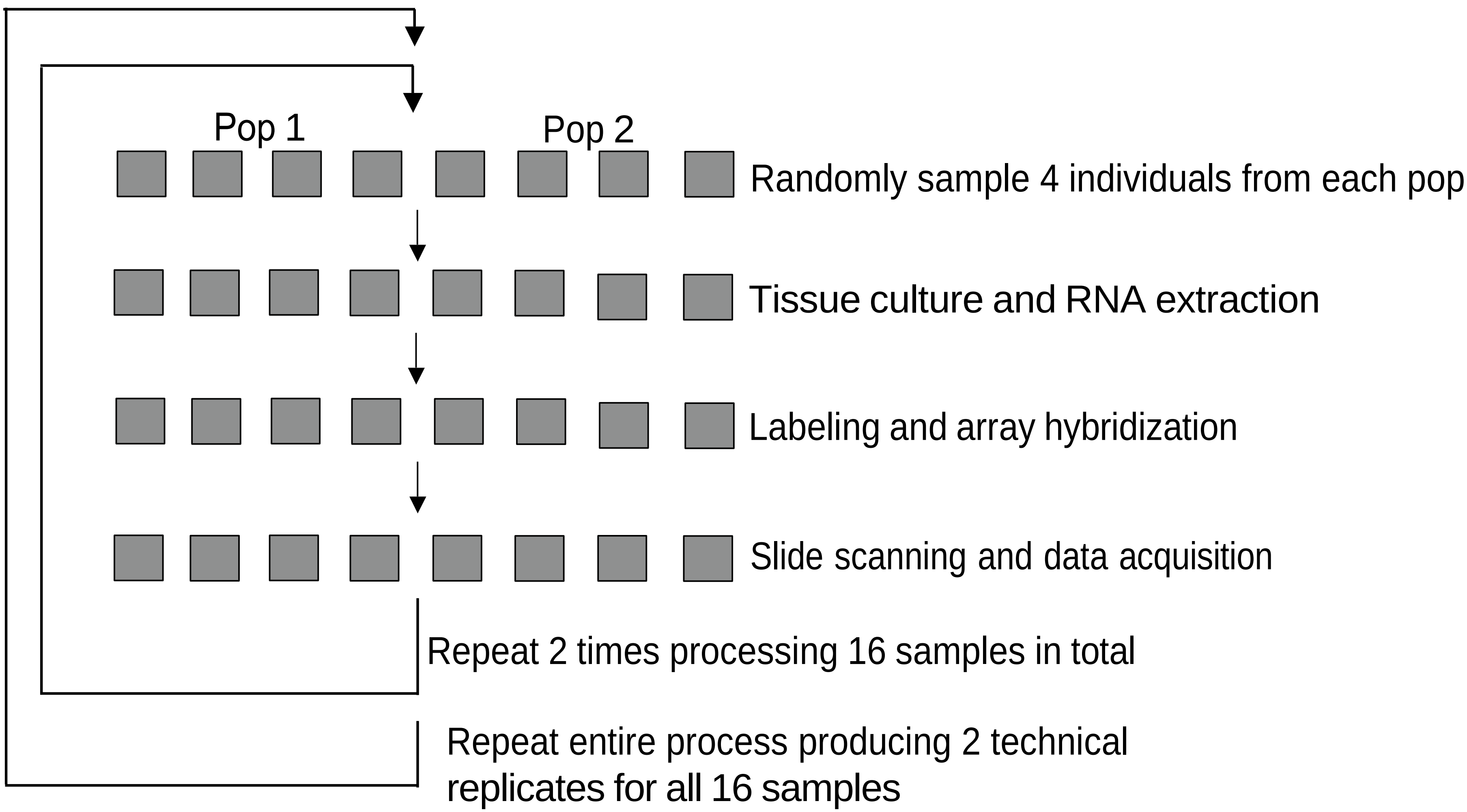
Descriptive Statistics and Exploratory Data Analysis

Further Thoughts on Experimental Design

- 16 Individuals (8 each from two populations) with replicates



Other Business

- Course web-site:

http://www.gs.washington.edu/academics/courses/akey/56008/index.htm

- Homework due on Thursday not Tuesday

- Make sure you look at HW1 soon and see either Shameek or myself with questions

Today

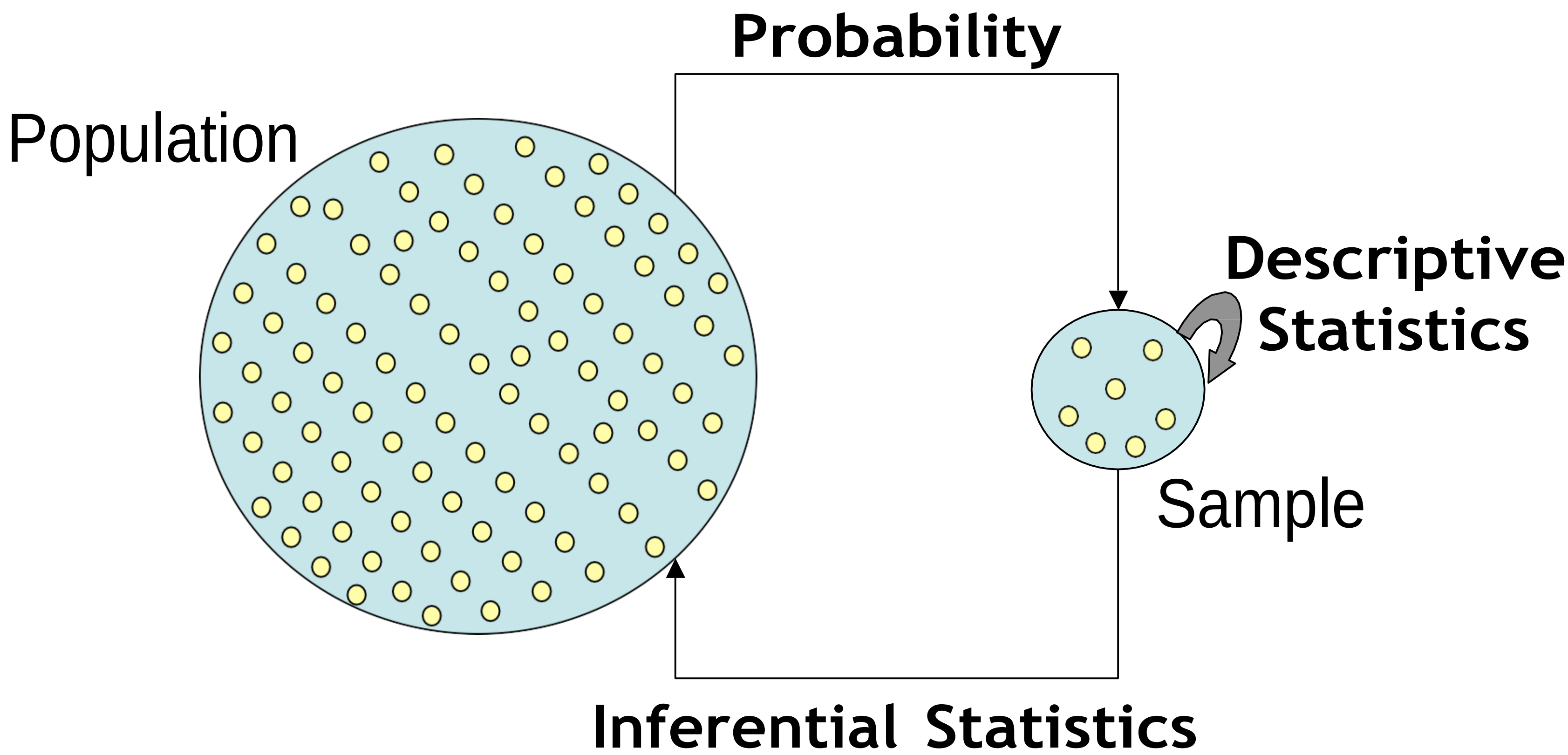
- What is descriptive statistics and exploratory data analysis?

- Basic numerical summaries of data

- Basic graphical summaries of data

- How to use R for calculating descriptive statistics and making graphs

“Central Dogma” of Statistics



EDA

Before making inferences from data it is essential to examine all your variables.

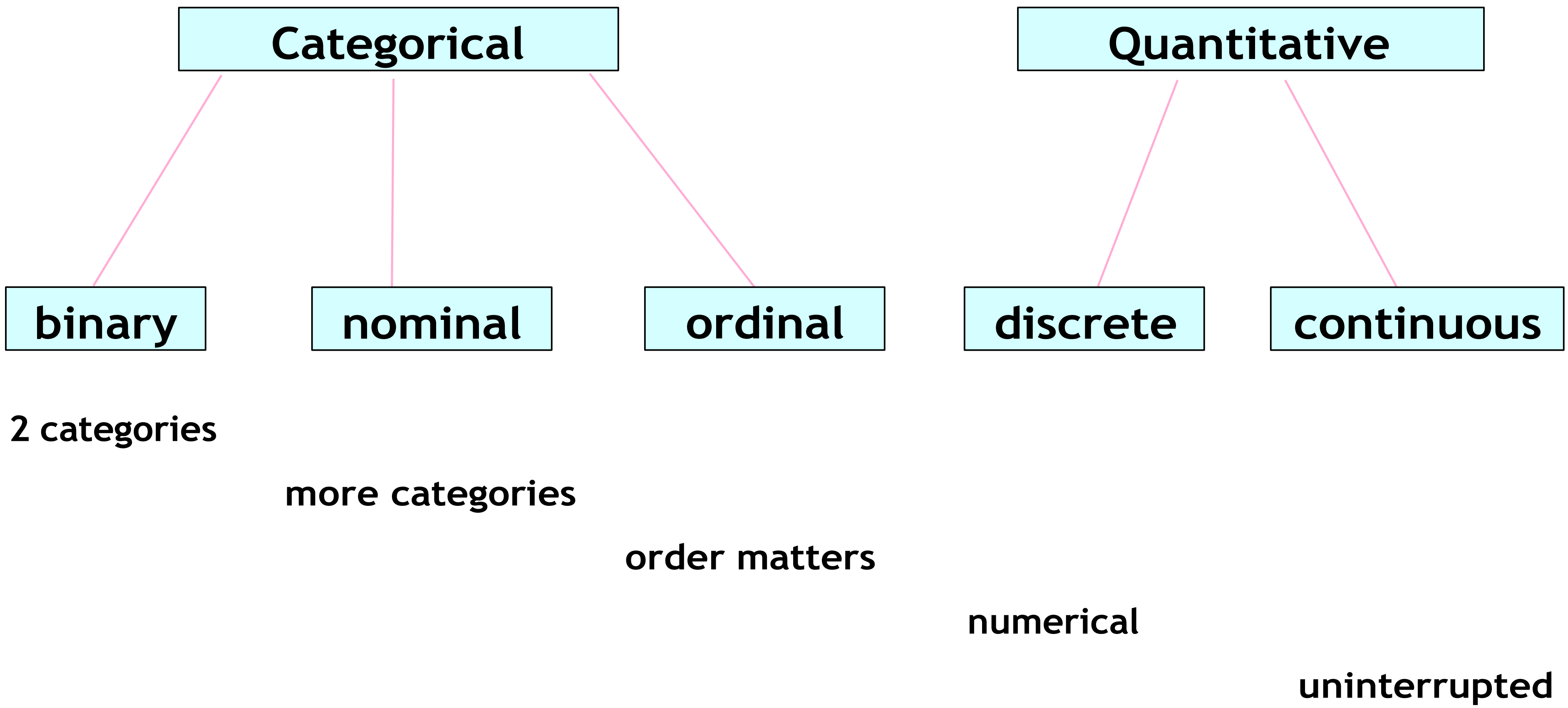
Why?

To ***listen*** to the data:

- to catch mistakes
- to see patterns in the data
- to find violations of statistical assumptions
- to generate hypotheses

…and because if you don't, you will have trouble later

Types of Data



Dimensionality of Data Sets

- **Univariate:** Measurement made on one variable per subject
- **Bivariate:** Measurement made on two variables per subject
- **Multivariate:** Measurement made on many variables per subject

Numerical Summaries of Data

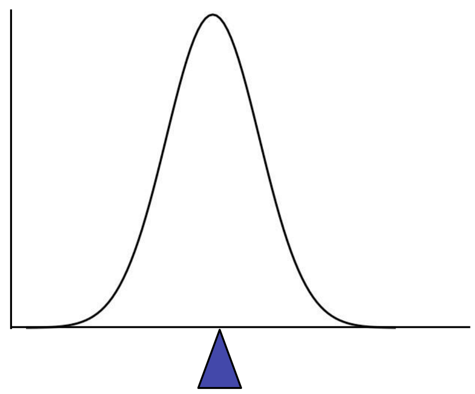
- ***Central Tendency measures***. They are computed to give a “center” around which the measurements in the data are distributed.
- ***Variation or Variability measures***. They describe “data spread” or how far away the measurements are from the center.
- ***Relative Standing measures***. They describe the relative position of specific measurements in the data.

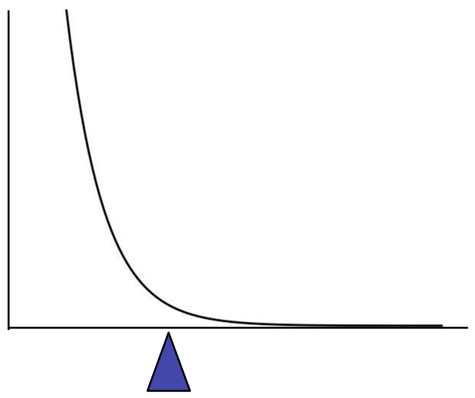
Location: Mean

1. The Mean

To calculate the average $\bar{x}$ of a set of observations, add their value and divide by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$





Other Types of Means

Weighted means:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed:

$$\bar{x} = \alpha$$

Geometric:

$$\bar{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Harmonic:

$$\bar{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Location: Median

- **Median** – the exact middle value

- **Calculation**:

  - If there are an odd number of observations, find the middle value

  - If there are an even number of observations, find the middle two values and average them

- **Example**

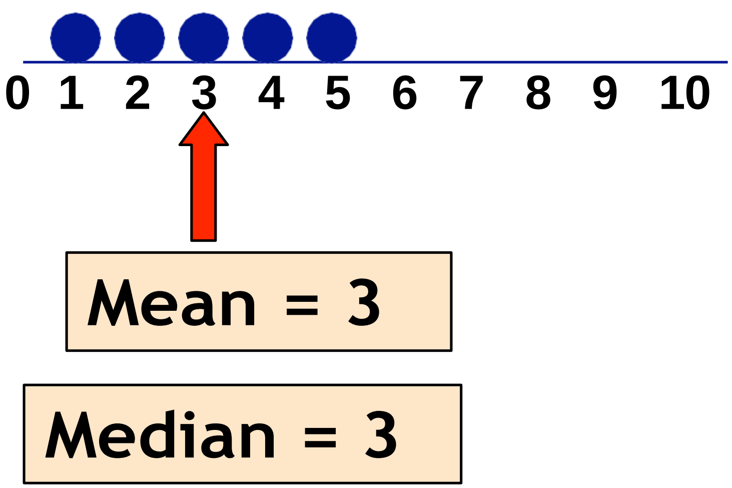
  <u>Some data</u>:

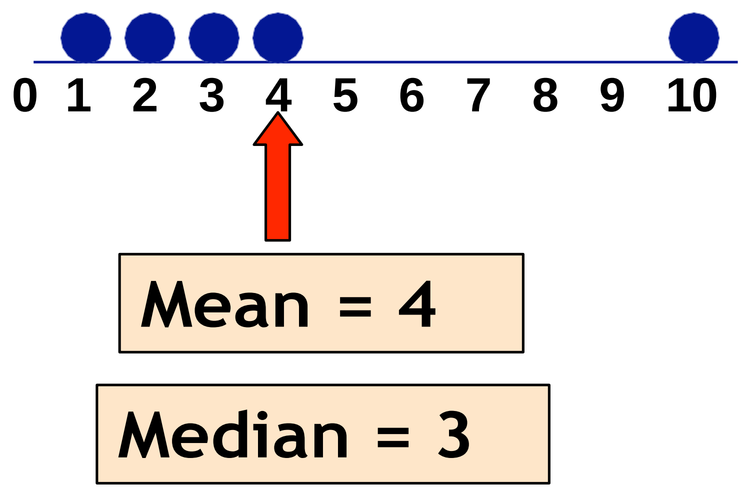
  Age of participants: 17 19 21 <u>22 23</u> 23 23 38

Median = (22+23)/2 = 22.5

Which Location Measure Is Best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers





Scale: Variance

- Average of squared deviations of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}$$

Why Squared Deviations?

- Adding deviations will yield a sum of ?
- Absolute values do not have nice mathematical properties
- Squares eliminate the negatives

- Result:
  - Increasing contribution to the variance as you go farther from the mean.

Scale: Standard Deviation

- Variance is somewhat arbitrary
- What does it mean to have a variance of 10.8? Or 2.2? Or 1459.092? Or 0.000001?
- Nothing. But if you could “standardize” that value, you could talk about any variance (i.e. deviation) in equivalent terms
- Standard deviations are simply the square root of the variance

Scale: Standard Deviation

$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}}$$

1. Score (in the units that are meaningful)
2. Mean
3. Each score’s deviation from the mean
4. Square that deviation
5. Sum all the squared deviations (Sum of Squares)
6. Divide by n-1
7. Square root – now the value is in the units we started with!!!

Interesting Theoretical Result

- Regardless of how the data are distributed, a certain percentage of values must fall within k standard deviations from the mean:

Note use of μ (mu) to represent "mean".

Note use of σ (sigma) to represent "standard deviation."

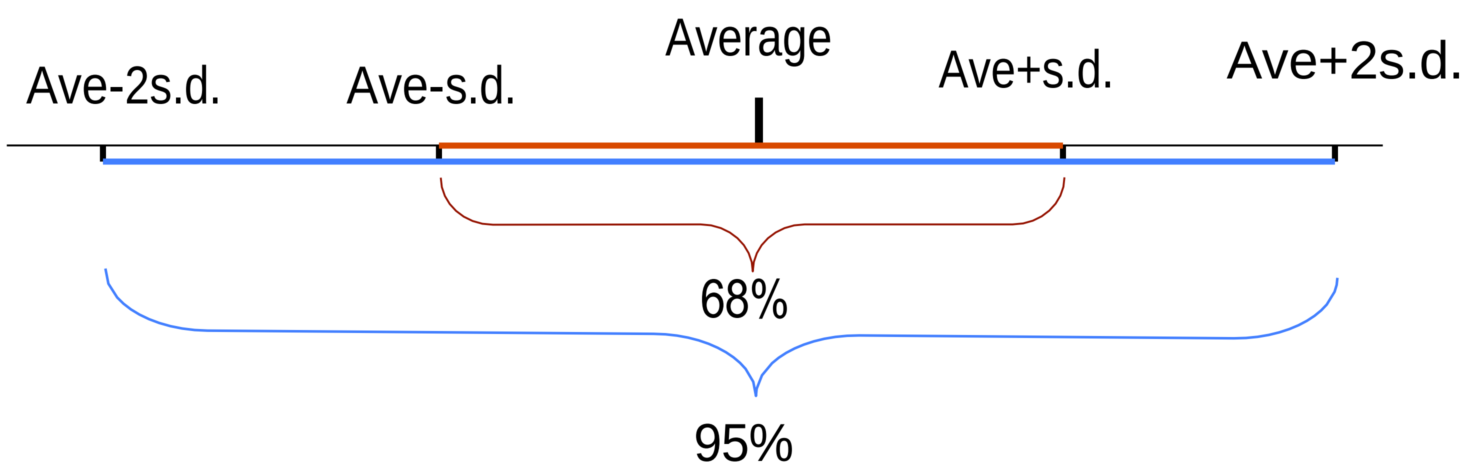
At least	within
$(1 - 1/1^2) = 0\%$	$k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$	$k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	$k=3$ $(\mu \pm 3\sigma)$

Often We Can Do Better

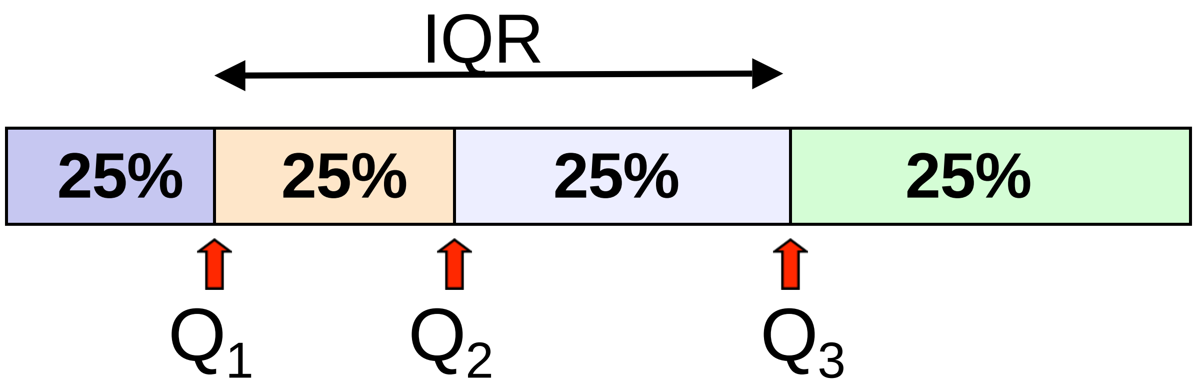
For many lists of observations – especially if <u>their histogram is bell-shaped</u>

1. Roughly 68% of the observations in the list lie within <u>1 standard deviation</u> of the average

2. 95% of the observations lie within <u>2 standard deviations</u> of the average



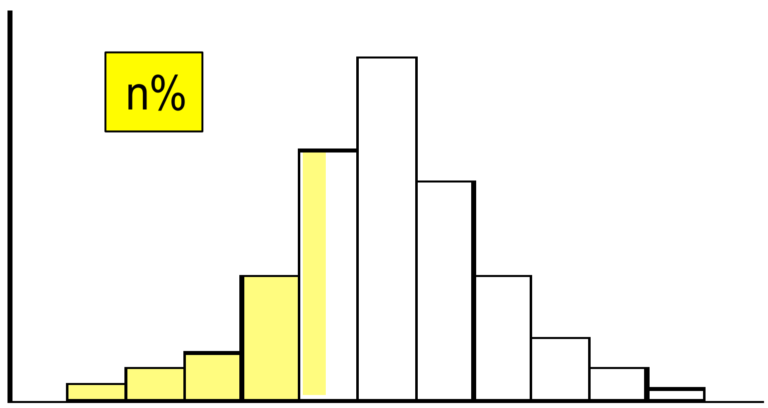
Scale: Quartiles and IQR



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Percentiles (aka Quantiles)

In general the **n^{th} percentile** is a value such that n% of the observations fall at or below or it



Q_1 = 25^{th} percentile

Median = 50^{th} percentile

Q_2 = 75^{th} percentile

Graphical Summaries of Data

A (Good) Picture Is
Worth A 1,000 Words

Univariate Data: Histograms and Bar Plots

- What's the difference between a histogram and bar plot?

Bar plot

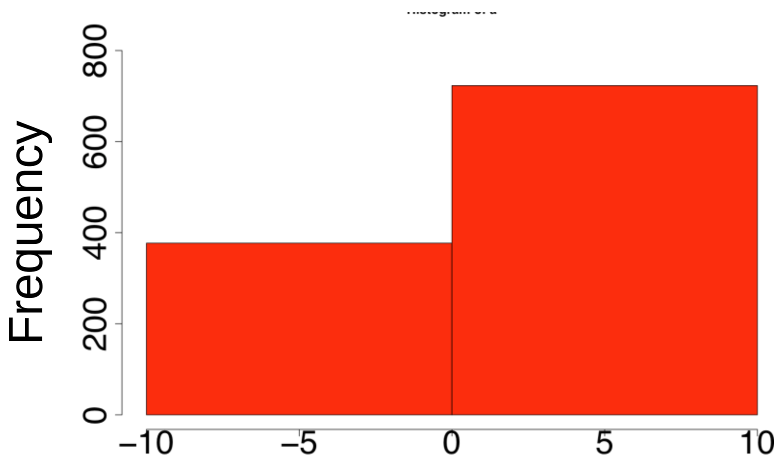
- Used for categorical variables to show frequency or proportion in each category.
- Translate the data from frequency tables into a pictorial representation…

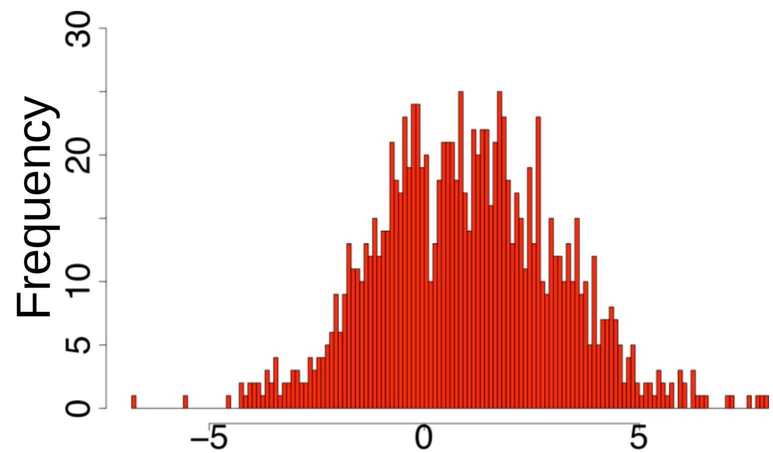
Histogram

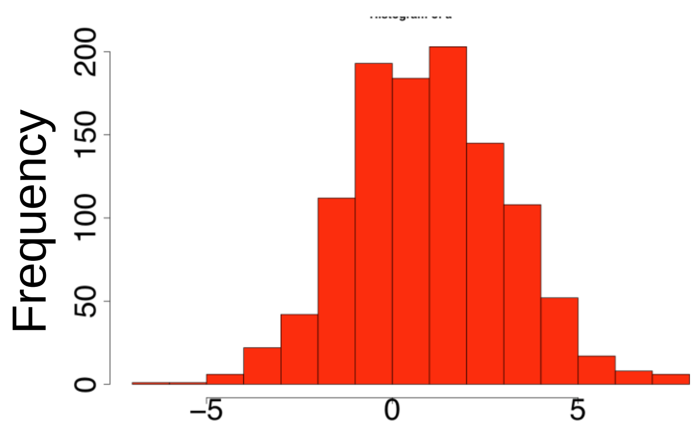
- Used to visualize distribution (shape, center, range, variation) of continuous variables
- "Bin size" important

Effect of Bin Size on Histogram

- Simulated 1000 N(0,1) and 500 N(1,1)

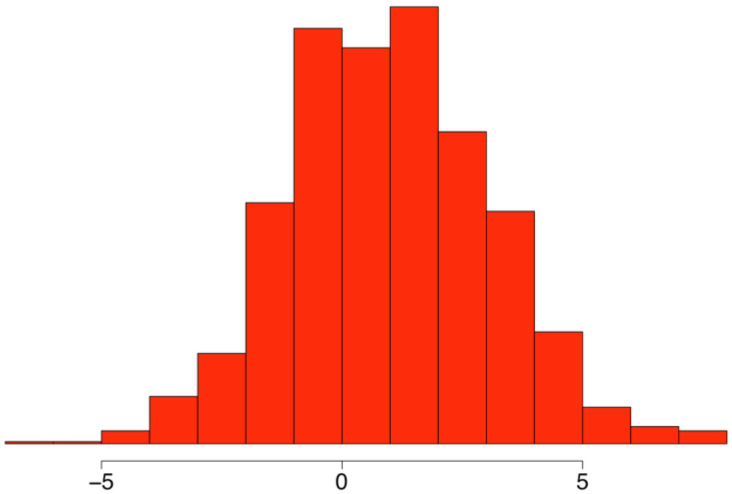


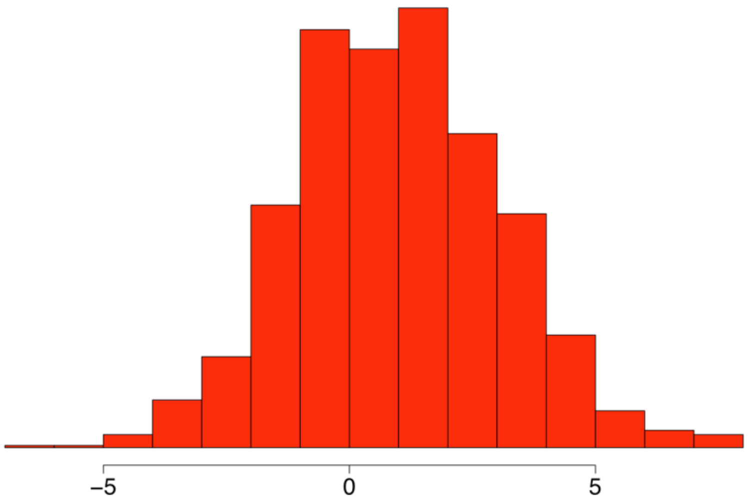




More on Histograms

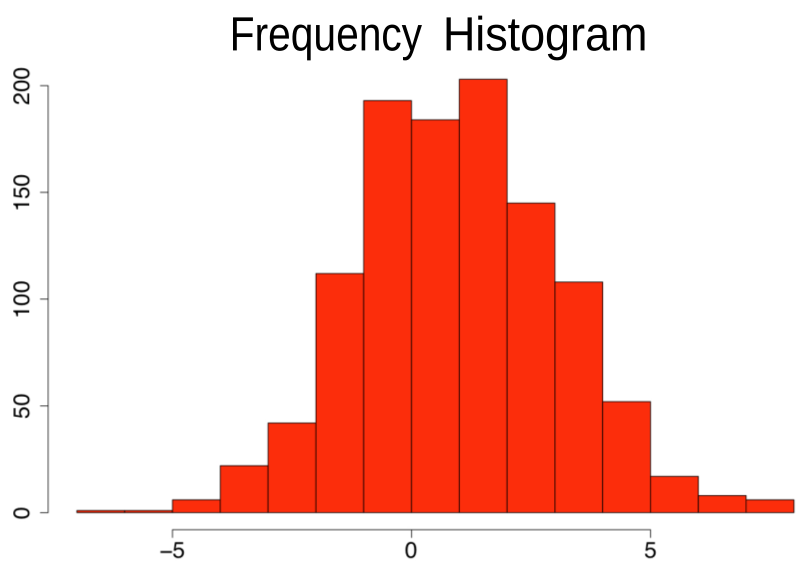
- What's the difference between a frequency histogram and a density histogram?

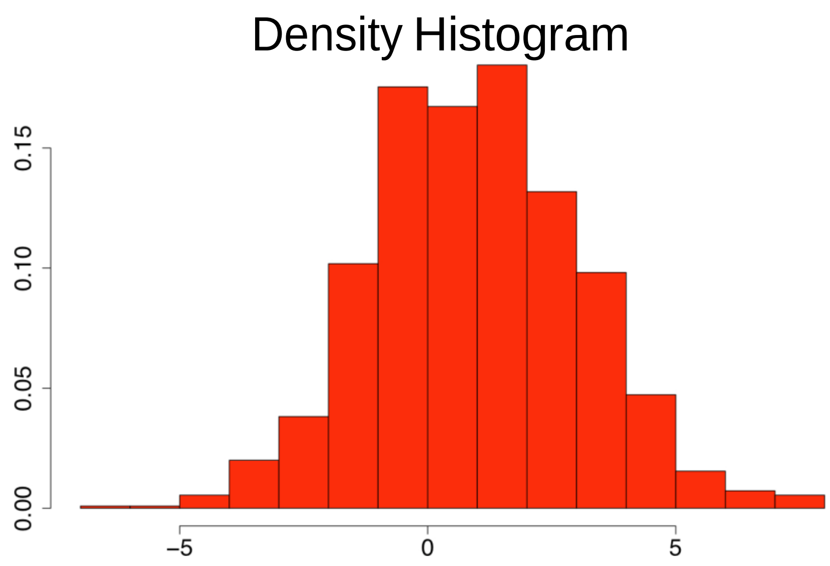




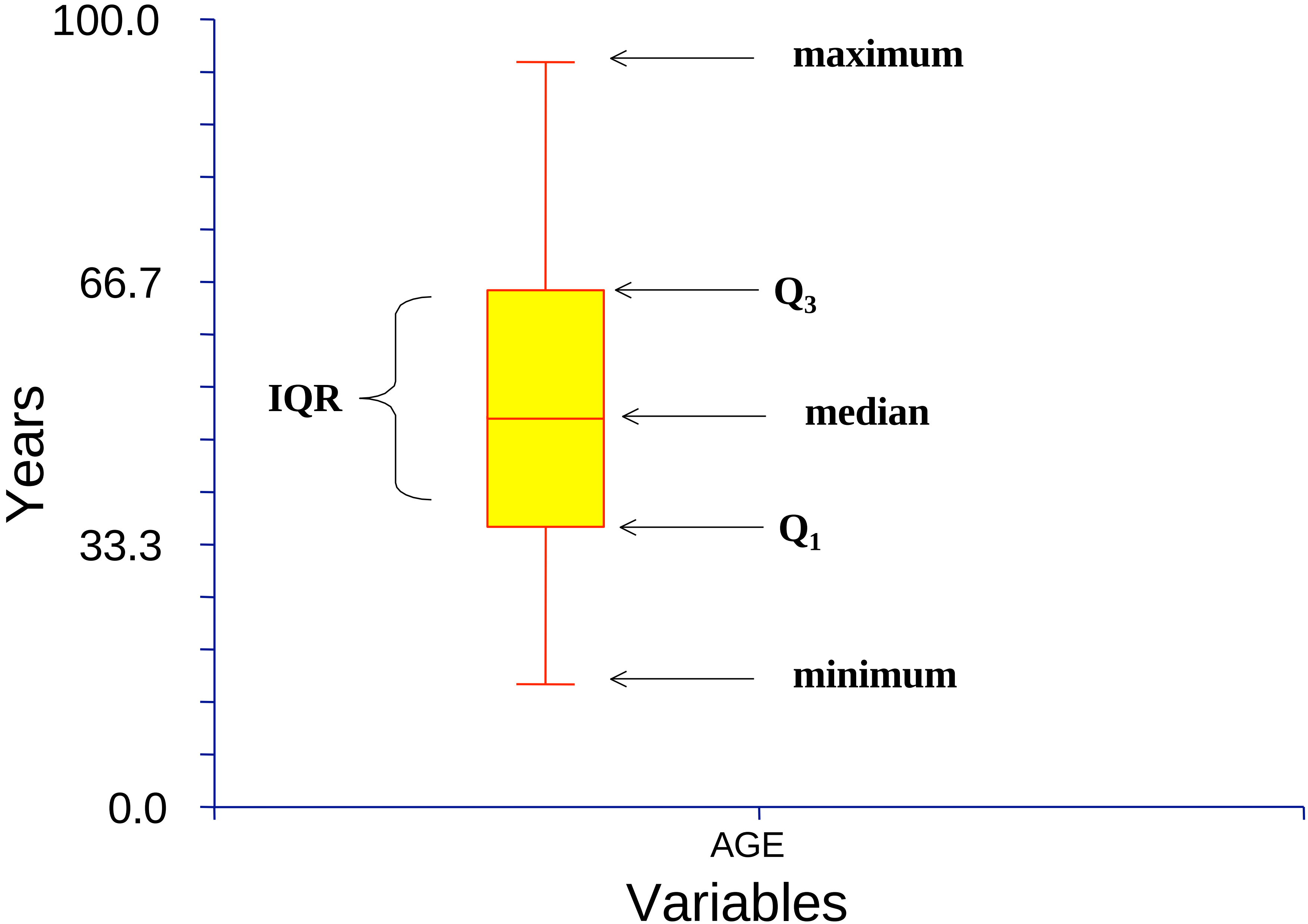
More on Histograms

- What's the difference between a frequency histogram and a density histogram?





Box Plots



100.0
66.7
33.3
0.0
Years
maximum
Q3
median
IQR
Q1
minimum
AGE
Variables

Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Box Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot Stacked Box Plot

Multivariate Data

Clustering

- Organize ***units*** into clusters
- Descriptive, not inferential
- Many approaches
- “Clusters” always produced

Data Reduction Approaches (PCA)

- Reduce n-dimensional dataset into much smaller number
- Finds a new (smaller) set of variables that retains most of the information in the total sample
- Effective way to visualize multivariate data

How to Make a Bad Graph

The aim of good data graphics:

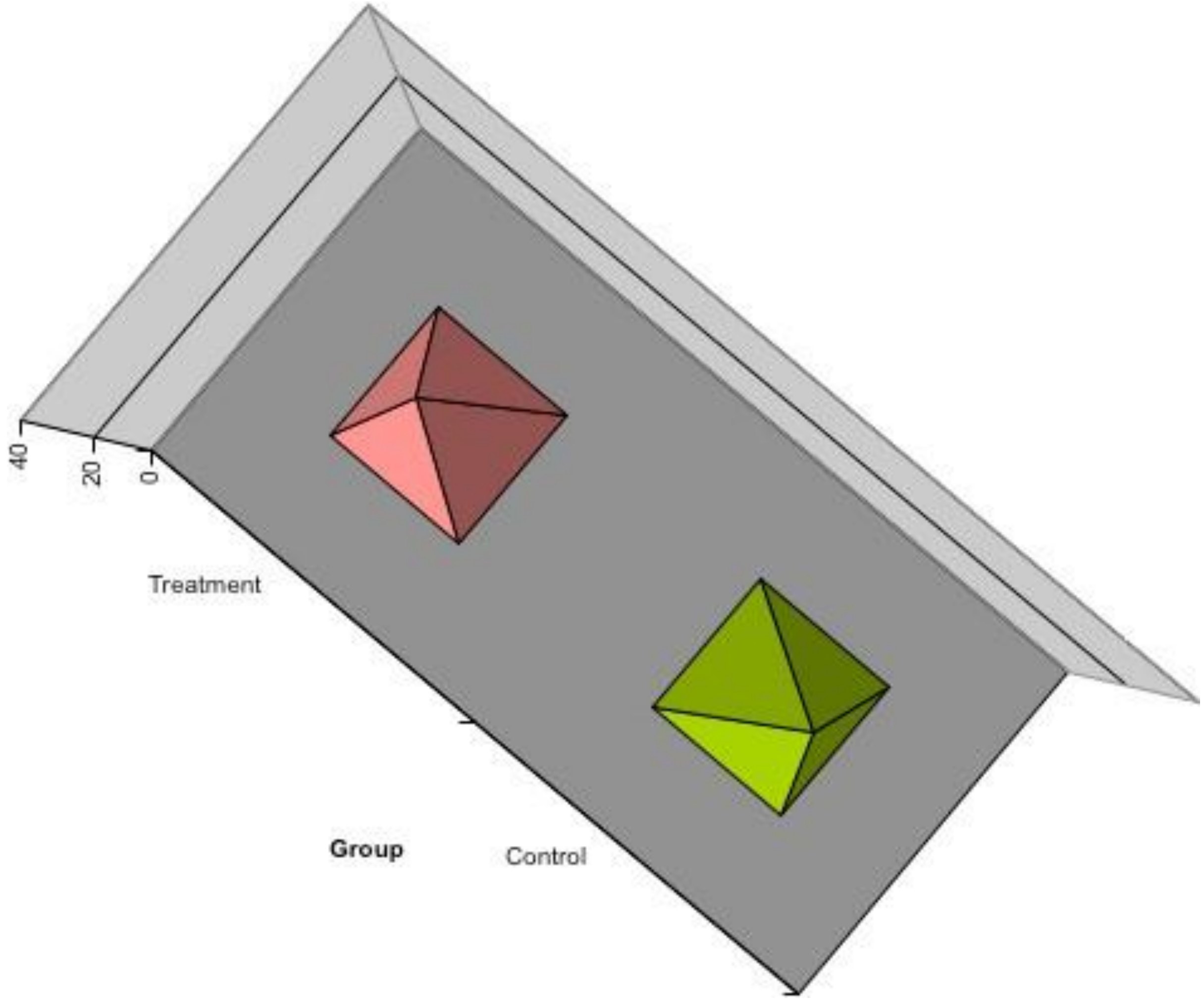
Display data accurately and clearly

Some rules for displaying data badly:

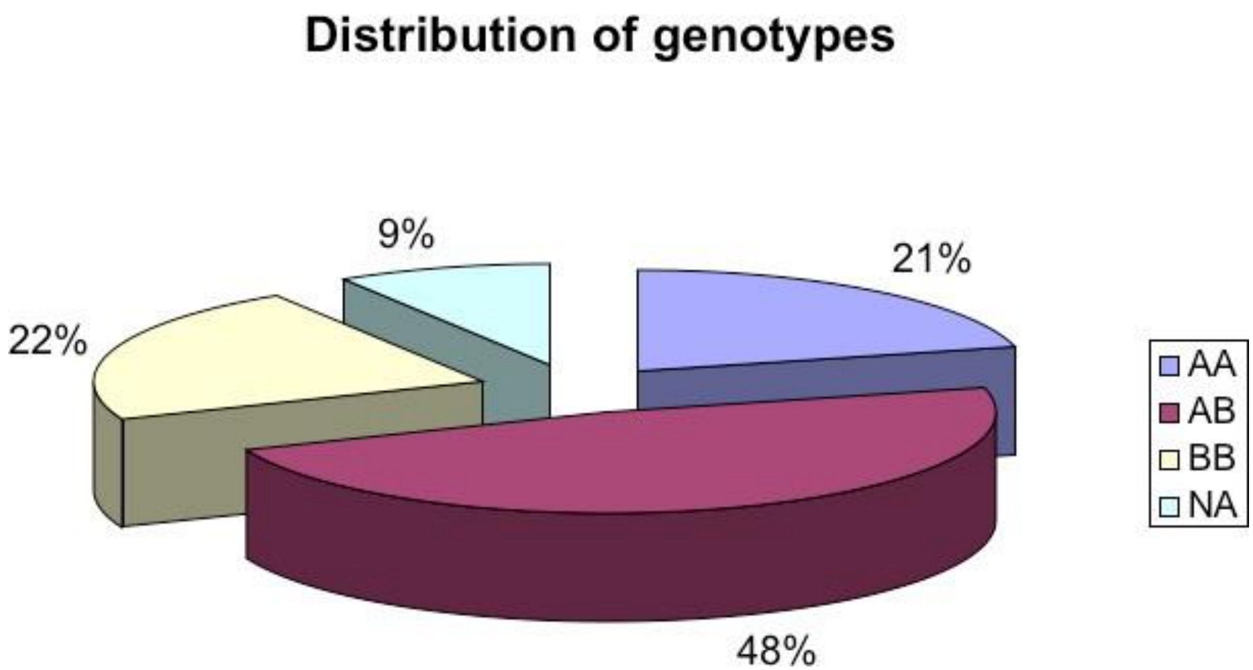
- Display as little information as possible
- Obscure what you do show (with chart junk)
- Use pseudo-3d and color gratuitously
- Make a pie chart (preferably in color and 3d)
- Use a poorly chosen scale

From Karl Broman: http://www.biostat.wisc.edu/~kbroman/

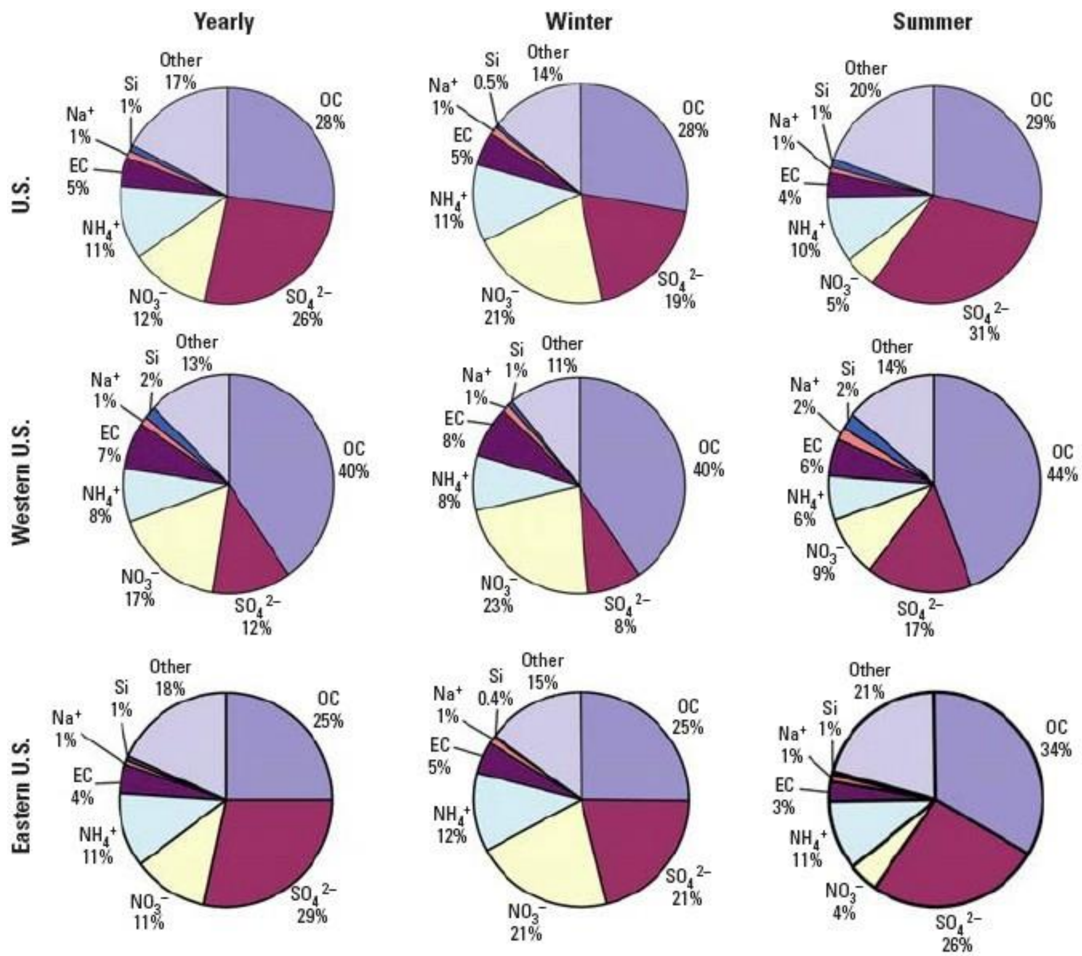
Example 1



Example 2

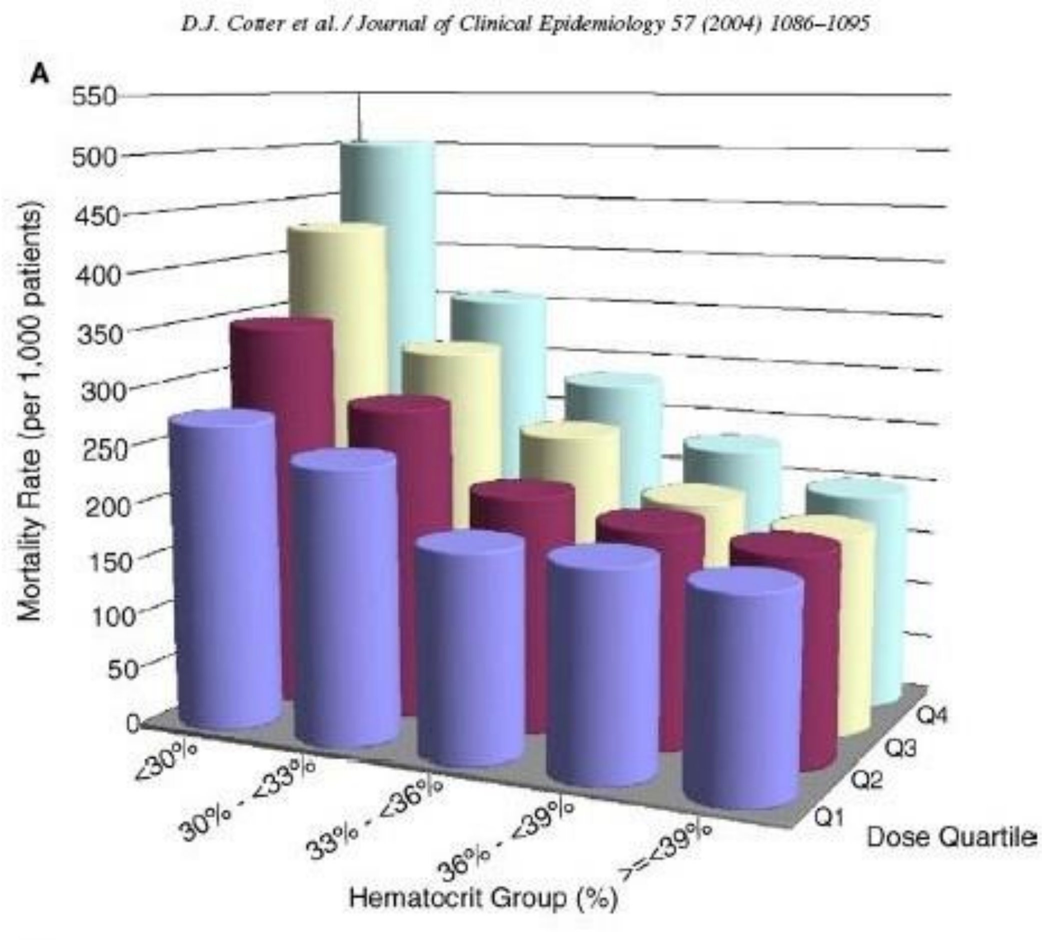


Example 3

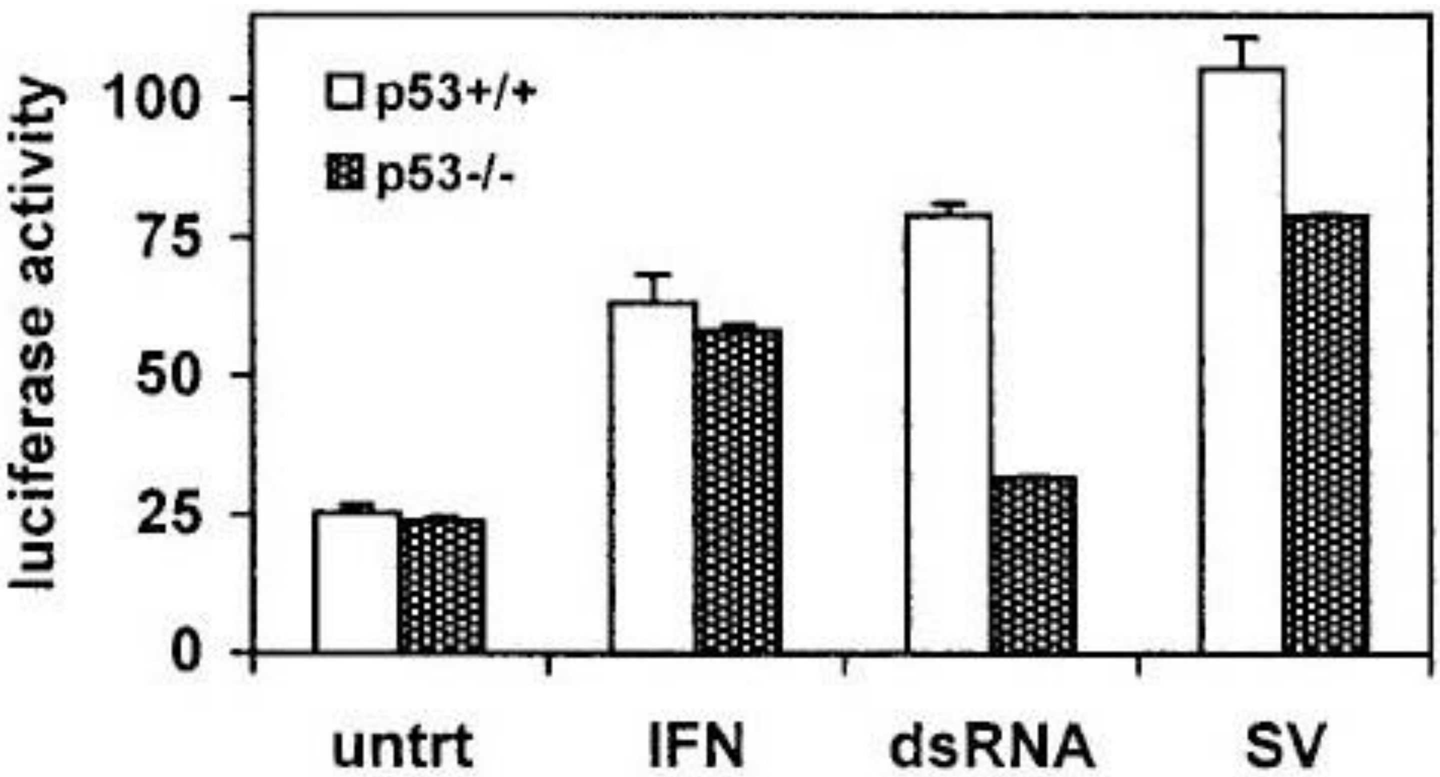


Yearly
Winter
Summer
U.S.
Western U.S.
Eastern U.S.
Other 17%
Si 1%
Na+ 1%
EC 5%
NH4+ 11%
NO3− 12%
SO4 2− 26%
OC 28%
Other 14%
Si 0.5%
Na+ 1%
EC 5%
NH4+ 11%
NO3− 21%
SO4 2− 19%
OC 28%
Other 20%
Si 1%
Na+ 1%
EC 4%
NH4+ 10%
NO3− 5%
SO4 2− 31%
OC 29%
Other 13%
Si 2%
Na+ 1%
EC 7%
NH4+ 8%
NO3− 17%
SO4 2− 12%
OC 40%
Other 11%
Si 1%
Na+ 1%
EC 8%
NH4+ 8%
NO3− 23%
SO4 2− 8%
OC 40%
Other 14%
Si 2%
Na+ 2%
EC 6%
NH4+ 6%
NO3− 9%
SO4 2− 17%
OC 44%
Other 18%
Si 1%
Na+ 1%
EC 4%
NH4+ 11%
NO3− 11%
SO4 2− 29%
OC 25%
Other 15%
Si 0.4%
Na+ 1%
EC 5%
NH4+ 12%
NO3− 21%
SO4 2− 21%
OC 25%
Other 21%
Si 1%
Na+ 1%
EC 3%
NH4+ 11%
NO3− 4%
SO4 2− 26%
OC 34%

Example 4


D.J. Cotter et al. / Journal of Clinical Epidemiology 57 (2004) 1086–1095
A
Mortality Rate (per 1,000 patients)
550
500
450
400
350
300
250
200
150
100
50
0
<30%
30% - <33%
33% - <36%
36% - <39%
>=<39%
Hematocrit Group (%)
Q4
Q3
Q2
Q1
Dose Quartile

Example 5



R Tutorial

- Calculating descriptive statistics in R
- Useful R commands for working with multivariate data (apply and its derivatives)
- Creating graphs for different types of data (histograms, boxplots, scatterplots)
- Basic clustering and PCA analysis