Notes on Applied Statistics

Contents

Introduction to Statistics

- Basic Concepts and Definitions
- Data and Variables

The word Statistics is derived from Latin word status meaning “**state**” . Early uses of statistics involved compilation of data and graphs describing various aspects of the state or country. The word statistics has two basic meanings. We sometimes use this word when referring to actual numbers derived from data and the other refers to as a method of analysis

Overarching Definition

- **Statistics**:=deals with{ collection(data), presentation(data), analysis(data), interpretation(data)}
  - **Collection** - gathering of information
  - **Organization** or **presentation** - summarizing data
  - **Analysis** - describing the data by statistical methods or procedures
  - **Interpretation** - making conclusions on the analysed data

- **Inferential Statistics** - make inferences about the sample
  - Generalizes from samples to population
  - Performs estimations and hypothesis tests
  - Determines the relationships among variables
  - Making predictions
  - Uses probability

- **Descriptive Statistics** - describe a situation
  - Collection of data
  - Organization of data
  - Summarization of data
  - Presentation of data

Why study statistics?

- You have to be a data literate professional
- Statistics is **basic to research**: designing experiments, collecting-, organizing-, analysing-, and summarizing data to make reliable predictions or forecasts.
- You can also use the knowledge gained from studying statistics to become better consumers and citizens

Basic Concepts and Definitions

- **Variable** - characteristic or attribute that can assume different values
- **Data** - the value of the variable
- **Random data** - values of the variables are determined by chance
- **Data set** - collection of data values
- **Data value (datum)** - each value in the data set
- **Population** - all subjects that are being studied
  - **Parameter -**numerical summary or any measurement coming from a **population**
- **Sample** - group of subject selected from a **population**
  - **Statistic** - measure of the **sample**

Data and Variables

- **Quantitative** - categorical
  - Data
  - Variable

- **Qualitative** - numerical
  - Data
  - Variable

- **Dependent variable** - a variable that is **influenced by** another variable
- **Independent variable** - one that affects or **influences** another variable

Levels of Measurement

- **Nominal** - names
- **Ordinal -** categories in scale e.g. ranking
- **Interval -** comparison between the numerical differences are meaningful but not the ratio of the measurements
- **Ratio -** one can compare both the differences between measurements of the variable and the ratio of the measurements meaningfully

Data Collection

- Methods of Data Collection
- Sampling Technique

Data - facts and statistics collected together for reference or analysis.

Types of Data:

- **Primary** data - collected from an original source
- **Secondary** data - collected from published or unpublished sources.

Methods of Data Collection

1. **Direct or Interview Method** - the researcher has a direct contact with the interviewee. The researcher obtains the information needed by asking questions and inquiries from the interviewee. This method gives precise and consistent information because clarifications can be made.
2. **Indirect or Questionnaire Method** - make use of a written questionnaire. The researcher cannot expect that all distributed questionnaires will be retrieved because some respondents simply ignore the questionnaires. In addition, clarification cannot be made if the respondent does not understand the question.
3. **Registration Method** - collecting data is governed by laws.
4. **Experimental Method** - usually used to find out the cause and effect relationships

Sampling Techniques

- Determining the sample size - typically we use samples of the population and not the population itself because of the tremendous cost of using the entire population. To determine the sample size from a given population size, the **Slovin's Formula** is used: $n = \frac{N}{1+Ne^2}$; where n= sample size, N = population size, e = margin of error.

- We want unbiased samples: each interesting features should have an unbiased

- **Simple Random Sampling**

  - **Subjects are selected by random numbers**
    - **Can be done by labelling each possible subject and randomly picking the subject**

- **Systematic Sampling**
  - **Number each subject of the population then selecting every k-th subject**
  - **Lecturer: we can select samples from a moving population**

- **Stratified Sampling**
  - Done by dividing the population into groups called strata according to some characteristic that is important to the study, then sampling from each group

- **Cluster Sampling**
  - The population is divided into groups called clusters by some means. Then the researcher randomly selects some of these clusters and uses all members of the selected cluster as the subject of the samples.
  - There's no guarantee that every clusters is represented

Observational and Experimental Studies

- There are different ways to classify statistical studies:

- **Observational** - researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
- **Experimental** - the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.
-

Data Presentation

- The Frequency Distribution Table
- Graphical Presentation of Data

The Frequency Distribution Table

- **Raw data** - information obtained by observing values of a variable
- **Qualitative Data** - obtained by conserving values of a qualitative variable
- **Quantitative Data** - obtained by conserving values of a quantitative variable
  - **Discrete Data**
  - **Continuous Data**

Methods of Data Presentation

- Data can be classified as grouped or ungrouped

1. **Ungrouped data** - not organized or if arranged could only be from highest to lowest or lowest to highest
2. **Grouped data** - are data that are organized and arranged into different classes or categories

Textual Method (ungrouped)

- Ungrouped data can be presented in paragraph form.
- Involves enumerating the important characteristics, giving emphasis on significant figures and identifying important features of the data

Tabular Method (grouped)

- **Table heading** - consists of the table number and the title
- **Column header** - describes data in each column
- **Row classifier** - shows the classes or categories
- **Body** - main part of the table
- **Source note** - placed below when the data written are not original

Frequency Distribution - most commonly used by tabular method

Frequency Distribution

- Organization of raw data in table form, using classes and frequencies
- For qualitative data lists all categories and the number of elements that belong to each of the categories.
- For quantitative data, the data is grouped according to some numerical of quantitative characteristics
  - **Class limits** - endpoints of a class interval
  - **Upper class limit** - represents the largest data value that can be included in the class.
  - **Lower class limit** - represents the lowest data value that can be included in the class.
  - **Class boundaries** - used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 30 and 31
  - **Lower boundary** - Lower limit - 0.5
  - **Upper boundary** - Upper limit +0.5
  - **Class width** - the difference between the boundaries for any class i.e. i=upper boundary - lower boundary or i=(upper limit-lower limit)+1
  - **Class mark** - the midpoint of the class

Frequency Distribution Table (FDT) - a statistical table showing the frequency or number of observations contained in each of the defined classes or categories.

Relative frequency - a category is obtained by dividing the frequency (f) for a category by the sum of all the frequencies (n). They are commonly expressed as percentages.

To **construct a frequency distribution**, follow these rules:

1) There should be between 5 and 20 classes

   a) Although there are no hard and fast rule for the number of classes contained in a frequency distribution, it is of utmost importance to have enough classes to present a clear description of the collected data

2) It is preferable but not absolutely necessary that the class width be an odd number

   a) Ensures the midpoint of each class has the same place value as the data

3) The classes must be mutually exclusive - since they have no overlapping class limits

4) The classes must be continuous

a) There must be no gaps in a frequency distribution

b) The only exception occurs when the class with a zero frequency is the first of last class since a class with a zero frequency on either end can be omitted

5) The classes must be exhaustive.

a) There should be enough classes to accommodate all the data.

6) The classes must be equal in width

a) Avoids the distorted view of the data

Constructing an FDT

1) Determine the classes

a) Find the highest and lowest values

b) Find the range $R = highest\ value - lowest\ value$

c) Select the number of classes desired (k)

d) Find the width by dividing the range by the number of classes and rounding up. $i = \frac{R}{k}$

e) Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits

f) Find the upper class limits

g) Find the boundaries

2) Tally the data

3) Find the frequencies from the tallies

Cumulative Frequency

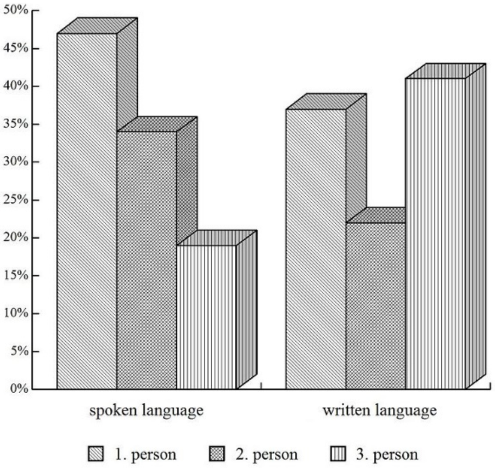
- Less than cumulative frequency (<cf) - total number of observations less than the upper boundary of a class interval
- greater than cumulative frequency (>cf) – total number of observations greater than the lower boundary of a class interval

Graphical Presentation of Data

- After you have organized the data into a frequency distribution, you can present them in a graphical form.

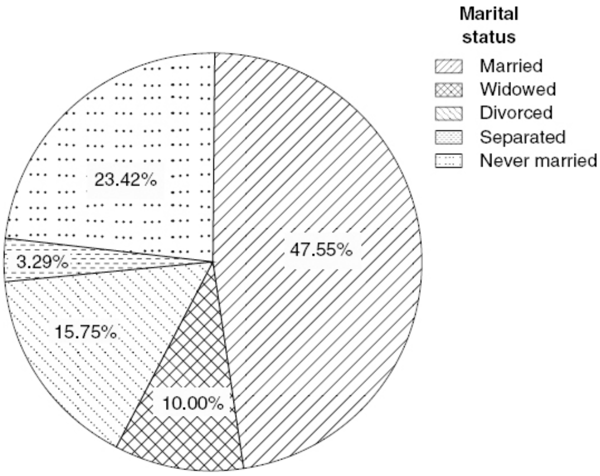
The Bar Graph

- Qualitative data

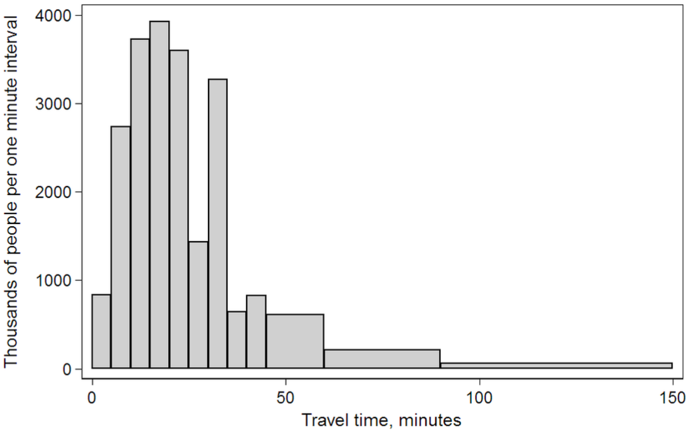


Pie Chart

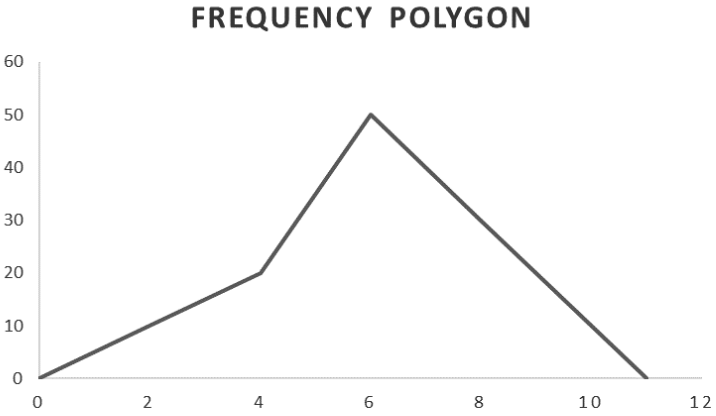
- Qualitative data



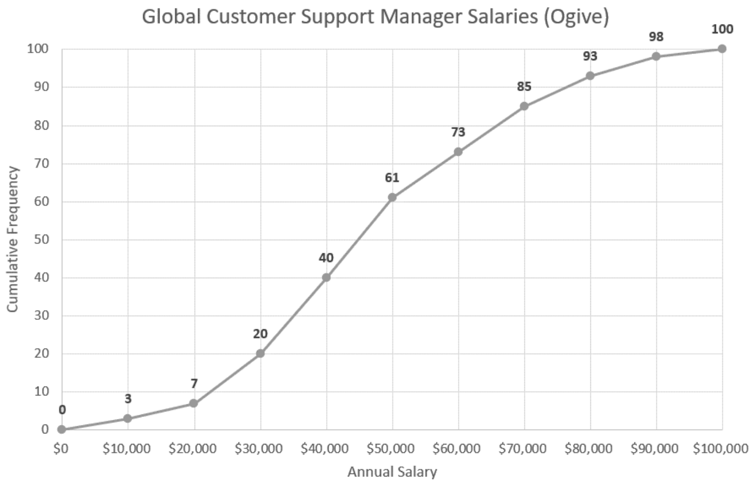
Histogram



Frequency Polygon



Ogive


Global Customer Support Manager Salaries (Ogive)
Cumulative Frequency
Annual Salary
0
10
20
30
40
50
60
70
80
90
100
$0
$10,000
$20,000
$30,000
$40,000
$50,000
$60,000
$70,000
$80,000
$90,000
$100,000
0
3
7
20
40
61
73
85
93
98
100

Measures of Central Tendency

- Measures of Central Tendency (MCT) of Ungrouped Data
- Measures of Central Tendency (MCT) for Grouped Data

Measures of Central Tendency Ungrouped

- MCT gives a single value that acts as a representative or average of the values of all the outcomes of your data set; describes the center of the distribution and represents the entire distribution to identify the single value that is the best representative for the entire set of data

Mean

- The most commonly used to measure central tendency; requires scores (or data types) that are numerical values measured on **an interval or ratio scale**.
  - ○ Obtained by $\overline{x} = \frac{Total\ SUM}{number\ of\ dataset} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{x} sample\ mean$, $\mu population\ mean$
  - ○ For Discrete quantitative data, tabulated in a frequency table, then if possible observations are $\{x_1, x_2, \ldots, x_k\}$ and these occur with frequencies $\{f_1, f_2, \ldots, f_k\}$ respectively, so that $\sum f_n = n$, then the mean is: $\overline{x} = \frac{1}{n}\sum x_i f_i$.
  - ○ For continuous data, the sample mean should be calculated from the original data if this is known.
    - ∗ If it is tabulated in a frequency table, and the original data is not known, the sample mean can be **estimated** by assuming that all observations in a given interval occurred at the midpoint of that interval i.e. the **class mark.** So, if the class marks of the interval are $\{m_1, m_2, \ldots m_k\}$ and the corresponding frequencies are $\{f_1, f_2, \ldots, f_k\}$ then the sample mean can be approximated by:

$$\overline{x} = \frac{1}{n}\sum m_i f_i$$

- Weighted Mean
  - ○ If k quantities $\{x_1, x_2, \ldots, x_k\}$ have weights $\{w_1, w_2, \ldots, w_k\}$ respectively, where the weights represents measures of relative importance, then the weighted mean is:

$$\overline{x_k} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum w_i}$$

- Combined mean
  - If k finite groups having $\{n_1, n_2, \ldots n_k$ measurements respectively, have means $\{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_k}\}$, the combined mean is:

$$\overline{x_c} = \frac{\sum_{i=1}^{k} n_i \; \overline{x_i}}{\sum n_i}$$

Median

- Data type: ranked
    - Obtained by
        - If even $\widetilde{x} = \frac{x_n + x_{n+1}}{2}$
        - If odd $\widetilde{x} = x_{\frac{n+1}{2}}$

Mode

- Most frequent value in a data set

Measures of Central Tendency of Grouped Data

Mean

$$\overline{x} = \frac{1}{n} \sum f x_m$$

where
f = *class frequency;*
x_m = class mark;
n=total number of observations

Median

$$\widetilde{x} = L + \left(\frac{\left(\frac{n}{2} - S_b\right)}{f_m} \right) i$$

where:
f_m = frequency of the median class;
x_i = class mark;
n = total number of observations
L = Lower boundary of the median class
S_b =< cf of the class before Median class

Mode

$$\hat{x} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

where:
Δ_1 = difference in the frequencies of the modal class and the next lower class
Δ_2 = differenc in the frequencies of the modal class and the next higher class
i = size of class interval
L = lower boundary of the modal class

Measures of Dispersion, Position, and Shapes

- Measures of Dispersion
- Measures of Position
- Measures of Shape

Measures of Dispersion

- Measures of variability or dispersion are measures of average distance of each observation from the center of the distribution;
- They summarize and describe the extent to which scores in a distribution differ from each other
- Tell us how spread out the scores are
  - A small dispersion would indicate that the data are:
    - Clustered closely around the mean
    - More homogeneous
    - Less variable
    - More consistent
    - More uniformly distributed

Classifications

Measures of absolute dispersion

- Are expressed in the units of the general observations; they cannot be used to compere variations of two data sets when the average of these sets differ a lot in value or when the observations differ in units of measurements

Absolute Dispersion:

- Range: difference between the highest and the lowest values; **simplest but most unreliable measure of dispersion** $Range = HV - LV$ where HV is the highest value, and LV is the lowest value

- Variance: the average of the squared deviation of each score from the **mean.**

  Population variance (raw data):

  $$\sigma^2 = \frac{1}{N}\sum f\left(x-\mu\right)^2$$

  Population Variance (grouped data):

  $$\sigma^2 = \frac{1}{N}\sum f\left(x_m-\mu\right)^2$$

  Sample variance (raw data):

  $$s^2 = \frac{1}{n-1}\sum \left(x-\overline{x}\right)^2$$

  Sample Variance (grouped data):

  $$s^2 = \frac{1}{n-1}\sum f\left(x_m-\overline{x}\right)^2$$

- Standard Deviation: the square root of the variance

Population Standard Deviation (raw data)

$$\sigma = \sqrt{\frac{1}{n}\sum \left(x-\mu\right)^2}$$

Sample Standard Deviation (raw data)

$$s = \sqrt{\frac{1}{n-1}\sum \left(x-\overline{x}\right)^2}$$

Or

$$s = \sqrt{\frac{n\sum fx^2 - \left(\sum fx\right)^2}{n\left(n-1\right)}}$$

Population Standard Deviation (grouped data)

$$\sigma = \sqrt{\frac{1}{N}\sum f\left(x_m-\mu\right)^2}$$

Sample Standard Deviation (grouped data)

$$s = \sqrt{\frac{1}{n-1}\sum f\left(x_m - \overline{x}\right)^2}$$

Measures of relative dispersion

- Are unit-less measures and are **used when one wishes to compare the scatter of the distribution with another distribution**

Relative Dispersion

- The Coefficient of Variation

Is the ration of the standard deviation to the man and is usually expressed in percentage; used to compare variability of two or more sets of data even when they are expressed in different units of measurements: $cv = \frac{1}{\overline{x}}$

1) Chebyshev's Theorem: at least the fraction of $1 - \frac{1}{k^2}$ of measurements of any set of data must lie within k standard deviations of the mean

Example:

- If the IQ's of a random sample of 1080 students at a large university have a mean score of 120 and standard deviation of 8,
  - Determine the interval containing at least 810 of the IQ's in the sample

    Solution: note that $810/1080 = 0.75$. Hence we want to determine the interval to which 75% of the IQ scores lie. By Chebyshev's theorem, 75% lie within $\mu \pm 2\sigma$, i.e. between

    $$\mu - 2\sigma = 120 - 2\left(8\right) = 104; \mu + 2\sigma = 120 + 2\left(8\right) = 136$$

    Therefore, at least 810 of the respondents score are $104 - 136$.

- In what range can we be sure that no more than 120 of scores fall?

  - Solution: Results from a tells at least 810 scored $104 - 136$ .Hence the remaining not more than 120 must have scored outside this interval, i.e. less than 104 or greater than 136

Measures of Position

- Used for locating a position of non-central piece of data relative to the entire set of data

z-score

- measures how many standard deviation an observation is above or below the mean

- Population: $z = (x - \mu)/\sigma$
- Sample: $z = (x - x\ )/s$

Fractiles or quantiles

- specific fraction or percentage of the observations in a given set must fall

1) Percentiles

    a) Ungrouped:

        (i) Arrange the data from lowest to highest

        (ii) Substitute into the formula $c = \frac{np}{100}$ where n=total number of values; p=percentile rank

        (iii) Either:

            (1) If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded up value

            (2) If c is a whole number, use the value halfway between the c-th and $(C + 1)$ st values when coming up from the lowest value

- Grouped:
  - The kth percentile on the class interval with at least $\frac{kn}{100}$ cumulative frequency. The k-th percentile is given by

$$P_k = L + \frac{\left(\frac{kn}{100} - S_b\right) i}{f_p}$$

Where:

- f_p = frequency of the percentile class
- n = total number of observations
- i = size of class interval
- L = lower boundary of the percentile class
- S_b =<cf of the class before percentile class

1) Quartiles

   a) Definition $Q_1 = P_{25}; Q_2 = P_{50}$

1) Deciles

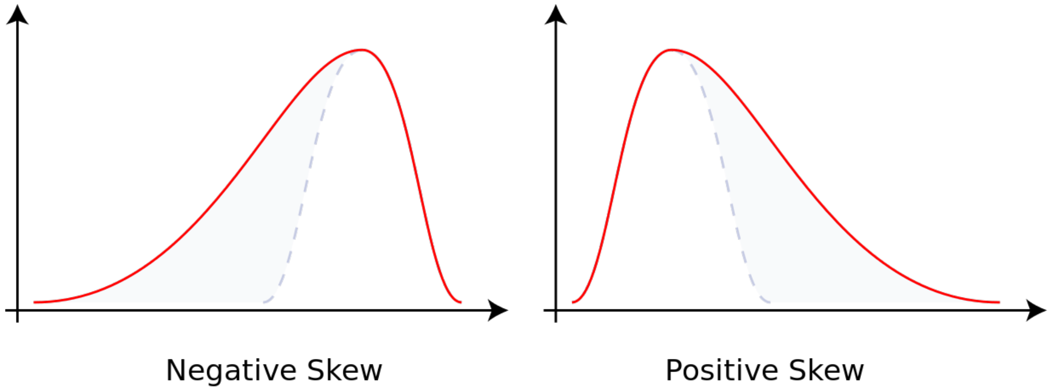
   a) Definition $D_1 = P_{10}; D_2 = P_{20}$

Measures of Shapes

Skewness

- Refers to the degree of symmetry and asymmetry of a distribution;
- The normal distribution is bell-shaped and symmetric through the mean; it has the property of mean=median=mode

1) **Skewed to the left** (negatively skewed)- mean is less than the median; the bulk of the distribution is on the right
2) **Skewed to the right** (positively skewed) - mean is greater than median; the bulk of the distribution is on the left

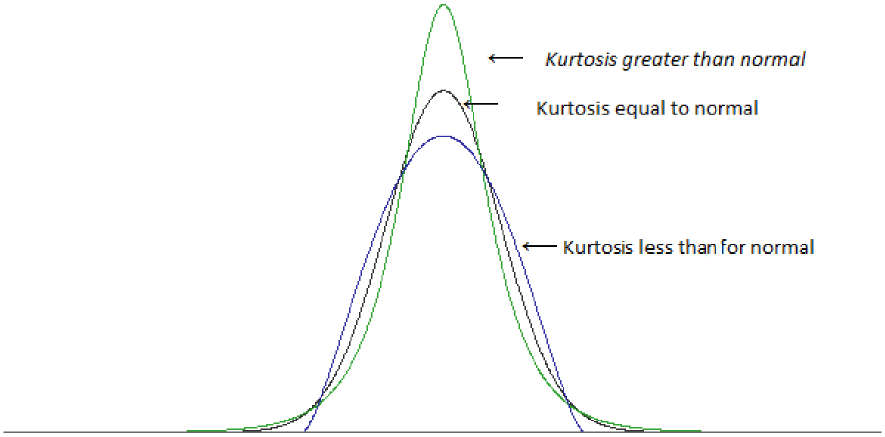


The extend of skewness can be obtained by getting the coefficient of skewness

$$SK = \frac{3\left(\overline{x} - \widetilde{x}\right)}{s} | s = standard\ dev$$

- If $SK = 0$, the distribution is normal,
- If $SK < 0$, the distribution is to the left,
- If $SK > 0$, the distribution is to the right

Kurtosis



- Mesokurtic - is a normal distribution
- Leptokurtic - more peaked than the normal distribution
- Platykurtic - flatter than the normal distribution

Ungrouped

$$Ku = \frac{\sum (x - \overline{x})^4}{ns^4}$$

Grouped

$$Ku = \frac{\sum f\,(x_m - \overline{x})^4}{ns^4}$$

- If $Ku = 3$, mesokurtic
- If $Ku < 3$, leptokurtic
- If $Ku > 3$, platykurtic