2 What is Statistics?

Data collection and analysis is fundamental to science in general and bioinformatics in particular.

statistics: a discipline useful for organizing, collecting, summarizing and using data to make predications and justify decisions.

The real world is subject to uncertainty and variation. Statistics provides an intelligent way to deal with this uncertainty and variation.

14 Example (Variation/Uncertainty)
How accurate is a DNA sequence? Has the sequencer made a mistake or have we detected a mutation?

15 Example (Basic Problems in Statistics)
How tall are homo sapiens?

(a) (Sampling Distribution) How much variation is there between one random sample and another random sample?

(b) (Confidence Intervals) What range of heights should we report so that we are 90%, 95%, 99% confident in our reported answer?

(c) (Hypothesis Testing)
H_0 : Men and women have equal heights.
H_1 : Men and women have different heights.
(p-value)

(d) (Regression) How does height depend on age?
(correlation coef. or R^2)

(e) (ANOVA) Does height vary by country?
(p-value)

population: a collection of objects we are interested in.

census data: data about every object in a population.

sample data: data about a subset of a population.

Most of the time, we only have access to information on a very small sample of the population we are interested in. Why? Often because of cost.

Proper sampling of a population is key to the successful application of statistics. You would not, for example, estimate the average height of men from a sample of men taken during basketball practice. (This would be a biased sample.)

One of the best ways to avoid sample bias is to use a random sample.

simple random sample of size n: sampling procedure where every subset of size n from the population has an equal chance of being selected.

sample of convenience: sampling procedure which is not random, but instead is based on convenience.

16 Example (Sampling)
Give examples of both simple random sampling and samples of convenience for various populations.

Solution:

Data comes in two forms, **categorical data** and **numerical data**.

Numerical data can be either **discrete** or **continuous**.

17 Example (Types of Data)

- categorical data: Male, Female.
- numerical data: age (discrete) G.P.A. (continuous)

Data is either **univariate**, **bivariate** or **multivariate**.

18 Example (Types of Data)
Consider the population of homo sapiens.

- univariate data set: (one variable data)
  height
- bivariate data: (two variable data set)
  (height, weight)
- multivariate data: (height, weight, age,...,sex)

3 Descriptive Statistics

Statistics has two branches:
descriptive statistics: summarizes data.
inferential statistics: uses data to make predictions.

Inferential statistics is postponed until we have covered **probability theory**.

Descriptive statistics answers questions like:

- What is a typical height of a homo sapien?
- How much variation is there in the height of a homo sapien?
- Are there gaps in the range of heights of homo sapiens?
- Are there any outliers?

We can summarize a data set using two numbers:

(i) (mean or median) a number which gives a typical value for the data
(ii) (standard deviation or interquartile range) a number which described the variability (level of scatter) of the data

sample mean: $\bar{x}$ where

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

sample median: is the middle data point of a data set. For a data set with an even number of data points, the median is define to be the average of the two middle data points.

19 Example (Median for Even Data Set)
Consider the data set:

$$1,\ 2,\ 4,\ 9$$

The median is given by $\dfrac{2+4}{2} = 3$ (even data set).

The median is less sensitive to the magnitude of individual data points than the mean (i.e. average) is. The median is particularly useful when the data has a skewed distribution.

One measure of the variability of a data set is the **range** of the data, i.e. the largest minus smallest data value, but this statistic is not as robust as some others.

sample variance: s^2 where

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

sample standard deviation: s where $s = \sqrt{s^2}$.

median: divides a data set into two equal halfs.

quartiles: divide a data set into four equal quarters. (sketch)

first (lower) quartile: is the median of lower half of the data.

third (upper) quartile: is the median of upper half of the data.

When computing quartiles, if the data set is odd, include the median in both the lower and upper half.

second (middle) quartile: equals the median

interquartile range (IQR): is equal to upper quartile minus lower quartile.

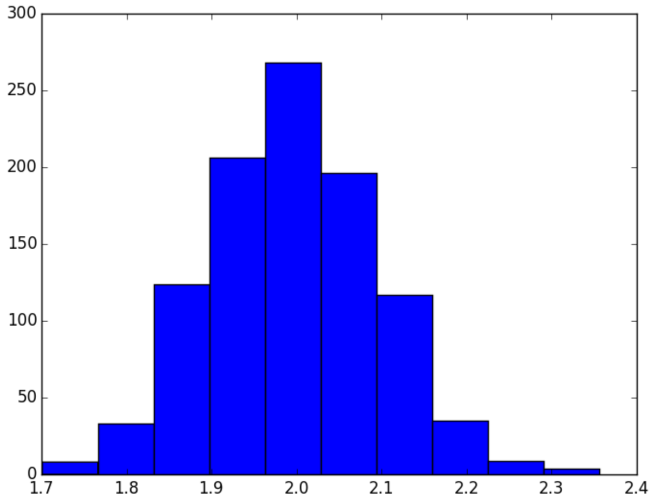
$$\text{IQR} = Q_3 - Q_1$$

The IQR is a measure of data variability/scatter. The IQR is a more robust measure of variability than the standard deviation is.

pth percentile: is the number that divides the data so that p% of the data are less than the number.

Histograms are good for representing large data sets.

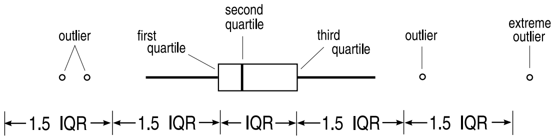
20



Heights of 1,000 Homo Sapiens.

Box plots provide a compact, high level, representation of a data set. Box plots are good for comparing data sets. To compute a box plot you need:

(1) the smallest and largest data point.
(2) the median.
(3) the lower and upper quartiles.



Whiskers extend to smallest/largest data point within 1.5 IQR of the box.

5 Lesson (Box Plots)

Ten measurements are given below. Construct a box plot for these measurements.

$$57,\ 65,\ 66,\ 68,\ 70,\ 70,\ 72,\ 73,\ 75,\ 89$$