## Descriptive statistics and visualizing information

Descriptive statistics
_ ways to summarize data with numbers and graphs
_ two most important function: communicate information, support reasoning about data

Pie chart, bar graph, and histograms
_ qualitative: pie chart and dot plot
_ quantitative: bar graph and histogram

The box plot (box-and-whisker plot) and scatter plot
_ box plot: min, 1st quartile, median, 3rd quartile, max (five-number summary)
— scatter plot: depicts data that come as pairs, visualizes the relationship between the two variables

Mean and median
_ mean and median are the same when the histogram is symmetric

Skewed to the right

Standard deviation
_ looks at the average of squared differences of each number from its average

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

_ Look at the interquartile range = 3rd quartile - 1st quartile (measures how spread out the data are) when concerned about skewness

## Statistical inference

Population
_ the entire group of subjects about which we want information (ex. all US voters)
Parameter
_ the quantity about the population we are interested in (ex. approval percentage among all US voters)
Sample
_ the part of the population from which we collect information (the 1,000 voters selected at random)
Statistic (estimate)
_ the quantity, we are interested in, as measures in the sample (approval percentage among the sampled voters)

Key point: even a relatively small sample (100 or 1,000) will produce an estimate that is close to the parameter of a very large population of 250 million subjects. This is the reason why statistics is so powerful.

Random Sampling
— Bias: this sampling will favor a certain outcome
    — selection bias: more likely to sample certain subjects than others
    — non-response bias: who choose to respond to the question my be different from the non-responders.
    — voluntary response bias: more likely to get responses from people who had very bad or very good experiences

— Simple Random Sample: selects subjects at “random” without replacement
— Stratified Random Sample: divides the population into groups of similar subjects called strata (e.g. urban, suburban)
    — Stratified random sampling can result in a more precise estimate than with simple random sampling
    — However, it is more complicated to execute

— Chance Error: Since the sample is drawn at random, the estimate will be different from the parameter due to unavoidable chance error (sampling error). Drawing another sample will result in a different chance error.

estimate = parameter + bias + chance error
Increasing sample size (n) results in lower chance error, but does not affect the magnitude of bias.

Observational Studies
— measures outcomes of interest and can be used to establish association (not causation, due to confounding factors)

Randomized Controlled Experiments
— To establish causation, an experiment is required:
    — A treatment is assigned to the treatment group, as opposed to the control group, at random. Therefore, influences other than the treatment operate equally on both groups, apart from differences due to chance.
    — The experiment must be double-blind: neither the subjects nor the evaluators know the assignments to treatment and control.

## The Interpretation of Probability

What is Probability?
— The Probability of an event = the proportion of times this event occurs in many repetitions.
    — It requires that it is possible to repeat this chance experiment many times.

Four Basic Rules:
— Complement rule: P(A does not occur) = 1 - P(A)
— Rule for equally likely outcomes: P(A) = number of outcomes in A / n
— Addition rule: If A and B are mutually exclusive, then P(A or B) = P(A) + P(B)
— Multiplication rule: If A and B are independent, then P(A and B) = P(A) P(B)

A and B are mutually exclusive if they cannot occur at the same time.
Two events are independent if knowing that one occurs does not change the probability that the other occurs.

Conditional Probability and Bayes' Rule

— The conditional probability of B given A = P(B|A) = P(A and B) / P(A)

-> General multiplication rule: P(A and B) = P(A) P(B|A)

-> In the special case where A and B are independent: P(A and B) = P(A) P(B)

## Conditional probability

Computing probabilities by total enumeration:
P(spam) = 20%. What is the probability that 'money' appears in an e-mail?

From data we know P(money | spam)= 8%, P(money | ham)= 1%.

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam or money appears and e-mail is ham

P(money appears)

=P(money and spam) + P(money and ham)

=P(money | spam) P(spam) + P(money | ham) P(ham)

$= 0.08 \times 0.2 + 0.01 \times 0.8$

$= 2.4\%$

## Bayes' rule

From data we know P(money appears in e-mail | e-mail is spam)= 8%, but what we need to build a spam filter is P(e-mail is spam | money appears in e-mail).

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(B \text{ and } A)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A)}$$

$$P(\text{spam} \mid \text{money}) = \frac{P(\text{money} \mid \text{spam})\,P(\text{spam})}{P(\text{money})} = \frac{0.08 \times 0.2}{0.024} = 67\%$$

Bayes' rule:

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

Bayes' rule:

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

$$= \frac{P(A|B)\,P(B)}{P(A|B)P(B) + P(A|\text{not } B)\,P(\text{not } B)}$$

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know P(D)= $1\%$ , P(+|D)= $95\%$ , P(+|no D)= $2\%$.

$$\text{Want } P(D|+) = \frac{P(+|D)\,P(D)}{P(+)}$$

$$= \frac{P(+|D)\,P(D)}{P(+|D)\,P(D) + P(+|\text{no D})\,P(\text{no D})}$$

$$= \frac{0.95\times 0.01}{0.95\times 0.01 + 0.02\times 0.99} = 32.4\%$$