

PLSC 503 – Fall 2024

Central Tendency and Variation

September 23, 2024

Link Of The Day



#!/usr/bin/env Rscript

@RandVegan

...

Post your favorite [#rstats](#) blog sites below please :)

I want to do something cool with the entries:



#!/usr/bin/env Rscript @RandVegan · 2h

the bigbookofR.com help me tremendously in my learning journey.

Realizing we need a blog equivalent to categorize and index all the blog posts over the years as there are so many hidden gems.

...

[Show more](#)

4:53 PM · Sep 22, 2024 · **909** Views



4



8



14



11

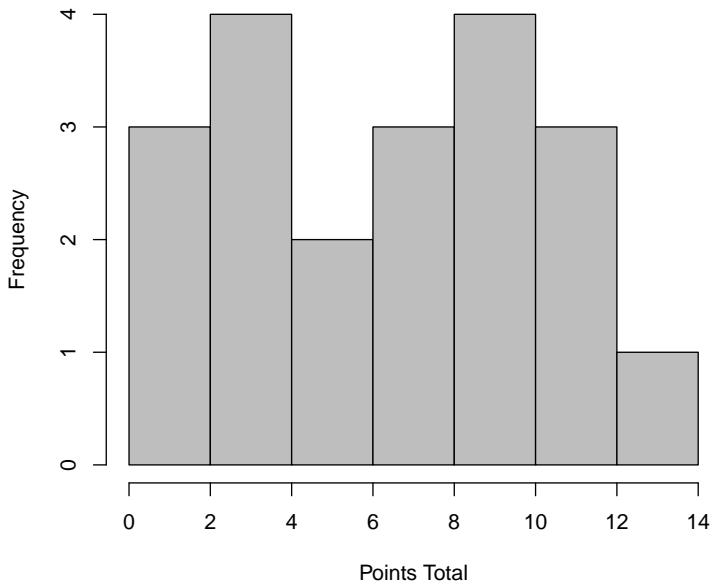


Link: <https://x.com/RandVegan/status/1837958470890348970>

Example: Today's Premier League

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

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:

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

\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:

$$\sum_{i=1}^N (2\mu - 2X_i) = 0$$

$$2N\mu - 2 \sum_{i=1}^N X_i = 0$$

$$2N\mu = 2 \sum_{i=1}^N X_i$$

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

Means from Sums of Frequencies

Frequency table:

Points	Frequency f_j
1	3
3	3
4	1
5	1
\vdots	\vdots
13	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 (arithmetic) 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 $w'_i = \frac{w_i}{\sum_{i=1}^N w_i}$.

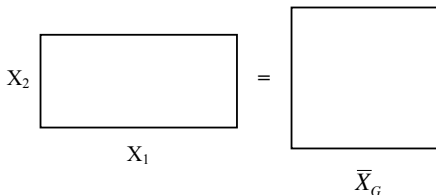
Geometric Mean

The *geometric* mean is:

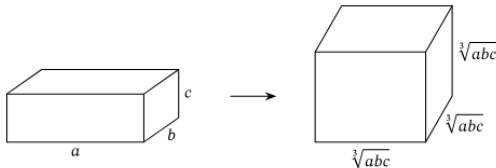
$$\begin{aligned}\bar{X}_G &= \left(\prod_{i=1}^N X_i \right)^{\frac{1}{N}} \\ &= \sqrt[N]{X_1 \cdot X_2 \cdot \dots \cdot X_N} \\ &= \exp \left[\frac{1}{N} \sum_{i=1}^N \ln X_i \right]\end{aligned}$$

Geometric Mean (graphically)

For example, with $N = 2$:



With $N = 3$:



Geometric Mean (continued)

Note: Geometric means don't like zero or 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:

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

```
> 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:

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

The harmonic mean:

- Is the reciprocal of the arithmetic mean of the reciprocals of X
- Can be calculated for negative values of X , *but not for zero values*

Harmonic Mean: Applications

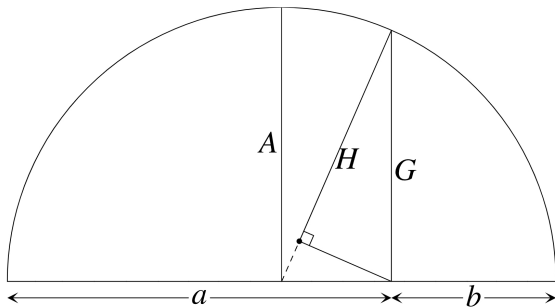
If you drive to a friend's house D miles away going 40 mph (X_1), but bike back at 10 mph (X_2), what is your average speed?

- Hint: It's not $(X_1 + X_2)/2 = (40 + 10)/2 = 25$ mph...
- Average speed is:

$$\begin{aligned}\frac{\text{Total Distance}}{\text{Total Time}} &= \frac{2D}{\frac{D}{X_1} + \frac{D}{X_2}} \\ &= \frac{2}{\frac{1}{X_1} + \frac{1}{X_2}} \\ &= \frac{2}{\frac{1}{40} + \frac{1}{10}} \\ &= 16 \text{ mph}\end{aligned}$$

When X is a *rate*, use the harmonic mean.

The Pythagorean Means



Note that:

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

The *arithmetic median*:

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

The median minimizes *absolute* distance:

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

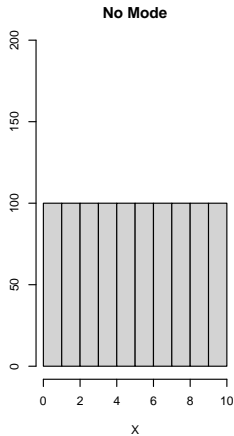
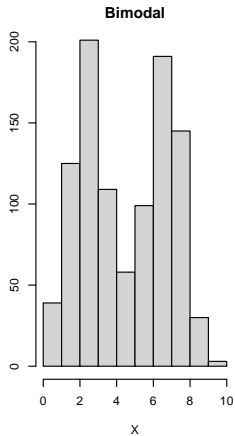
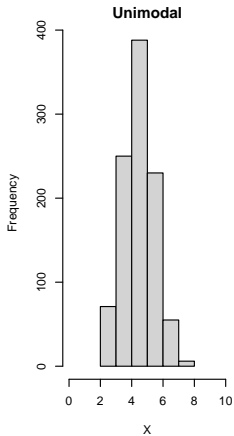
Some Fun Facts About Medians

1. Median observations need not be unique; more than one observation can be the “median” observation. Likewise...
2. By convention, when N is even, \check{X} is the arithmetic mean of the two “middle observations” ...
 - This means for N even, there is *no* observation in the data that is the median observation. Conversely,
 - A *medoid* is the observation in the data that is the most similar / least distant from the others.
3. Medians can be calculated on *discrete* ordinal data (e.g., ranks).
4. Medians can be calculated on *censored* data (e.g., durations).
5. If a distribution has finite variance (see below), the difference between the median and the mean can be no more than one standard deviation.

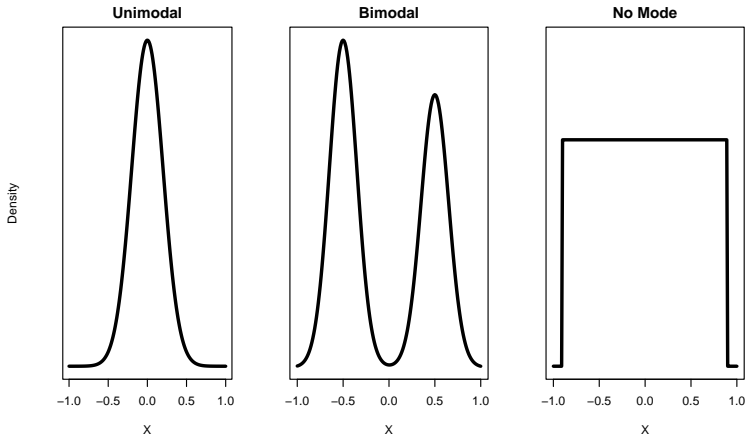
The mode of X is “the value of X that appears most frequently in the data.”

- 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

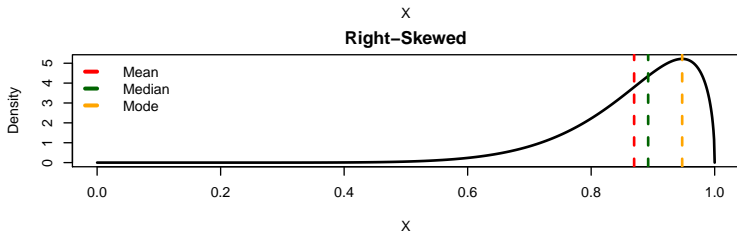
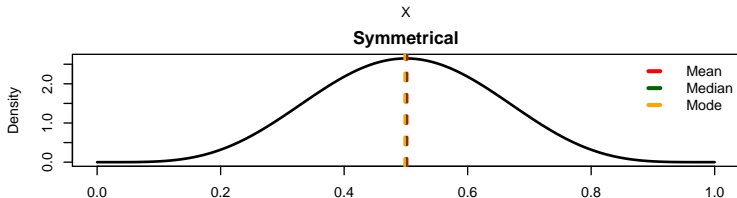
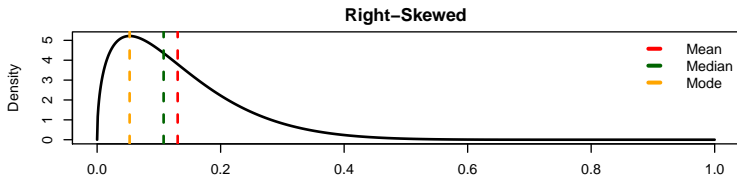
Modes: Discrete X



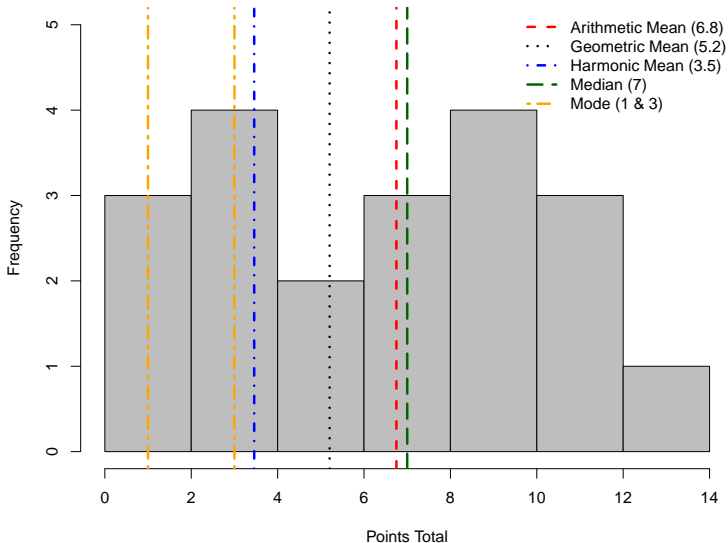
Modes: Continuous X



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 = $\text{minimum}(X)$
- 100th percentile = $\text{maximum}(X)$

More Percentiles

- *Quartiles* = {25th, 50th, 75th percentiles}
- *Interquartile Range* (IQR):

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

- *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”
→ 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	7

This gives:

$$\bar{X} = \frac{14}{1} = 7 \quad \text{and} \quad RMSD = \sqrt{\frac{(7-7)^2}{1}} = 0$$

For $N = 2$:

		Team Points
Tottenham	Hotspur	7
	Everton	1

we get:

$$\bar{X} = \frac{7+1}{2} = 4 \quad \text{and} \quad RMSD = \sqrt{\frac{(7-4)^2 + (1-4)^2}{2}} = 3$$

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.
1.00	3.00	7.00	6.75	10.00	13.00

```
> var(PL$Points)
```

```
[1] 15.7
```

```
> sd(PL$Points)
```

```
[1] 3.96
```

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	median	min	max	range	se
Won	1	20	1.75	1.45	2.0	0	4	4	0.32
Drew	2	20	1.50	1.05	1.0	0	3	3	0.24
Lost	3	20	1.75	1.33	2.0	0	4	4	0.30
GoalsFor	4	20	6.80	2.78	6.5	2	13	11	0.62
GoalsAgainst	5	20	6.80	3.27	6.5	1	14	13	0.73
GoalDifference	6	20	0.00	5.30	0.5	-9	9	18	1.19
Points	7	20	6.75	3.96	7.0	1	13	12	0.89

```
> PL.Z<-scale(PLSmall)
```

```
> describe(PL.Z,trim=0,skew=FALSE)
```

	vars	n	mean	sd	median	min	max	range	se
Won	1	20	0	1	0.17	-1.21	1.56	2.77	0.22
Drew	2	20	0	1	-0.48	-1.43	1.43	2.85	0.22
Lost	3	20	0	1	0.19	-1.31	1.69	3.00	0.22
GoalsFor	4	20	0	1	-0.11	-1.72	2.23	3.95	0.22
GoalsAgainst	5	20	0	1	-0.09	-1.77	2.20	3.98	0.22
GoalDifference	6	20	0	1	0.09	-1.70	1.70	3.40	0.22
Points	7	20	0	1	0.06	-1.45	1.58	3.03	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 k th *raw moment* is:

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

The k th *central moment* is:

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

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:[†]

$$\mu = E(X)$$

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

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

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

Third central moment:

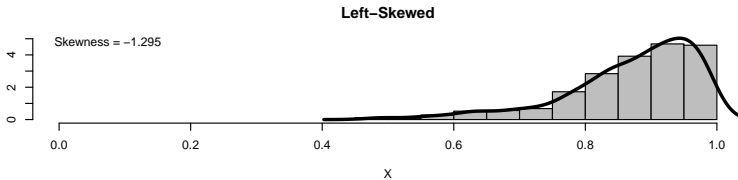
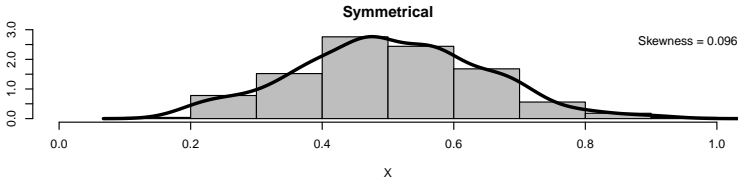
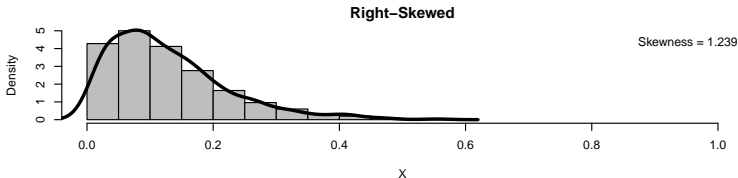
$$M_3 = E[(X - \mu)^3]$$

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:

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

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

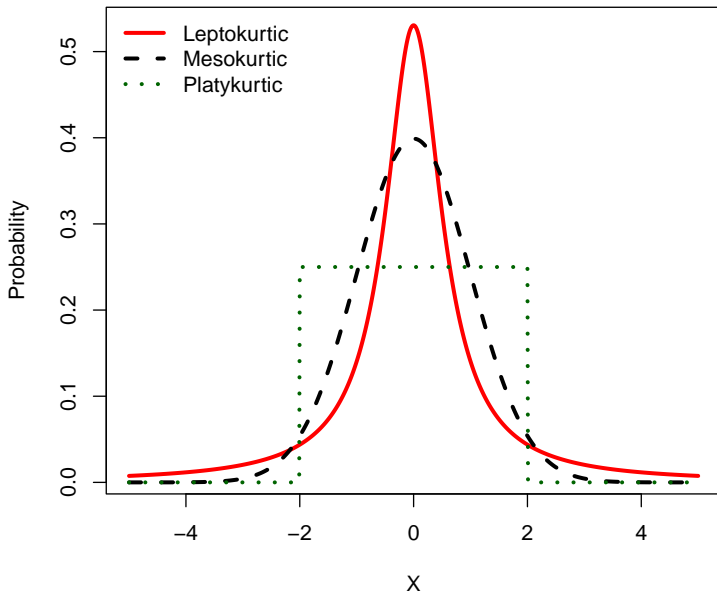
Kurtosis is “the average of the standardized X raised to the fourth power (minus three).”

- 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:

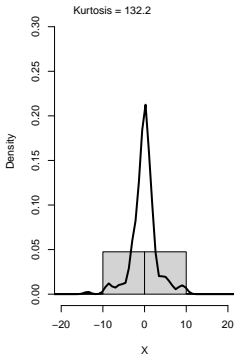
- “Fat-tailed” = *leptokurtic*: μ_4 is positive.
- “Medium-tailed” = *mesokurtic*: μ_4 is close to zero.
- “Thin-tailed” = *platykurtic*: μ_4 is negative.

Kurtosis Illustrated

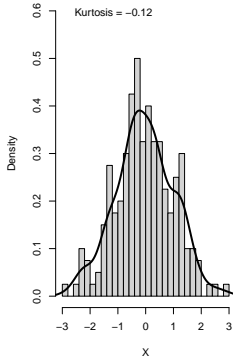


Kurtosis Examples

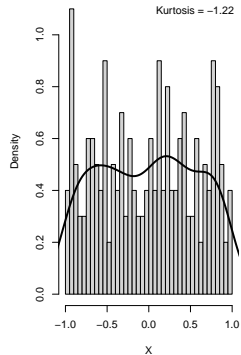
Leptokurtic



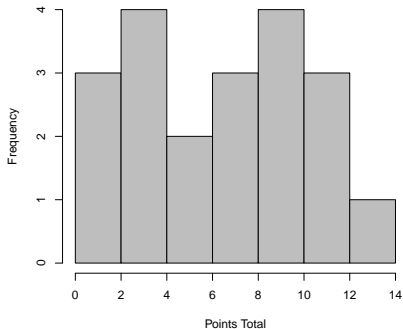
Mesokurtic



Platykurtic



PL Points Data



```
> library(moments)
```

```
> skewness(PL$Points)  
[1] -0.0447
```

```
> kurtosis(PL$Points)-3  
[1] -1.29
```

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:

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

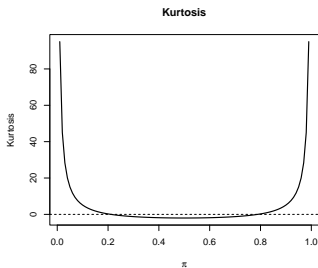
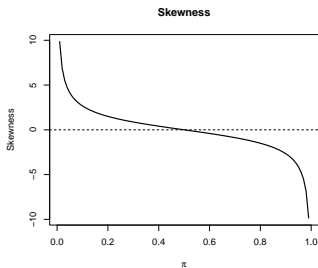
Binary Variables (continued)

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. :0.00	Min. :0.00	Min. : 2.00	Min. : 1.00
1st Qu.:0.00	1st Qu.:1.00	1st Qu.:1.00	1st Qu.: 5.00	1st Qu.: 5.00
Median :2.00	Median :1.00	Median :2.00	Median : 6.50	Median : 6.50
Mean :1.75	Mean :1.50	Mean :1.75	Mean : 6.80	Mean : 6.80
3rd Qu.:3.00	3rd Qu.:2.25	3rd Qu.:2.25	3rd Qu.: 8.25	3rd Qu.: 8.25
Max. :4.00	Max. :3.00	Max. :4.00	Max. :13.00	Max. :14.00

GoalDifference	Points
Min. : -9.00	Min. : 1.00
1st Qu.: -3.25	1st Qu.: 3.00
Median : 0.50	Median : 7.00
Mean : 0.00	Mean : 6.75
3rd Qu.: 4.00	3rd Qu.:10.00
Max. : 9.00	Max. :13.00

Better: describe (in psych)

```
> describe(PLSmall)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Won	1	20	1.75	1.45	2.0	1.69	1.48	0	4	4	0.12	-1.39	0.32
Drew	2	20	1.50	1.05	1.0	1.50	1.48	0	3	3	0.26	-1.31	0.24
Lost	3	20	1.75	1.33	2.0	1.69	1.48	0	4	4	0.31	-1.08	0.30
GoalsFor	4	20	6.80	2.78	6.5	6.69	2.22	2	13	11	0.40	-0.60	0.62
GoalsAgainst	5	20	6.80	3.27	6.5	6.50	2.22	1	14	13	0.66	0.05	0.73
GoalDifference	6	20	0.00	5.30	0.5	0.06	5.19	-9	9	18	-0.10	-1.11	1.19
Points	7	20	6.75	3.96	7.0	6.75	5.19	1	13	12	-0.04	-1.46	0.89

Reporting Summary Statistics

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

Table: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Won	20	1.750	1.450	0	4
Drew	20	1.500	1.050	0	3
Lost	20	1.750	1.330	0	4
Goals For	20	6.800	2.780	2	13
Goals Against	20	6.800	3.270	1	14
Goal Difference	20	0.000	5.300	-9	9
Points	20	6.750	3.960	1	13

Packages / commands for summary statistics:

- `summary` – basic summaries by variable
- `describe` (in `psych`) – flexible summary statistics
- `describe` (in `Hmisc`) – more information than you probably want
- `stat.desc` (in `pastecs`) – like `psych::describe`, but sideways

Packages / commands for making pretty tables:

- `stargazer`
- `tinytable`
- `gt`
- `kable` / `kableExtra`
- `flextable`
- `huxtable`