PLSC 502 – Autumn 2016
Variation

September 15, 2016

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"

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

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

Mean Squared Deviation

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

team	points
Bears	14

team	points
Bears	14
Giants	20

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:

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

NFL Points Data

```
> with(NFL, summary(points))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    16.0    23.0    22.4    28.2    39.0

> with(NFL, var(points))
[1] 87.4

> with(NFL, sd(points))
[1] 9.35
```

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

kth moment:

$$M_k = \mathsf{E}[(X-\mu)^k]$$

- First moment = mean: $\mu = \mathsf{E}(X)$.
- Second moment - variance: $\sigma^2 = \mathsf{E}[(X-\mu)^2]$.

Skewness

Third moment = *skewness*:

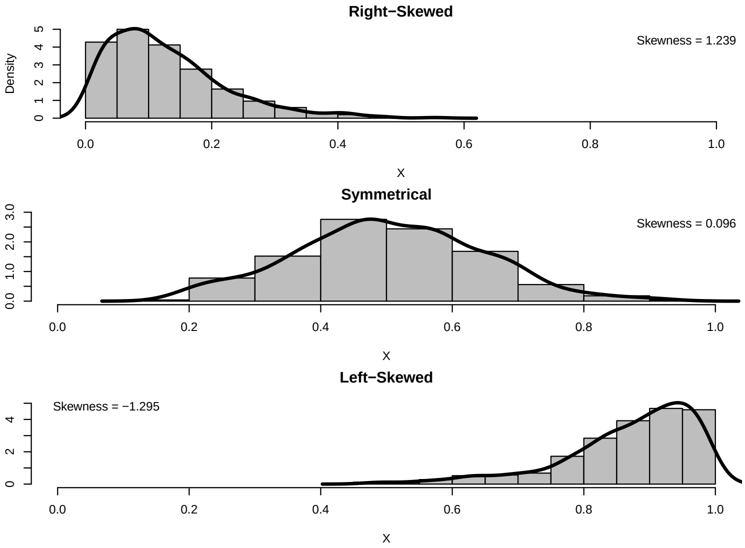
$$M_3 = \mathrm{E}[(X-\mu)^3]$$

More typically:

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

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

Skewness Illustrated



Symmetry

If a distribution is *symmetrical*, then:

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

Kurtosis

Fourth moment = *kurtosis*:

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

More typically ("excess kurtosis"):

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

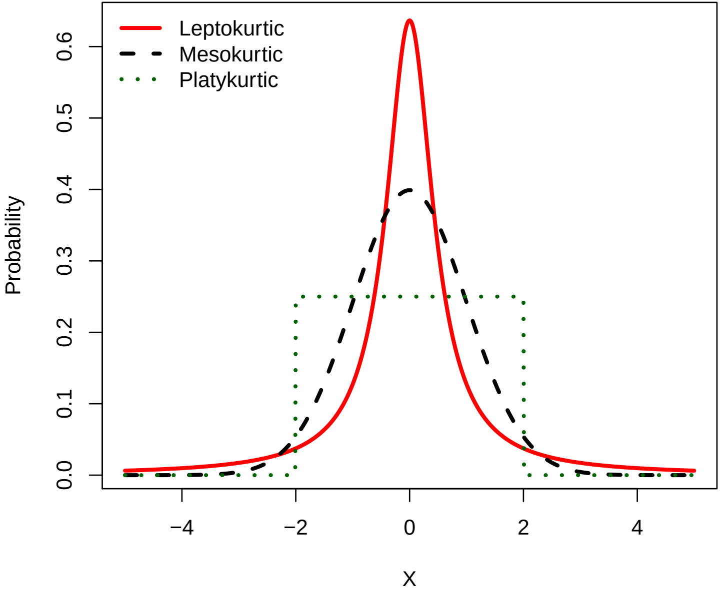
Note that:

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

Kurtosis Explained

- Fat-tailed/"Peaked" $=$ *leptokurtic*: μ_4 is large / (much) greater than zero.
- Medium-tailed $=$ *mesokurtic*: μ_4 is close to zero.
- Thin-tailed/"Flat" $=$ *platykurtic*: μ_4 is small / negative.

Kurtosis Illustrated



NFL Points Data

```
> library(moments)

> with(NFL, skewness(points))
[1] -0.229

> with(NFL, kurtosis(points))
[1] 2.66
```

Dichotomous Variables

Variance:

$$\sigma_D^2 = \bar{D} \times (1 - \bar{D})$$

and so the standard deviation is:

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

Implies:

- $\sigma_D > \sigma_D^2$
- $\max(\sigma_D^2) \;\leftrightarrow\; \bar{D} = 0.5$

Best Practices...

Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
Assassination	0.01	0.09	0	1
Previous Assassinations Since 1945	0.45	0.76	0	4
GDP Per Capita / 1000	5.83	6.04	0.33	46.06
Political Unrest	0.01	1.01	-1.67	20.11
Political Instability	-0.03	0.92	-4.66	10.08
Executive Selection	1.54	1.34	0	4
Executive Power	3.17	2.39	0	6
Repression	1.67	1.19	0	3

Note: $N = 5614$. Statistics are based on all non-missing observations in Model X.