Exploratory data analysis

- Also called *descriptive statistics*, this term is used to describe the process of 'looking at the data' prior to formal analysis
- In this phase of analysis, data are examined for *quality* and 'cleaned' as well as *displayed* to provide an overall impression of results
- We will look at two types of summaries:
  - Graphical summaries
  - Numerical summaries
- Necessary to use *statistical software*

Why R?

- Powerful, flexible, and extensible statistical computing language and environment
- Wide range of built-in statistical functions and add-on packages available, including a growing number specifically for microarray data analysis
- High quality, customizable graphics capabilities
- Available for Unix/Linux, Windows, Mac
- All this and … R is free!

Variables (I)

- Statisticians call characteristics which can differ across individuals *variables*
- Types of variables
  - *Categorical* (also called *qualitative*)
    - Examples: eye color, favorite television program
  - *Numerical* (also called *quantitative*)
    - Examples: height, number of children, fluorescence intensity

Variables (II)

- Categorical variables may be
  - *Nominal* – the categories have names, but no ordering (e.g. eye color)
  - *Ordinal* – categories have an ordering (e.g. ‘Always’, ‘Sometimes’, ‘Never’)
- Numerical variables may be
  - *Discrete* – possible values can differ only by fixed amounts (most commonly counting values)
  - *Continuous* – can take on any value within a range (e.g. any positive value)

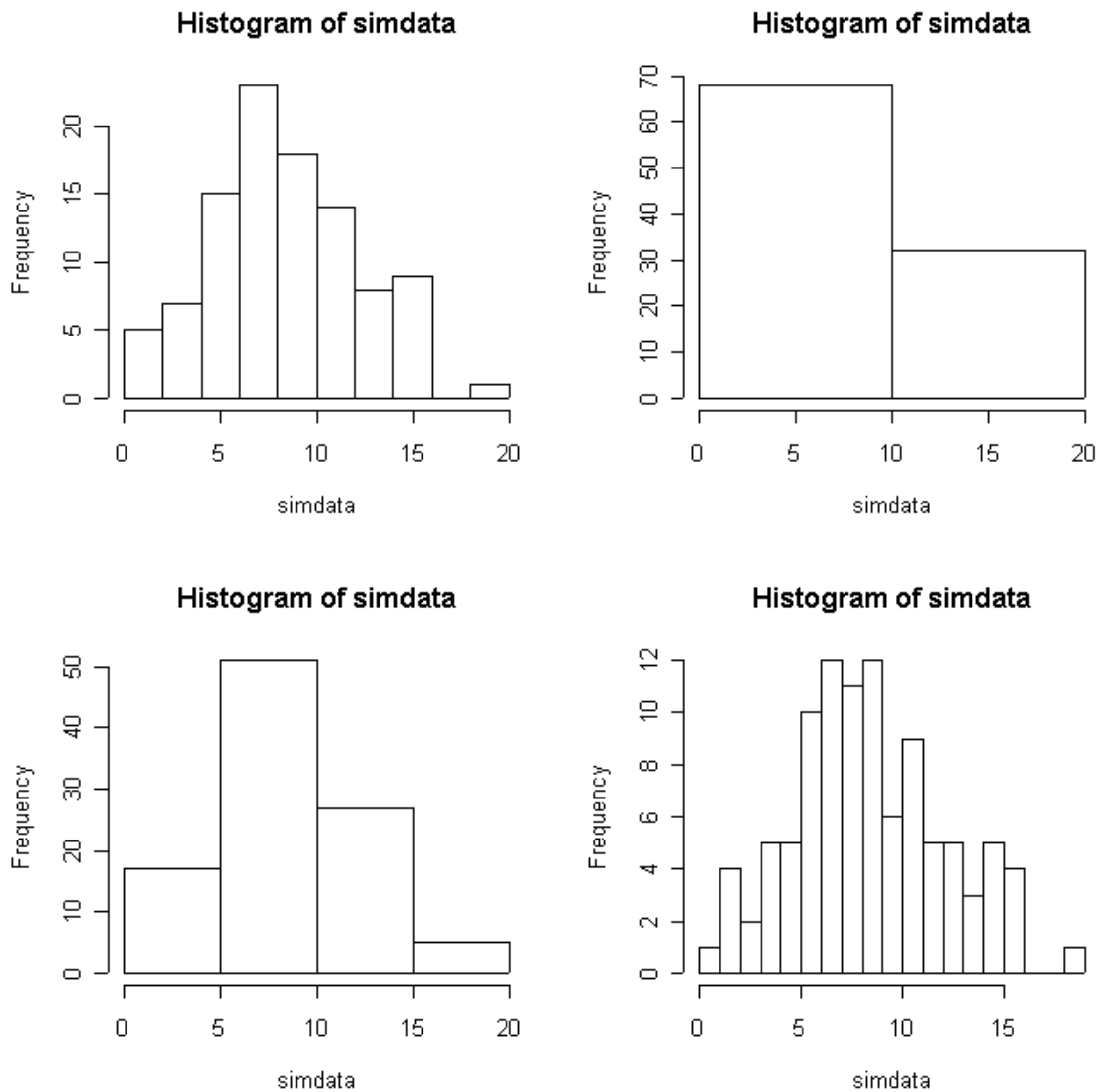
Univariate Data

- Measurements on *a single (continuous)* variable *X*
- Summarizing *X*
  - Graphically:
    - Distribution: histogram, QQ plot, dotplot, boxplot
    - Quality: cluster analysis, PCA, spatial plots
  - Numerically:
    - Distribution: quantiles
    - Center: mean, median
    - Spread: SD, IQR, MAD

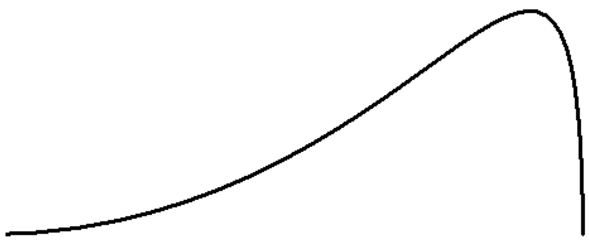
Bivariate / Multivariate Data

- *Bivariate (or multivariate) data* – data with measurements on *two (or more)* variables
- Here, we will look at two *continuous* variables
- Want to explore the *relationship* between the two variables
  - Graphically: scatterplot
  - Numerically: correlation coefficient

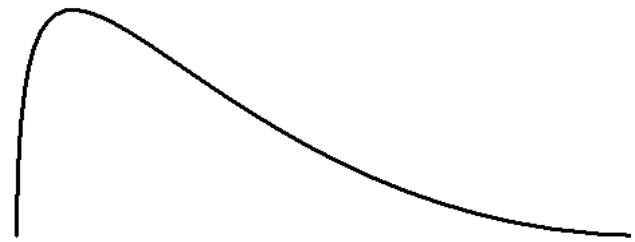
Histogram: same data



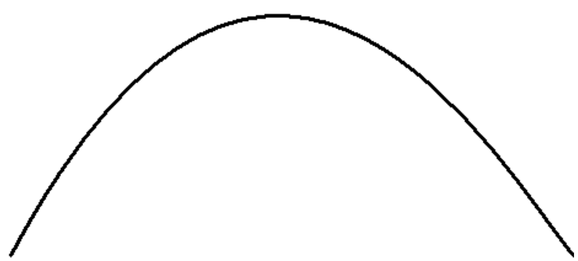
Some general histogram forms



left-skewed

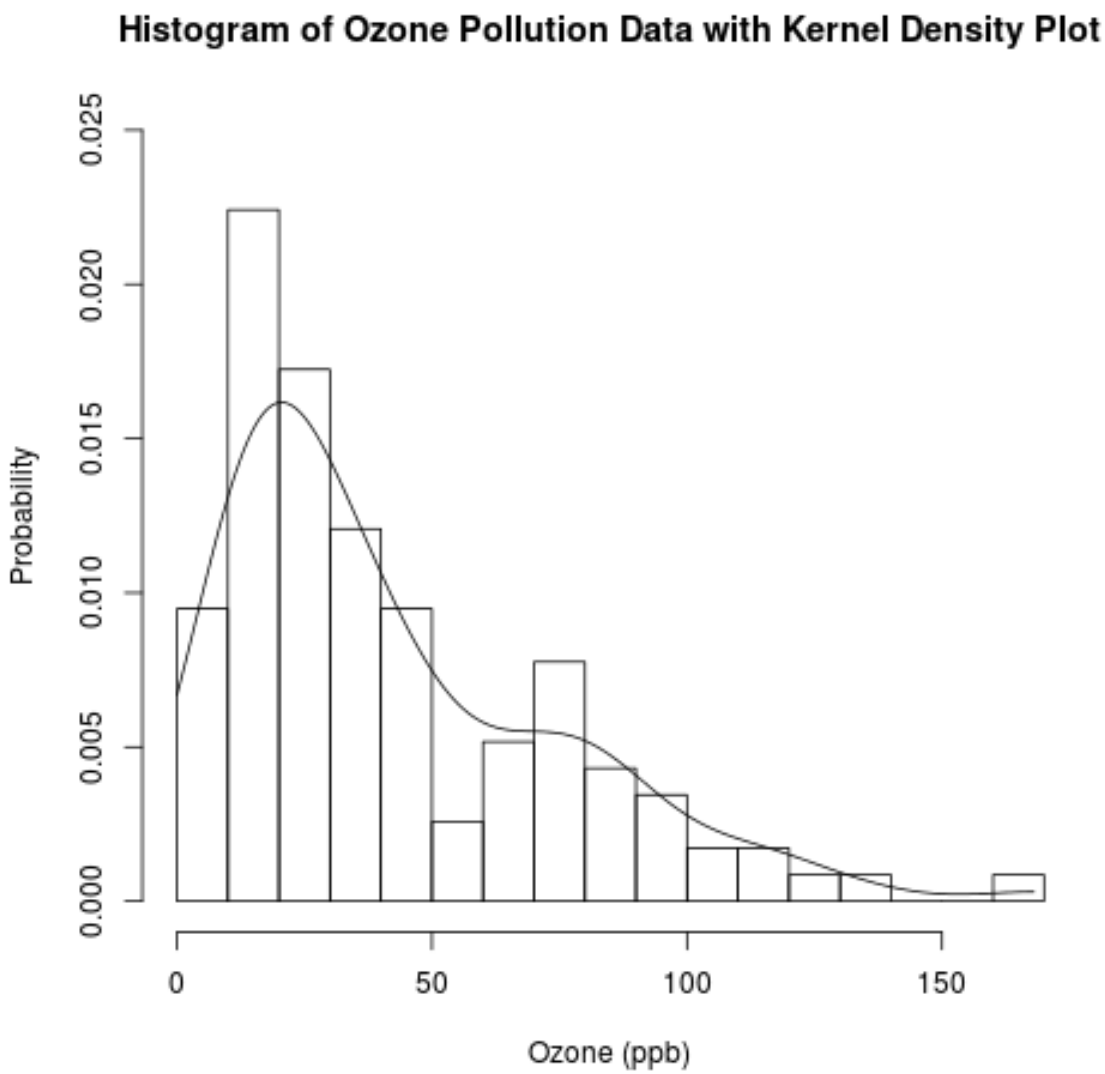


right-skewed

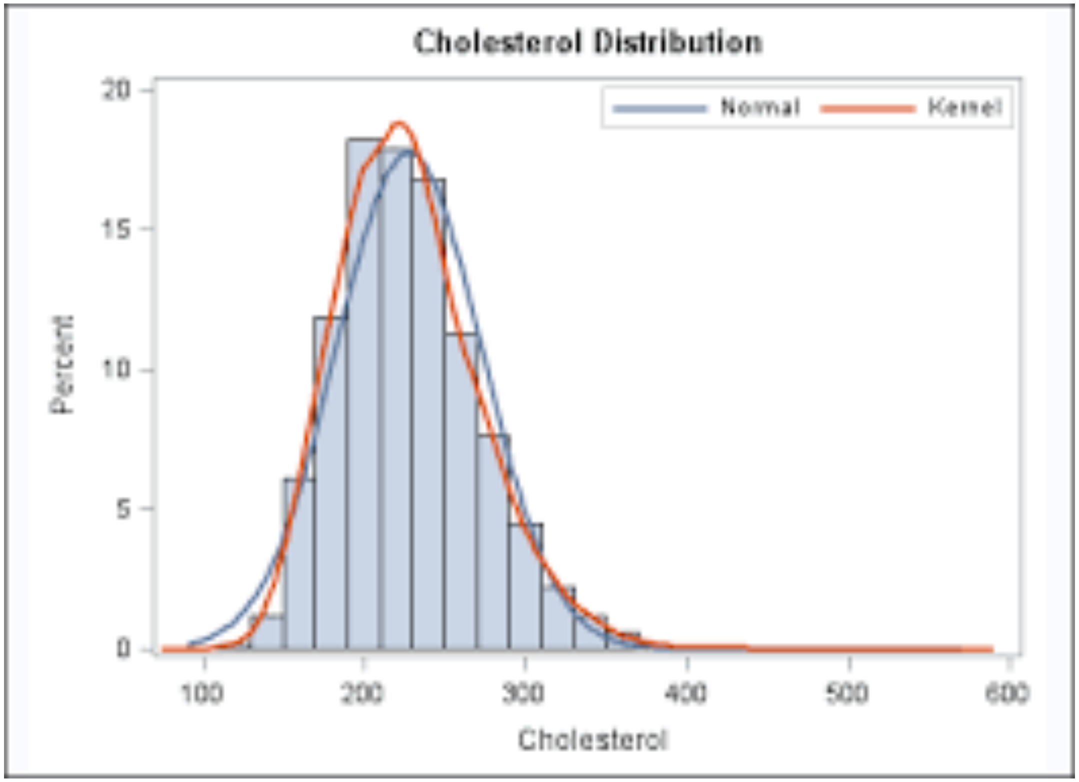


symmetric

Histogram: bars and smoothed



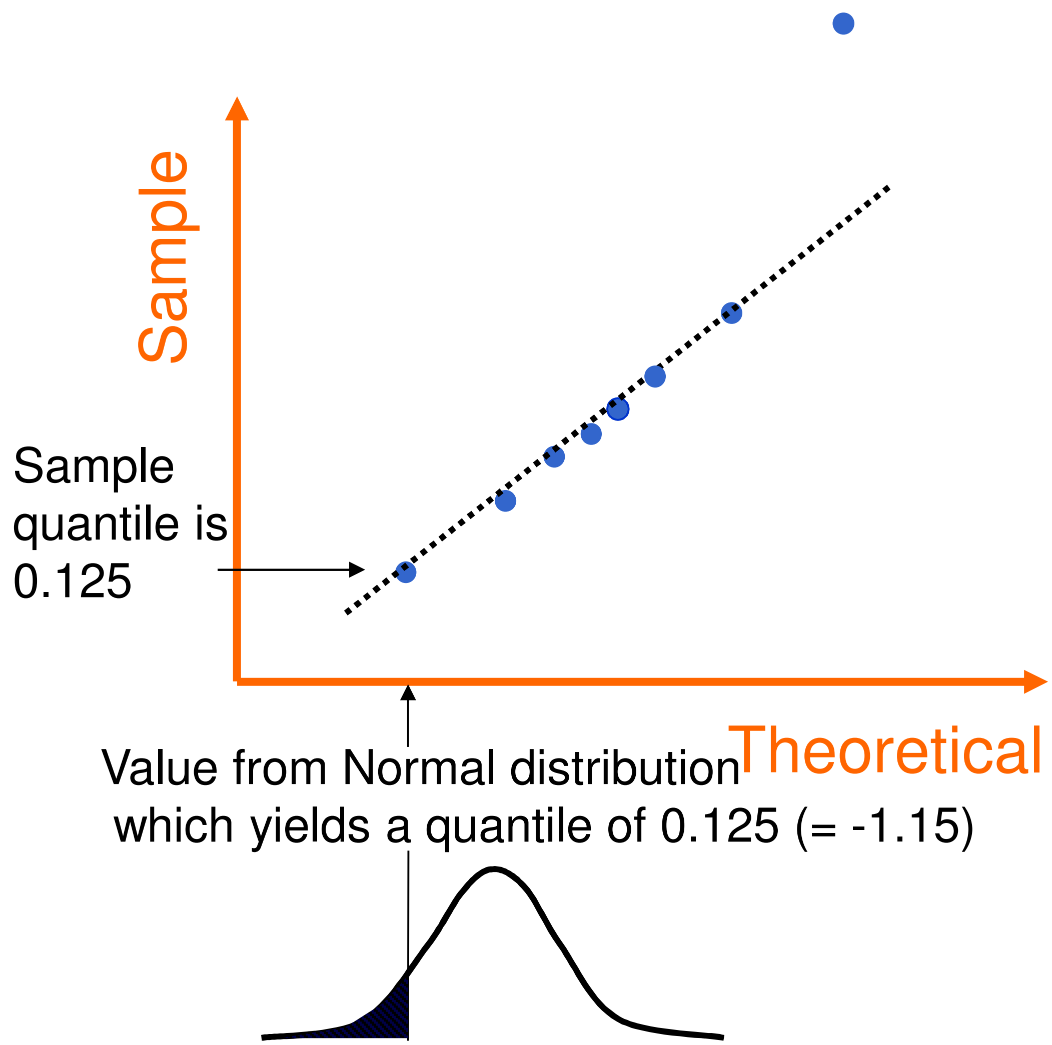
Histogram: comparing distributions



- Histogram, smoothed histogram (kernel), normal density
- NOT the best way to compare distributions (use QQ plot)

QQ-Plot

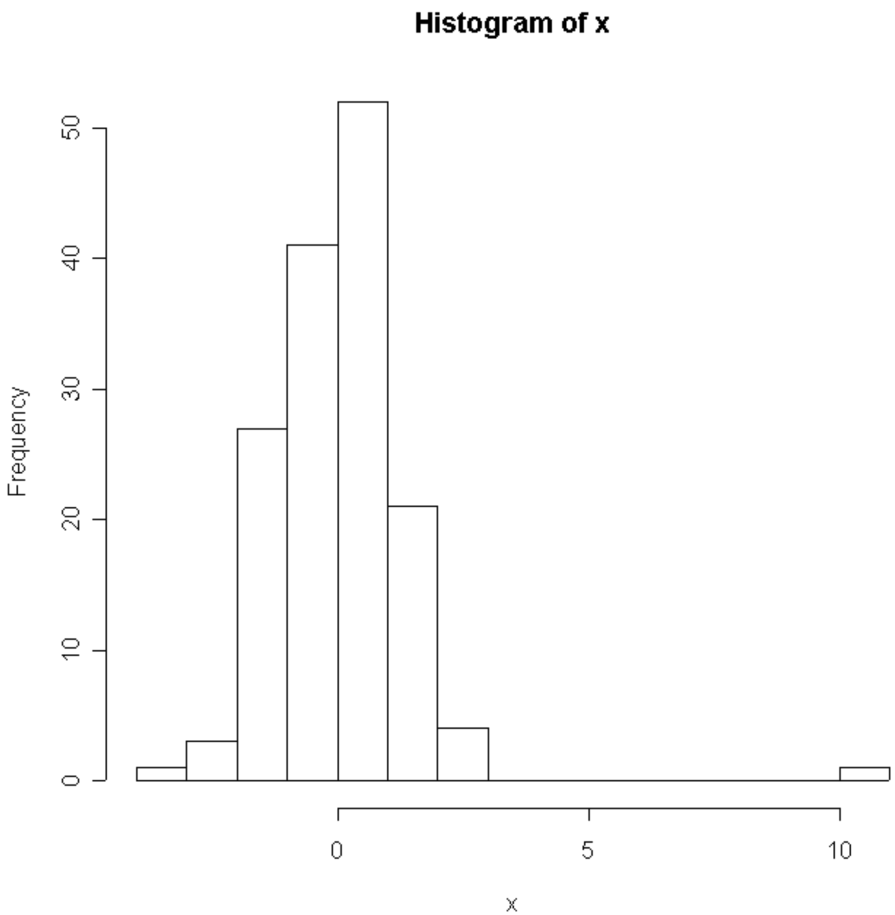
- Quantile-quantile plot
- Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples)
- A method for looking for outliers when data are mostly normal

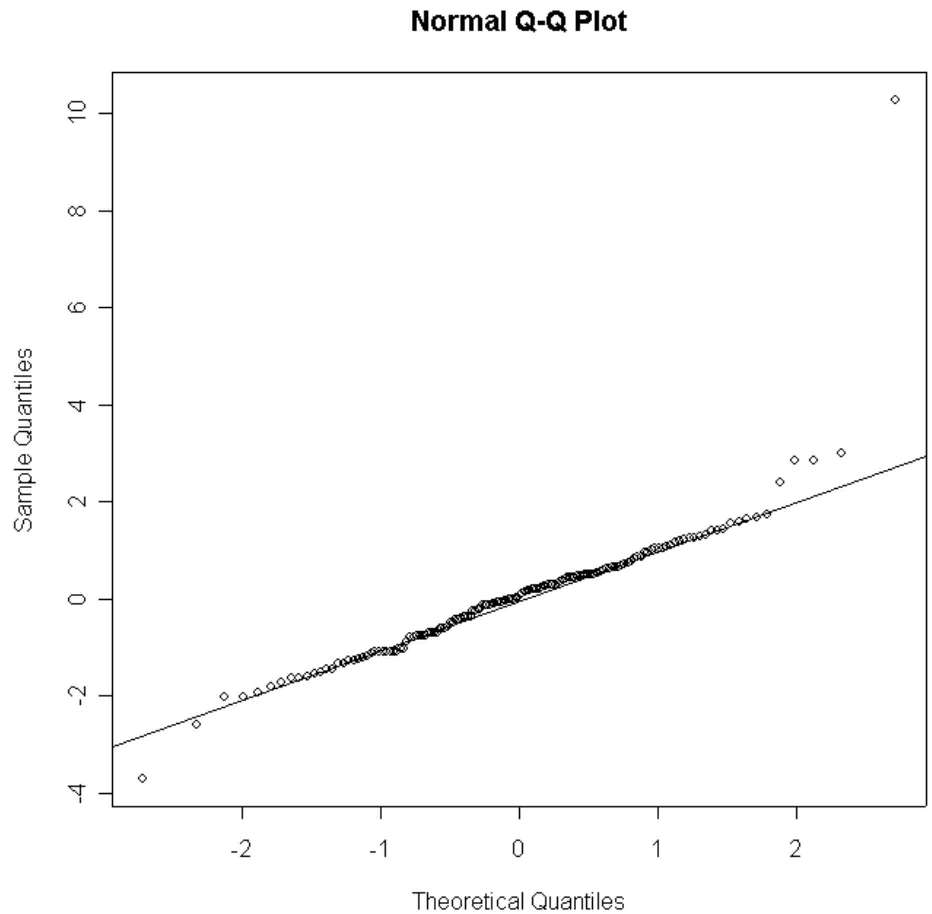


Typical deviations from straight line patterns

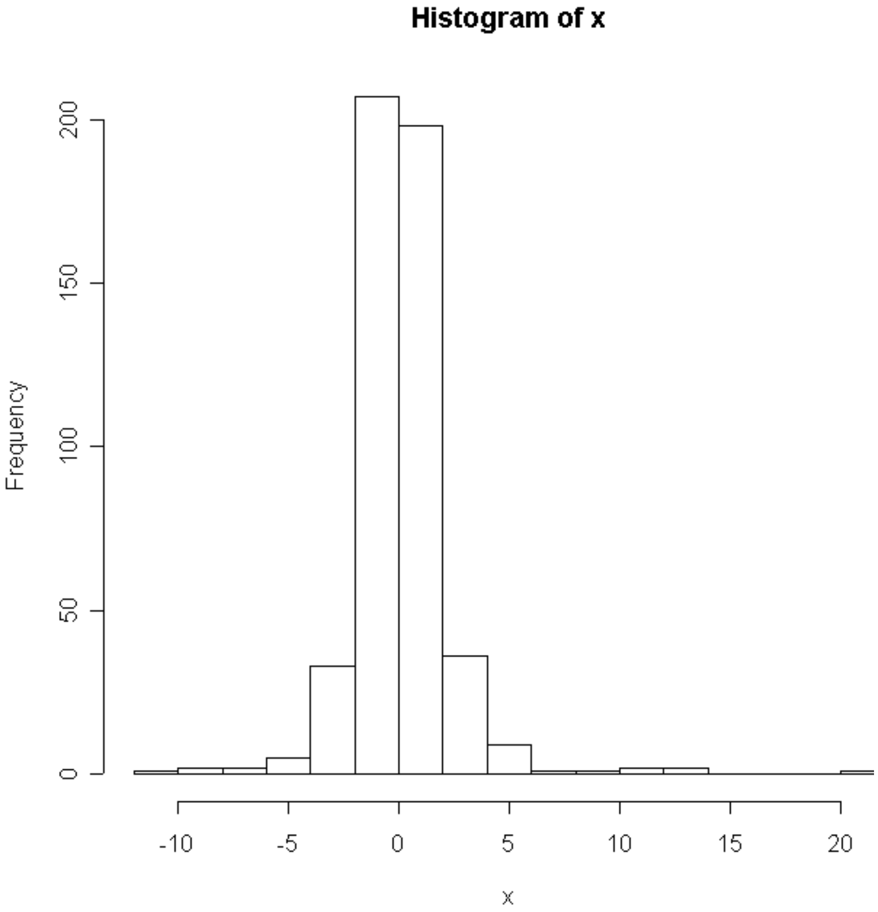
- Outliers
- Curvature at both ends (long or short tails)
- Convex/concave curvature (asymmetry)
- Horizontal segments, plateaus, gaps

Outliers

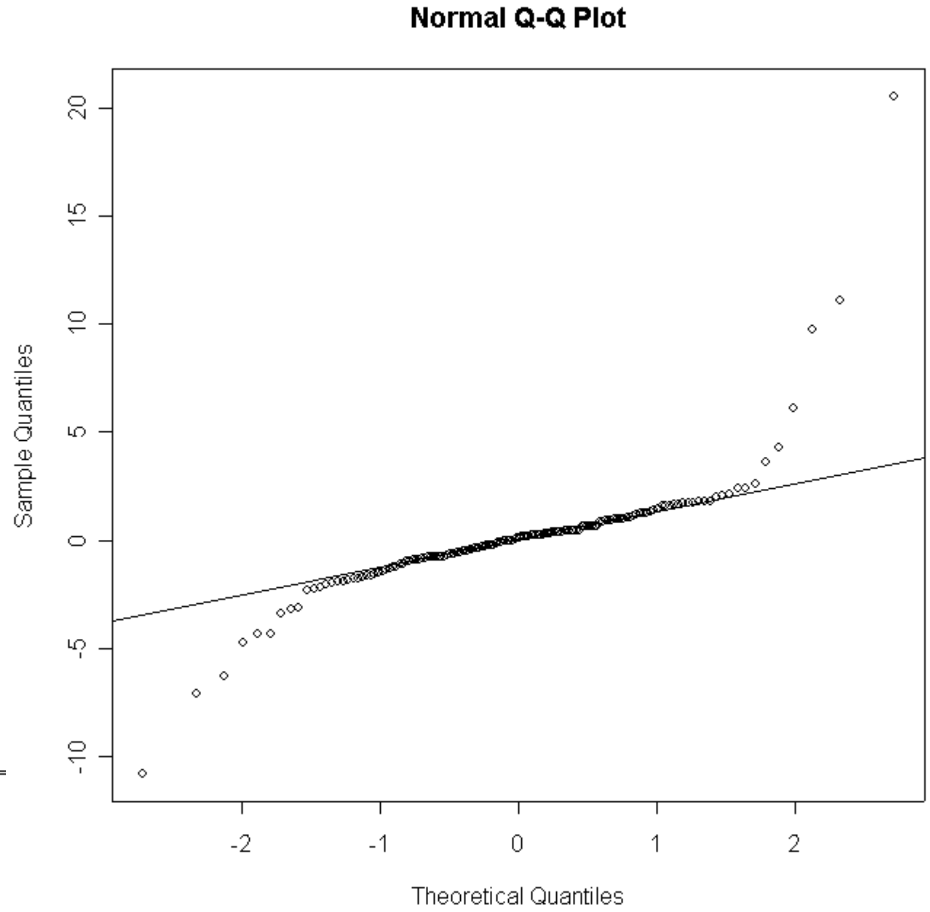

Histogram of x
Frequency
0
10
20
30
40
50
0
5
10
x


Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
6
8
10
-2
-1
0
1
2
Theoretical Quantiles

Long Tails

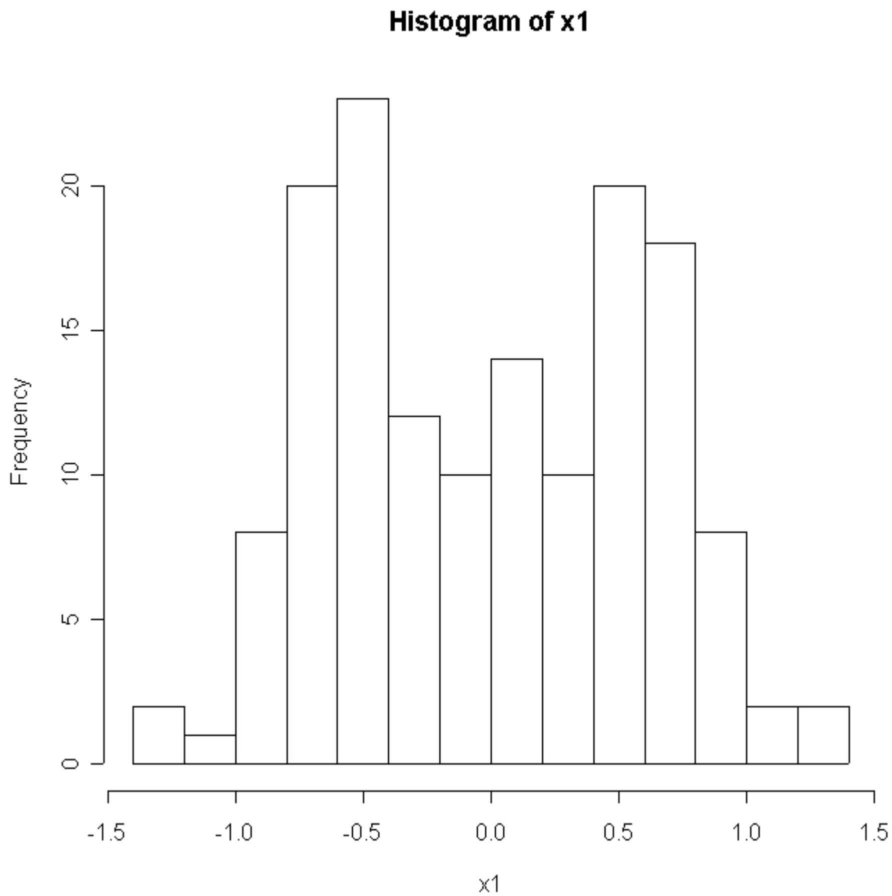


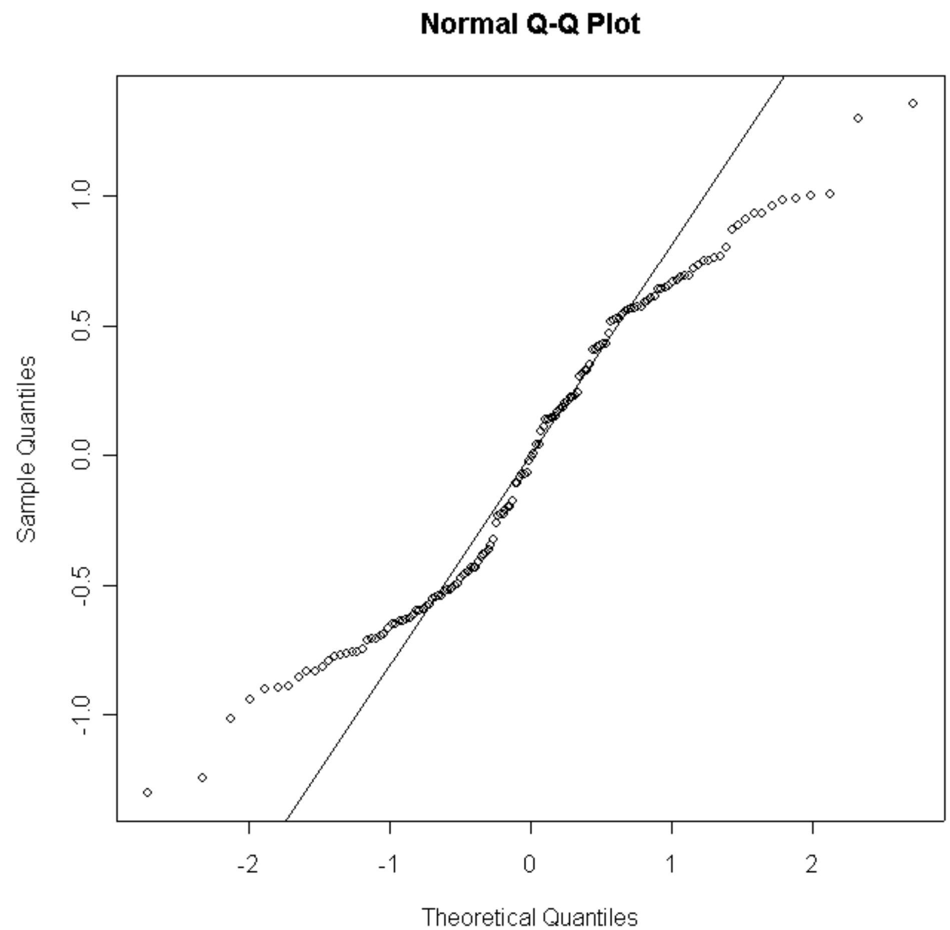
Histogram of x
Frequency
0
50
100
150
200
-10
-5
0
5
10
15
20
x



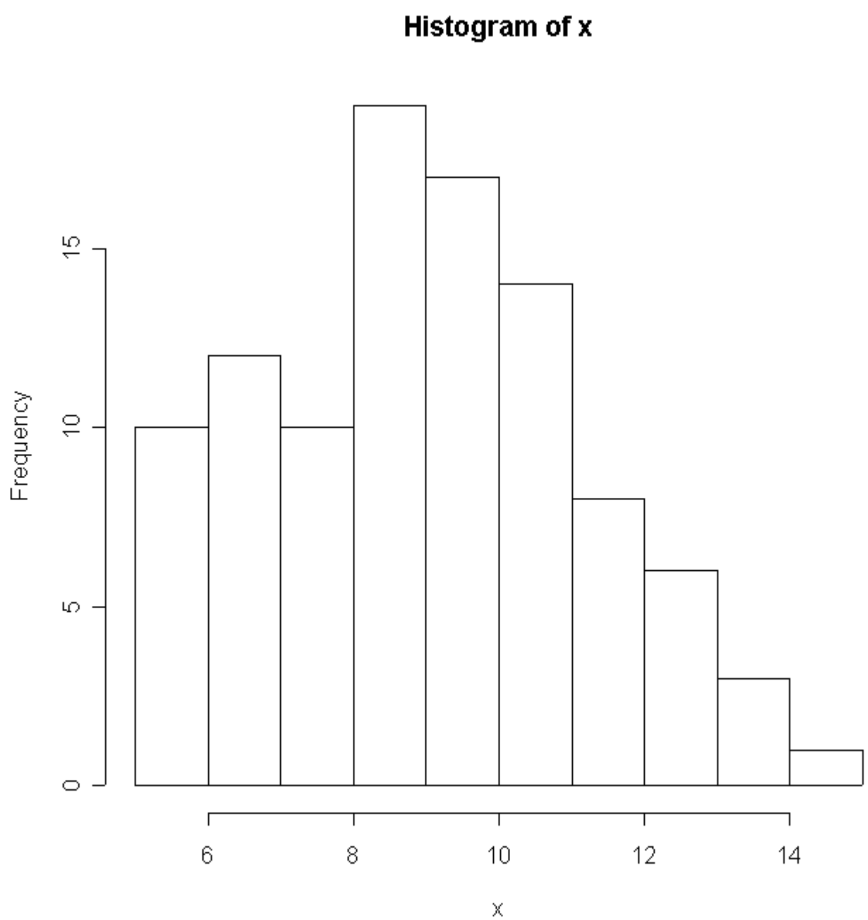
Normal Q-Q Plot
Sample Quantiles
-10
-5
0
5
10
15
20
-2
-1
0
1
2
Theoretical Quantiles

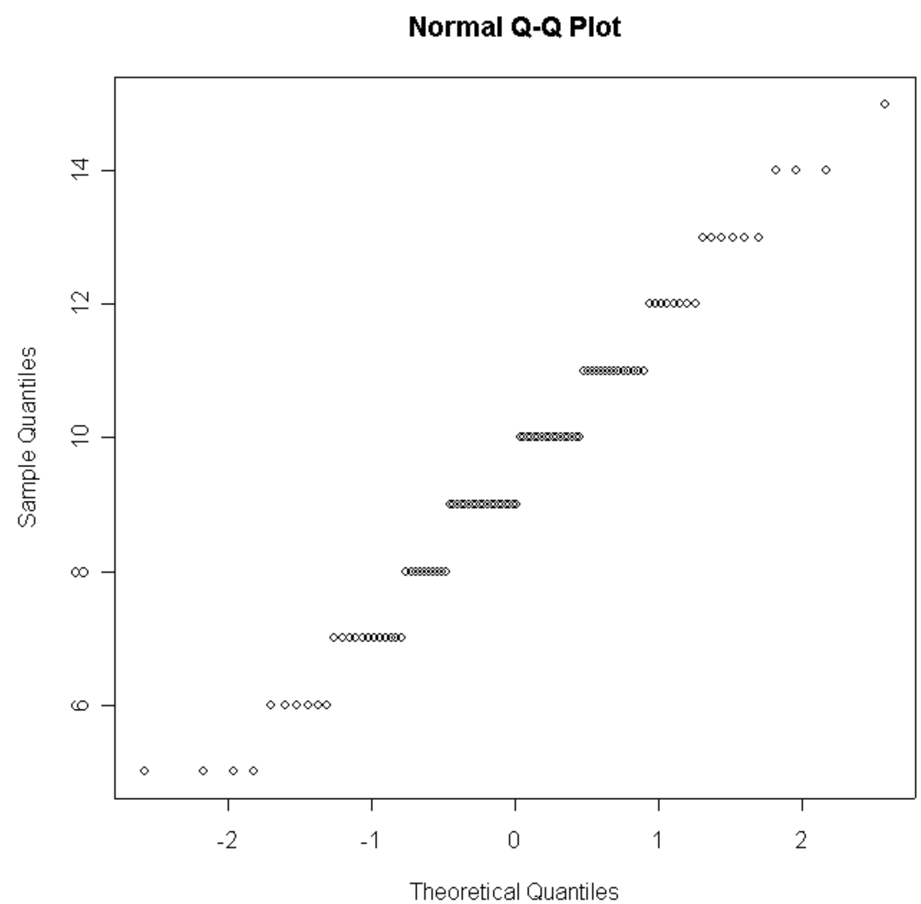
Short Tails


Histogram of x1
Frequency
0
5
10
15
20
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x1

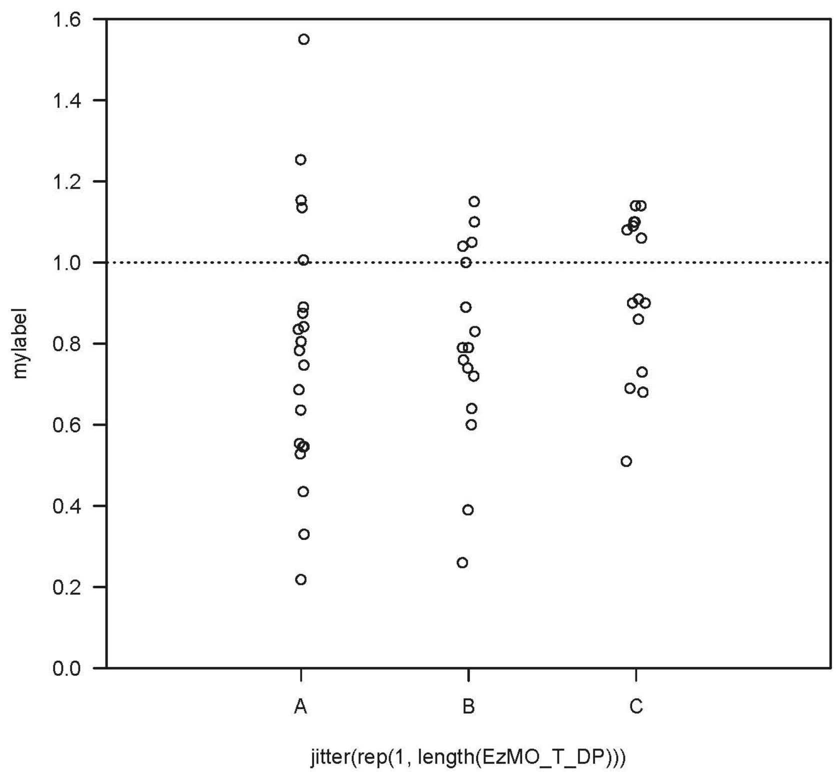

Normal Q-Q Plot
Sample Quantiles
-1.0
-0.5
0.0
0.5
1.0
-2
-1
0
1
2
Theoretical Quantiles

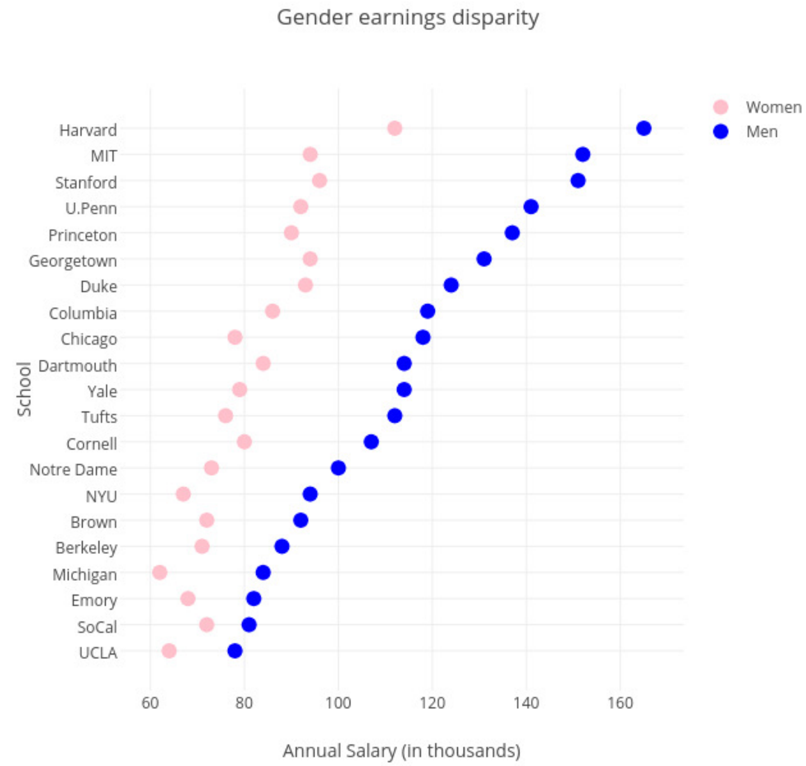
Plateaus/Gaps





Dot plot





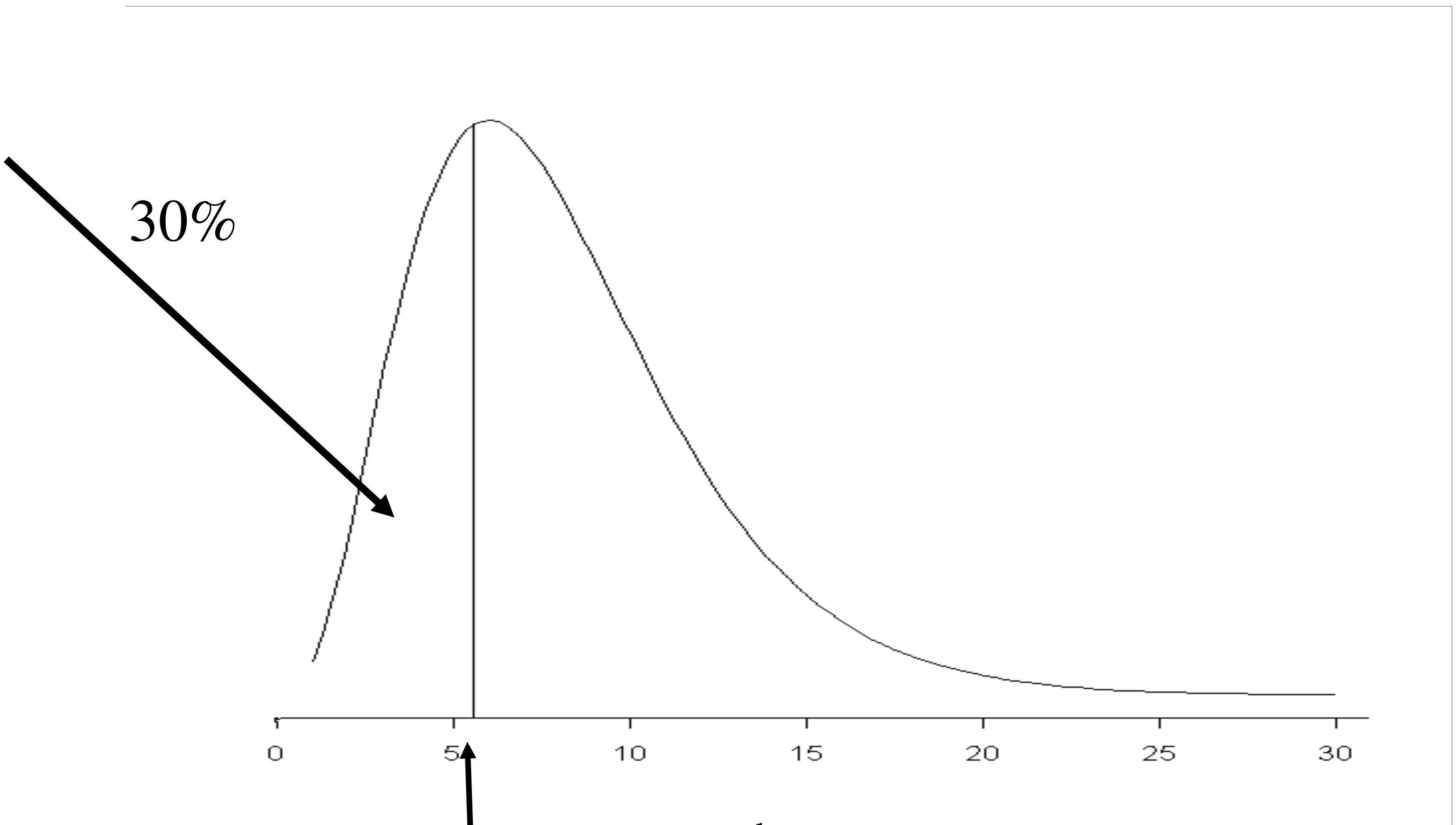
- *Values* plotted separately (as dots) for each group
- Most useful when there *aren't too many* observations

Numerical Summaries

- To provide *objectivity* (put in same objects to same methods, get out same classification
  - This is in contrast to *experts* deciding
- To provide *stability*
  - Would like classification to be ‘robust’ to a wide variety of additions of objects, or characteristics
- Categorical/Qualitative variables
  - frequency table
- Numerical/Quantitative variables
  - Distribution: quantiles
  - Center: mean, median
  - Spread: SD, IQR, MAD

Quantiles

- The p^{th} *quantile* is the number that has the proportion p of the data values smaller than it

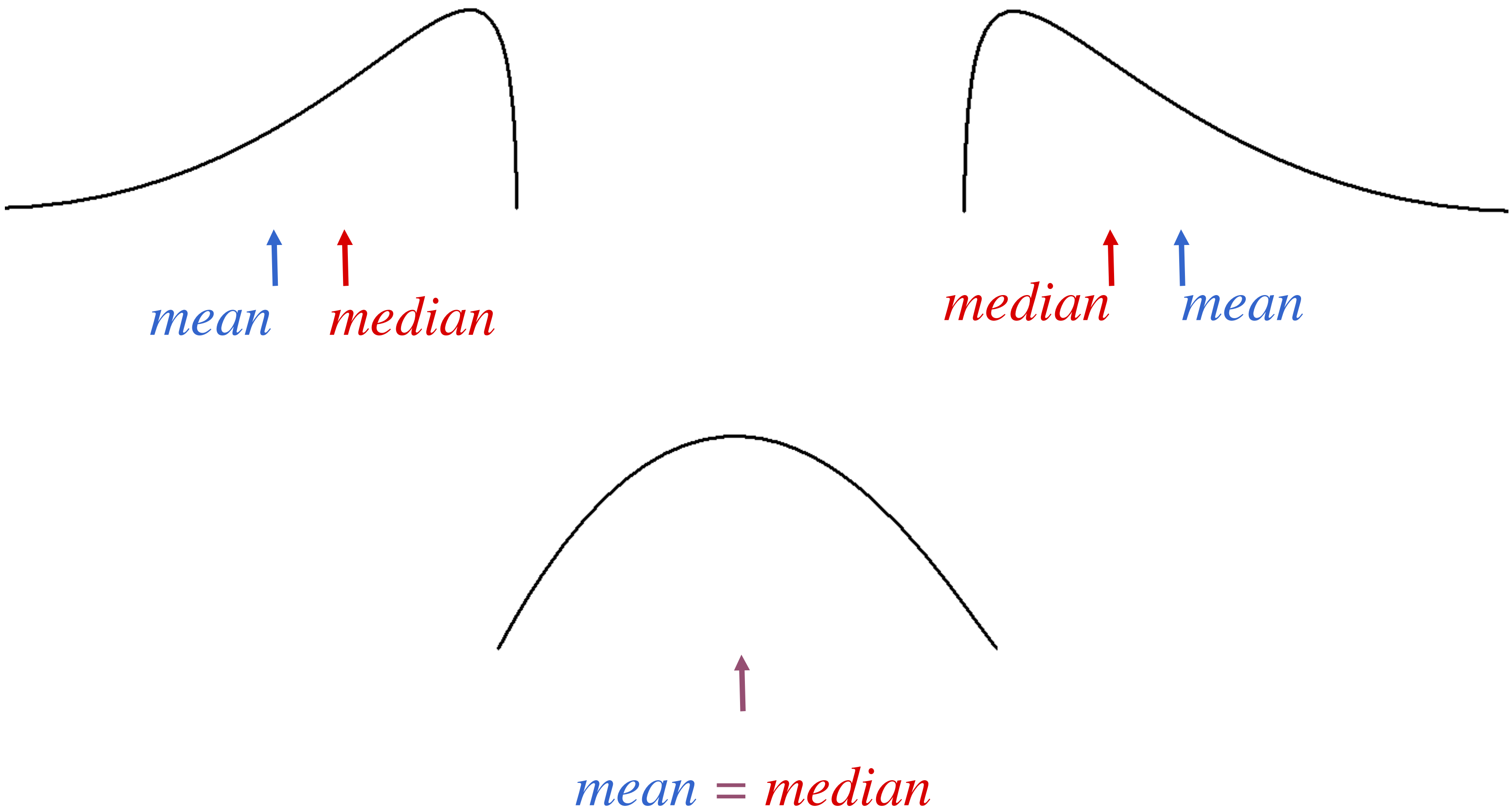


5.53 = 30th percentile

Measures of center

- Mean
  - Total of the values divided by the number of values
  - Appropriate for distributions that are fairly *symmetric*
  - *Sensitive* to outliers (since all values contribute equally)
  - ‘Balance-point’ for a histogram
- Median
  - The *median* value of a variable is the ‘middlemost number: 50% (half) of the values are smaller than it, 50% bigger
  - NOT sensitive to outliers (since it ‘ignores’ most values)
  - Appropriate summary for *skewed distributions*

Relative location of mean and median



Measures of spread

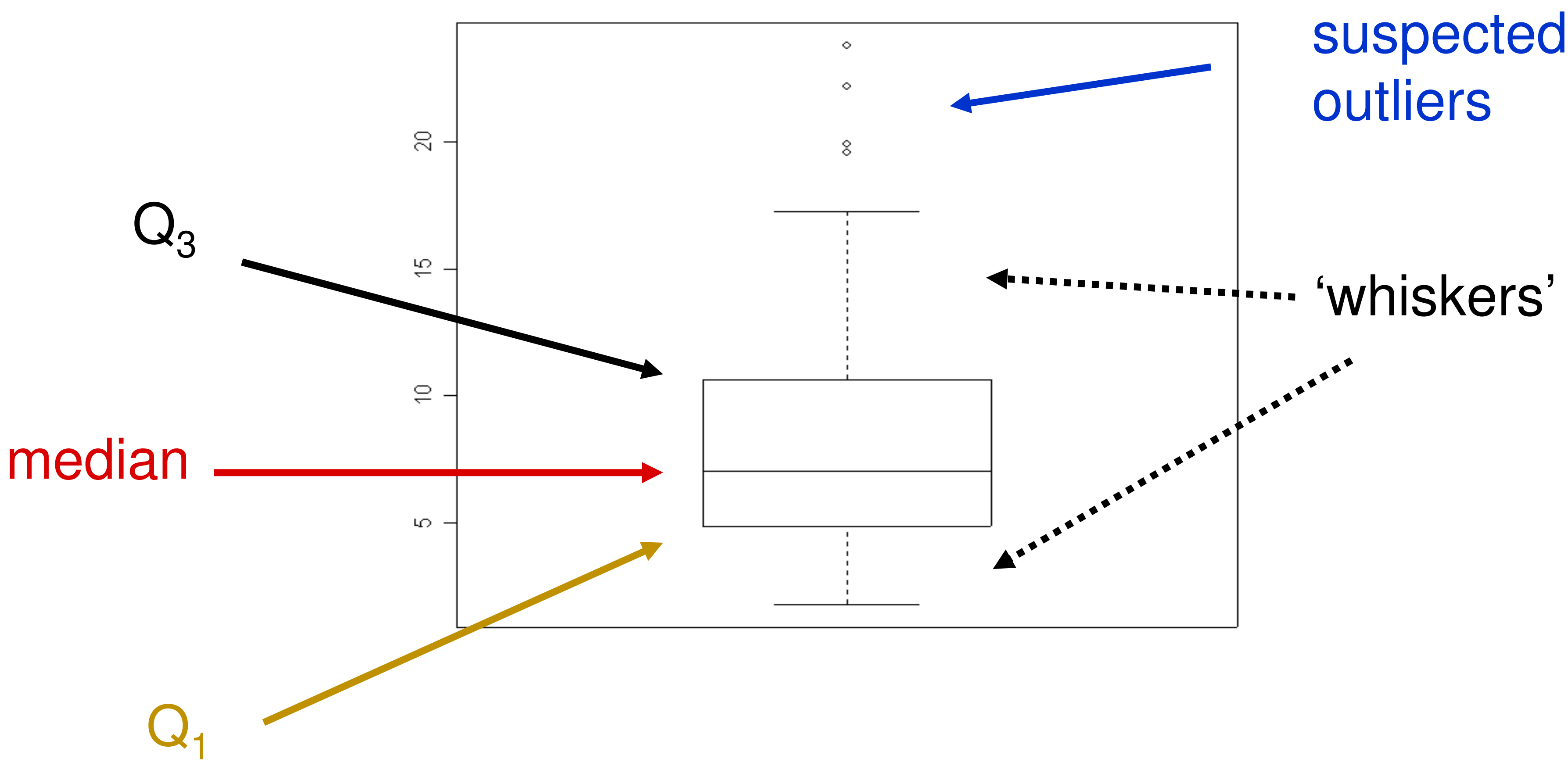
- Standard deviation (SD)
  - Square root of the average* of squared deviations from mean
  - Appropriate when center measured with the *mean*
- Interquartile range (IQR)
  - Distance between 25th (Q_1) and 75th (Q_3) percentiles:

$$IQR = Q_3 - Q_1$$

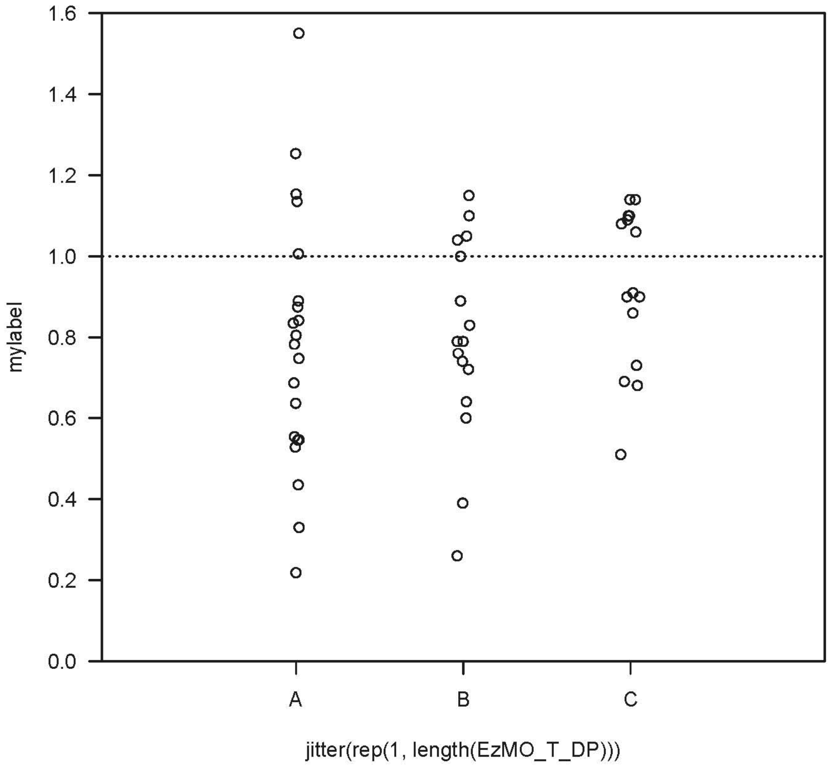
  - One measure of spread when center measured with *median*
- Median Absolute Deviation (MAD)
  - *Median* of *absolute* values of *deviations* from median
  - More *robust* measure of spread than SD
  - Another way (besides IQR) to measure spread when center measured with *median*

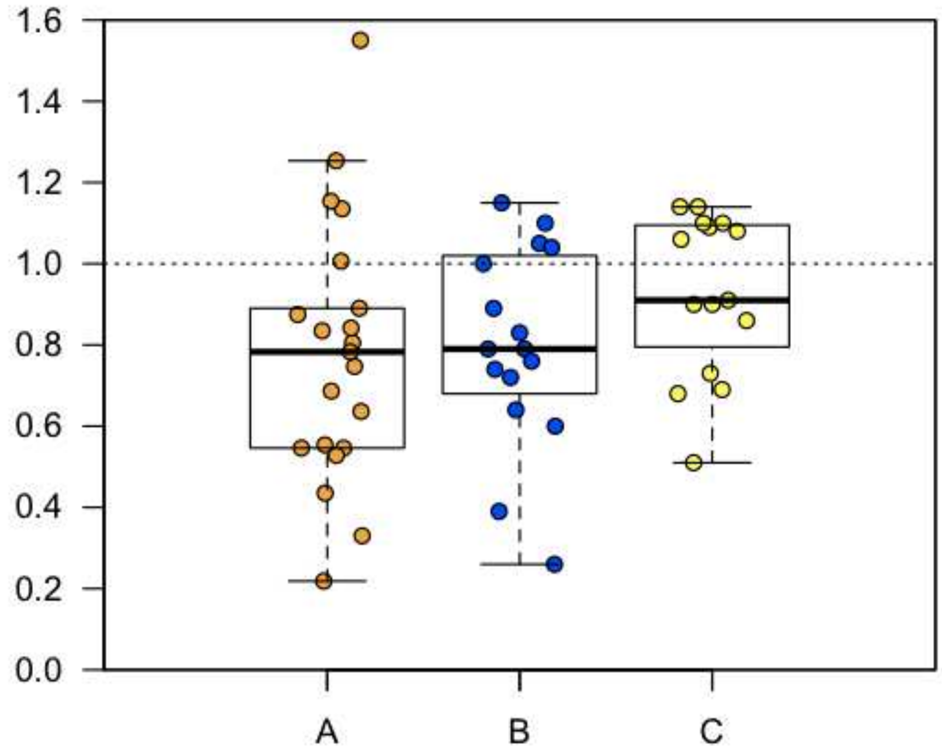
Five-number summary and boxplot

- Overall summary of the distribution: Min, Q_1, Median, Q_3, Max
- A *boxplot* provides a visual summary:



Box plot combined with dot plot





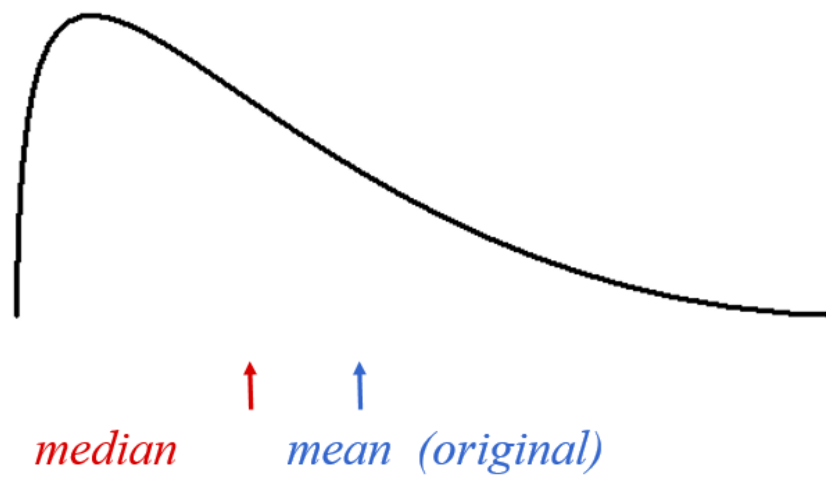
- ‘*jitter*’, *size* and *color* aid in the comparison of groups

Robustness and resistance

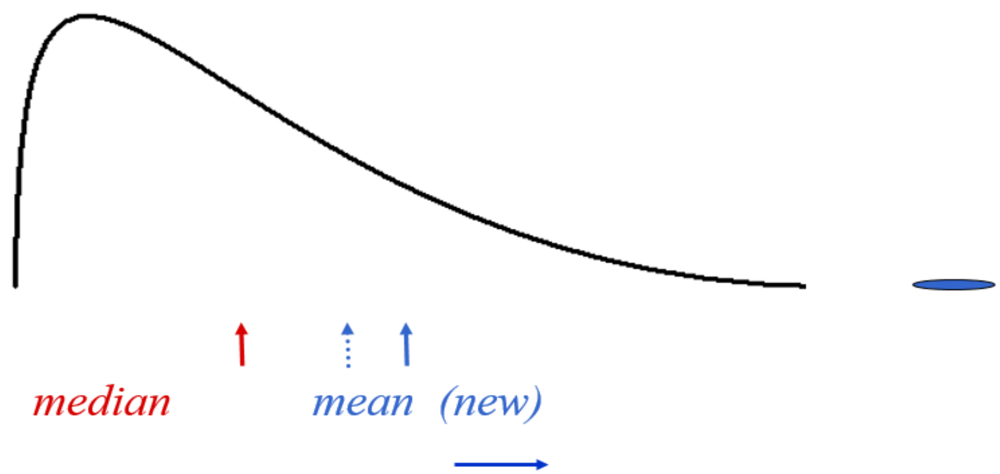
- These concepts refer to *lack of sensitivity* to assumed distributions and effects of a small number of values or outliers
- These qualities are *desirable*: you don’t want inferences to be strongly influenced by only a small part of the data set
- The mean is very sensitive to outlying values, the median is very resistant

Robustness of mean, median

Just us:



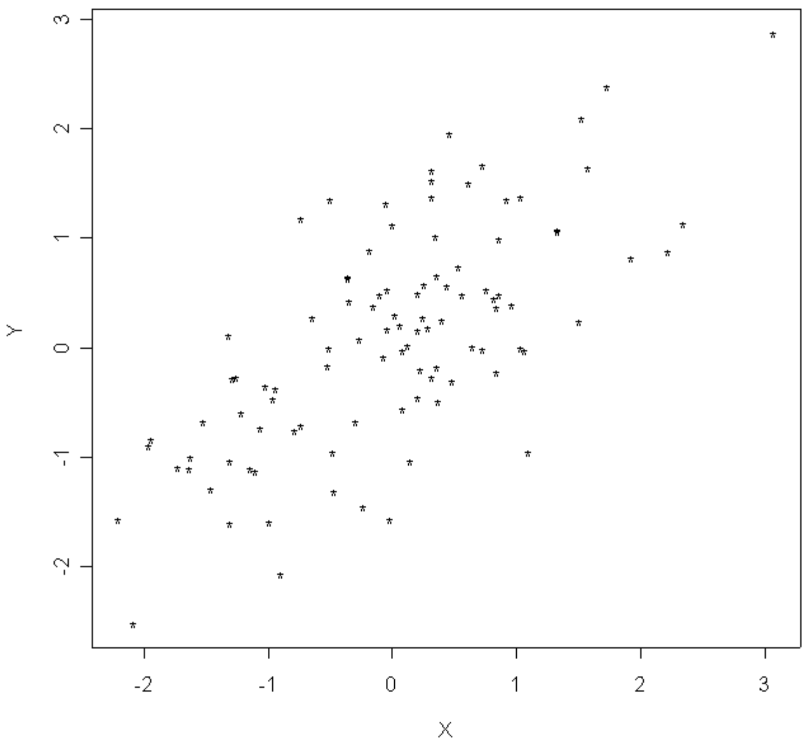
With Mark:

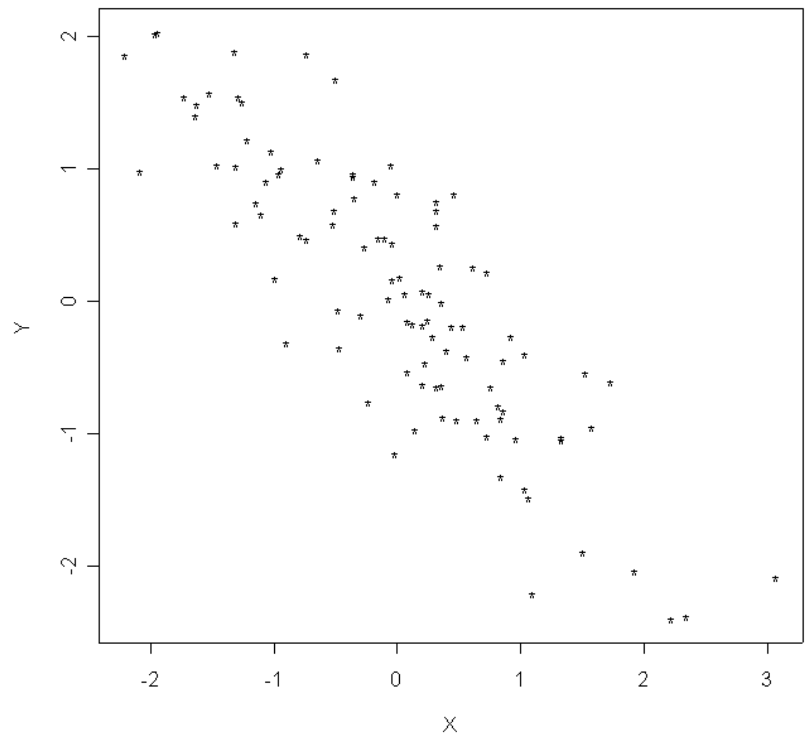


Scatterplot

- We can graphically summarize a bivariate data set with a *scatterplot* (also sometimes called a *scatter diagram*)
- Plots values of one variable on the horizontal axis and values of the other on the vertical axis
- Can be used to see how values of 2 variables tend to move with each other (*i.e.* how the variables are *associated*)

Scatterplots

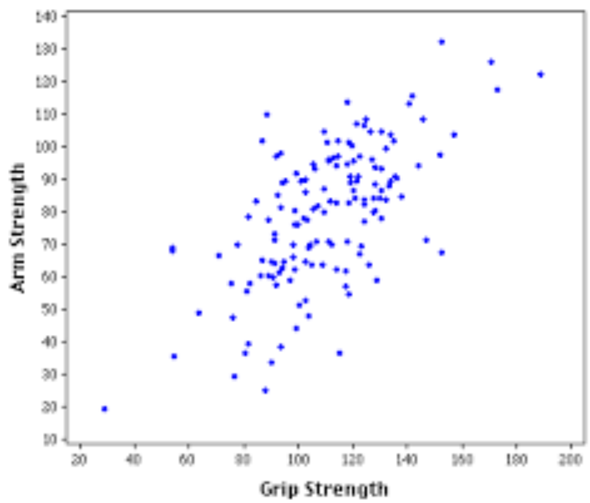


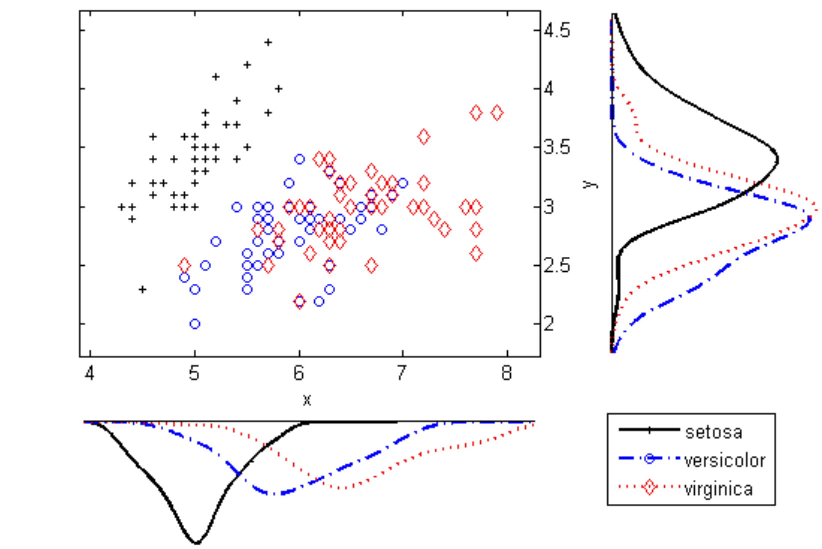


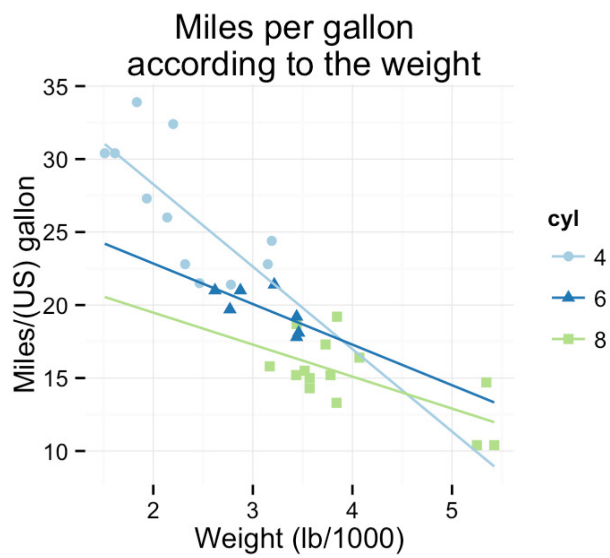
positive association

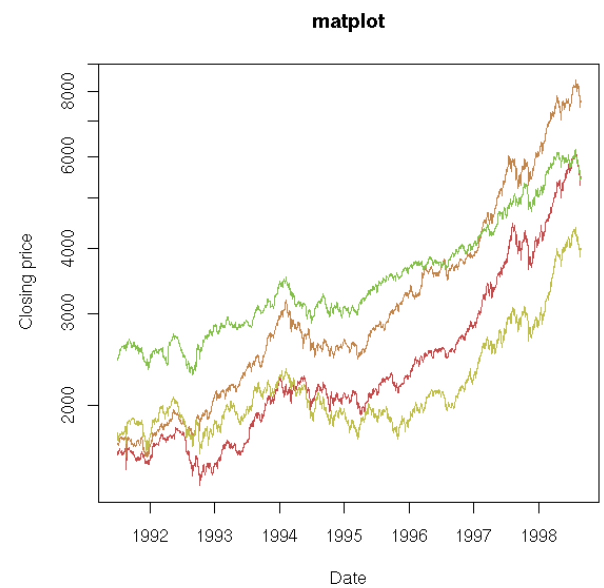
negative association

Scatterplots: customized

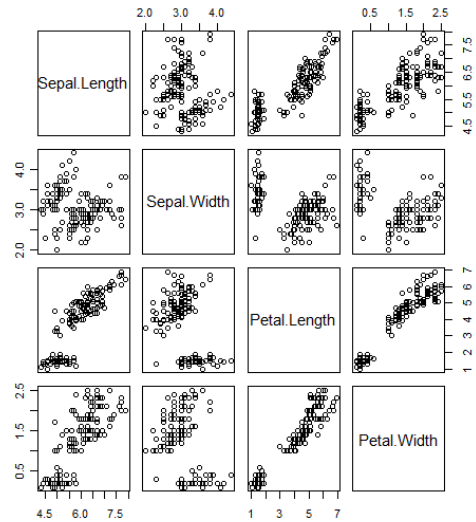


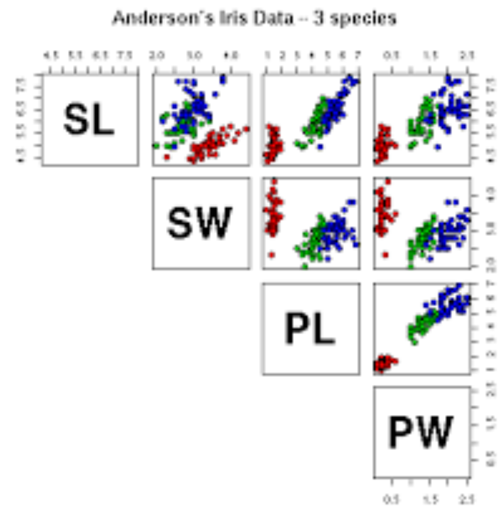


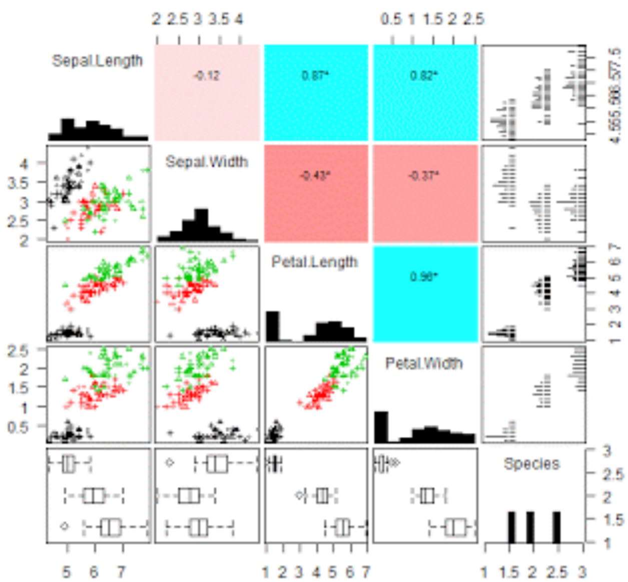


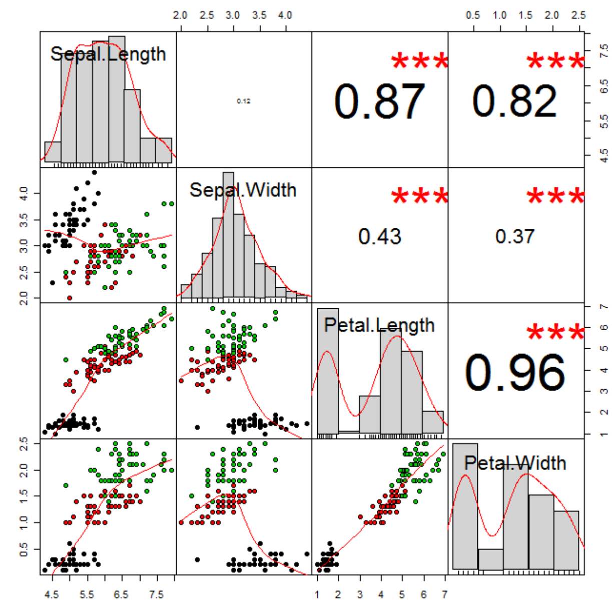


All pairwise plots: pairs / splom









Numerical Summary

- Typically, a bivariate data set is summarized numerically with 5 *summary statistics*
- These provide a fair summary for scatterplots with the same general shape as we just saw, like an oval or an ellipse
- We can summarize each variable *separately* : *X* mean, *X* SD; *Y* mean, *Y* SD
- But these numbers don't tell us how the values of *X* and *Y* vary together

Correlation Coefficient

- The (sample) *correlation coefficient* ***r*** is defined as the average value of the product

  (X in SUs)*(Y in SUs)

- [SU = standard units = (value-mean)/SD]
- *r* is a *unitless* quantity
- $-1 \leq r \leq 1$
- *r* is a measure of *LINEAR ASSOCIATION*

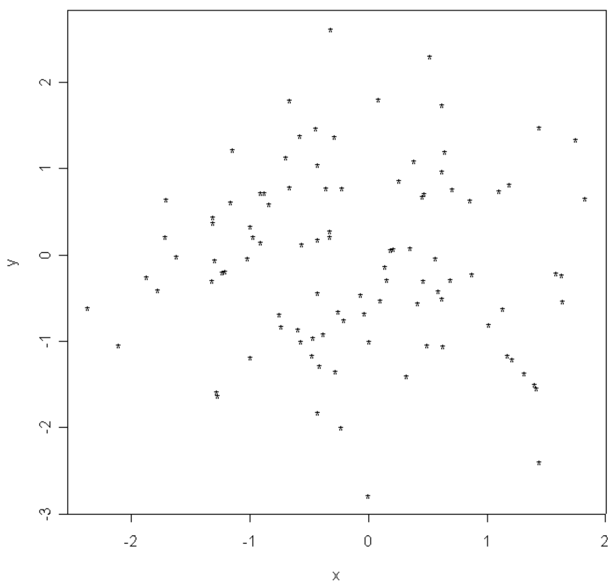
What r is...

- r is a measure of *LINEAR ASSOCIATION*
- The closer r is to -1 or 1, the more tightly the points on the scatterplot are clustered around a line
- The sign of r (+ or -) is the same as the sign of the slope of the line
- When $r = 0$, the points are not *LINEARLY ASSOCIATED* – this does *NOT* mean there is *NO ASSOCIATION*

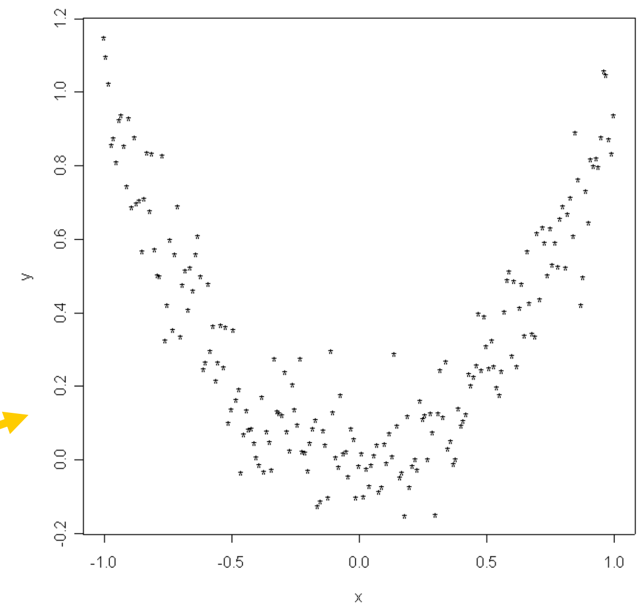
...and what *r* is *NOT*

- *r is* a measure of *LINEAR ASSOCIATION*
- *r* does *NOT* tell us if *Y* is a function of *X*
- *r* does *NOT* tell us if *X causes Y*
- *r* does *NOT* tell us if *Y causes X*
- *r* does *NOT* tell us what the scatterplot looks like

$r \approx 0$

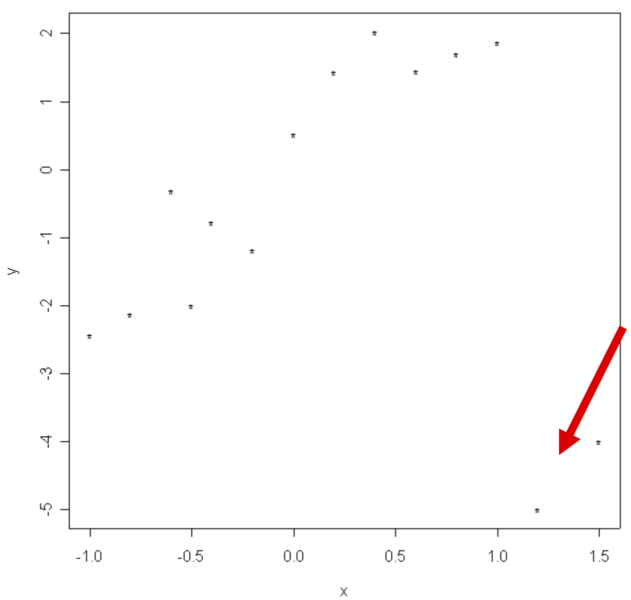


random scatter

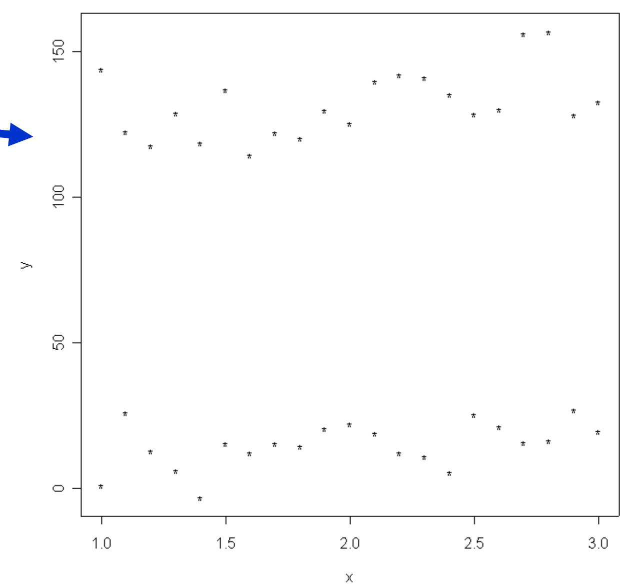


curved pattern

parallel lines



outliers



Categorical data

- So far, we have been looking at *continuous* response variables
- Sometimes, the response is *categorical*
  - male/female
  - yes/no
- In this case, we are often interested in questions dealing with *proportions* (rather than means)

Two-way tables

- Table below is from a blind 5 year randomized study of physicians testing whether regular aspirin use reduces mortality from cardiovascular disease
- Every other day, participants took an aspirin or a placebo

	MI		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Table layout

- Tables often better than words to convey quantitative data
- Avoid too many decimal places
- Usually better to use *space* to separate columns (rather than lines):

Subject	Time 1	Time 2
Joe	3.67390	2.79495
Mary	4.75435	1.23578
Nancy	3.96456	2.84379

Subject	Time 1	Time 2
Joe	3.67	2.79
Mary	4.75	1.24
Nancy	3.96	2.84

Presenting results

- *Communicating results* is an important part of science
- There is no magic ‘formula’ for how to present results!
- You need to think carefully about the message you wish to give and how to present it *clearly* and *convincingly*
- Avoid excessive computer output

Edward Tufte on graphics

- ‘Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency’; should
  - show the data
  - make the reader think about substance
  - avoid data distortion
  - present many numbers in a small space
  - encourage the eye to make comparisons
  - reveal several levels of detail
  - serve a clear purpose
- See also work by Karl Broman

Graphical display tips

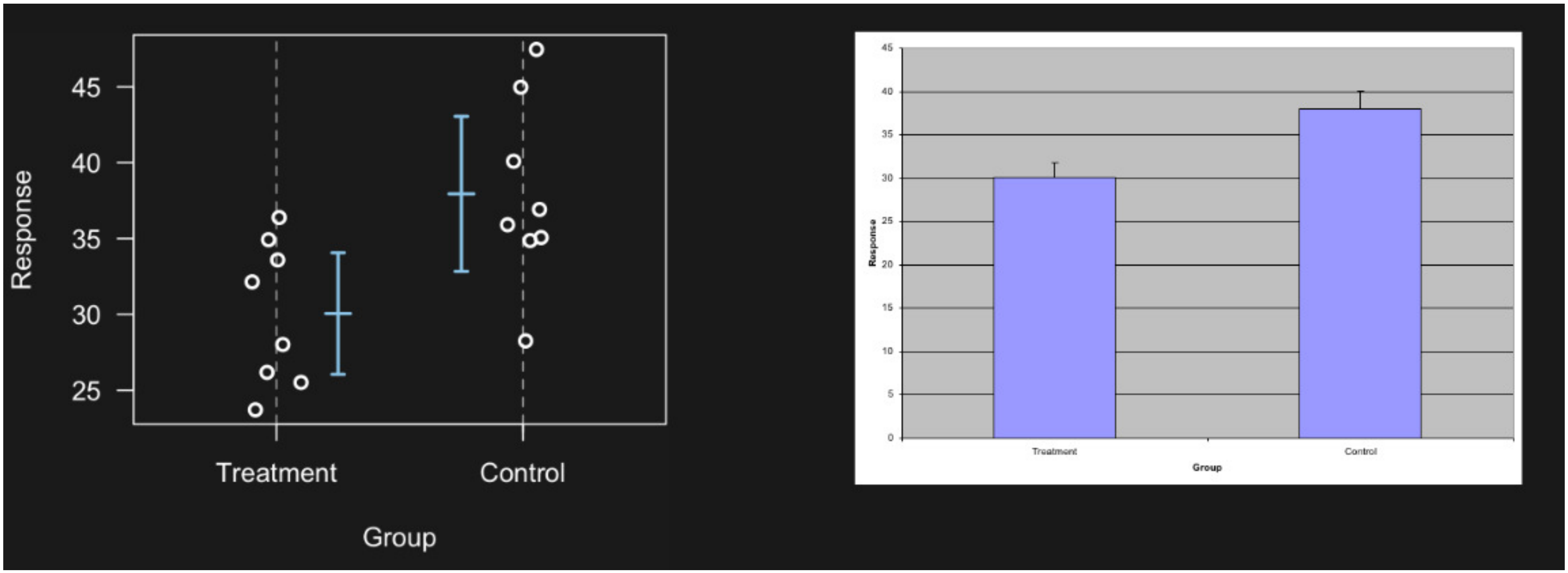
- Show the data **(!!)**
- **Don’t use pie charts**
- Consider logs
- Take differences
- Ease comparisons
  - Things to be compared should be adjacent
  - Align vertically
  - Common axes
  - Labels not legends (where possible)
  - Should sorting really be alphabetical?
  - Consider whether the 0 is needed

More graphical display tips

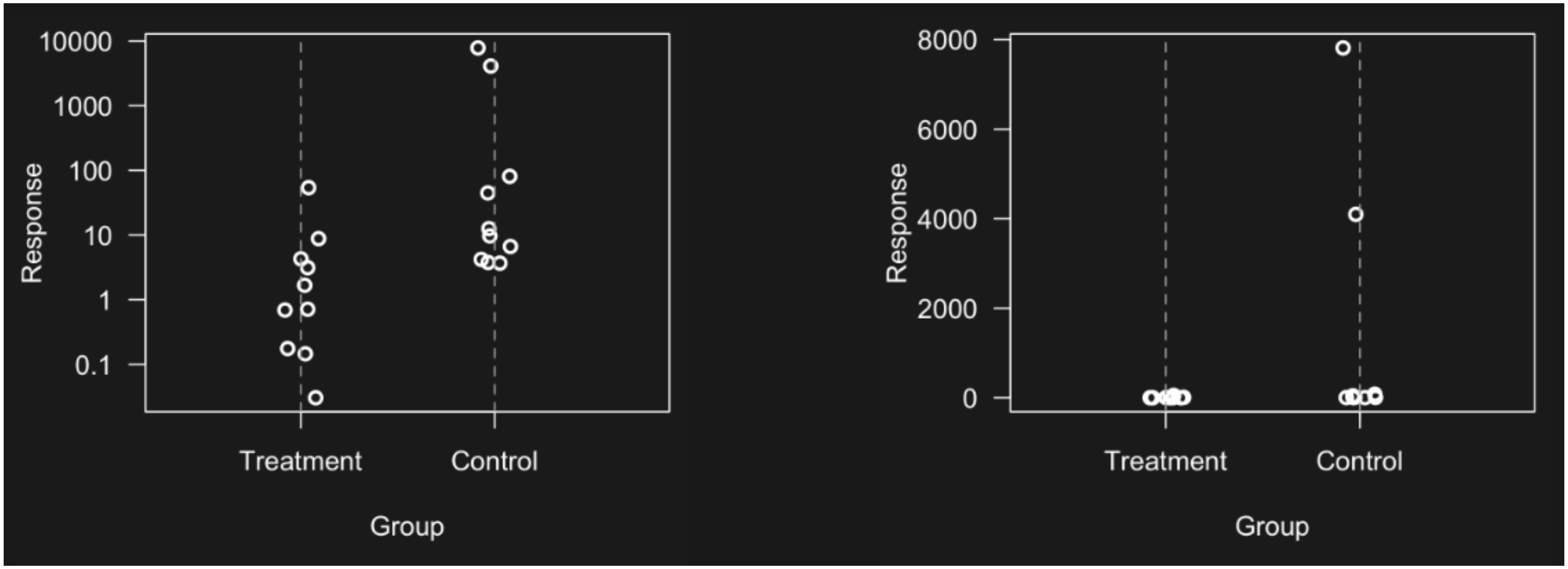
- Data density – for example, number of data points per square centimeter
- Avoid ‘chartjunk’ – decoration that provides no data
- Use color to convey information
- Use appropriate dimensionality
- Did I say **Don’t use pie charts** ?? ☺

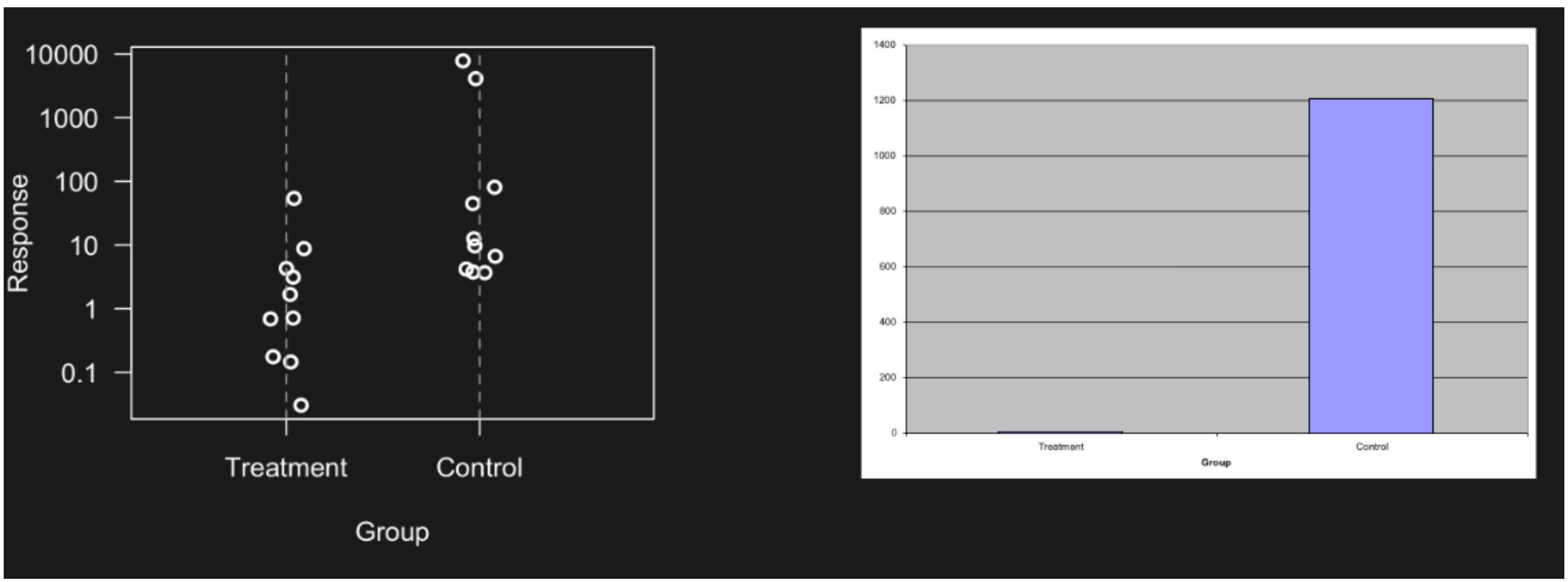
- And now: a *graphics tour* for discussion ...

Show the data

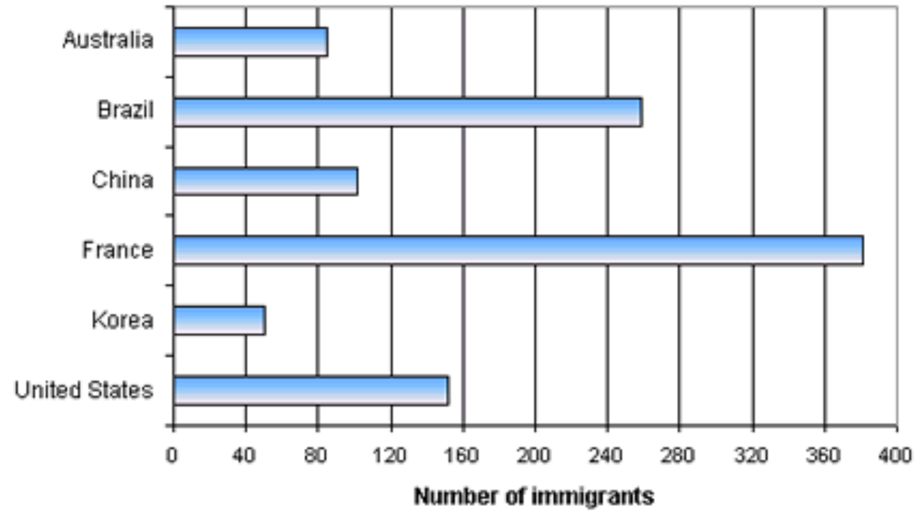


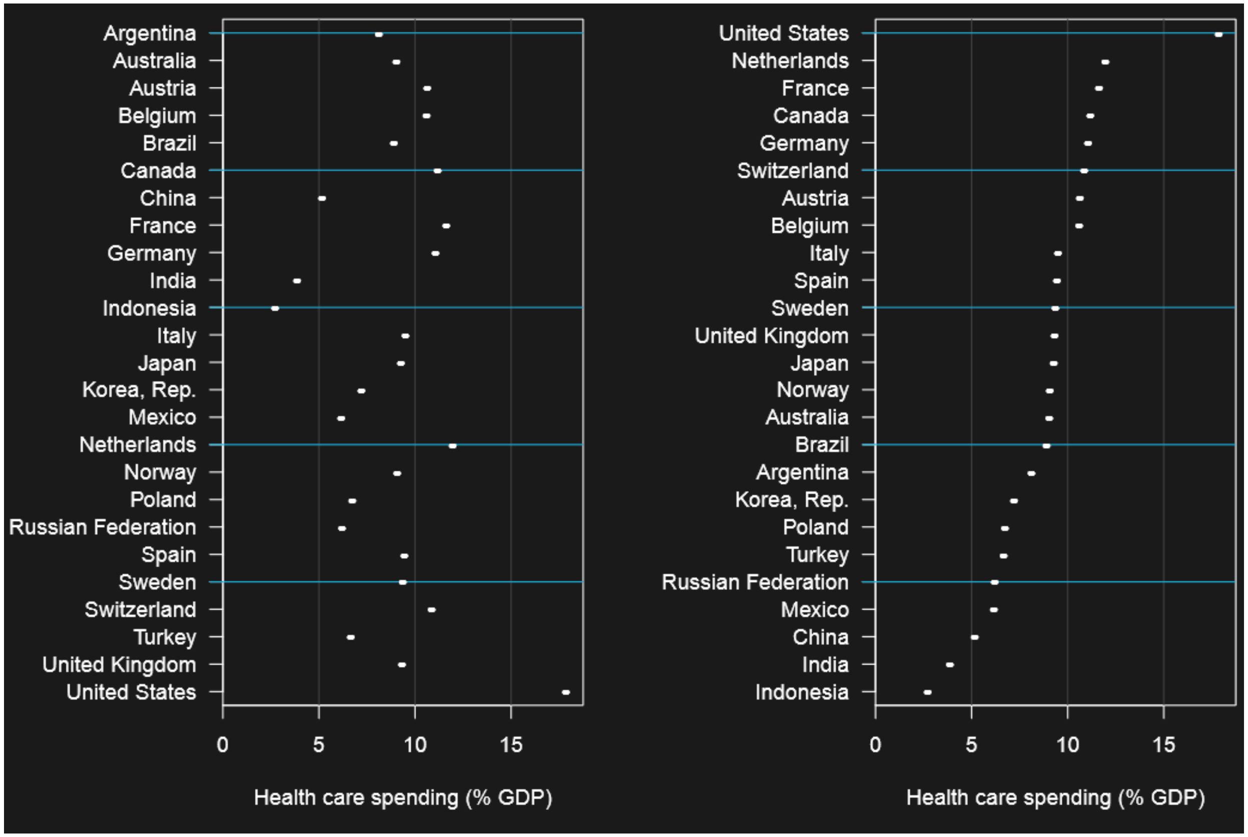
Consider logs



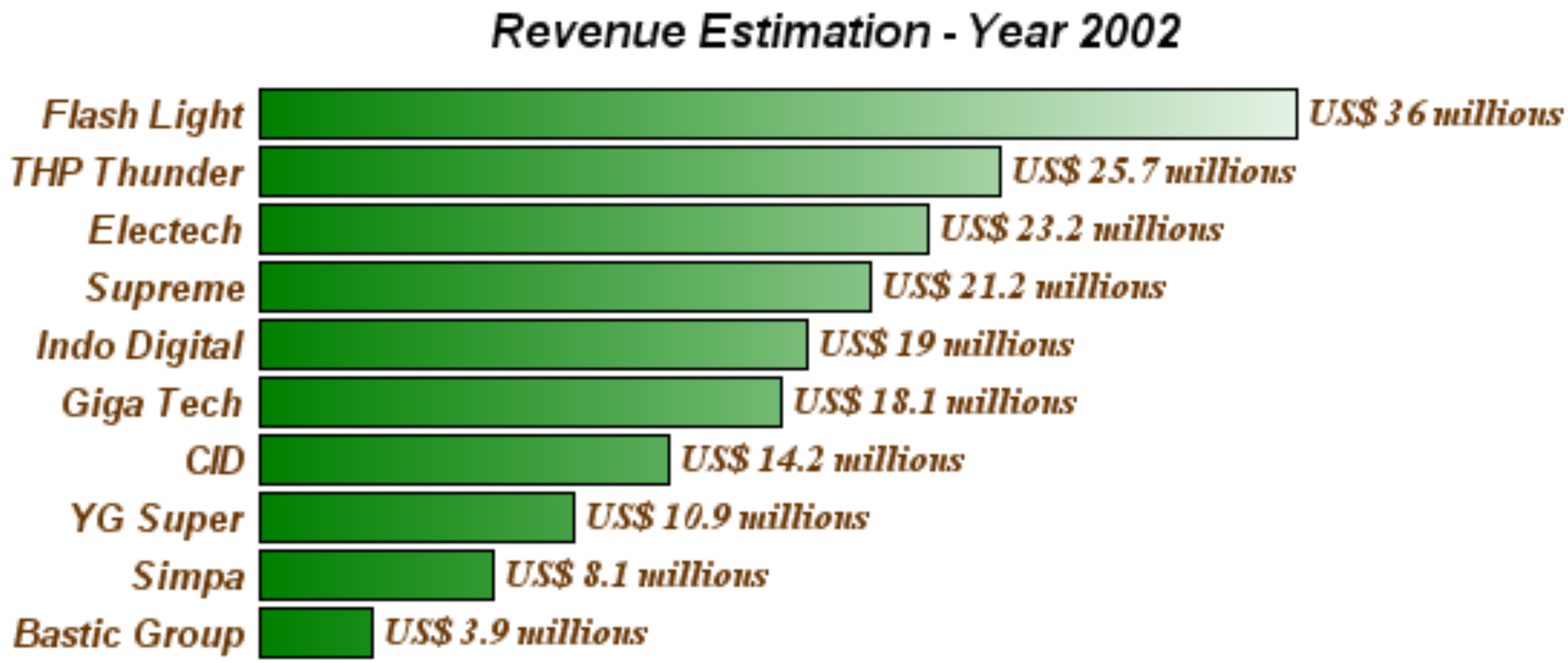


Alphabetical?

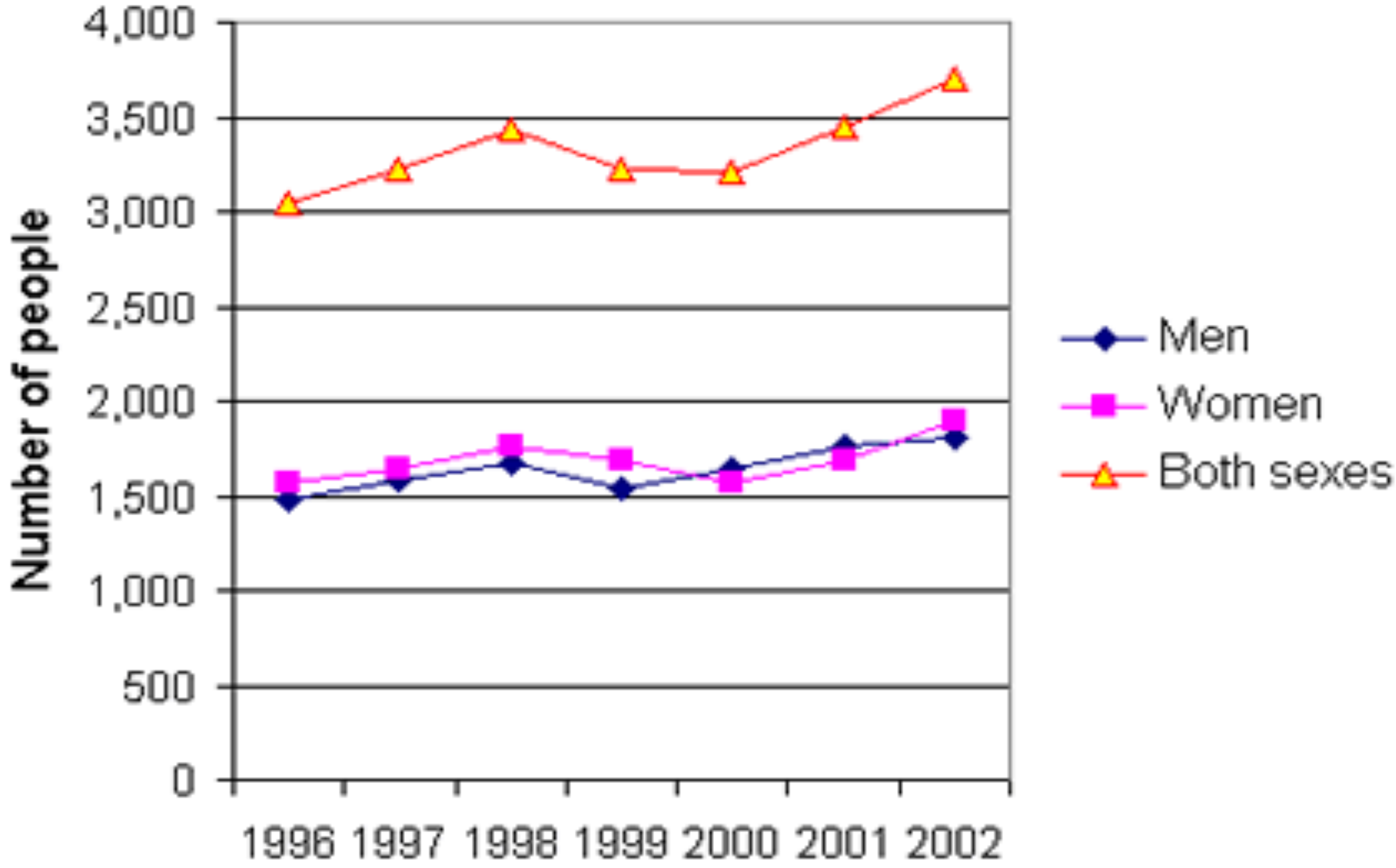




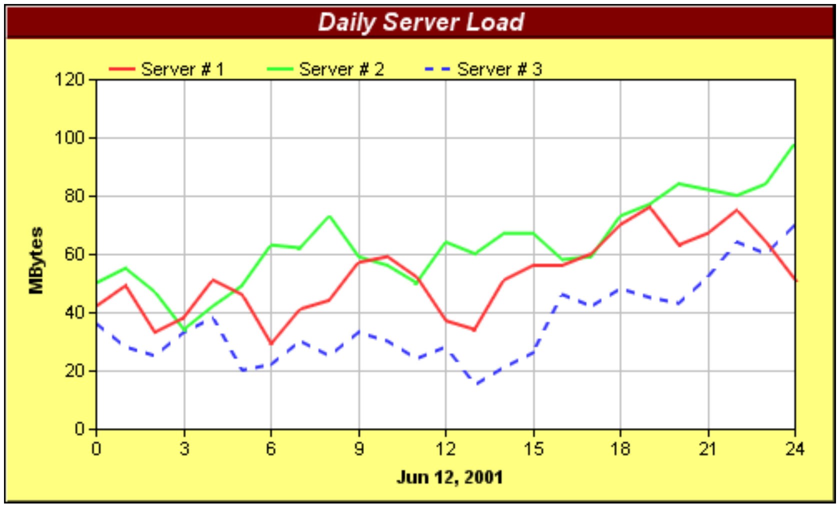
Do we really need color here?

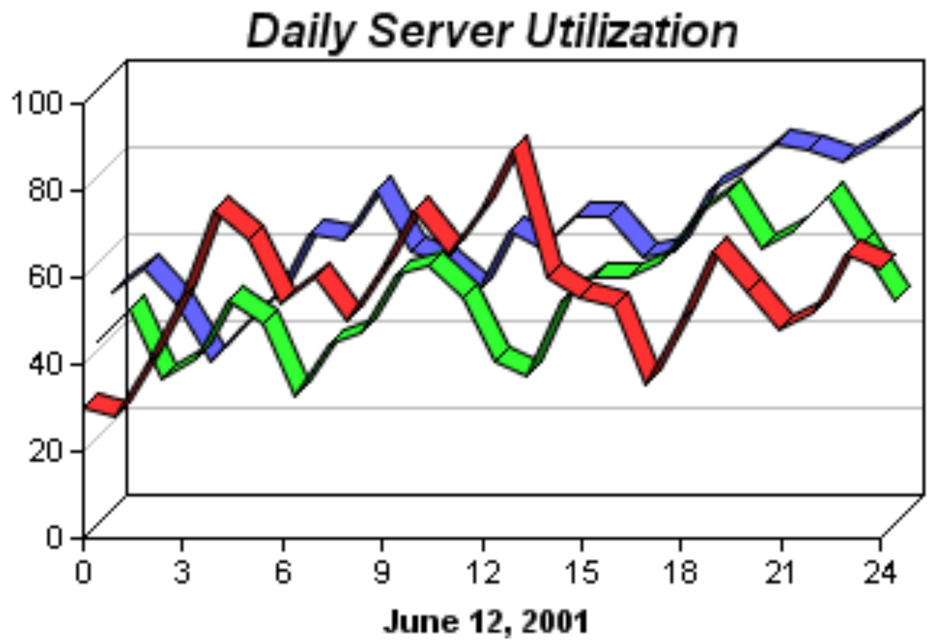


3 lines?


4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Number of people
1996 1997 1998 1999 2000 2001 2002
Men
Women
Both sexes

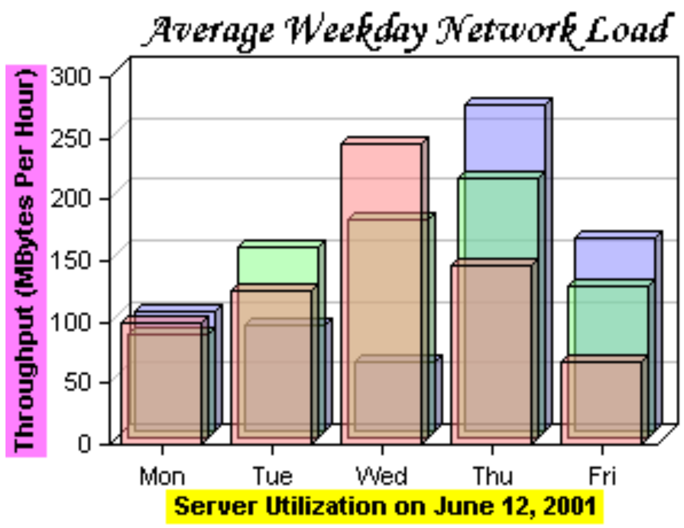
More about lines

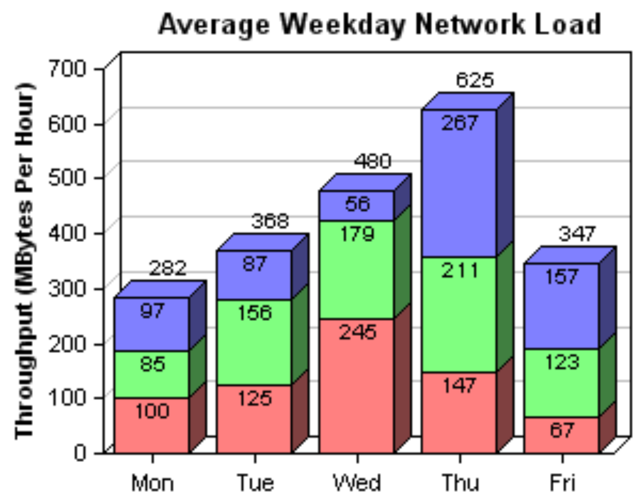


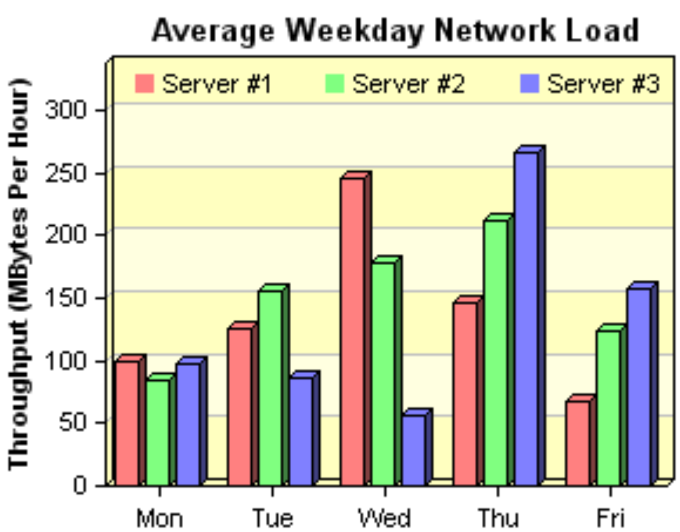


- Different types (solid, dotted)?
- Colors?
- 3D??

What the *^*$%# are these saying?





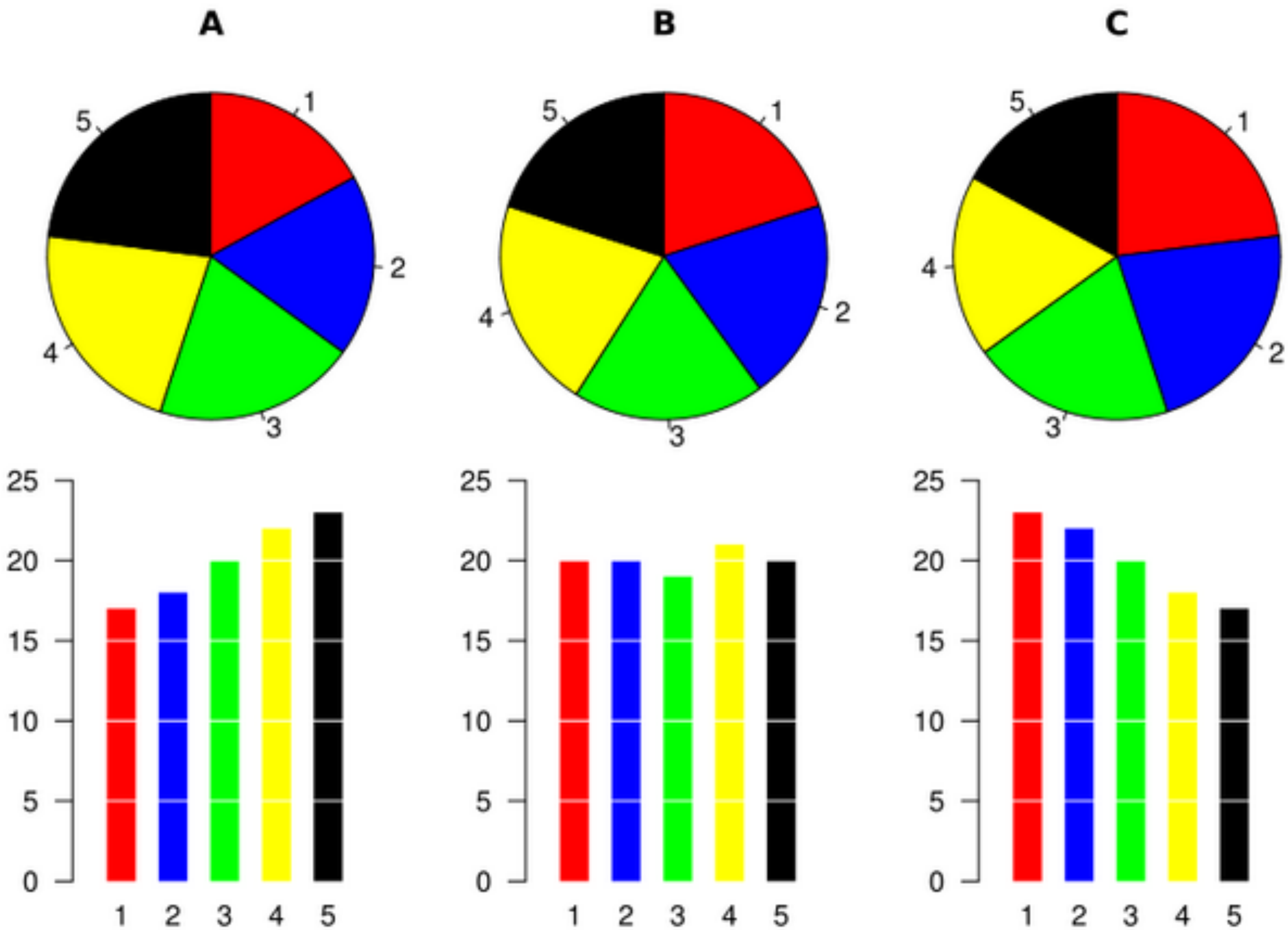


What improvements might be made?

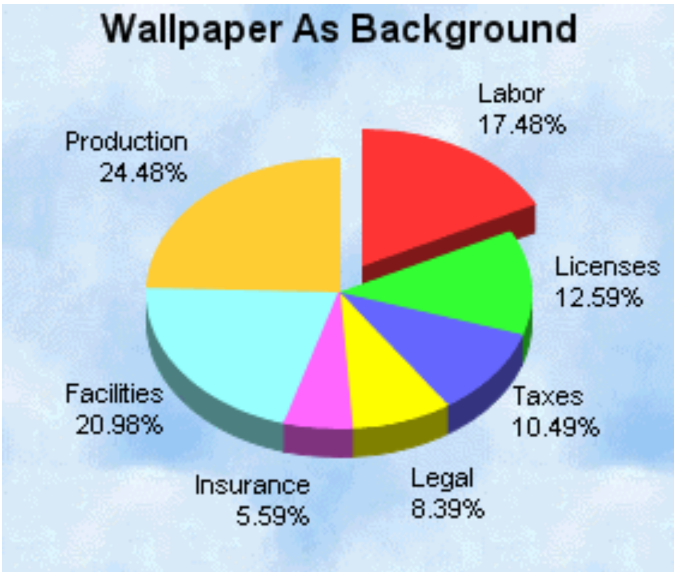
Pie Charts: JUST SAY NO !!!

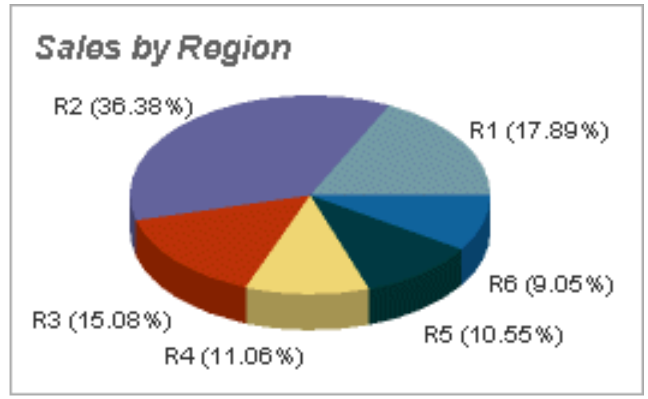
- Pie charts are a ***bad way*** to display information
- The eye is
  - **good** at judging *linear measures* and
  - **bad** at judging *relative areas, volumes or angles*
- A pie chart is *never necessary* - data that can be shown by pie charts *always* can be shown by a dot plot (or bar chart, or table)
- 3D version even worse!

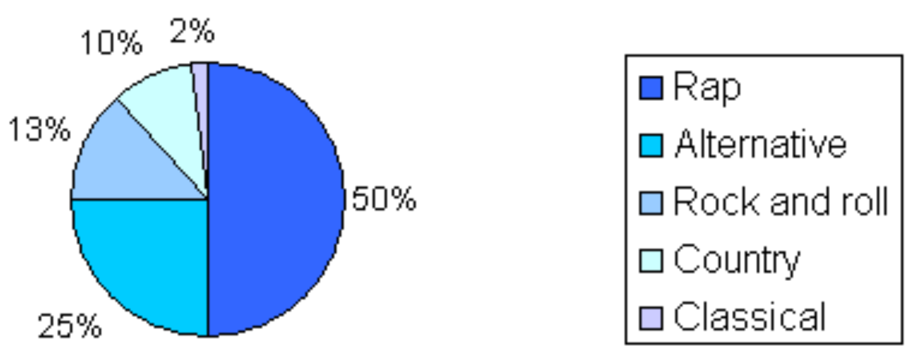
Spot the differences: pie vs. bar

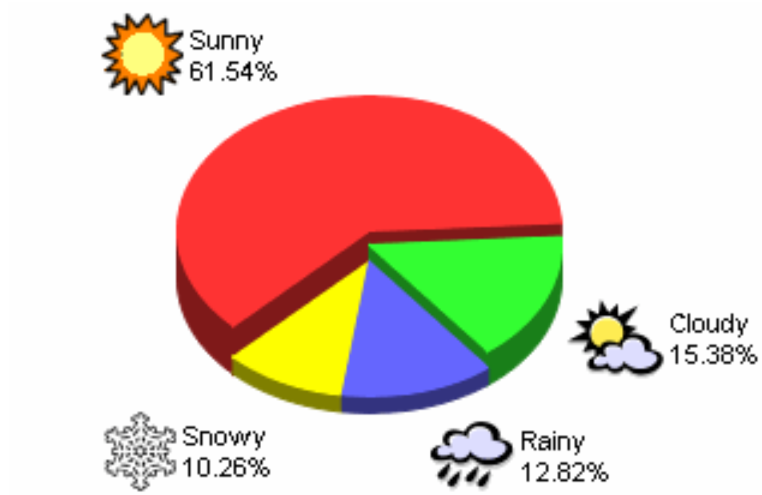


Even worse examples of pie charts

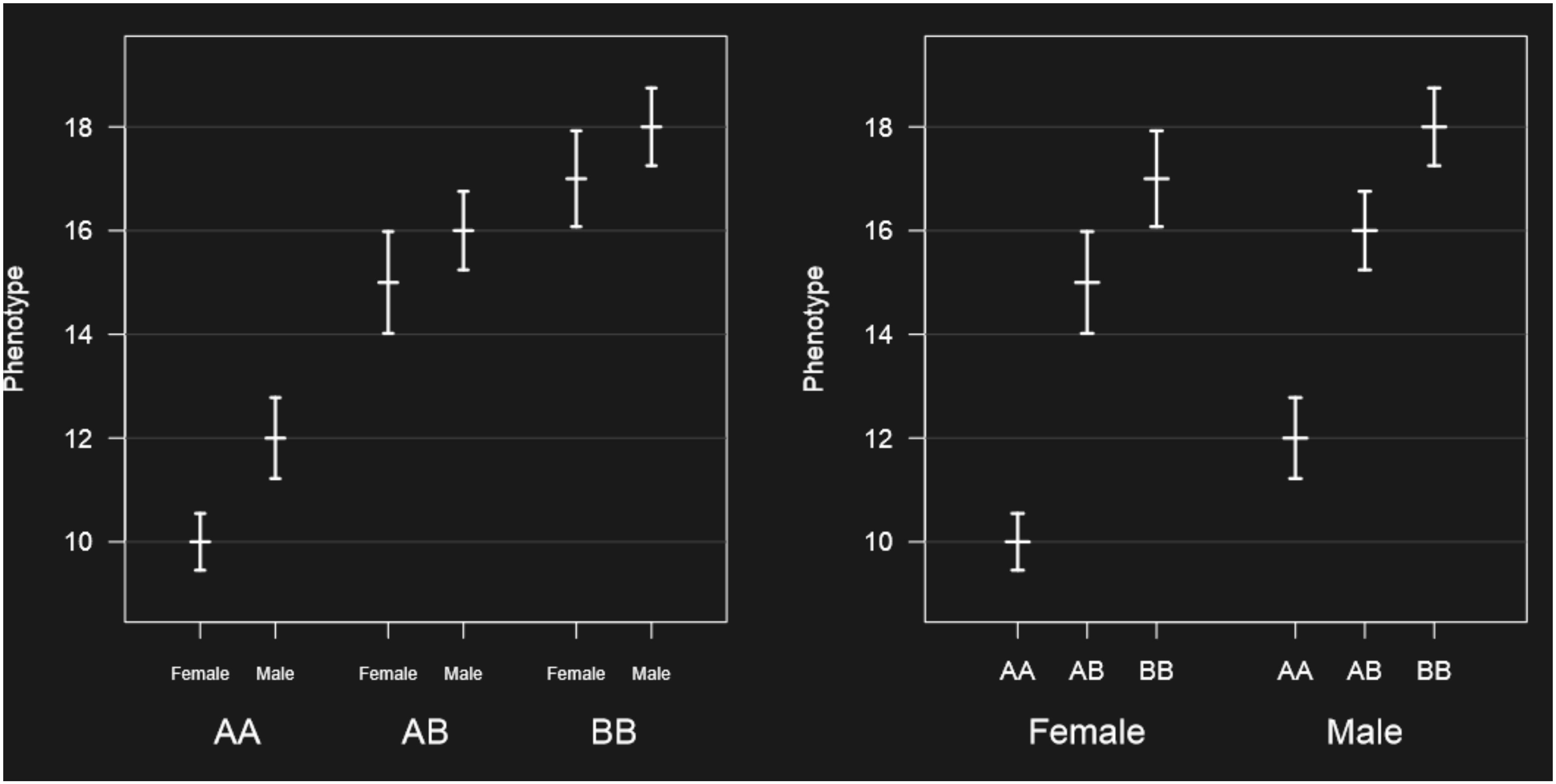




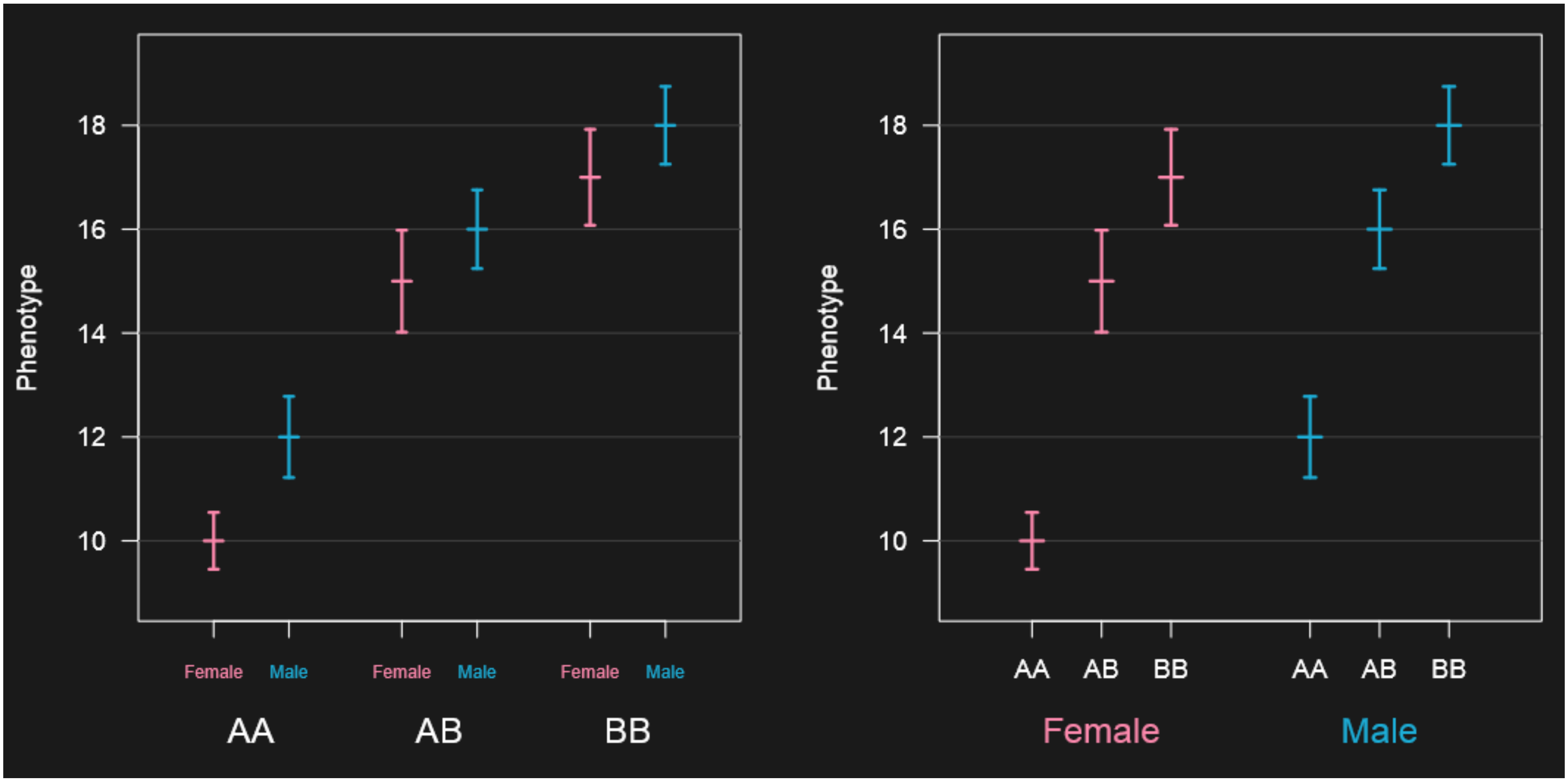




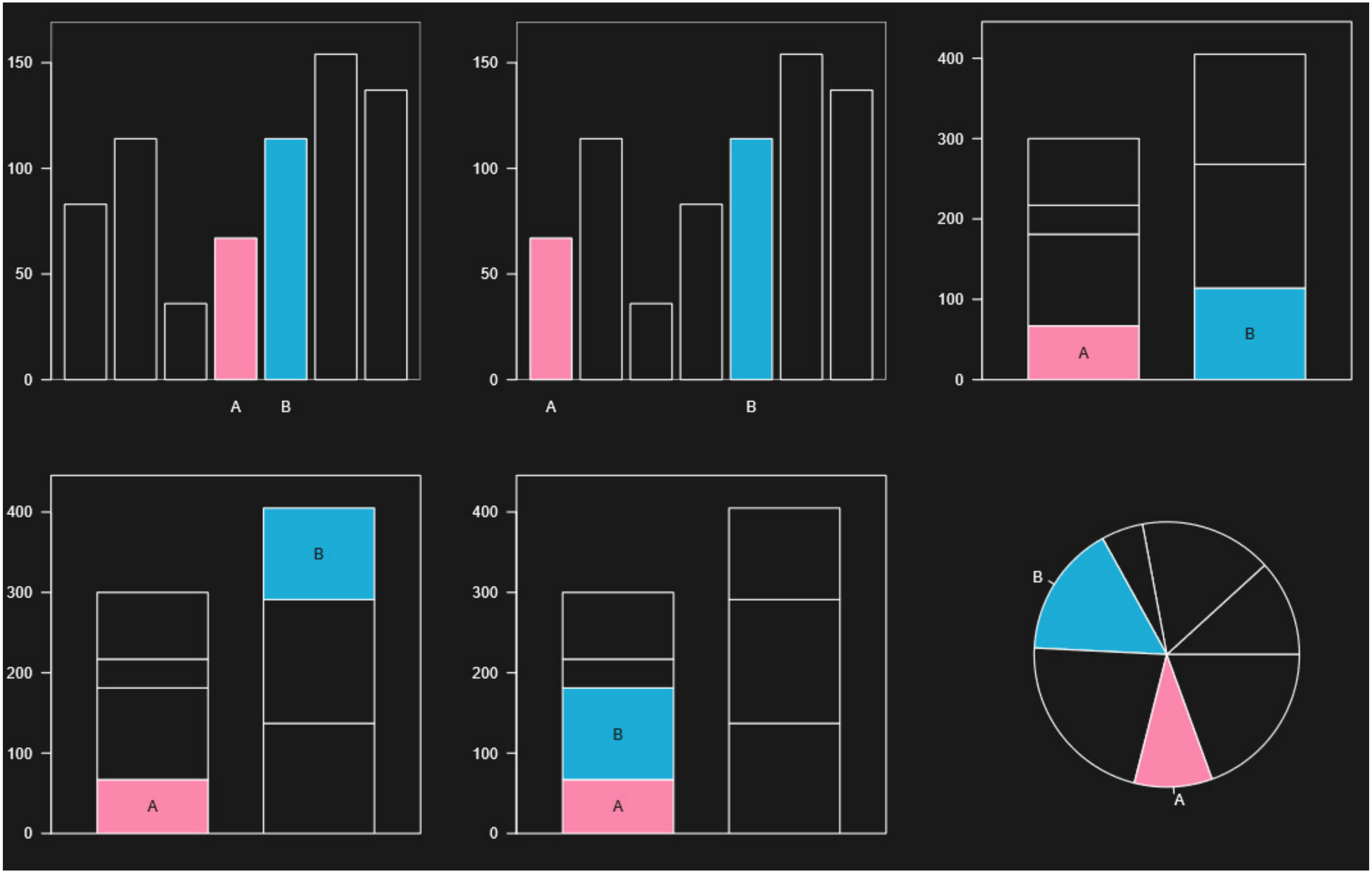
Things to be compared: adjacent



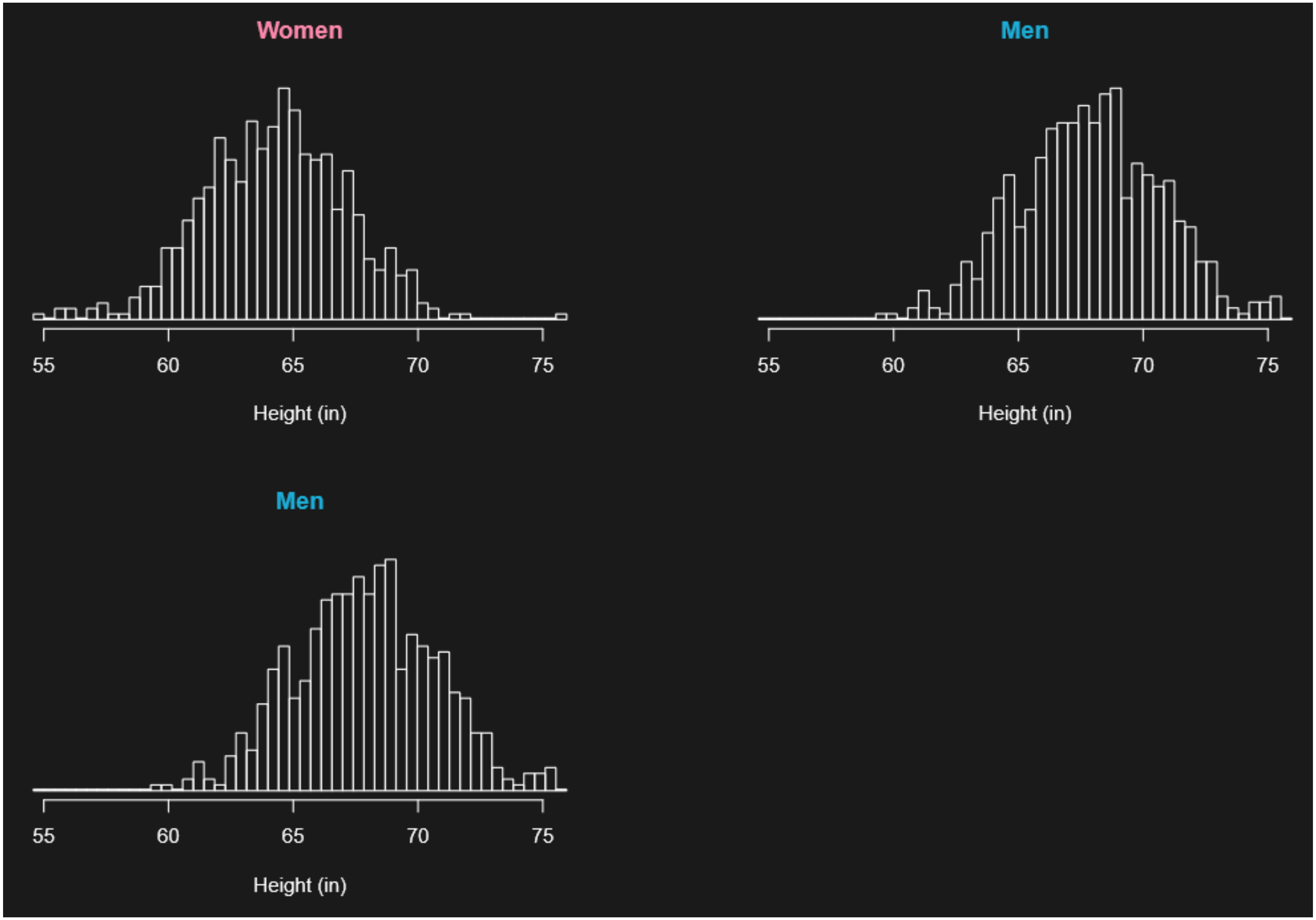
Use color where helpful



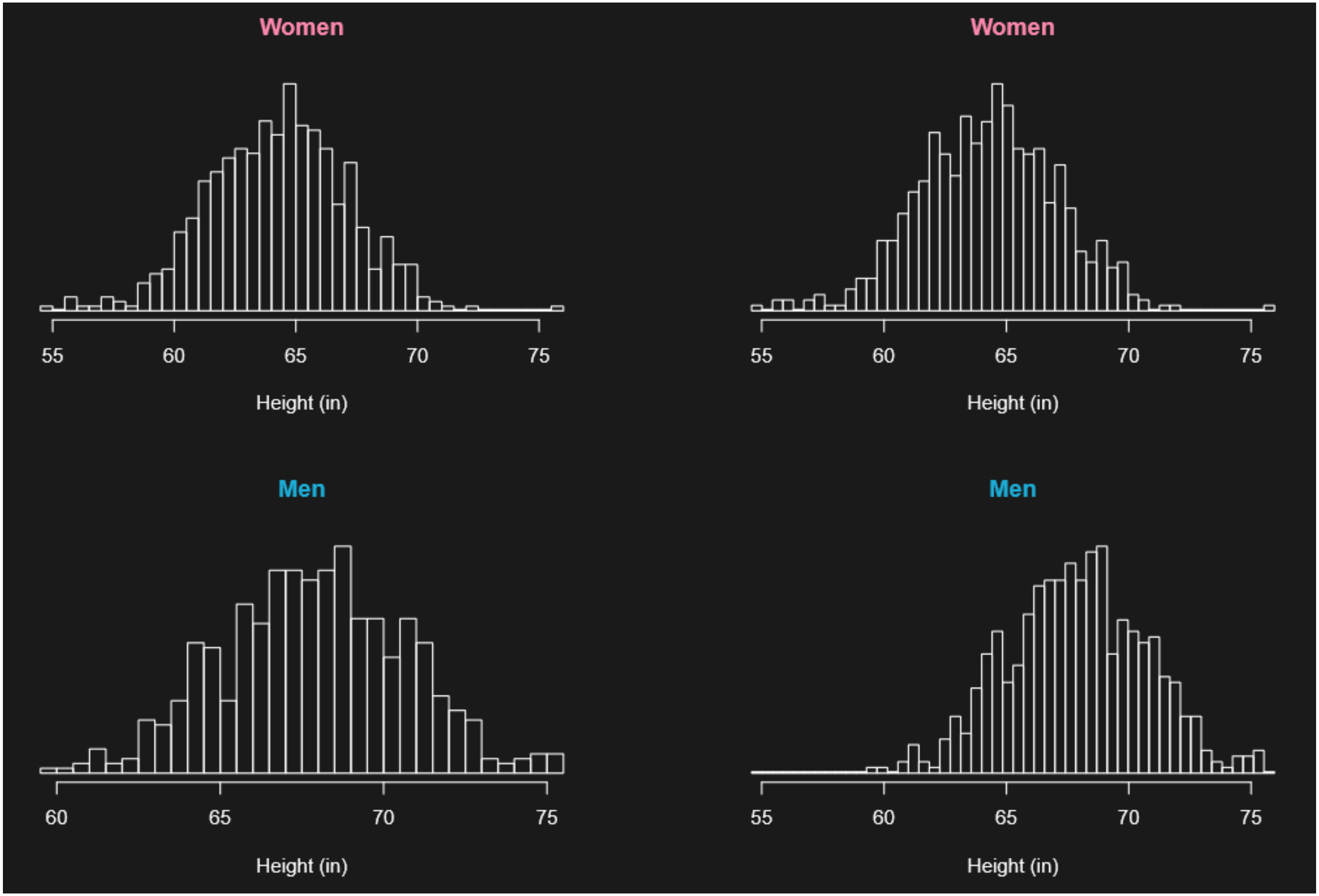
Where easiest to compare A and B?



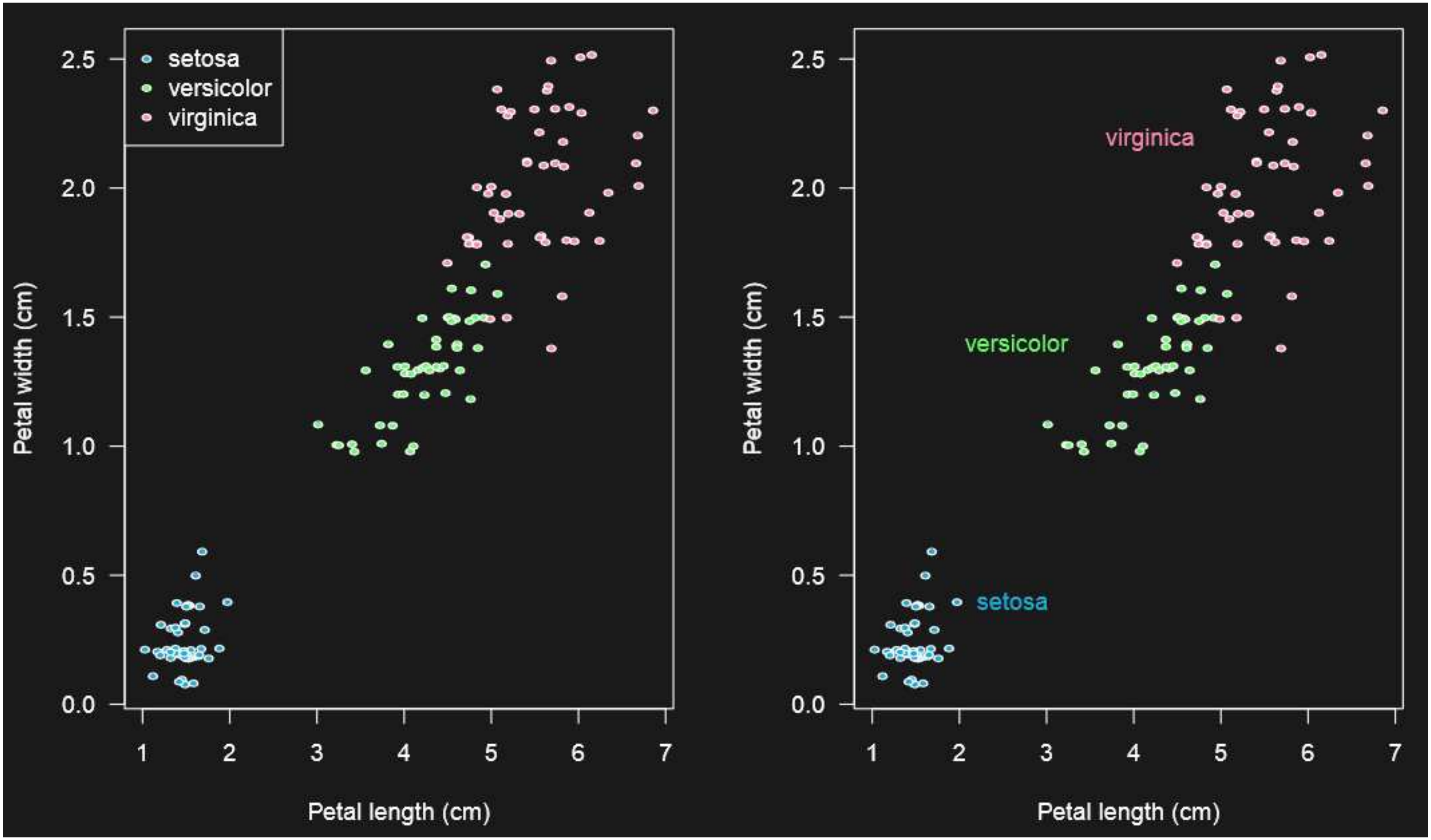
Easier to compare vertical aligned



Use common axes

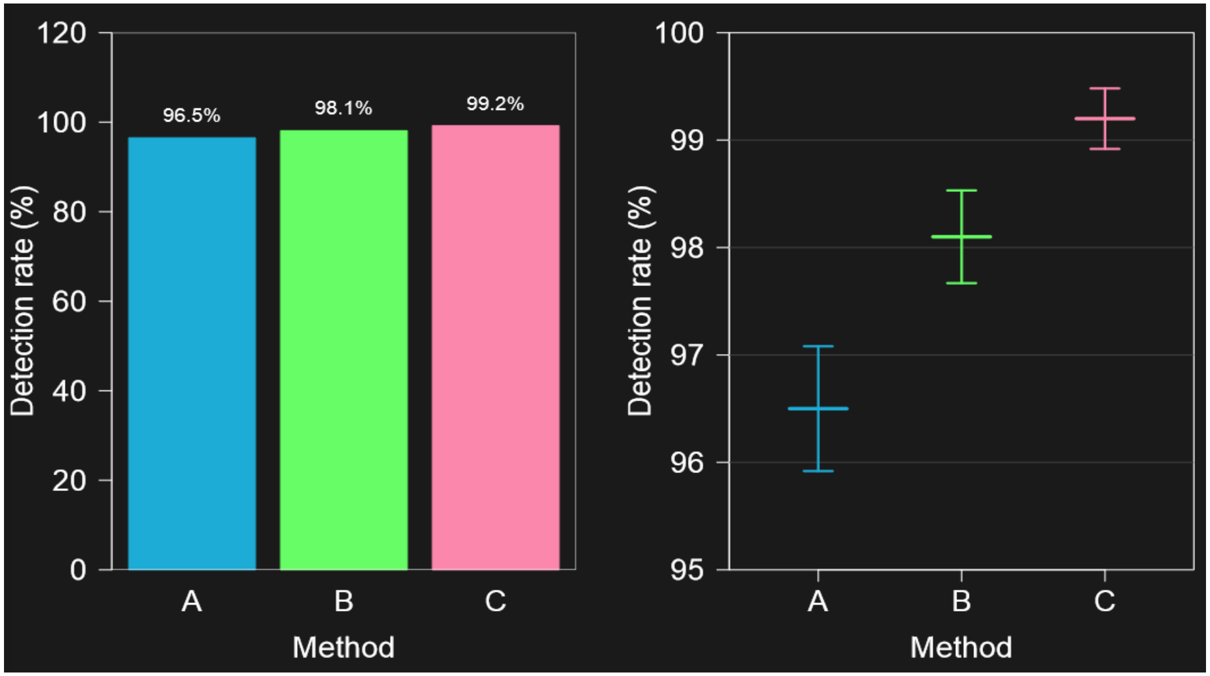


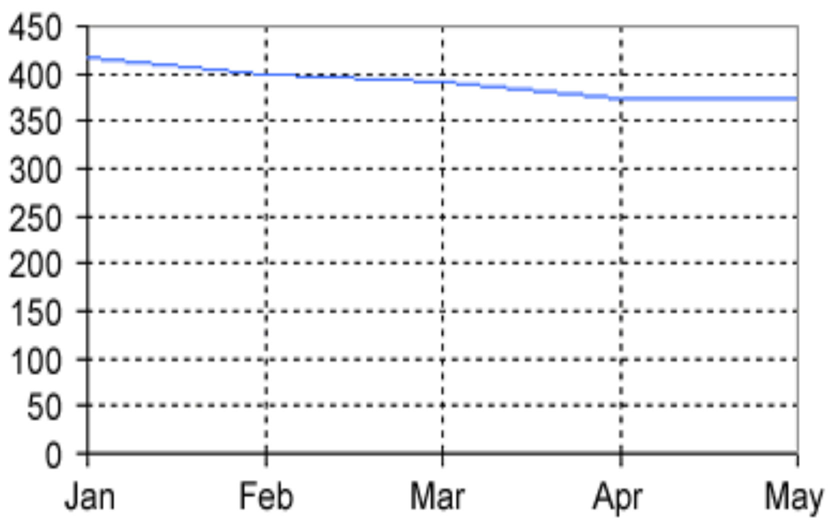
Use labels not legends *

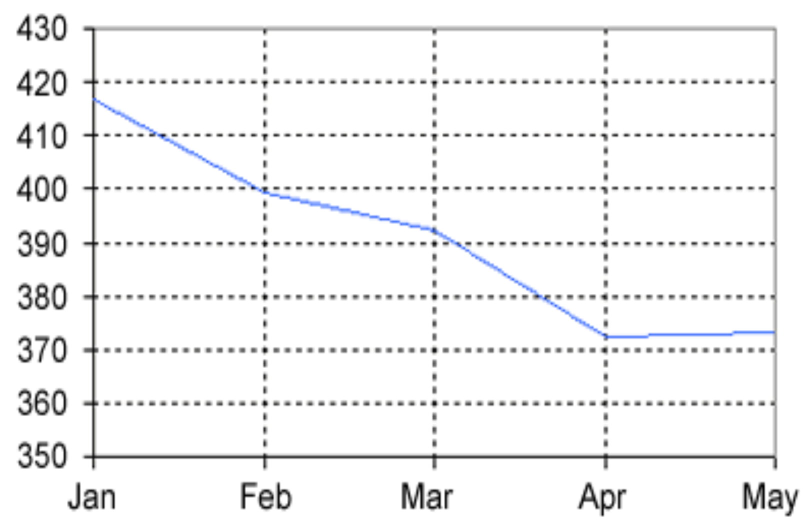


* Where possible

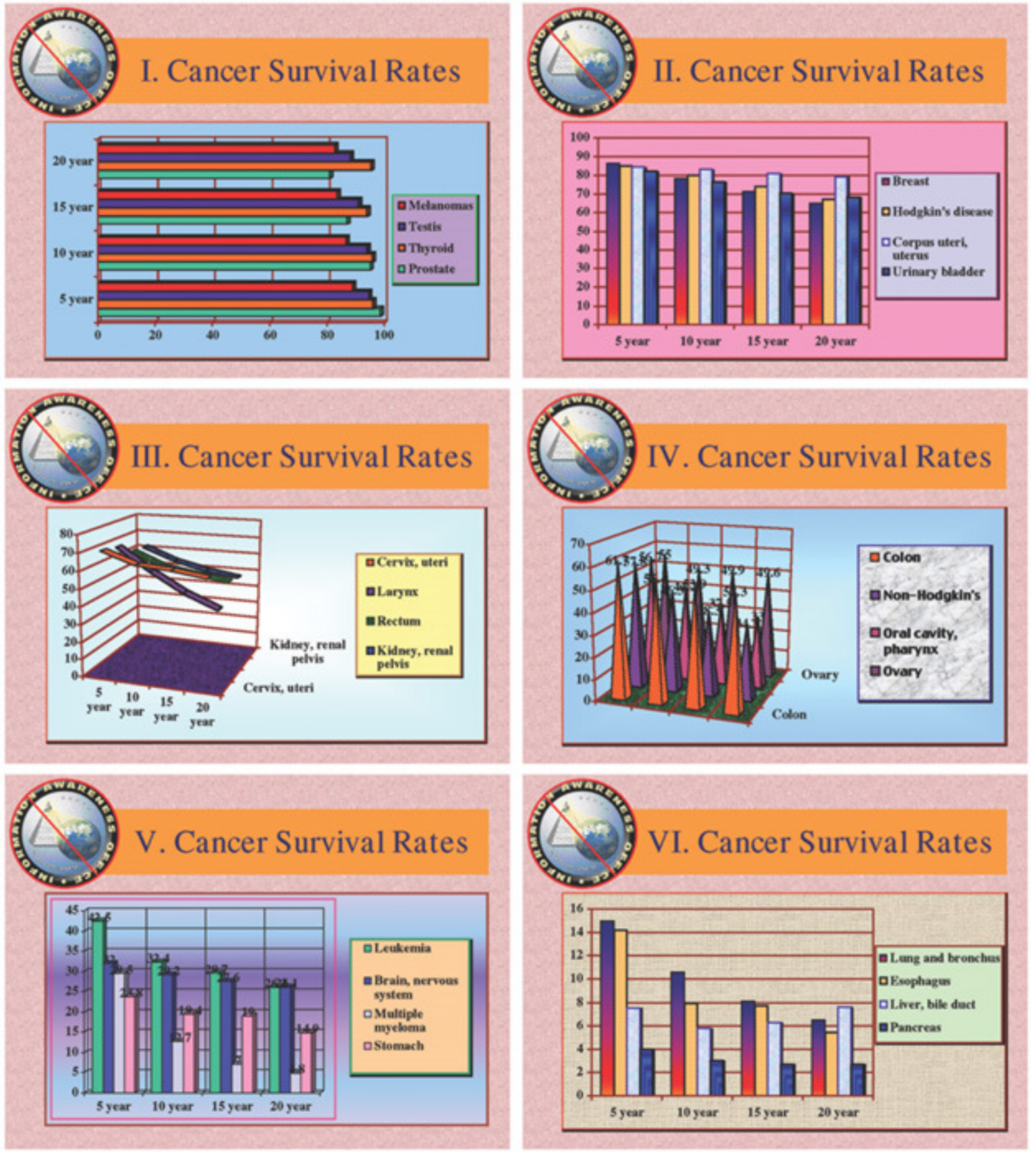
Consider whether you need 0







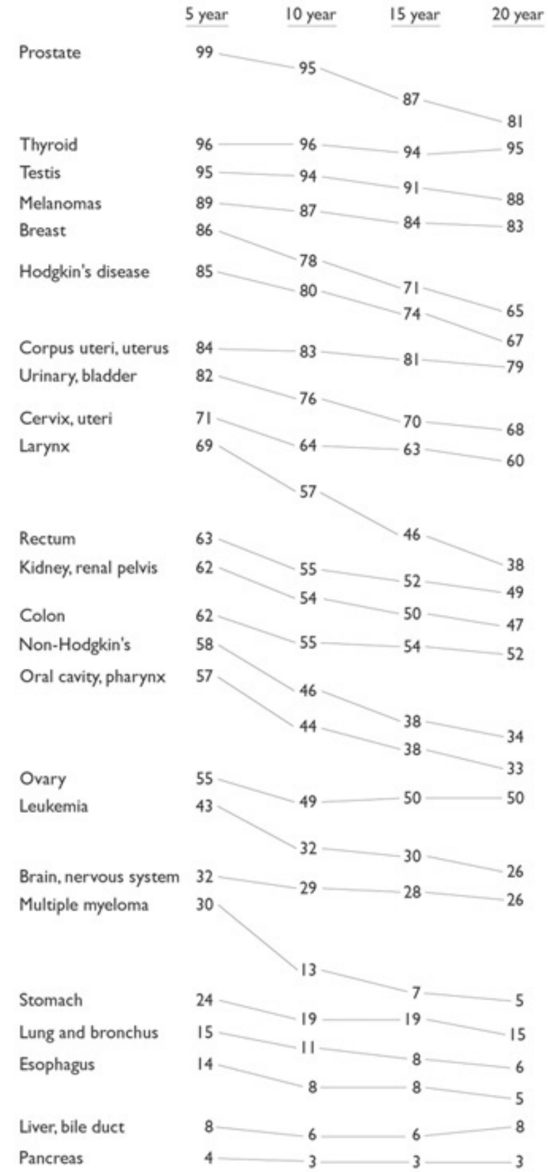
Several types of problems



The same data

Estimates of relative survival rates, by cancer site

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8



More advanced techniques

- Cluster analysis
  - Leads to readily interpretable figures
  - Can be helpful for identifying patterns in time or space
  - Can be used for exploratory purposes
  - Used to find groups of objects when not already known
- Principal components analysis
  - Often used as exploratory tool
  - Dimensionality reduction
- Useful for EDA and quality assessment of high-dimensional datasets