1. Introduction/Overview

Readings - Rosner: 1, 2.7-12, 6.4, 13.1-2
SAS: PROC GPLOT, PROC UNIVARIATE
PROC GCHART, PROC FREQ
R: Hmisc: describe, factor, plot, hist;
ggplot2: qplot, stem, plot

Overview

- Statistics as a discipline arose from the need to use data to answer scientific questions in the face of uncertainty

- Statistical concepts are at the heart of scientific inquiry in the health sciences

- Your mastery of fundamental concepts will facilitate:
  - a better understanding of published research
  - a better understanding of how to structure effective scientific research
  - interpretation and presentation of results
  - collaboration with other biostatisticians and scientific investigators

(see last page for guide to critique of literature/research)

- Along the way you'll learn about common pitfalls, such as:
  - Threats to validity and reproducibility - e.g. lack of denominators, control groups
  - Association implies causation – need to apply scientific method and Hill's classic criteria (**http://www.edwardtufte.com/tufte/hill**)

- Multiplicity – leads to *selection* of results and this has an *effect* on observed ability to replicate results (see 2005 *JAMA* paper by Ioannidis on course website)
  - Multiple variables, endpoints, time points, subgroups, comparisons
  - Multiple testing, multiple looks
  - Multiple models and adjustments
  - Fishing expeditions, mountains of output without *a priori* thought and justification - …*what exactly was my (their) research hypothesis or question…???*

Compelling examples that underscore the need for careful design and analysis:

- Retrolental fibroplasia (*Scientific American* 236:100-107; 1977); http://www.neonatology.org/classics/parable/notes.html
- Beta carotene for prevention of Lung Cancer (*NEJM*, 330: 1029-35; 1994); http://www.nejm.org/doi/full/10.1056/NEJM199404143301501
- HRT for prevention of cardiovascular disease in postmenopausal women (*JAMA*, 288:321-333; 2002); http://jama.ama-assn.org/cgi/content/abstract/288/3/321

Let's get started then …

A) What is/are Statistics?
B) Samples and sampling
C) Designing and executing studies
D) Data, variables
E) Displaying and graphing Data

An Example: 45 non-obese men and women completed weighed food intake records. Describe their eating patterns: On average, what was the kcal intake per day? How did intake vary from person to person? Do men eat more than women? Do heavier people eat more, etc?

```
data diet;
  input sex fdwt3 kcal3 prot3gm fat3gm cho3gm ncal3gm pctfat3 pctcho3 pctpro3;
  datalines;
  1  782  1780    59.0   56.1  287.1   379.8    28.4    64.5    13.3
  1  963  2150    64.8   73.9  318.0   506.3    30.9    59.2    12.1
  1 1432  2754   110.1  119.6  312.6   889.7    39.1    45.4    16.0
  1 2366  4403   120.8  135.7  694.2  1415.3    27.7    63.1    11.0
  1 2986  4475   243.4  213.7  423.5  2105.4    43.0    37.9    21.8
  1 1430  2716    88.1   90.6  395.2   856.1    30.0    58.2    13.0
  1 1857  2244   165.4  107.0  153.5  1431.1    42.9    27.4    29.5
  1 1111  1482   114.4   32.2  178.9   785.5    19.6    48.3    30.9
  1 1046  1799    88.2   40.6  278.3   638.9    20.3    61.9    19.6
  1 1576  2538   175.6   88.4  267.8  1044.2    31.3    42.2    27.7
  1 1269  2292    83.5  113.8  250.5   821.2    44.7    43.7    14.6
  1  611  2280    72.0   45.0  420.0    74.0    17.8    73.7    12.6
  1 3006  3312   224.4   49.9  506.4  2225.3    13.6    61.2    27.1
  1 1409  2651   109.0   90.3  354.1   855.6    30.7    53.4    16.4
  0 1017  1855    57.2   64.5  271.8   623.5    31.3    58.6    12.3
  0  988  1340    35.8   24.5  256.6   671.1    16.5    76.6    10.7
  0 1864  3086    93.5   84.5  512.3  1173.7    24.6    66.4    12.1
  0  874  1196    72.7   56.2  114.5   630.6    42.3    38.3    24.3
  0 1493  2313   131.5   59.4  314.2   987.9    23.1    54.3    22.7
  0  848  1708    77.8   89.4  146.3   534.5    47.1    34.3    18.2
  0 1420  2821    93.8  124.9  349.1   852.2    39.8    49.5    13.3
  0 1518  2212    68.4   43.1  401.6  1004.9    17.5    72.6    12.4
  0 1056  1588    69.1   33.9  262.1   690.9    19.2    66.0    17.4
  1 1405  2059    63.5   67.1  316.3   958.1    29.3    61.4    12.3
  1  738  1435    42.1   67.0  169.5   459.4    42.0    47.2    11.7
  1 1790  4352   151.2  169.3  562.9   906.6    35.0    51.7    13.9
  1  788  1936    86.5   88.5  207.1   405.9    41.1    42.8    17.9
  0  998  1902    77.7   79.5  218.8   622.0    37.6    46.0    16.3
  1 1293  2547    95.4   78.1  368.6   750.9    27.6    57.9    15.0
  0  521   836    39.5   17.7  137.0   326.8    19.1    65.6    18.9
  0 1120  1760    85.6   40.0  278.7   715.7    20.5    63.3    19.5
  1 1187  2037   118.1   69.8  240.8   758.3    30.8    47.3    23.2
  1 1453  2715   150.8  152.0  185.1   965.1    50.4    27.3    22.2
  1 1574  2916   125.6  127.7  319.8  1000.9    39.4    43.9    17.2
```

```
  1 2452  4204   139.7  110.2  686.0  1516.1    23.6    65.3    13.3
  1 1613  2522   136.3   57.0  382.6  1037.1    20.3    60.7    21.6
  1 1553  2518    62.4   79.7  393.9  1017.0    28.5    62.6     9.9
  1 2066  4317   174.3  181.1  524.3  1186.3    37.8    48.6    16.2
  1 1798  3254   135.3  108.3  432.5  1121.9    30.0    53.2    16.6
  0  667  1352    90.5   30.8  174.1   371.6    20.5    51.5    26.8
  1 1719  3064    75.2  174.9  305.0  1163.9    51.4    39.8     9.8
  1 1294  2572   124.9   81.1  352.3   735.7    28.4    54.8    19.4
  1 1599  2585    92.0  131.6  274.8  1100.6    45.8    42.5    14.2
  1  705  1863    44.6   79.8  247.8   332.8    38.6    53.2     9.6
  1 1157  3103   128.4  164.1  293.1   571.4    47.6    37.8    16.6
 ; run;
```

A) What is/are Statistics?

Statistics is a collection of methods for collecting, classifying, summarizing, analyzing (quantifying relationships), and presenting data. It involves using a ***sample*** to make ***inferences*** about a ***population***, i.e., data are available for a group of individuals/subjects that we are willing to consider as ***representative*** of some larger population of individuals. We usually start the analytic process by ***describing*** the sample.

Two main areas (overlapping):

1) Mathematical: development of new methods of statistical inference that requires detailed knowledge of abstract mathematics for its implementation

2) Applied: Application of methods of mathematical statistics to specific subject areas, e.g. econometrics--economics, psychometrics -- psychology

Biostatistics: application (and development) of statistical methods to problems in biology, medicine and public health

Two main aspects of statistics:

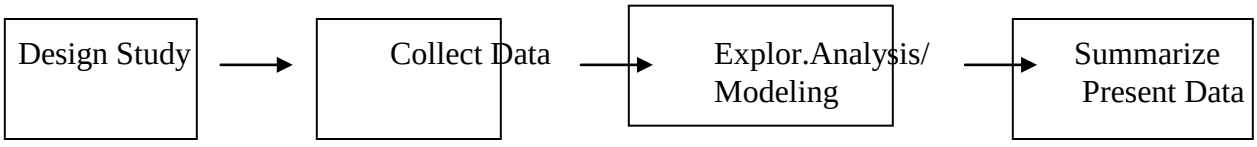
1) *Descriptive*: characterize data using graphs, tables, numerical summaries

2) *Inferential*: draw conclusions about populations of interest based on samples using appropriate analytical strategies

B) Samples and sampling

Samples provide partial information about a population. If drawn at ***random*** (and correctly) they can be assumed to represent the population. The variability in the information contained from (possible) sample to (possible) sample is called ***sampling variability*** and is a function of the underlying (natural, biological) variability in what is being measured in the sample, the sample size, and, for small populations, the size of the population. The incomplete nature of random samples, as representatives of a population, means that, ***by chance***, our inferences to the population could be wrong.

C) Designing and executing studies/research



It's important always to consider all of the above steps:

Study design: What is the question or hypothesis to be addressed? To what population do we want to make an inference? What are the strengths and weaknesses of various study designs for answering the question? Will the data collected really answer the research question? What are the outcomes of interest, predictors, potential confounders (biases)? Will the sample be large enough to provide meaningful results?

Data collection: Have well-standardized measures been used? Are they ***valid*** and ***reliable***? Have all the necessary variables been collected? What is the nature and extent of missing data?

Validity – *Does it measure what it's supposed to measure?*

Reliability – *Are repeated measurements approximately the same?*

Exploratory analysis: Edit the data, checking for errors; perform quality control. Make heavy use of statistical graphics to look for patterns, check assumptions. Obtain basic descriptive statistics by examining averages or percents or graphing data. Look for unusual patterns or observations. Identify outliers. Check sample sizes.

Modeling: Use models that make sense based on the research question, study design, results of the exploratory analysis. Use models you can understand and interpret. Check that the results make sense and follow from the descriptive data analysis.

Reporting results: Create useful tables and graphs (see Edward Tufte on this! - **http://www.edwardtufte.com/tufte/**). Use

simple language to convey results. Avoid logical leaps and over-interpretation.

Some common types of studies seen in epidemiology, medicine, public health, biology: (listed by increasing validity)

Observational: researcher observes what exists

Case series – sample of outcomes only, no controls, good for hypothesis generation

Cross-sectional – single point in time, no temporal sequence established, hypothesis generating

Retrospective – sampling based on outcome, backward look at exposure, confounding must be accounted for, great potential for recall bias, must choose controls carefully; good for rare outcomes; hypothesis testing. *Case-control*

Prospective – longitudinal – sampling based on "exposure", which precedes outcome in time, confounding must be accounted for since exposure is not randomly assigned; hypothesis testing. *Cohort*

e.g. food intake example is observational

Experimental: researcher randomly assigns cases to treatment, intervention, etc. Confounding minimized or eliminated. Temporal sequence can be established. Hypothesis testing – scientific method.

Randomized/ Controlled/Clinical/Community Trial

Crossover study

e.g.

There are advantages and disadvantages to both main types of studies. Experiments are usually more difficult (sometimes impossible!), but give more valid information about possible cause-effect relationships.

The text lists a variety of practical problems in executing and analyzing studies, including difficulties in getting accurate data and data management. Statisticians are, ideally, involved in all phases of research studies.

D) Data, variables: What is/are data? – Measured or observed characteristics of a person, animal, plant, entity, or phenomenon.

Dataset: file containing the information collected from a study; usually stored on a computer or on paper, organized with a column for each variable and a row for each case ("flat file")

Cases: subjects, the items that are measured

e.g. 45 obese men (sex = 1) and women (sex = 0)

Variables: the quantities or qualities that are measured

e.g. weighed food intake in kcal/day (kcal3)

Types of Variables: In general, the type of measurement scale determines both the descriptive and the inferential statistics used.

Nominal: not ordered, identified by name; qualitative, categorical, class

e.g. hair color, gender

Ordinal: ordered categories, numbers may not have meaning, spacing may not be equal (linear)

e.g. never, sometimes, frequently, always

Numerical scales -

Discrete: ordered values with inherent gaps between successive observable values, i.e., between any two successive points on the scale there are no other observable points

e.g., number of teeth

Range of possible continuous values might be condensed into discrete groupings

e.g., shoe size is continuous measured discretely

Continuous: ordered values such that between any two observable values there are an infinite number of observable values - e.g. 0.1, 0.15. 0.157 …; equal differences between values have equal quantitative meaning

e.g. height, weight, blood pressure

Often reported in discrete form reflecting the precision of the measurements.

Subtypes of continuous scales:

Interval – no meaningful zero

e.g.

Ratio – meaningful zero

http://books.google.com/books?id=NT8eiiYhIpoC&pg=PA66&lpg=PA66&dq=continuous+ratio+scale+kelvin&source=bl&ots=6YaGLLd4WH&sig=UGSwmXcoHKFKlO1R8pGzPPqvOcA&hl=en&ei=tB

2gSv2BJI2aMYSWjfQP&sa=X&oi=book_result&ct=result&resnum=1#v=onepage&q=continuous%20ratio%20scale%20kelvin&f=false

e.g.

Ranks: ordering of values of a numerical variable

e.g. 8, 4 , 1 , 6, 9 ⇒ 4, 2 , 1 , 3 , 5

E) Displaying and Graphing data: Tables, bar graphs, histograms, stem-and-leaf plots, box plots

Purpose of graphical displays:

1. Exploration: quick overview of data; better study and understanding
2. Presentation: better communication

For each case (subject, individual, etc.) various quantities or qualities (variables) are presented visually. Variables may be continuous or discrete (some overlap here) – there are different ways to represent these.

1) Tables: Provide a numerical summary of frequencies, percents, summary statistics, etc.

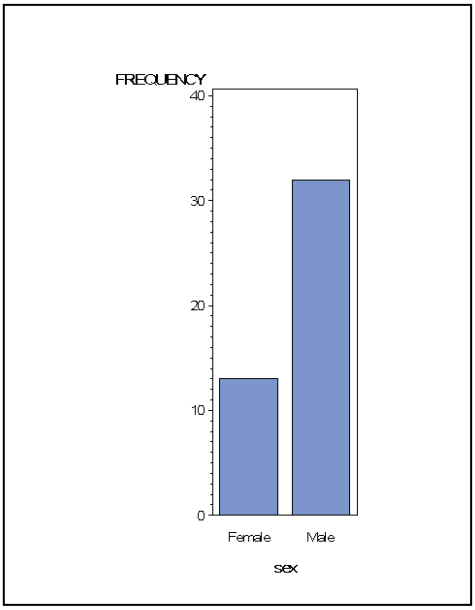
e.g. diet data

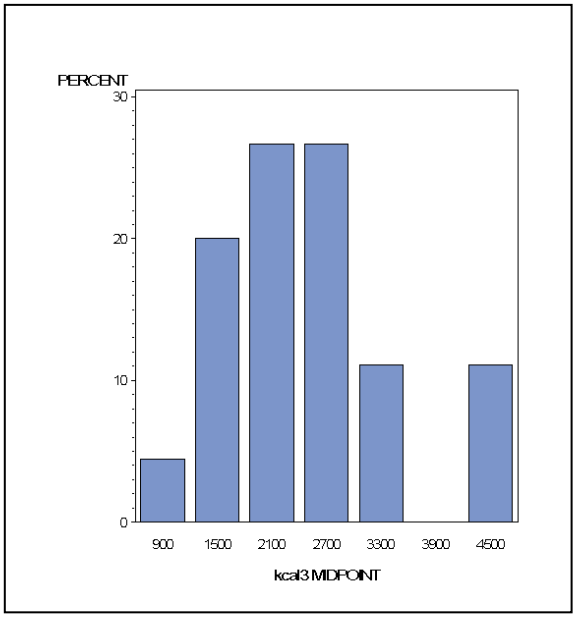
```
ODS PDF;

   PROC FREQ DATA=diet;
    TABLE sex;
   RUN;

ODS PDF CLOSE;
```

The FREQ Procedure

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	13	28.89	13	28.89
Male	32	71.11	45	100.00

2) Bar Graphs: good for representing measured/observed values for nominal variables. Data are divided into groups and for each group, a rectangle is constructed with a base of constant width and height proportional to the frequency within that group. Rectangles are generally equally spaced and not contiguous.






```
   ODS PDF;

PROC GCHART DATA=diet;
 VBAR sex / DISCRETE TYPE=freq;
RUN;

   ODS PDF CLOSE;
```

```
   ODS PDF;

   PROC GCHART DATA=diet;
    VBAR kcal3 / TYPE=percent;
   RUN;

   ODS PDF CLOSE;
```

3) Histograms: good for representing measured/observed values for discrete or continuous variables. Data are divided into groups and for each group, a rectangle is constructed with a base

location corresponding to the position of the ends of the group interval along the x-axis and area proportional to the frequency within that group. The scale used along either axis should allow all rectangles to fit into the space allotted for the graph.

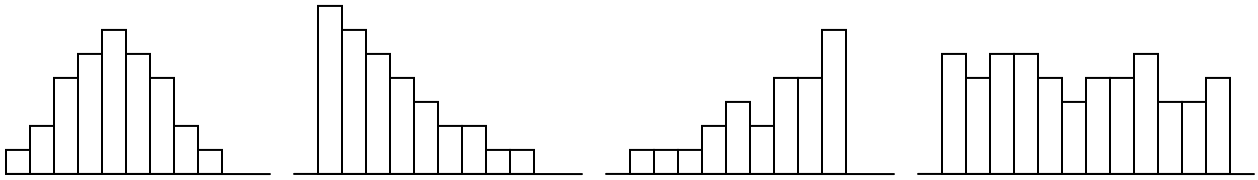
e.g. systolic blood pressure, birthweight

Frequencies and Histograms: PROC FREQ in SAS

Must choose classes (# and limits); must adjust for unequal class sizes (Rosner p. 30)

e.g. blood pressure groupings with corresponding frequencies (60-69--3, 70-6, etc)

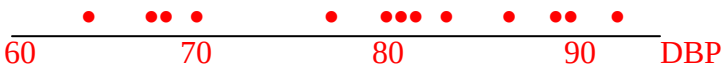
Useful for picturing location and shape of a variable's distribution in a large sample.



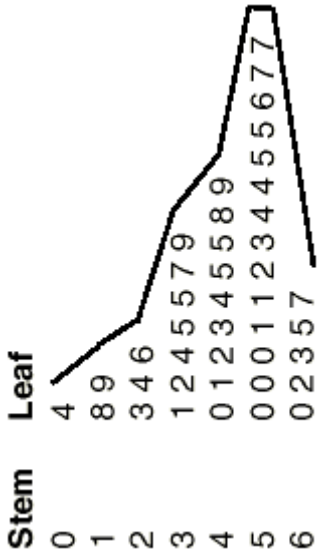
Symmetric Skewed Positive Skewed Negative Uniform

```
ODS PDF;

      PROC UNIVARIATE DATA=diet;
       CLASS sex;
       HISTOGRAM kcal3;
       VAR kcal3;
       TITLE 'Calorie intake by gender';
      RUN;

ODS PDF CLOSE;
```

4) Dot Plot: dots on axis. If comparing several groups, use same scale for all groups. Easy to do by hand; good for small samples. Look for groupings, outliers, skewness…



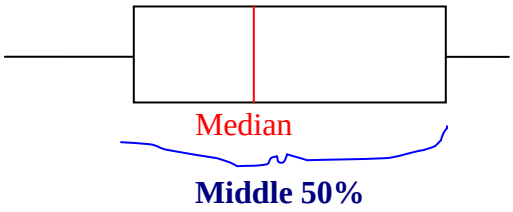
5) Stem and Leaf Plot: see description in text. Another way of representing frequency distribution.



6) Box Plot: see description in text. Uses the relationships among the median, upper quartile and lower quartile to visually portray symmetry vs. skewness of a distribution. A robust

display of a distribution that is useful for identifying outliers and comparing several distributions at once.

e.g. daily temperature in Melbourne

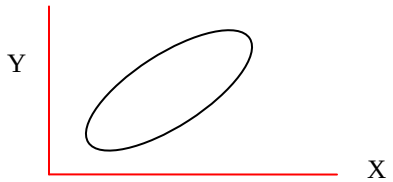


e.g. diet data

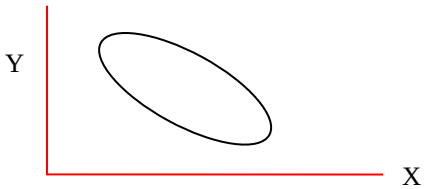
```
PROC UNIVARIATE DATA=diet freq PLOT;
 VAR kcal3;
RUN;
```

```
The UNIVARIATE Procedure
Variable:  kcal3
Stem Leaf                     #  Boxplot
44 08                         2     0
42 025                        3     0
40                                  |
38                                  |
36                                  |
34                                  |
32 51                         2     |
30 690                        3     |
28 22                         2  +-----+
26 5225                       4  |     |
24 224578                     6  |  +  |
22 14891                      5  *-----*
20 465                        3  |     |
18 06604                      5  +-----+
16 168                        3     |
14 489                        3     |
12 045                        3     |
10                                  |
8 4                           1     |
----+----+----+----+
Multiply Stem.Leaf by 10**+2
```

7) Scatterplot: describes the relationship between two numerical variables.

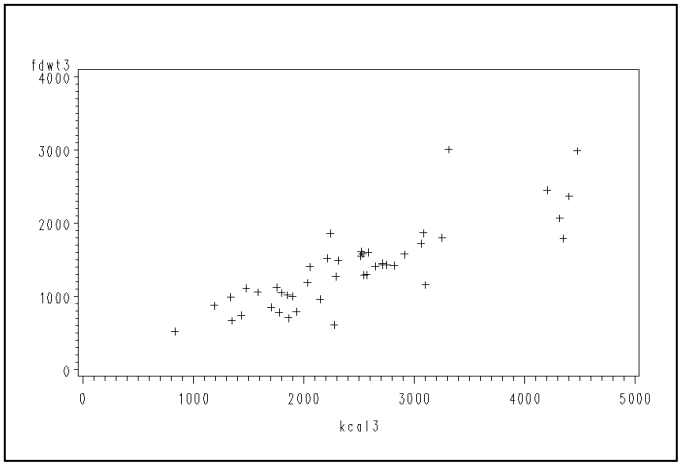


Positive Correlation



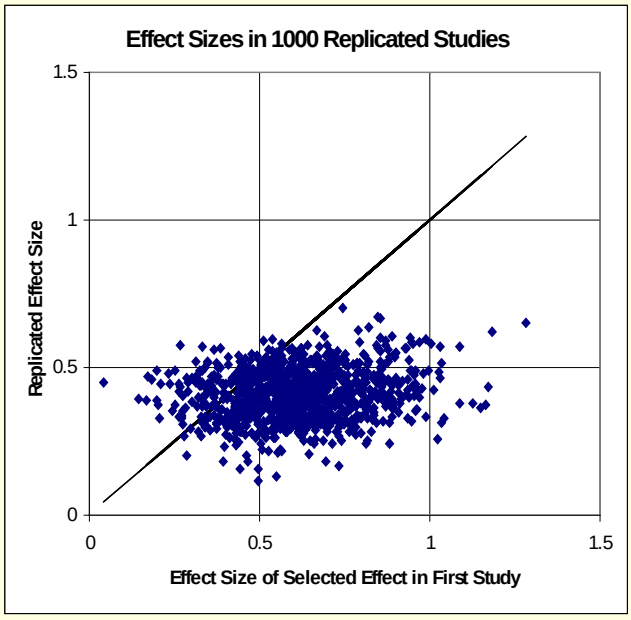
Negative Correlation

```
ODS PDF

PROC GPLOT DATA=diet;
 PLOT fdwt3*kcal3;
RUN;

ODS PDF CLOSE;
```



e.g. What can we conclude from the following simulation based on the *JAMA* study of non-replication of highly cited research?

The Scientific Concern

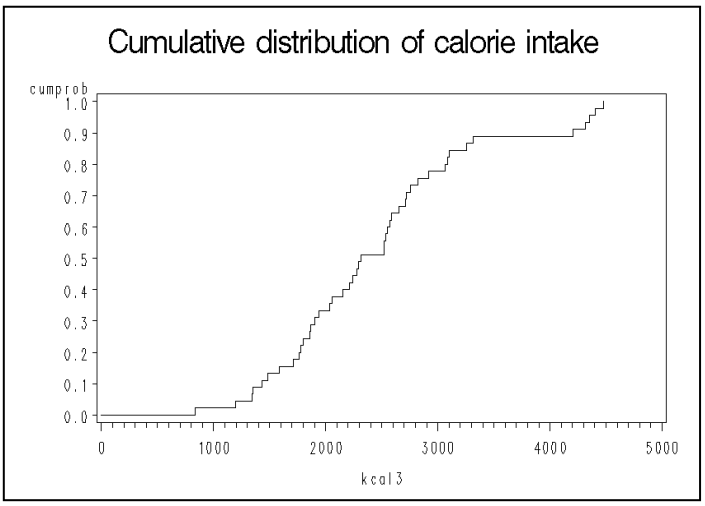


10

This phenomenon is called ***Regression to the Mean***. You'll see it occur in many contexts. Beware! Source: Peter Westfall – A Course in Multiple Comparisons and Multiple Tests. Joint Statistical Meetings, Salt Lake City, August 30, 2007

8) Cumulative Distributions: see text. Assist in determining the percentiles of a sample. Good for survival data with censored observations.

e.g. diet data

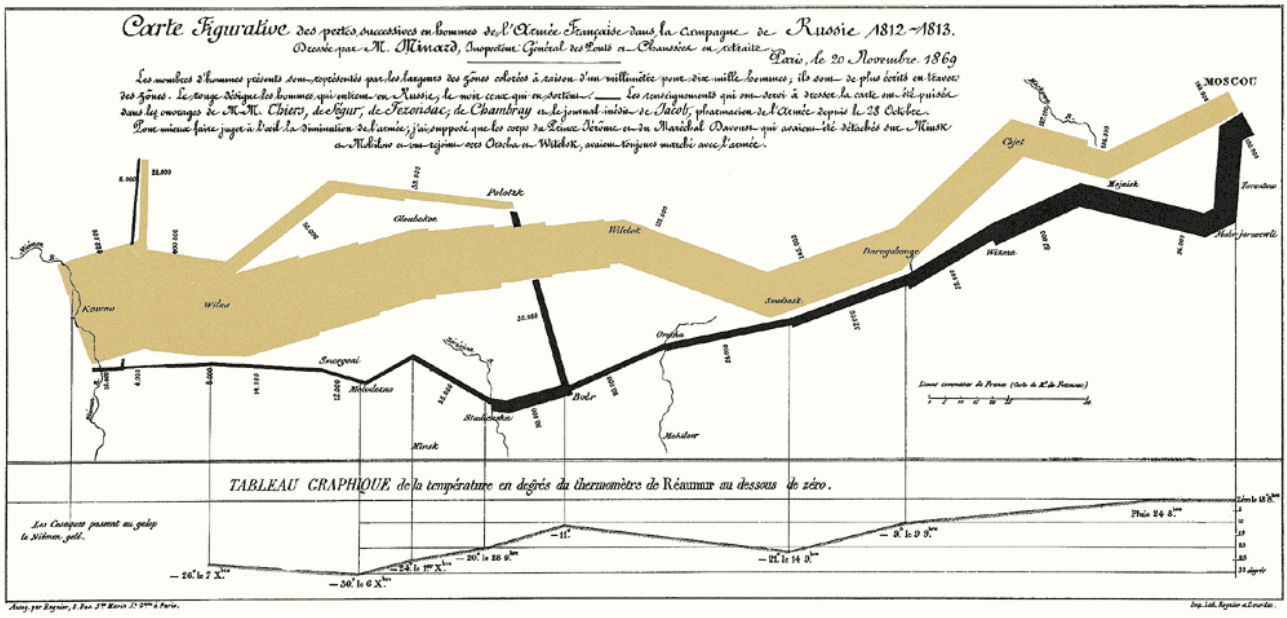


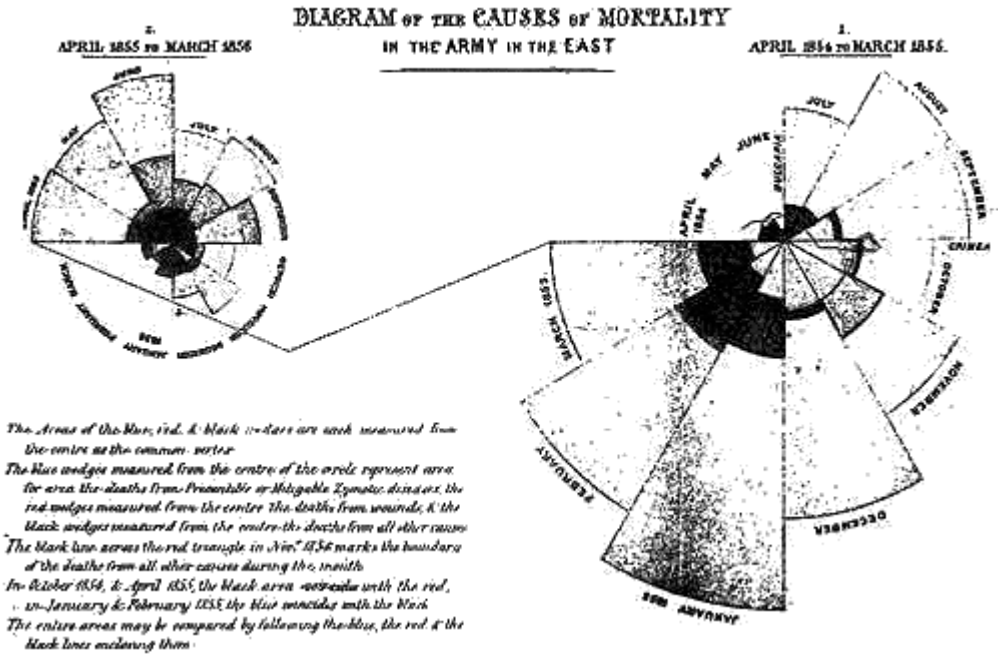

```
ODS PDF;

          PROC LIFETEST DATA=diet OUTSURV=a;
           TIME kcal3*censor(0);
          RUN;

          DATA a;
           SET a;
           cumprob = 1 - survival;
          RUN;

          PROC GPLOT DATA=a;
           PLOT cumprob*kcal3;
           SYMBOL i = stepjl;
           TITLE 'Cumulative distribution of calorie intake';
          RUN;

ODS PDF CLOSE;
```

Note: Graphics are powerful – worth a thousand words! Use them liberally to explore data and convey results.

Depending on the audience, you may choose to present data in different ways.

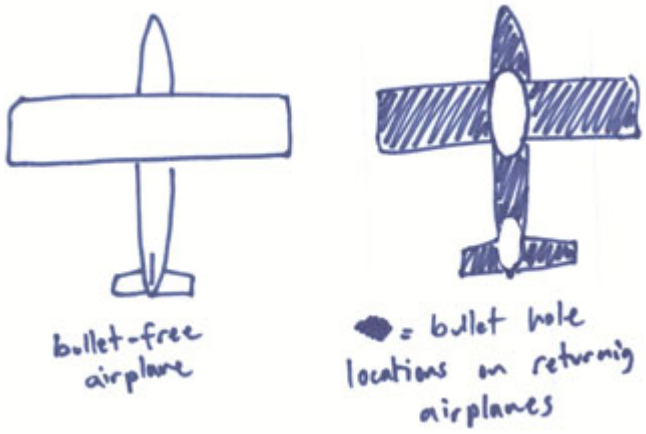
Classic examples of excellent graphics:





See also:
http://www.csiss.org/classics/content/8

The absence of presence or the presence of absence?



See also:
http://digitalroam.typepad.com/digital_roam/2006/03/the_hole_story_.html

Rutstein's/Colton's "Outline for Critique of a Medical Report"

I. Object or hypothesis

A. What are the objectives of the study or the questions to be answered?

B. What is the population to which the investigators intend to refer their findings?

*C. Is the title appropriate?

II. Design

A. Was the study an experiment, planned observations, or an analysis of records? *or a laboratory experiment?

B. How was the sample selected? Are there possible sources of selection which would make the sample atypical or nonrepresentative? If so, what provision was made to deal with this bias?

C. What is the nature of the control group or standard of comparison?

*D. If a clinical trial, was it approved by an IRB and was informed consent obtained?

III. Observations

A. Are there clear definitions of the terms used, including diagnostic criteria, measurements made, and criteria of outcome?

B. Was the method of classification or of measurements consistent for all the subjects and relevant to the objectives of the investigation? Are there possible biases in measurement and if so, what provision was made to deal with them?

C. Are the observations reliable and reproducible?

*D. If an animal experiment which species was used and why?

*E. If a laboratory experiment, was enough detail given for you to replicate the experiment?

IV. Presentation of findings

A. Are the findings presented clearly, objectively, and in sufficient detail to enable the reader to judge them for him/herself?

B. Are the findings internally consistent, i.e. do the numbers add up properly, can the different tables be reconciled, etc.?

V. Analysis

A. Are the data worthy of statistical analysis? If so, are the methods of statistical analysis appropriate to the source and nature of the data and is the analysis correctly performed and interpreted?

B. Is there sufficient analysis to determine whether significant differences may in fact be due to lack of comparability among the groups in sex or age distribution, in clinical characteristics, or in other relevant variables?

*C. Was there a power analysis to determine sample size and a discussion of a "clinically important" difference?

VI. Conclusion

Which conclusions are justified by the findings? Which are not? Are the conclusions relevant to the questions posed by the investigators?

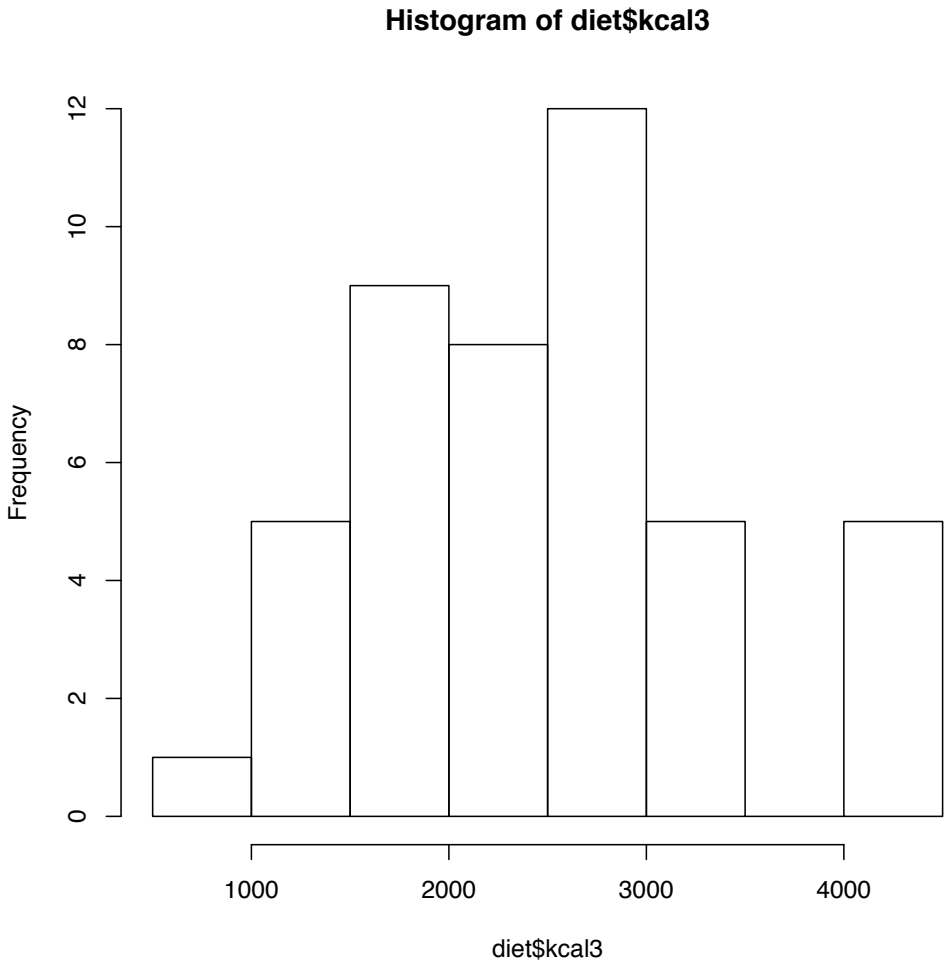
VII. Constructive suggestions

Assume you are planning an investigation to answer the questions put in this study. If they have not been clearly put by the authors, frame them in an appropriate manner. Suggest a practical design, criteria for observations, and type of analysis that would provide reliable and valid information relevant to the questions under study.

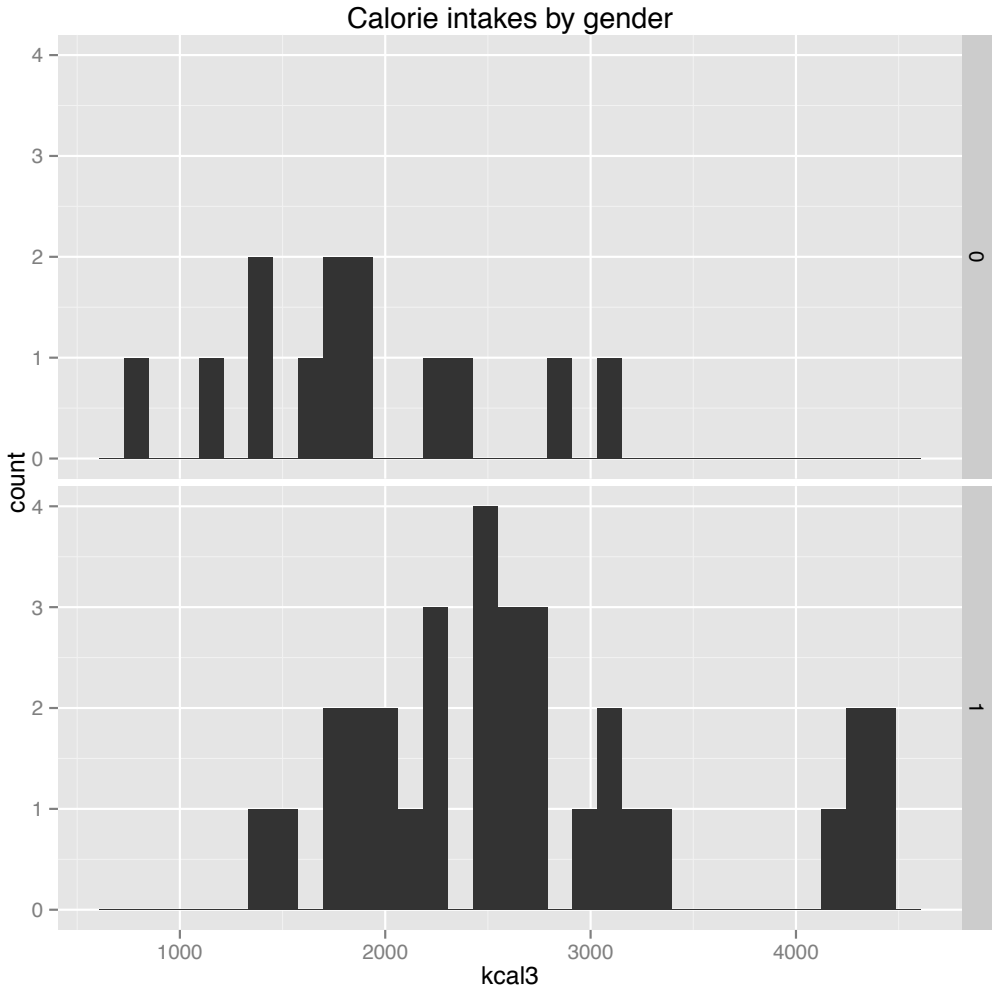
Source: Colton, T. Statistics in Medicine. Boston: Little, Brown and Co., 1974.

(* = added by H. James Norton)

```
hist(diet$kcal3)
```

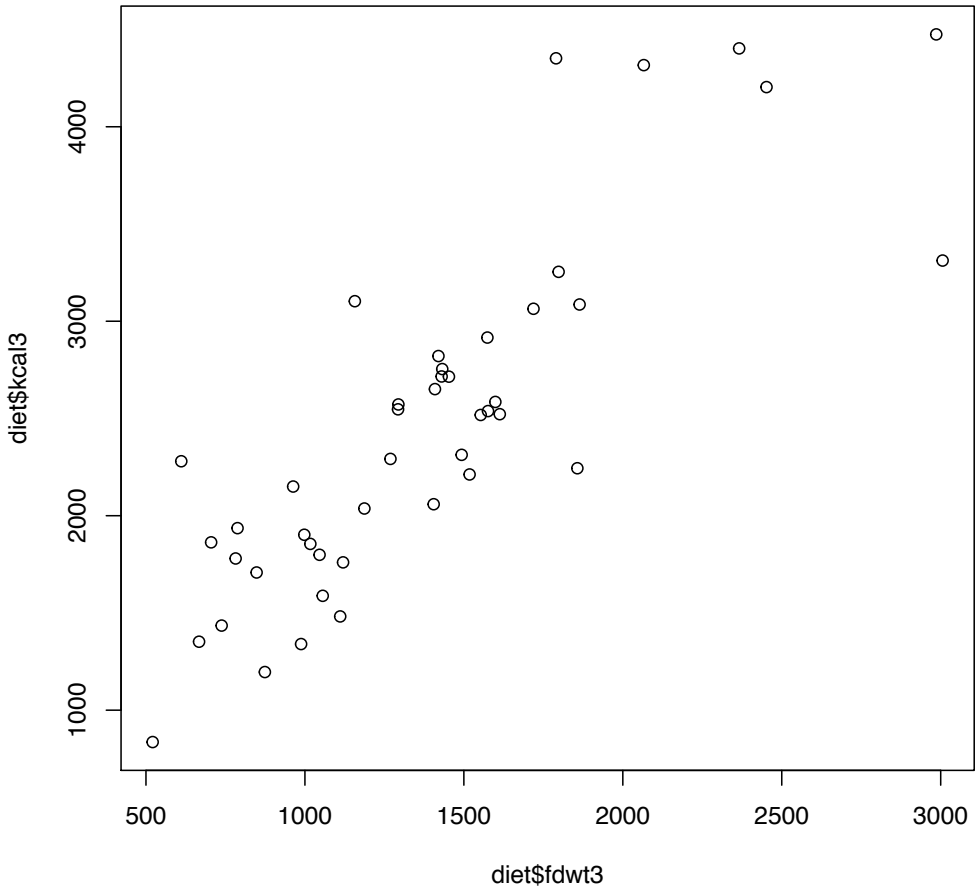



```
library(ggplot2)
qplot(kcal3, data = diet, facets = sex ~ .) + ggtitle("Calorie intakes by gender")
```




```
stem(diet$kcal3)


  The decimal point is 3 digit(s) to the right of the |

  0 | 8
  1 | 2344
  1 | 5678889999
  2 | 01222333
  2 | 555566777889
  3 | 11133
  3 |
  4 | 2344
  4 | 5
```

```
plot(diet$fdwt3, diet$kcal3)
```



Appendix: Code

```
diet<-read.table(header=T, con <- textConnection('
sex fdwt3 kcal3 prot3gm fat3gm cho3gm ncal3gm pctfat3 pctcho3 pctpro3
  1   782  1780    59.0   56.1  287.1   379.8    28.4    64.5    13.3
  1   963  2150    64.8   73.9  318.0   506.3    30.9    59.2    12.1
  1  1432  2754   110.1  119.6  312.6   889.7    39.1    45.4    16.0
  1  2366  4403   120.8  135.7  694.2  1415.3    27.7    63.1    11.0
  1  2986  4475   243.4  213.7  423.5  2105.4    43.0    37.9    21.8
  1  1430  2716    88.1   90.6  395.2   856.1    30.0    58.2    13.0
  1  1857  2244   165.4  107.0  153.5  1431.1    42.9    27.4    29.5
  1  1111  1482   114.4   32.2  178.9   785.5    19.6    48.3    30.9
  1  1046  1799    88.2   40.6  278.3   638.9    20.3    61.9    19.6
  1  1576  2538   175.6   88.4  267.8  1044.2    31.3    42.2    27.7
  1  1269  2292    83.5  113.8  250.5   821.2    44.7    43.7    14.6
  1   611  2280    72.0   45.0  420.0    74.0    17.8    73.7    12.6
  1  3006  3312   224.4   49.9  506.4  2225.3    13.6    61.2    27.1
  1  1409  2651   109.0   90.3  354.1   855.6    30.7    53.4    16.4
  0  1017  1855    57.2   64.5  271.8   623.5    31.3    58.6    12.3
  0   988  1340    35.8   24.5  256.6   671.1    16.5    76.6    10.7
  0  1864  3086    93.5   84.5  512.3  1173.7    24.6    66.4    12.1
  0   874  1196    72.7   56.2  114.5   630.6    42.3    38.3    24.3
  0  1493  2313   131.5   59.4  314.2   987.9    23.1    54.3    22.7
  0   848  1708    77.8   89.4  146.3   534.5    47.1    34.3    18.2
  0  1420  2821    93.8  124.9  349.1   852.2    39.8    49.5    13.3
  0  1518  2212    68.4   43.1  401.6  1004.9    17.5    72.6    12.4
  0  1056  1588    69.1   33.9  262.1   690.9    19.2    66.0    17.4
  1  1405  2059    63.5   67.1  316.3   958.1    29.3    61.4    12.3
  1   738  1435    42.1   67.0  169.5   459.4    42.0    47.2    11.7
  1  1790  4352   151.2  169.3  562.9   906.6    35.0    51.7    13.9
  1   788  1936    86.5   88.5  207.1   405.9    41.1    42.8    17.9
  0   998  1902    77.7   79.5  218.8   622.0    37.6    46.0    16.3
  1  1293  2547    95.4   78.1  368.6   750.9    27.6    57.9    15.0
  0   521   836    39.5   17.7  137.0   326.8    19.1    65.6    18.9
  0  1120  1760    85.6   40.0  278.7   715.7    20.5    63.3    19.5
  1  1187  2037   118.1   69.8  240.8   758.3    30.8    47.3    23.2
  1  1453  2715   150.8  152.0  185.1   965.1    50.4    27.3    22.2
  1  1574  2916   125.6  127.7  319.8  1000.9    39.4    43.9    17.2
  1  2452  4204   139.7  110.2  686.0  1516.1    23.6    65.3    13.3
  1  1613  2522   136.3   57.0  382.6  1037.1    20.3    60.7    21.6
  1  1553  2518    62.4   79.7  393.9  1017.0    28.5    62.6     9.9
  1  2066  4317   174.3  181.1  524.3  1186.3    37.8    48.6    16.2
  1  1798  3254   135.3  108.3  432.5  1121.9    30.0    53.2    16.6
  0   667  1352    90.5   30.8  174.1   371.6    20.5    51.5    26.8
  1  1719  3064    75.2  174.9  305.0  1163.9    51.4    39.8     9.8
  1  1294  2572   124.9   81.1  352.3   735.7    28.4    54.8    19.4
  1  1599  2585    92.0  131.6  274.8  1100.6    45.8    42.5    14.2
  1   705  1863    44.6   79.8  247.8   332.8    38.6    53.2     9.6
  1  1157  3103   128.4  164.1  293.1   571.4    47.6    37.8    16.6
')
)

install.packages("Hmisc")
diet <- read.csv("~/Dropbox/6611METHODS/6611/diet.csv")
library(Hmisc)
describe(factor(diet$sex))
plot(factor(diet$sex))
hist(diet$kcal3)

install.packages("ggplot2")
library(ggplot2)
```

```
qplot(kcal3,data=diet,facets=sex~.)+ggtitle("Calorie intakes by gender")

stem(diet$kcal3)
plot(diet$fdwt3,diet$kcal3)
```