Random Variables

In the last section we talked about the length of the longest run of heads in the data if we flipped a coin 10 times.

- The length of the longest run of heads varied from trial to trial of the experiment.
- In this section, we will introduce some notation for such variables associated to experiments.
- We will talk about their distributions and measures of their expected value and their variance.

Random Variables

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. e.g. If I roll a pair of dice, one red and one green, and record the pair of numbers on the uppermost faces. Let X be the sum of the numbers on the uppermost faces.

- The value of X varies from trial to trial.
- Each outcome has a corresponding value of X.
- For example if the outcome is $(1, 1)$, the corresponding value of X is $1 + 1 = 2$.
- If the outcome is $(4, 5)$, the corresponding value of X is 9.

Example: Random Variables

If I roll a pair of dice, one red and one green, and record the pair of numbers on the uppermost faces. Let X be the sum of the numbers on the uppermost faces. What are the possible values of X?

- The outcomes of the experiment are given by:

$$\{(1,1) \quad (1,2) \quad (1,3) \quad (1,4) \quad (1,5) \quad (1,6)$$
$$(2,1) \quad (2,2) \quad (2,3) \quad (2,4) \quad (2,5) \quad (2,6)$$
$$(3,1) \quad (3,2) \quad (3,3) \quad (3,4) \quad (3,5) \quad (3,6)$$
$$(4,1) \quad (4,2) \quad (4,3) \quad (4,4) \quad (4,5) \quad (4,6)$$
$$(5,1) \quad (5,2) \quad (5,3) \quad (5,4) \quad (5,5) \quad (5,6)$$
$$(6,1) \quad (6,2) \quad (6,3) \quad (6,4) \quad (6,5) \quad (6,6)\}$$

- The possible values of X are $\{2,3,4,5,6,7,8,9,10,11,12\}$.

Example: Random Variables

If I flip a coin 20 times and let X be the number of runs (total number of runs of heads and tails) in the data, then X is a random variable.
If the outcome is

$$HHTTTHTTTTHHHTHHHHHH,$$

what is the value of X?

- we show runs of tails in red and runs of Heads in white:

$$HH\,TTT\,H\,TTTT\,HHH\,T\,HHHHHH$$

- We have 7 runs in this outcome, so the value of X corresponding to this outcome is 7.

More than One Random Variable

We can have more than one random variable associated to an experiment.
Example: An experiment consists of flipping a coin 4 times and observing the result ion sequence of heads and tails. The outcomes in the sample space are $\{HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT\}$.

- (a) Let Z denote the number of runs observed. What are the possible values of Z?
- The possible values of Z are $\{1, 2, 3, 4\}$.
- (b) We could also define another random variable associated to this experiment. Let X denote the number of heads observed. What are the possible values of X?
- The possible values of X are $\{0, 1, 2, 3, 4\}$.

Discrete Random Variables

For some random variables, the possible values of the variable can be separated and listed in either a finite list or an infinite list. These variables are called discrete random variables. Some examples are shown below:

Experiment	**R. Var. , X**	**Poss. values of X**
Roll a pair of six-sided dice	Sum of the numbers	$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
Toss a coin 5 times	Number of tails	$\{0, 1, 2, 3, 4, 5\}$
Flip a coin until you get a tail	The number of coin flips	$\{1, 2, 3, \ldots, \}$
Flip a coin 50 times	Longest run of heads	$\{0, 1, 2, 3, \ldots, 50\}$

Continuous Random Variables

On the other hand, **a continuous random variable** can assume any value in some interval. Some examples are:

Experiment	**Random Variable, X**
Choose an NFL Quarterback at random	Height
Choose an NCAA Shot Putter at random	Arm Length
Choose a Track and Field athlete at random	Their best time for 100 meters

Probability Distributions for Discrete Random Variables

For a discrete random variable with finitely many possible values, we can calculate the probability that a particular value of the random variable will be observed by adding the probabilities of the outcomes of our experiment associated to that value of the random variable (assuming that we know those probabilities). This assignment of probabilities to each possible value of X is called the probability distribution of X.

$$\{(1,1)\ (1,2)\ (1,3)\ (1,4)\ (1,5)\ (1,6)$$
$$(2,1)\ (2,2)\ (2,3)\ (2,4)\ (2,5)\ (2,6)$$
$$(3,1)\ (3,2)\ (3,3)\ (3,4)\ (3,5)\ (3,6)$$
$$(4,1)\ (4,2)\ (4,3)\ (4,4)\ (4,5)\ (4,6)$$
$$(5,1)\ (5,2)\ (5,3)\ (5,4)\ (5,5)\ (5,6)$$
$$(6,1)\ (6,2)\ (6,3)\ (6,4)\ (6,5)\ (6,6)\}$$

X	P(X)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Probability Distributions for Discrete Random Variables

If a discrete random variables has possible values $x_1, x_2, x_3, \ldots, x_k$, then a **probability distribution** $P(X)$ is a rule that assigns a probability $P(x_i)$ to each value x_i. More specifically,

$$0 \leq P(x_i) \leq 1 \text{ for each } x_i.$$

- and $P(x_1) + P(x_2) + \cdots + P(x_k) = 1$.

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ HHHT,\ HHTH,\ HHTT,\ HTHH,\ HTHT,\ HTTH,\ HTTT,$
  $THHH,\ THHT,\ THTH,\ THTT,\ TTHH,\ TTHT,\ TTTH,\ TTTT\}$.

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

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = # Runs	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ HHHT,\ HHTH,\ HHTT,\ HTHH,\ HTHT,\ HTTH,\ HTTT,$
  $THHH,\ THHT,\ THTH,\ THTT,\ TTHH,\ TTHT,\ TTTH,\ \underline{TTTT}\}$.

X= # Heads	**P(X)**
$\underline{0}$	$\underline{1/16}$
1	4/16
2	6/16
3	4/16
4	1/16

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = # Runs	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ HHHT,\ HHTH,\ HHTT,\ HTHH,\ HTHT,\ HTTH,\ \underline{HTTT},$
  $THHH,\ THHT,\ THTH,\ \underline{THTT},\ TTHH,\ \underline{TTHT},\ \underline{TTTH},\ TTTT\}$.

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

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = # Runs	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ HHHT,\ HHTH,\ HHTT,\ HTHH,\ \underline{HTHT},\ \underline{HTTH},\ \underline{HTTT},$
  $THHH,\ THHT,\ \underline{THTH},\ \underline{THTT},\ TTHH,\ \underline{TTHT},\ \underline{TTTH},\ TTTT\}$.

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

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = # Runs	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ HHHT,\ \underline{HHTH},\ \underline{HHTT},\ \underline{HTHH},\ HTHT,\ HTTH,\ HTTT,$
  $THHH,\ \underline{THHT},\ THTH,\ THTT,\ \underline{TTHH},\ TTHT,\ TTTH,\ TTTT\}$.

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

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = **# Runs**	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ \underline{HHHT},\ HHTH,\ \underline{HHTT},\ HTHH,\ HTHT,\ HTTH,\ \underline{HTTT},$
  $\underline{THHH},\ THHT,\ THTH,\ THTT,\ \underline{TTHH},\ TTHT,\ \underline{TTTH},\ TTTT\}$.

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

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = # Runs	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

Example: Probability Distributions for Discrete R. V.s

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. The random variable Y is the length of the longest run of heads in the sequence and the random variable Z is the total number of runs in the sequence (of both H's and T's). Find the probability distributions for X, Y and Z.

- The equally likely sample space is:
  $\{HHHH,\ HHHT,\ \underline{HHTH},\ HHTT,\ \underline{HTHH},\ HTHT,\ \underline{HTTH},\ HTTT,$
  $THHH,\ \underline{THHT},\ THTH,\ \underline{THTT},\ TTHH,\ \underline{TTHT},\ TTTH,\ TTTT\}$.

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

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

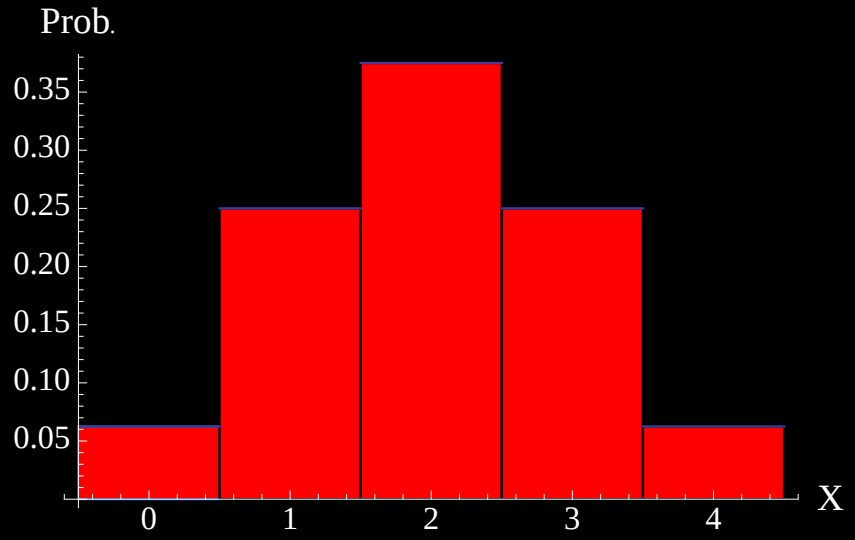
-

Z = # Runs	**P(Z)**
1	2/16
2	6/16
<u>3</u>	<u>6/16</u>
4	2/16

Graphical Representation

We can also represent a probability distribution for a discrete random variable with finitely many possible values **graphically** by constructing a bar graph.

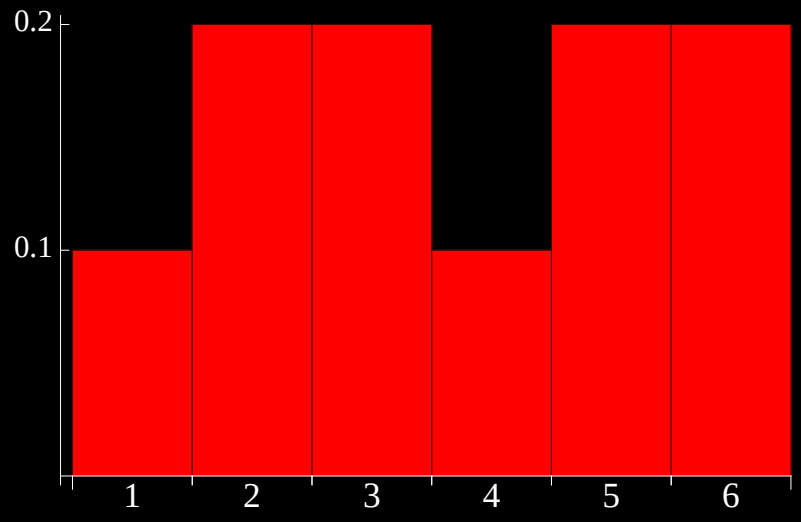
- We form a category for each value of the random variable centered at the value which does not contain any other possible value of the random variable.
- We make each category of equal width and above each category we draw a bar with height equal to the probability of the corresponding value.
- If the possible values of the random variable are integers, we can give each bar a base of width 1.
- Example: If we flip a coin 4 times and let X denote the number of heads in the observed sequence. The following is a graphical representation of the probability distribution of X.



Using The Graphical Representation

By Making all bars of equal width, we ensure that the graph adheres to the **area principle** in that the probability that any set of values will occur is equal to the area of the bars above those values. The total area of the distribution is 1.

- **Example** The following is a probability distribution histogram for a random variable X.

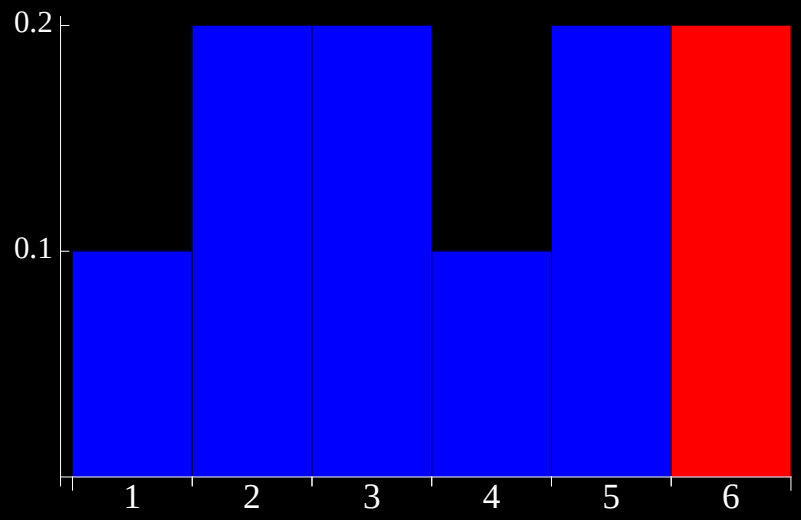


What is $P(X \leq 5)$?

Using The Graphical Representation

By Making all bars of equal width, we ensure that the graph adheres to the **area principle** in that the probability that any set of values will occur is equal to the area of the bars above those values. The total area of the distribution is 1.

- **Example** The following is a probability distribution histogram for a random variable X.



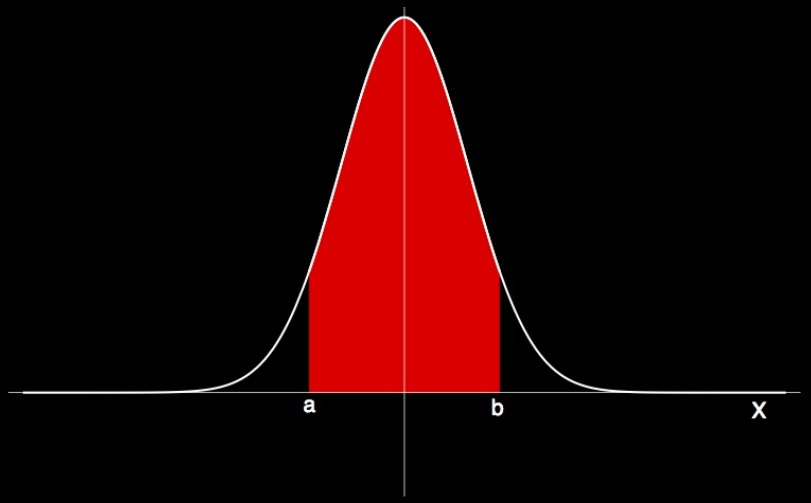
What is $P(X \leq 5)$?

- $P(X \leq 5)$ is equal to the sum of the areas of the blue rectangles shown above, which is $0.1 + 0.2 + 0.2 + 0.1 + 0.2 = 0.8$
- Notice that since the total area of the distribution is 1, we can also calculate $P(X \leq 5)$ as $1 - P(X = 6) = 1 - 0.2$.

Calculating Probability for Continuous Variables

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 total area under the entire probability density curve is 1.

Calculating Probability for Continuous Variables

The picture below taken form the website
http://datascopeanalytics.com/what-we-think/2009/11/23/height-differences-among
It shows three probability density functions for the height (in inches) of NFL, NBA and NHL players respectively(Compiled in November 2009).

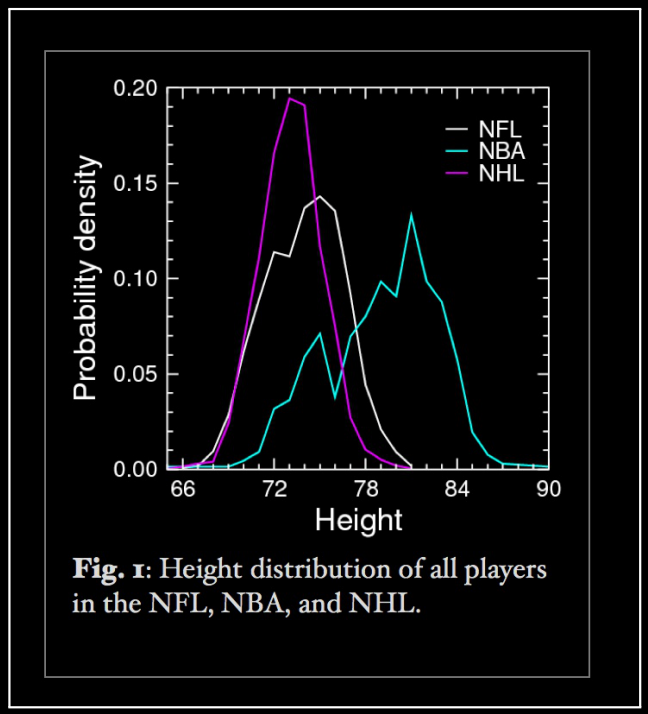


Fig. 1: Height distribution of all players in the NFL, NBA, and NHL.

Estimate the probability that an NFL player chosen at random from the group will have a height greater than 77 inches.

Calculating Probability for Continuous Variables

The picture below taken form the website
http://datascopeanalytics.com/what-we-think/2009/11/23/height-differences-among
It shows three probability density functions for the height (in inches) of NFL, NBA and NHL players respectively(Compiled in November 2009).

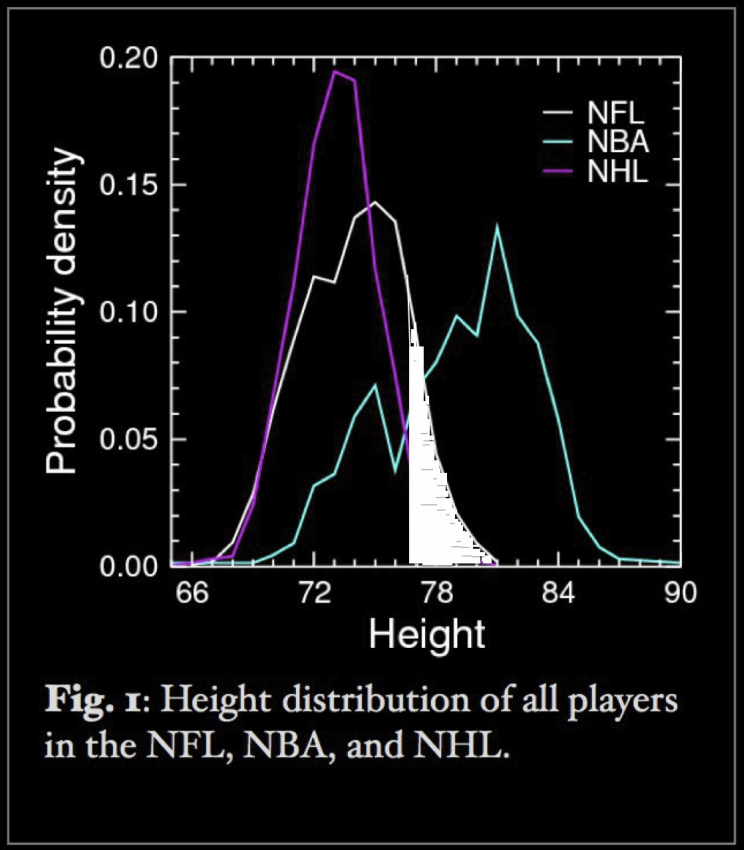


Fig. 1: Height distribution of all players in the NFL, NBA, and NHL.

Estimate the probability that an NFL player chosen at random from the group will have a height greater than 77 inches.

- We must estimate the area under the white curve(height of NFL players) to the right of 77. This area is shaded in white on the left.
- The total area under the white curve is 1.
- About 10% of that area is in the shaded region, thus the probability is approx. 0.1.

Average of a set of observations of a R. Variable X.

Suppose that we perform 20 trials of the experiment "roll a fair six sided die" and get the following outcomes: $1, 6, 3, 2, 5, 2, 3, 2, 4, 6, 4, 6, 2, 6, 3, 6, 2, 6, 3, 5$.

- We can calculate the average of these outcomes by adding the numbers and dividing by twenty:
- $\bar{x} = \text{Average} = \frac{1+6+3+2+5+2+3+2+4+6+4+6+2+6+3+6+2+6+3+5}{20} = \frac{77}{20}$
- For a very large number of trails, it would be better to organize the data in a frequency table first:

Outcomes (O_i)	Frequency (f_i)
1	1
2	5
3	4
4	2
5	2
6	6

$\bar{x} = \frac{1(1)+2(5)+3(4)+4(2)+5(2)+6(6)}{20} = \frac{77}{20}$

$= \frac{1(f_1)+2(f_2)+3(f_3)+4(f_4)+5(f_5)+6(f_6)}{20}$

we can rewrite this as the sum of the outcomes times their relative frequencies:

$\bar{x} = 1\frac{f_1}{20} + 2\frac{f_2}{20} + 3\frac{f_3}{20} + 4\frac{f_4}{20} + 5\frac{f_5}{20} + 6\frac{f_6}{20}$

- For a very large number of trials N, we would expect the relative frequency of each outcome to be roughly equal to its probability. Since each outcome has probability, $1/6$ and we would expect the average to be roughly $\boxed{\bar{x} \approx 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5.}$

Expected Value of a Discrete Random Variable

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

$$\boxed{\mu = E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.}$$

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

- 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

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

Example: Expected Value of a Discrete Random Variable

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. Find $E(X)$ (the expected value of X).

- We've already worked out the probability distribution of this Random Variable:

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

Example: Expected Value of a Discrete Random Variable

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. Find $E(X)$ (the expected value of X).

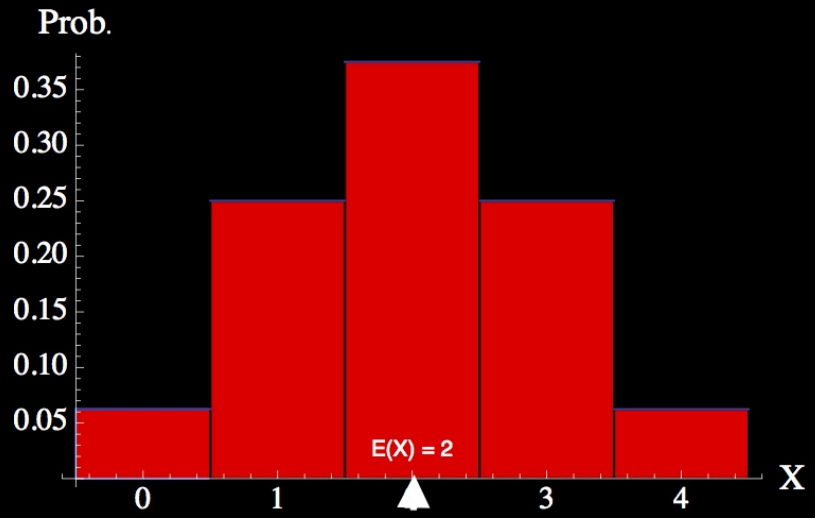
- We've already worked out the probability distribution of this Random Variable:

X= # Heads	**P(X)**	XP(X)
0	1/16	0
1	4/16	4/16
2	6/16	12/16
3	4/16	12/16
4	1/16	4/16
	E(X) = Total − >	32/16 = 2

Graphical Interpretation of the Expected Value

Graphically the expected value of a random variable X corresponds to the **balance point of the graphical representation of the distribution of X.**

- Here is the graphical representation of the distribution of $X = \#$ Heads from the previous example where $E(X) = 2$;



- If the variable is discrete but has infinitely many possible outcomes, we can use infinite summation to calculate the expected value, however this is beyond the scope of this course.

More Examples

For the experiment consisting of flipping a coin 4 times and observing the sequence of heads and tails, we also figured out the probability distribution of the two random variables: Y = length of the longest run of heads in the sequence and Z = total number of runs in the sequence (of both H's and T's).

- The distribution of these variables are:

Y = longest run H's	**P(Y)**
0	1/16
1	7/16
2	5/16
3	2/16
4	1/16

Z = # Runs	**P(Z)**
1	2/16
2	6/16
3	6/16
4	2/16

More Examples

For the experiment consisting of flipping a coin 4 times and observing the sequence of heads and tails, we also figured out the probability distribution of the two random variables: Y = length of the longest run of heads in the sequence and Z = total number of runs in the sequence (of both H's and T's).

- We calculate the expected value as before as the sum of the products of the outcomes and their probabilities:

Y = longest run H's	**P(Y)**	**YP(Y)**
0	1/16	0
1	7/16	7/16
2	5/16	10/16
3	2/16	6/16
4	1/16	4/16
	Total$-> = E(Y)$	27/16 ≈ 1.69

Z = # Runs	**P(Z)**	**ZP(Z)**
1	2/16	2/16
2	6/16	12/16
3	6/16	18/16
4	2/16	8/16
	Total$-> = E(Z)$	40/16 $= 2.5$

Expected Value of A Continuous R.V.

For a continuous random variable, X, we can use a method from calculus called integration to calculate the expected value. This is beyond the scope of this course, however $E(X)$ can be thought of geometrically as the balance point of the probability density function in this case.

- Example: We can find the average height of an NFL player chosen at random from the population of NFL players from the picture of the probability density function shown below. We estimate the point of balance of the white distribution to get roughly 75 inches as the average height of an NFL player.

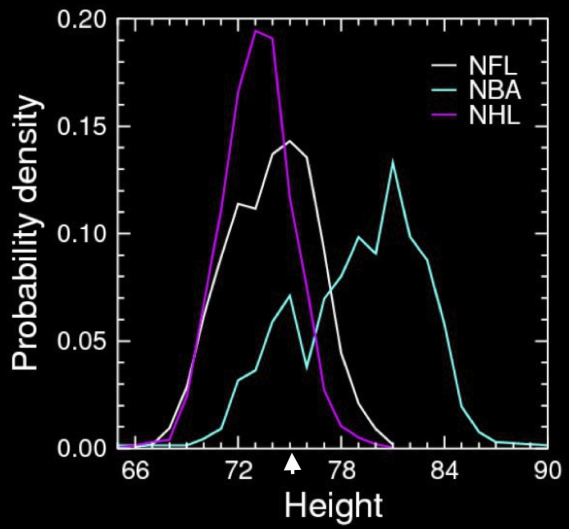


Fig. 1: Height distribution of all players in the NFL, NBA, and NHL.

- As with discrete variables, we can interpret the expected value of a continuous random variable as the number we would expect to get if we calculated the average of the observations of the variable over many trials of the experiment.

Measures of Variability for a R.V.

When we analyzed the length of the longest run of heads in K flips of a coin, we saw that we could expect some variation in the length of the longest run in randomly generated data.

We also saw that in order to make good decisions about the data, we needed some measure of the variation.

- In this section, we will look at two related measures of variation for a random variable; the variance and the standard deviation.
- The variance of a random variable can be viewed as the average squared distance of the outcomes from the mean (expected value).

Variance and St. Dev. of a Discrete R.V.

Let us reconsider our experiment of rolling a fair six-sided die and let X denote the number on the uppermost face. We saw already that $\mu = E(X) = 3.5$.

- The variance is the (weighted) average squared distance of the outcomes from the expected value of X. For any outcome, x, its squared distance from $\mu = E(X)$ is given by $(x - \mu)^2$
-

X	**P(X)**	XP(X)	(X - μ)	$(X - \mu)^2$	$(X - \mu)^2 P(X)$
1	1/6	1/6	-2.5	6.25	(6.25)/6
2	1/6	2/6	-1.5	2.25	(2.25)/6
3	1/6	3/6	-0.5	0.25	(0.25)/6
4	1/6	4/6	0.5	0.25	(0.25)/6
5	1/6	5/6	1.5	2.25	(2.25)/6
6	1/6	6/6	2.5	6.25	(6.25)/6
	μ = Total $->$	21/6 = 3.5		σ^2 = Sum $->$	(17.5)/6 = 35/12

- If we roll the die many times, we would expect the average squared distance of the outcomes from the mean to be approximately $\sigma^2 = 35/12 \approx 2.92$.
- The Standard Deviation of X denoted by σ is the square root of the variance. $\sigma = \sqrt{\sigma^2} = \sqrt{35/12} \approx 1.71$.

General Variance and St. Dev. of a Discrete R.V.

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

$$\boxed{\textbf{Standard Deviation} = \sigma(X) = \sqrt{\text{Variance}}}.$$

- To compute the variance, we can proceed as in the previous example:

$\mathbf{x_i}$	$\mathbf{p_i}$	$\mathbf{x_i p_i}$	$(\mathbf{x_i} - \mu)$	$(\mathbf{x_i} - \mu)^2$	$\mathbf{p_i}(\mathbf{x_i} - \mu)^2$
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** $= \sigma^2(X)$

Example: Variance and St. Dev. of a Discrete R.V.

An experiment consists of flipping a coin 4 times and observing the sequence of heads and tails. The random variable Z is the number of runs in the sequence. Find $E(Z)$ and the standard deviation, $\sigma(Z)$

-

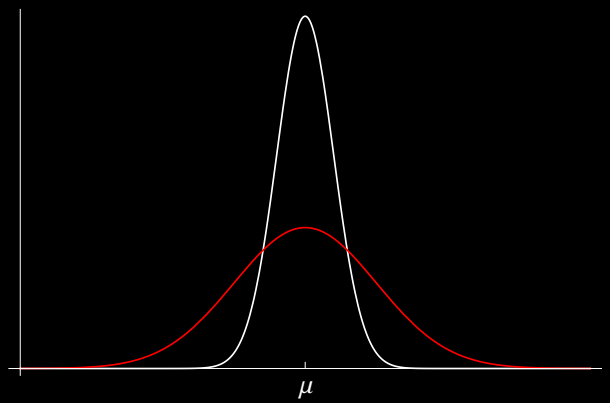
Z	**P(Z)**	ZP(Z)	(Z - μ)	$(Z-\mu)^2$	$(Z-\mu)^2 P(Z)$
1	2/16	2/16	-1.5	2.25	0.281
2	6/16	12/16	-0.5	0.25	0.094
3	6/16	18/16	0.5	0.25	0.094
4	2/16	8/16	1.5	2.25	0.281
	$\mu =$ Total $->$	$40/16 = 2.5$		$\sigma^2 \approx$ Sum $->$	0.75

- The standard deviation is given by $\sigma(Z) = \sigma = \sqrt{\sigma^2} \approx \sqrt{0.75} \approx 0.866$.

The Standard Deviation For Continuous Random Variables

The calculation of the variance for a continuous random variable is beyond the scope of this course.

- However, the variance for a continuous random variable X has the same interpretation as in the discrete case, it should give a good approximation to the average squared distance of the outcomes from the mean if we have many independent observations of the random variable X.
- The standard deviation of a continuous random variable is the square root of the variance.
- If two continuous random variables have the same mean but different standard deviations, then the one with the larger standard deviation will have greater variation in its observations.
- For the symmetric distributions of the variables X(in red) and Y(in white) shown below, we have $E(X) = E(Y) = \mu$ and $\sigma(X) > \sigma(Y)$.

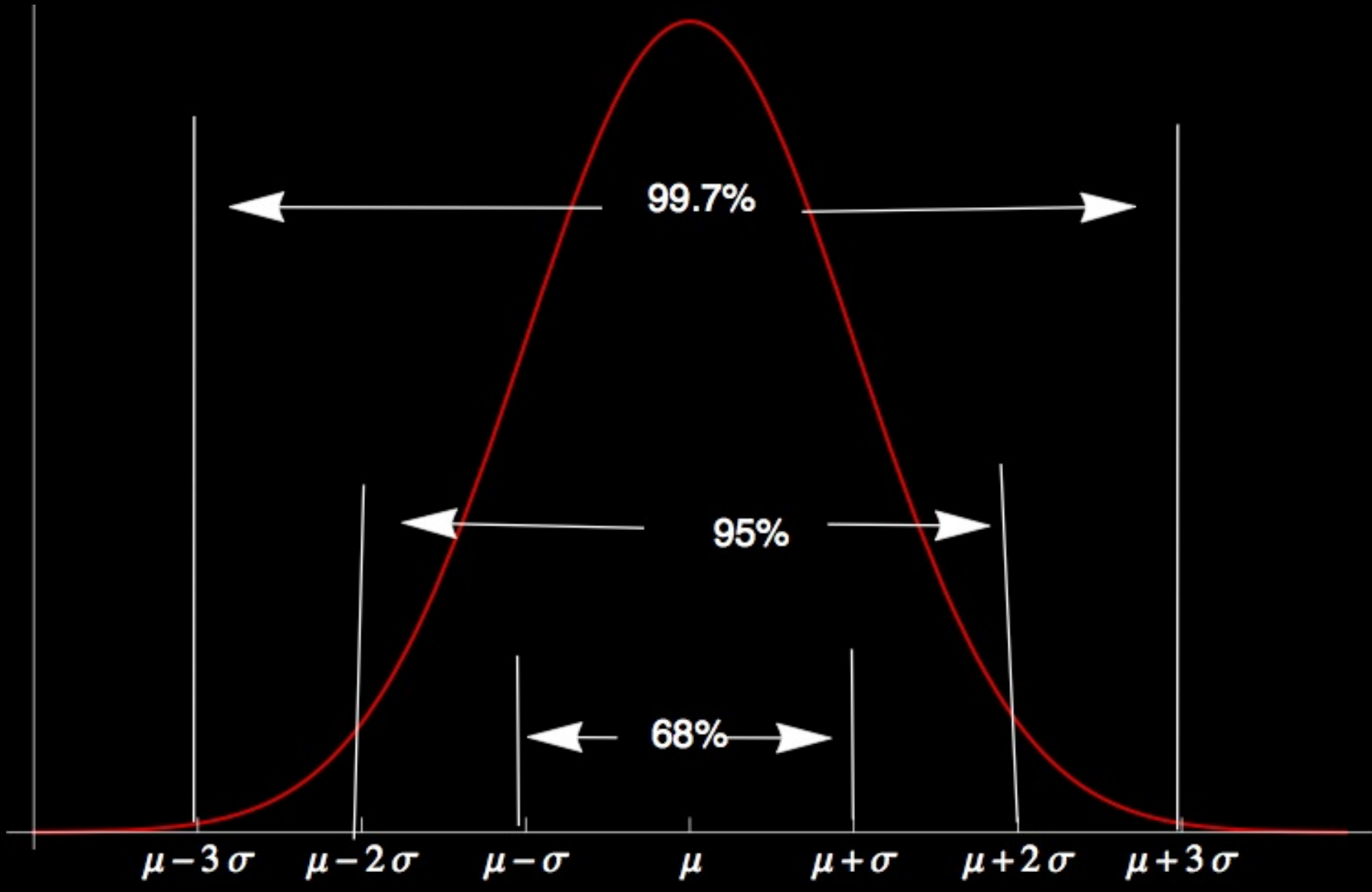


The Empirical Rule for Bell Shaped Densities

Bell curves or Normal Density functions frequently give good approximations to the densities of random variables we observe in everyday life. There are tables available for precise calculation of probabilities for these distributions. In this course, we will use a rule of thumb called the Empirical Rule: If a random variable has a probability distribution which is bell shaped or approximately bell shaped, we have the following

- The probability of getting an outcome within one standard deviation of the mean on any given trial of the experiment is approximately 0.68. That is $P(\mu - \sigma, \mu + \sigma) \approx 0.68$
- The probability of getting an outcome within two standard deviations of the mean on any given trial of the experiment is approximately 0.95. That is $P(\mu - 2\sigma, \mu + 2\sigma) \approx 0.95$
- The probability of getting an outcome within three standard deviations of the mean on any given trial of the experiment is approximately 0.997. That is $P(\mu - 3\sigma, \mu + 3\sigma) \approx 0.997$.

The Empirical Rule for Bell Shaped Densities



Z-Scores, Numerical Measures of relative Standing

Quite often when interpreting an observation of a random variable, such as height or weight or a performance statistic, we are interested in how it compares to the rest of the population; is it close to average or outstanding in some respect?

- One measure of relative standing of an observation x is called a Z-score and it measures the number of standard deviations that the observation lies from the mean of the population.
- For an observation x of a random variable X with mean $\mu = E(X)$ and standard deviation $\sigma = \sigma(X)$, the Z-score for the observation x is given by

$$z = \frac{x - \mu}{\sigma}.$$

Using Z-Scores

One way in which we can use Z-scores is to standardize scores for comparison of performance.

- Example: The NFL combine is a week-long showcase where college football players perform physical and mental tests in front of National Football League coaches, general managers, and scouts. On this webpage: http://matlabgeeks.com/sports-analysis/nfl/nfl-draft-running-the-40-and-bench-pressing, the statistics for the performance of the top 750 prospects for the NFL draft at the NFL yearly combine over a 7 year period from 2005 to 2011. The average time for a wide receiver for the 40-yard dash was $\mu_{40} = 4.51$ and the standard deviation was $\sigma_{40} = 0.1$. The average time for a wide receiver for the cone test was $\mu_c = 6.96$ and the standard deviation was $\sigma_c = 0.2$. Notre dame player Golden Tate took part in the 2010 NFL combine. His time for the 40 yard dash was 4.42 seconds and his time for the cone test was 7.12 seconds. In which test did he have a better performance?

Using Z-Scores

Let's calculate the Z-scores for comparison

	W.R. 40 yd. dash $\mu_{40} = 4.51$ $\sigma_{40} = 0.1$	W.R Cone Test $\mu_c = 6.96$ $\sigma_c = 0.2$
G.Tate	4.42	7.12

- G. Tate's z-score for the 40 yard dash was $z_{40} = \frac{4.42-4.51}{0.1} = -0.9$.
- G. Tate's z-score for the cone test was $z_c = \frac{7.12-6.96}{0.2} = 0.8$.
- In the 40-yd dash G.Tate's time was 0.9 standard deviations below the average for wide receivers and on the cone test his time was 0.8 standard deviations above the mean. Therefore he had a better relative performance on the 40-yd. dash.

Using Z-Scores to make Decisions

We could also use z-scores to make better decisions about whether a string of baskets and misses was randomly generated, something we considered in the previous section. Our reasoning would go like this:

- If a player took K shots with a constant probability p of making a basket on every shot, where K is large;
- the expected length of the longest run of baskets is approximately $\mu = \frac{-\ln((1-p)K)}{\ln(p)}$ and the standard deviation is approximately $\sigma = \frac{-\pi}{\sqrt{6}\ln(p)}$. (see Schilling's paper for details).
- The chances that the length of the longest run of baskets in a randomly generated sequence will be outside the interval $(\frac{-\ln((1-p)K)}{\ln(p)} - 2(\frac{-\pi}{\sqrt{6}\ln(p)}),\ \frac{-\ln((1-p)K)}{\ln(p)} + 2(\frac{-\pi}{\sqrt{6}\ln(p)}))$ is about 5%.
- If the longest run in the data for large K is outside this interval, I will say that the sequence is not randomly generated, otherwise, I will say there is not enough evidence to say that the sequence is not randomly generated. There is a 5% chance that I will be wrong when I say a sequence is not randomly generated.
- This is a crude example of something called Hypothesis testing.

Using Z-Scores to make Decisions; Example

We could also use z-scores to make better decisions about whether a string of baskets and misses was randomly generated, something we considered in the previous section. Our reasoning would go like this:

- If a player took K shots with a constant probability p of making a basket on every shot, where K is large;
- the expected length of the longest run of baskets is approximately $\mu = \frac{-\ln((1-p)K)}{\ln(p)}$ and the standard deviation is approximately $\sigma = \frac{-\pi}{\sqrt{6}\ln(p)}$. (see Schilling's paper for details).
- The chances that the length of the longest run of baskets in a randomly generated sequence will be outside the interval $(\frac{-\ln((1-p)K)}{\ln(p)} - 2(\frac{-\pi}{\sqrt{6}\ln(p)}),\ \frac{-\ln((1-p)K)}{\ln(p)} + 2(\frac{-\pi}{\sqrt{6}\ln(p)}))$ is about 5%.
- If the longest run in the data for large K is outside this interval, I will say that the sequence is not randomly generated, otherwise, I will say there is not enough evidence to say that the sequence is not randomly generated. There is a 5% chance that I will be wrong when I say a sequence is not randomly generated.
- This is a crude example of something called Hypothesis testing.

Example: Does this player have unusually long/short "longest run of Baskets"?

Last day we looked at data for a string of 548 consecutive shots taken by Dion Waiters who has a career FG% of $p = 0.417$. We see that the longest run of baskets in the data has length 5 (in fact there are 3 such runs in the data). Is this consistent with what we would expect from a player who takes 548 consecutive shots with a probability of 0.417 of making a basket on each shot?

- For such a player, the expected value of the longest run of baskets in the data is approximately $\mu = \frac{-\ln((0.583)548)}{\ln(0.417)} \approx 6.88 \approx 7$
- The standard deviation is approximately $\sigma = \frac{-\pi}{\sqrt{6}\ln(0.417)} \approx 1.47$
- Our rule tells us that if the longest run is outside the interval $(6.68 - 2(1.47), 6.68 + 2(1.47)) = (3.74, 9.62)$, we decide that the sequence is not a result of 548 Bernoulli trials with $p = 0.417$, otherwise we say we have no reason to believe it is not the result of such an experiment.
- In this case, the observed value of the longest run is 5 which is not outside the interval, so we have no reason to believe that the sequence is not a result of 548 Bernoulli trials with $p = 0.417$.

Wald Wolfowitz Runs Test

The relatively easy to use test for randomness given below is due to Wald and Wofowitz.

Given a sequence with two values, success (S) and failure (F), with N_s success' and N_f failures, let X denote the number of runs (of both S's and F's). Wald and Wolfowitz determined that for a random sequence of length N with N_s success' and N_f failures (note that $N = N_s + N_f$), the number of runs has mean and standard deviation given by

$$E(X) = \mu = \frac{2N_sN_f}{N} + 1, \qquad \sigma(X) = \sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}.$$

The distribution of X is approximately normal if N_s and N_f are both bigger than 10.

- Using our crude form of hypothesis testing to test if a sequence of Baskets and Misses was a result of K Bernoulli trials with constant probability, we would decide that the sequence was not randomly generated if the observed number of runs fell outside the interval $(\frac{2N_sN_f}{N} + 1 - 2(\sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}), \frac{2N_sN_f}{N} + 1 - 2(\sqrt{\frac{(\mu-1)(\mu-2)}{N-1}}))$.

Example: Wald Wolfowitz Runs Test

Use the Wald Wolfowitz Runs Test to test the following sequence of 58 consecutive baskets and misses for basketball player J R Smith for randomness:

MMBBMMMMMBMMMBBMMBMBMMBMMBMBMM
BBMMMMMBBMBMBBBBMMBMBMMBBBBMMM

- We have the number of baskets is $N_B = 25$ and the number of misses is $N_M = 33$. The number of runs is 31.
- The expected number of runs is $\mu = \frac{2N_B N_M}{N} + 1 = \frac{2(33)(25)}{58} + 1 \approx 29$
- The standard deviation of the number of runs in randomly generated data of this type is $\sigma = \sqrt{\frac{(\mu-1)(\mu-2)}{N-1}} = \sqrt{\frac{(29-1)(29-2)}{58-1}} \approx 3.64$
- If the observed number of runs is outside the interval $(29 - 2(3.64), 29 + 2(3.64)) = (21.72, 36.28)$, we will conclude that this sequence was not randomly generated, otherwise, we say that there is not sufficient evidence to make such a conclusion.
- In this case, the observed number of runs is 31 which is in the interval $(21.72, 36.28)$, so we say that there is not sufficient evidence to occlude that the sequence of baskets and misses is not randomly generated as a sequence of Bernoulli trials with constant probability.