Statistics and Probabilities for Computing (CEN1602)

Unit-1

Introduction to Statistics and Data Description

Dr. Ahmad Alzahrani

Note: These slides are adapted from material by **Prof. Jonathan Marchini** and modified for this course.

Outline

- **What is Statistics?**
- **Why we need Statistics?**
  - The scientific process
- **Different types of data (variable values)**
  - Discrete/Continuous
  - Quantitative/Qualitative
- **Graphical descriptions of data**
  - Bar charts, Histograms, Dot plots, Scatter plots, Box plots
- **Numerical descriptions of data**
  - Location: Mean, Median, Mode
  - Dispersion: IQR, SIQR, MAD, variance, standard deviation

What Is Statistics?

- **Statistics** is the science of working with data
- It involves:
  1. **Collecting** data from real situations
  2. **Organizing** data into meaningful forms
  3. **Summarizing** data using numbers and graphs
  4. **Analyzing** data to detect patterns and trends
  5. **Interpreting** data to draw conclusions
  6. **Supporting** decisions under uncertainty
- In short: Statistics turns data into meaningful information.

Why Statistics Matters in Computing?

- **Statistics helps us understand large sets of computing data** (server logs, sensor readings)
- **It helps summarize and understand system performance** (average response time, CPU usage)
- **It supports analyzing user behavior and usage patterns** (click frequency, session duration)
- **It helps detect unusual or suspicious activity** (unusual logins, abnormal traffic spikes in networks)
- **It is essential for evaluating machine learning models** (accuracy rate, prediction error)
- **It provides evidence for software testing and quality** (failure rate, bug occurrence)
- **It helps on improving system and efficiency** (throughput improvement, memory usage)
- **It supports evidence-based decision making, not guesses** (algorithm selection, deciding best system settings)

Two Main Branches of Statistics

1. Descriptive statistics:

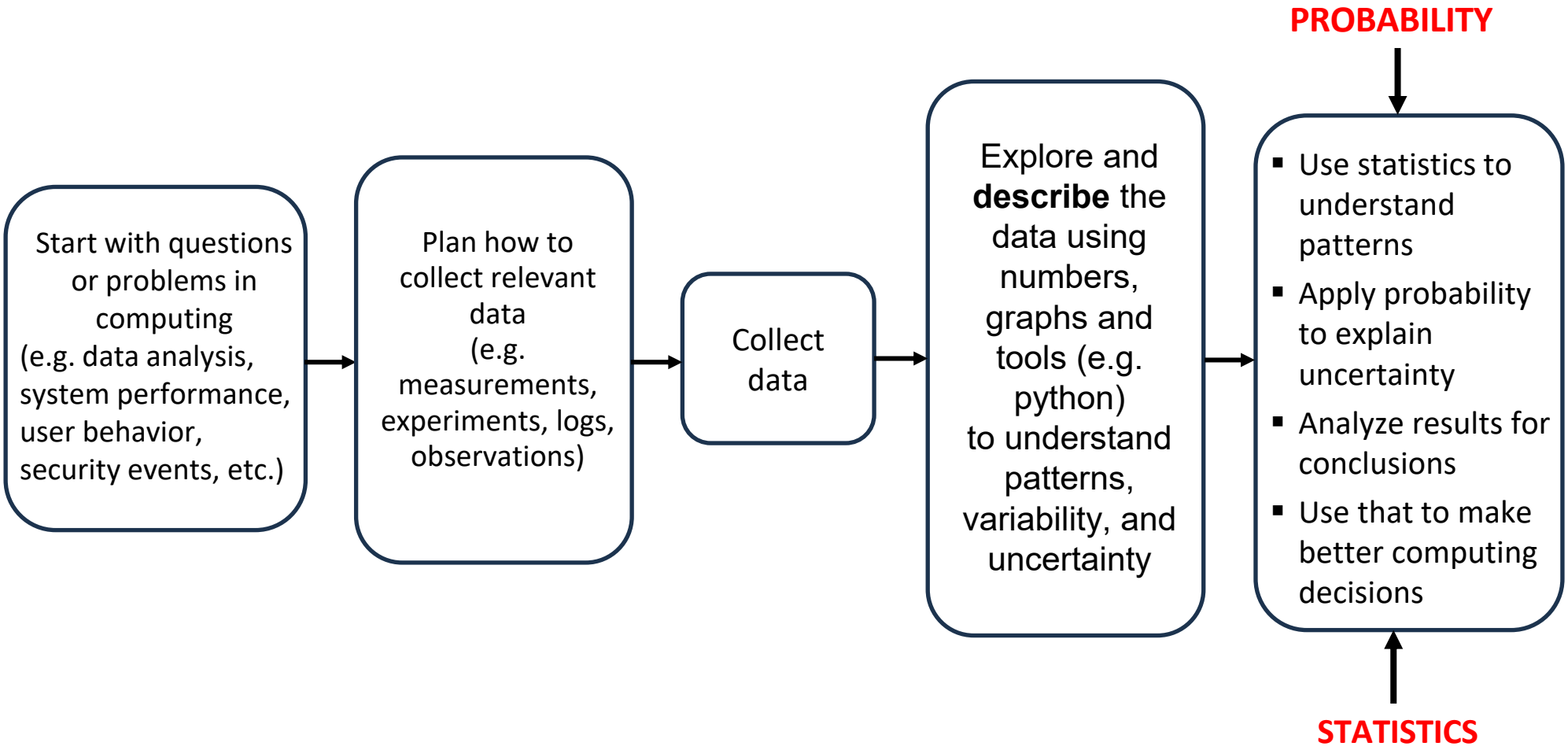
- It focuses on summarizing and organizing data
- It uses tables, charts, and summary measures
- It answers: “What does the data show?”

2. **Inferential statistics:**

- It uses samples to make conclusions about populations
- It answers: “What can we conclude beyond the data?”

Both branches are used together in real applications

The Statistical Process



In this lecture, we focus on the step of describing and summarizing data.

An Example: Applying the Statistical Process

- Engineers and computer scientists are often interested in the relationship between **system load** and **performance**.
- A focused question might involve studying a specific performance metric (e.g., **response time**) and how it changes as the number of users increases.
- To assess whether load is a good predictor of slowdowns, we need to measure **load level** and **response time** for a sample of requests from the system.
- We must decide how to collect the data fairly (peak vs off-peak, background tasks).
- Once data is collected, we use statistics to analyze patterns and variation
- This helps us decide whether the system slows down due to load or random effects.

Datasets Consist of Variables

- Datasets consist of values recorded for one or more variables.
- A **variable** is any characteristic that can take different values.
- Variables describe properties of objects, people, or events.
- Examples of variables: age, height, number of siblings, exam score, eye color, satisfaction level, etc.
- In most tabular datasets, each row corresponds to one observation, and each column corresponds to one variable.
- **Data** are the actual values that **variables** take.
- In this course, we use datasets to learn key concepts in statistics and probability.

Two main types of variables

1. Measurement (Quantitative) variables occur when we 'measure' things (e.g., height or weight.)

2. Categorical (Qualitative) variables occur when we assign objects into labelled groups or categories. e.g., when we group people according to hair colour or race.

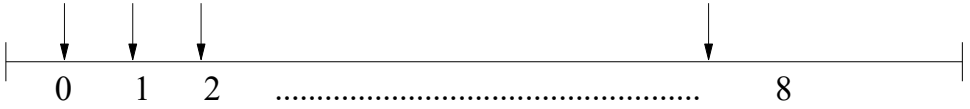
a. **Ordinal variables:** categories that can be ranked in a meaningful order (e.g., gold/silver/bronze medal)

b. **Nominal variables:** do not have a natural ordering (e.g., gender)

Discrete and Continuous Variables

Discrete variables

No. of students late for a lecture

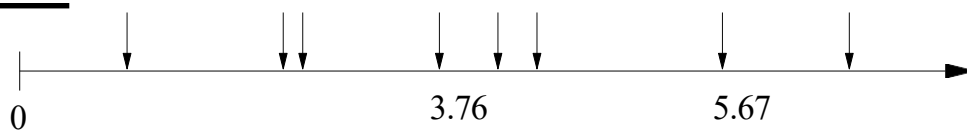


There are only a limited set of distinct values/categories

i.e. we can't have exactly 2.23 students late, only integer values are allowed.

Continuous variables

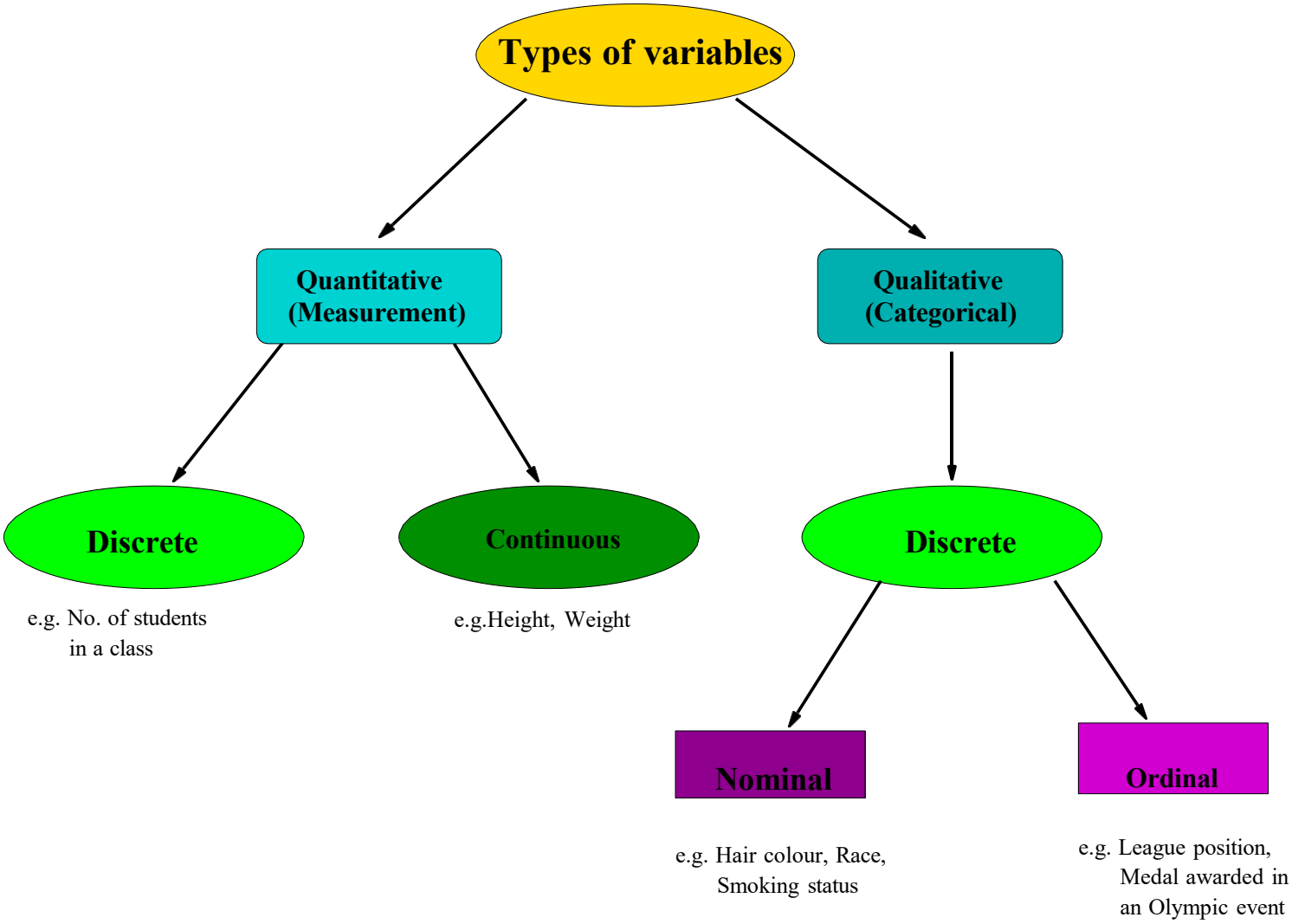
Time spent studying statistics (hrs)



In theory there are an unlimited set of possible values!

There are no discrete jumps between possible values.

Summary of Variable Types



Graphical Description of Data

- Plotting data helps us understand patterns, trends, and unusual or outlier values
- Graphs reveal important features that numbers alone may hide
- Visualizing data early helps detect errors and unexpected behavior
- Plots help us understand variation in data
- Visualization helps explain uncertainty in results
- Plotting supports choosing the right statistical methods
- Graphs make complex data easier to interpret
- We should always visualize data before drawing conclusions

Example Dataset: Birth Records (Illustrative)

- A dataset consisting of birth records observed over time is collected.
- For each record, several properties are measured, such as time of birth, birth weight, and gender.
- The dataset is shown on the next slide and is used as an example to demonstrate statistical concepts and graphs.
- Even though this dataset was collected for a different study, it can still help us explore questions such as:
    - Do girls weigh more than boys at birth?
    - What is the distribution of the number of births per hour?
    - Is birth weight related to the time of birth?
    - Is gender related to the time of birth?
    - Is there an equal chance of being born a girl or a boy?

Time	Gender	Weight
5	1	3837
64	1	3334
78	2	3554
1:15	2	3838
1:77	2	3625
2:45	1	2208
2:47	1	1745
2:62	2	2846
2:71	2	3166
4:28	2	3520
4:55	2	3380
4:92	2	3294
4:94	1	2576
5:49	1	3208
6:35	2	3521

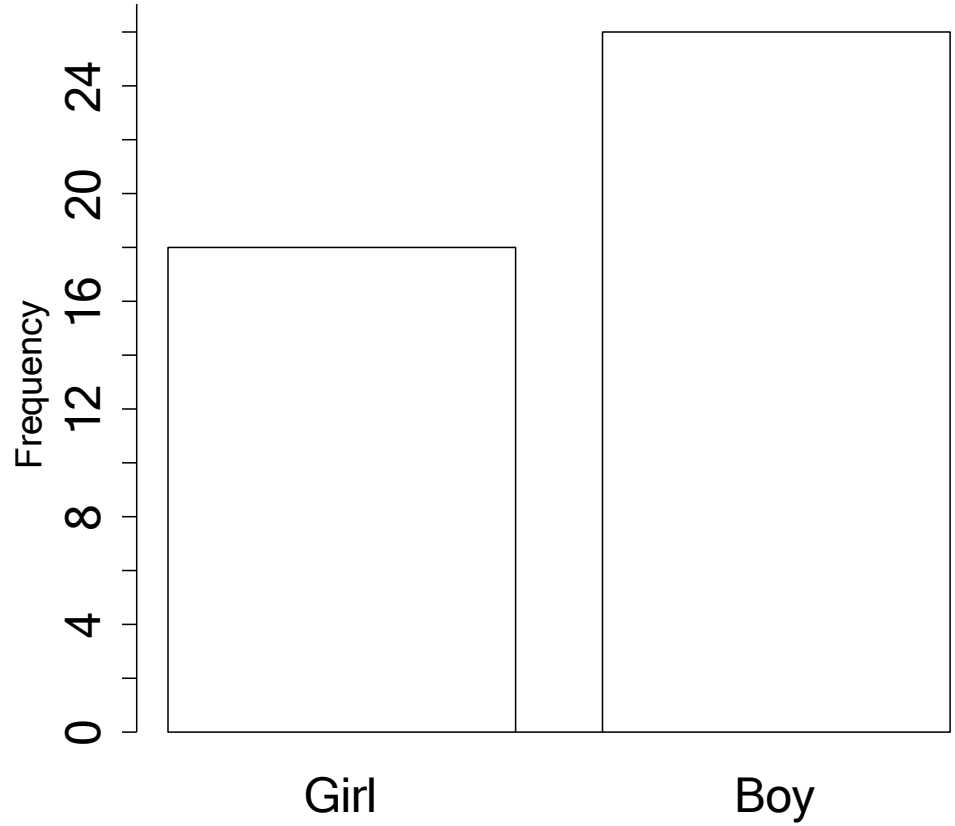
Time	Gender	Weight
6:49	1	3746
6:53	1	3523
6:93	2	2902
7:29	2	2635
7:76	2	3920
7:85	2	3690
8:46	1	3430
8:47	1	3480
8:73	1	3116
8:86	1	3428
9:14	2	3783
9:91	2	3345
10:17	2	3034
10:62	1	2184
10:87	2	3300

Time	Gender	Weight
11:05	1	2383
11:34	2	3428
11:49	2	4162
11:87	2	3630
11:89	2	3406
11:91	2	3402
12:10	1	3500
12:37	2	3736
12:51	2	3370
12:64	2	2121
12:83	2	3150
13:37	1	3866
14:07	1	3542
14:35	1	3278

Time here is measured in hours, birth weight is measured in grams, and gender is coded as 1 = male, 2 = female.

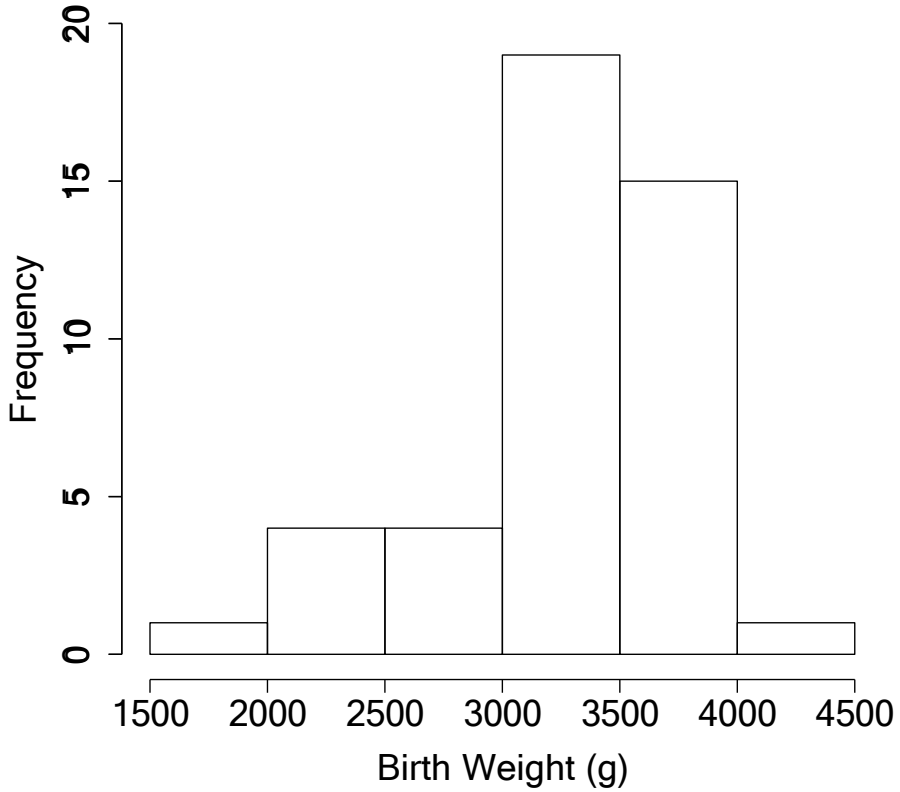
Bar Charts

- A Bar Chart is a useful method of summarising Categorical Data.
- We represent the counts/frequencies/percentages in each category by a bar.



Histograms

- Bar charts are used for categorical data, while histograms are used for numerical measurement data.



Constructing Histograms (an example)

- For the birth records dataset we can draw a histogram of the birth weights.

- To draw the histogram, we find the smallest and largest values

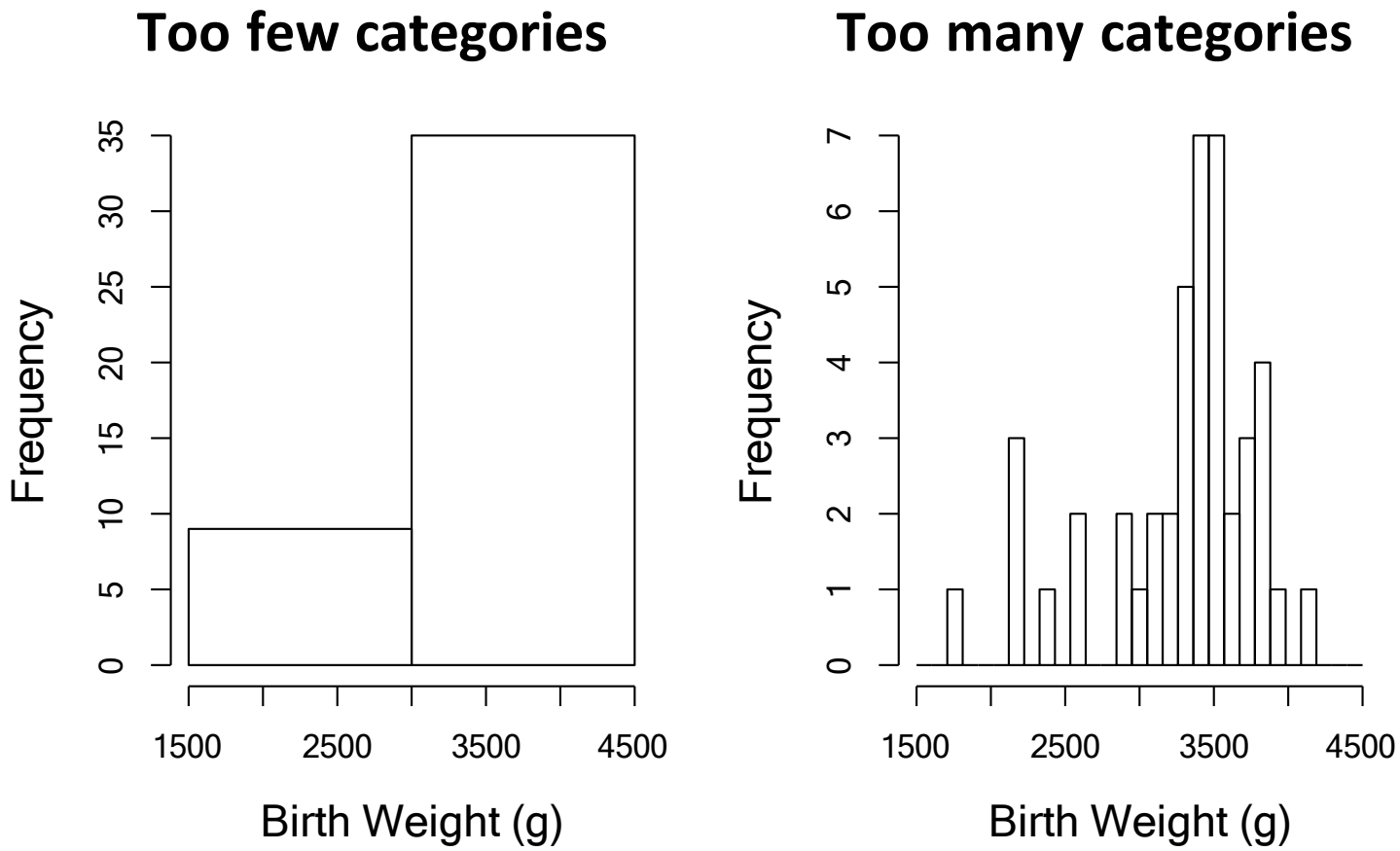
smallest = 1745 largest = 4162

There are only 44 weights, so we decide on 6 equal sized categories

Interval	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
Frequency	1	4	4	19	15	1

- Using these categories works well, the histogram shows us the shape of the distribution, and we notice that distribution has an extended left '**tail**'.

Constructing Histograms (an example)

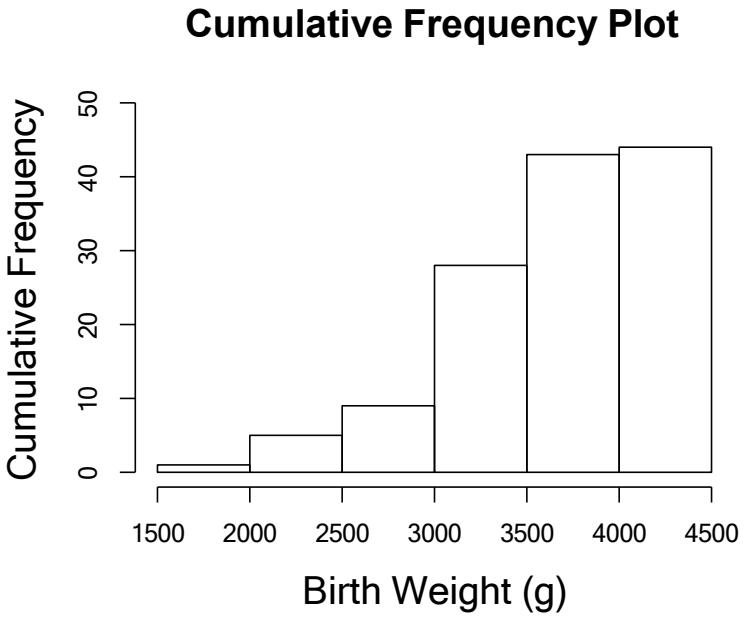


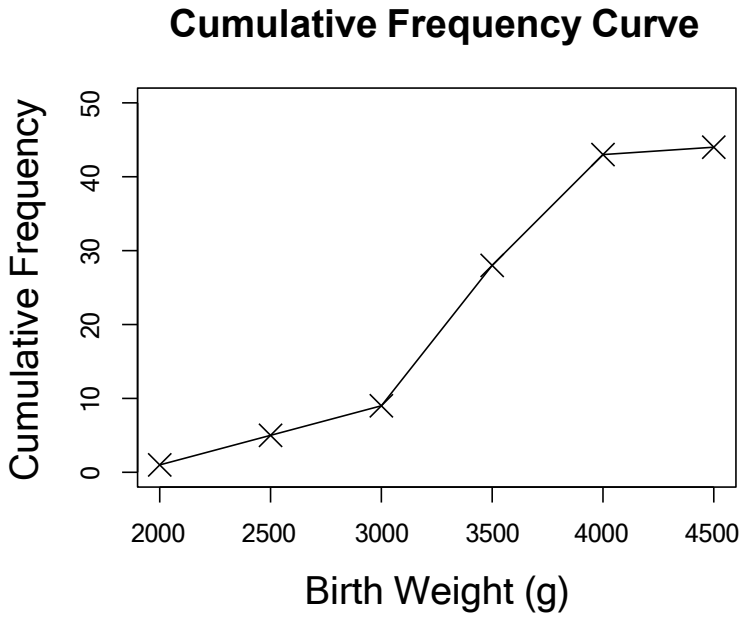
- Too few categories and the details are lost. Too many categories and the overall shape is obscured by too many details

Cumulative Frequency Plots and Curves

- The cumulative plot shows how many values are below a given value.
- It is made by adding the frequencies in the table step by step as shown.
- The median is found at the point where half of the data has been counted

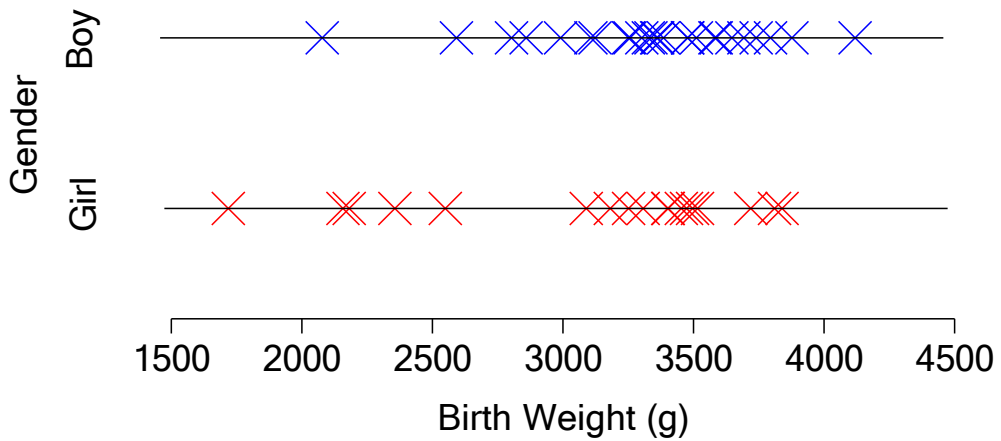
Interval	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
Frequency	1	4	4	19	15	1
Cumulative Frequency	1	5	9	28	43	44





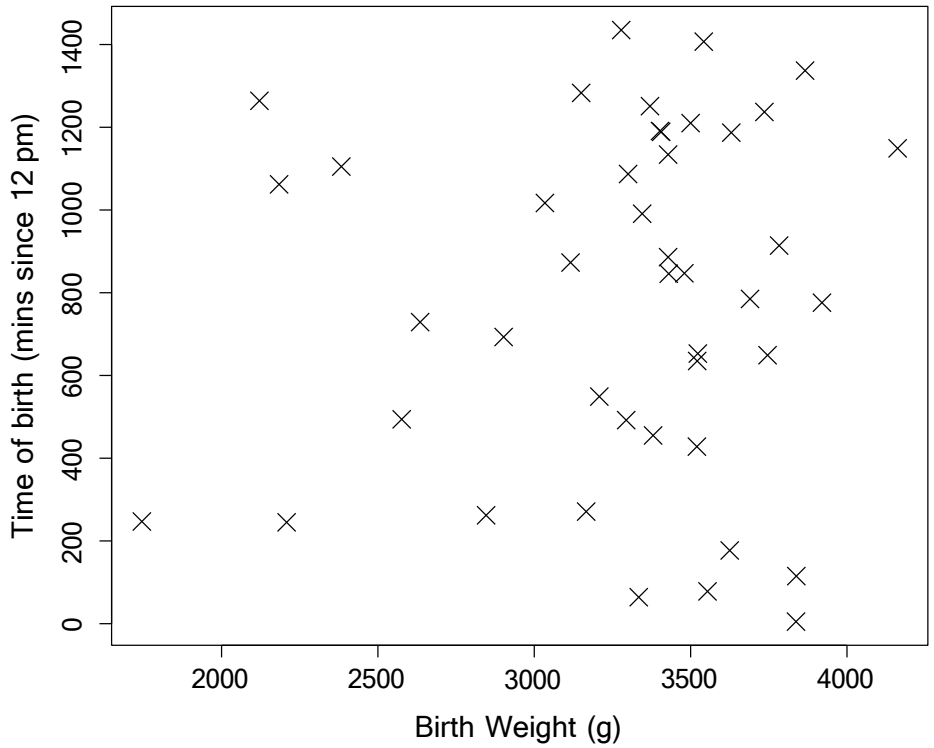
Dot plots

- A dot plot is a simple and quick way of visualizing a dataset
- A dot plot places a dot on a number line for each observation in the dataset
- Each dot represents one observation. Dots may be stacked when the same value occurs more than once
- This makes repeated values (frequency) easy to see
- Helps quickly compare differences between groups
- The dot plot below displays birth weight observations for boys and girls.

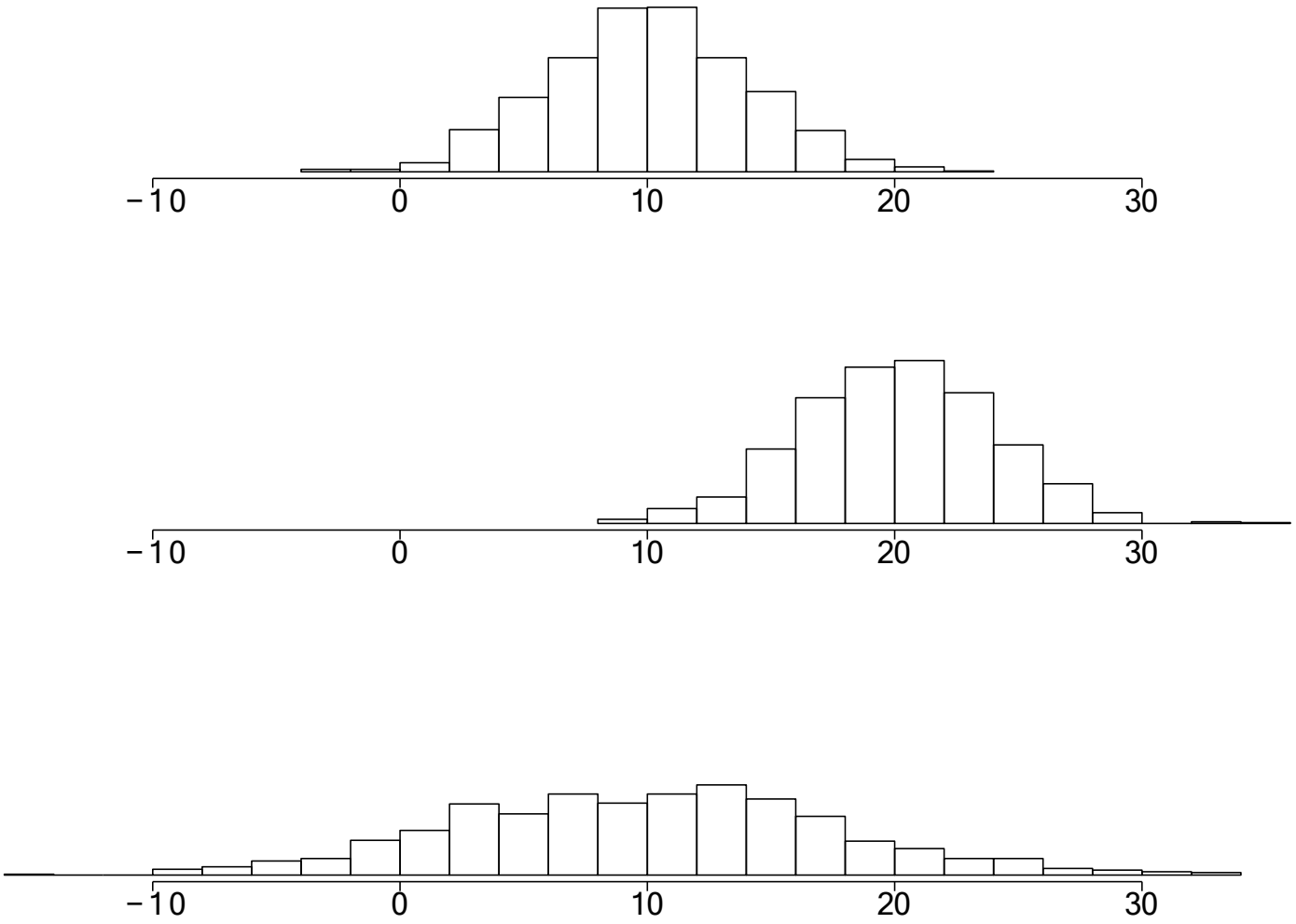


Scatter Plots

- A scatter plot displays pairs of numerical values as points on a graph
- Used to visualize the relationship between two quantitative variables
- Each point represents one observation from the dataset
- From the pattern in the plot, we can tell if there is a relationship between the variables



Comparing Histograms



Numerical Description of Data

- Numerical measures summarize data using numbers
- They help describe key features such as **location** and **spread**
- Measures of **location** (also called **center**) show typical values in the data. Examples include mean, median, and mode
- Measures of **spread** describe how data values differ.
- Examples of measures of **spread** include range, IQR, variance, and standard deviation
- **Spread**, **variability**, and **dispersion** mean the same idea
- **Numerical** summaries provide precise values, while **graphical** summaries help reveal patterns and trends

Numerical Description of Data

There are 3 main measures of location:

1. The Mode
2. The Median
3. The **Sample** Mean ($\bar{x}$) and **Population** mean (μ)

There are 5 main measures of spread:

1. The Interquartile Range (IQR) and Semi-Interquartile Range (SIQR)
2. The Mean Deviation
3. The Mean Absolute Deviation (MAD)
4. The **Sample** Variance (s^2) and **Population** Variance (σ^2)
5. The **Sample** Standard Deviation (s) and **Population** Standard Deviation (σ)

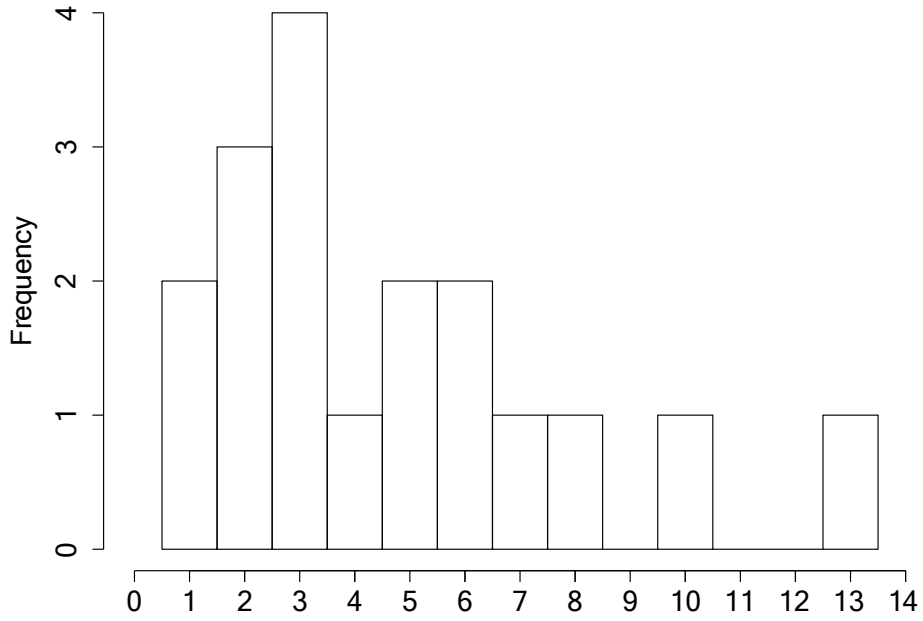
Keep in mind: By convention in statistics, **population size** is denoted by capital **N**, and **sample size** is denoted by lowercase **n**.

The Mode

- **The Mode** of a set of numbers is simply the most common value.

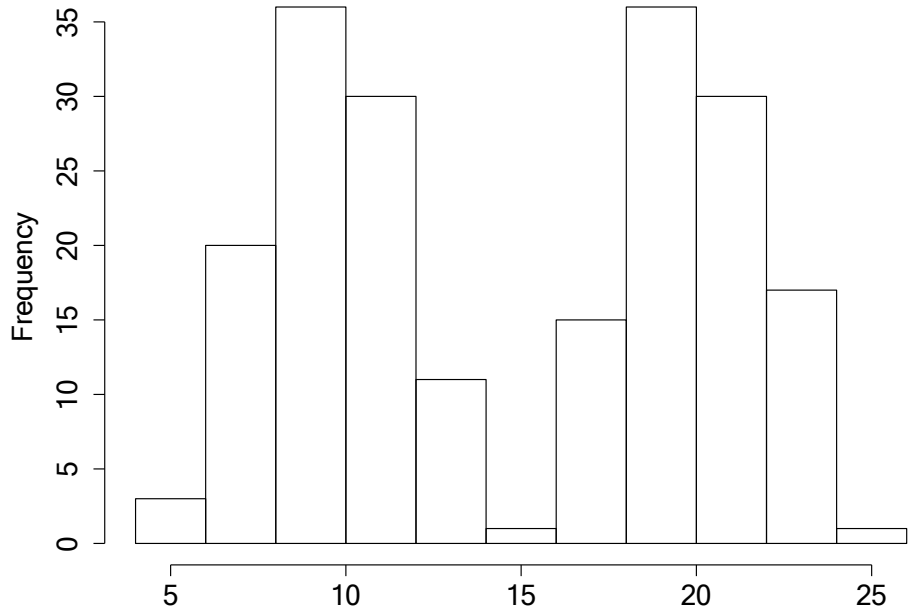
- Example: the mode of the following set of numbers is 3

1, 1, 2, 2, 2, **3, 3, 3, 3**, 4, 5, 5, 6, 6, 7, 8, 10, 13



The Mode

- **Advantage** The Mode always a value that occurs in the data and can be used with categorical data that are labels rather than numbers(such as gender or type).
- **Disadvantage** There may be two or more values that share the largest frequency. In the case of two modes, we would report both and refer to the distribution as *bimodal*.



The Median

- **The Median** can be thought of as the 'middle' value i.e. the value for which 50% of the data fall below when arranged in numerical order.
- For example, consider the numbers

15, 3, 9, 21, 1, 8, 4,

- When arranged in numerical order

1, 3, 4, 8, 9, 15, 21

- we see that the median value is 8.
- If there were an even number of scores, e.g.

1, 3, 4,8, 9, 15

then we take the midpoint of the two middle values. In this case the median is (4 + 8)/2 = 6.

The Median

- In general, if we have *n* data points then the **median location** is defined as follows:

$$\text{Median Location} = \frac{n+1}{2}$$

- For example, the median location of 7 numbers is (7 + 1)/2 = 4 and the median of 8 numbers is (8 + 1)/2 = 4.5, i.e. between observation 4 and 5 (when the numbers are arranged in order).

- **Advantage** The median is unaffected by extreme scores (a point it shares with the mode). We say the median is **resistant** to outliers. For example, the median of the numbers

1, 3, 4, **8**, 9, 15, 99999

is still 8. This property is very useful in practice as *outlying* observations can and do occur (Data collected in practice is often messy).

The Mean

- **The Mean** is calculated by adding all values and dividing by how many values there are. For example, the mean of

$$1, 3, 4, 8, 9, 15 \quad \text{is} \quad \frac{1 + 3 + 4 + 8 + 9 + 15}{6} = 6.667 \quad \text{(to 3 dp)}$$

- In mathematical notation, the mean of a sample of n numbers $x_1, \ldots, x_n$ is denoted by $\bar{x}$ where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{or} \qquad \bar{x} = \frac{\sum x}{n}$$

The symbol Σ (sigma) means “add up”.

For a **population**, the mean is denoted by **μ** where:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \text{or} \qquad \mu = \frac{\sum x}{N}$$

The Mean

- **Advantage** The mean is the most widely used measure of location. Historically, this is because statisticians can write down equations for the mean and derive nice theoretical properties for the mean, which are much harder for the mode and median.

- **Disadvantage** The mean is *not resistant* to outlying observations. For example, the mean of

  - 1, 3, 4, 8, 9, 15, 99999

  is 14323.57, whereas the median (from before) is 8.

The Mean

- Sometimes discrete measurement data are presented in the form of a frequency table in which the frequencies of each value are given.

Data (x)	1	2	3	4	5	6
Frequency (f)	2	4	6	7	4	1

- We calculate the sum of the data as

$$(2 \times 1) + (4 \times 2) + (6 \times 3) + (7 \times 4) + (4 \times 5) + (1 \times 6) = 82$$

- and the number of observations as

$$2 + 4 + 6 + 7 + 4 + 1 = 24$$

- The the mean is given by

$$\bar{x} = \frac{82}{24} = 3.42 \text{ (rounded to two decimal places)}$$

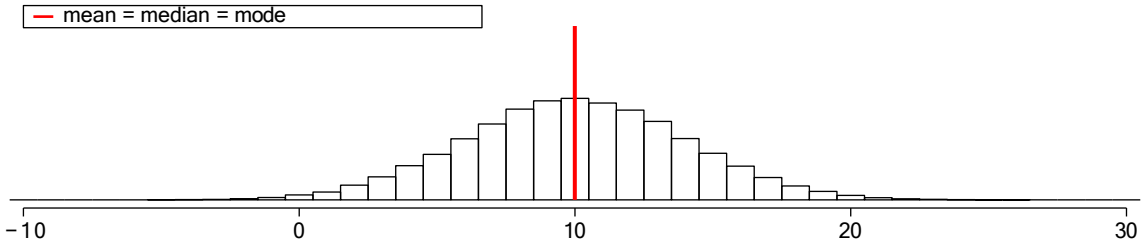
The Mean

In mathematical notation the formula for the mean of frequency data is given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad \text{or} \qquad \bar{x} = \frac{\Sigma fx}{\Sigma f}$$

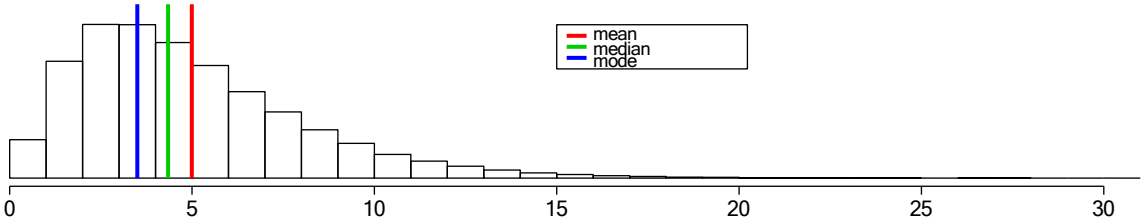
The relationship between the mean, median and mode

Symmetric Data



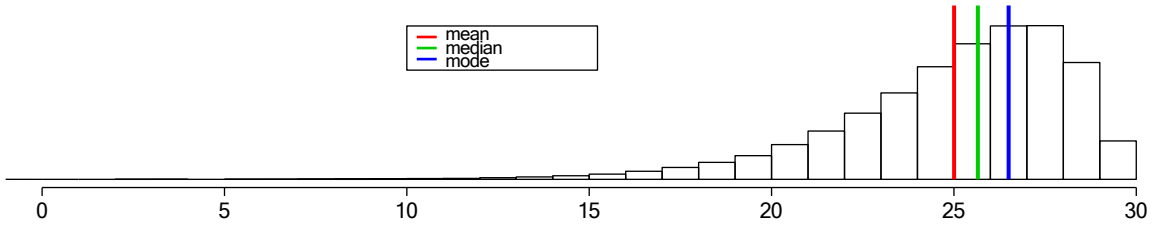
mean=median=mode

Positively Skewed (right skewed) data



mean>median>mode

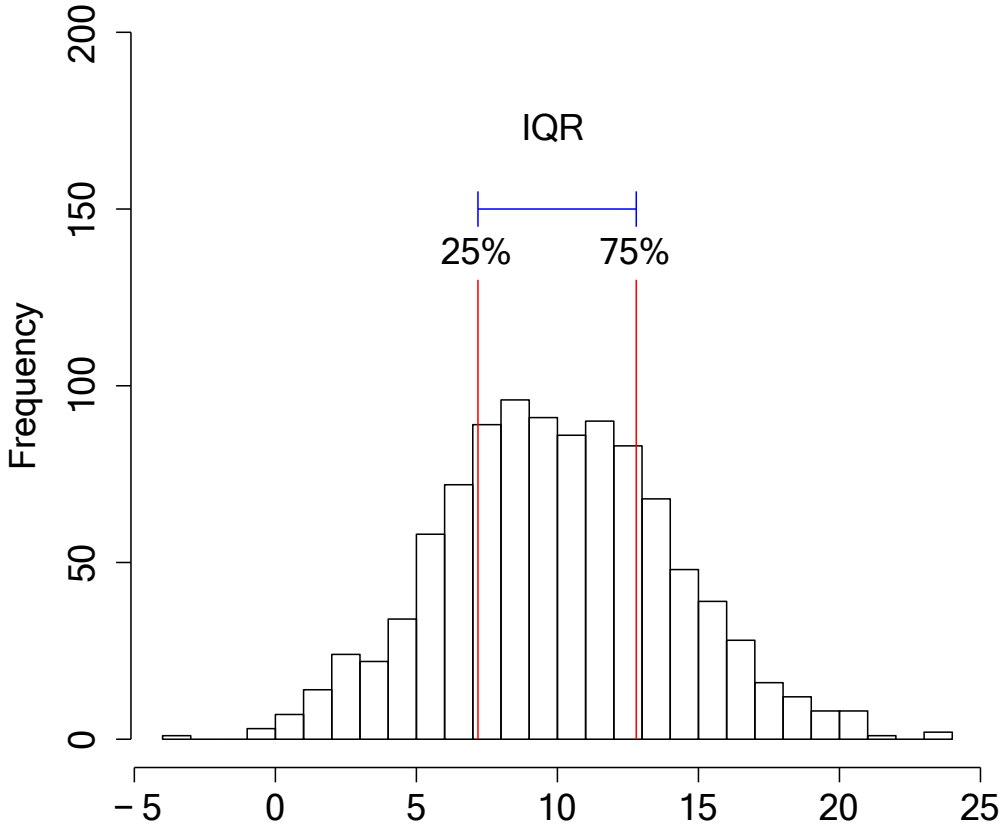
Negatively Skewed (left skewed) data



Mean<median<mode

IQR and SIQR

- **IQR** stands for **Interquartile Range**
- **SIQR** stands for **Semi-Interquartile Range**
- The IQR is the range of the middle 50% of the data. The SIQR is simply half the IQR.



IQR and SIQR

We calculate the IQR in the following way:

1. Calculate the 25% point (**1st quartile, or Q1**) of the dataset. The location of the 1st quartile is defined to be the $\frac{n+1}{4}$ th data point.
2. Calculate the 75% point (**3rd quartile, or Q3**) of the dataset. The location of the 3rd quartile is defined to be the $\frac{3(n+1)}{4}$ th data point.
3. Calculate the IQR as

$$\text{IQR} = Q3 - Q1$$

Example:

Consider the set of 11 numbers (which have been arranged in order)

$$10,\ 15,\ 18,\ 33,\ 34,\ 36,\ 51,\ 73,\ 80,\ 86,\ 92$$

The 1st quartile is the

$$\frac{11+1}{4} = 3rd\ data\ point = 18$$

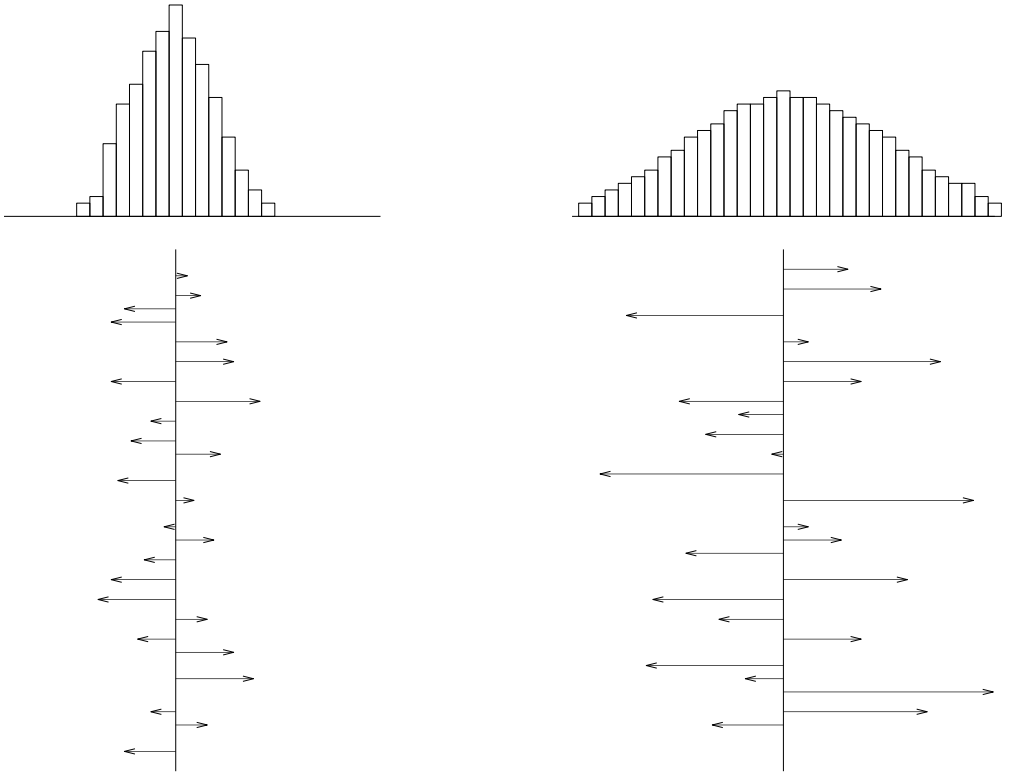
The 3rd quartile is the

$$\frac{3(11+1)}{4} = 9th\ data\ point = 80$$

$$\Rightarrow IQR = 80 - 18 = 62 \qquad \Rightarrow \qquad SIQR = \frac{62}{2} = 31$$

The Mean Deviation

- To measure the spread of a dataset it seems sensible to use the ‘deviation’ of each data point from the mean of the distribution. The deviation of each data point from the mean is simply the data point minus the mean.



small spread = small deviations

large spread = large deviations

The Mean Deviation

Data x	Deviations $x - \bar{x}$
10	10 - 48 = -38
15	15 - 48 = -33
18	18 - 48 = - 30
33	33 - 48 = -15
34	34 - 48 = -14
36	36 - 48 = -12
51	51 - 48 = 3
73	73 - 48 = 25
80	80 - 48 = 32
86	86 - 48 = 38
92	92 - 48 = 44
$\Sigma x = 528$	$\Sigma (x - \bar{x}) = 0$

n=11

The mean is $\bar{x} = \frac{528}{11} = 48$

The Mean Deviation of a set of numbers is simply mean of deviations.

In practice, the mean deviation is *always* zero.

Mean Absolute Deviation (MAD)

- We solve the problem of the deviations summing to zero by considering the **absolute values** of the deviations.

- The absolute value of a number is the value of that number with any minus sign removed, e.g. $|-3| = 3$.

- We then measure spread using the mean of the absolute deviations, denoted (MAD).
- This can be written in mathematical notation as

$$\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \quad \text{or} \quad \frac{\sum |x - \bar{x}|}{n}$$

Mean Absolute Deviation (MAD)

Data x	Deviations $x - \bar{x}$	|Deviations| $\|x - \bar{x}\|$
10	10 - 48 = -38	38
15	15 - 48 = -33	33
18	18 - 48 = - 30	30
33	33 - 48 = -15	15
34	34 - 48 = -14	14
36	36 - 48 = -12	12
51	51 - 48 = 3	3
73	73 - 48 = 25	25
80	80 - 48 = 32	32
86	86 - 48 = 38	38
92	92 - 48 = 44	44
$\Sigma x = 528$	$\Sigma (x - \bar{x}) = 0$	$\Sigma \|x - \bar{x}\| = 284$

$$\text{MAD} = \frac{284}{11}$$

$$= 25.818$$

(rounded to 3dp)

The Sample Variance (s^2) and Population Variance (σ^2)

- Another way to ensure the deviations don't sum to zero is to look at the squared deviations. Thus, another way of measuring the spread is to consider the mean of the squared deviations, called the **Variance.**
- If our dataset consists of the whole **population** (a rare occurrence) then we can calculate the population variance σ^2 (said 'sigma squared') as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{N} \qquad \text{or} \qquad \sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$$

- When we just have a **sample** from the population (most of the time) we can calculate the sample variance s^2 as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad \text{or} \qquad s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$$

Note: We divide by $n - 1$ when calculating the sample variance as then s^2 is a '**better estimate**' of the population variance σ^2 than if we had divided by n.

The Sample Variance (s^2) and Population Variance (σ^2)

Data x	Deviations $x - \bar{x}$	Deviations^2 $(x - \bar{x})^2$
10	10 - 48 = -38	1444
15	15 - 48 = -33	1089
18	18 - 48 = - 30	900
33	33 - 48 = -15	225
34	34 - 48 = -14	196
36	36 - 48 = -12	144
51	51 - 48 = 3	9
73	73 - 48 = 25	625
80	80 - 48 = 32	1024
86	86 - 48 = 38	1444
92	92 - 48 = 44	1936
$\Sigma x = 528$	$\Sigma (x - \bar{x}) = 0$	$\Sigma (x - \bar{x})^2 = 9036$

$$s^2 = \frac{9036}{11 - 1} = 903.6$$

The Sample and Population Standard Deviation (s and σ)

- Notice how the sample variance in our example is much higher than the SIQR and the MAD.

$$\text{SIQR} = 31 \quad \text{MAD} = 25.818 \quad s^2 = 903.6$$

- This happens because we have squared the deviations transforming them to a quite different scale.
- We can recover the scale of the original data by simply taking the square root of the sample (population) variance.
- The square root of the variance is called the **standard deviation**, and it is the **most widely used measure of spread** in statistics.

The Sample and Population Standard Deviation (s and σ)

- Thus, we define the sample standard deviation s as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

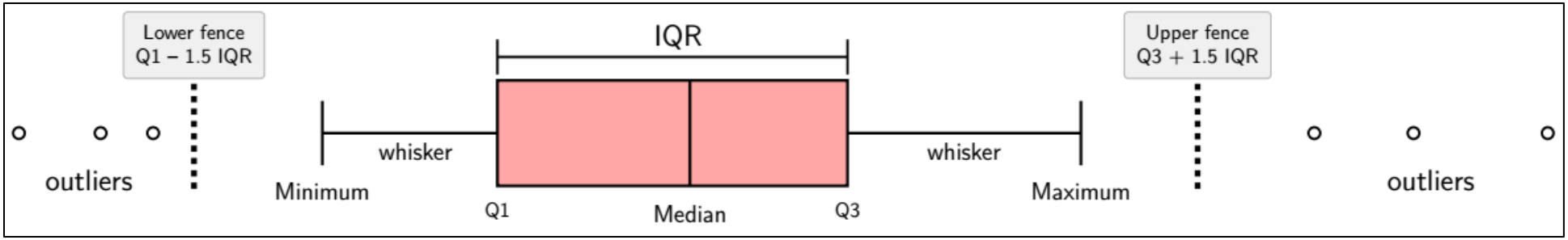
- and we define the population standard deviation σ as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{\mathrm{N}}(x_i - \mu)^2}{\mathrm{N}}}$$

- Returning to Example 1 the sample standard deviation is
  - $s = \sqrt{903.6} = 30.05$ (to 2dp)

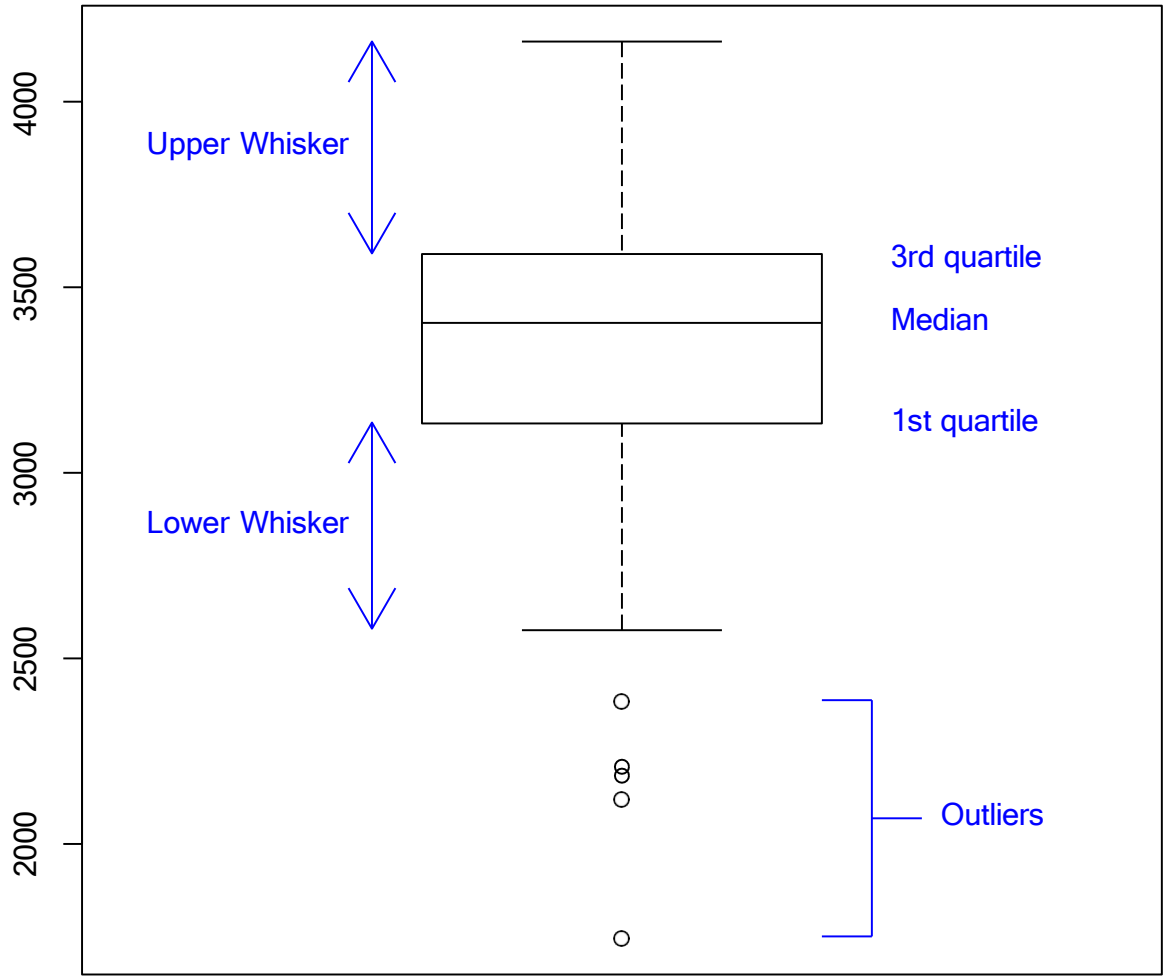
which is comparable with the SIQR and the MAD.

Box Plot



- Also known as the **five-number summary** in statistics
- Includes **minimum, Q1, median, Q3, and maximum**
- The **box** spans from **Q1 to Q3** (the IQR)
- The **lower whisker** ends at the smallest value that is **not an outlier**
- The **upper whisker** ends at the largest value that is **not an outlier**
- Outliers are identified using the **lower and upper fences**
  - Lower fence=$Q1 - 1.5 \times IQR$
  - Upper fence=$Q3 + 1.5 \times IQR$
- Any data value beyond the whiskers is classified as an **outlier**
- Commonly used to **compare distributions between groups**

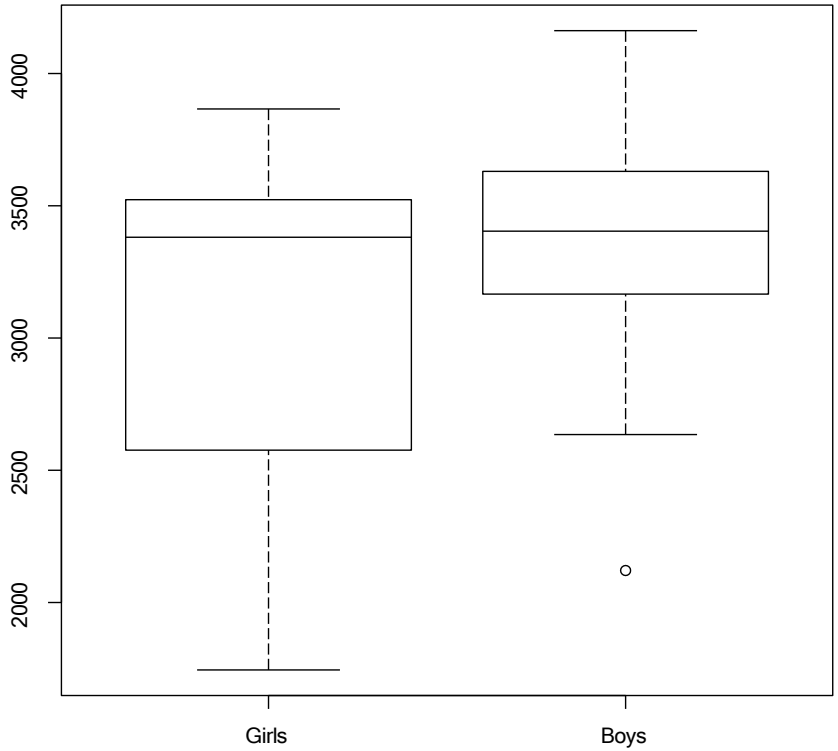
Box Plot of Birth Weight (All Newborns)



- This boxplot summarizes birth weights for all newborns, combining boys and girls
- It shows the overall center and spread of birth weights in the dataset
- The plot shows variability and outliers in birth weight

Boxplot Comparison of Birth Weights by Gender

Side-by-side boxplots compare birth weight distributions for boys and girls



Observations:

- The girls' boxplot shows left skewness, as the median is not centered.
- Differences in medians are small, suggesting typical birth weights of boys and girls are very similar
- The boxes overlap, indicating similar typical ranges, but girls have greater variability.
- We see that a low outlier is present in the boys' birth weight data

Summary

- **Statistics** is the science of collecting, organizing, analyzing, and interpreting data
- Two branches of statistics: **Descriptive** and **Inferential**
- The **statistical process** moves from questions → data → analysis → conclusions
- **Data** are recorded values from observations or experiments
- **Variables** are characteristics that can take different values
- Variable types: **Quantitative** (numerical) and **Qualitative** (categorical)
- Quantitative variables can be **Discrete** or **Continuous**
- Qualitative variables can be **Nominal** or **Ordinal**
- **Population vs Sample**: **N** = population size, **n** = sample size
- Measures of **location (center)**: **mean, median, mode**
- Measures of **spread (variability / dispersion)**: **range, IQR, variance, standard deviation**
- Data can be summarized using **numerical measures** and **graphs** (bar charts, histograms, scatter plots, box plots)

Assessment Problems

1. Classify each variable: CPU usage (%), operating system type, number of login attempts, response time (ms).

2. Given response times 180, 120, 130, 130, 150, 300 ms, find the mean, median, mode, & IQR.

3. Choose the best graph for:
   (a) Count of different error types
   (b) How long programs take to run
   (c) file size vs. download speed
   (d) Comparing exam scores between two sections of a class

4. The following dataset represents exam scores of 10 students:

$$45,\ 52,\ 60,\ 62,\ 65,\ 70,\ 72,\ 75,\ 90,\ 95$$

   a) Find the five-number summary
   b) Identify any outliers using the 1.5×IQR rule
   c) Draw a boxplot for the data and label Q1, median, Q3, whiskers, and outliers