# Statistical Concepts

## Session I

Vik Gopal

Jan 13, 2017

# Exploratory Data Analysis



- In this session, we shall cover mostly numerical summaries of data.
- Since we observed a fair number of tables in the data, we shall also provide some guidelines on summarising them, and on comparing proportions.
- We shall introduce some graphical displays, and some suggestions on what to say when we see them.
- Some examples in this session are slightly more pedagogical than practical; others have been taken directly from UNESCAP reports.

# Variables



**Definition 1 (Variable)**

A **variable** is any characteristic observed in a study.

- The term *variable* highlights that the data values vary, either from year to year, from region to region, or just over repeated experiments.
- If there were no variability in data, we would not need statistics.
- If we wanted to investigate inflation, we could record the following variables and study their change over time:
  - Cost of a loaf of bread every year.
  - Cost of a meal at McDonald's every year.
- Other examples are the number of tuberculosis cases in a year, the literacy rates in different countries, etc.

# Categorical and Quantitative Variables



**Definition 2 (Categorical and Quantitative Variables)**

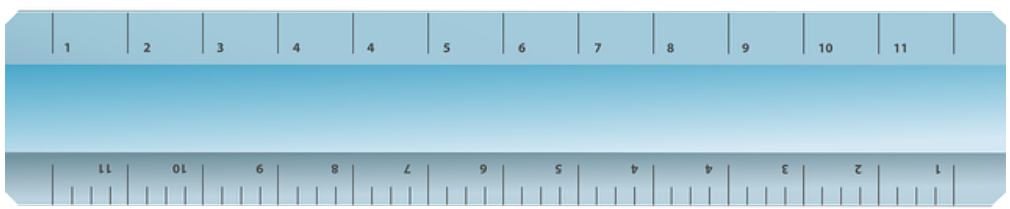
1. A variable is called **categorical** if each observation belongs to one of a set of categories.
2. A variable is called **quantitative** if observations on it take on numerical values that represent different magnitudes of the variable.

- Examples of categorical variables are gender, religion, race, migrant status, etc.
- Examples of quantitative variables are age, GDP, income, etc.

# Distinguishing Between Quantitative and Categorical Data



- You can identify whether a data is categorical or not simply by asking if there is a meaningful distance between any two points in the data. If such a distance is meaningful, then you have quantitative data.
- For instance, it makes sense to compute the difference in blood pressure between subjects in a study.
- However, it does not make sense to consider the mathematical operation ("literate" - "illiterate").
- Being able to classify our data as categorical or quantitative is not a huge deal in itself, but it does inform us about which summaries and graphics are appropriate.



# Quantitative Variables



Quantitative variables can be further sub-divided.

**Definition 3 (Discrete and Continuous Variables)**

A quantitative variable is **discrete** if its possible values form a set of separate numbers such as 0, 1, 2, 3, ...
A quantitative variable is **continuous** if its possible values form an interval.

- Discrete variables are usually counts.
- Examples are
  - Number of children in a home
  - Number of TB cases in a country
- Continuous variables have a continuum of infinitely many possible values.
- Examples of continuous variables are
  - Time taken until Gini Index doubles.
  - Life expectancy of Japanese.

# Categorical Variables



Categorical variables can also be sub-divided.

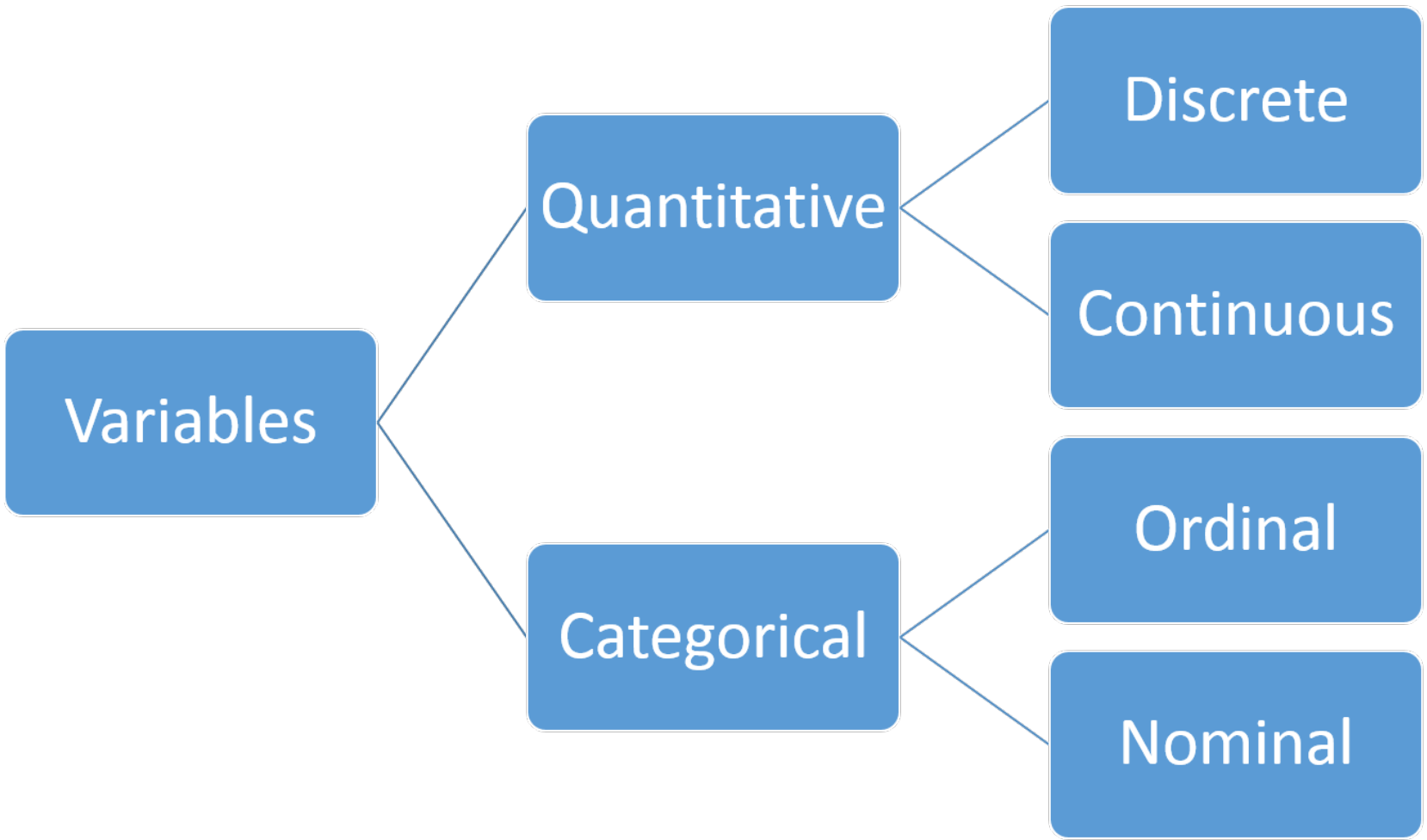
**Definition 4 (Nominal and Ordinal Variables)**

A categorical variable is **ordinal** if the observations can be ordered, but do not have specific quantitative values.
A categorical variable is **nominal** if the observations can be classified into categories, but the categories have no specific ordering.

- Examples of ordinal random variables are
  - Level of education - lower primary, upper primary, etc.
  - Answers to survey questions - strongly agree, agree, etc.
- Examples of nominal random variables are
  - Gender
  - Ethnic group
  - Disease status

# Overview of Variable Types





# Summarising Data with Tables



- Usually, the first step in summarising data is to look at the possible values, and count how often each one occurs.
- For categorical variables, the number of times each category turns up is counted and displayed in a table.
- The category with the highest frequency is the **modal category**.
- Note that it is possible to coerce quantitative data into categories and then display them in a table too.

# Frequency Tables



Definition 5 (Frequency Tables)

1. A **frequency table** is a listing of possible values for a variable, together with the number of observations for each value.
2. The **proportion** of observations that fall in a certain category is the count of observations in that category divided by the total number of observations.
3. The **percentage** is the proportion multiplied by 100.
4. Proportions and percentages are also known as **relative frequencies**.

# Breast Cancer Data



Example 6 (Breast Cancer)

- In 1990, 1200 post-menopausal women were recruited for a study to investigate the effect of a Post-Menopausal Hormone (PMH) on the incidence of breast cancer.
- The women were continually examined until 2000 for the incidence of breast cancer.
- Consider the following variables:

| | |
|---|---|
| Age: | The age of the woman in 1990. *(Quantitative, continuous)* |
| PMH: | Whether or not the woman used PMH. *(Categorical, nominal)* |
| BBD: | The incidence or otherwise of benign breast cancer. *(Categorical, nominal)* |
| BMI: | The BMI (Body Mass Index) of the woman in 1990. *(Quantitative, continuous)* |

# Frequency Tables for Categorical Variables



Here is a summary of the incidence of BBD over the 10 years:

**Breast Cancer Incidence**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Absent | 803 | 66.9 | 66.9 | 66.9 |
| | Present | 397 | 33.1 | 33.1 | 100.0 |
| | Total | 1200 | 100.0 | 100.0 | |

Similarly, here is a summary of the usage of PMH among the study participants:

**PMH User Status**

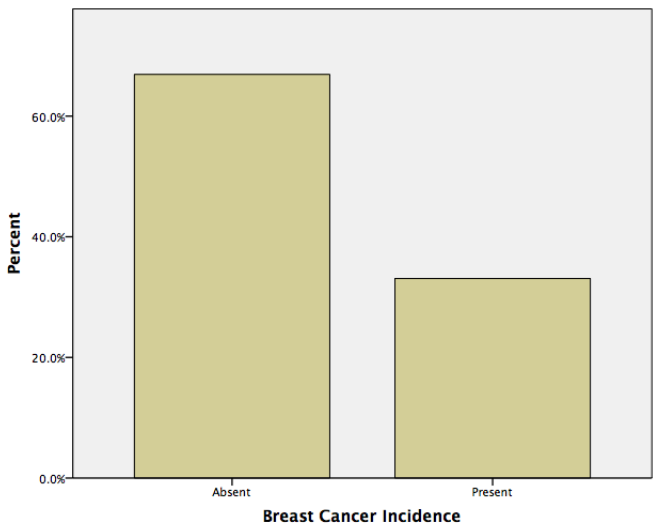
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Non-user | 1000 | 83.3 | 83.3 | 83.3 |
| | User | 200 | 16.7 | 16.7 | 100.0 |
| | Total | 1200 | 100.0 | 100.0 | |

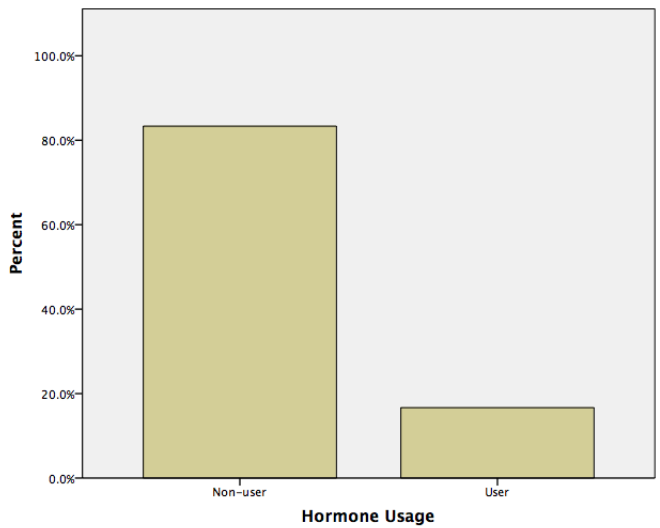
# Bar Plots



- A bar plot is a common way to display a single categorical variable.
- It consists of a vertical bar for each possible category that could occur, with the height proportional to the frequency of that particular category.
- We can think of a bar plot as a visual representation of a frequency table for a single variable.

Here are the bar plots for the categorical variables that were introduced on slide 15.



# What to Say



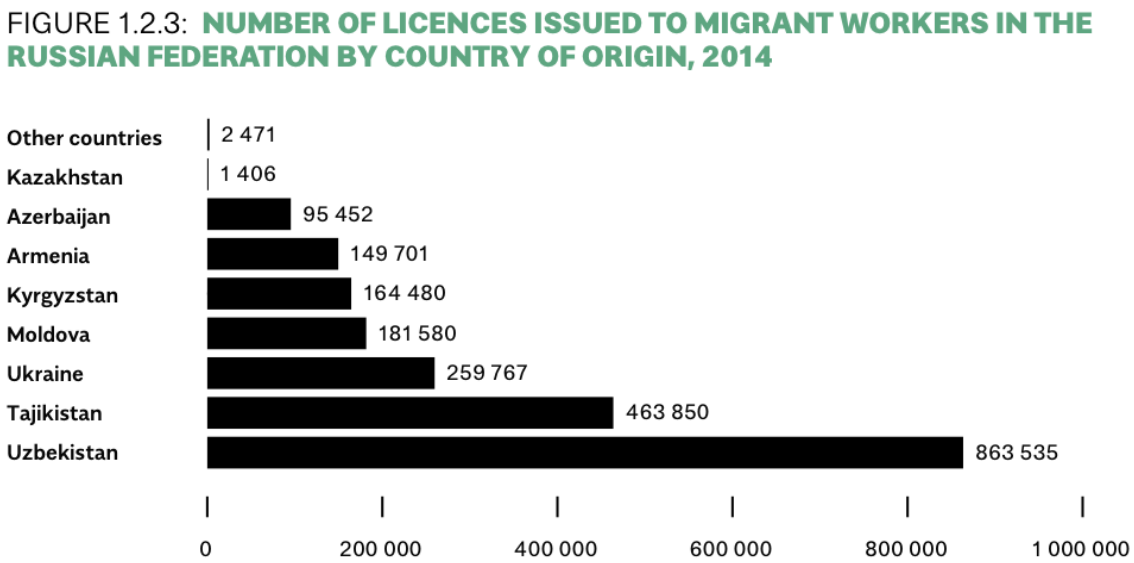
When asked to summarise a bar plot, typical things to mention would be:

- The modal category.
- The proportion for this category. If there are other categories that are of known interest, be sure to compare the proportion of the modal category to that category via their difference.
- Mention if there are groups of categories with high/low proportions. Such a pattern is usually easier to see visually.
- If there is an ordering to the categories, mention if there is any apparent trend in proportions.

# Migrant Licenses in Russian Federation



Example 7 (Migrant Licenses in Russian Federation, 2014)




- The above chart was taken from *The Role of Labour Migration in the Development of the Economy of the Russian Federation*, Figure 1.2.3.

# Summary of Migrant Licenses



- Ordering the categories by the number of licenses has made it easier to compare.
- Sometimes, it is more efficient to use percentages or proportions instead of the raw counts, in which case we could point out that
  - Uzbekistan, the modal country, took up 40% of the licenses.
  - The next highest country was Tajikistan, which only took up approximately half as many.
  - Tajikistan and Uzbekistan account for 60% of all licenses.
  - A personal feeling is that it would not have made much difference to lump Kazhakstan with "Other countries".

# Histograms



**Definition 8 (Histogram)**

A **histogram** is a graph that uses bars to portray the frequencies or relative frequencies of the possible outcomes for a quantitative variable.

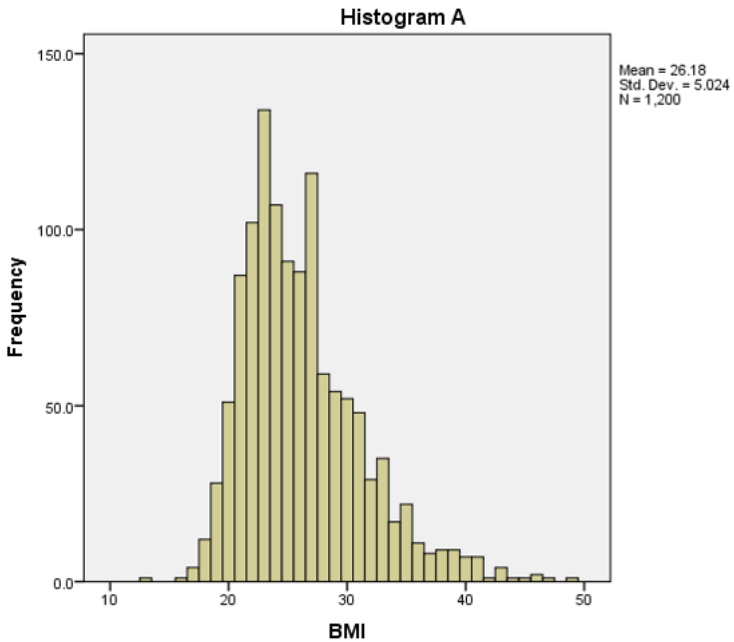
Given a sample, a histogram is constructed as follows:

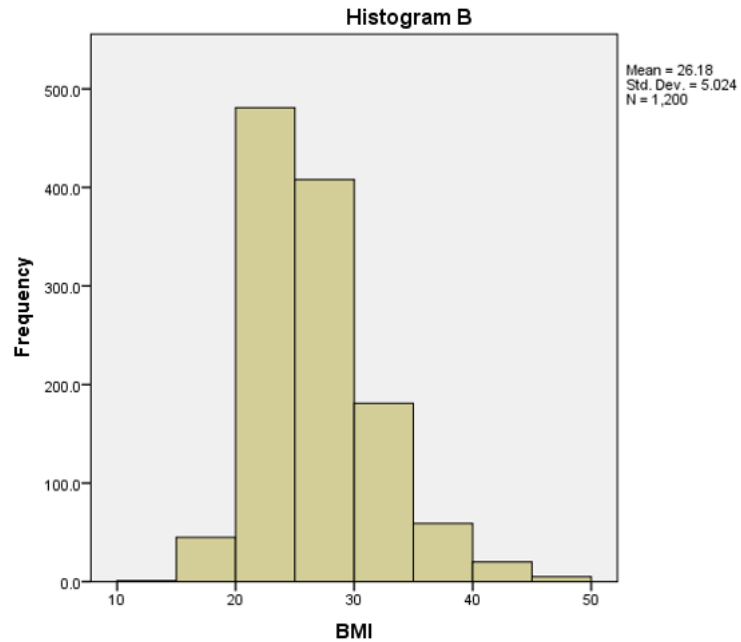
1. Divide the range of the data into intervals of equal width. For a discrete variable with only a few values, use the actual possible values.
2. Count the number of observations that fall within each interval, forming a frequency table.
3. Label the intervals on the $x$-axis, and draw a bar over each interval, that has height equal to its frequency or relative frequency.

# Histograms for BMI Variable



We construct histograms for the BMI data from slide 15.




- Histogram A was created using intervals of length 1, while Histogram B was created using intervals of length 5.
- From B, we can see that there were about 500 participants whose age was in the interval (20, 25].
- From A, we know that there about 50 participants whose age fell within (19.5, 20.5].

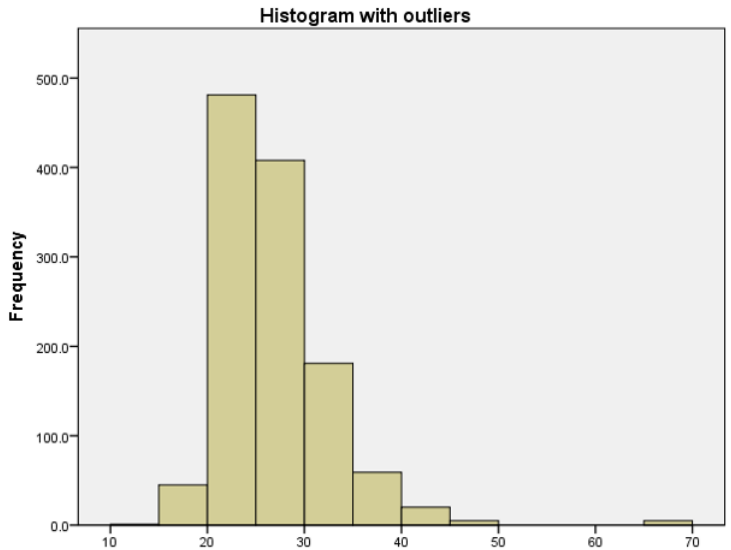
# What to Say



- Histograms are very similar to bar plots, which are visual representations of frequency tables.
- What do we look for in a histogram?
  - The overall pattern. Do the data cluster together, or is there a gap such that one or more observations deviate from the rest?
    - If they are stragglers, we would consider them to be outliers. We shall see how to identify them using boxplots in the next session.
    - If there are many of them, it could be indicative of a different population.
  - Do the data have a single mound? This is known as a **unimodal** distribution. Data with two are known as bimodal, and data with many mounds are referred to as **multimodal**.
  - Is the distribution **symmetric** or **skewed**?
- Let us investigate what the terms in bold mean in the next few slides.

# A Histogram With Outliers



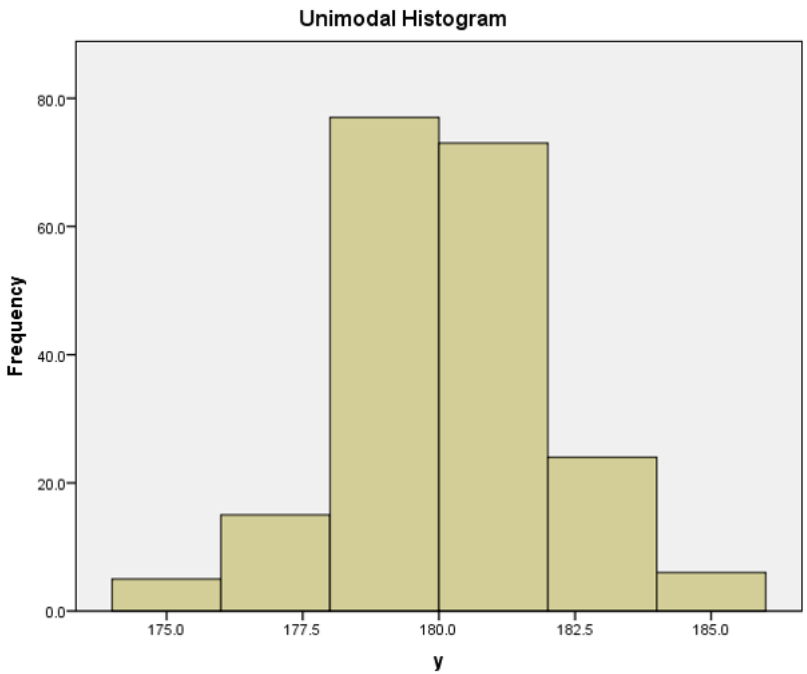


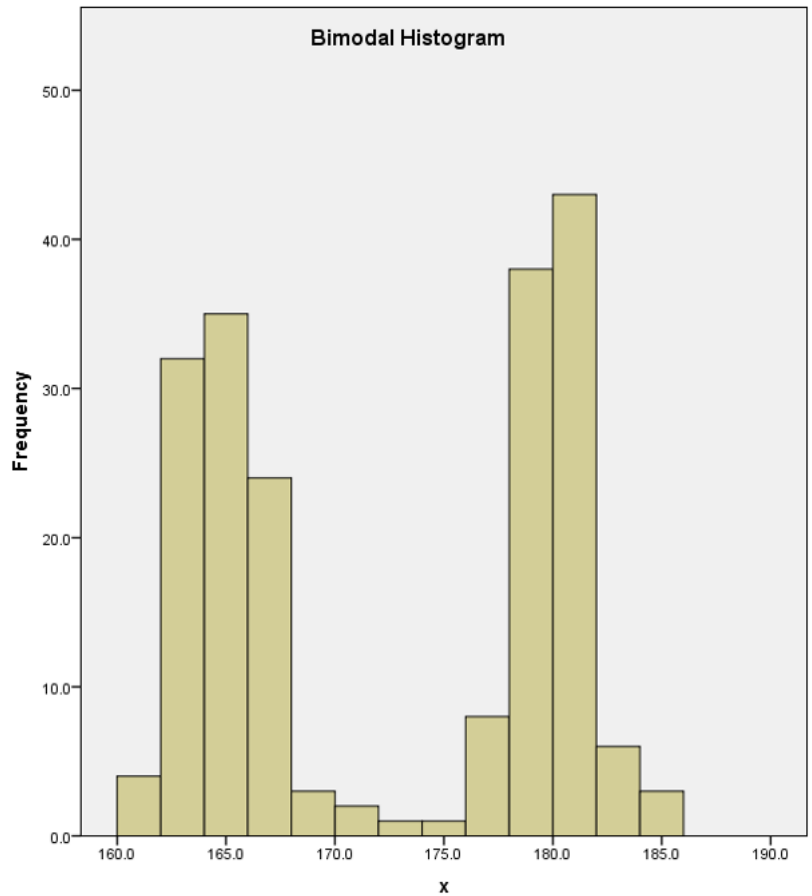

- Here's an example of a histogram with outliers (artificial data).
- The bar for those few observations are separate from the rest.

# Unimodal and Bimodal Histograms



Here are examples of histograms of unimodal and bimodal distributions (artificial data).



# Skewness



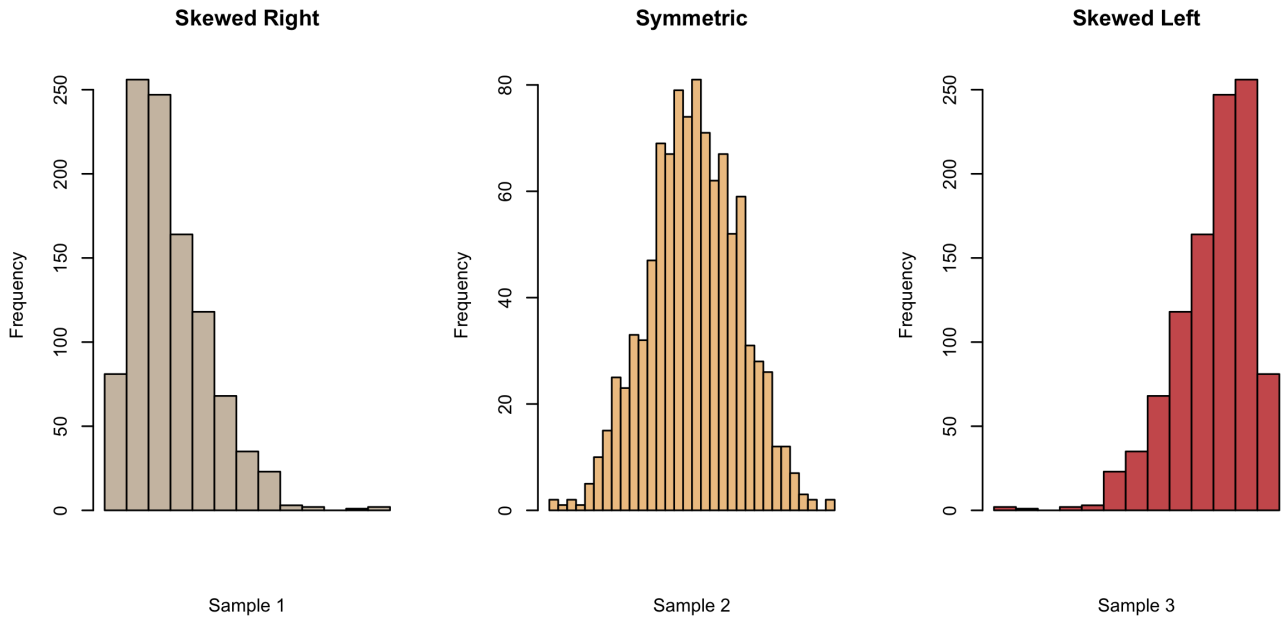
Definition 9 (Skew)

- To **skew** is to pull in one direction.
- A distribution is **skewed to the left** if the left tail is longer than the right tail.
- A distribution is **skewed to the right** if the right tail is longer than the left tail.

# Examples of Histograms of Skewed Data



Example 10 (Histograms of Skewed and Symmetric Datasets)



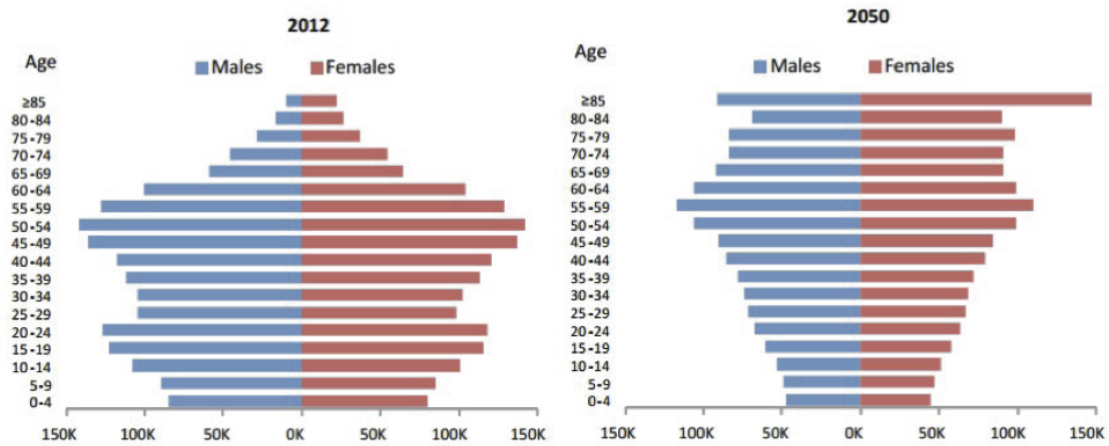

- Income is typically right-skewed. This means that there are few people who earn a huge amount. Most others earn much less.
- IQ is typically symmetric.
- Life-span is typically left-skewed.

# Changing Age Structure, Singapore



## Example 11 (Changing Age Structure in Singapore)

FIGURE 2. CHANGING AGE STRUCTURE, SINGAPORE, 2012, 2050




**Source:** National Population and Talent Division, Prime Minister's Office (2013).

- The above chart was taken from *Ageing Long Term Care Singapore*, Figure 2.

# Summary of Changing Age Structure



- We in fact have four histograms of age counts – males and females in 2012, and the projected counts in 2050.
- We start by comparing males in 2012 to males in 2050:
  - In 2012, the distribution of age is right-skewed and bimodal with modes at 50-54 and 20-24.
  - In 2050, the distribution is left-skewed, with a single mode at 55-59. In addition, there is a spike of individuals aged above 85; this could be because it aggregates people from more than one category.
  - It is striking to focus on the extreme categories - the 0-4 category has reduced by a third, while the 85- category seems to have a ten-fold increase.
- How does the change in females compare?
- Is there anything worth mentioning when comparing males to females directly?

# Describing the Center



- In trying to identify the center of a distribution, we are attempting to identify what a "typical" observation looks like.
- The two common measures that we shall learn about are the mean and the median.
- Bear in mind that reducing a sample of $n$ points to **one** number inevitably masks some details about the data.
- The mean and median report the *center* from different points of view, so be wary.

# Mean



**Definition 12 (Mean)**

The **mean** is the sum of all the observations, divided by the number of observations. It is denoted by $\bar{X}$, which is read as "X-bar".

- The mean is the balance point of the distribution.
- It is also referred to as the average.

# Mean Computation



**Dataset 1**

Consider the following observations of a quantitative variable.

$$-2.0,\ -1.0,\ 1.2,\ 1.8$$

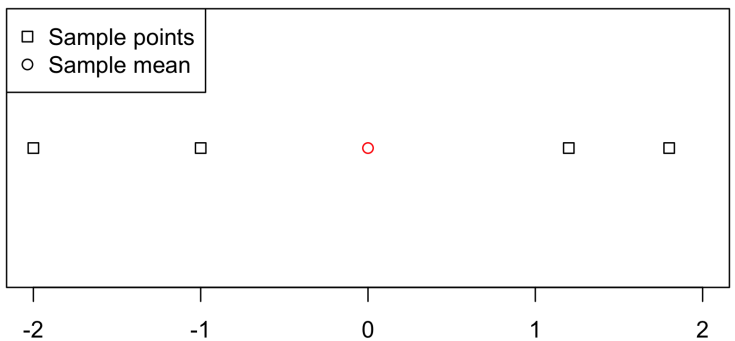
The mean is given by

$$\frac{(-2.0) + (-1.0) + 1.2 + 1.8}{4} = 0$$

# Visualising the Mean



Here is a plot of the observations, together with the sample mean.




The mean identifies the "center" of a set of values with the balance point.

# Mean Computation



**Dataset 2**

Now consider the following dataset:

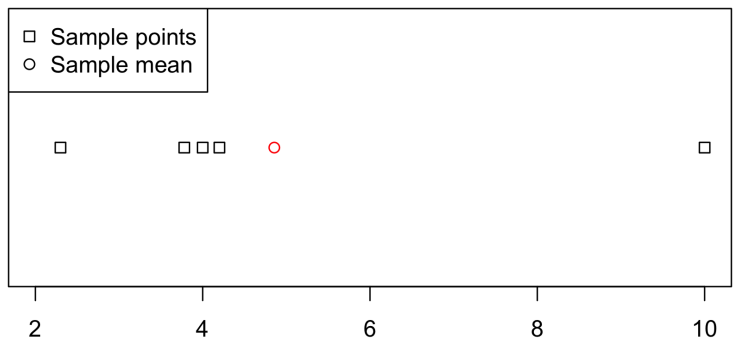
$$-4.0,\ 2.3,\ 3.78,\ 10.0,\ 4.2$$

The mean is given by

$$\frac{(-4.0) + 2.3 + 3.78 + 10.0 + 4.2}{5} = 4.856$$

# Visualising the Mean






Again, the mean identifies the "center" of a set of values with the balance point; it does not split the dataset into two groups of equal size.

# Median



**Definition 13 (Median)**

The **median** is the middle value of the observations when the observations are ordered from smallest to the largest. In a sample of $n$ observations, the median is defined to be

1. The $\left(\frac{n+1}{2}\right)$-th largest observation if $n$ is odd.
2. The average of the $\left(\frac{n}{2}\right)$-th and $\left(\frac{n}{2}+1\right)$-th largest observation if $n$ is even.

We shall denote the median by $X_{(0.5)}$.

- The median is the number that divides the sample points into two equal groups. Approximately half of the sample points will be less than the median, and the other half will be greater than the median.

# Median Computation



**Dataset 1**

Consider once again the following observations of a quantitative variable.

$$-2.0,\ -1.0,\ 1.2,\ 1.8$$

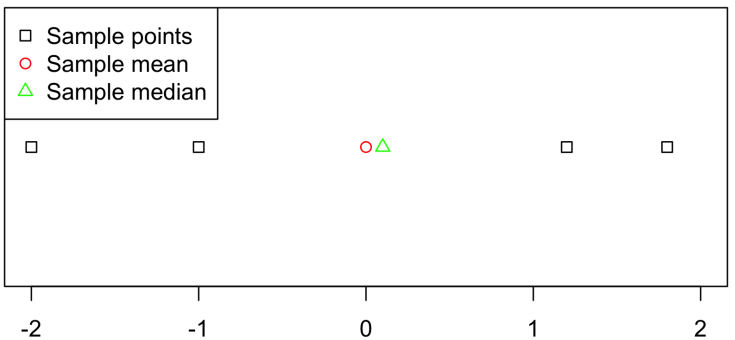
The mean was earlier computed to be 0.

The dataset is already ordered. Since there are 4 observations, the median is

$$\frac{-1.0 + 1.2}{2} = 0.1$$

# Median and Mean Visualisation






- There are two observations on either side of the median.
- In this dataset the mean and the median are close to each other.

# Median Computation



**Dataset 2**

Now consider the following dataset once more:

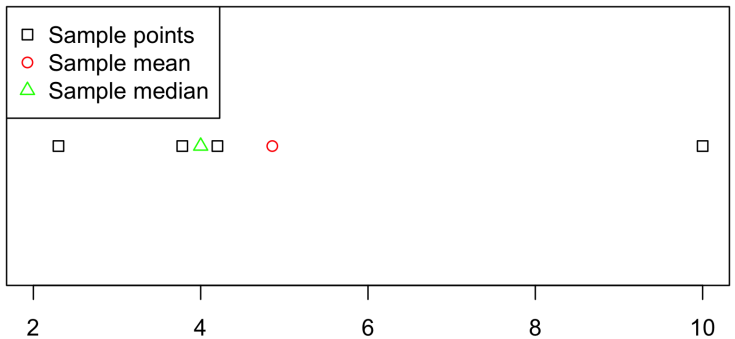
$$4.0,\ 2.3,\ 3.78,\ 10.0,\ 4.2$$

The mean was earlier found to be 4.856.

- The ordered dataset is
$$2.3,\ 3.78,\ 4.0,\ 4.2,\ 10.0$$
- Since there are 5 observations, the median is the 3rd point, which is 4.

# Mean and Median Visualisation






The sample median is in fact one of the sample points (since $n$ is odd). It "divides" the data in a more natural way than the sample mean. There are 2 sample points on either side of the median.

# Comparing the Mean and the Median



- The mean is sensitive to extreme observations, whereas the median is not.

Dataset 2

Consider this dataset once more:

$$4.0,\ 2.3,\ 3.78,\ 10.0,\ 4.2$$

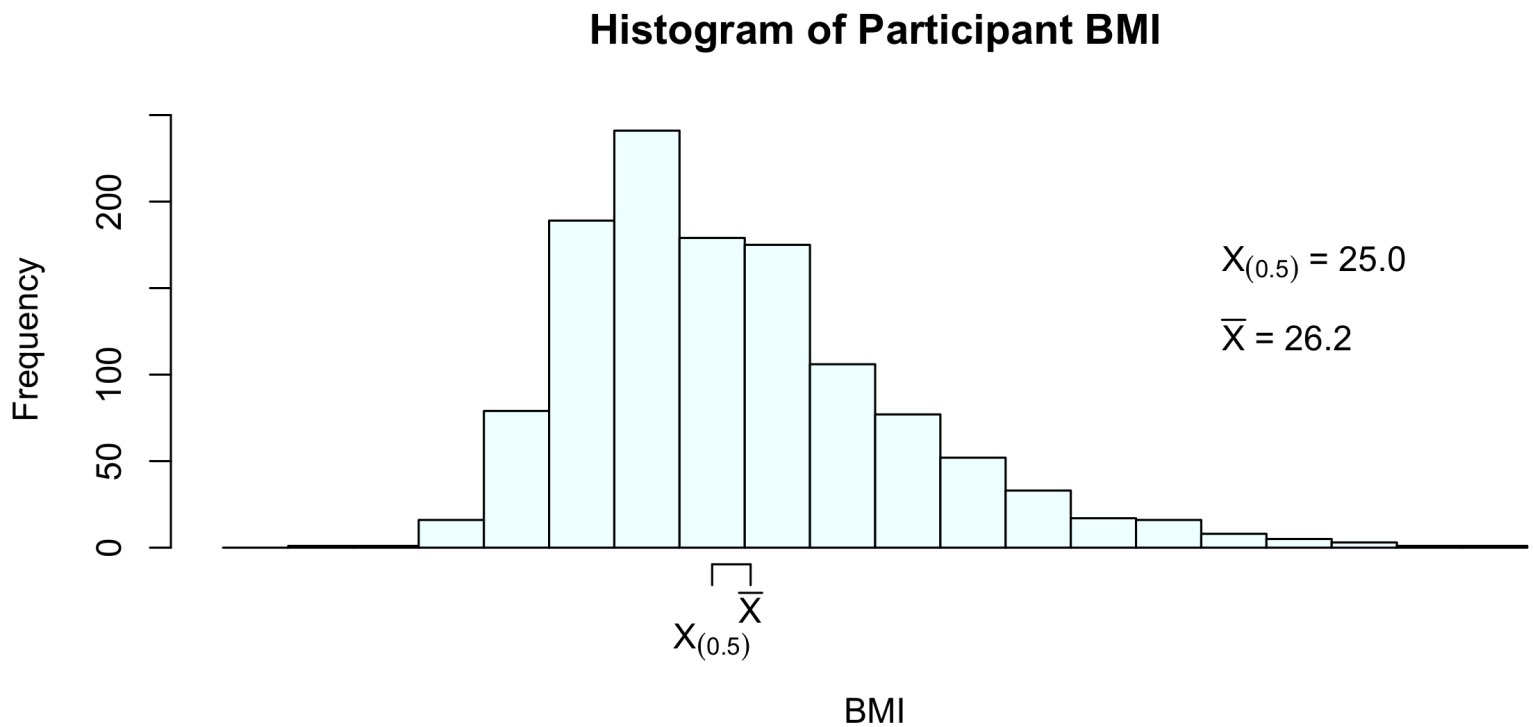
The mean is 4.856, and the median is 4.0.

- If the largest observation was a 100 instead of a 10, how would the mean and median be affected?
- The mean would now become 22.856, but the median would still be 4.0.
- We say that the median is *robust* to extreme observations.
- When a dataset is highly skewed, we report the median. Otherwise, if it is symmetric and bell-shaped, we report the mean since it uses all the observations.

# Measures of Center in a Graphical Context



Let us overlay these measures for the BMI data from slide 15.



# Regarding the Spread or Variability of Data



- Reporting the center alone of a sample of quantitative data is not sufficient.
- For instance, in the breast cancer dataset, if we reported that the mean BMI is 26.18, it is not enough for a good understanding of the women in the study.
  - Are most of the observations within 5 units of 26.18?
  - Would a BMI of 35 be considered high in this sample?
- Moreover, two samples with the same mean or median could nonetheless look very different; in particular, their variability could be very different.
- Measures of spread attempt to quantify this variability in different ways.

Dataset 3

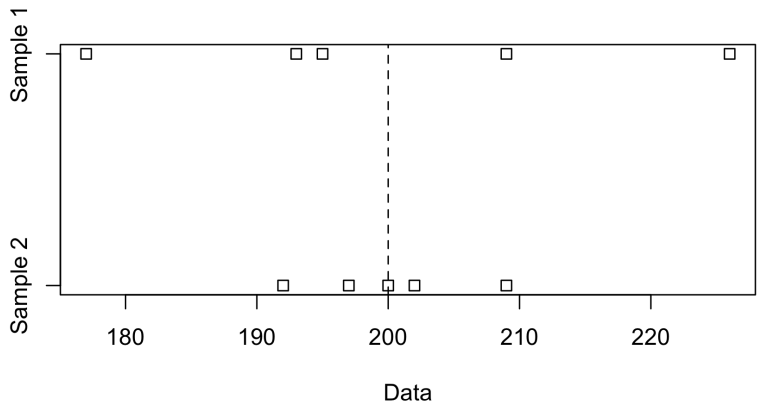
Consider two datasets, which we shall refer to as Sample 1 and Sample 2:

$$\text{Sample 1:} \quad 177,\ 193,\ 195,\ 209,\ 226$$

$$\text{Sample 2:} \quad 192,\ 197,\ 200,\ 202,\ 209$$

# Differing Spreads






The mean is 200 for both samples. However, it is clear that they both have different "spreads" or variability.

# The Range of a Dataset



**Definition 14 (Range)**

The **range** is the difference between the largest and smallest observations in a dataset. It is denoted by $R$.

- For Sample 1, we have that $R = 49$.
- For Sample 2, we have that $R = 17$.
- The strength of $R$ as a measure of spread is that it is very easy to compute.
- The main weakness is that it is sensitive to extreme observations.

# Variance and Standard Deviation



**Definition 15 (Variance and Standard Deviation)**

The **variance** of a dataset, $s^2$, is defined to be the average of the squared deviations from the mean. The **standard deviation** $s$ is defined to be square root of the variance.

- Roughly speaking, the standard deviation (or just sd) represents a typical distance or an average distance of an observation from the mean.
- If it is larger, it means that the data is more spread out.
- Variance is not in the same units as the observations, but the standard deviation is.

# An Empirical Rule Using Standard Deviation



If a distribution is bell-shaped, then approximately

- 68% of the observations fall within 1 standard deviation of the mean, i.e. between the values $\bar{X} - s$ and $\bar{X} + s$.
- 95% of the observations fall within 2 standard deviations of the mean. $(\bar{X} \pm 2s)$.
- All or nearly all the observations fall within 3 standard deviations of the mean $(\bar{X} \pm 3s)$.

The rules-of-thumb above can be used to identify intervals that contain "almost all" of the data.

# Quantiles



Let $p$ be a value between 0 and 1.

Definition 16 (Quantile or Percentile)

The $p$-**th quantile**, $\hat{q}_p$, is a value such that $p$ percent of the observations fall below or at that value. Quantiles are also known as percentiles.

- $\hat{q}_{0.25}$ is the value such that 25% of the observations are at or below this value. This value has a special name - it is also known as the lower, or first quartile.
- $\hat{q}_{0.5}$ is the value such that 50% of the observations are at or below this value. This value is in fact equal to $X_{(0.5)}$, the median.
- $\hat{q}_{0.75}$ is the value such that 75% of the observations are at or below this value. This value has a special name - it is also known as the upper, or third quartile.

# Inter-quartile Range (IQR)



**Definition 17 (IQR)**

The **IQR** is the distance between the upper and lower quartiles.

- It follows that 50% of the observations fall within the IQR.
- The IQR gives us an idea of how spread out the "middle" of the sample is.
- The IQR is robust to outliers and extreme observations.

# Computing Quantiles



**Data**

Consider the following set of numbers:

$$1, 2, 3, \ldots 198, 199, 200$$

The corresponding quantiles can be easily computed:

- $\hat{q}_{0.25} = 50$.
- $\hat{q}_{0.5} = 100.5$.
- $\hat{q}_{0.75} = 150$.
- The IQR is equal to 150 - 50 = 100.

Are these values unique?

# A Five-Number Summary of the Sample



**Definition 18 (Five-Number Summary)**

The **five-number summary** of a dataset consists of the minimum, lower quartile, median, upper quartile and the maximum.

- It gives a good indication of the center and variability of a dataset.
- It reduces huge datasets (typically in the size of thousands of observations) to just five numbers.

# Which Measure of Spread to Use?



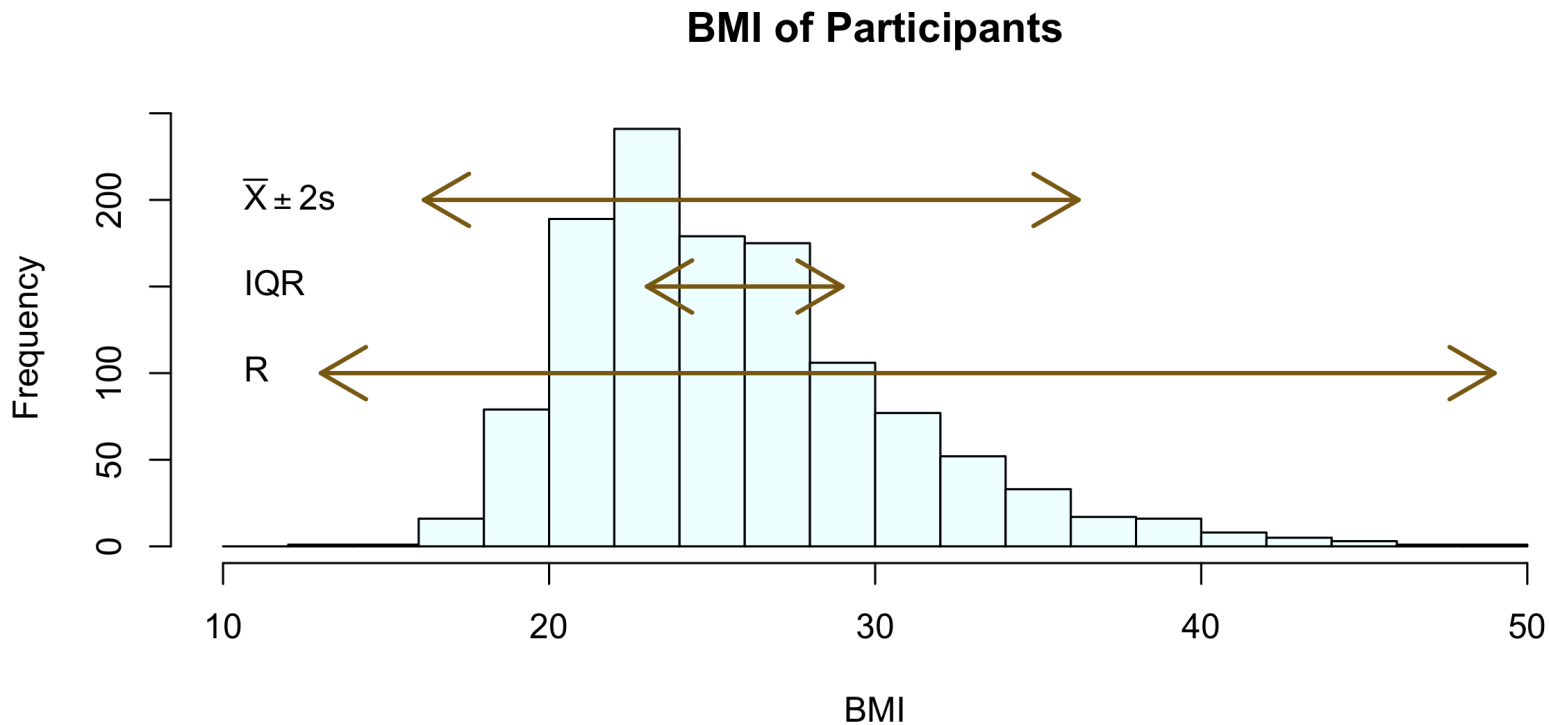
We have learnt the following measures of spread for numerical data:

- Range
- Variance and sd. This typically is used in conjunction with the sample mean to summarise the data.
- IQR. This typically is used in conjunction with the sample median to summarise the data, since they all rely on percentiles.

# Measures of Spread in a Graphical Context



Let us superimpose the measures of spread that we studied earlier for the BMI data in the Breast Cancer example.



# Measures of Spread in a Graphical Context

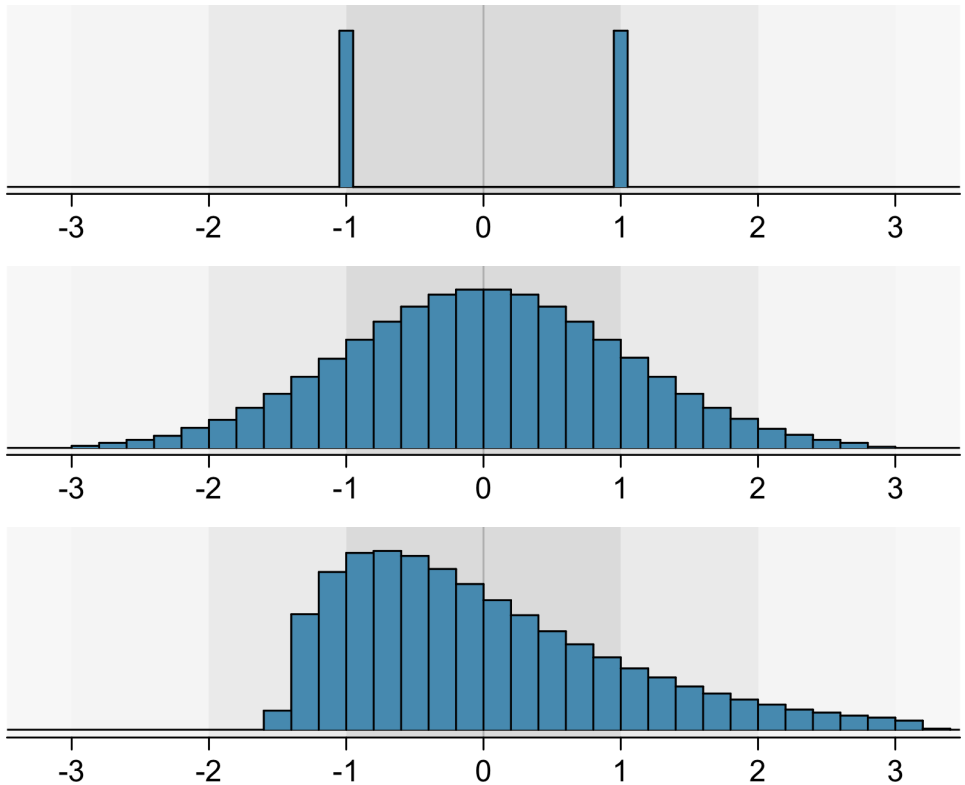


- Recall that the interval defined by $R$ encompasses 100% of the data.
- Recall that the interval defined by $IQR$ encompasses 50% of the data.
- Recall that the interval defined by $\bar{X} \pm 2s$ encompasses approximately 95% of the data (for bell-shaped data).
- These three measures are not to be compared with each other, but with their corresponding counterparts in another sample.
- The larger these intervals are, the more variability there is in the data - the less peaked the histogram will be.

# The Inadequacies of Numerical Summaries



No matter how many of the summary measures we report, nothing beats a picture..




All 3 samples had a sample mean of 0 and a sample variance of 1.

# Associations



- In the previous section, we covered exploratory techniques for summarizing a single variable at a time.
- However, many times variables are related or associated with others (see the example on slide 29).
- An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.
- An association between variables does not mean that one variable is causing the other.

Definition 19 (Response and Explanatory Variables)

- The **response variable** is the variable on which comparisons are made.
- The **explanatory variable** is any variable you believe the response depends on. If it is categorical, it defines the groups to be compared.

# Contingency Tables



Definition 20 (Contingency Table)

- A **contingency table** is a display for two categorical variables.
- Its rows list the categories for one variable and its columns list the categories of the other variable.
- Each entry in the table is the number of observations in the sample at a particular combination of categories of the two variables.

# $2 \times 2$ Contingency Table



Here is a summary table for the use of PMH and the incidence of Breast Cancer.

**Hormone Usage * Breast Cancer Incidence Crosstabulation**

| | | | Breast Cancer Incidence | | |
|---|---|---|---|---|---|
| | | | Absent | Present | Total |
| Hormone Usage | Non-user | Count | 682 | 318 | 1000 |
| | | % within Hormone Usage | 68.2% | 31.8% | 100.0% |
| | User | Count | 121 | 79 | 200 |
| | | % within Hormone Usage | 60.5% | 39.5% | 100.0% |
| Total | | Count | 803 | 397 | 1200 |
| | | % within Hormone Usage | 66.9% | 33.1% | 100.0% |

- It appears that the rate of breast cancer is higher for PMH users than non-users. (39.5% versus 31.8%)
- This is precisely what we mean by "association". A higher proportion of breast cancer is **associated** with PMH users.

# Independence and Dependence



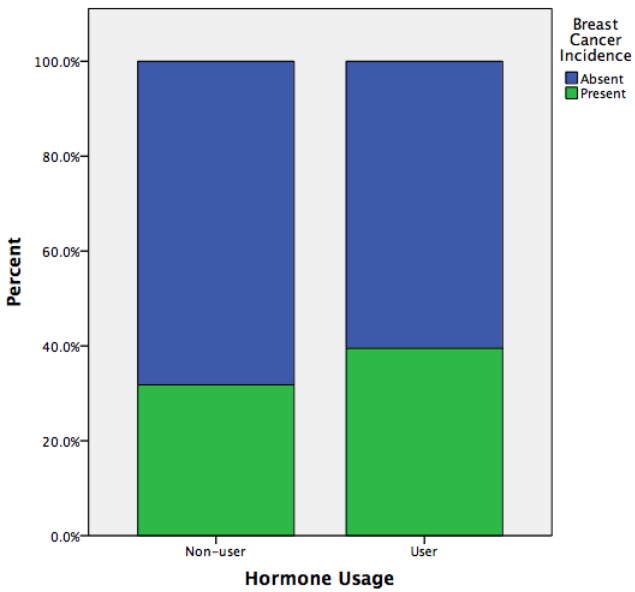
**Definition 21 (Independence and Dependence (Association))**

- Two categorical variables are **independent** if the conditional proportions for one of them are identical at each category of the other.
- The variables are **dependent**, or associated, if the conditional proportions are not identical.

- In a contingency table, it is possible to compute what we "expect" to see under the assumption that the column and row totals are known.
- Statistical tests use these expected values to assess the deviation from independence, i.e. the strength of the association.

# A Stacked Bar Plot






- We can visualise a table using a stacked bar chart like the one on the left.
- In session II, we shall see an improvement on this chart, known as a mosaic plot, that also displays the deviation from independence.
- For now, let's think about how we can describe the difference between PMH users and non-users.

# Describing the Difference - Difference



We let $\hat{p}_1$ and $\hat{p}_2$ be the proportions from the two groups. For instance, in the PMH user example, $\hat{p}_1 = 0.318$ and $\hat{p}_2 = 0.395$.

- The difference of proportion is defined to be:

$$\hat{p}_1 - \hat{p}_2$$

- Thus we can say that the difference in proportions is 0.077, or 7.7%.
- Note that the difference in proportions can take any value between -1 and 1.
- A value of 0 corresponds to "no association".
- It does not matter whether we take $\hat{p}_1 - \hat{p}_2$ or $\hat{p}_2 - \hat{p}_1$. The strength of the association does not change; only the sign of this metric does.

# Describing the Difference - Relative Risk



- The relative risk is defined to be
$$\hat{p}_1/\hat{p}_2$$
- The relative risk can take on values between 0 and infinity. A value of 1 corresponds to "no association".
- Analogous to the difference in proportions, it does not matter whether we take $\hat{p}_1/\hat{p}_2$ or $\hat{p}_2/\hat{p}_1$. A relative risk of 0.25 is the same as a relative risk of 4.0.
- It is preferrable to use relative risk when both proportions are close to 0. For instance, suppose that in the PMH users example, we had observed that $\hat{p}_1 = 0.01$ and $\hat{p}_2 = 0.03$. Which statement below conveys more insight?
  - The proportion of breast cancer incidence for PMH users is 3 times higher than that for PMH non-users.
  - The difference in proportions between PMH users and non-users is 0.02 or 2%.

# Definition of Odds



Definition 22 (Odds)

- For a categorical variable with 2 possible values, define one of them to be the "success" and the other to be the "failure"
- Let $p$ be the probability of success, and $1 - p$ be the probability of failure.
- Then the **odds of success** is defined to be

$$\text{odds} = \frac{p}{1-p}$$

- Odds equal to 0 corresponds to probability of success equal to 0.
- Odds equal to 1 corresponds to probability of success equal to 0.5.
- Odds equal to $\infty$ corresponds to probability of success equal to 1.

# Example of Odds in Football



Example 23 (Odds of Brazil Winning)

- Just before the 2014 FIFA World Cup began, the bookmakers listed the odds of Brazil winning the trophy as 3/1.
- This just means the odds were 3.
- When we convert it to a probability, what they are saying is that the probability of Brazil winning the World Cup was 0.25.
- This can be obtained by solving for $p$ in the equation

$$\begin{aligned} 3 &= \frac{p}{1-p} \\ 3-3p &= p \\ 3 &= 4p \\ \frac{3}{4} &= p \end{aligned}$$

# An Aside on Logarithms



- The logarithm of a value $x$ is defined to by $y$ such that

  $$x = 10^y$$

  We write it as

  $$y = \log_{10} x$$

- In computations, the most important property of logs is that

  $$\log_{10}(x_1 x_2) = \log_{10} x_1 + \log_{10} x_2$$

- Thus the log of a ratio is the difference of the logs of numerator and denominator.

  $$\log_{10}(x_1/x_2) = \log_{10} x_1 - \log_{10} x_2$$

  This property enables us to deal with subtraction instead of division.

- Another important property of logs is that it de-emphasises large differences. This is best understood from the graph; it enables us to make a distribution of values more symmetric.

# Describing the Difference – Odds Ratios and Log-Odds



- In statistics, the **odds ratio** - the ratio between the odds of success between two groups is used quite often because, due to the sampling design, it is not possible to estimate the relative risk.
- In our case, we consider taking a transformation of the odds and then comparing the transformed odds.
- The transformation is the log-transformation. It accentuates differences when the proportions for both groups are close to 0 or 1. For instance,
  - Suppose $\hat{p}_1 = 0.99$ and $\hat{p}_2 = 0.97$. Then the log odds for group 1 is 1.996 and the log odds for group 2 is 1.590.
- On the log-odds scale,
  - Log-odds equal to $-\infty$ corresponds to probability of success equal to 0.
  - Log-odds equal to 0 corresponds to probability of success equal to 0.5.
  - Log-odds equal to $+\infty$ corresponds to probability of success equal to 1.

# "Starting" The Counts



- When we observe a count of 0, we should be wary.
- Extreme values should be shrunk towards the middle slightly when used in estimation or prediction.
- Starting the count refers to adding 1/6 to both "success" and "failure" categories before computing the proportions.
- This has two benefits:
  - The extreme values of 0 and 1 are made less so.
  - We can now distinguish ties better. Consider a couple of groups where $\hat{p}_1 = 1/10$ and $\hat{p}_2 = 1000/10000$. The proportion for group two is clearly a more reliable estimate since the sample size is larger. By starting the counts, we estimate $\hat{p}_1 = 0.113$ and $\hat{p}_2 = 0.100$.

# Starting Counts in Literacy Rates



Example 24 (Literacy Rates)

- In 2015, there were 9.581 million males above the age of 15 in North Korea.
- They were all classified as literate.
- Instead of computing the log-odds as

$$\log_{10} \frac{1}{0} = \text{undefined}$$

we start the counts, and compute it as

$$\log_{10} \frac{9581000 + 1/6}{9581000 + 1/3}$$

# Two Quantitative Variables

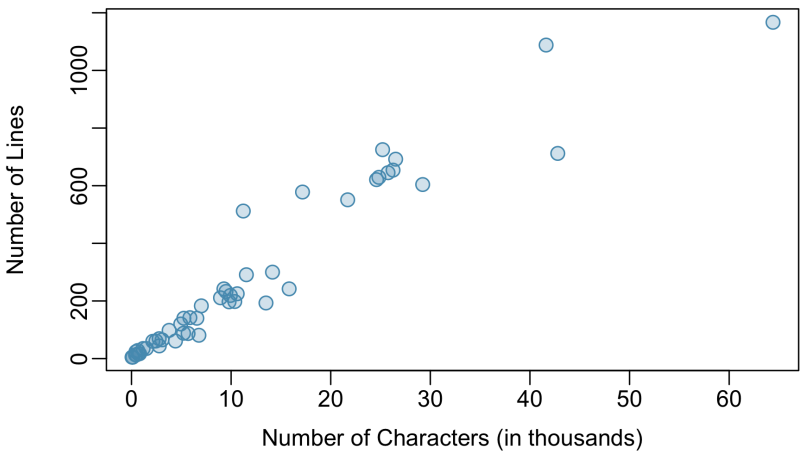


- Generally speaking, there are a couple of scenarios in which we could be interested in the relationship between two quantitative variables:
  (1) Neither variable is time, and the points on the plot are not related to each other. For instance, consider the association between military spending and social unrest. Each point on the plot would represent a country.
  (2) One of the variables is time, and we are interested in the progression of, say, Consumer Product Index (CPI) over time. Every point on the plot would be from the same country, and points would be related to each other. For instance, knowing the CPI in 2000 gives us a good idea about the CPI in 2001.
- In scenario (1), ideally, we would hope to see a straight line relationship between the two variables. We shall discuss this scenario here in this Session.
- In scenario (2), we do not expect to see a straight line, but we wish to explain or summarise the trend over time. We shall take a look at this (smoothing) in Session II.

# Scatterplots



## Example 25 (Email characters)




- On the left, we have a scatterplot from a dataset regarding 50 emails.
- The number of line breaks was counted for each email (y-axis), along with the number of characters (x-axis)
- Would you say there is a relationship between the two variables?

# What to Say



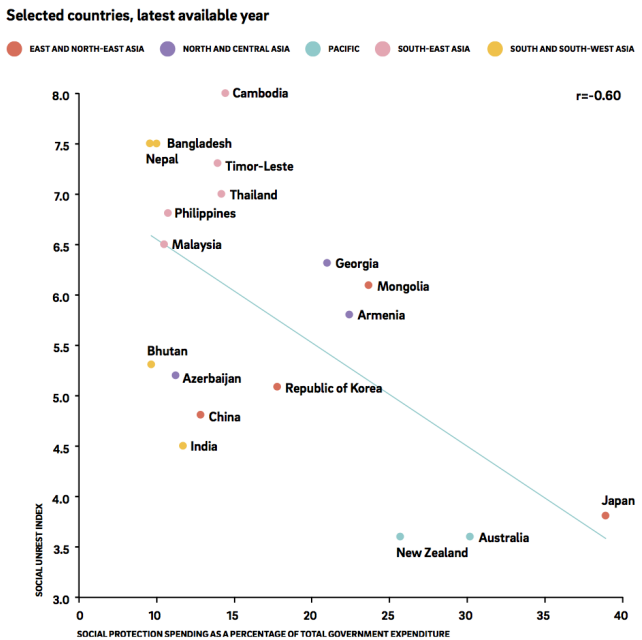
Typically, when we are faced with a scatterplot of two variables, we would try to mention the following points:

- Does there seem to be a positive, negative or no association?
- Can the trend be approximated reasonably well by a straight line? If so,
  - How do the points vary about that line?
  - In particular, is the variability consistent at all points?
  - Can we approximate and quote the rate of change?
- Are some observations unusual? For instance,
  - Are any points very far away from the postulated straight line?
  - Are any points far away from their neighbours?

# Social Protection and Unrest



## Example 26 (Social Protection and Unrest)




- The chart on the left plots Social Unrest Index on the $y$-axis against Social Protection Spending (as a percentage) on the $x$-axis.
- It is taken from Figure 2.3 of the *Time for Equality* report.
- What can we say about the relationship?

# Scatterplot Summary



- There appears to be a negative relationship between percentage of social protection spending and social unrest.
- The gradient is approximately -0.1, meaning that every 1% increase in social protection spending is associated with a 0.1 point decrease in the social unrest index.
- India and Cambodia deviate greatly from this relationship. The former has a surprisingly low social unrest index given it's expenditure, whereas the latter has an extremely high social unrest index.

# A Hidden Relationship



- At times, the relationship is not obviously linear. In such cases, a transformation of the $y$-axis or $x$-axis might help.
- The idea is similar to the purpose of taking log-odds in proportions.
- We wish to emphasize/de-emphasize differences at one end of the scale in order to "straighten" out our plot.
- Some might claim that this reduces interpretability, but as an analyst we will be missing the information in the data if we do not try transformations.
- We shall see an example about this in the next Session, but it may be worth trying some of the following transformations:
  - $\log y, \log x, -\frac{1}{x}, -\frac{1}{y}, x^2, y^2$
- For instance, in the example on slide 79, it might be worth trying $\log y$ or $\log x$.

# Lifting the Lid



**A Quote from George Box, 1987**

*Essentially, all models are wrong, but some are useful.*

- A model that we fit could be as simple as a mean or a median to different groups, or as complex as a straight line fitted to transformed $x$- and $y$- variables.
- No matter the complexity of the model (which we always regard as incomplete), let us assume that our data is of the form

$$\text{observed} = \text{fit} + \text{residual}$$

- The fit is what our model predicts; the residual is what it cannot explain.
- By studying the residuals after a fit, we can understand more about our data; we have lifted the lid and peeked underneath.
- We shall use this approach as we embark on Session II.

# Comparing Two Numbers



- Suppose we encounter two individuals, John and James. It is not uncommon to hear the following sort of descriptions:
  - John is a head taller than James.
  - John is twice as big as James.
- When we compare two numbers, we usually compare them via their difference or their ratio. *Recall the difference in proportions and relative risk metrics*
- Logarithms were in fact invented to make working with ratios as easy as working with differences.
- Let's keep these ideas in mind when making our summaries/insights from now on.

# Resource Distribution



We hear anecdotes such as the following all the time:

- The poorest 20% of the people on Earth earn only 1% of the income.
- A mere 20% of the people on Earth consume 86% of the consumer goods.
- Only 3% of the US population owns 95% of the privately held land.

These are all statements pertaining to how equally a resource is distributed throughout a population.

# Lorenz Curves



- A Lorenz curve is an instrument for visualising the distribution of a quantity in a population.
- Suppose that a quantity $Q$ is distributed in a population.
- Line up the population by increasing order of their shares of $Q$.
- Then for any $p$ between 0 and 1, the people in the first fraction $p$ of the line represent the $Q$-poorest $100p\%$ of the population.
- We denote the fraction of $Q$ owned by that fraction of the population, as $L(p)$.

*The Lorenz curve for a resource $Q$ is the curve $L(p)$ ($y$-axis) against $p$ ($x$-axis). It depicts the fraction of total $Q$ that the poorest $100p\%$ of the population possess.*

# Using Lorenz Curves to Describe Resource Distribution



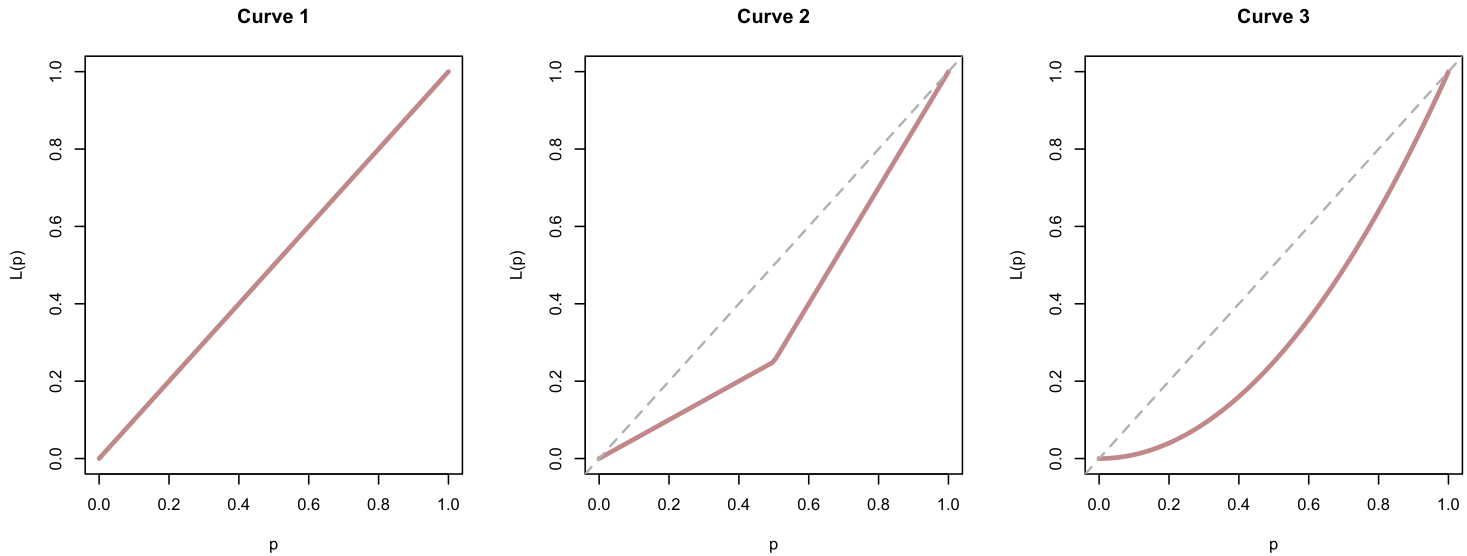
We hear anecdotes such as the following all the time:

- The poorest 20% of the people on Earth earn only 1% of the income.
  - $L(0.2) = 0.01$
- A mere 20% of the people on Earth consume 86% of the consumer goods.
  - $L(0.2) = 0.86$
- Only 3% of the US population owns 95% of the privately held land.
  - $L(0.97) = 0.05$

These are all statements pertaining to how equally a resource is distributed throughout a population.

# Examples of Lorenz Curves






Curve 1: Every individual has the *same income*.

Curve 2: The poorest half of the population accounts for 25% of all the income. Each individual within the poorest 25% has the same income.

Curve 3: Within the population, *all incomes are equally likely*.

# The Gini Index



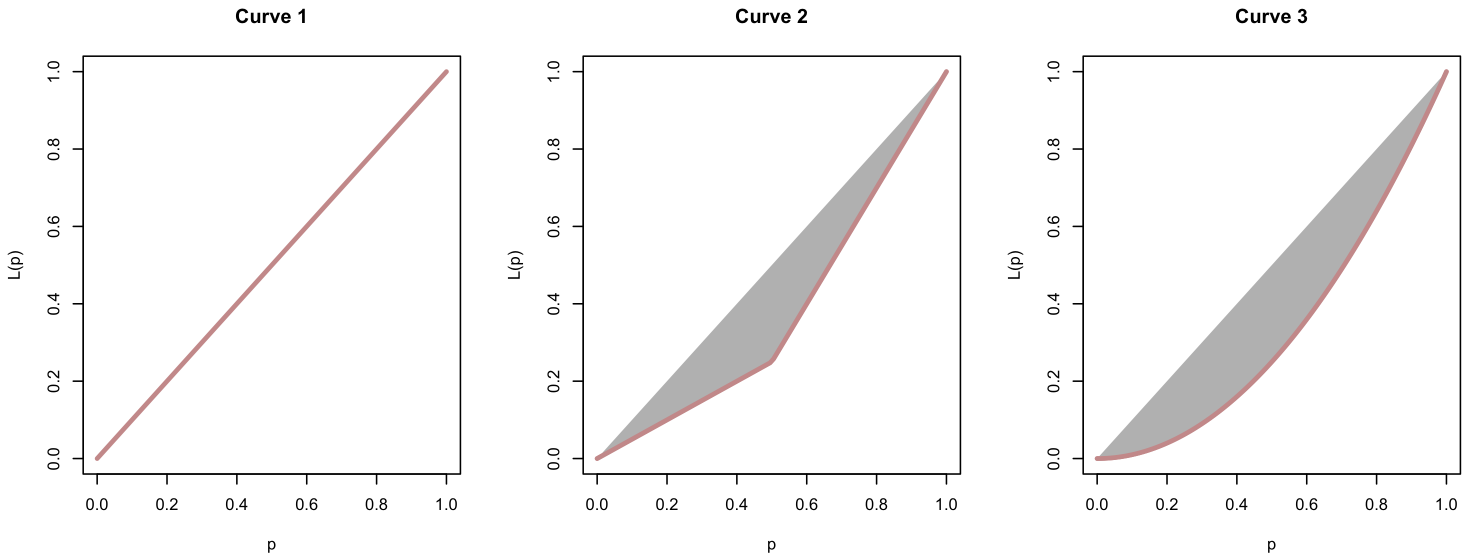
- The Lorenz curve contains all details about the income distribution of a population.
- However, it is a curve. We typically need to compare the income distributions across several nations, and hence we need a single number to summarise it.
- The Gini index has been frequently used. It is computed as

$$2 \times \text{ Area between } L \text{ and Curve 1}$$

- It is a number between 0 and 1
  - 0 denotes absolute income equality.
  - 1 denotes that all the wealth is concentrated in a single individual.

# Examples of Gini Indices






Curve 1: $G = 0$.

Curve 2: $G = 0.25$.

Curve 3: $G = 0.67$

# Interpreting the Gini Index



Although it has its criticisms, here are a few ways to interpret the Gini index in an intuitive way.

- Suppose we throw all the money earned by the population into a huge pot, and then pick one dollar at random. Let us call refer to this as the "average dollar'. Then it can be shown that the percentile of this dollar is

$$\frac{G+1}{2}$$

- Suppose that the mean income is $\mu$. Suppose that we pick two incomes at random, and then record the minumum and maximum incomes as $X_{(1)}$ and $X_{(2)}$. Then on average,

$$X_{(1)} = (1-G)\mu \text{ and } X_{(2)} = (1+G)\mu$$

which gives us an idea of the spread of incomes.

# Index