Lecture 2: Descriptive Statistics

Supplementary Reading: Pagano/Gauvreau; Chapters 2-3

Descriptive statistics

- Method for organizing and summarizing data
- Method for detecting important features/patterns of a dataset in order to extract useful information
- Characterize "regularities" of measurements that are naturally "variable"
- Important tool in communicating final results of a study
- Basic elements: tables, graphs, numerical summary measures

Basic concepts

- **Data**: "numbers" resulting from measuring "subjects"
- **Data sources**: routine medical records, surveys, experiments, cell phones, etc.
- **Variable**: characteristic that takes on different values for different subjects
- **Random variable**: values of a variable that can not be determined exactly in advance
- **Population**: largest collection of subjects of interest
- **Sample**: subset of a population (usually much smaller)
- **Descriptive statistics**: means of summarizing and organizing data
- **Inferential statistics**: methods to determine if the differences in data are real or due to chance

Types of data

- **Qualitative**: *does not* take on numerical values; e.g., marital status, diagnosis of a patient
  - **Nominal**: no order; magnitude not important; categories (e.g., 1=female 0=male; 0=no disease 1=disease; etc.)
- **Quantitative**: *does* take on numerical values; e.g., body weight, number of tumors, blood pressure

  - **Ordinal**: order matters; magnitude not important (e.g., 1=fatal 2=severe 3=moderate 4=minor 5=no injury)

  - **Discrete**: order and magnitude important; integer valued (e.g., Number of people hospitalized on 4 days: Day 1 = 10, Day 2 = 16, Day 3 = 8, Day 4 = 13)

  - **Continuous**: real valued; *any* conceiveable value – in theory (e.g., height, weight, blood pressure, etc.)

Displaying and visualizing data

- **Tables**
  - Frequency
  - Relative frequency
- **Graphs**
  - Bar charts
  - Histograms
  - Scatterplots
  - Boxplots
  - Stem and leaf plots
  - Maps

Example: Lead exposed children

Data from a study investigating the psychological and neurological effects on children exposed to lead.

Of 124 children who lived near a lead smelter in El Paso, TX, 46 had blood lead levels $\geqslant 40$ micrograms per ml (high blood lead levels)

For those 46 children, information was recorded for their gender (dichotomous) and IQ (discrete or continuous?)

Gender (1 = male and 2 = female)

1 2 1 1 1 2 1 2 2 1 1 1 1 2 1 1 1 1 2 1 2 1 2
1 1 1 1 2 1 2 1 2 1 2 1 1 1 1 2 1 2 2 1 2 1 1

For nominal and ordinal data, a frequency distribution is a nice way to summarize data. For example, in the lead data:

Gender	Freq	Rel. Freq
Male	30	65.2%
Female	16	34.8%
Total	46	100%

Relative frequency is the percentage of times each value occurs.

For continuous data, values are often grouped into non-overlapping intervals, usually of equal width.

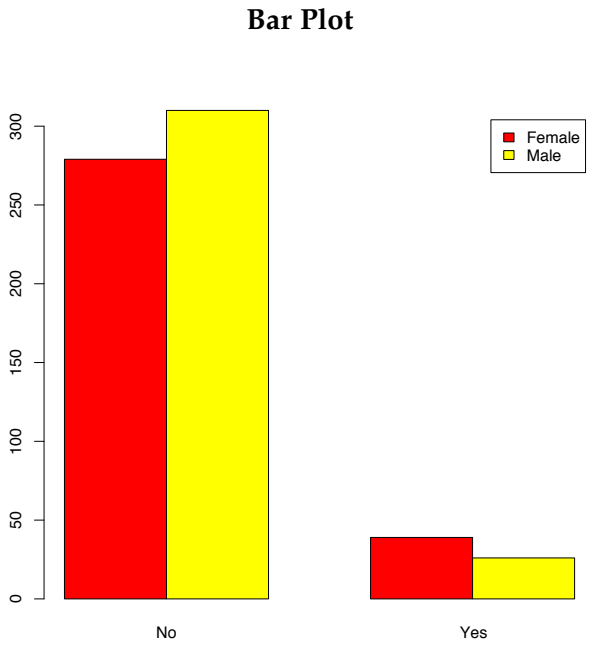
IQ	Freq	Rel. Freq (%)	Cum Rel Freq (%)
40-49	1	2.2	2.2
50-59	0	0	2.2
60-69	0	0	2.2
70-79	9	19.6	21.7
80-89	15	32.6	54.3
90-99	13	28.3	82.6
100-109	5	10.9	93.5
110-119	3	6.5	100.0

Cumulative relative frequency for an interval is the percentage of the total number of observations that have a value in or below that interval.

Example: Smoking status by gender

Sex by smoking status in a data set consisting of 654 subjects upon whom forced expiratory volume (FEV) was measured.

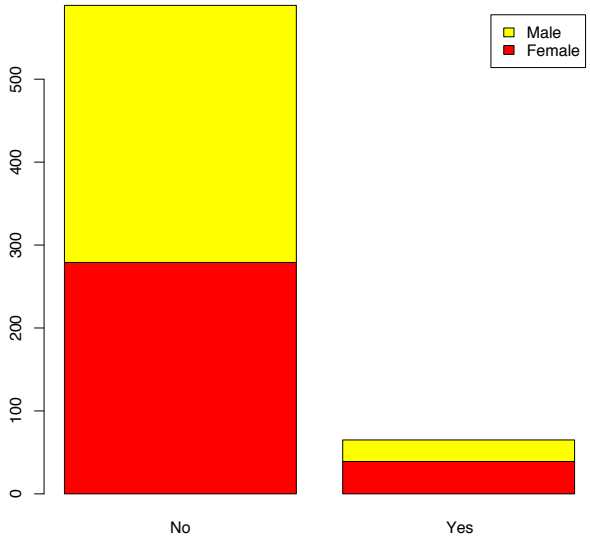
	Smoking status		
Sex	No	Yes	Total
Female	279	39	318
Male	310	26	336
Total	589	65	654



R command:

```
barplot(smoke_stat, main="Bar Plot", col=c("red","yellow"),
        legend = rownames(smoke_stat))
```

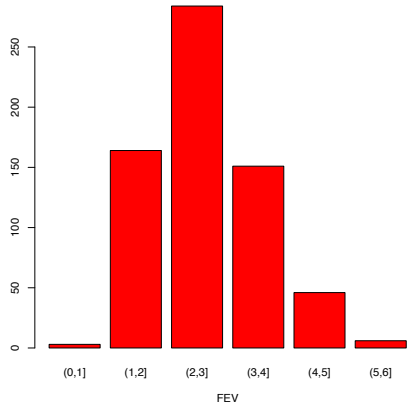
Bar Plot (Stacked)



R command:

```
barplot(smoke_stat, main="Bar Plot (Stacked)", col=c("red","yellow"),
        legend = rownames(smoke_stat), beside=TRUE)
```

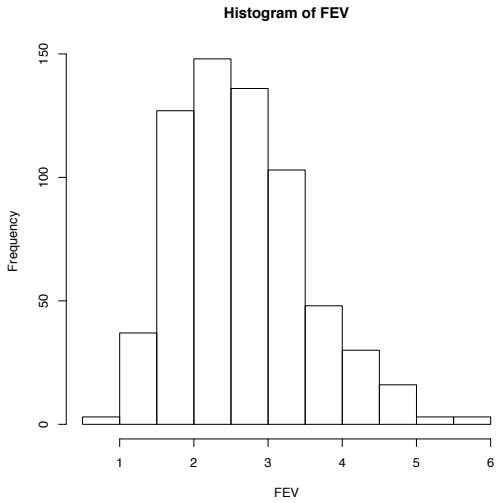
Bar Plot: FEV (l)



R command:

```
barplot(fev, xlab="FEV")
```

Histogram: FEV (l)



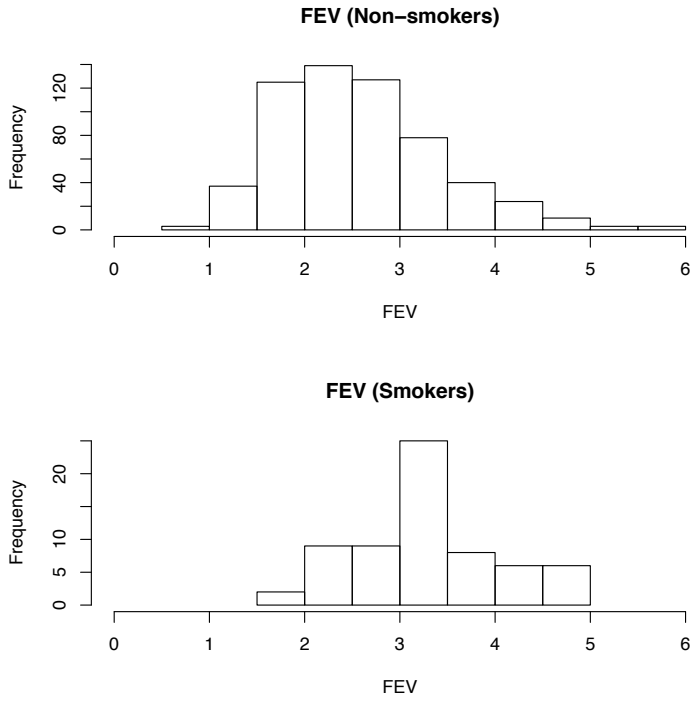
R command:

```
hist(fev, main="Histogram of FEV")
```

Histogram

- Graphical display of tabulated frequencies
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height (a crucial distinction when the categories are not of uniform width)

Histogram: FEV for Nonsmokers and Smokers

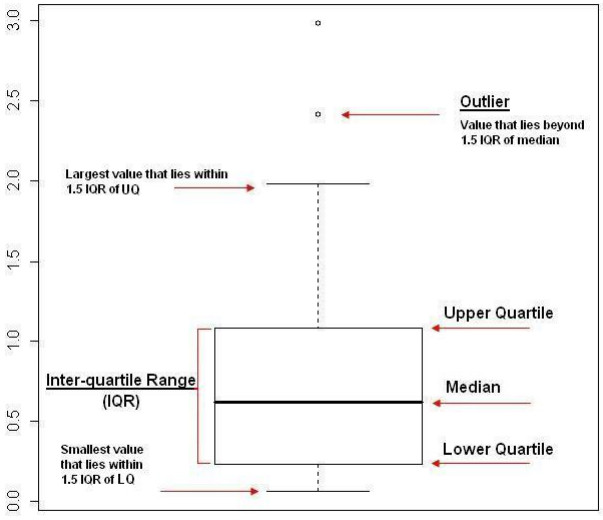


R command:

```
par(mfrow=c(2,1))
hist(nosmoke, main="FEV (Non-smokers)")
hist(smoke, main="FEV (Smokers)")
```

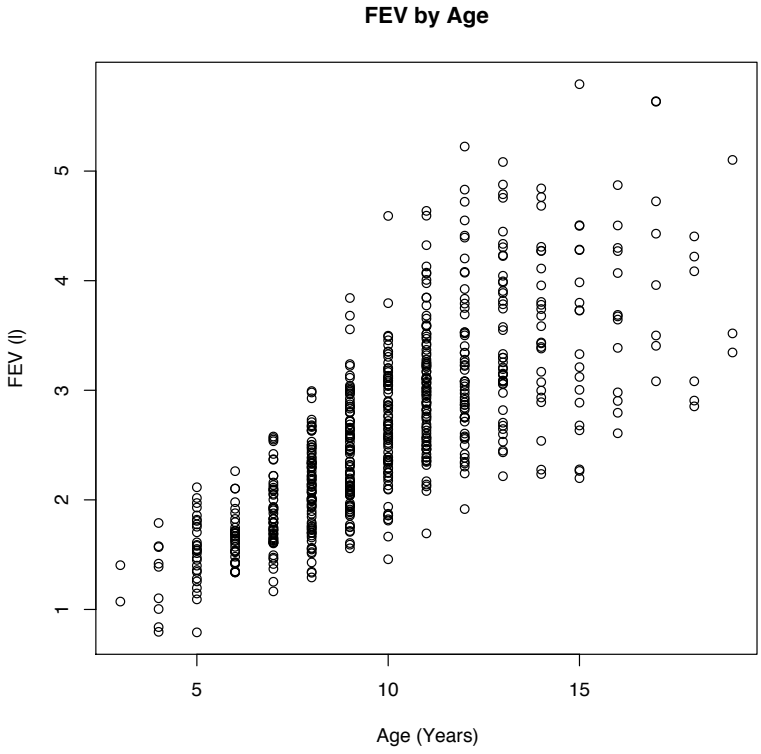
Box Plot (Box-and-Whisker Diagram)

- Graphical display of *five-number summary*
  - *Minimum*: smallest observation
  - *Maximum*: largest observation
  - *Median*: value above which lie half of the observations and below which lie half of the observations
  - *Upper quartile/75th Percentile*: value above which lie 25% of the observations and below which lie 75% of the observations
  - *Lower quartile/25th Percentile*: value above which lie 75% of the observations and below which lie 25% of the observations
  - *Inter-quartile Range (IQR)*: the difference between the upper (75%) and lower (25%) quartiles
  - *Outlier*: values which fall more than 1.5 times the interquartile range above the third quartile or below the first quartile



Scatter Plot

- Used to depict the relationship between two different *continuous* measurements
- Each point on the graph represents a pair of values



R command:

```
plot(age, fev1, main="FEV by Age", xlab="Age (Years)", ylab="FEV(I)")
```

Numerical summaries

Numerically quantify characteristics of a dataset. Suppose our dataset consists of the numbers $x_1, x_2, \ldots, x_n$.

- Central tendency - quantify the "middle" of the data
  - **Mean**: average (arithmetic) value; $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$
  - **Median**: middle value (50th percentile); value m such that half the measurements lie above and half the measurements lie below m
  - When $mean < median$ then data is *left* or *negatively* skewed
  - When $mean > median$ then data is *right* or *positively* skewed
  - When $mean = median$ then data is *symmetric*
  - **Mode**: most frequent value

- Variability/Spread
  - **Range**: Maximum - Minimum
  - **IQR**: Interquartile Range; 75th %ile - 25th %ile; encompasses inner 50% of the observations; the pth percentile is the value that is greater than or equal to p% of the observed values
  - **Variance**: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
  - **Standard Deviation**: the square root of the variance

Most commonly used is the **standard deviation**, which is a measure of dispersion or spread of the data, and is denoted as s or σ (expressed in same units as the mean)

Variance is the "average" squared deviation of each observation from the mean

Using s, we try to "unsquare" the distances (expressed in same units as the mean)

Measuring "center"

Measures of "central tendency" describe the mid/balance point of a set of observations.
Set of data: $x_1, x_2, \ldots, x_n$

Example: Hospital data

Let x represent hospital admission percents for female children with cystic fibrosis.

$$\begin{aligned} x_1 &= 3.10 \\ x_2 &= 3.45 \\ x_3 &= 6.25 \\ x_4 &= 1.20 \\ x_5 &= 6.19 \\ x_6 &= 5.00 \\ x_7 &= 8.75 \\ x_8 &= 6.22 \\ x_9 &= 5.00 \\ x_{10} &= 2.50 \\ x_{11} &= 25.00 \end{aligned}$$

Mean :

$$\begin{aligned} \overline{x} &= n^{-1} \sum_{i=1}^{n} x_i \\ \overline{x} &= (3.10 + 3.45 + 6.25 + 1.20 + 6.19 + 5.00 + 8.75 + 6.22 + 5.00 + 2.50 + 25.00)/11 = 6.61 \end{aligned}$$

Compute the mean without the observation 25: $\overline{x} = 4.77$

R command:

```
mean(cfp)
```

Mean is sensitive to unusually small or large values, it is not "robust".

Median :

- More robust, not as "sensitive" to high and low values
- 50*th* percentile
- Middle value if odd number of observations
- Average of two middle values if even number of observations

With the hospital rates: 1.20, 2.50, 3.10, 3.45, 5.00, 5.00, 6.19, 6.22, 6.25, 8.75, 25
Median = 5.00

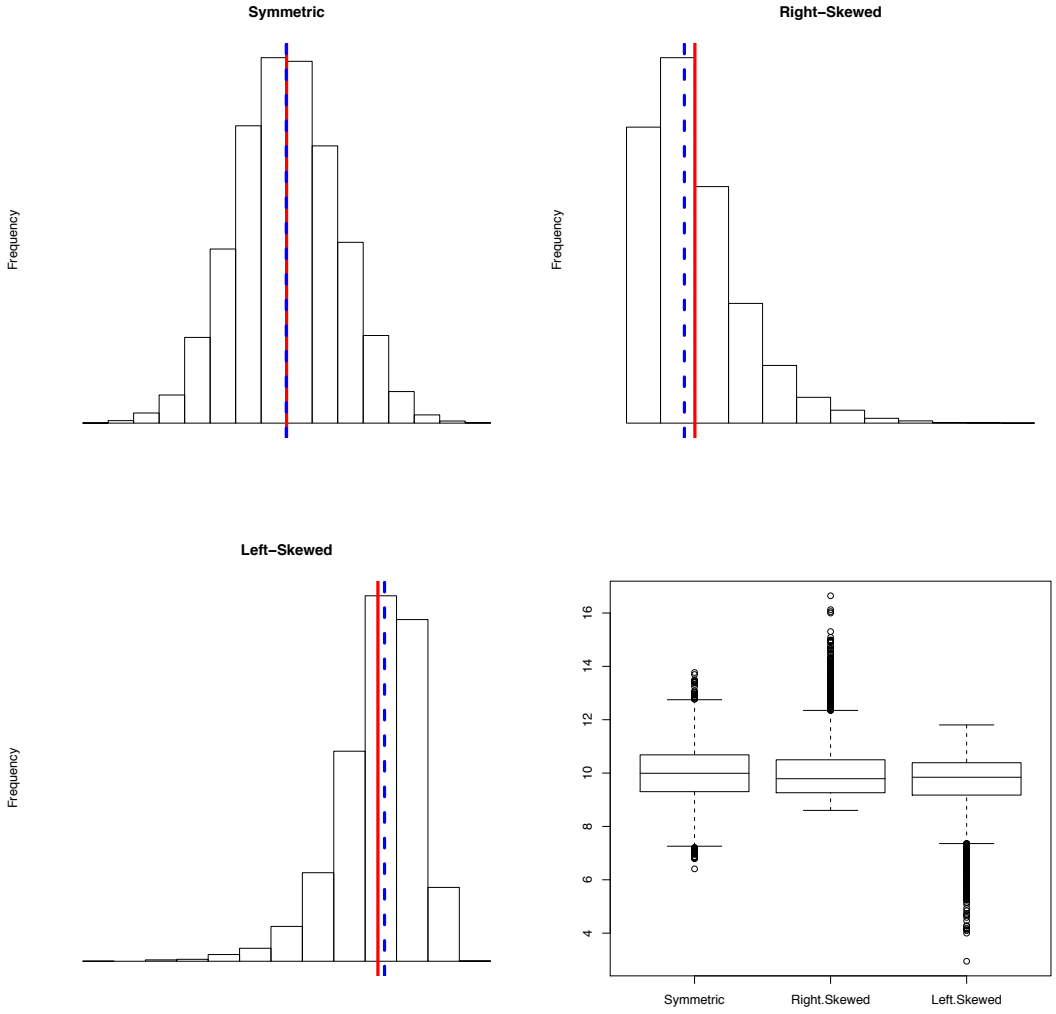
R command:

```
median(cfp)
```

If 25 is replaced with a larger number, say 100, the median is still 5.00.

Which to use: use both! The mean performs best with a symmetric distribution, but if the distribution is skewed to the right or left, the median is better. However, it is usually best to report both.

Symmetric, Right-Skewed, and Left-Skewed Distributions



Example: Asthma

Study examining patients with severe asthma.

Data collected for 10 subjects who arrived at the hospital in a state of respiratory arrest: their breathing had stopped and individuals were unconscious upon arrival.

Heart rates (bpm): 167, 150, 125, 120, 150, 145, 40, 136, 120, 150

What is a 'typical' heart rate for these patients?

First, formally represent measurements by $x_1, ..., x_{10}$ for each of the 10 subjects.

Mean:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{10}(x_1 + x_2 + ... + x_{10}) \\ &= 130.3 \text{ beats per minute} \end{aligned}$$

Deleting the 40 bpm measurement, $\bar{x} = 140.3$ bpm. A 10 bpm increase!

Median: 50th percentile

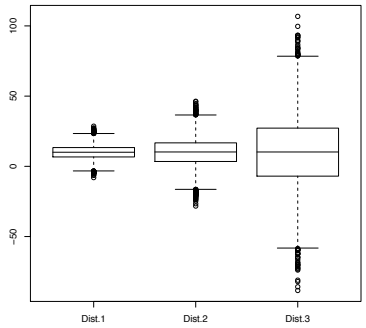
Order observations from smallest to largest: 40, 120, 120, 125, 136, 145, 150, 150, 150, 167

Median is $[(n+1)/2]$th largest value. If n is odd, then the median is the average of the $[n/2]$th and $[n/2+1]$th.

For the heart rate example, $n = 10$ is even and the median is the average of the 5th and 6th largest measurements: $(136 + 145)/2 = 140.5$ bpm

Again, removing unusual 40 bpm: median = 145 (5th largest observation)

When looking at a dataset, the measure of center alone can be misleading. For example, the following 3 distributions have the same mean, median and mode.



We also need a measure of variability or spread.

Variability in hospital data (children with cystic fibrosis)

Rate	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$
1.20	-5.41	29.22
2.50	-4.11	16.85
3.10	-3.51	12.29
3.45	-3.16	9.96
5.00	-1.61	2.58
5.00	-1.61	2.58
6.19	-0.42	0.17
6.22	-0.39	0.15
6.25	-0.36	0.13
8.75	2.14	4.60
25.00	18.39	338.36
Total	0.00	416.88

Standard Deviation:

$$s = \sqrt{Variance} = \sqrt{416.88/10} = 6.46$$

R command:

```
sd(cfp)
```

Heart rate data

Range: $167 - 40 = 127$ bpm. Sensitive to "extreme" values (function of the 2 most extreme values!)

Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1,239.3$ bpm^2

Standard deviation: $s = \sqrt{s^2} = 35.2$ bpm

Typical presentation patterns:

Range appears with median

IQR sometimes appears with the median as well

SD appears with mean

For nominal and ordinal data, a table is often more effective than numerical summary measures

*Note: the R command for a summary of your data that includes the mean, median, 75th and 25th percentiles, min, and max is:

```
summary(variable)
```