Descriptive Statistics

Bijay Lal Pradhan

Descriptive Statistics

- Describing data (frequency) with *tables* and *graphs* (quantitative or categorical variables)
- Numerical descriptions of *center, variability, Normality (skewness, kurtosis)*
- *Bivariate and Multivariate* descriptions

Frequency

Frequency tells you **how often something happened**. The frequency of an observation tells you the number of times the observation occurs in the data.

Frequency distribution is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

Frequency Table

State	Wheat Production (000 metric tons)
I	65
II	94
III	108
IV	95
V	80
VI	65
VII	40

Gender	Nos of person
Male	150
Female	86

Discrete Distribution		***Continuous Distribution***	
No. of children	*No. of families*	*Weight (kg)*	*No. of students*
0	10	50-55	7
1	30	55-60	20
2	40	60-65	25
3	80	65-70	37
4	100	70-75	9
5	40	75-80	2
Total	300	Total	100

Multivariate table

Year / *Trade union*	*2018*			*2019*		
	Male	*Female*	*Total*	*Male*	*Female*	*Total*
Members	1175	25	1200	1290	290	1590
Non–members	375	175	550	180	100	280
Total	1550	200	1750	1470	390	1860

Cross tabs (SPSS) / Pivot table (Excel)

Stem-and-leaf plot (John Tukey, 1977)

Example: Exam scores (*n* = 40 students)

Stem	Leaf
3	6
4	
5	37
6	235899
7	011346778999
8	00111233568889
9	02238

Numerical Descriptives

Center	Mean	Mode, median
Spread	Variance (standard deviation)	Range, Quartile Deviation
Skewness	Skewness	
Peakedness	Kurtosis	

Central Tendency

Let *y* denote a quantitative variable, with observations $y_1, y_2, y_3, \ldots, y_n$

Describing the *center*

Median: Middle measurement of ordered sample

Mean:

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\Sigma y_i}{n}$$

Mode: most repeated value (which has highest frequency)

Also: Geometric mean and Harmonic mean

Which central tendency should be used??

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Dispersion

Range= L-S

Quartile Deviation = (Q3-Q1)/2

Mean Deviation = $\frac{\Sigma f|X - \bar{X}|}{N}$

Standard Deviation = $\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}} = \sqrt{\frac{\Sigma X^2}{n} - \left\{\frac{\Sigma X}{n}\right\}^2}$

Measure of position

p^{th} *percentile*: *p* percent of observations below it, (100 - *p*)% above it.

- *p* = 50: *median*
- *p* = 25: *lower quartile* (LQ)
- *p* = 75: *upper quartile* (UQ)
- *Interquartile range* IQR = UQ - LQ

Boxplot and Outliers

Box plots have box from LQ to UQ, with median marked. They portray a *five-number summary* of the data:

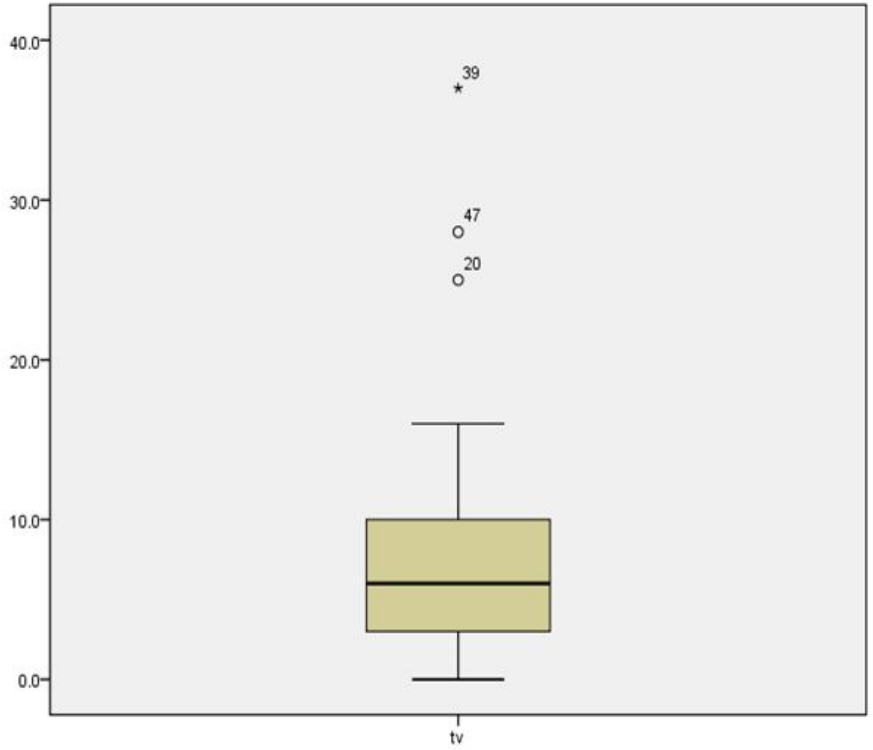
Minimum, LQ, Median, UQ, Maximum

except for outliers identified separately

Outlier = observation falling

below LQ – 1.5(IQR)

or above UQ + 1.5(IQR)



Skewness Kurtosis

Skewness measures departure from symmetry and is usually characterized as being left or right skewed.

Kurtosis measures “peakedness” of a distribution and comes in three forms, platykurtic, Mesokurtic and leptokurtic.

Skewness

$$Skewness = \frac{\bar{x} - median}{s}$$

Pearson's Skewness Coefficient

Fisher's Measure of Skewness has complicated formula but most software packages compute it.

Fisher's Skewness

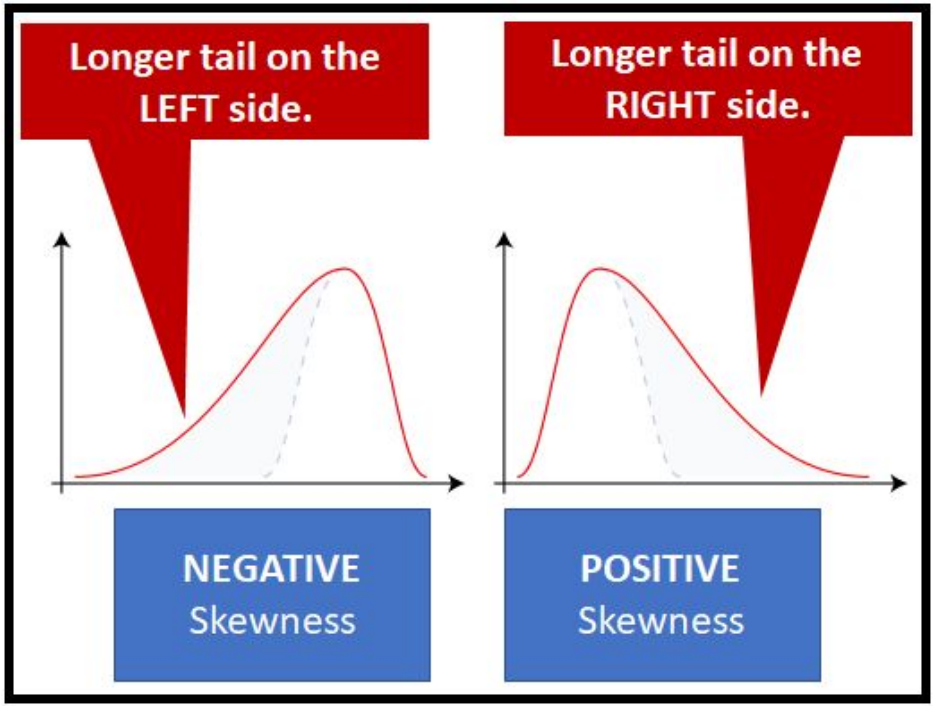
> 1.00 moderate right skewness

> 2.00 severe right skewness

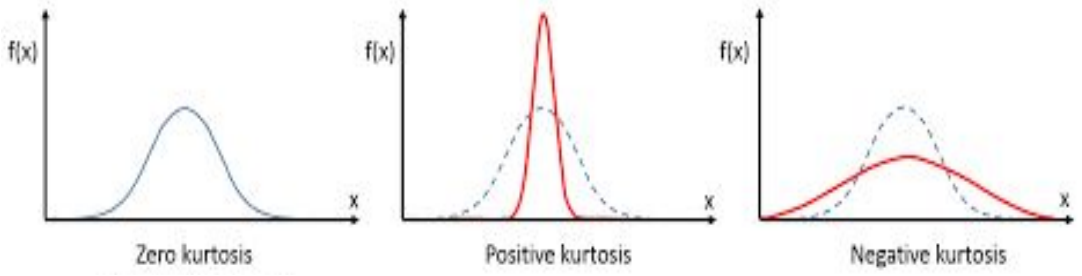
Fisher's Skewness

< -1.00 moderate left skewness

< -2.00 severe right skewness



Kurtosis



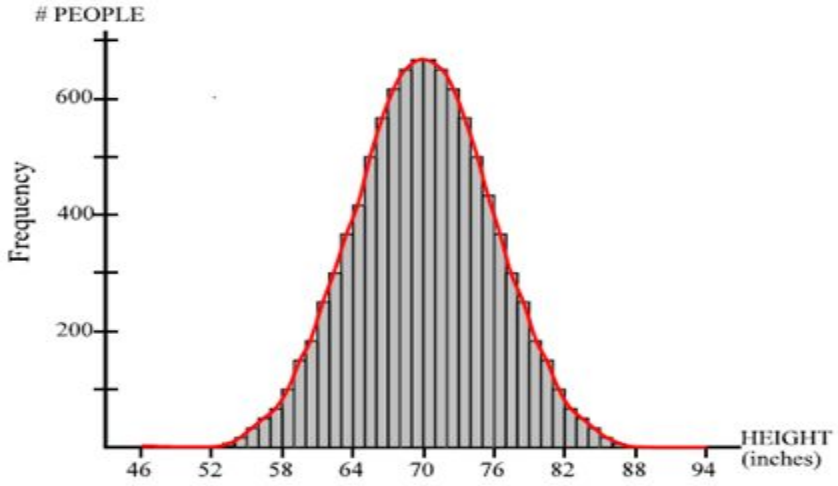
Measured the peakedness of the distribution

Normal distribution has Kurtosis = 0.

Leptokurtic distributions are more peaked than normal with fatter tails, Kurtosis > 0

Platykurtic distributions are less peaked (squashed normal) than normal, Kurtosis < 0

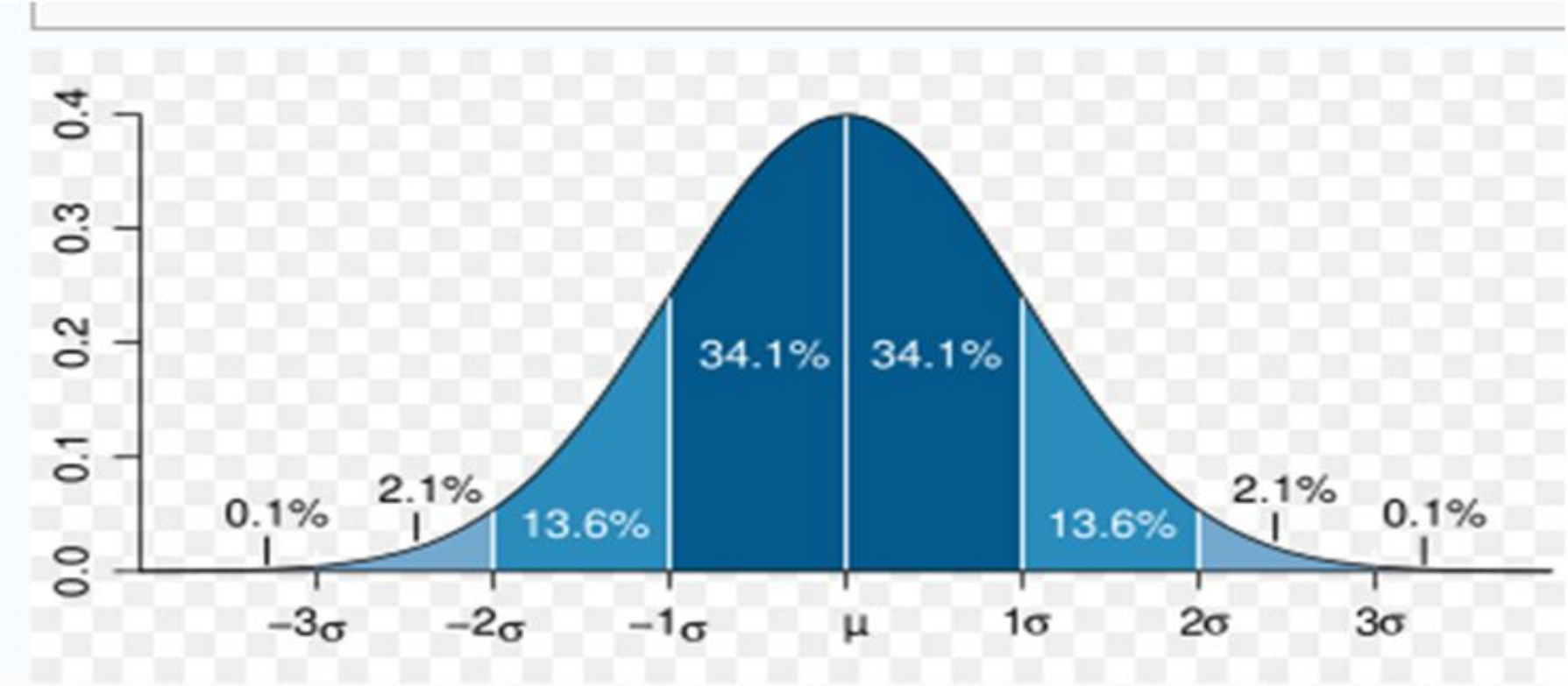
Normal Distribution



- Skewness = 0
- Kurtosis = 0

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

Normal Distribution



Test of skewness kurtosis

$$Z = \frac{\text{Skewness}}{\text{SE(Skewness)}}.$$

$$Z = \frac{\text{Kurtosis}}{\text{SE(Kurtosis)}}.$$

Compare with 1.96 for 5% level of significance

Normality Test

There are several different tests that can be used to test the following hypotheses:

Ho: The distribution is normal
H1: The distribution is NOT normal

Common tests of normality include:

Shapiro-Wilk **Kolmogorov-Smirnov**

Anderson-Darling **Lillefor's**

What if the data is not normal

• Transform the dependent variable (repeating the normality checks on the transformed data): Common transformations include taking the log or square root of the dependent variable.
• Use a non-parametric test: Non-parametric tests are often called distribution free tests and can be used instead of their parametric equivalent

Parametric test	What to check for normality	Non-parametric test
Independent t-test	Dependent variable by group	Mann-Whitney test
Paired t-test	Paired differences	Wilcoxon signed rank test
One-way ANOVA	Residuals/ dependent variable by group	Kruskal-Wallis test
Repeated measures ANOVA	Residuals at each time point	Friedman test
Pearson's correlation coefficient	Both variables should be normally distributed	Spearman's correlation coefficient
Simple linear regression	Residuals	N/A