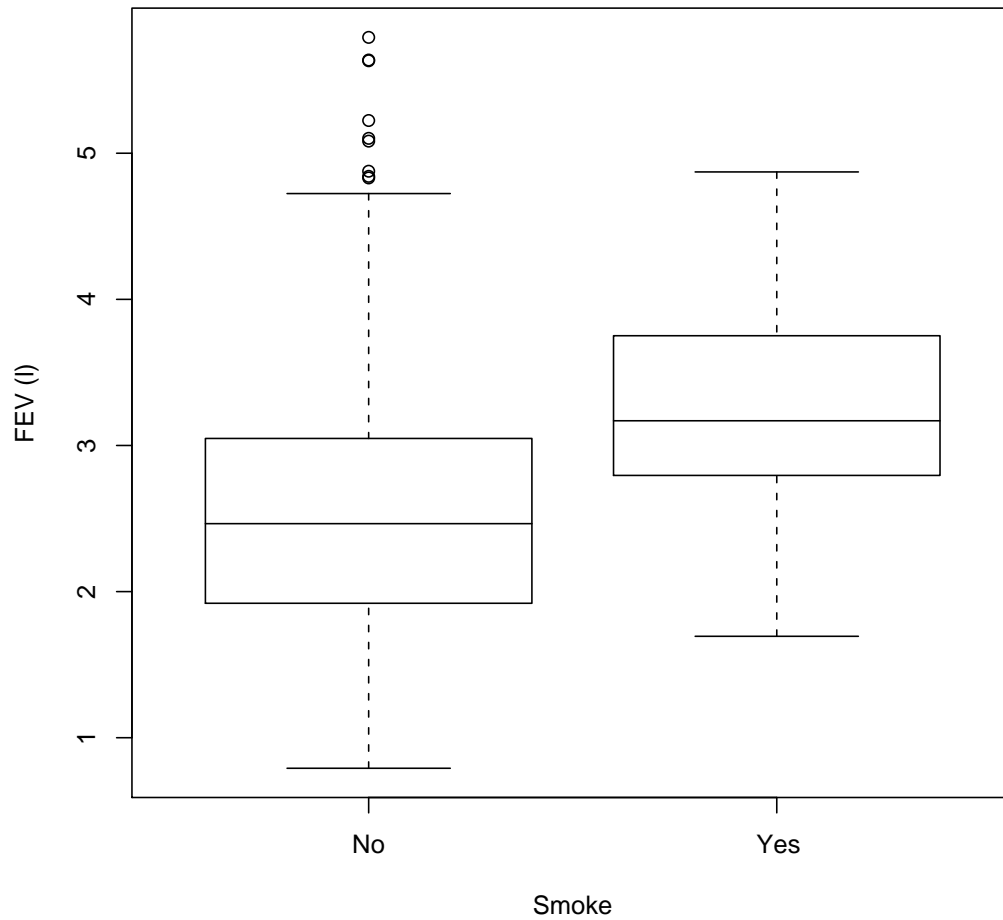


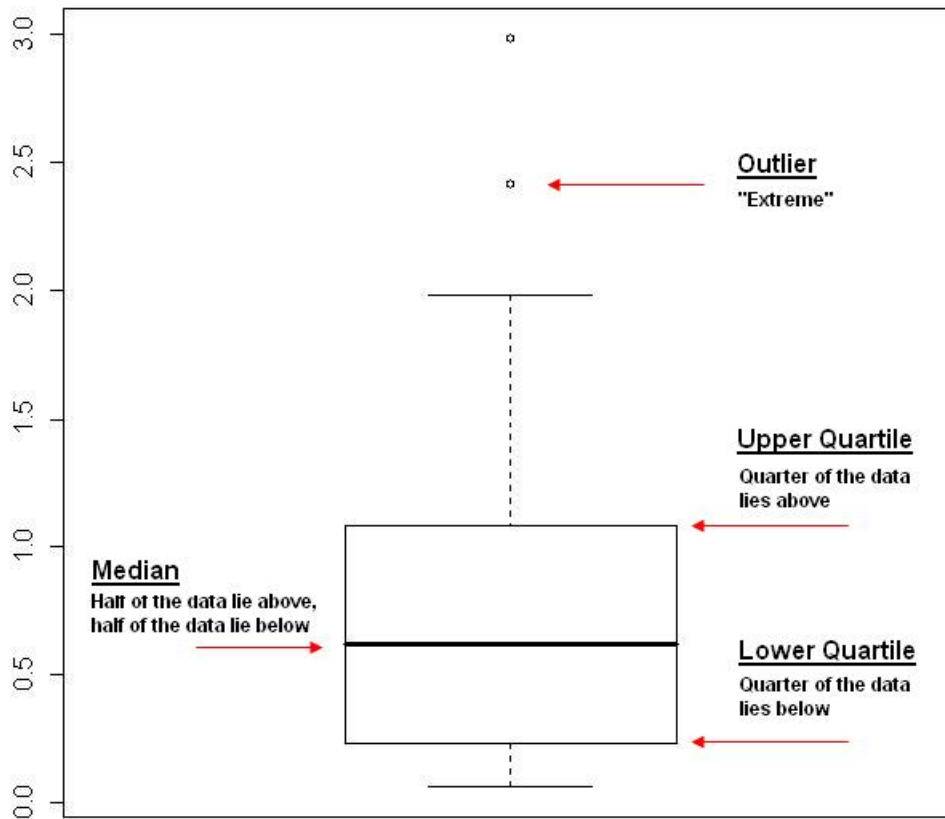
Lecture 1: Intro to Biostatistics

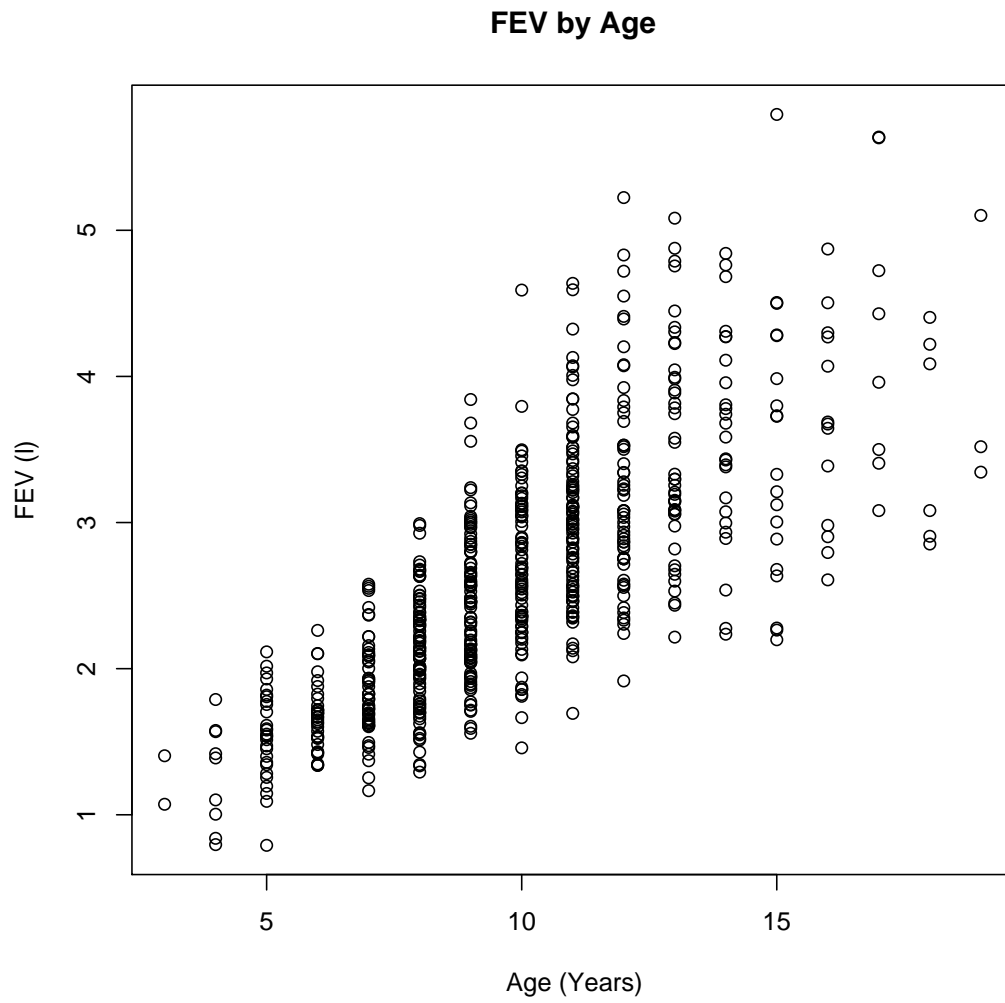
Smoking: hazardous?



Box Plot

a.k.a “box-and-whisker diagram” or “candlestick chart”





How do numbers tell us what we want to know?

Overview

What is *Biostatistics*?

Sub-discipline of statistics focused on analysis of biological, medical, and public health science data

Methods in common with statistics:

- *Descriptive statistics*: methods for describing, displaying, organizing, and summarizing data
- *Inferential statistics*: methods and procedures for inferring “truth” from data. Are apparent features real or are they simply due to chance?
- *Experimental design*: methods for gathering data; methods for creating “fair” experiments that will provide sufficient data to make inferences (at minimal cost)
- *Prediction*: methods for predicting future values from existing data

Issues unique to biostatistics:

- Extremely noisy data and high potential for bias
- Length of time required to conduct experiments
- Time can play a fundamental role in analysis
- Integration with bioinformatic databases

Primary tool for inference, design, and prediction: probability theory (laws of chance)

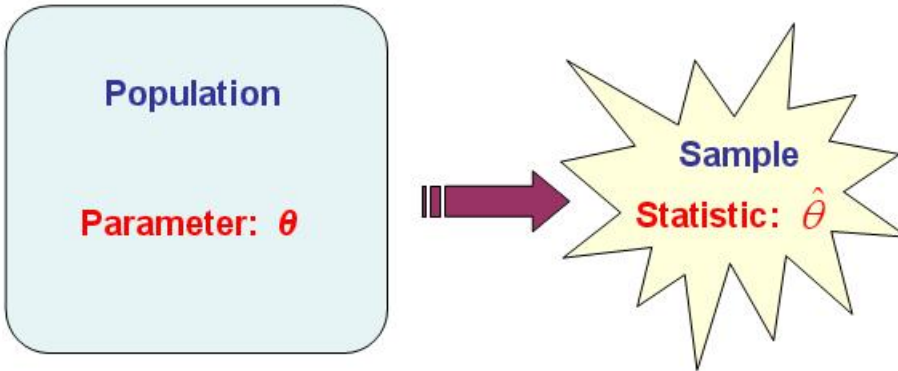
The game:

- We want to know something about a target population (e.g. test a scientific hypothesis)...
- We collect data from a “representative” sample of the population,
- then organize and summarize the data,
- then use the laws of probability to make inference and subsequent predictions.

Examples of hypotheses:

- Two subpopulations (e.g. treatment groups) are different (e.g. in drug response)
- Multiple subpopulations are different
- Two attributes are associated (e.g. behavior and presence of disease)
- A trend exists between two variables (e.g. chemical exposure and probability of disease)

General Statistical Inference



What is the distribution of $\hat{\theta}$?

→ Confidence Intervals, Hypothesis Tests

Applications

- Clinical trials
 - Evaluating toxicity/safe dosage: *Phase I*
 - Preliminary evaluation of efficacy: *Phase II*
 - Formal comparison against an established standard treatment: *Phase III*
- Epidemiology
 - Determining distribution of and risk factors for disease
- Public health
 - Screening for chronic disease
 - Public health policy
 - Environmental risk assessment
- Laboratory research
 - Experimental design
 - Drug discovery
 - Toxicology

Where is biostatistics used?

- Industry (e.g. pharmaceutical companies)
- Academia (e.g. medical research, public health research)
- Government (e.g. regulatory agencies)

Deciding on proper analysis

Depends on **outcome** (and covariates)

- Continuous & interested in comparing mean response between 2 groups \Rightarrow *t-test*
- Continuous & interested in the effect of covariates on mean response \Rightarrow *regression*
- Continuous & interested in comparing mean response between 2 or more groups adjusting for other covariates \Rightarrow *regression*
- Dichotomous & interested in the effect of covariates on mean response \Rightarrow *logistic regression*
- Counts & interested in the effect of covariates on mean response \Rightarrow *Poisson regression*
- Survival times & interested in the effect of covariates on “hazard” (instantaneous risk of dying) \Rightarrow *Cox regression* (“*survival analysis*”)
- etc...

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, 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

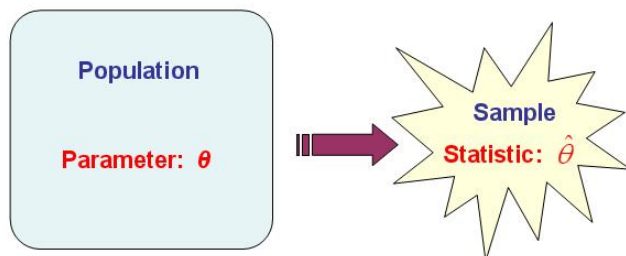
Population – largest collection of subjects of interest

Sample – subset of 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

General Statistical Inference



What is the distribution of $\hat{\theta}$?

→ Confidence Intervals, Hypothesis Tests

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* conceivable value – in theory (e.g., height, weight, blood pressure, etc.)

Table 1: Count data summarized in a table: Number of males and females that smoke in the FEV data

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

Sex by smoking status in a data set consisting of 654 subject upon whom forced expiratory volume (FEV) was measured. See Rosner (2000), page 440, for details.

In general,

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

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

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

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%

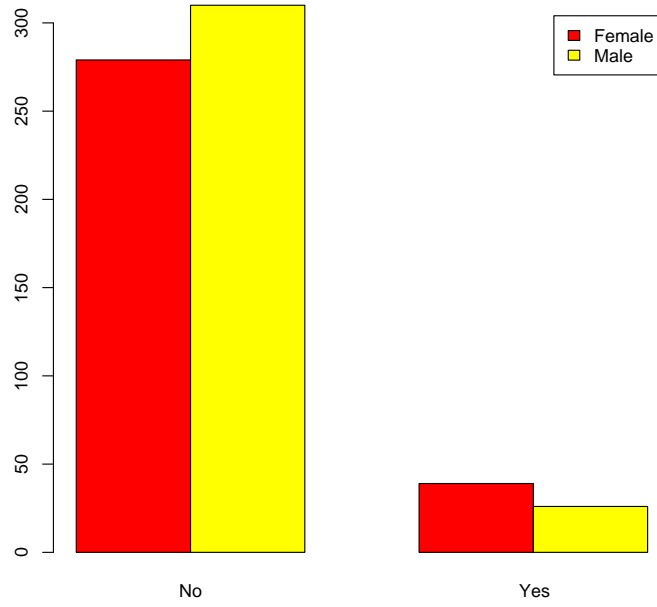
Relative frequency is the percentage of times each value occurs

For discrete and continuous data, values are often grouped on non-overlapping intervals, usually of equal width

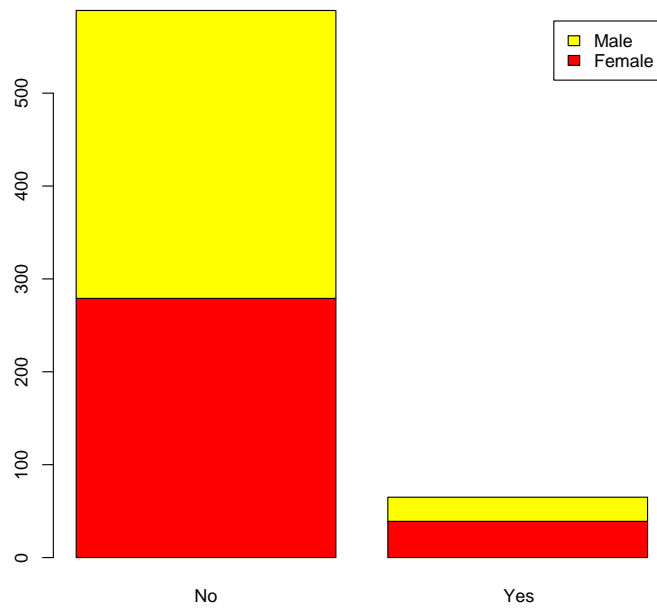
IQ	Freq	Rel. Freq (%)	Cum Rel Freq (%)
40-49	1	2.2	2.2
50-59	0	0	2.2
60-69	0	0	2.2
70-79	9	19.6	21.7
80-89	15	32.6	54.3
90-99	13	28.3	82.6
100-109	5	10.9	93.5
110-119	3	6.5	100.0

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

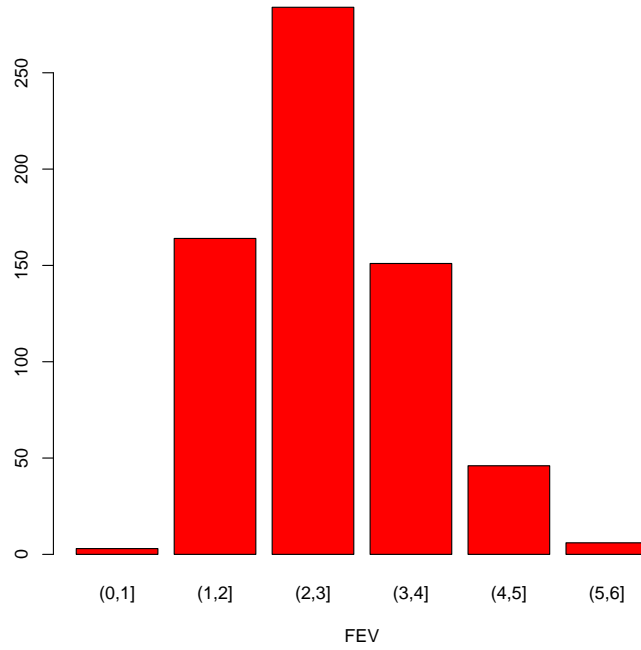
Bar Plot



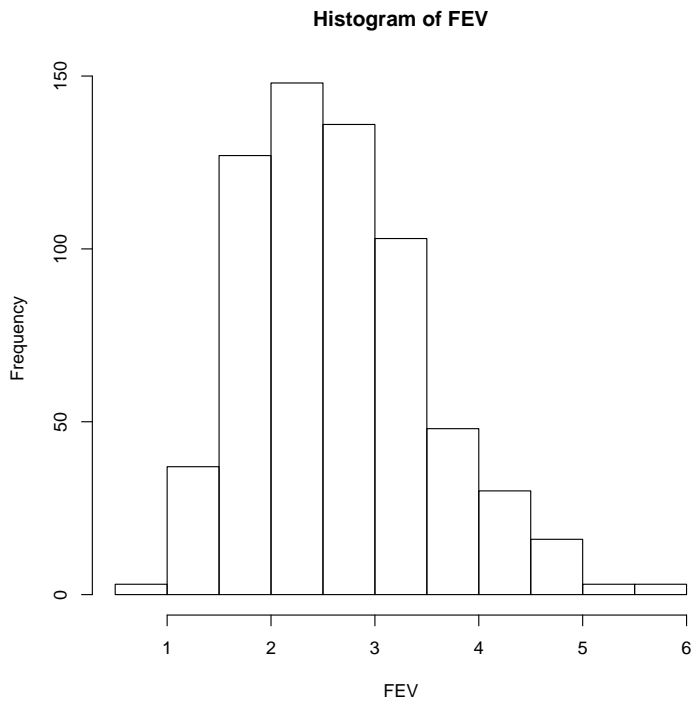
Bar Plot (Stacked)



Bar Plot: FEV (1)



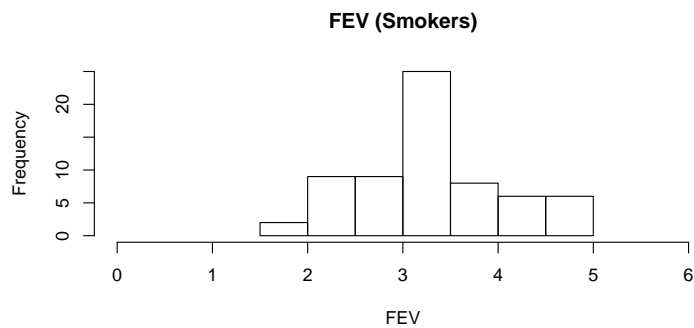
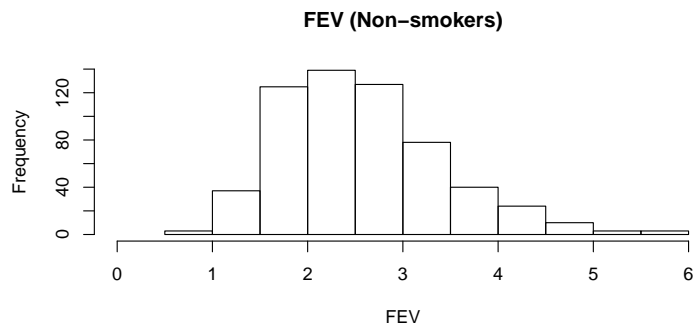
Histogram: FEV (1)



Histogram

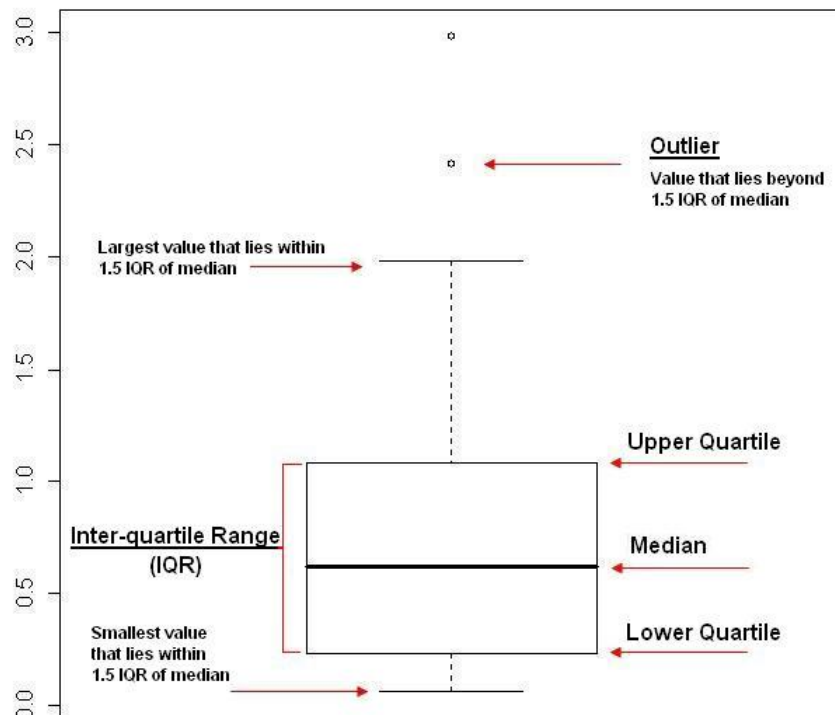
- Graphical display of tabulated frequencies
- differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height (a crucial distinction when the categories are not of uniform width)

Histogram: FEV for Nonsmokers and Smokers



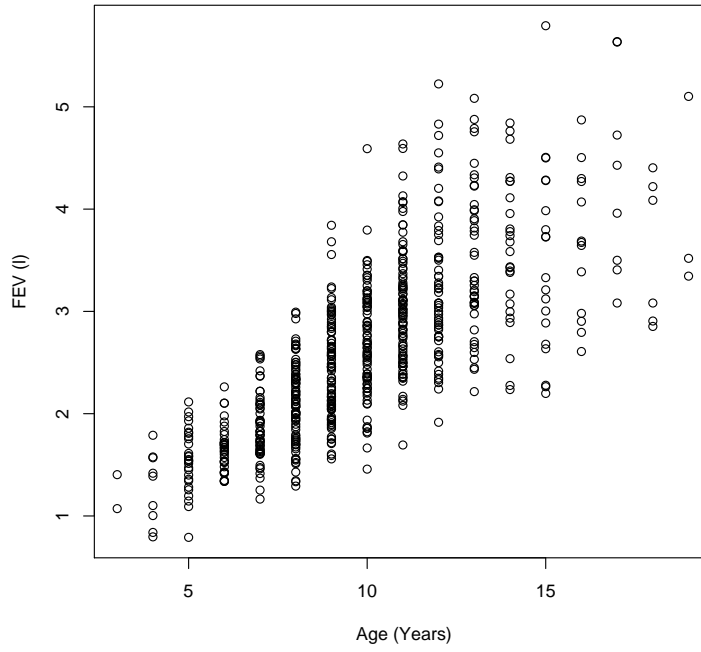
Box Plot (Box-and-Whisker Diagram)

- 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 25% of the observations and below which lie 75% of the observations



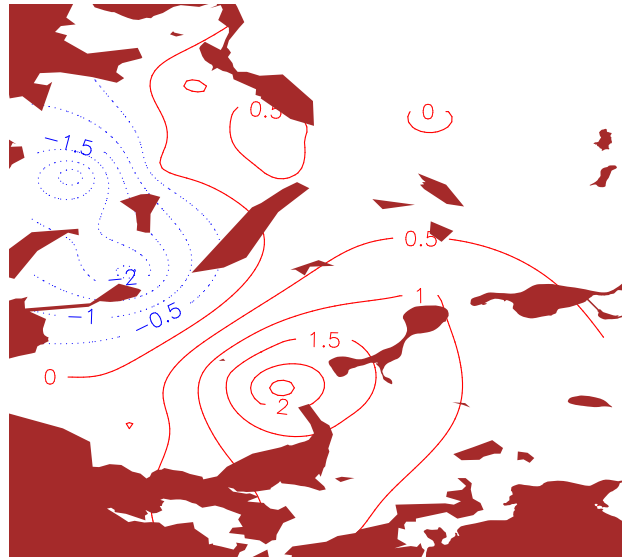
Scatter Plot

FEV by Age

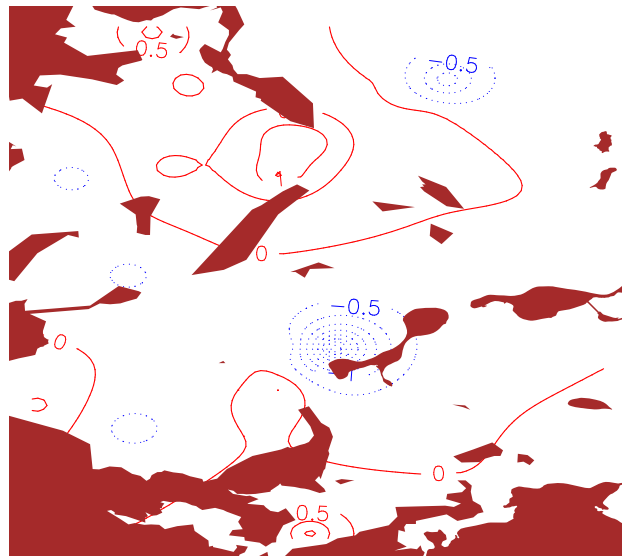


Other Graphical Summaries: Maps

Spatial Heterogeneity with NITP Active



Spatial Heterogeneity with NITP Inactive



Numerical summaries

Numerically quantify characteristics of a dataset

- 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; value m such that half the measurements lie above and half the measurements lie below m
 - When $mean < median$ then data *negatively* skewed
 - When $mean > median$ then data *positively* skewed
 - When $mean = median$ then data *symmetric*
 - **Mode**; most frequent value
- Variability/Spread
 - **Range**; Max - Min = Range
 - **IQR**; 75th %ile - 25th %ile; encompasses inner 50% of the observations; the p th percentile is the value that is greater than or equal to $p\%$ of the observed values
 - **Variance**; $s^2 = (n - 1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Most common is the **standard deviation** which is a measure of dispersion or spread of the data, denoted s (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)

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

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

Measuring “center”

Measures of “central tendency” describe the mid/balance point of set of observations

Set of data : x_1, x_2, \dots, x_n

Example: Hospital data

Mean

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

Compute the mean without 25 :

$$\bar{x} = 4.77$$

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

Median

- more robust, but often not as “sensitive”
- 50th percentile
- middle value if odd number of observations
- average of two middle values if even number of observations

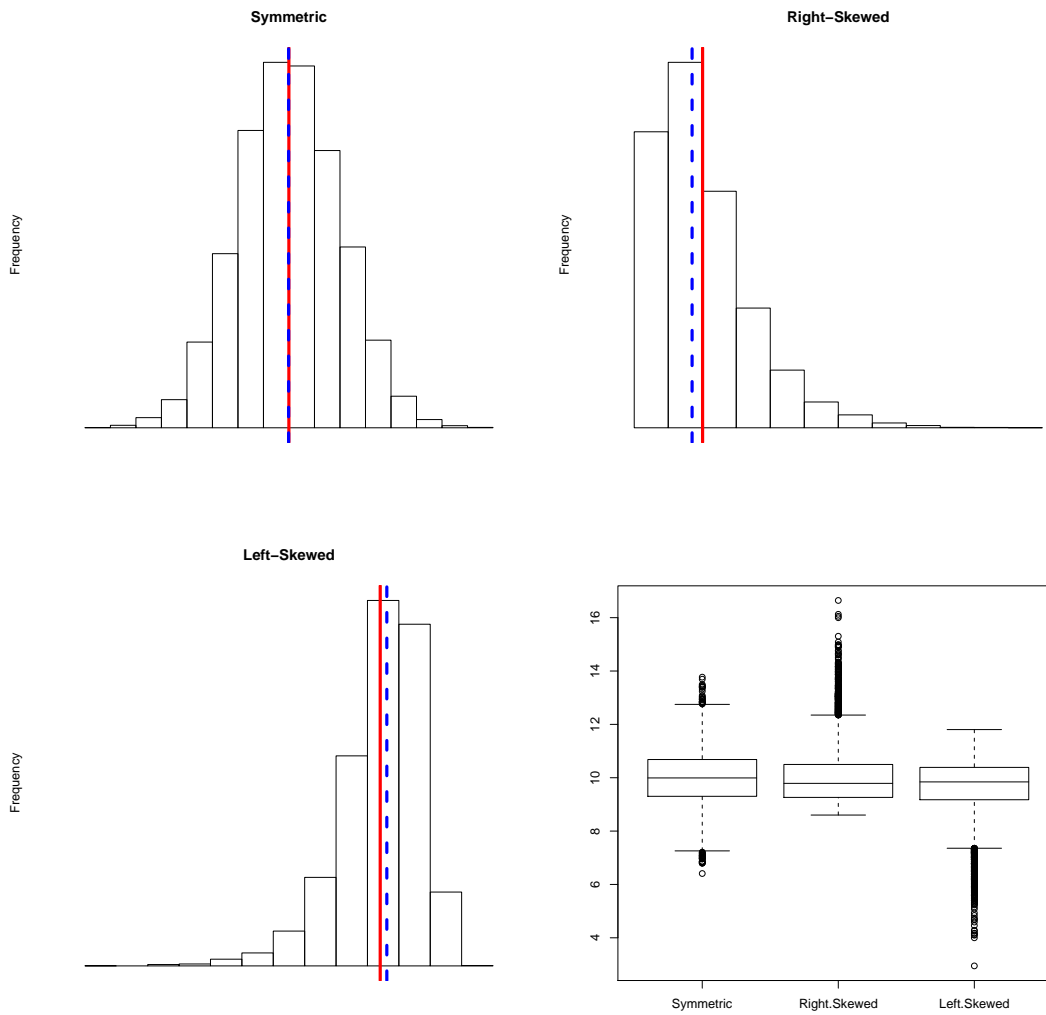
With the hospital rates: 1.20, 2.50, 3.10, 3.45, 5.00, 5.00, 6.19, 6.22, 6.25, 8.75, 25

Median = 5.00

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

Which to use: use both; mean performs best with a symmetric distribution. If the distribution is skewed to the right or left use the median.

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 – 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, \dots, x_{10} ; for each of the 10 subjects.

Mean:

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

Deleting the 40 bpm measurement $\bar{x} = 140.3$ bpm; 10 bpm reduction!

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)

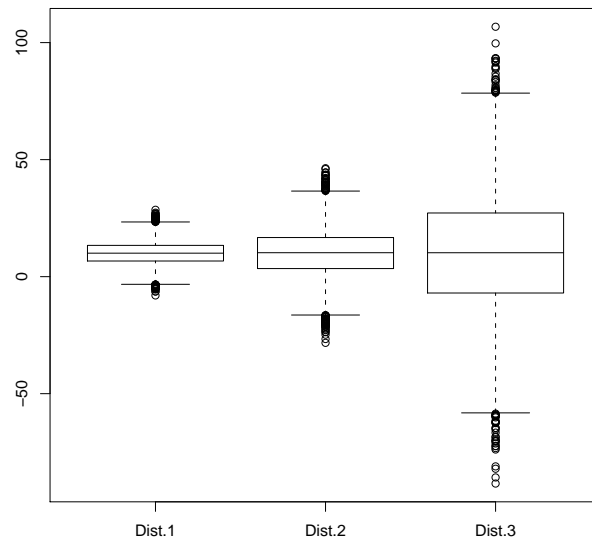
Which measure of central tendency is best for a given dataset??

Depends on type of data and the way values are distributed

Mean: continuous or discrete data; useful for “regular” data (no extreme values)

Median: more often used for continuous data; useful for data with outliers and extreme values.

When looking at dataset, measure of center alone can be misleading. For example the following 2 distributions have the same mean, median and mode.



Also need measure of variability or spread.

Variability in hospital data

Rate	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
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:

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

Heart rate data

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

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

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

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

Properties of the mean

Example: Interested in a certain desert region where temperatures recorded in °C for 30 days in June

The mean temperature was 42.8°C

What would be the mean temperature for the region in °F?

Logic: Given a set of n observations $(x_1 \dots x_n)$, a “translated” sample has the form, $y_1 = x_1 + c_1, y_2 = x_2 + c_1, \dots, y_n = x_n + c_1$ where c_1 is a constant. In general $y_i = x_i + c_1$ Mean of “translated” sample:

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{n} \sum_{i=1}^n (x_i + c_1) \\ &= \frac{1}{n} \left[\left(\sum_{i=1}^n x_i \right) + nc_1 \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^n x_i \right] + \left(\frac{1}{n} \right) (nc_1) \\ &= \bar{x} + c_1\end{aligned}$$

Consider “rescaling” sample: $y_i = c_2x_i$

Mean of rescaled sample is

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{n} \sum_{i=1}^n c_2x_i \\ &= \bar{x}c_2\end{aligned}$$

In general, for $y_i = c_2x_i + c_1$, then

Mean of new sample is $\bar{y} = c_2\bar{x} + c_1$. So returning to the temperature example, converting from °C to °F has the following form:

$$y_i = \frac{9}{5}x_i + 32$$

so,

$$\begin{aligned}\bar{y} &= \frac{1}{30} \sum_{i=1}^{30} y_i \\ &= \frac{1}{30} \sum_{i=1}^{30} \left(\frac{9}{5} x_i + 32 \right) \\ &= \frac{9}{5} \bar{x} + 32 \\ &= \frac{9}{5} 42.8 + 32 \\ &= 109.0\end{aligned}$$

Properties of the variance

Consider the translated sample, $y_i = x_i + c_1$, then $s_y^2 = s_x^2$

Rescaled sample: $y_i = x_i c_2$, then $s_y^2 = c_2^2 s_x^2$

In general, same rescaling and translation properties do not hold for the variance.