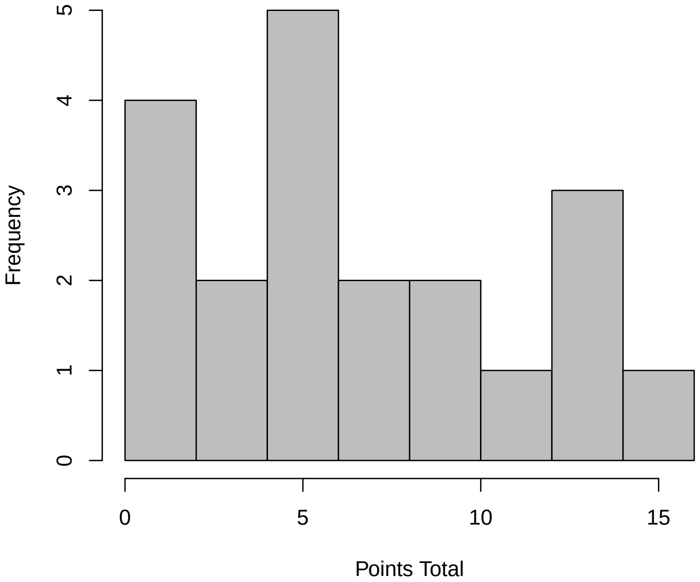
PLSC 503 – Fall 2023
Central Tendency and Variation

September 18, 2023

Current Premier League Statistics

Team	Won	Drew	Lost	GoalsFor	GoalsAgainst	GoalDifference	Points
Manchester City	5	0	0	14	3	11	15
Tottenham Hotspur	4	1	0	13	5	8	13
Liverpool	4	1	0	12	4	8	13
Arsenal	4	1	0	9	4	5	13
Brighton and Hove Albion	4	0	1	15	7	8	12
West Ham United	3	1	1	10	7	3	10
Aston Villa	3	0	2	11	10	1	9
Crystal Palace	2	1	2	6	7	-1	7
Fulham	2	1	2	5	10	-5	7
Brentford	1	3	1	8	6	2	6
Newcastle United	2	0	3	8	7	1	6
Nottingham Forest	2	0	2	6	6	0	6
Manchester United	2	0	3	6	10	-4	6
Chelsea	1	2	2	5	5	0	5
Bournemouth	0	3	2	4	8	-4	3
Wolverhampton Wanderers	1	0	4	5	11	-6	3
Sheffield United	0	1	4	5	9	-4	1
Everton	0	1	4	2	9	-7	1
Burnley	0	0	3	3	11	-8	0
Luton Town	0	0	4	2	10	-8	0

Premier League Points: Histogram



The Arithmetic Mean

The "mean":

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

implies that:

$$\sum_{i=1}^{N} X_i = N\bar{X}$$

and so:

$$\sum_{i=1}^{N}(X_i - \bar{X}) = 0$$

$\bar{X}$ Minimizes Squared Deviations

Find the value of X μ that minimizes the sum of squared deviations...

$$
\begin{aligned}
f(X) &= \sum_{i=1}^{N}(X_i - \mu)^2 \\
&= \sum_{i=1}^{N}(X_i^2 + \mu^2 - 2\mu X_i) \\
\frac{\partial f(X)}{\partial \mu} &= \sum_{i=1}^{N}(2\mu - 2X_i)
\end{aligned}
$$

$\bar{X}$ Minimizes Squared Deviations

Solve:

$$\begin{aligned}
\sum_{i=1}^{N}(2\mu - 2X_i) &= 0 \\
2N\mu - 2\sum_{i=1}^{N} X_i &= 0 \\
2N\mu &= 2\sum_{i=1}^{N} X_i \\
\mu &= \frac{1}{N}\sum_{i=1}^{N} X_i \equiv \bar{X}
\end{aligned}$$

Means from Sums of Frequencies

Frequency table:

Points	Frequency f_j
0	2
1	2
3	2
5	1
$\vdots$	$\vdots$
15	1

For J different unique values of X:

$$\bar{X} = \frac{1}{N} \sum_{j=1}^{J} f_j X_j$$

Weighted Means

For "weights" w_i, the *weighted mean* is:

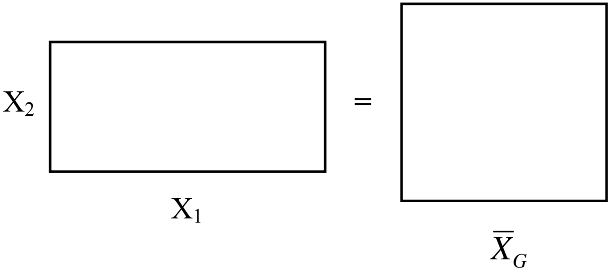
$$\bar{W} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

Things to remember:

- If $w_i = \frac{1}{N} \; \forall \; i$, then $\bar{W} = \bar{X}$
- If $w_i = w \; \forall \; i$, then $\bar{W} = w\bar{X}$
- Weighted means are simpler if $\sum_{i=1}^{N} w_i = 1.0$...
- ... we can normalize any set of weights by $w_i' = \frac{w_i}{\sum_{i=1}^{N} w_i}$.

Geometric Mean

$$
\begin{aligned}
\bar{X}_G &= \left(\prod_{i=1}^{N} X_i \right)^{\frac{1}{N}} \\
&= \sqrt[N]{X_1 \cdot X_2 \cdot \ldots \cdot X_N} \\
&= \exp \left[\frac{1}{N} \sum_{i=1}^{N} \ln X_i \right]
\end{aligned}
$$



Geometric Mean (continued)

Note: Geometric means don't like negative values...

- Formally, $\bar{X}_G$ is defined only if $X_i > 0 \; \forall \; i$
- R's `geometric.mean()` defaults to removing them before calculation...
- If *all* values of X are negative, the geometric mean will be `NaN`.

Consider percentage changes:

$$\{ +12\%, +5\%, \text{-}9\%, +2\%, \text{-}10\% \}$$

```
> geometric.mean(c(12,5,-9,2,-10))
[1] 4.932424

Warning message:
In log(x) : NaNs produced

> geometric.mean(c(1.12,1.05,0.91,1.02,0.90))
[1] 0.9964563
```

Harmonic Mean

The harmonic mean is:

$$
\begin{aligned}
\bar{X}_H &= \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}} \\
&= \frac{1}{\left(\overline{\frac{1}{X}}\right)}
\end{aligned}
$$

Note that:

$$\bar{X}_H \leq \bar{X}_G \leq \bar{X}$$

The Median

Median:

$$\begin{aligned} \check{X} &= \text{“middle observation” of } X \\ &= \text{50th } \textit{percentile} \text{ of } X. \end{aligned}$$

Minimizes *absolute* distance:

$$\check{X} = \min \left(\sum_{i=1}^{N} |X_i - c| \right).$$

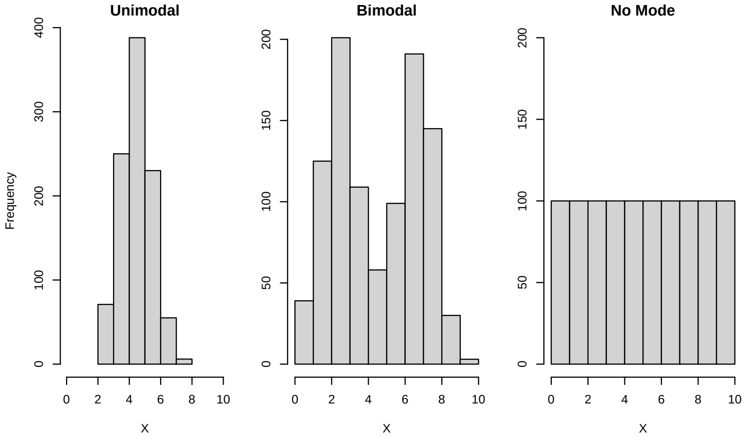
The Mode

The <u>mode</u> of X is "the value of X that appears most frequently in the data."

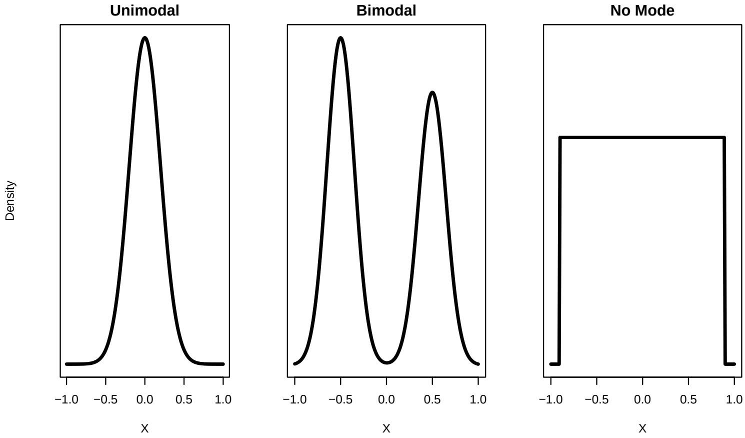
- That works fine for discrete variables...
  - There can be zero, one, two, or more modes,
  - If (say) two values of X have *nearly* the same number of cases, we often refer to that as "bimodal" data.

- For continuous variables:
  - There is often no mode (no two observations have *exactly* the same values of X)
  - Modes are usually defined as any *local maximum* of the probability density function of X

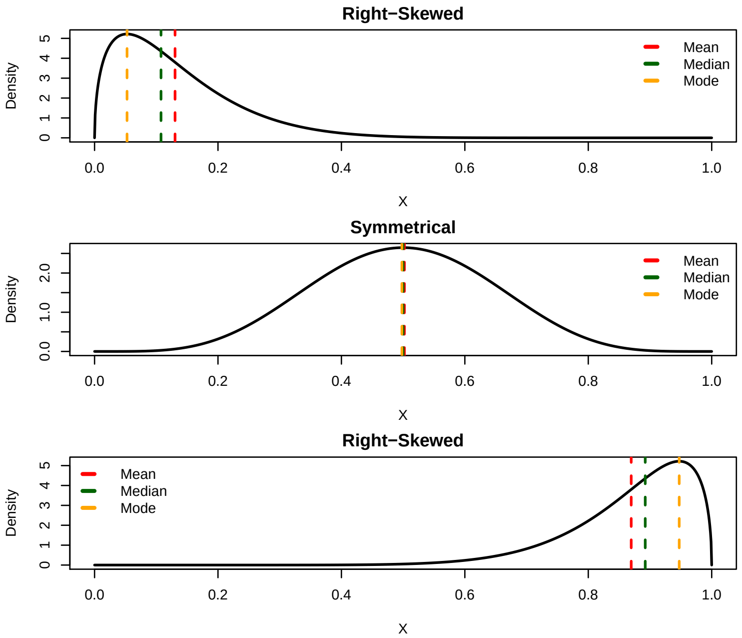
Modes: Discrete X



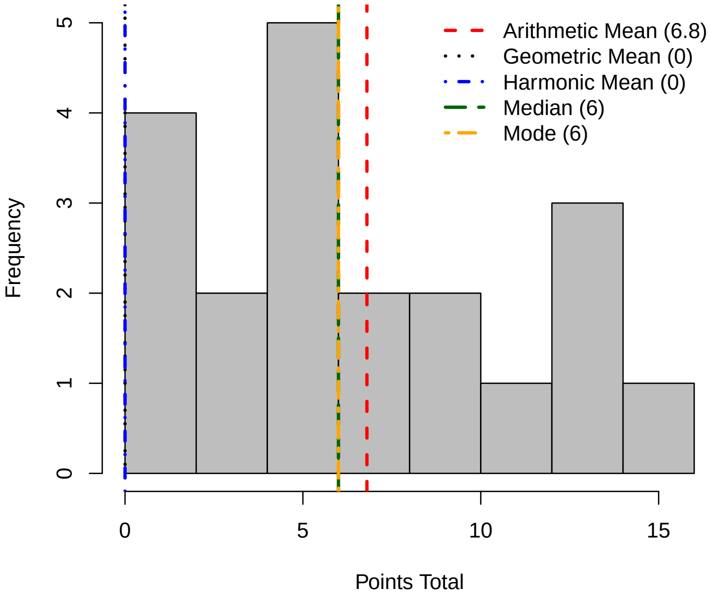
Modes: Continuous X



Means, Medians, Modes, and Skewness



Central Tendencies: Premier League Data



Variation

Range and Percentiles

Range:

$$\text{Range}(X) = \max(X) - \min(X)$$

The k**th percentile** is the value of the variable below which k percent of the observations fall.

- 50th percentile $= \check{X}$
- 0th percentile $=$ minimum(X)
- 100th percentile $=$ maximum(X)

More Percentiles

- *Quartiles* $= \{$25th, 50th, 75th percentiles$\}$

- *Interquartile Range* (IQR):

$$\text{IQR}(X) = \text{75th percentile}(X) - \text{25th percentile}(X)$$

- *Deciles* $= \{$10th, 20th, 30th, etc. percentiles$\}$

"Mean Deviation"

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}).$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X}) &= \frac{1}{N}\left[\left(\sum_{i=1}^{N} X_i\right) - N\bar{X}\right] \\
&= \frac{1}{N}\left[\sum_{i=1}^{N} X_i - N\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)\right] \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} X_i\right) = \frac{1}{N}(0) \\
&= 0
\end{aligned}$$

Squared Deviation

Mean squared deviation:

$$\text{MSD} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Also known as *mean squared error* ("MSE") in regression models...

Note that MSD is "average squared difference from the mean" $\rightarrow$ expressed in "squared" units of X...

A more useful quantity is "root mean squared deviation":

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

An Important Fact

Consider $N = 1$:

Team	Points
Tottenham Hotspur	14

This gives:

$$\bar{X} = \frac{14}{1} = 14 \quad \text{and} \quad RMSD = \sqrt{\frac{(14-14)^2}{1}} = 0$$

For $N = 2$:

Team	Points
Tottenham Hotspur	14
Leeds United	8

we get:

$$\bar{X} = \frac{14+8}{2} = 11 \quad \text{and} \quad RMSD = \sqrt{\frac{(14-11)^2 + (8-11)^2}{2}} = 3$$

You cannot learn about more characteristics of data than you have observations.

Variance:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Standard deviation:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

"Geometric" Standard Deviation:

$$\sigma_G = \exp\left[\sqrt{\frac{\sum_{i=1}^{N}(\ln X_i - \ln \bar{X}_G)^2}{N}}\right]$$

PL Points Data

```
> summary(PL$Points)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     3.0     6.0     6.8    10.5    15.0

> var(PL$Points)
[1] 22.1

> sd(PL$Points)
[1] 4.7
```

Standardizing Variables

Sometimes useful to put variables on a common scale... ("z-scores")...

Typically:

$$Z_i = \frac{X_i - \bar{X}}{\sigma}$$

A standardized variable Z has:

- A mean of zero, and
- A standard deviation (and therefore variance) of 1.0

Standardizing Example

```
> library(psych)

> PLSmall<-PL[,4:10]
> describe(PLSmall,trim=0,skew=FALSE)
               vars  n mean   sd min max range   se
Won               1 20 2.00 1.62   0   5     5 0.36
Drew              2 20 0.80 0.95   0   3     3 0.21
Lost              3 20 2.00 1.41   0   4     4 0.32
GoalsFor          4 20 7.45 3.94   2  15    13 0.88
GoalsAgainst      5 20 7.45 2.48   3  11     8 0.55
GoalDifference    6 20 0.00 5.80  -8  11    19 1.30
Points            7 20 6.80 4.70   0  15    15 1.05

> PL.Z<-scale(PLSmall)
> describe(PL.Z,trim=0,skew=FALSE)
               vars  n mean sd   min  max range   se
Won               1 20    0  1 -1.23 1.85  3.08 0.22
Drew              2 20    0  1 -0.84 2.31  3.15 0.22
Lost              3 20    0  1 -1.41 1.41  2.83 0.22
GoalsFor          4 20    0  1 -1.38 1.92  3.30 0.22
GoalsAgainst      5 20    0  1 -1.79 1.43  3.22 0.22
GoalDifference    6 20    0  1 -1.38 1.90  3.27 0.22
Points            7 20    0  1 -1.45 1.75  3.19 0.22
```

Absolute Deviations and MAD

Median Absolute Deviation ("MAD"):

$$\text{MAD} = \text{median}[|X_i - \check{X}|]$$

Mean Absolute Deviation:

$$\text{Mean Absolute Deviation} = \frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|$$

Moments

Moments are functions of distributions that characterize their shape...

For a random variable X, the kth *raw moment* is:

$$m_k = \begin{cases} \sum f(X)\,\text{Pr}(X) \text{ if } X \text{ is discrete} \\ \int f(X)\,\text{Pr}(X) dX \text{ if } X \text{ is continuous.} \end{cases}$$

The kth *central moment* is:

$$M_k = \begin{cases} \text{E}[(X-\mu)^k] \text{ for discrete } X \\ \int_{-\infty}^{+\infty} (X-\mu)^k f(X)\,dX \text{ for continuous } X \end{cases}$$

A distribution for X can be completely characterized by its non-zero moments...

Why Might We Care?

The first (raw) moment of a variable is the mean:[†]

$$\mu = \mathsf{E}(X)$$

The second (central) moment of a variable is its variance:

$$\sigma^2 = \mathsf{E}[(X - \mu)^2]$$

[†] The first central moment is zero (why?)...

Third central moment:

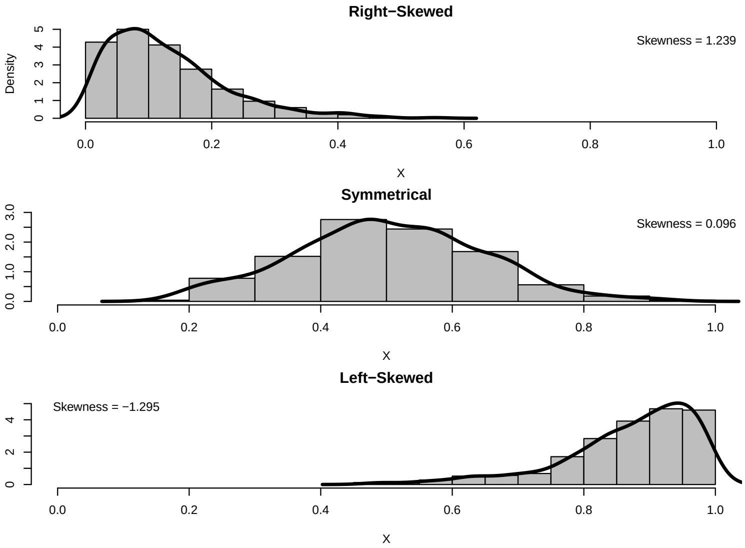
$$M_3 = \mathrm{E}[(X - \mu)^3]$$

More typically, we use the third *standardized moment* (usually called *skewness*):

$$\begin{aligned}\mu_3 &= \frac{M_3^2}{\sigma^3} \\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right]^{3/2}}\end{aligned}$$

- Skewness $= 0 \rightarrow$ symmetrical
- Skewness $> 0 \rightarrow$ "positive" (tail to the right)
- Skewness $< 0 \rightarrow$ "negative" (tail to the left)

Skewness Illustrated



If a distribution is *symmetrical*, then:

- $\mu_3 = 0$
- $\check{X} = (Q_{25} + Q_{75})/2$,
- $\text{MAD} = \frac{\text{IQR}}{2}$

Note that:

- Both discrete and continuous variables can be symmetrical or asymmetrical;
- Every distribution with *no mode* is symmetrical, but
- Unimodal, bimodal, etc. distributions can be symmetrical or asymmetrical.

Fourth moment:

$$M_4 = \mathrm{E}[(X-\mu)^4]$$

More typically, *kurtosis* ("excess kurtosis"):

$$\begin{aligned}\mu_4 &= \frac{M_4}{\sigma^4} - 3 \\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right]^2} - 3\end{aligned}$$

Note that:

$$\frac{M_4}{\sigma^4} \geq \left(\frac{M_3}{\sigma^3}\right)^2 + 1$$

Kurtosis Intuition

Kurtosis is "the average of the standardized X raised to the fourth power (minus three)."

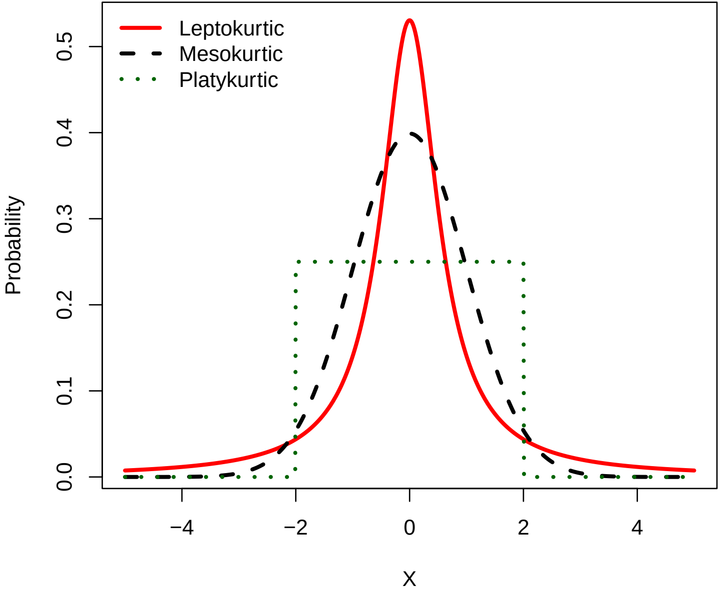
- Values of standardized variables within one σ of $\bar{X}$ have $|X| \leq 1$
- Taking X^4 when $|X| \leq 1$ gives values very close to 0
- $\rightarrow$ only those values on the "tail" of the distribution contribute significantly to kurtosis

Kurtosis Explained

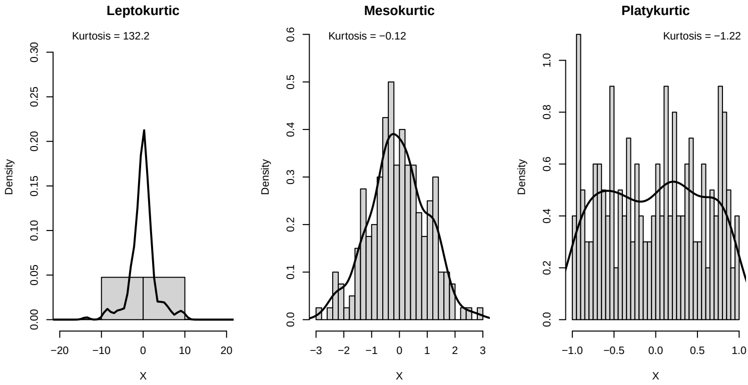
Kurtosis:

- "Fat-tailed" $=$ *leptokurtic*: μ_4 is positive.
- "Medium-tailed" $=$ *mesokurtic*: μ_4 is close to zero.
- "Thin-tailed" $=$ *platykurtic*: μ_4 is negative.

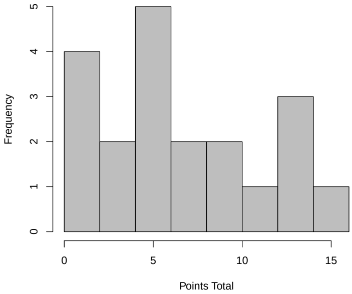
Kurtosis Illustrated



Kurtosis Examples



PL Points Data




```
> library(moments)

> skewness(PL$Points)
[1] 0.2408431

> kurtosis(PL$Points)-3
[1] -0.8326687
```

Binary Variables

For a Bernoulli (binary) variable D:

- $\text{mode}(D) = \check{D}$ (why?)
- The mean of D is:

$$\begin{aligned} \bar{D} &= \frac{1}{N}\sum D_i \\ &= \pi \quad [\equiv \Pr(D = 1)] \end{aligned}$$

- The variance is:

$$\sigma_D^2 = \pi \times (1 - \pi)$$

- and so the standard deviation is:

$$\sigma_D = \sqrt{\pi \times (1 - \pi)}$$

Implies:

- $\sigma_D > \sigma_D^2$
- $\max(\sigma_D^2) \;\leftrightarrow\; \pi = 0.5$

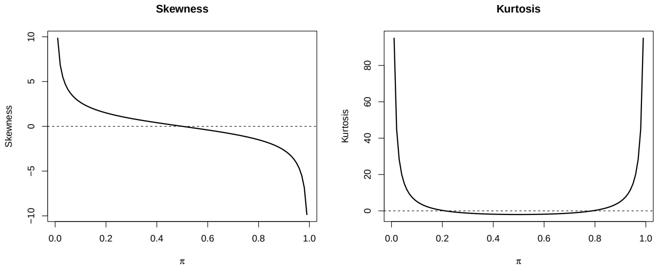
Binary Variables (continued)

For a binary variable, skewness is:

$$\mu_3 = \frac{1 - 2\pi}{\sqrt{\pi(1-\pi)}}$$

and the (excess) kurtosis is:

$$\mu_4 = \frac{1 - 6\pi(1-\pi)}{\pi(1-\pi)}$$



Getting Summary Statistics

Good: summary

```
> summary(PLSmall)
      Won             Drew          Lost       GoalsFor      GoalsAgainst
 Min.   :0.00   Min.   :0.0   Min.   :0   Min.   : 2.00   Min.   : 3.00
 1st Qu.:0.75   1st Qu.:0.0   1st Qu.:1   1st Qu.: 5.00   1st Qu.: 5.75
 Median :2.00   Median :1.0   Median :2   Median : 6.00   Median : 7.00
 Mean   :2.00   Mean   :0.8   Mean   :2   Mean   : 7.45   Mean   : 7.45
 3rd Qu.:3.25   3rd Qu.:1.0   3rd Qu.:3   3rd Qu.:10.25   3rd Qu.:10.00
 Max.   :5.00   Max.   :3.0   Max.   :4   Max.   :15.00   Max.   :11.00
 GoalDifference     Points
 Min.   :-8.00   Min.   : 0.0
 1st Qu.:-4.25   1st Qu.: 3.0
 Median : 0.00   Median : 6.0
 Mean   : 0.00   Mean   : 6.8
 3rd Qu.: 3.50   3rd Qu.:10.5
 Max.   :11.00   Max.   :15.0
```

Better: describe (in psych)

```
> describe(PLSmall)
               vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
Won               1 20 2.00 1.62      2    1.94 2.22   0   5     5  0.21    -1.34 0.36
Drew              2 20 0.80 0.95      1    0.62 1.48   0   3     3  1.07     0.21 0.21
Lost              3 20 2.00 1.41      2    2.00 1.48   0   4     4  0.00    -1.33 0.32
GoalsFor          4 20 7.45 3.94      6    7.25 3.71   2  15    13  0.43    -1.11 0.88
GoalsAgainst      5 20 7.45 2.48      7    7.50 2.97   3  11     8 -0.14    -1.34 0.55
GoalDifference    6 20 0.00 5.80      0   -0.19 6.67  -8  11    19  0.29    -1.20 1.30
Points            7 20 6.80 4.70      6    6.75 5.19   0  15    15  0.15    -1.28 1.05
```

Reporting Summary Statistics

```
> stargazer(PLSmall,title="Summary Statistics")
```

Table: Summary Statistics

Statistic	N	Mean	Std. Dev.	Min	Max
Won	20	2.000	1.620	0	5
Drew	20	0.800	0.951	0	3
Lost	20	2.000	1.410	0	4
Goals For	20	7.450	3.940	2	15
Goals Against	20	7.450	2.480	3	11
Goal Difference	20	0.000	5.800	−8	11
Points	20	6.800	4.700	0	15