Question: What can we learn about a *population* from data collected from a *sample* of the population?

Terminology:

- **Data:** observations (e.g., heights, temperatures, genders, cultural backgrounds, household sizes, etc.) that have been collected.
- **Population:** The set of *all* members of the group being studied.
- **Census:** The collection of (relevant) data from the entire population.
- **Sample:** A subset of the population.
- **Statistics:** The scientific field concerned with the collection, organization, description and analysis of data.

- **Parameter:** A number that describes a characteristic of the *population*.
- **Statistic:** A number that describes a characteristic of a *sample*.

$\Longrightarrow$ Parameters and statistics are calculated from data.

Data Types:

- **Quantitative data:** data that are naturally represented by numbers, e.g., measurements or counts.
- **Qualitative data:** non-numerical data that can be separated into different categories, e.g., gender, nationality, species. Qualitative data is also called *categorical data*.

Types of quantitative data:

- **Discrete:** The data comes from a specific finite or *countable* set. Typically arises from *counting* something.

  Examples:

  - The number of children in a family.
  - The length of the line at a checkout counter.
  - The number of coin tosses before we see '*heads*' for the first time.

- **Continuous:** The data can take any value from some interval on the real line. Typically arises from *measuring* something.

  Examples:

  - The height of a randomly selected UCSC student.
  - The temperature at noon at the corner of Pacific and Laurel.
  - The weight of an adult elephant seal.

Four levels of measurement:

- **Nominal:** data that consists of *names* or categories that cannot be arranged in an (objective) order.

  Examples: color, cola brand, movie type.

- **Ordinal:** the data can be arranged in some order — lowest to highest — but the differences between the values cannot be determined or have no real meaning.

  Examples: Grades in a class, belt-rank in a martial art.

(*) Ordinal data allow for comparisons between observations, but do not allow for a meaningful interpretation of the *size* of the difference. We typically do *not* use ordinal data in calculations (with some exceptions).

- **Interval:** Similar to ordinal, but with *meaningful* interpretations of *differences* between different values. But no meaningful interpretation for the *ratio* of two values.

  Examples: Years in which the U.S. went to war (1917, 1941, 1950, 1964, 1990, 2003).

- **Ratio:** Like interval, but with meaningful interpretations of the ratios of different values.

  Examples: Heights of brown bears; household sizes in Santa Cruz.

Very Important: The way that sample data is collected can have a ***very big*** impact on the interpretive value of the data. If we want to draw conclusions about a population from sample data, then the sample should be *representative* of the population.

Terminology:

(*) **Observational study:** observations are made and measurements taken **without modifying** the subjects.

Example: 1200 adults are selected at random and surveyed. Their cola preference, Pepsi or Coke, is observed.

(*) **Experiment:** a *'treatment'* is applied to subjects and the effects are observed.

Example: School children are assigned to two groups. One group is given a polio vaccine and the other is given a placebo. The number of cases of polio in both groups is observed.

Types of observational studies:

(*) **Cross-sectional:** Data are observed and collected at one point in time.

(*) **Retrospective:** Data are collected from the past (by studying existing records).

(*) **Prospective:** Data are (will be) collected at future times from groups sharing common traits.

Comment: In both experiments and observational studies, researchers should control for the effect of **confounding variables**.

(*) This is often trickier to do in observational studies.

Experiments often involve two (or more) groups so the effects of different treatments can be measured. A group of subjects receiving a treatment is called a ***treatment group*** and the group of subjects receiving no treatment is called the ***control group***.

Key point: The treatment and control groups should be similar in every way ***but the treatment*** to limit the effect of confounding variables.

Experimental design elements:

- **Blinding.** Subjects in the treatment and control groups **do not know** which group they are in.
- **Double blinding.** Neither the subjects nor the clinicians observing the subjects know which group the subject is in.
- **Blocks.** Subjects are divided into different blocks or *experimental units*. Each block is split into treatment and control groups. The blocks consists of subjects who share a trait that might affect the outcome of the treatment.

- **Randomization.** Subjects are assigned to treatment and control groups randomly. This means that each subject is just as likely to be in the control group or the treatment group.
- **Randomized block design.** Subjects are first separated into blocks and then each block is randomized separately into treatment and control groups.
- **Sample size.** If the treatment groups are very small, the natural variations in responses to the treatments can disguise the effect of the treatment. The sample sizes should be big enough to avoid this.
- **Replication.** This is repetition of the experiment with different treatments. Replication is effective when we have enough subjects to differentiate between the responses to different treatments.

Sample types.

- **Random sample.** A sample from a population is *random* if every subject from the population has the same chance of being selected for the sample.
- **Simple random sample.** Every possible sample ***of the same size*** has the same chance of being drawn.
- **Systematic sample:** A starting point is selected (at random) and then every k^{th} subject from the population.
- **Convenience sample:** Data is collected in a convenient (easy) manner.
- **Stratified sample:** The population is divided into different subsets (strata) sharing the same characteristic(s), then a sample is selected from each subset (stratum).
- **Cluster sample.** The population is divided into clusters, randomly select some of the clusters and sample *all* subjects in those clusters.

Error types. The difference between a sample statistic and the parameter in the population that it is trying to estimate can be a result of...

(*) **Sampling error:** these are errors due to chance variation between different samples (and between the samples and the population).

(*) **Nonsampling error:** these are errors due to sample data that are incorrectly collected, recorded or analyzed. 'Incorrectly collected' includes data collected from **biased** samples — samples that are different from the population in a significant way.

> ***Data collected from a poorly chosen sample may be completely useless for making meaningful inferences about the population from which it came.***

Describing data

Frequency distributions.

A *frequency distribution* lists the data values together with the frequency (number of times) the values occur.

(*) The data values can be listed individually when the number of different values is relatively small. This is typical for categorical data and discrete quantitative data.

(*) When there is a large set of possible values, the data are typically collected in *classes* (intervals) and the frequency for each class is listed. This is typical for continuous data.

(*) Frequency distributions can be displayed in tables or in graphs (histograms)

Example.

Heights of adult Women

Height (cm)	Frequency
56.0 – 57.9	1
58.0 – 59.9	4
60.0 – 61.9	8
62.0 – 63.9	12
64 – 65.9	7
66.0 – 67.9	7
68.0 – 69.9	1

(*) The intervals of values are called *classes* or *bins*.

Example. Instead of listing the frequency, we can also list the *relative frequency*, as below.

Height (cm)	Relative Frequency
56.0 – 57.9	2.5%
58.0 – 59.9	10%
60.0 – 61.9	20%
62.0 – 63.9	30%
64 – 65.9	17.5%
66.0 – 67.9	17.5%
68.0 – 69.9	2.5%

Example. Another way to list frequencies is *cumulatively*, as below.

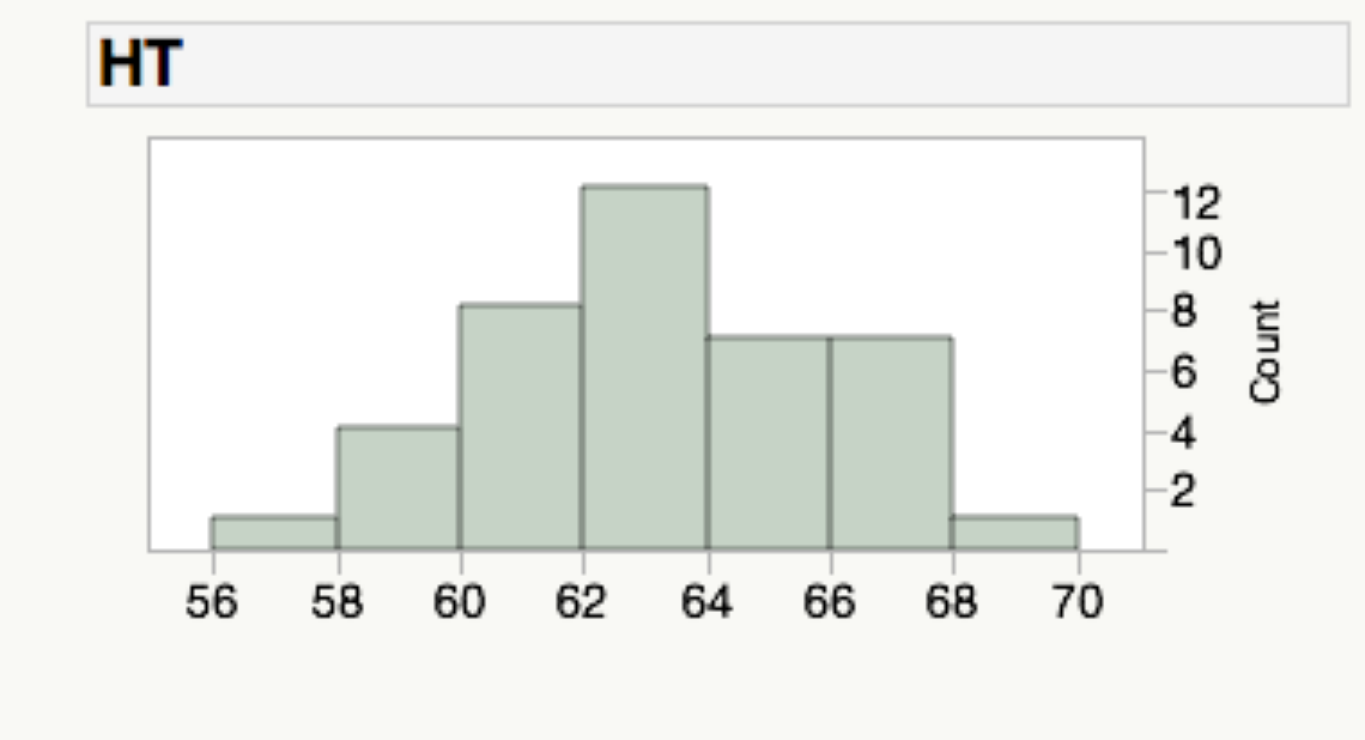
Height (cm)	Cum. Frequency	Cum. Rel. Frequency
56.0 – 57.9	1	2.5%
58.0 – 59.9	5	15%
60.0 – 61.9	13	33.5%
62.0 – 63.9	25	62.5%
64 – 65.9	32	80%
66.0 – 67.9	39	97.5%
68.0 – 69.9	40	100%

Visualizing data.

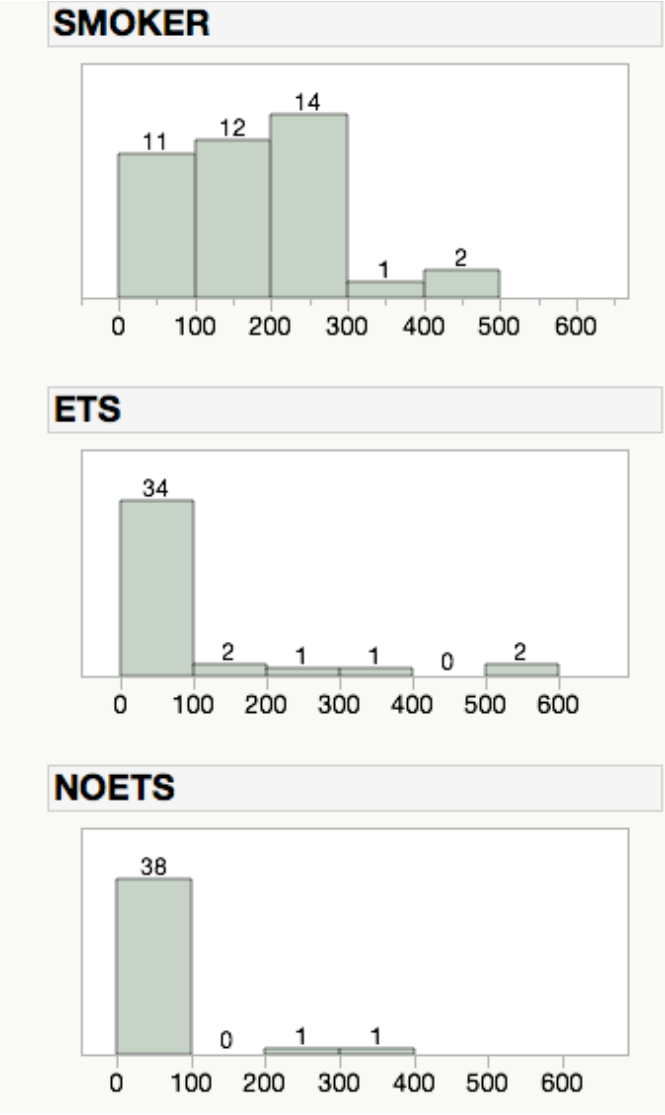
Histograms. A histogram is a visual representation of a frequency distribution.

Example.

Women's heights:



Example. Cotinine levels in smokers, nonsmokers exposed to secondhand smoke and nonsmokers not exposed to secondhand smoke.

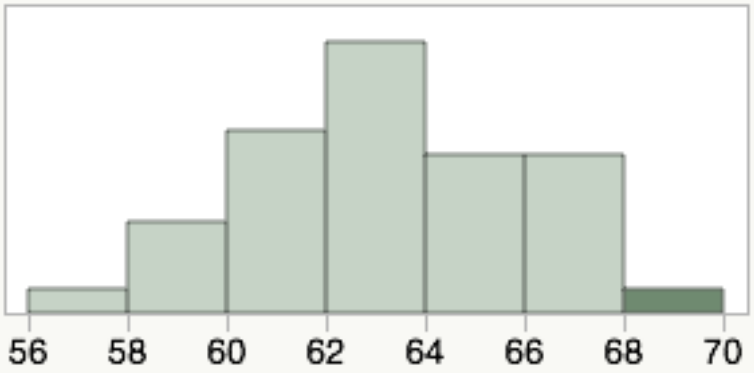


Stem and Leaf plot. Another way of visualizing a frequency distribution.

Example. Women's heights.

Distributions

HT



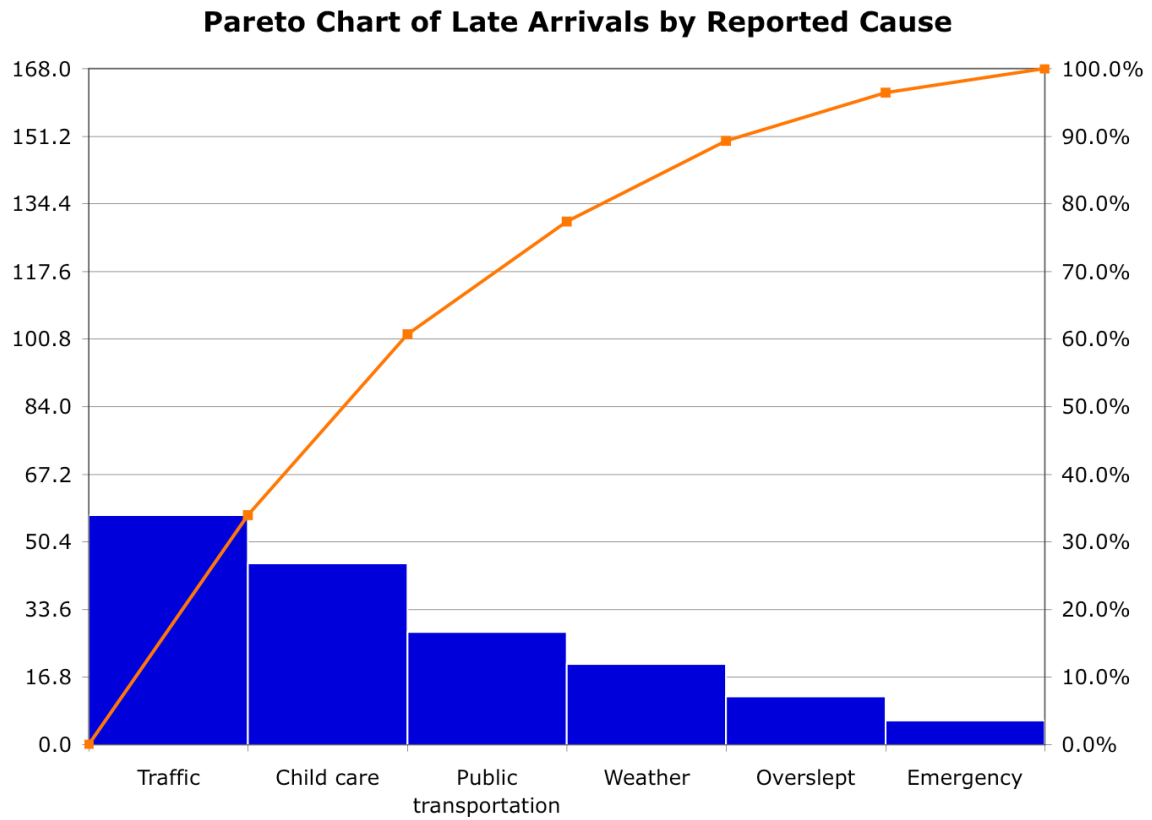
Stem and Leaf

Stem	Leaf	Count
68	**0**	1
67	069	3
66	4778	4
65	01	2
64	13378	5
63	12344456	8
62	3367	4
61	3489	4
60	2567	4
59	68	2
58	26	2
57	0	1

57|0 represents 57.0

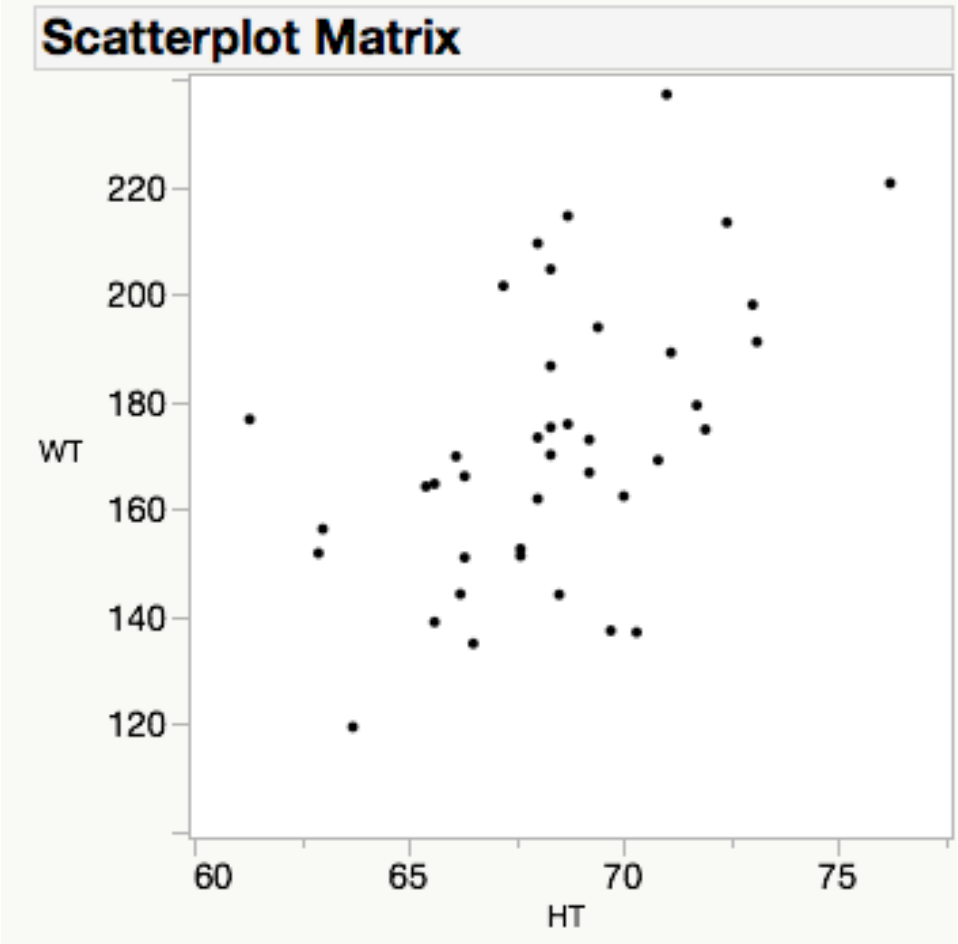
Pareto chart. Gives frequencies of different categories in descending order.

Example. Reasons for arriving late from work (hypothetical, from Wikipedia).

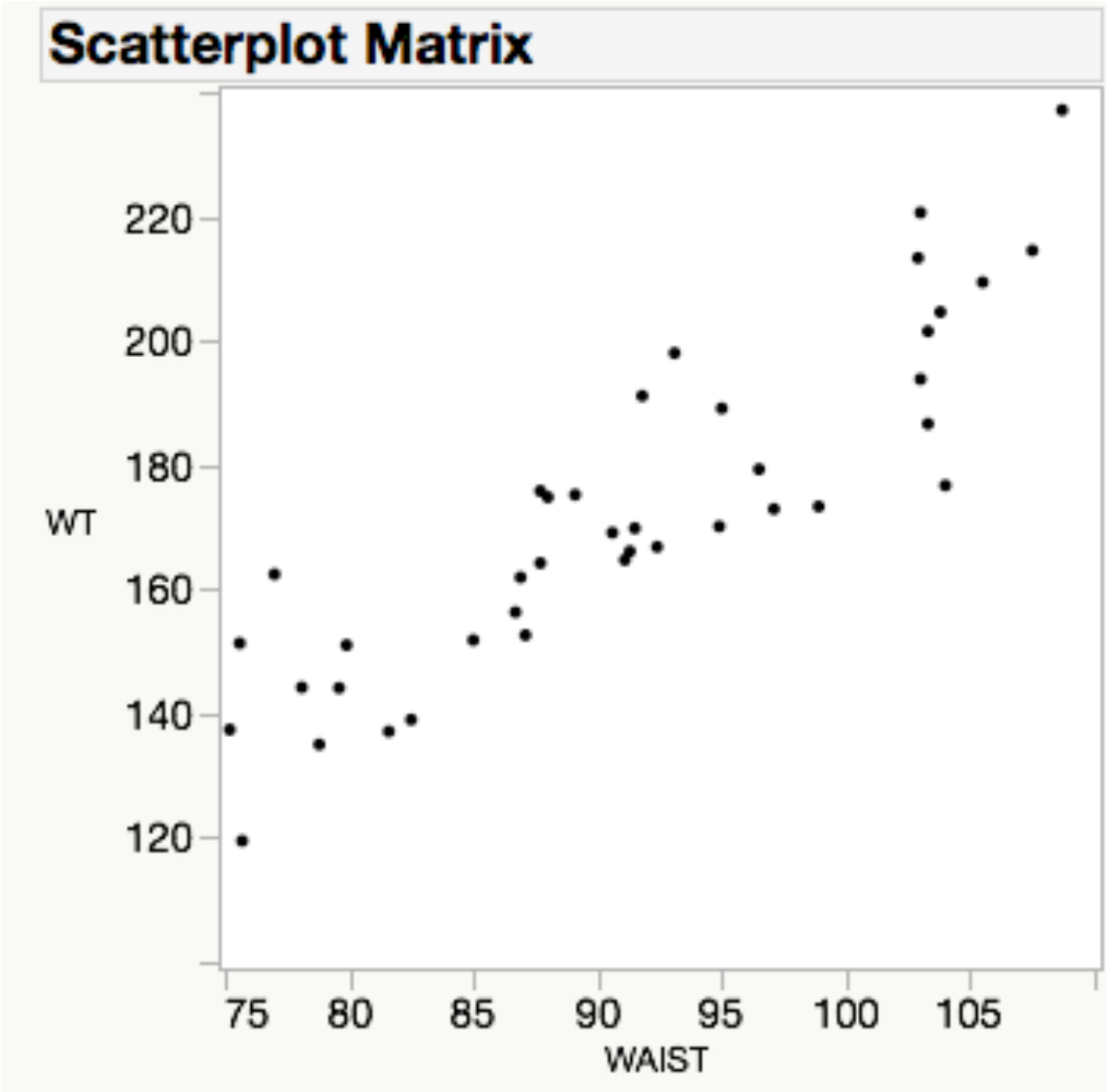


Scatter diagrams. For studying the relation between two variables. Each observation yields a pair (x, y) of values which is plotted as a point in a coordinate system.

Example. Men's heights (inches) and weights (lbs).



Example. Men's waists (inches) and weights (lbs) .



Example. Men's pulses (beats/minute) and weights (lbs).