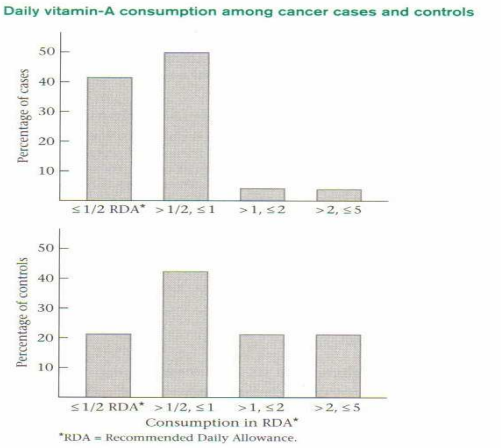
BINF702 FALL 2006

CHAPTER 2 – Descriptive Statistics

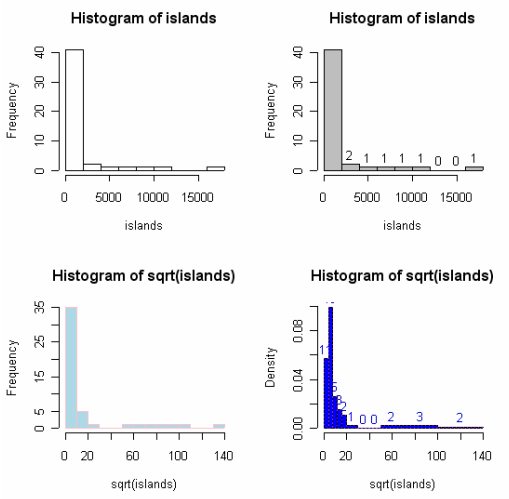
Section 2.1 - Introduction

- If our set of observations is small we may study them via enumeration. This is usually not possible though.
- Ex. 2.1 – Some investigators have proposed that consumption of vitamin A prevents cancer. To test this theory, a dietary questionnaire to collect data on vitamin-A consumption among 200 hospitalized cancer cases and 200 controls might be used. The controls would be matched on age and sex to the cancer cases and would be in the hospital at the same time for an unrelated disease. What should be done with these data after they are collected?

Barplot of the Smokers Data



hist in R

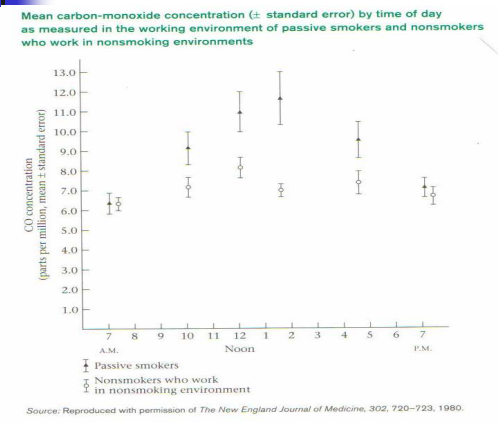



```
op <- par(mfrow=c(2, 2))

hist(islands)

utils::str(hist(islands,
col="gray", labels =
TRUE))

 hist(sqrt(islands), br
= 12, col="lightblue",
border="pink")

##-- For non-equidistant
breaks, counts should
NOT be graphed unscaled:

 r <-
hist(sqrt(islands), br =
c(4*0:5, 10*3:5, 70,
100, 140), col='blue1')

text(r$mids, r$density,
r$counts, adj=c(.5, -
.5), col='blue3')

sapply(r[2:3], sum)
sum(r$density *
diff(r$breaks)) # == 1
lines(r, lty = 3, border
= "purple") # ->

lines.histogram(*)
```

Section 2.1 - Introduction

- Ex. 2.2 – Medical researchers have often suspected that passive smokers-people who themselves do not smoke but who live or work in an environment where other where others smoke-might have impaired pulmonary function as a result. In 1980 a research group in San Diego published results indicating that passive smokers did indeed have significantly lower pulmonary function than comparable nonsmokers who did not work in smoky environments. As supporting evidence, the authors measured the carbon-monoxide (CO) concentrations in the working environment of passive smokers and of nonsmokers (where no smoking was permitted in the workplace) to see if the relative CO concentration changed over the course of the day.

Section 2.1 - Scatterplot of the Smokers Data With Confidence Bounds



It would be an interesting pedagogical exercise to try to write an R function to produce a plot like this. The inputs would be a set of x y values.

Discussions on Graphics Etiquette

- Dan Carr
  - Several EDA and Visualization courses at GMU
    - Statistical Graphics and Data Exploration (STAT663/CSI773)
    - Scientific and Statistical Visualization (CSI703/IT875)
- Bill Cleveland
  - **Visualizing Data (Hardcover)** by William S. Cleveland
- Naomi Robbins
  - **Creating More Effective Graphs (Paperback)** by Naomi B. Robbins
- Edward R. Tufte
  - **The Visual Display of Quantitative Information (Hardcover)** by Edward R. Tufte

Section 2.2 – Measurements of Location

- The Central Problem of Statistics – Consider a sample of data $x_1, x_2, \ldots, x_n$ drawn from a population P what inferences or conclusions about P can be gleaned from the sample?
- Section 2.2.1 – Arithmetic Mean
- Summation Notation

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n$$

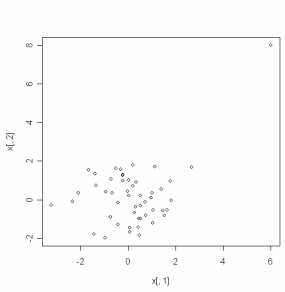
$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$$

- Def. 2.1 – The arithmetic mean $\overline{x}$ is defined as

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

Section 2.2 – The Arithmetic Mean

- The arithmetic mean is sensitive to the presence of outliers.
- Def. – An outlier is a point that differs in some manner from the rest of the data points.



Section 2.2 The Arithmetic Mean

- The mean in R
- mean(x, …)

```
> x = matrix(rnorm(100), nrow=50, ncol=2)
> x[50,1] = 60
> x[50,2] = 880
> plot(x[,1],x[,2])
> apply(x,2,mean)
[1] 1.259362 1.648327
> apply(x,2,mean,trim=0.1)
[1] 0.1475340 0.1076194
```

The fraction of observations to be removed from each end of the dataset.

Section 2.2 – The Median

- Def. 2.2 – The sample median is given by

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observation if n is even.

Section 2.2.1 – The Median in R

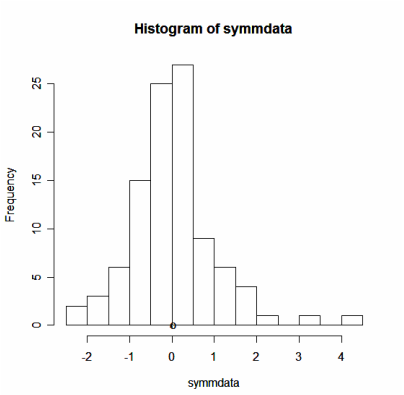
- median(x, na.rm = FALSE) ← What does this do?

- Ex. 2.6

```
tbl2.2  = c(7, 35, 5, 9, 8, 3, 10, 12, 8)
> median(tbl2.2)
[1] 8
```

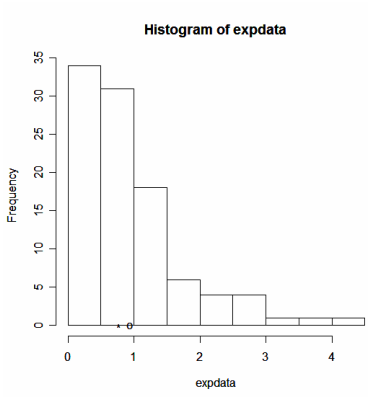
Section 2.2 – Comparison of the Arithmetic Mean and the Median (Symmetric Distribution)

```
> symmdata = rnorm(100)
> hist(symmdata,nclass=10)
> points(mean(symmdata),
   0, pch = 'o')
> points(median(symmdata),
   0, pch = '*')
```



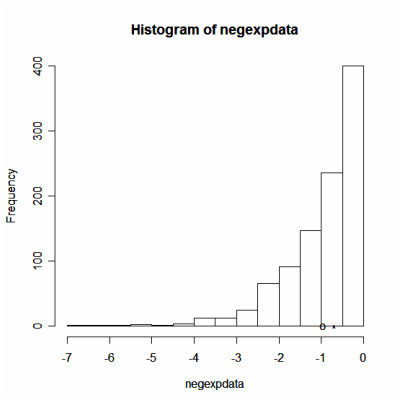
Section 2.2 – Comparison of the Arithmetic Mean and the Median (Asymmetric Distribution Right Tailed)

```
> expdata = rexp(100)
> hist(expdata, nclass =
   15)
> points(mean(expdata), 0,
   pch = 'o')
> points(median(expdata),
   0, pch = '*')
```

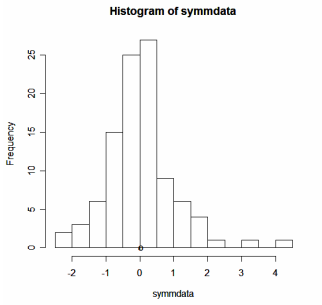


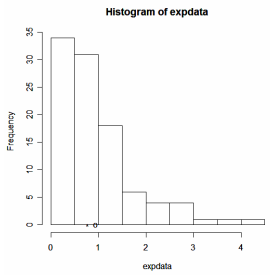
Section 2.2 – Comparison of the Arithmetic Mean and the Median (Asymmetric Distribution Left Tailed)

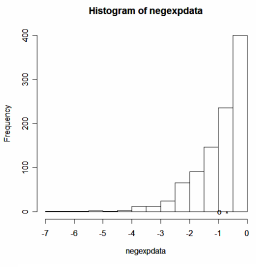
```
> negexpdata = -rexp(1000)
> hist(negexpdata, nclass
   = 20)
> points(mean(negexpdata),
   0, pch = 'o')
>
   points(median(negexpdata
   ), 0, pch = '*')
```



Section 2.2 – Comparison of the Arithmetic Mean and the Median (All Three Distributions)







Section 2.2 – Measures of Location (The Mode)

- Def. 2.3 – The mode is the most frequently occurring value among all the observations in a sample.
  - Can there be more than one mode?
  - Does there have to be a mode?
  - Does a mode have to be a data value?
  - What do you think a unimodal distribution is?
  - How about a multimodal distribution?
- Writing a function for the mode in R would be an interesting pedagogical exercise.

Section 2.2 – Measures of Location (The Geometric Mean)

Def. 2.4 - The geometric mean is the antilogarithm of $\overline{\log x}$, where

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

- There is currently no geometric mean in R.

2.3 – Some Properties of the Arithmetic Mean (Effect of Translation)

Eq. 2.1 - If $y_i = x_i + c, i = 1, \cdots, n$

then $\bar{y} = \bar{x} + c$

```
> x = rnorm(1000)
> mean(x)
[1] 0.02741812
> mean(x+5)
[1] 5.027418
```


2.3 – Some Properties of the Arithmetic Mean (Effect of Scaling)

Eq. 2.2 - If $y_i = cx_i, \ i = 1, \cdots, n$

then $\bar{y} = c\bar{x}.$

```
> x = rnorm(1000)
> mean(x)
[1] 0.02741812
> mean(2*x)
[1] 0.05483624
```

2.3 – Some Properties of the Arithmetic Mean (Effect of Translation and Scaling)

Eq. 2.3 - Let $x_1, \cdots, x_n$ be the original sample of data and let $y_i = c_1 x_i + c_2,\ i = 1, \cdots, n$ represent a transformed sample obtained by multiplying each orinigal sample point by a factor c_1 and then shifting over by a constant c_2

then $\overline{y} = c_1 \overline{x} + c_2$.

```
> x = rnorm(1000)
> mean(x)
[1] 0.02741812
> mean(2*x+5)
[1] 5.054836
```


2.4 – Measures of Spread

- Ex.
  - These two datasets have the same mean
  - c(rep(10, 10))
  - C(rep(0,9),100)
  - Are they the same?

2.4 – Measures of Spread (The range)

- Def. 2.5 – The range is the difference between the largest and the smallest observation in the sample.
- In R we merely write

```
range
```


2.4 – Measures of Spread (Quantiles)

- The pth percentile V_p is the value such that p% of the data lies to the left of this value. Formally we write.
- Def. 2.6 – The pth percentile is defined by
  - The (k+1)th largest sample point if np/100 is not and integer (where k is the largest integer less than np/100)
  - The average of the (np/100)th and (np/100 + 1)th largest observation sif np/100 is an integer.
- Percentiles are sometimes designated quartiles.

2.4 – Measures of Spread (quantiles in R)

- quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE, names = TRUE, type = 7, ...)

```
> bw = c(3265, 3260, 3245, 3484, 4146, 3323,
  3649, 3200, 3031, 2069, 2581, 2841, 3609, 2838,
  3541, 2759, 3248, 3314, 3101, 2834)

> quantile(bw, probs =c(.1,.9), type = 2)
 10%  90%
2670 3629
```


2.4 – Measures of Spread (The Variance and Standard Deviation)

- Hey let's try this

$$d = \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{n}$$

- Eq. 2.4 – The sum of the deviations of the individual observations of a sample about the sample mean is always 0 (bummer)

2.4 – Measures of Spread (The Variance and Standard Deviation)

Def. 2.7 - The sample variance, or variance , is defined as follows:

$$S^2 = \frac{\sum_{I=1}^{N} (x_i - \bar{x})^2}{n-1}$$

Def. 2.8 - The sample standard deviation, or standard deviation, is defined as follows:

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

2.4 – Measures of Spread (The Variance and Standard Deviation) in R

```
> x = rnorm(100)
> var(x)
[1] 1.031008
> sd(x)
[1] 1.015386
> sqrt(var(x))
[1] 1.015386
```

2.5 Some Properties of the Variance and Standard Deviation (Effect of Translation)

Eq. 2.5 Suppose there are two samples

$x_1, \cdots, x_n$ and $y_1, \cdots, y_n$

were $y_i = x_i + c, i = 1, \cdots, n$

If the respective sample variances of the two samples are denoted by

s_x^2 and s_y^2

then $s_y^2 = s_x^2$

2.5 Some Properties of the Variance and Standard Deviation (Effect of Scaling)

Eq. 2.6 Suppose there are two samples

$x_1, \cdots, x_n$ and $y_1, \cdots, y_n$

were $y_i = cx_i, i = 1, \cdots, n, c > 0$

If the respective sample variances of the two samples are denoted by

s_x^2 and s_y^2

Then $s_y^2 = c^2 s_x^2$; $s_y = cs_x$

2.6 – The Coefficient of Variation

Def. 2.9 The coefficient of variation (CV) is defined by

$$100\% * \left(\frac{s}{\bar{x}} \right)$$

- Since it is a unitless ratio, you can compare the CV of variables expressed in different units.
- It only makes sense to report CV for variables where zero really means zero, such as mass or enzyme activity.
- Don't calculate CV for variables, such as temperature, where the definition of zero is somewhat arbitrary.

http://www.graphpad.com/articles/interpret/Analyzing_one_group/descr_stats.htm

2.6 – The Coefficient of Variation and R

- There is no COV implemented in R but it would be trivial to do so.

2.7 Grouped Data

- Some times we have to many observations to examine all of them individually.
- De. 2.10 – A **frequency distribution** is an ordered display of each value in a data set together with its **frequency**; that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given.
- In R we have

```
xx = c(1,1,1,1,1,2,2,3,3,3,4,4,4,4,4)
> table(xx)
xx
1 2 3 4
5 2 3 5
```


2.7 Grouped Data in R (hist) - I

```
hist(x, ...)

   ## Default S3 method:
   hist(x, breaks = "Sturges", freq = NULL,
  probability = !freq,
         include.lowest = TRUE, right = TRUE,
         density = NULL, angle = 45, col = NULL,
  border = NULL,
         main = paste("Histogram of" , xname),
         xlim = range(breaks), ylim = NULL,
         xlab = xname, ylab,
         axes = TRUE, plot = TRUE, labels = FALSE,
         nclass = NULL, ...)
```

2.7 Grouped Data in R (hist) - II

```
Arguments:

      x: a vector of values for which the histogram is desired.

  breaks: one of:

            *  a vector giving the breakpoints between histogram
               cells,

            *  a single number giving the number of cells for the
               histogram,

            *  a character string naming an algorithm to compute the
               number of cells (see Details),

            *  a function to compute the number of cells.

          In the last three cases the number is a suggestion only.
```


2.7 Grouped Data in R (hist) - III

```
 freq: logical; if 'TRUE', the histogram graphic is a representation
          of frequencies, the 'counts' component of the result; if
          'FALSE', probability densities, component 'density', are
          plotted (so that the histogram has a total area of one).
          Defaults to 'TRUE' _iff_ 'breaks' are equidistant (and
          'probability' is not specified).

probability: an _alias_ for '!freq', for S compatibility.

include.lowest: logical; if 'TRUE', an 'x[i]' equal to the 'breaks'
          value will be included in the first (or last, for 'right =
          FALSE') bar.  This will be ignored (with a warning) unless
          'breaks' is a vector.

   right: logical; if 'TRUE', the histograms cells are right-closed
          (left open) intervals.
```

2.7 Grouped Data in R (hist) - IV

density: the density of shading lines, in lines per inch. The default value of 'NULL' means that no shading lines are drawn. Non-positive values of 'density' also inhibit the drawing of shading lines.

angle: the slope of shading lines, given as an angle in degrees (counter-clockwise).

col: a colour to be used to fill the bars. The default of 'NULL' yields unfilled bars.

2.7 Grouped Data in R (hist) - V

```
border: the color of the border around the bars.  The default
   is to
          use the standard foreground color.

main, xlab, ylab: these arguments to 'title' have useful
   defaults here.

xlim, ylim: the range of x and y values with sensible
   defaults. Note
          that 'xlim' is _not_ used to define the histogram
   (breaks),
          but only for plotting (when 'plot = TRUE').

    axes: logical.  If 'TRUE' (default), axes are draw if the
   plot is
          drawn.
```

2.7 Grouped Data in R (hist) - VI

```
plot: logical.  If 'TRUE' (default), a histogram is plotted,
          otherwise a list of breaks and counts is returned.

  labels: logical or character.  Additionally draw labels on
   top of
          bars, if not 'FALSE'; see 'plot.histogram'.

  nclass: numeric (integer).  For S(-PLUS) compatibility only,
   'nclass'
          is equivalent to 'breaks' for a scalar or character
   argument.

     ...: further graphical parameters passed to
   'plot.histogram' and
          their to 'title' and 'axis' (if 'plot=TRUE').
```


2.7 Grouped Data in R (hist) - VII

```
The definition of "histogram" differs by source (with
    country-specific biases).  R's default with equi-spaced breaks
    (also the default) is to plot the counts in the cells defined by
    'breaks'.  Thus the height of a rectangle is proportional to the
    number of points falling into the cell, as is the area _provided_
    the breaks are equally-spaced.

    The default with non-equi-spaced breaks is to give a plot of area
    one, in which the _area_ of the rectangles is the fraction of the
    data points falling in the cells.

    If 'right = TRUE' (default), the histogram cells are intervals of
    the form '(a, b]', i.e., they include their right-hand endpoint,
    but not their left one, with the exception of the first cell when
    'include.lowest' is 'TRUE'.
```

2.7 Grouped Data in R (hist) - VIII

```
For 'right = FALSE', the intervals are of the form '[a, b)', and
     'include.lowest' really has the meaning of "_include highest_".

     A numerical tolerance of 1e-7 times the median bin size is applied
     when counting entries on the edges of bins.

     The default for 'breaks' is '"Sturges"': see 'nclass.Sturges'.
     Other names for which algorithms are supplied are '"Scott"' and
     '"FD"' / '"Friedman-Diaconis"' (with corresponding functions
     'nclass.scott' and 'nclass.FD'). Case is ignored and partial
     matching is used. Alternatively, a function can be supplied which
     will compute the intended number of breaks as a function of 'x'.
```


2.7 Grouped Data in R (hist) – IX (values)

```
an object of class '"histogram"' which is a list with
  components:

  breaks: the n+1 cell boundaries (= 'breaks' if that
   was a vector).

  counts: n integers; for each cell, the number of
   'x[]' inside.

 density: values f^(x[i]), as estimated density values.
   If
          'all(diff(breaks) == 1)', they are the
   relative frequencies
          'counts/n' and in general satisfy sum[i;
   f^(x[i])
          (b[i+1]-b[i])] = 1, where b[i] = 'breaks[i]'.
```

2.7 Grouped Data in R (hist) – X (values)

```
intensities: same as 'density'. Deprecated, but
  retained for
          compatibility.

    mids: the n cell midpoints.

   xname: a character string with the actual 'x'
  argument name.

equidist: logical, indicating if the distances
  between 'breaks' are all
          the same.
```


2.7 Grouped Daa in R (hist)

An example (table 2.10 pg. 27)

```
> bw = c(58, 120, …)
> bw.hist = hist(bw, breaks = c(29.5, 69.5, 89.5, 99.5, 109.5,
   119.5, 129.5, 139.5, 169.5), freq=TRUE,include.lowest=TRUE)
> attributes(bw.hist)
$names
[1] "breaks"      "counts"      "intensities" "density"
   "mids"        "xname"       "equidist"

$class
[1] "histogram"

> bw.hist$counts
[1]  5 10 11 19 17 20 12  6
```

2.8 Graphic Methods

- Your author's definition of a bar graph really matches R's implementation of a histogram via the hist command.
- Stem-and-leaf plot
  - Carefully read your author's construction guidelines on pg. 28
  - We will talk about stem-and-leaf plots on the next slide

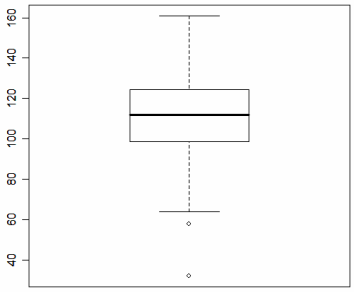
2.8 Graphics Methods (stem-and-leaf plot in R)

- An example (see Fig. 2.6 of your text)

```
> stem(bw, scale = 2)

  The decimal point is 1 digit(s) to the right of the |

   3 | 2
   4 |
   5 | 8
   6 | 478
   7 |
   8 | 3556788999
   9 | 12344568889
  10 | 0123444445567888899
  11 | 00122235555556889
  12 | 01112222344445567788
  13 | 222334557888
  14 | 0146
  15 | 5
  16 | 1
```

2.8 Graphics Methods (boxplot)

> boxplot(bw)



2.9 Case Study 1: Effects of Lead Exposure on Neurological and Psychological Function in Children - I

```
> lead = read.csv("C:/Documents and Settings/Owner/My
   Documents/Work/Beth/gmu/fall06/binf702/soltn_man/Data sets/ASCII -
   .txt comma/lead.txt",header=TRUE)
> dim(lead)
[1] 124  38
> lead[1,]
   Id Area  Age Sex Iqv_inf Iqv_comp Iqv_ar Iqv_ds Iqv_raw Iqp_pc
   Iqp_bd Iqp_oa Iqp_cod Iqp_raw HH_index Iqv Iqp Iqf Iq_type
1 101    3 1101   1       3        4      3      5      15     10
   8      8       5      31       77  61  85  70       1
  Lead_type Ld72 Ld73 Fst2yrs Totyrs Pica Colic Clumsi Irrit Convul
   X2Plat_r X2Plat_l Visrea_r Visrea_l Audrea_r Audrea_l
1         1   25   18       2     11    2     2      2     2      2
   16       16       36       38       27       25
  FWT_r FWT_l Hyperact
1    72    52       99
```

2.9 Case Study 1: Effects of Lead Exposure on Neurological and Psychological Function in Children - II

```
>boxplot(lead$Iqf[lead$Lead_t
  ype == 1],
  lead$Iqf[lead$Lead_type
  ==2])
```