Topic 16 Samples vs. Populations

In previous lectures we looked at full populations of data (e.g. all NCAAD1 mens basketball games in the 2018/2019 season) and samples of data collected from underlying populations (e.g. the data for the Notre Dame Combine participants in the years 2015_2016 could be regarded as a sample from the population of Notre Dame Combine participants since the year 2000, which in turn could be regarded as a (biased) sample from all Combine participants since the year 2000).

Our data on the population of NFL players in 2014 could be regarded as the entire population data for that year or could be regarded as a (biased) sample of players from the population of all NFL players ever.

Also the performance statistics for Jeremy Lin and Stephen Curry for the 2015/2016 season could be viewed as a sample of their overall performance over the many years or as a population of performance statistics fro that season.

```
> CombineND<-read.csv("combineND.csv",header = TRUE)
> PPM<-read.csv("PPM.csv",header = TRUE)
> NFL<-read.csv("NFL.csv",header = TRUE)
```

Samples vs. Populations

We used population data in Basketball to attempt to make predictions. If we know the data distribution for a variable, we can estimate the chances that our next observation of that variable will take a particular value. If we do not have all of the population data available, we can use a sample to estimate the distribution of a variable in the underlying population. The sample distribution may or may not reflect the underlying population distribution depending on the amount of bias in the selection process for the sample, and the degree of bias that occurs from randomness. The latter tends to be minimized when we work with large samples, the former when we try to make our selection process as random as possible.

Traditionally it has often been too costly or difficult to collect all of the data from a population to figure out what the entire population distribution looks like for some variables which requires a census or election. Recently advances in technology have made the processing of some data of this kind easier and more accessible, however we are still reliant on the traditional methods of estimating the population statistics and distribution from a sample for much of the data we are interested in. **From our sample we build a model of the underlying population distribution of the variable in question**.

- The data points in our sample (hopefully selected as randomly as possible) are just a subset of all possible values of the variable in question.

- Their relative frequencies in the sample may or may not reflect the corresponding relative frequencies in the population.
- The sample mean will give us an estimate of the population mean but it is unlikely that it will coincide exactly with the population mean.

Example (Sample Statistics vs. Population Statistics) Suppose we view our data on the population of NFL players as all of the data for the population of NFL players in 2014. Suppose also we did not have all of the data available and we wanted to estimate he average height of a player in the NFL that year. We could choose a random sample of players of size 20 from the population and use the sample average to estimate the population average. We can use the `sample()` function in R to take a random sample from the population of players' heights.

```
> mean(NFL$HT) #True population mean

[1] 74.02001

> nrow(NFL) #population size

[1] 1699

> s<-sample(NFL$HT, 20) #random sample of size 20 without replacement
> s

 [1] 75 73 75 73 79 72 73 78 71 80 72 78 71 72 70 78 74 77 75 68

> mean(s) #estimate for population mean

[1] 74.2

> s1<-sample(NFL$HT, 20) #different sample gives different estimate
> mean(s1)

[1] 74.05

> s2<-sample(NFL$HT, 10) #smaller sample gives worse estimate in general
> mean(s2)

[1] 74.3

> s3<-sample(NFL$HT, 5)
> mean(s3)
```

```
[1] 74.2
```

So **before we make statistical inferences about a population variable and population statistics from a sample** we must think

1. about the different types of population distributions that the sample might be coming from and
2. the variation in samples that we might get from such a population.
3. Then we can turn the question around and say "our sample looks like ...., what are the chances the population looks like .....".

The Experiment Framework: We can consider our data collection as many repeats of the same experiment "choose an element of the population at random and record the value of the variable in question". It turns out **this framework of experiment, outcome, sample space, event and random variable adds a lot of clarity to how we think about probability and it is important to keep it in mind when conducting research with statistics.**

- An **Experiment** is an activity or phenomenon under consideration. The experiment can produce a variety of observable results called **outcomes**. The theory of probability makes most sense in the context of activities that can be repeated or phenomena that can be observed a number of times. We call each observation or repetition of the experiment a **trial**.
- A **sample space** for an experiment is the set of all possible outcomes of the experiment. Elements of the sample space are sometimes simply called **outcomes** or if there is a risk of confusion they may be called a **sample points** or **simple outcomes**.
- **An Event** E is a subset or sub collection of the simple outcomes of the sample space S. We denote the probability that an event E will occur by $P(E)$.
- Given two events A and B, we can define a new event $A \cap B$ (called the **intersection of** A **and** B) which is the event that both A and B occur. The simple outcomes in the set corresponding to $A \cap B$ is the set of all simple outcomes which are in both events.
- Two events A and B are **independent** if the fact that A has occurred has no bearing on the probability that B will also occur. In this case $P(A \cap B) = P(A)P(B)$.

- If $E_1, E_2, \ldots, E_n$ are events in a sample space, we say they are independent if the probability of any one of them does not depend in any way on what combination of the other events happen. In this case, we have $P(E_1 \cap E_2 \cap E_3 \cdots \cap E_n) = P(E_1)P(E_2)P(E_3)\ldots P(E_n)$.

- **Trials of an experiment are independent** if what happens on one trial is not influenced in any way by what happens on the other trials.

- **A Random Variable** is a rule that assigns a number to each outcome of an experiment. There may be more than one random variable associated with an experiment.

Example

Experiment: Flip a coin 4 times and observe the sequence of Heads and Tails.
Sample Space: $\{HHHH, HTHT, THTH, \ldots\}$ (16 sequences)
Example of an event: A = "the event that I see a head on the first flip" = $\{HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT\}$
Example of another event: B = "the event that I see a head on the last flip" = $\{HHHH, THHH, HTHH, TTHH, HHTH, THTH, HTTH, TTTH\}$
$A \cap B = \{HHHH, HHTH, HTHH, HTTH\}$
Random Variable associated to this experiment: X = the number of H's observed. (if the outcome is HHTH, the value of X observed is 3)
Random Variable associated to this experiment: Y = the number of T's observed.
Both variables X and Y are discrete.

Example

Experiment: Roll a pair of fair six sided dice and observe the numbers on the uppermost face.
Sample Space:

$$\begin{array}{cccccc}
\{(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\
(2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\
(3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\
(4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\
(5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\
(6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\}
\end{array}$$

Event: A = the event that I see a six on the first roll = $\{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$.
Event: B = the event that I see a six on the second roll = $\{(1,6), (2,6), (3,6), (4,6), (5,6), (6,6)\}$.
$A \cap B$**:** $\{(6,6)\}$.

Random Variable: X = sum of the numbers on the uppermost faces. (discrete random variable).

Example

Experiment: Record the time for the FortyYD for a participant in the NFL Combine. **Sample Space:** $(0, \infty)$.
Event: A = event that the time recorded is less than 4.5 seconds. $A = (0, 4.5)$.
Random Variable The observed time itself; (continuous random variable).

Example:
Experiment: Choose the data of one NFL player at random from the data stored in the data set `NFL`.

```
> N<-nrow(NFL)
> N

[1] 1699

> v<-1:N
> s1<-sample(v, 1)
> s1

[1] 853

> NFL[s1,]

      Team X.                   Name First.Name     Last.Name POS HT  WT AGE
853 Saints 33 Jean-Baptiste, Stanley    Stanley Jean-Baptiste  CB 75 218  26
       BDAY EXP  College AVG.Annual.Salary Birth.City Birth.State  Race
853 4/12/90   0 Nebraska            909183  Miami, FL     Florida Black
                     HS   HS.City HS.State Pro.Bowler Champ X..Of.Teams Heisman
853 Central (Miami, FL) Miami, FL  Florida          0    NO           1      NO
    X
853
```

Sample Space: The sample space here is all possible rows in the data set (or equivalently all possible players in the data set)
Random Variables: There are a number of random variables associated to the outcome of this experiment (all of which are recorded in the data set); `HT`, `WT`, `AGE`,`Salary`,`EXP`.
Example of an Event: A = the event that we choose a player with height above 75 inches.

Example of an Event: B = the event that we choose a player with age less than 30.
Example of an Event: $A \cap B$ = the event that we choose a player with age less than 30 and height greater than 75 inches.

Discrete Random Variables and Their Probability Distributions

When we assign probabilities to the outcomes in our sample space, we want those probabilities to reflect the relative frequency with which the outcomes would occur if the experiment was repeated many times. Sometimes, we use logic to assign such probabilities and sometimes we use experience or repeated experimentation to observe relative frequencies of outcomes and assign propbailities accordingly. The **Law of Large Numbers** says that both of these methods give the same answer if we repeat the experiment many many times. However one should also be aware of The chaos of small numbers, namely in a small number of trials of an experiment, the relative frequencies of outcomes in a sample space may not resemble their logical probabilities.

Example(Law of Large Numbers in action:) If our experiment is to roll a fair six sided die and observe the number on the uppermost face, we can derive by logic that the probability of a ”6” is 1/6. If we repeat the experiment 10 times, then the relative frequency of a ”6” in the data may not be close to 1/6, but if we repeat the experiment 1000 times, the relative frequency of a ”6” in the data should be close to $1/6 \approx 0.1666667$. We can simulate many rolls of a die with the `sample()` function in R:

```
> k<-1:6 #vector of outcomes
> p<-c(1/6,1/6,1/6, 1/6,1/6,1/6) #vector of probabilities
> s<-sample(k,size=10,prob=p, replace=TRUE) #10 rolls
> s

 [1] 4 3 2 5 5 3 1 5 4 2

> table(s)

s
1 2 3 4 5
1 2 2 2 3

> table(s)/10 #estimates of probabilities using relative frequencies

s
  1   2   3   4   5
0.1 0.2 0.2 0.2 0.3
```

```
> s1<-sample(k,size=1000,prob=p, replace=TRUE) #1000 rolls
> table(s1)

s1
  1   2   3   4   5   6
177 168 157 175 174 149

> table(s1)/1000 #estimates of probabilities using relative frequencies

s1
    1     2     3     4     5     6
0.177 0.168 0.157 0.175 0.174 0.149

> s2<-sample(k,size=10000,prob=p, replace=TRUE) #10,000 rolls
> table(s2)

s2
   1    2    3    4    5    6
1724 1656 1653 1691 1625 1651

> table(s2)/10000 #estimates of probabilities using relative frequencies

s2
     1      2      3      4      5      6
0.1724 0.1656 0.1653 0.1691 0.1625 0.1651
```

Assigning Probabilities to Outcomes in The Sample Space.

When assigning probabilities to discrete outcomes in a sample space, we adhere to the following rules in keeping with the intention that **the probabilities assigned should reflect the relative frequency of each outcome in many repeated trials of the experiment.** If the outcomes are listed as $\{o_1, o_2, o_3, \dots\}$ and the respective probabilities as $p_1, p_2, p_3, \dots$, then we make the following rules for $p_1, p_2, p_3, \dots$:

- $0 \leq p_i \leq 1$ for all i,
- $p_1 + p_2 + \cdots = 1$.

Equally Likely Outcomes If we can use logic to determine that all outcomes in a finite sample space are equally likely, then combining our rules and the fact that all outcomes are equally likely, we get the probability of each outcome is 1/(number outcomes in S.S.).

The probability of an event A, denoted $P(A)$, is the sum of the probabilities of the outcomes in the event. In an equally likely sample space the probability of an event A is the number of outcomes in A divided by the number of outcomes in the sample space.

$$P(A) = \frac{\text{\# outcomes in A}}{\text{\# outcomes in S.S.}}$$

Example If I roll a die twice, I get 36 equally likely outcomes.

$$\text{Sample Space:} = \begin{matrix} \{(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\} \end{matrix}$$

Consider the event E = "The sum of the numbers on the uppermost faces is 7". The number of outcomes in E is 6, therefore $P(E) = 6/36 = 1/6$.
Consider also the two events A = "get a six on the first roll", B = "get a six on the second roll". A and B are independent events since the probability that we get a six on the second roll is not influenced in any way by what happens on the first roll. The event that we get a six on both rolls is $A \cap B = \{(6,6)\}$. We could caluclate the probability of $A \cap B$ with our formula for the probabilities of events in equally likely sample space but because they are independent events we can also multiply the probabilities of A and B to get $P(6,6) = 1/36$.

Example:
Experiment: Choose the data of one NFL player at random from the data stored in the data set `NFL`.

```
> N<-nrow(NFL)
> N

[1] 1699

> v<-1:N
> s1<-sample(v, 1)
> s1

[1] 772

> NFL[s1,]

     Team X.         Name First.Name Last.Name POS HT  WT AGE    BDAY EXP
772 Bills 92 Wynn, Jarius     Jarius      Wynn  DE 75 285  30 8/29/86   6
    College AVG.Annual.Salary  Birth.City Birth.State  Race          HS
```

```
772 Georgia            795000 Augusta, GA     Georgia Black Lincoln Co.
           HS.City HS.State Pro.Bowler Champ X..Of.Teams Heisman X
772 Lincolnton, GA  Georgia         0     1           5      NO
```

Sample Space: The sample space here is all possible rows in the data set (or equivalently all possible players in the data set)
Since each line of data is a an outcome in our sample space the probability that we choose a particular line of data (say that of Tom Brady) with our random sampling method is $1/N$, where $N =$ the number of rows of data.

Probability of an Event: A = the event that we choose a player with height above 75 inches.

$$P(A) = \frac{\text{\# outcomes in A}}{\text{\# outcomes in S.S.}} = \frac{\text{\# players with H} > 75}{\text{\# rows of data}}$$

```
> P_A<-length(NFL$HT[NFL$HT>75])/nrow(NFL)
> P_A

[1] 0.3090053
```

Probability distribution of a random variable:
The probability distribution of a discrete random variable is just a list of the probabilities we assign to each possible value of the random variable. It can be presented in a number of ways. For a finite discrete random variable, we can present the distribution as a table consisting of a list of the possible values of the variable X alongside their probabilities or as a bar graph.

Example An experiment consists of flipping a coin 4 times and observing the sequence of heads and tails. The random variable X is the number of heads in the observed sequence. Use the equally likely sample space

$$S.S. = \{HHHH,\ HHHT,\ HHTH,\ HHTT,\ HTHH,\ HTHT,\ HTTH,\ HTTT,$$
$$THHH,\ THHT,\ THTH,\ THTT,\ TTHH,\ TTHT,\ TTTH,\ TTTT\}.$$

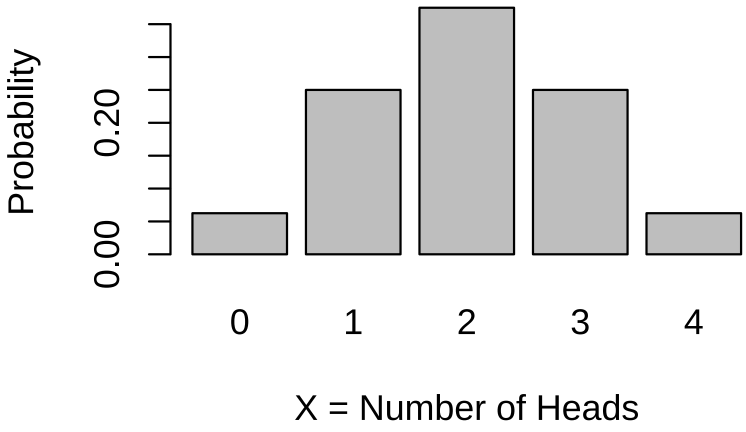
to fill in probabilities for each possible value of X in the table below.

X (# Hs)	**P(X)**
0	
1	
2	
3	
4	

We can draw a barplot showing this distribution with our `barplot()` command:

```
> x<-c(0,1,2,3,4)
> px<-c(1/16,4/16,6/16,4/16,1/16)
> names(px)<-x

> barplot(px, xlab="X = Number of Heads",ylab="Probability" )
```



Example(using relative frequencies to assign probabilities) If we consider the following experiment:
Experiment: Pick an NFL player from the NFL data set at random and record their height, H.
We can figure out the frequencies of players with each height in the data with

`table()` command and use the relative frequencies as probabilities for the possible values of H. We could also use this sample as empirical evidence to estimate the distribution of H for all NFL players (not just those playing in 2014).

```
> nrow(NFL)
```

```
[1] 1699
```

```
> HT.freq<-table(NFL$HT)
> HT.freq
```

```
 66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81
  2   3  24  47 111 126 179 208 223 251 213 164  92  38  15   3
```

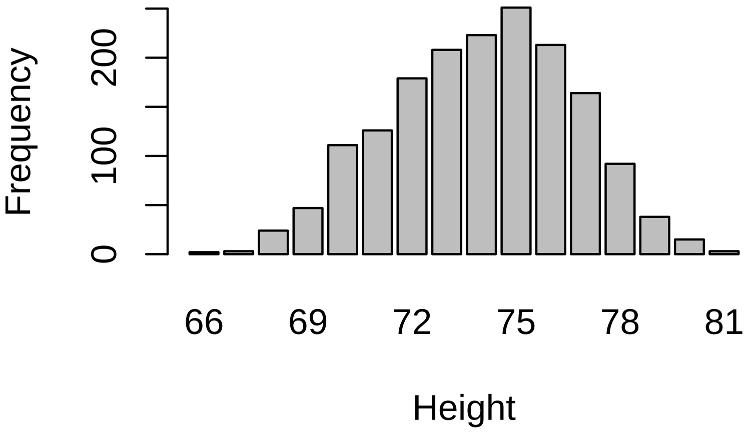
The table below shows the probability distribution of our random variable H.

H	66	67	68	69	70	71	72	73
P(H)	2/1699	3/1699	24/1699	47/1699	111/1699	126/1699	179/1699	208/1699

H	74	75	76	77	78	79	80	81
P(H)	223/1699	251/1699	213/1699	164/1699	92/1699	38/1699	15/1699	3/1699

We can also present the distribution of this random variable as a barplot:

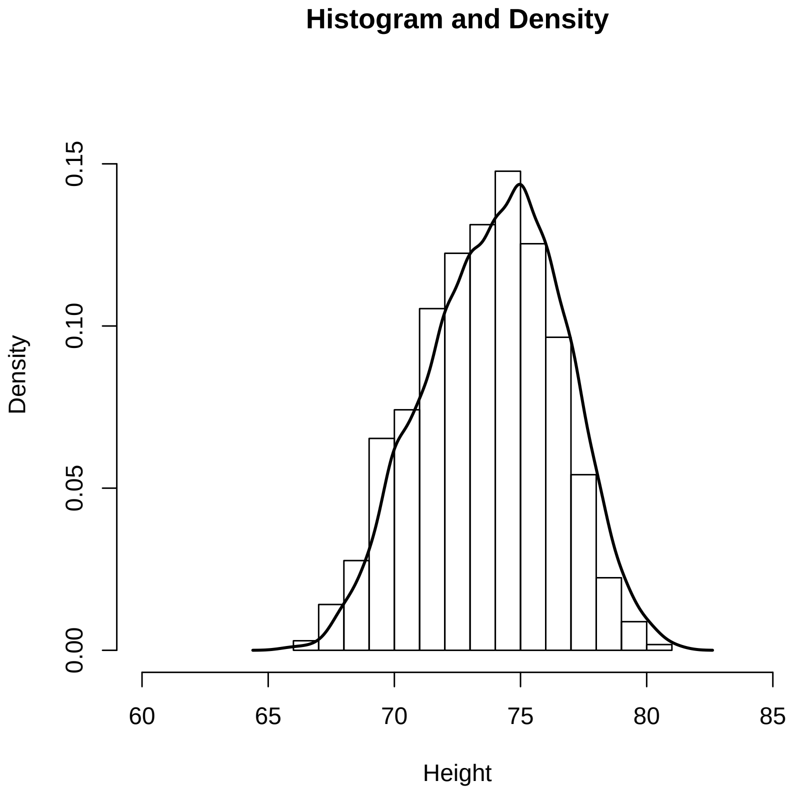
```
> barplot(HT.freq, xlab="Height",ylab="Frequency", )
```



Continuous Random Variables and their probability distributions: The probability distribution of a continuous random variable cannot be represented in a table since the possible values of the variable cannot be separated. The distribution is represented using the graphical method as a continuous curve and is called a **probability density function**. Probabilities are calculated for intervals instead of particular values. The probability that the value of a random variable will fall in the interval $[a, b]$, denoted $P(a \leq X \leq b)$ is given by the area under the probability density function above that interval. The area under the entire probability density curve is 1. We have already created density functions from our census data for the NFL players in 2014 using R.

Since Height is normally considered to be a continuous random variable, it would be more appropriate to represent the distribution of heights with a density curve.

```
> hist(NFL$HT, main="Histogram and Density",xlab="Height",
+       probability=TRUE, xlim=range(c(60,85)),ylim=range(c(0,0.17)) )
> lines(density(NFL$HT,na.rm=TRUE),lwd=2)
```

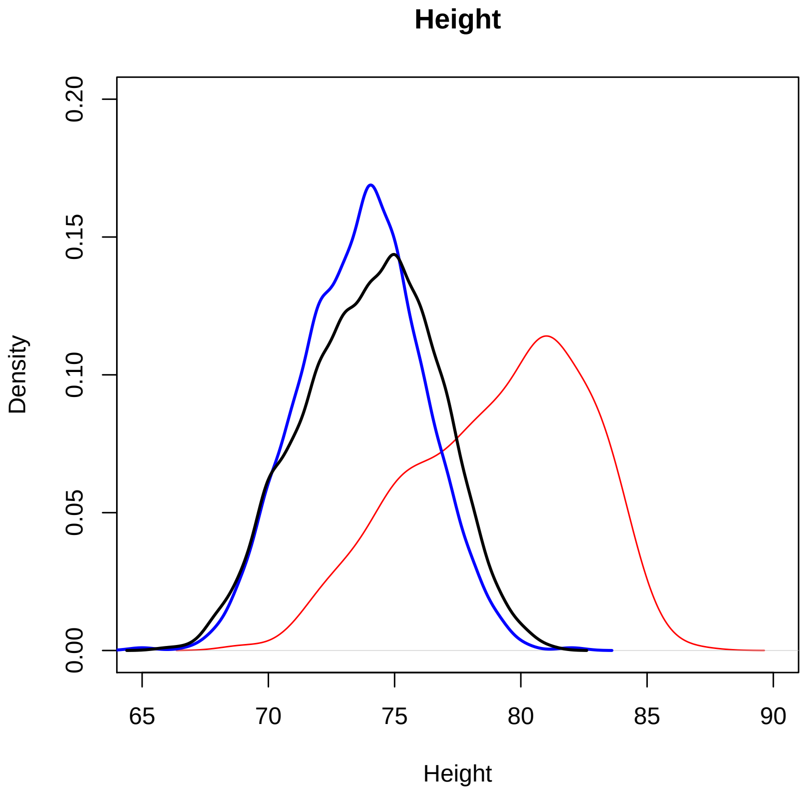


R finds the density (a continuous curve with area beneath it = 1) which best fits our data. To find the actual area underneath the curve over an interval, we would need some methods from calculus. Most of the densities that we deal with, will be well documented densities, and we can calculate the probabilitioes required using a computer or calculator.

Below we create plots of the densities for the heights of players in the NFL, MLB and the NBA. You should load the MLB and NBA data files to replicate this.

```
> NBA<-read.csv("NBA.csv",header = TRUE)
> MLB<-read.csv("MLB.csv",header = TRUE)

> plot(density(NBA$Ht,na.rm=TRUE),xlim=range(c(65,90)),
+       ylim=range(c(0,0.2)),main="Height",xlab="Height",col="red" )
> lines(density(MLB$Height,na.rm=TRUE),lwd=2,col="blue")
> lines(density(NFL$HT,na.rm=TRUE),lwd=2)
```

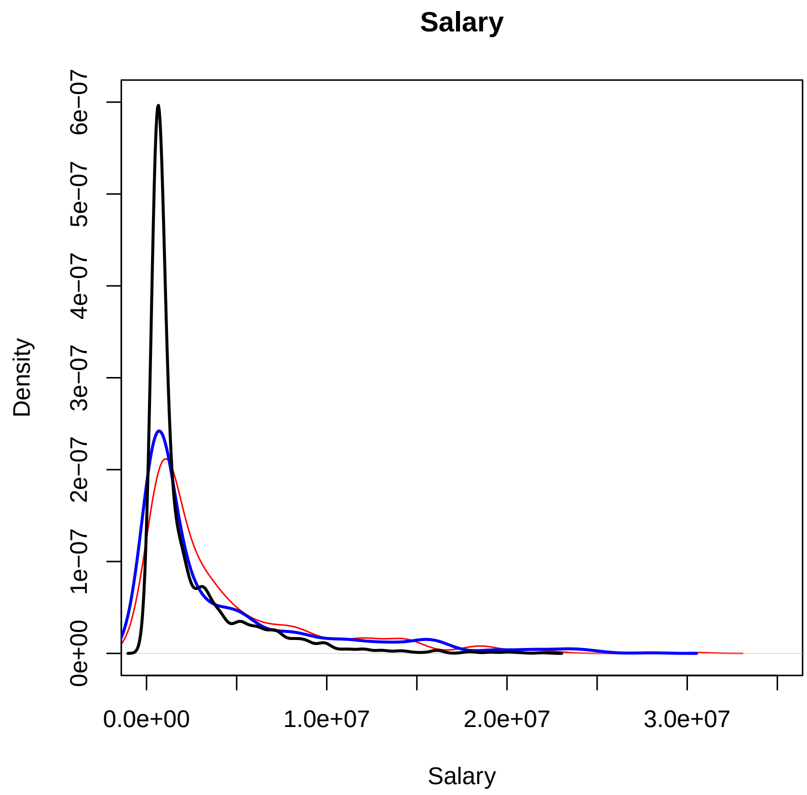


If M denotes the height of a MLB player chosen at random from the data, the probability that their height is between 75 and 80 inches, $P(75 \leq M \leq 80)$ is the area under the blue curve between 75 and 78. This is approximately 0.371. If B

denotes the heights of NBA players then $P(75 \leq B \leq 80)$ is the area under the red curve over that interval which is larger (in fact it is approximately 0.477).

The densities for salaries of the three groups are shown below.

```
> plot(density(NBA$Salary,na.rm=TRUE),xlim=range(c(0,35000000)),
+       ylim=range(c(0,0.0000006)),main="Salary",xlab="Salary",col="red" )
> lines(density(MLB$SALARY,na.rm=TRUE),lwd=2,col="blue")
> lines(density(NFL$AVG.Annual.Salary,na.rm=TRUE),lwd=2)
```



Mean and Standard Deviation of a Random Variable We have used the mean $\bar{x}$ and standard deviation s to measure center and spread of a data set. The population mean μ and population standard deviation σ are used for the same purpose, but have slightly different methods of calculation and hence the different notation.

Expected value of a Random Variable As with our data sets, the expected value of a random variable is the balance point of the graphical representation of the distribution of the variable. It is a weighted average. To calculate the expected value of a finite discrete random variable X, we can draw on our knowledge of the relative frequency version of the formula for the mean of a sample. If we replace the relative frequency by probabilities we get the formula for the expected value of the random variable X shown below.

If X is a random variable with a finite number of possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p_1, p_2, \ldots, p_n$, the **expected value of** X, denoted by $E(X)$ or μ, is

$$\boxed{\mu = E(X) = x_1p_1 + x_2p_2 + \cdots + x_np_n.}$$

Outcomes	**Probability**	**Out.** × **Prob.**
X	**P(X)**	**XP(X)**
x_1	p_1	x_1p_1
x_2	p_2	x_2p_2
$\vdots$	$\vdots$	$\vdots$
x_n	p_n	x_np_n
		Sum $= E(X) = \mu$

Note: If we run a **large number of trials of the experiment**, say N, and observe the value of the random variable X in each, $x_1, x_2, x_3, \ldots, x_N$, we should have that $\mu = E(X) \approx \frac{x_1+x_2+x_2+\cdots+x_N}{N} = \bar{x}$ or

$$\boxed{E(X)N \approx x_1 + x_2 + x_2 + \cdots + x_N.}$$

Example An experiment consists of flipping a coin 4 times and observing the sequence of heads and tails. The random variable X is the number of heads in the sequence. Find $\mu = E(X)$.

X	**P(X)**
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16

To calculate the **expected value of a continuous random variable** X, also denoted by μ or $E(X)$, we need to use integration (from calculus). However it **has the same geometric meaning as its sample counterpart as the balance point of the population distribution** and thus we can get an estimate by looking at the density function. For the most part we consider it as a fixed but unknown entity and explore how it relates to the estimate we get from the sample; namely the sample mean $\bar{x}$.

Population variance and standard deviation The population variance for a random variable X is defined as

$$\sigma^2(X) = VAR(X) = E((X-\mu)^2) = E(X^2) - \mu^2.$$

The population standard deviation σ is the square root of the population variance, $\sigma = \sqrt{\sigma^2}$. For a continuous random variable, we need integration to calculate its value, however we will for the most part be estimating the population variance for these variables with our sample variance.

If X is a random variable with values $x_1, x_2, \ldots, x_n$, corresponding probabilities $p_1, p_2, \ldots, p_n$, and expected value $\mu = E(X)$, then

$$\boxed{\textbf{Variance} = \sigma^2(X) = p_1(x_1 - \mu)^2 + p_2(x_2 - \mu)^2 + \cdots + p_n(x_n - \mu)^2}$$

and

$$\boxed{\textbf{Standard Deviation} = \sigma(X) = \sqrt{\sigma^2(X)}}.$$

$\mathbf{x_i}$	$\mathbf{p_i}$	$\mathbf{x_i p_i}$	$(\mathbf{x_i} - \mu)$	$(\mathbf{x_i} - \mu)^\mathbf{2}$	$\mathbf{p_i}(\mathbf{x_i} - \mu)^\mathbf{2}$
x_1	p_1	$x_1 p_1$	$(x_1 - \mu)$	$(x_1 - \mu)^2$	$p_1(x_1 - \mu)^2$
x_2	p_2	$x_2 p_2$	$(x_2 - \mu)$	$(x_2 - \mu)^2$	$p_2(x_2 - \mu)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	p_n	$x_n p_n$	$(x_n - \mu)$	$(x_n - \mu)^2$	$p_n(x_n - \mu)^2$
		Sum $= \mu$			**Sum** $= \sigma^2(X)$

One may also calculate $E(X)$ as $E(X^2) - \mu^2$.

Example An experiment consists of flipping a coin 4 times and observing the sequence of heads and tails. The random variable X is the number of heads in the sequence. Find $\sigma(X)$.

X	**P(X)**
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16