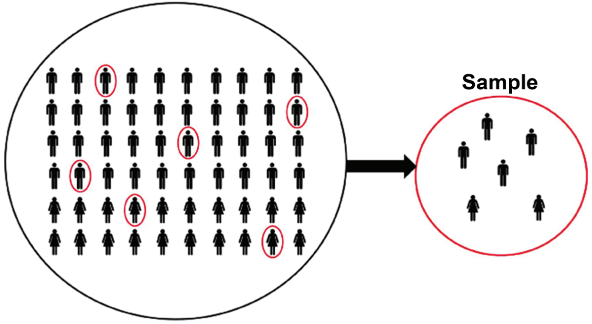
## 8. Some Basic Statistical Concepts (very brief)

It is rarely feasible to collect data for all elements in a group. We often use a sample, a “representative” subset.



- **Population** is all elements in a group and **sample** is a subset of the population.

**Statistics and parameters**

- A parameter is a descriptive measure computed from the data of the population.
- A statistic is a descriptive measure computed from the data of the sample.

**Statistics is a field of study with two major areas:**

- **Descriptive Statistics**
  - collection, organization, summarization, visualisation, and analysis of data
- **Inferential Statistics**
  - drawing of conclusions/inferences about a large body of data (population) when only a part of the data is observed (sample).

- **Data** – the information we generate or gather with experiments, observations, and with surveys.
- **Variable** – a characteristic or attribute that can assume different values in different persons, places, or things, etc.
- A **random variable** (r.v.) is one that cannot be predicted in advance because it is influenced by too many factors (or chance).
  - For instance, height cannot be predicted at birth.
- Are the following r.v. Qualitative or Quantitative? If Quantitative, are they discrete or continuous ?
  - Number of car accidents a person had
  - Marital status of patients
  - Height
  - Age

**Measures of central tendency:**

If you have some quantitative data, you could compute:

- The **mean**:
  - The arithmetic mean (typically referred to as simply the **mean**,) is the "average" which is obtained by adding all the values in a sample or population and dividing by the number of values.
  - Generally, this value is not part of the data set.
  - The mean is a “balance point” of the data.
  - Not robust to outliers (extreme values).
  - Typically denoted:
    - $\bar{x}$ for a sample mean and
    - $\mu$ for a population mean.

- The **mode**, the most frequently occurring value.
  - This center measure can also be used for qualitative/categorical data too.
  - Must be a value in the data set.
- The **median**, the value in the middle of the sorted data
  - The value in the middle of the data set.
  - or average of the two middle values if even number of observations.
  - Robust to outliers.
- The **midrange**, the average of the minimum and the maximum value.
  - Uses only two values to compute so a very crude measure.

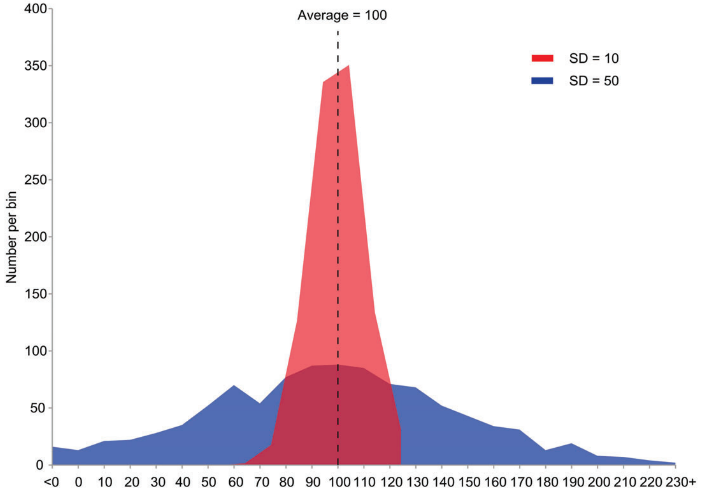
- All four measures of center give a single number, a **typical** or **center value**, to represent the entire dataset.

### Measures of Dispersion

- Each measure of dispersion gives a a positive real number summarizing the dispersion or spread of the data about the center.
  - The value of the dispersion measure **increases as the variability** of the data values increases.
- The 3 most common measures of dispersion(absolute) are **range**, **variance**, and **standard deviation.**
- Dispersion measure will be **zero** if all data values are the same.

### Standard Deviation

- By far the most used dispersion measure.
  - Intuitively it is conveying the “average” difference between the data points and the mean.
  - Gauss used to call this measure the mean error.
- Standard Deviation (SD) is with the same units as the data.
- As with any dispersion measure, the **higher** the value of the standard deviation, the **more spread out** the data values in the sample.
- Conversely, the **lower** the value of the standard deviation, the more **closely clustered to the mean** are the data values in the sample.



J.R.Brown, https://commons.wikimedia.org/w/index.php?curid=10777712

- The (definition) formula to compute the sample standard deviation, ***s***, is:

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

  - The formula involves the squared differences (squared deviations).
  - For technical reasons, for sample standard deviation, $s$, we divide by $n - 1$, the sample size minus one.
  - We take a square root and get back to the original units.

- **Variance** is the **standard deviation squared.**
  - Conversely, the standard deviation is calculated by taking the square root of the variance.

- **Range** is the **difference** between the **max** and the **min,** the highest and the lowest data values.
  - To compute the range, you ignore most of the data.
  - Unusually high or unusually low values will influence the range very much (sensitive to outliers).
  - Range is mainly reported in addition to the standard deviation.

**Two Main Types of Statistical Studies:**

- **Observational study** – researcher observers what is happening or what has happened in the past and tries to draw conclusions based on these observations
  - Example: smoking status
- **Experimental study** – researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.
  - Example: Drug studies: effects of taking aspirin on heart attacks

In an experiment there are at least two groups: the treatment group(s) and a control group.

- **Treatment Group** – A group in the sample that receives a treatment or experimental condition.

- **Control Group** – A group in the sample who are treated identically in all respects except that they don’t receive active treatment.
  - Using a control group allows us to see what would have happened to the response variable if treatments had not been applied.

- **Placebo** – deals with drugs – looks like a real drug but has no active ingredient.
    - In an experiment 1/2 of the people receive real treatment and 1/2 receive the placebo without knowing who is taking what.

- **Placebo Effect** – when people take the placebo, and it works almost like the drug studied.
    - This is usually because of psychological reasons…our minds are powerful!
    - Note: In true experimental studies, the subjects are randomly assigned to groups and the treatment is randomly assigned to the groups.

- **Independent Variable** – the variable that is being manipulated by the researcher (also called the explanatory variable).
- **Dependent Variable** – the response to the independent variable or the result to the explanatory variable (also called the response or outcome variable).
    - Example: Time spent studying outside of class and grades
        - What is the explanatory (independent) variable?
        - What is the response (dependent) variable?
- A **confounding variable** is one that influences the dependent or outcome variable but cannot be separated from the independent variable.

**Advantages of Experiments over Observational Studies:**

- More accurate study of explanatory variable on response variable.
- Researchers have control over participants, groups, and independent variables.
- Needed to establish cause and effect relationship.
- Experiments may not always be practical.

**Disadvantages of Experiments:**

- May occur in unnatural settings.
- Subjects may change behavior (Hawthorne Effect).
- Not all variables can be controlled for.

**Frequency Distribution**

Large data sets can be summarized, and we can gain some insight into the nature of data.

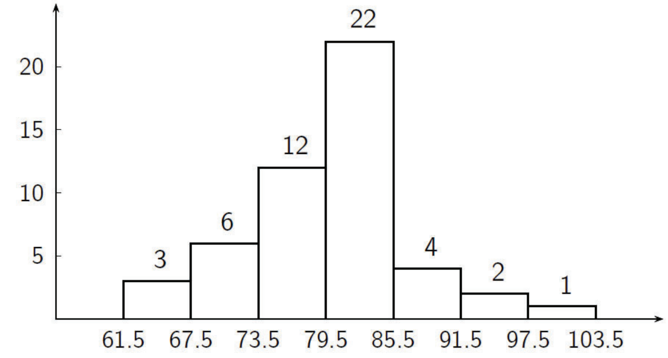
- Basis for constructing graphs.
- Groups the data in classes (bins)

Example: The heart rates (beat/min.) of 50 students are given below, construct a frequency distribution.

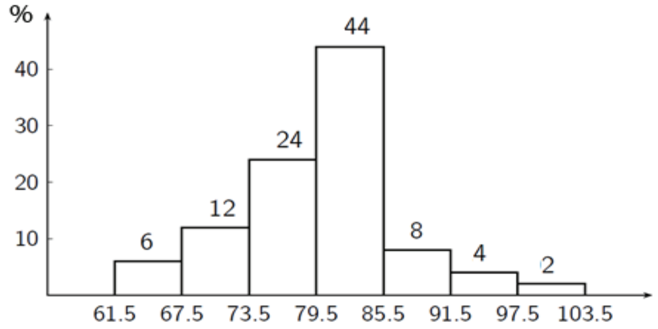
77 84 79 90 67 84 82 74 69 81 94 68 65 86 78 79 83 83 84 82 93 80 81 80 62
98 77 83 82 80 82 73 77 79 81 70 72 85 84 80 83 77 80 70 75 74 85 87 79 88

| Class Limits | Frequency | Class Limits | Frequency |
|---|---|---|---|
| 62 – 67 | 3 | 86 - 91 | 4 |
| 68 – 73 | 6 | 92 - 97 | 2 |
| 74 - 79 | 12 | 98 - 103 | 1 |
| 80 - 85 | 22 | | |

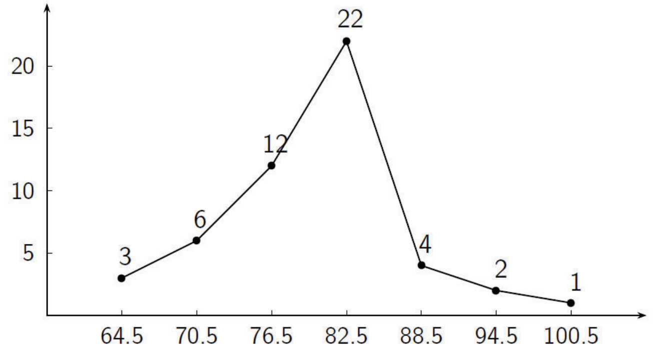
**Histogram** plots frequencies (counts) vs. class boundary.



**Relative frequency histogram** plots relative frequencies (percentages) vs. class boundary.



**Frequency polygon** plots frequency vs. class midpoint with connected line segment.



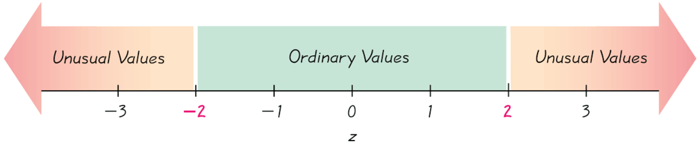
**You should always use a histogram or frequency polygon to see the "shape" of the data.**

**Measures of Relative Standing**

- A **standard value**, or ***z*-value** (sometimes called **standardized value**), is the number of standard deviations that a given data point *x* is above or below the mean.
- Z-value for a sample or a population is

$$\boldsymbol{Z-value = \frac{Value - Mean}{St.deviation}}$$

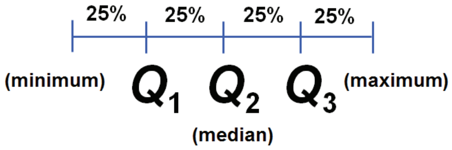
- Some books define:
  - ○ "Unusual" values: z < 2 or z > 2
  - ○ "Ordinary" values: $-2 \leq z \leq 2$



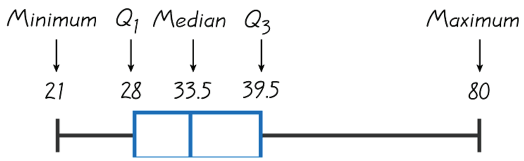
## The Quartiles Divide Sorted Data into Four Equal Parts

(equal number of data values)

- The **first quartile** denoted by $\mathbf{Q_1}$ (also called **the lower quartile**), separates the bottom 25% of all values from the top 75%. It is the median of the lower half of the data.
- The median is also the **second quartile**, denoted by $\mathbf{Q_2}$.
- The **third quartile**, denoted by $\mathbf{Q_3}$ (**the upper quartile**), separates the top 25% of all values from the bottom 75%. It is the median of the upper half of the data.



- 5-Number Summary consists of the Min, Q1, Med, Q3, and Max. A quick way to visualize these numbers is the boxplot.
- A **boxplot** (or **box-and-whisker diagram** or **box-and-whiskers plot**) is a graph that consists of:
  - a line extending from the Min to the Max, and a box with lines drawn at the Q1, Med, and Q3.



- **Interquartile Range (IQR) = $\mathbf{Q_3 - Q_1}$**

- An **outlier** is a value that is located very far away from almost all of the other values. Relative to the other data, an outlier is not a typical value.
  - An outlier can have a dramatic effect on the mean, on the standard deviation, and on the scale of the graphs so that the true nature of the distribution is obscured.
- There are different "rules" for finding outliers. Here is one:
  1. First, compute the IQR.
  2. Any value greater than Q3+(1.5*IQR) or less than Q1−(1.5*IQR) will be labeled outlier.
- We can draw a **modified boxplot**, with the outliers indicated with × or *.

## Normal Distribution

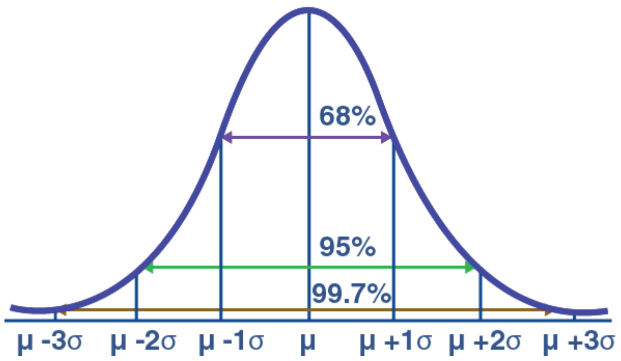
- a continuous **probability distribution** (think shape of the data) associated with many sets of **real-word data**.
  - Generally, data for phenomena that are influenced by many different factors behave approximately normal.
- Importantly, the Central Limit Theorem
  - for a "large enough" sample size the **distribution of the sample mean** is approximately normal.
    - Thus, we can infer characteristics of a population mean without making assumption for the population distribution.

- The formula defining the distribution is:

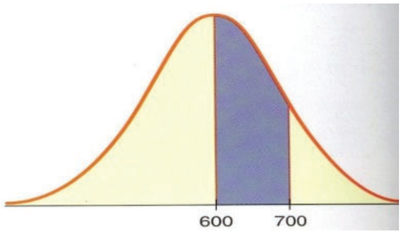
$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where $\boldsymbol{\mu}$ is the **mean**, the center of the data,
and $\boldsymbol{\sigma}$, the **standard deviation**,
defines the “spread” of the data.

- Physicists and Engineers often call it the "**Gaussian** distribution."
- In Social Sciences it is common to refer to it as the "**bell curve**".



- The shape of the distribution is indeed like a bell; it is symmetric (centered around the mean).
- The bulk of the data coming from such a distribution is within 2 standard deviations from the mean.



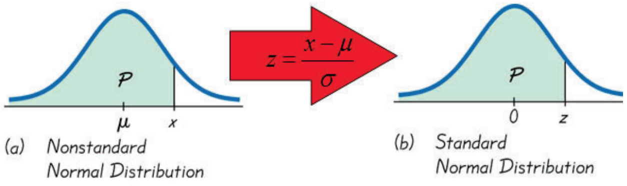
Def. A density curve is a graph of a continuous probability distribution.

Properties: 1. The total area under the curve and the horizonal axis must equal 1.

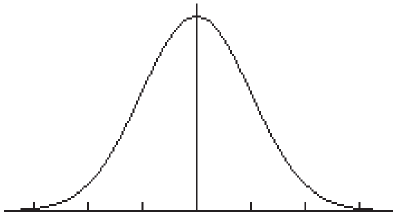
2. Every point on the curve must have a vertical height that is 0 or greater.

- Because the total area under the density curve is equal to 1, there is a correspondence between **area** and **probability**.
- The **probability** that a random selection falls within that interval **equals** the **area under the curve** over the interval.

- Traditionally, Introductory Statistics courses introduce conversion to standard normal for technical reasons.

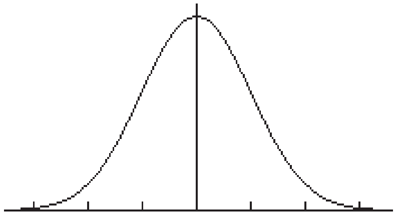

P
μ
x
(a) Nonstandard Normal Distribution
$z = \frac{x-\mu}{\sigma}$
P
0
z
(b) Standard Normal Distribution

- When using R (or even a calculator) there is actually no need to do this conversion.



Ex.: The sitting height (from seat to top of head) of drivers must be considered in the design of a new car model. Suppose that males have sitting heights that are normally distributed with a mean of 36.0" and a standard deviation of 1.4". Engineers have provided plans that can accommodate males with sitting heights up to 38.8", but taller males cannot fit comfortably.

If a male is randomly selected, find the probability that he has a sitting height less than 38.8". Based on that result, is the current engineering design feasible?



Ex.: In designing seats to be installed in commercial aircraft, engineers want to make the seats wide enough to fit 98% of all males. Men have hip breadths that are normally distributed with a mean of 14.4" and a standard deviation of That is, find the hip breadth of men that separates the bottom 98% from the top 2%.

## 9. Basic Plotting

(More plotting to come later)

Good graphics are essential in data analysis.

- They help us avoid mistakes.
- They help us decide on a model.
- They help communicate the results of our analysis.
- Graphics can be more convincing than text many times.

*"A picture is worth a thousand words."*

- The plotting capabilities of R are one of its very attractive features.
- It is relatively simple to construct histograms, (parallel) boxplots, scatterplots, etc. with the built-in functions.
- A histogram is created using the `hist` function.
- A boxplot is created using the `boxplot` function.
- A scatterplot is created using the `plot` function.

Many additional R packages exist for "fancier" graphics. For example, `ggplot2` package is often recommended, and we will do examples with `ggplot2` soon.

Built-in R functions are pretty good too…