mean to the change of curvature point on either side of the mean





Note: The bigger σ is, the fatter looking the bell-curve becomes.

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

- 68% of observations fall within 1 standard deviation of the mean .

- 98% of observations fall within 2 standard deviation of the mean.

- 99.7% of observations fall within 3 standard deviation of the mean.

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.

$\mu - 2\sigma = 12 - 2(3) = 6$

95% of observations of x fall between 6 and 18 or (6, 18)

$\mu + 2\sigma = 12 + 2(3) = 18$

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

(A) 68% (B) 95% [✓ (C) 99.7%]

$\mu - 3\sigma = 12 - 3(3) = 3$

$\mu + 3\sigma = 12 + 3(3) = 21$

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 12, the command:

quantile(X,0.25)

(generic command for any data vector x)

25th percentile

any percentile

x is a vector of observations

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.

```
sampleData = read.csv("variable.X.Sample.csv")
```

Step 1: Determine if the variable is approximately normal.

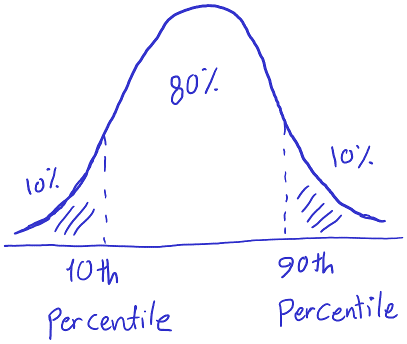
```
X=sampleData$X
hist(X, main="Distribution of X", xlab="Values of X", prob=TRUE)
lines(density(X), col="red")
```

We notice: The distribution appears to be roughly symmetric and bell-shaped so we can say the variable X is approximately normal

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: Since 80% is in the middle, the remaining 20% will be equally split between the 2 tails (10% each)

Sketch:



command:

```
quantile(X, 0.10)
quantile(X, 0.90)
```

quantile(X, 0.10) → 10 the percentile
quantile(X, 0.90) → 90 the percentile

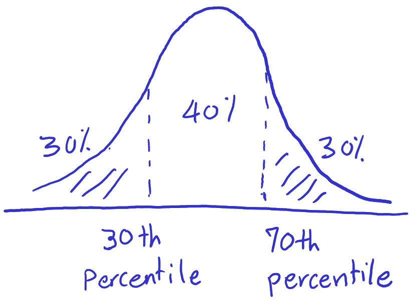
or

```
quantile(X, c(0.10, 0.90))
```

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) $(35.57, 37.40)$

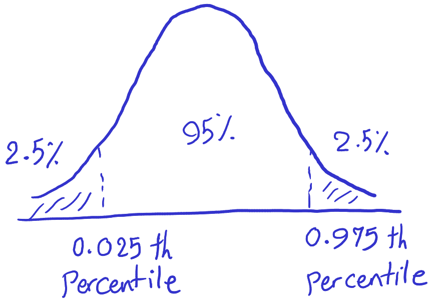
```
quantile(X, c(0.3, 0.7))
```



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) $(31.33, 41.50)$]

```
quantile(X, c(0.025, 0.975))
```



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

Question: Is the above interval the same as a 95% confidence interval?

Answer: No, a confidence interval corresponds to an estimate of a population parameter. For example, a 95% confidence interval for the mean is an interval that we are 95% confident contains the true value of the population mean.

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). μ, σ is unknown

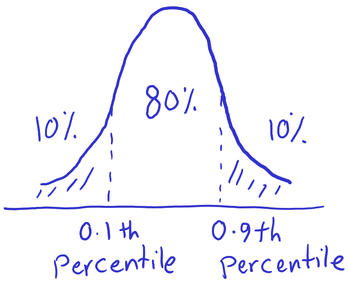
we use $\bar{x}$ and s to estimate

- 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.

```
xbar=mean(X)
s=sd(X)
```

- 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:

```
qnorm(c(0.1, 0.9), xbar, s)
```



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

quantile() function makes no assumptions about the distribution of the data set (empirical).

qnorm() function assumes data set comes from a normal distribution

Use qnorm() careful. (Justify data set comes from a normal distribution)

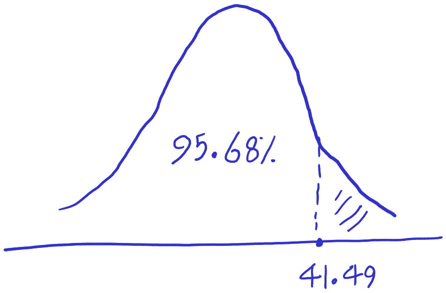
Determining the quantile/percentile of an observation: What if we want to go in the other direction? That is, what if we want to know what percentage of observations fall below a given observation?

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.

✓ pnorm (41.49, xbar, s)
✓ pnorm (41.49, mean = xbar, sd =s)
✓ pnorm (41.49, sd = s, mean xbar)
✗ pnorm (41.49, s, xbar)



(b) Determine the percentile of the observation 36.41.

pnorm(36.41, xbar, s)

Question: What types of variables are normally distributed?

Answer:

- Variables which have a symmetric, bell-shaped distribution. Examples of types of variables that are often normally distributed are:

  height, weight, student grades (sometimes)

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

  sample mean, sample proportion are approximately normal for "large" sample size.

  $\bar{x}$ and $\hat{p}$ are measurement based on samples.
  There also have distributions and they are called sampling distribution

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

Answer: If we were to take many samples, each of the same sample size n, and then compute the statistic and record the observed value of the statistic.
If we use those observed value and plot a histogram. This histogram will resemble a normal bell- shaped curve.

Note: Population distribution needs not to be normal.

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.

I will post the code for this demo in Brightspace after class.

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:

mean of $\bar{x} = \mu$

$$sd(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

the average of all possible sample means (from samples of size n) should be population mean, standard deviation of $\bar{x}$ is the true standard deviation of the population divided by $\sqrt{n}$

Definition: The **standard error** of a statistic is the standard deviation of the statistic.

The standard error (SE) of a statistic is the approximate standard deviation of a standard sample population.

standard error for $\overline{x}$ is $\frac{\sigma}{\sqrt{n}}$

standard error for $\hat{p}$ is $\sqrt{\frac{P(1-P)}{n}}$

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

estimated standard error for $\overline{X}$ is $\frac{S}{\sqrt{n}}$

estimated standard error for $\hat{p}$ is $\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$

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

```
s=sd(X)
n=length(X)
ese=s/sqrt(n)
```

Question: Why is this useful?

Answer: Recall the formula for a confidence interval is:

$$\text{estimate} \pm \underbrace{(\text{Critical value})\left(\begin{array}{l}\text{standard error or}\\ \text{estimated standard error}\end{array}\right)}_{\text{margin of error}}$$

We know how to compute the estimate (by computing the observed value of the statistic) and now we know how to estimate the standard error (for the sample mean or the sample proportion).

Quesiton: How do you determine the critical value?

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

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

qnorm(c(0.025, 0.975))

-1.959964, +1.959964

critical value is 1.959964 ≈ 1.96

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

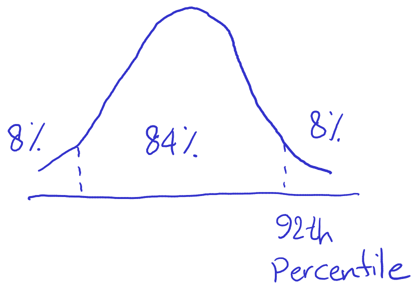
-qnorm(0.05)

abs(qnorm(0.05))

qnorm(0.95)

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]

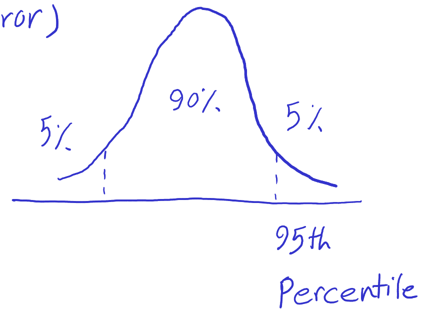


Putting it Together: Now we know how to compute confidence intervals for the population mean and population proportion.

Example: Determine a 90% confidence interval for the mean of variable X.

estimate ± (90% critical value) (estimated standard error)

$\bar{x} \pm (90\% \text{ critical value})\left(\frac{s}{\sqrt{n}}\right)$



Using R:

xbar + qnorm(0.95) * s/sqrt(n) → upper bound

xbar - qnorm(0.95) * s/sqrt(n) → lower bound

Answer: (36.38398, 37.1342)

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?

Population	Variable of Interest	Parameter
People who apply to Berely University	Whether or not an individual is accepted	Proportion of population who get accepted

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

Three variables. All categorical.

1. Admit: Admitted, Rejected
2.Geneder: Male, Female
3. Dept: A, B, C, D, E, F

3. 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).

```
totalAddmisions = sum(UCBAdmissions[1, ,])

1755
```

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

```
 totalRejections = sum(UCBAdmissions[2, ,])
or
totalRejections = sum(UCBAdmissions["Rejected", ,])
 2771
```

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

```
totalApplicants= totalAddmisions+ totalRejections
 4526
```

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

```
phat= totalAddmisions/totalApplicants

0.3877596
```

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

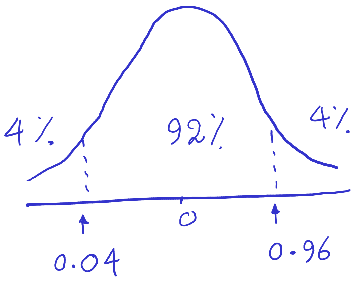
ese = sqrt(phat*(1-phat)/totalApplicants) $\rightarrow$ estimated standard error $= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

0.007242442

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

cv= qnorm(0.96) $\longrightarrow$ Critical value $\left\{ \begin{array}{l} -\text{qnorm}(0.04) \\ \text{or} \\ \text{qnorm}(0.96) \end{array} \right.$

1.750686



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

moe= cv*ese

0.01254007

10. Compare that to result of the approximate margin of error formula we learned earlier in the course.

Approximate moe feom previous formula in chapter 3: $\frac{1}{\sqrt{n}} \simeq \frac{1}{\sqrt{4526}} = 0.01486424$

Note: Don't use $\frac{1}{\sqrt{n}}$ for moe unless you are being asked.

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

Upper_bound= phat+moe

0.4004389

Lower_Bound= phat-moe

0.3750804