Descriptive Statistics Overview

When given a set of data values, to summarize the data we will construct graphical and numerical summaries. In discussing the distribution of the values, comment on **shape**, **center**, **spread**, and any **outliers** or deviations from the general pattern.

- **Graphical Summaries for Quantitative data**
  - *Stem and leaf plots*
    - Shows actual data values
    - May need to split stems or truncate values
  - *Histograms*
    - If bins are equal width then height of bar equals number of observations
    - Include left hand endpoint
    - Frequency vs. relative frequency

- **Graphical Summaries for Categorical data**
  - *Bar graphs*
    - Usually separate the bars
    - Height is count or percentage in the category

- **Numerical Summaries for Quantitative data**
  - Center
    - *Mean* ($\bar{x}$, μ)= numerical average of all the values
      - "Balance point" of the distribution
      - Pulled towards the tail of a skewed distribution
    - *Median* ($\tilde{x}$, $\tilde{\mu}$) = middle value
      - "Typical" value
      - Resistant to outliers
    - *Trimmed mean*, e.g. 10% trimmed mean is the average of middle 80%
  - Spread
    - *Range* = max – min
    - *Standard deviation* (s, σ)
      - Measures "average deviation" from the mean (how far do the values tend to fall from the mean)
    - *Fourth spread* (f_s)
      - Divide the data set into quarters
      - Resistant to outliers
  - *Five number summary* = min, lower fourth, median, upper fourth, max
    - *Boxplots* (most useful for comparing data sets)
  - Checking for outliers
    - Any values below Q_1-1.5(Q_3-Q_1) or above Q_3+1.5(Q_3-Q_1)

- **Numerical Summaries for Categorical data**
  - x/n = count/number of observations = $\hat{p}$ (p for population)
  - Proportion is usually more informative than count

Brief Review of Set Operations and Properties

A <u>set</u> is a collection of elements. For our purposes, these elements will be outcomes of a (random) experiment, i.e. an event. Sets are typically denoted with capital letters.

OPERATION	NOTATION	MEANING
Union of two events	$A \cup B$	A or B
Intersection of two events	$A \cap B$ [or AB]	A and B
Complement of an event	A' [or A^c or $\overline{A}$]	not A
Finite union of events	$A_1 \cup A_2 \cup \ldots \cup A_n$	at least one of the A_i's
Finite intersection of events	$A_1 \cap A_2 \cap \ldots \cap A_n$	all of the A_i's

DEFINITION	NOTATION	MEANING
A is a subset of B	$A \subset B$	A is contained in B
A, B mutually exclusive	$A \cap B = \phi$	no shared outcomes

PROPERTY	STATEMENT
Commutative	$A \cup B = B \cup A$
	$A \cap B = B \cap A$
Associative	$(A \cup B) \cup C = A \cup (B \cup C)$
	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributive	$A \cup (B \cap C) = (A \cup B) \cap (B \cup C)$
	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
DeMorgan's Laws	$(A \cup B)' = A' \cap B'$
	$(A \cap B)' = A' \cup B'$

An event is a set, while a probability is a number.
One calculates probabilities of events (and therefore of sets), but probabilities are numbers. The following <u>meaningless</u> statements are examples of nonsensical confusions of sets and numbers:

$P(A) \cap P(B)$ $(P(A))'$ $P(1\text{-}A)$ $P(A+B)$

Examples of <u>meaningful</u> statements about events and probabilities include:

$P(A \cap B)$ $P(A')$ $1\text{-} P(A)$ $P(A)+P(B)$

It's very useful for later understanding if you force yourself to explain each step in detail and always use the correct notation.

Some other handy translations:

$A = (A \cap B) \cup (A \cap B')$, which says that A is composed of its part that intersects together with its part that does not intersect B

$A \cup B = (A \cap B) \cup (A \cap B') \cup (A' \cap B)$, which says that A union B is composed of three mutually exclusive pieces

"exactly one of the two events A and B" = $(A \cap B') \cup (A' \cap B)$