Lecture 2: Box, bag and violin?

Today

We begin our discussion of data description, both numerical as well as visual -- We will expand on some of the selections made by your book and ignore others (notably, stem-and-leaf diagrams)

We will work with data from a large annual survey administered by the Centers for Disease Control and Prevention -- Our goal at this point is not inference (reasoning from this sample to the U.S. population) but instead data description

Turning Information Into Public Health

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based system of health surveys that collects information on health risk behaviors, preventive health practices, and health care access primarily related to chronic disease and injury. For many states, the BRFSS is the only available source of timely, accurate data on health-related behaviors.

BRFSS was established in 1984 by the Centers for Disease Control and Prevention (CDC); currently data are collected monthly in all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. More than 350,000 adults are interviewed each year, making the BRFSS **the largest telephone health survey in the world**. States use BRFSS data to identify emerging health problems, establish and track health objectives, and develop and evaluate public health policies and programs. Many states also use BRFSS data to support health-related legislative efforts.

California has used the system to

- Monitor the impact of tobacco control legislation based on Proposition 99.
- Assess the quality of life of adult Californians with arthritis in order to focus the California Arthritis Partnership Program resources appropriately.
- Assess the prevalence of diagnosed osteoporosis in California in order to target the California Osteoporosis Prevention and Education Program resources most effectively.
- Evaluate gun ownership and storage practices in California.
- Estimate potential lead exposure among adults and households with children in California.
- Estimate adult and child asthma prevalence in California to help study differences in asthma rates by various demographics.
- Evaluate the health care status and the health of uninsured California residents.
- Assess alcohol consumption and dependence.
- Assess the use of cancer screening tests.
- Identify Californians who are at risk for being overweight.
- Estimate sun avoidance and protection practices among California's adults and children.
- Assess the knowledge of certain sexually transmitted diseases (STDs) in order to target the efforts of STD awareness programs.
- Monitor progress toward *Healthy People 2010* goals for flu and pneumococcal immunization in California's seniors and high-risk individuals.

New York has used the system to

- Demonstrate the prevalence of disability and the care received by the elderly population in New York State.
- Provide information to the New York State Office for Aging use in of supporting budget requests.
- Monitor the effectiveness of the Performance Outcomes Measures Project, a national demonstration project funded by the Administration on Aging in response to a congressional mandate of the Government Performance Results Act.
- Provide measures for hypertension, high cholesterol, tobacco use, poor nutrition, physical inactivity, obesity, and diabetes to support program planning for the Cardiovascular Disease Program and for local health department's independent initiatives.
- Support efforts of the Tobacco Control Program to reduce smoking prevalence in youth and adults, prevent initiation of smoking in youth, reduce exposure to environmental tobacco smoke, and reduce disparities in tobacco use among affected groups.
- Provide data to the Tobacco Control Program, BRFSS and state and local programs to use in program planning.
- Provide data on adult tobacco use for an independent evaluation of the Tobacco Control Program.
- Support a grant application to monitor oral health status risk factors and use of oral health services.
- Provide data to the Bureau of Sexually Transmitted Disease Control (STD) to assess and monitor STD behavioral risks in New York State.
- Provide baseline statistics and impact evaluation measures for the Diabetes Surveillance and Evaluation Program to monitor progress toward the six national diabetes objectives proposed by the CDC.
- Provide data for various quarterly reports, annual summary reports, and funding continuation applications.

Description

Again, today our emphasis will be on the tools to examine a data set and not on formal inference -- In future lectures, we will focus on what this survey tells us about the larger U.S. population

In fact, for most of the quarter, we will examine design strategies for collecting data that will make inferences easy!

CDC
SAFER • HEALTHIER • PEOPLE™
GENTE•SEGURA•SALUDABLE™

Office of Surveillance, Epidemiology, and Laboratory Services
Behavioral Risk Factor Surveillance System

▶ TECHNICAL INFORMATION

- Overview
- BRFSS Datasets (downloads and documentation)
- Chronic Disease & the Environment
- Summary Data Quality Reports
- User's Guide
- BRFSS Forms

▶ BRFSS CONTENTS

- Prevalence and Trends Data
- SMART: City and County Data
- BRFSS Maps
- Web Enabled Analysis Tool (WEAT)
- Chronic Disease Indicators (CDI)
- About the BRFSS
- BRFSS Datasets (downloads and documentation)

BRFSS Annual Survey Data

Survey Data and Documentation

Annual Survey Data

Access the survey data and documentation for any BRFSS survey year. The documentation provides technical and statistical information regarding the BRFSS, such as comparability, sample information, and more. For the corresponding annual questionnaires, see the Questionnaires section of this site. For other data sets, see the SMART and BRFSS Maps (GIS) sections of this site.

2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000

1999	1998	1997	1996	1995	1994	1993	1992	1991	1990

1989	1988	1987	1986	1985	1984

Recommended citation: Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, **[appropriate data year or years]**.

Reference Documentation

BRFSS Weighting Formula
The computational formula and description of factors that can be taken into account when weighting a state's data.

Data Files

There are 451,075 records for 2010. More information on participation is available in the states conducting surveillance, by year table. The data files are provided in ASCII and SAS Transport formats. The May 25th update corrected one variable on five records in the Colorectal Cancer Screening section. The August 18th update corrects an issue with _FINALWT for AZ and DC.

2010 BRFSS Data (ASCII)
Data updated August 18, 2011

This file is in ASCII format. It has a fixed record length of 1518 positions. .zip 52MB

2010 BRFSS Data (SAS Transport Format)
Data updated August 18, 2011

This file was exported from SAS V8.2 in the XPT transport format. This file contains 397 variables. This format can be imported into SPSS or STATA. Please note: some of the variable labels get truncated in the process of converting to the XPT format so they may be slightly different from what is on the SASOUT10.SAS program. .zip 96MB

Variable Layout

Format information on variable name by column position. Web Page

---

2010 BRFSS Multiple Questionnaires

The multiple questionnaire data sets contain the data from the states which conducted dual questionnaires and used optional modules in 2010. Web Page

---

Data in the wild

In this class you will have an opportunity to collect data from various sources -- You are going to have to understand something of the formats these data are published in and how easy or hard it may be to work with these formats

When an organization like the CDC publishes data, they are, in effect, broadcasting the use cases they have in mind for the data -- Here we see the BRFSS is published in a SAS format (an established data analysis package) that can be read into SPSS and STATA (two other packages)

Interestingly, different software packages are often popular with different research communities -- SPSS (Statistical Package for the Social Sciences) finds traction in, well, sociology and economics

Of course, we can read the SAS format into R as well! To make your life easier, we have created a smaller subset of the BRFSS data and have made it available in a special binary format for R as well as in CSV (comma separated values) for those of you who might try some other software system

Our data

The BRFSS consists of responses from **hundreds of thousands of people** -- In this lecture and in your first homework, we will look at a subset of 20 thousand people (figuring that was big enough to be fun but small enough to avoid headaches)

Here are the first ten responses in our data set -- Each row refers to **a different person** and each column to **their response to a given question**

```
> head(cdc,n=25)
```

	state	genhlth	physhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender
1	Louisiana	good	0	0	1	0	70	175	175	77	m
2	Massachusetts	good	30	0	1	1	64	125	115	33	f
3	California	good	2	1	1	1	60	105	105	49	f
4	California	good	0	1	1	0	66	132	124	42	f
5	Ohio	very good	0	0	1	0	61	150	130	55	f
6	Pennsylvania	very good	0	1	1	0	64	114	114	55	f
7	California	very good	0	1	1	0	71	194	185	31	m
8	Texas	very good	1	0	1	0	67	170	160	45	m
9	California	good	2	0	1	1	65	150	130	27	f
10	Texas	good	3	1	1	0	70	180	170	44	m
11	Connecticut	excellent	4	1	1	1	69	186	175	46	m
12	Michigan	fair	30	1	1	1	69	168	148	62	m
13	New Mexico	excellent	0	1	0	1	66	185	220	21	m
14	Massachusetts	excellent	0	1	1	1	70	170	170	69	m
15	New Jersey	fair	3	1	0	0	69	170	170	23	m
16	California	good	0	1	1	1	73	185	175	79	m
17	California	good	0	0	0	1	67	156	150	47	m
18	Nebraska	fair	30	0	1	1	71	185	185	76	m
19	Alabama	good	0	1	1	1	75	200	190	43	m
20	Nevada	very good	0	1	1	0	67	125	120	33	f
21	Connecticut	very good	0	1	1	0	69	200	150	48	f
22	Washington	good	0	1	1	1	65	160	140	54	f
23	New York	very good	0	0	1	1	73	160	160	43	m
24	Ohio	good	2	1	1	1	67	165	158	30	m
25	New York	very good	0	0	0	1	64	105	120	27	f

Variables

Below we list off the variables we have included in this reduced data set -- Keep in mind that the original survey has literally hundreds of questions!

`state` where the respondent lives

`height` in inches

`weight` in pounds

`wtdesire` desired weight in pounds

`age` in years

`gender` labeled “f” and “m”

Variables

genhlth

Respondents were asked to evaluate their general health with choices being excellent, very good, good, fair and poor

physhlth

The number of days out of the last 30 that the respondent was in poor health

exerany

1 if the respondent exercised in the last month and 0 otherwise

hlthplan

1 if the respondent has some form of health coverage and 0 otherwise

smoke100

1 if the respondent has smoked at least 100 cigarettes in their entire life and 0 else

```
> head(cdc,n=25)
```

	state	genhlth	physhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender
1	Louisiana	good	0	0	1	0	70	175	175	77	m
2	Massachusetts	good	30	0	1	1	64	125	115	33	f
3	California	good	2	1	1	1	60	105	105	49	f
4	California	good	0	1	1	0	66	132	124	42	f
5	Ohio	very good	0	0	1	0	61	150	130	55	f
6	Pennsylvania	very good	0	1	1	0	64	114	114	55	f
7	California	very good	0	1	1	0	71	194	185	31	m
8	Texas	very good	1	0	1	0	67	170	160	45	m
9	California	good	2	0	1	1	65	150	130	27	f
10	Texas	good	3	1	1	0	70	180	170	44	m
11	Connecticut	excellent	4	1	1	1	69	186	175	46	m
12	Michigan	fair	30	1	1	1	69	168	148	62	m
13	New Mexico	excellent	0	1	0	1	66	185	220	21	m
14	Massachusetts	excellent	0	1	1	1	70	170	170	69	m
15	New Jersey	fair	3	1	0	0	69	170	170	23	m
16	California	good	0	1	1	1	73	185	175	79	m
17	California	good	0	0	0	1	67	156	150	47	m
18	Nebraska	fair	30	0	1	1	71	185	185	76	m
19	Alabama	good	0	1	1	1	75	200	190	43	m
20	Nevada	very good	0	1	1	0	67	125	120	33	f
21	Connecticut	very good	0	1	1	0	69	200	150	48	f
22	Washington	good	0	1	1	1	65	160	140	54	f
23	New York	very good	0	0	1	1	73	160	160	43	m
24	Ohio	good	2	1	1	1	67	165	158	30	m
25	New York	very good	0	0	0	1	64	105	120	27	f

Variables

A variable is a characteristic of a person or thing that can be assigned a number or a category -- We often distinguish between two kinds of variables

Qualitative data “arise when individuals may fall into separate” categories which may not have a numerical relationship (gender and state in the BRFSS) -- Qualitative data may be ordinal in the sense that there is a natural order to the categories (general health in the BRFSS)

Quantitative data, on the other hand, are numerical, “arising from counts and measurements” -- These data can, in turn, be loosely described as being continuous (able to take any number in some range) or discrete (integers, say, or numerical values with just a small number of unique entries)

We go through the effort of noting these differences because they often inform the **kinds of summaries or displays** that are appropriate for a variable and will influence the way they are **used in models** later in the quarter

Frequency graphs

A frequency distribution is a display of the frequency (or count) of all the values in a sample -- It is often tabular or graphical

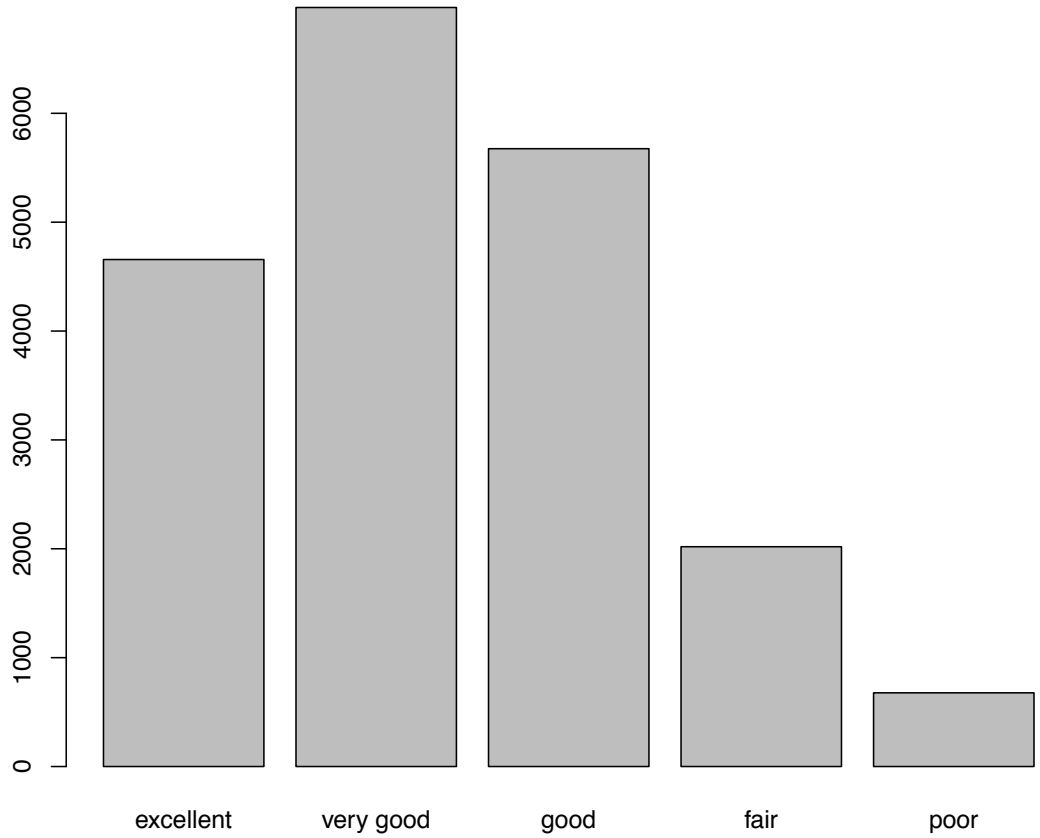
For qualitative or discrete quantitative variables, this idea makes sense -- For continuous variables we consider grouping the data in some way first

Here (next slide) are tables exhibiting the frequencies of some of the qualitative variables recorded for the BRFSS respondents in our data set

```
> table(cdc$gender)

    m     f
 9569 10431

> table(cdc$exerany)

    0     1
 5086 14914

> table(cdc$smoke100)

    0     1
10559  9441

> table(cdc$genhlth)

excellent very good      good      fair      poor
     4657      6972      5675      2019       677

> barplot(table(cdc$genhlth))
```

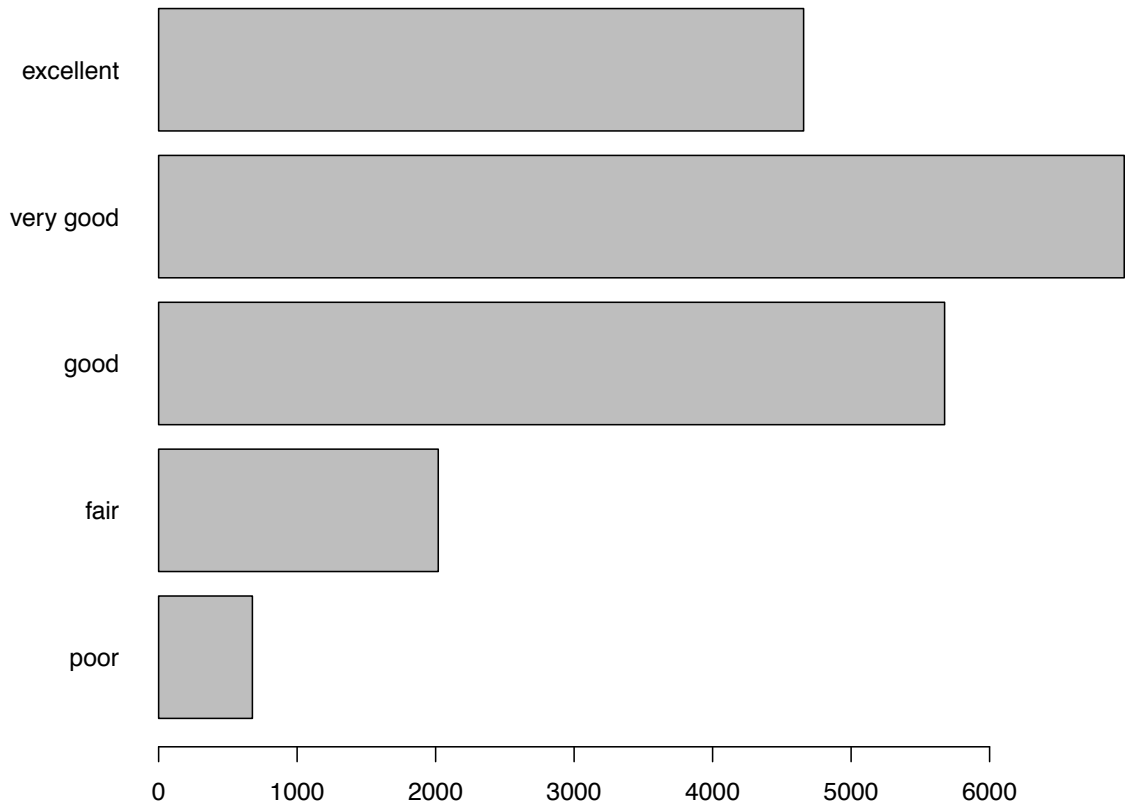
Graphical displays

A **barplot** can be formed to make comparisons easier

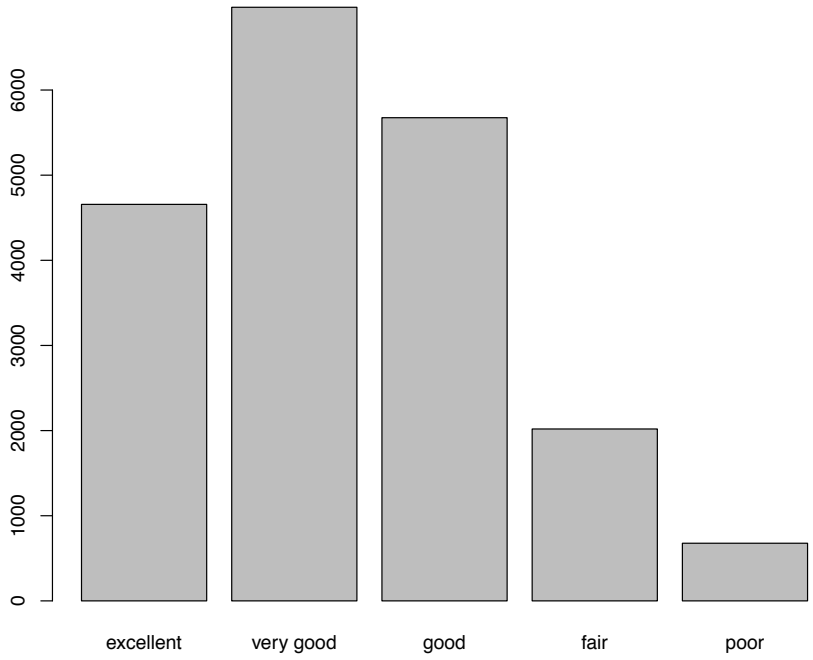


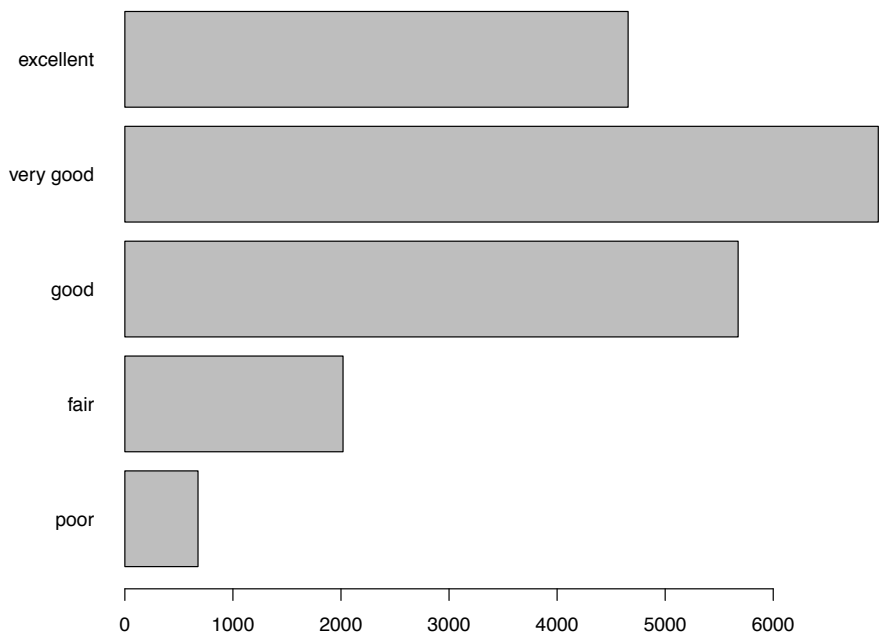
Graphical displays

Some have argued that comparisons are better made when the bars run horizontally*



* Cleveland, W. S. (1993), Visualizing Data, Hobart Press





What do you think?

Questions

While one-dimensional summaries are interesting, what more would we like to see from the data? Consider the variable list again...

```
state

height weight wtdesire

age gender

genhlth physhlth exerany

hlthplan smoke100
```

Questions

One-dimensional summaries can't address some basic questions we might have of the data -- For example, do respondents who exercise report being in better overall health than those who don't?

For this, we might consider a two-way table (also referred to as a contingency table) that reports the frequency in each pair of categories -- What do you notice?

```
> table(cdc$exerany,cdc$genhlth)

   poor fair good very good excellent
0   384  857 1731      1352       762
1   293 1162 3944      5620      3895
```

Admittedly, it can be hard to "see" what's going on here directly (although we can certainly improve the good old tabular display considerably) -- What might we try graphically here?

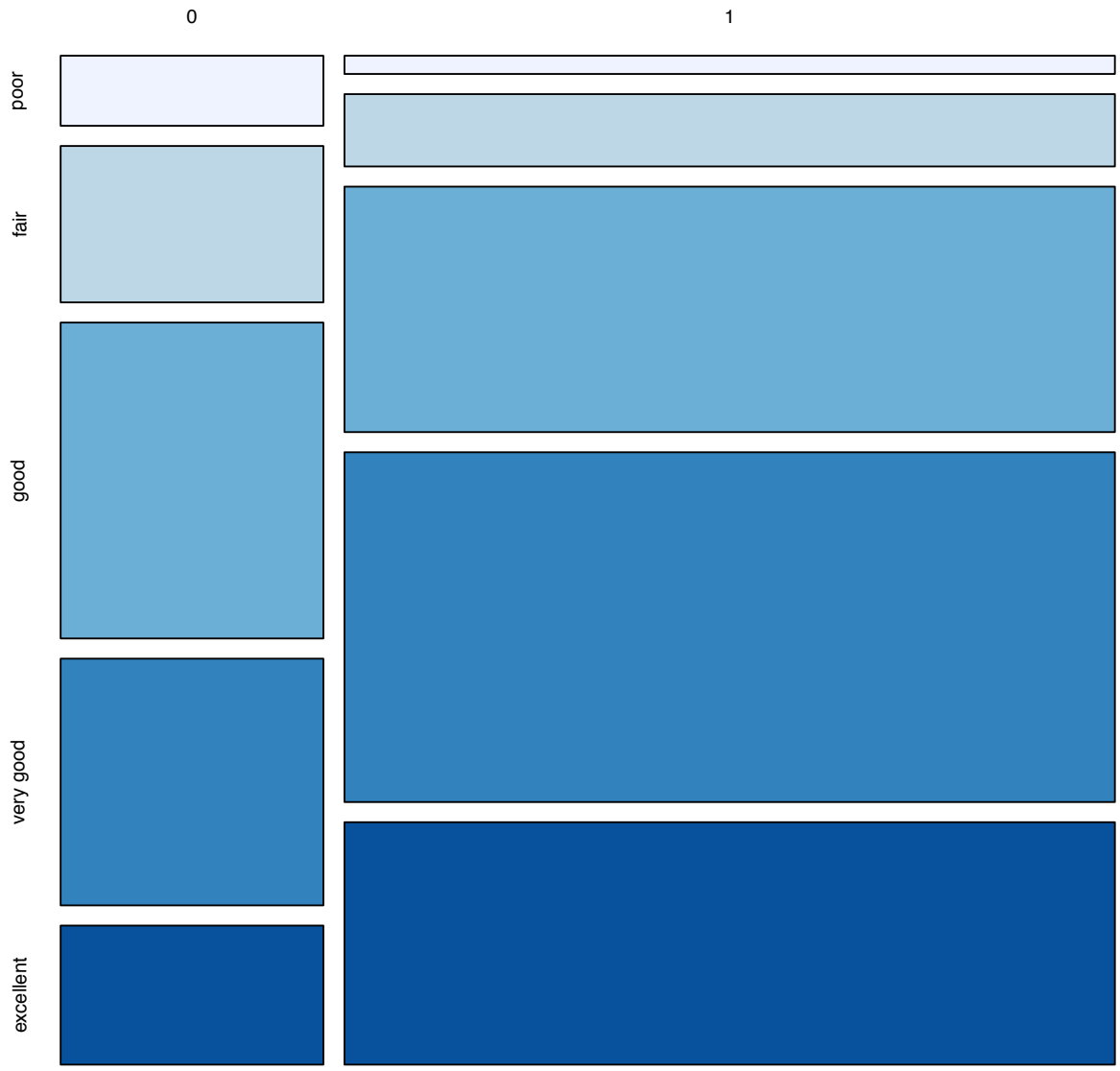
Mosaic plots

These displays represent the counts in a contingency table by tiles whose size (area) is proportional to the cell count or frequency*

```
# and now the plot -- we pass two arguments, the table
# we want to visualize and a title for the plot

> mosaicplot(table(cdc$exerany,cdc$genhlth),
             main="Exercise and general health")
```

It is also possible to extend these displays to tabulations with more than two variables -- How might this work?

* Hartigan, J.A., and Kleiner, B. (1984) A mosaic of television ratings. *The American Statistician*, 38, 32-35


Exercise and general health
0
1
poor
fair
good
very good
excellent

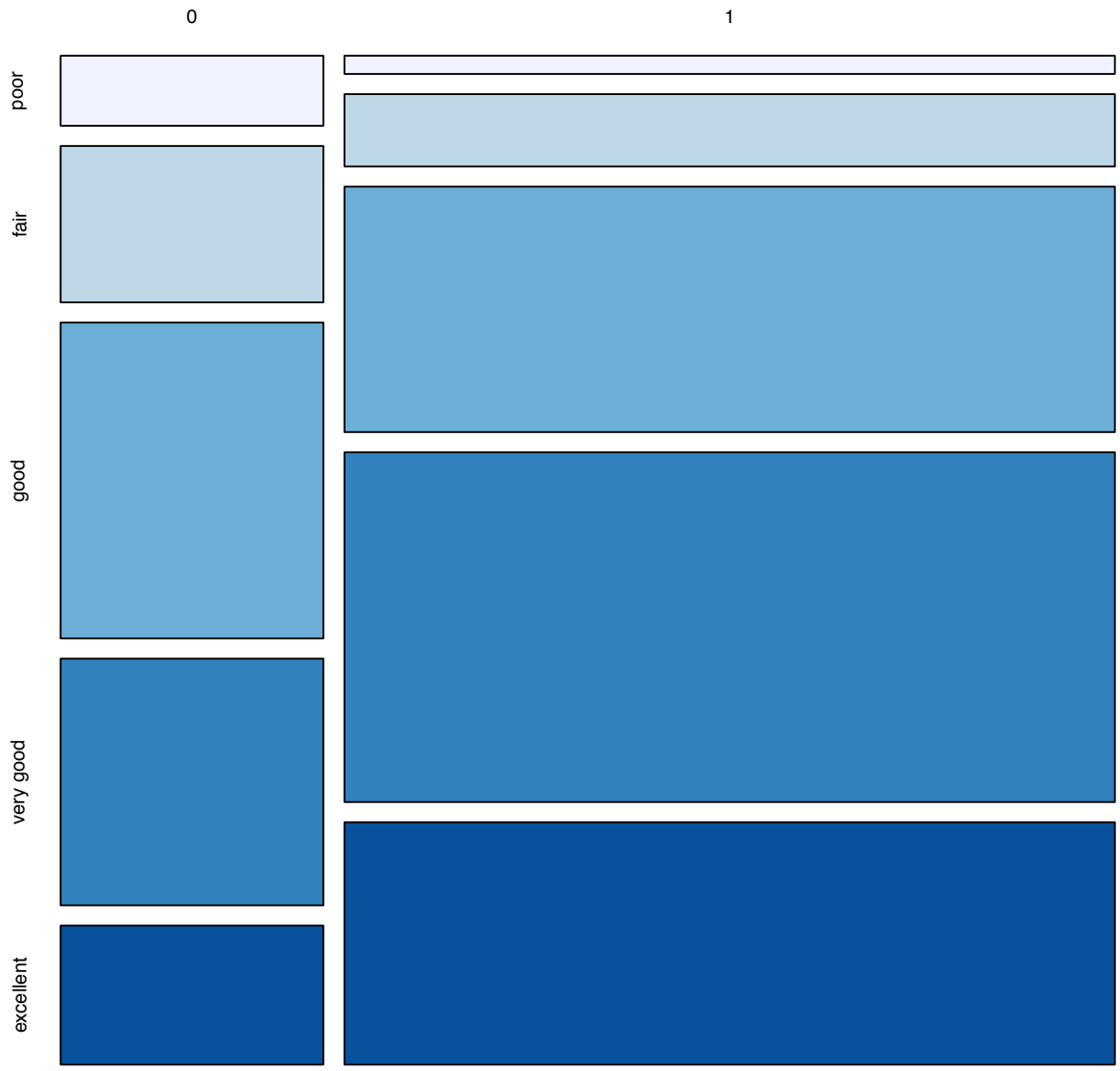
Mosaic plots

As a kind of multidimensional barplot, the mosaic plot makes the height and width of the boxes proportional to the counts in the corresponding cell of the table -- It does this in two passes

In our case, the width of the boxes are chosen proportional to the counts of respondents who said they had or had not exercised in the last 30 days (about 75% had)

	excellent	very good	good	fair	poor	
0	762	1352	1731	857	384	5086
1	3895	5620	3944	1162	293	14914
	4657	6972	5675	2019	677	

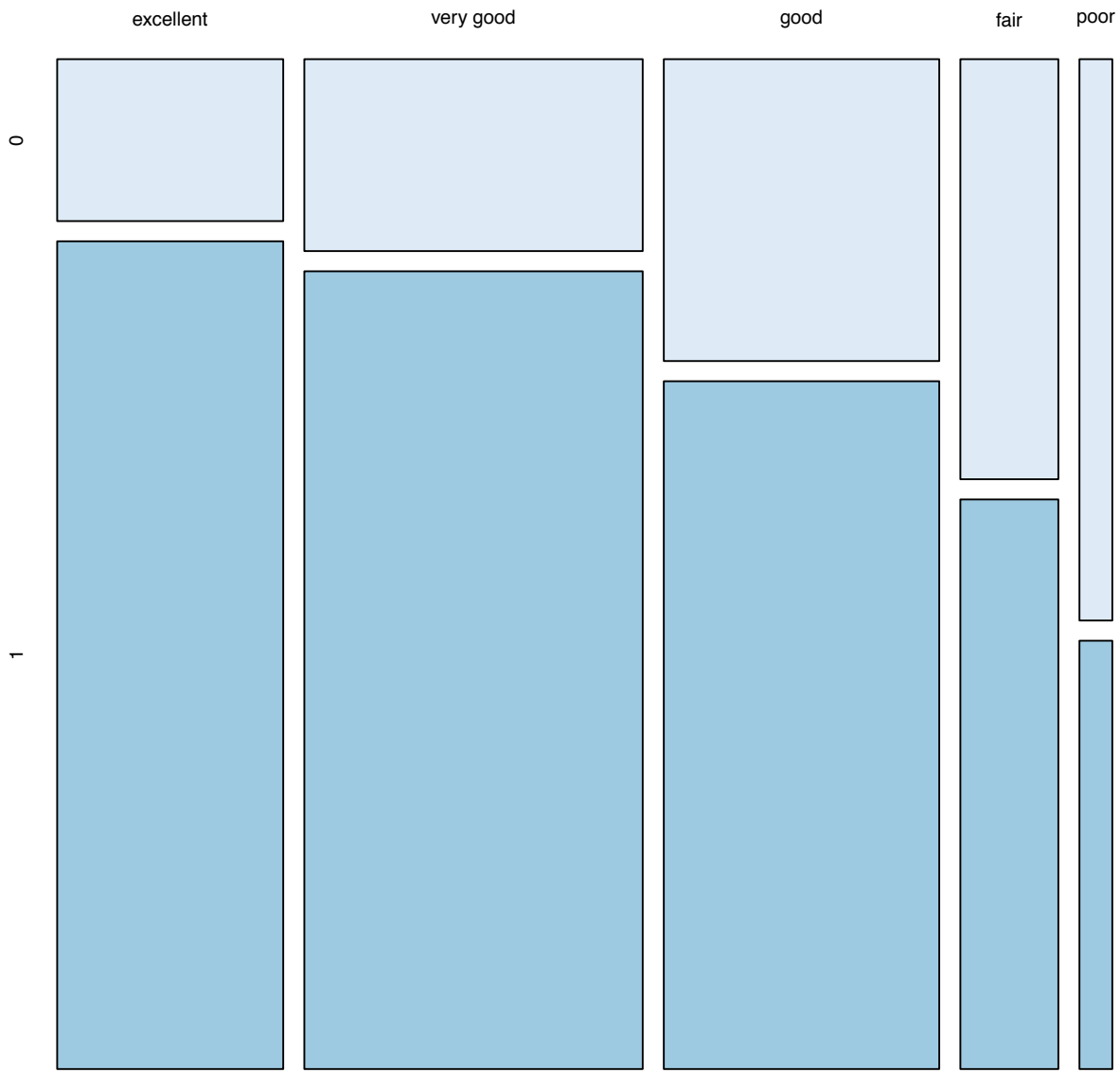
Then, within each category (0 or 1), we adjust the height of each box so that it is proportional to the counts of respondents who were in excellent, very good, good, fair and poor health


Exercise and general health
0
1
poor
fair
good
very good
excellent

Mosaic plots

Given this two-pass construction, there are actually two mosaic plots that can be formed from a single table -- On the next page we “transpose” the original table

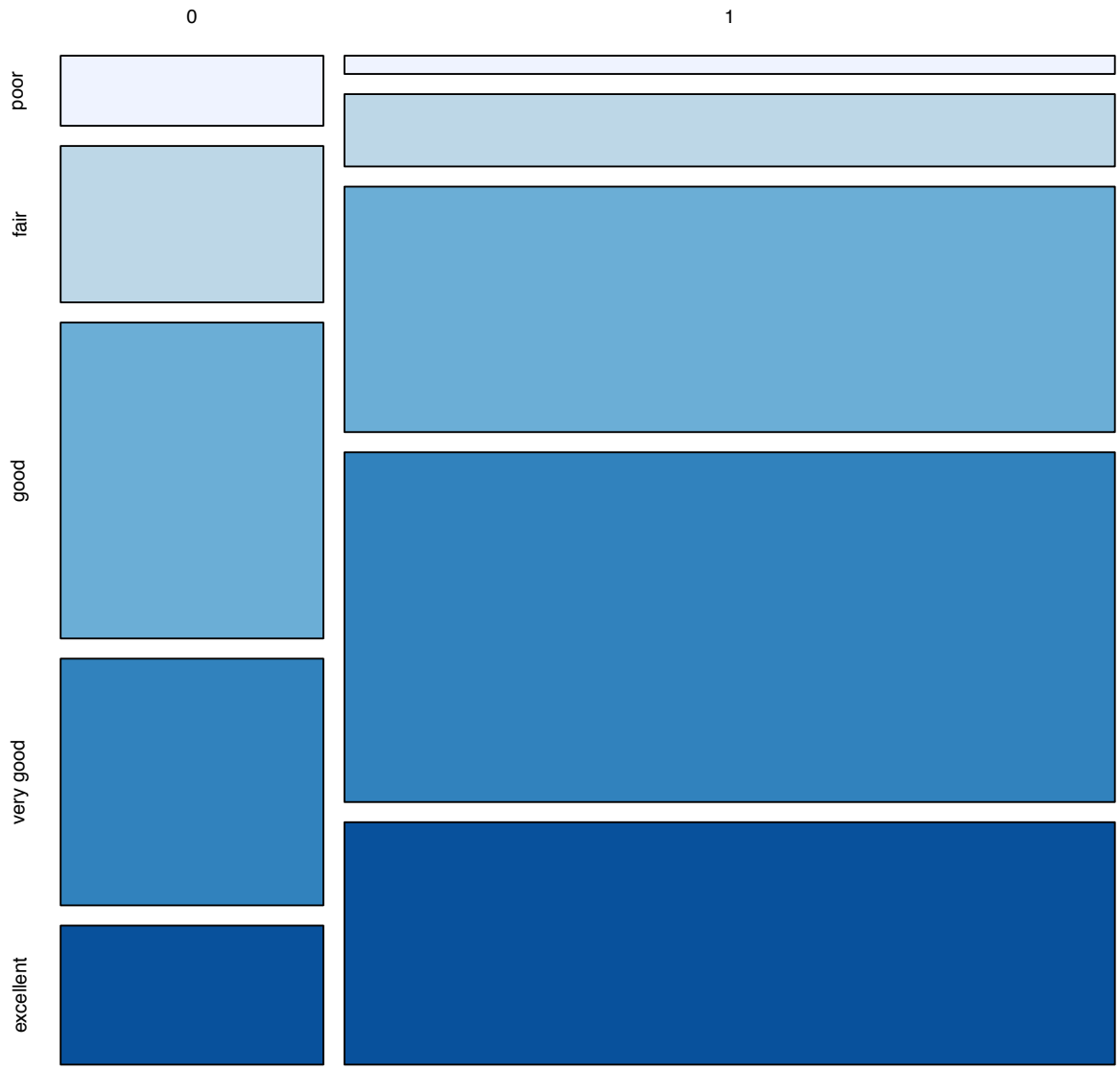
	0	1
excellent	762	3895
very good	1352	5620
good	1731	3944
fair	857	1162
poor	384	293

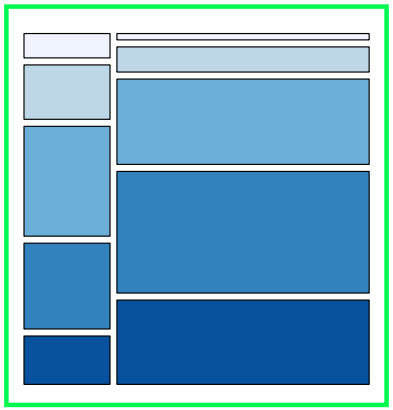
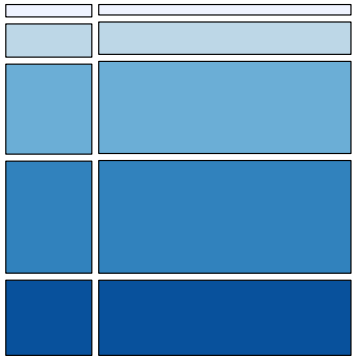
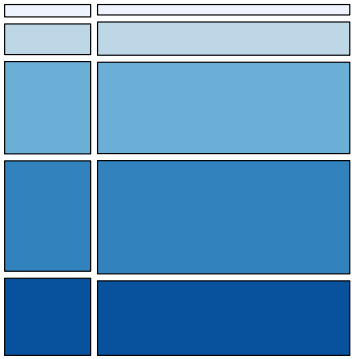
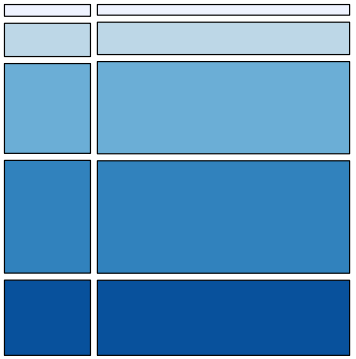
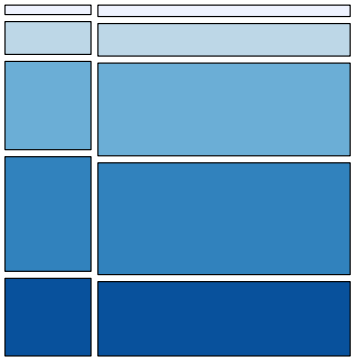
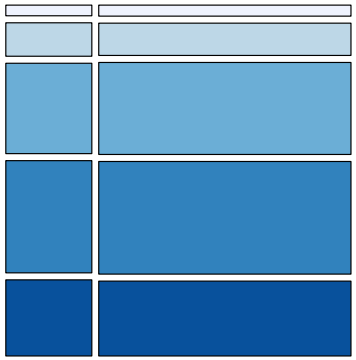
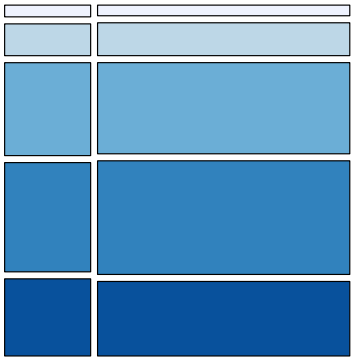
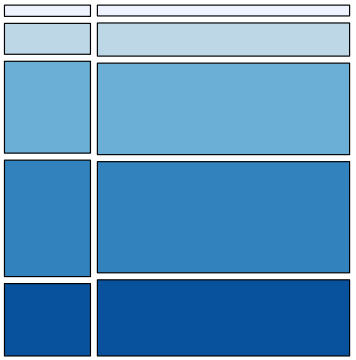
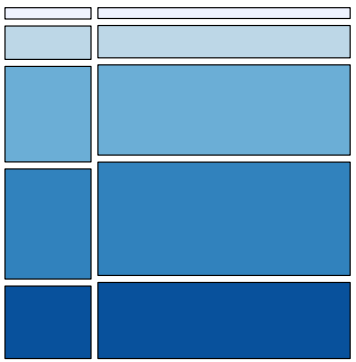
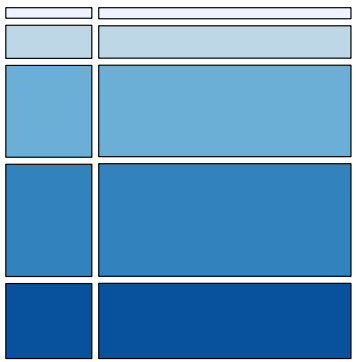
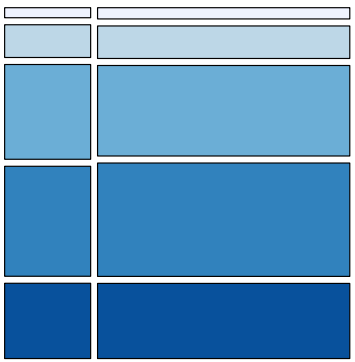
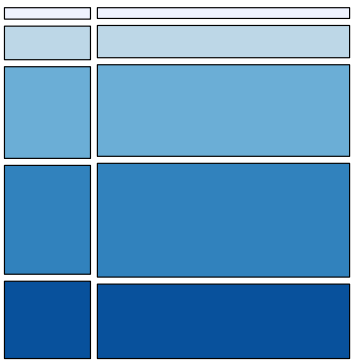

Exercise and general health
excellent
very good
good
fair
poor
0
1

Association or not?

There seems to be some association between exercise and general health -- That makes intuitive sense but you're relative newbs when it comes to "area" plots and it's worthwhile asking **what "no association" would look like**

Any ideas?


Exercise and general health
0
1
poor
fair
good
very good
excellent

Simulation from a null model

By permuting the data associated with each person, we are generating data that, by definition, **should exhibit no pattern of association** -- For the moment we are using this device to help us calibrate our eyes so that we understand the tool we're working with

Notice that **there isn't much variability in the plots that we simulated because our sample size is so large** that by permuting things, we tend to get pretty stable proportions in each category -- This will change dramatically with smaller sample sizes

In the next lecture, we'll see that plots like these can also be used as a kind of visual hypothesis test (sexy!)

Extensions?

What would we do if we wanted to assess a three-way association, say between general health, exercise and gender?

```
> table(cdc$genhlth,cdc$exerany,cdc$gender)

, ,  = m


              0    1
  excellent  335 1963
  very good  606 2776
  good       723 1999
  fair       340  544
  poor       145  138

, ,  = f


              0    1
  excellent  427 1932
  very good  746 2844
  good      1008 1945
  fair       517  618
  poor       239  155
```

Creating new variables

BMI (Body Mass Index) is defined to be

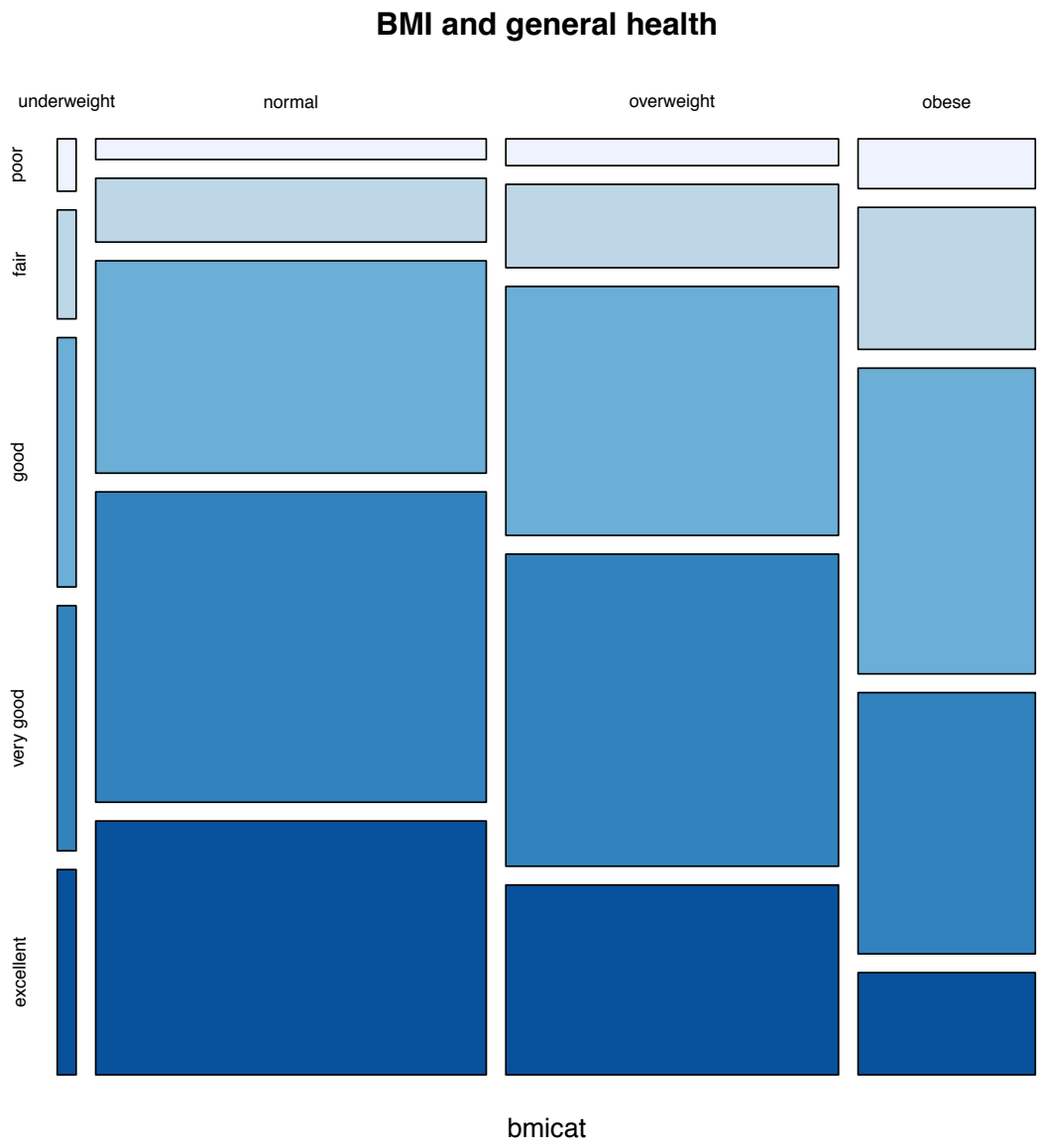
$$\text{BMI} = 703 \times \frac{\text{weight in pounds}}{(\text{height in inches})^2}$$

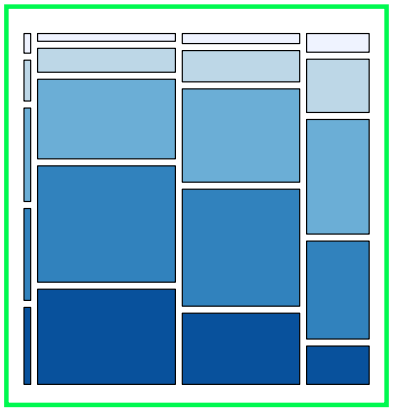
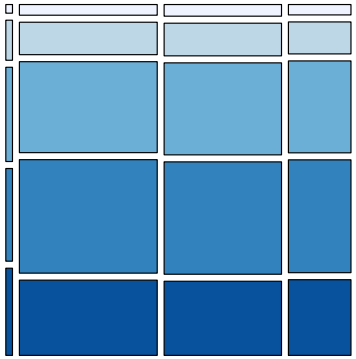
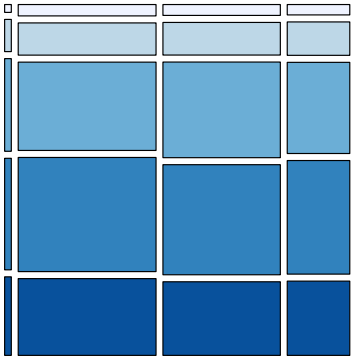
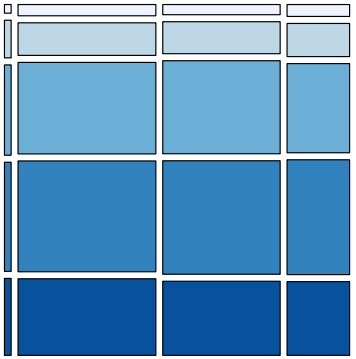
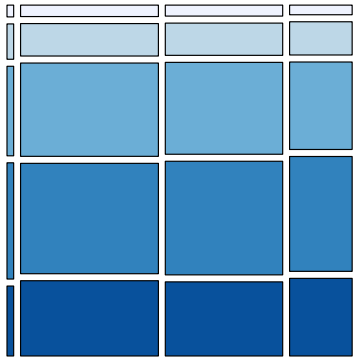
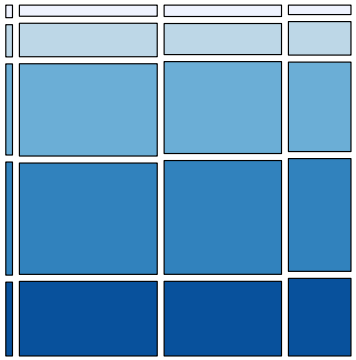
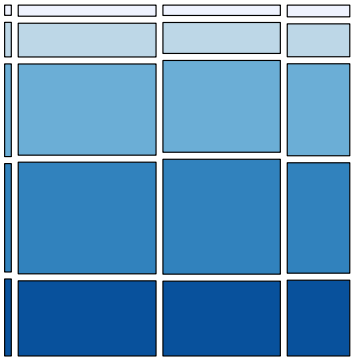
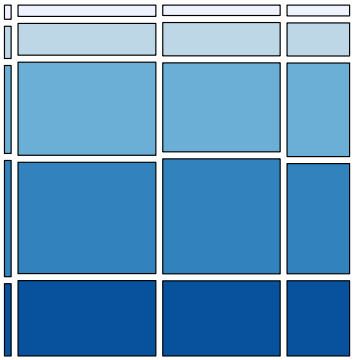
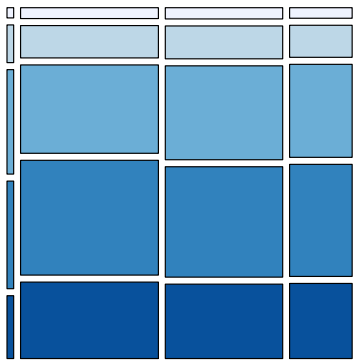
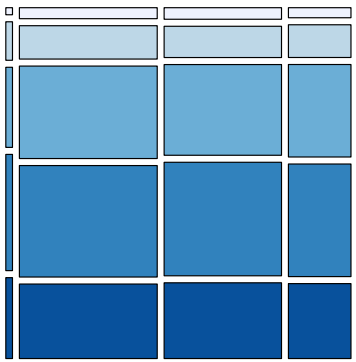
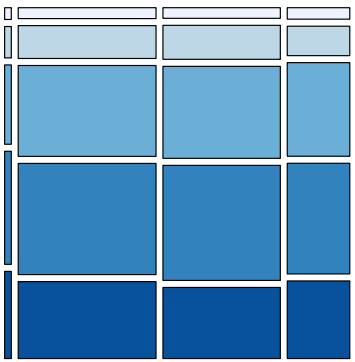
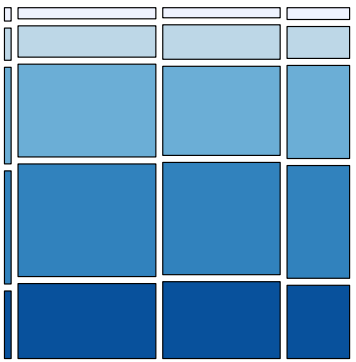
We can derive this from our data set and create a new quantitative variables

The CDC interprets these limits as follows

BMI	**Weight Status**
Below 18.5	Underweight
18.5 – 24.9	Normal
25.0 – 29.9	Overweight
30.0 and Above	Obese

```
> bmi <- 703*cdc$weight/(cdc$height)^2

# summary() is a handy function that will return a summary of the
# object you feed it -- Here we get a 6-number summary of the bmi values

> summary(bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.40   22.71   25.60   26.31   28.89   73.09

# now carve them into levels (this is advanced so ignore it on first reading)

> bmicat <- cut(bmi,c(0,18.5,25,30,100))
> levels(bmicat) <- c("underweight","normal","overweight","obese")

# and now summary gives you a table of the different category counts

> summary(bmicat)
underweight      normal  overweight       obese
        411        8496        7237        3856

# barplot!

> barplot(table(bmicat),horiz=T,main="BMI categories")

# and a mosaic plot!

> table(bmicat,cdc$genhlth)

bmicat        poor fair good very good excellent
  underweight   25   52  119       117        98
  normal       204  630 2095      3062      2505
  overweight   225  701 2092      2623      1596
  obese        223  636 1369      1170       458

> mosaicplot(table(bmicat,cdc$genhlth), main="BMI and general health")
```


BMI and general health
underweight
normal
overweight
obese
poor
fair
good
very good
excellent
bmicat

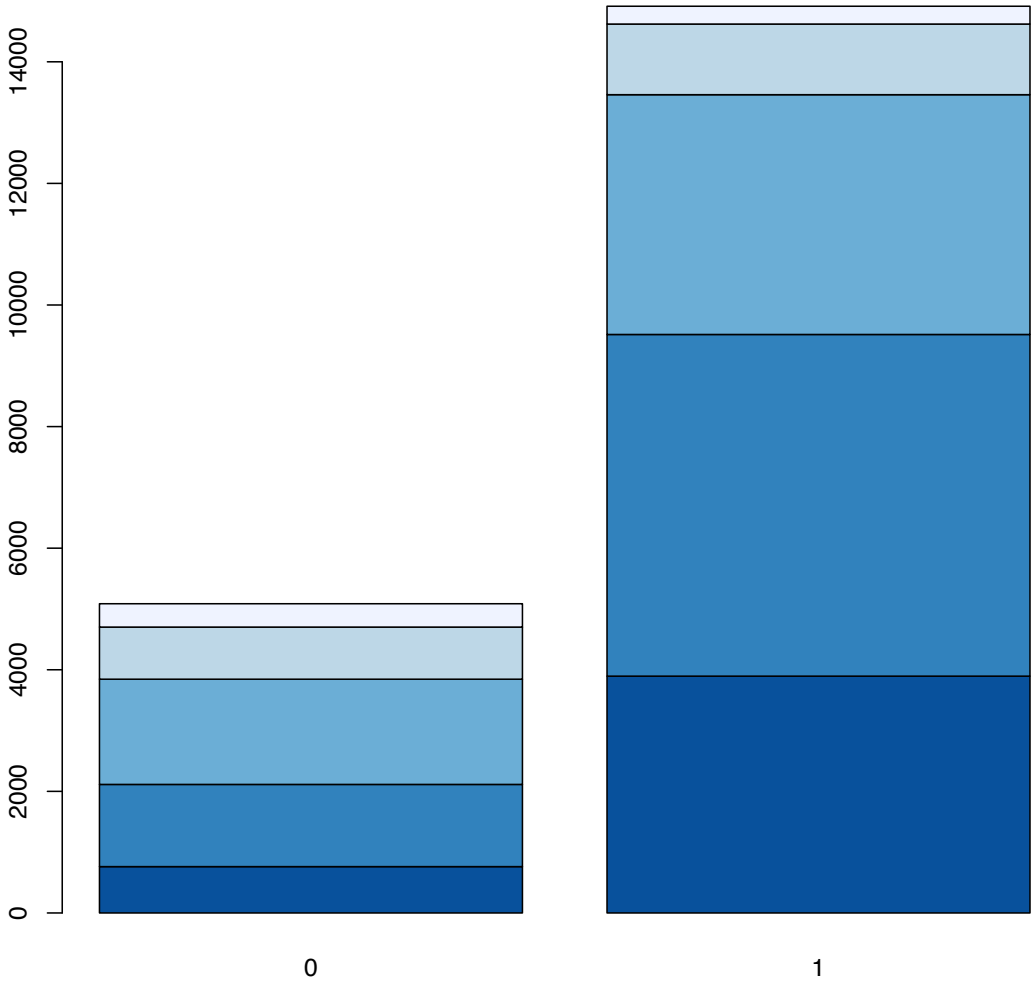













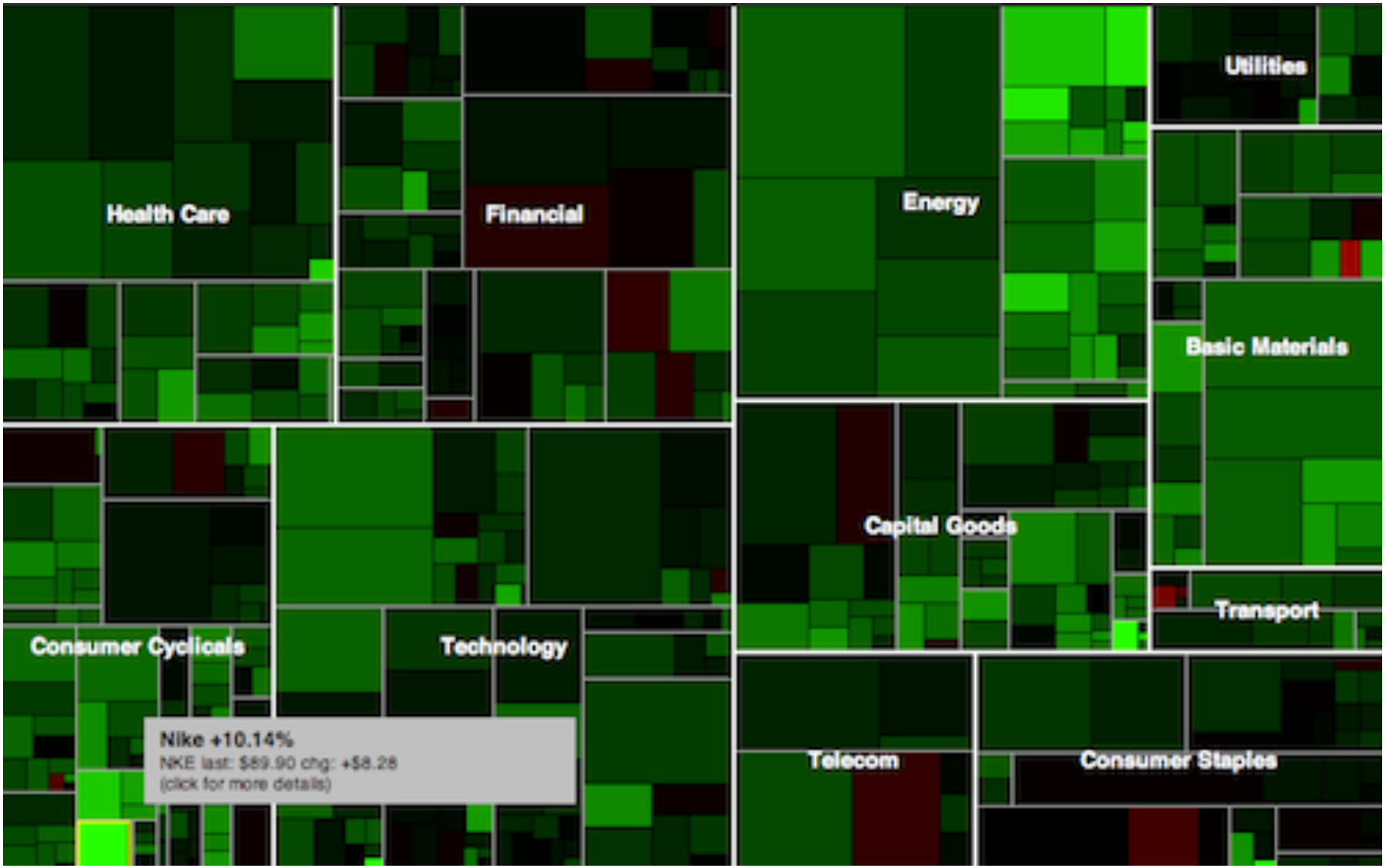
Mosaic plots and beyond

The essential features of a mosaic plot (representing counts of disjoint events with rectangles having areas proportional to the associated count) can be orchestrated in a variety of different ways

Stacked barplots (next page) are maybe more traditional and then there's the more recent treemaps -- Hadley Wickham* has a nice review article and associated R package

*Hadley Wickham, Heike Hofmann. Product plots. *IEEE Transactions on Visualization and Computer Graphics* (Proc. InfoVis '11), 2011.


Exercise and general health
0
2000
4000
6000
8000
10000
12000
14000
0
1


Health Care
Financial
Energy
Utilities
Basic Materials
Capital Goods
Transport
Consumer Cyclicals
Technology
Telecom
Consumer Staples
Nike +10.14%
NKE last: $89.90 chg: +$8.28
(click for more details)

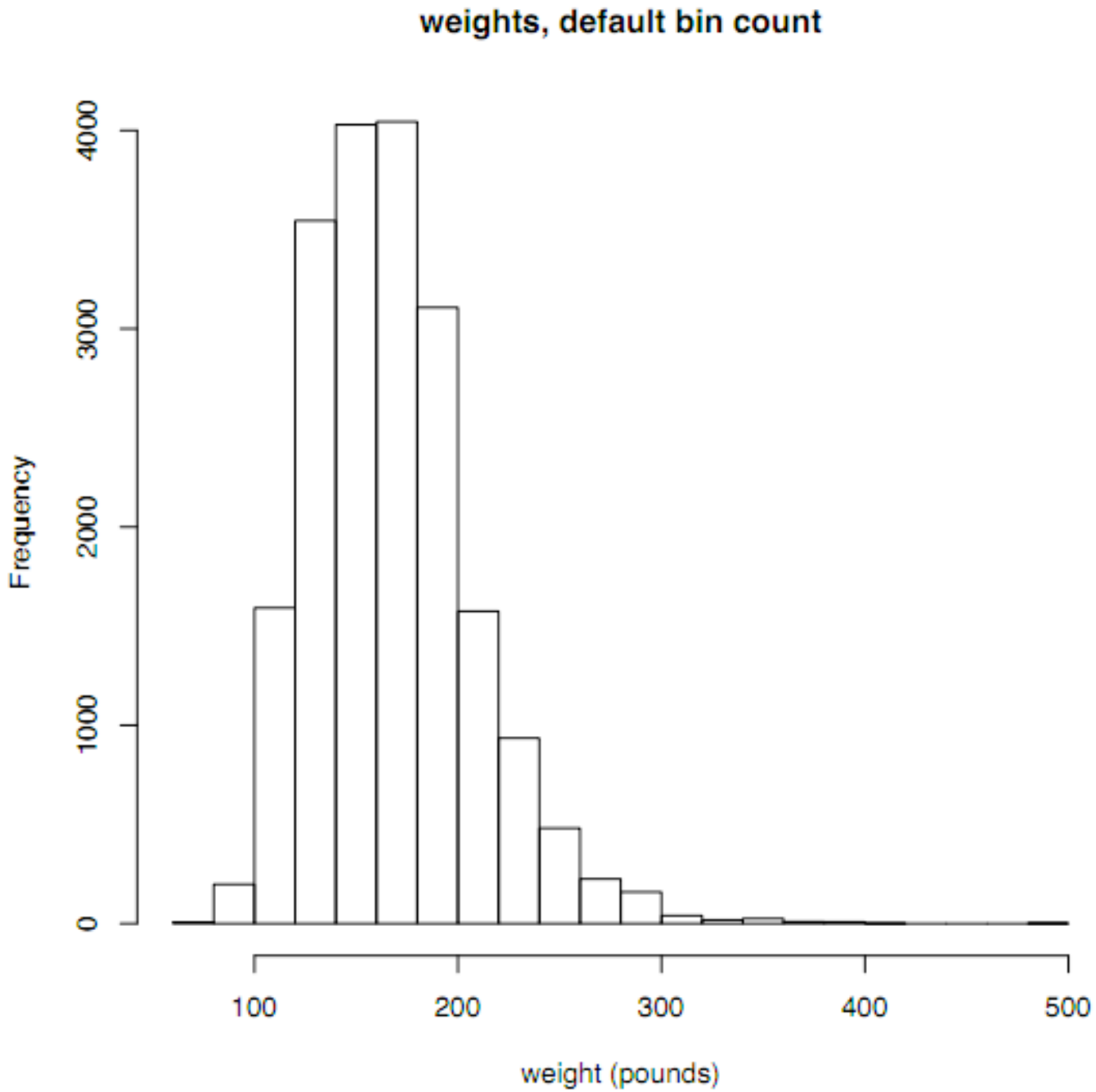
Histograms

A histogram groups or bins the data and, like a barplot, presents the number of data points that fall into each group

The display involves a “tuning parameter” -- That is, we are free to choose how many bins we want to use in the display (of course statisticians have thought for decades about how to choose this parameter in a more automatic fashion)

In situations like this, it is always good to vary the number of bins and examine the plot for any structure that emerges -- In doing so, we want to get a sense of the “shape” of the data

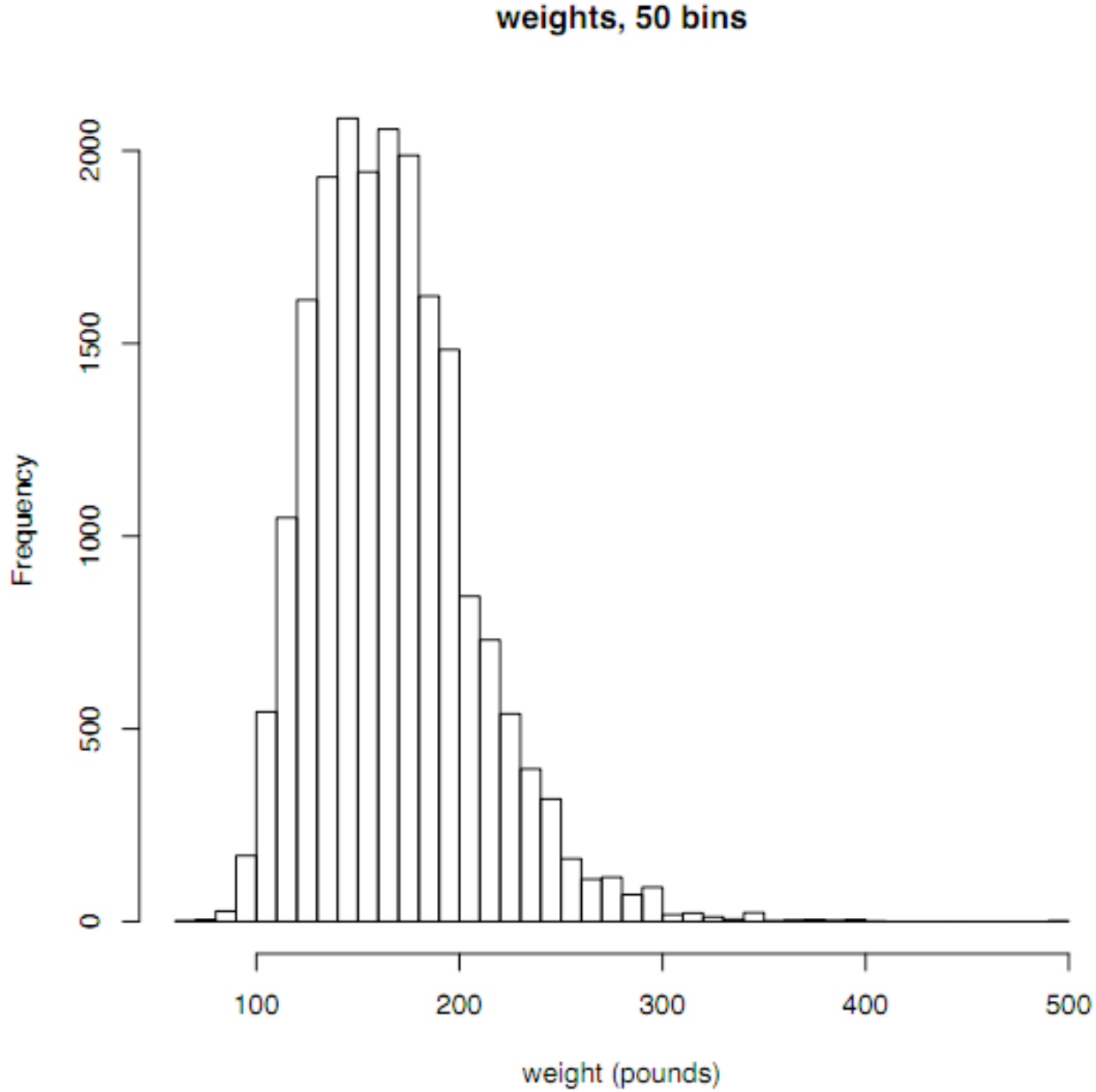
```
> hist(cdc$weight)
```

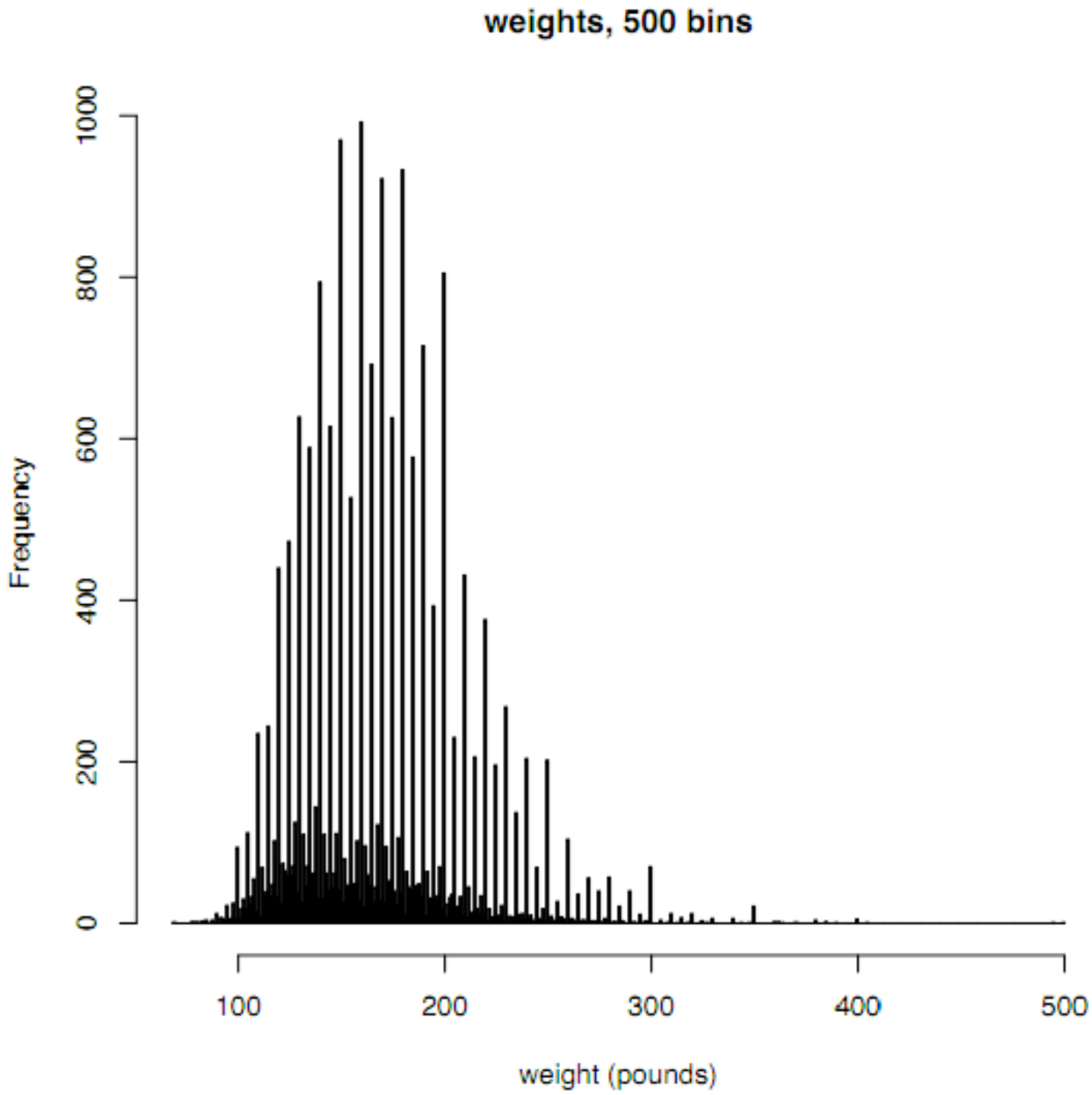

weights, default bin count
Frequency
0
1000
2000
3000
4000
100
200
300
400
500
weight (pounds)

And, varying bin widths...

With R, varying the size of our bins is as easy as

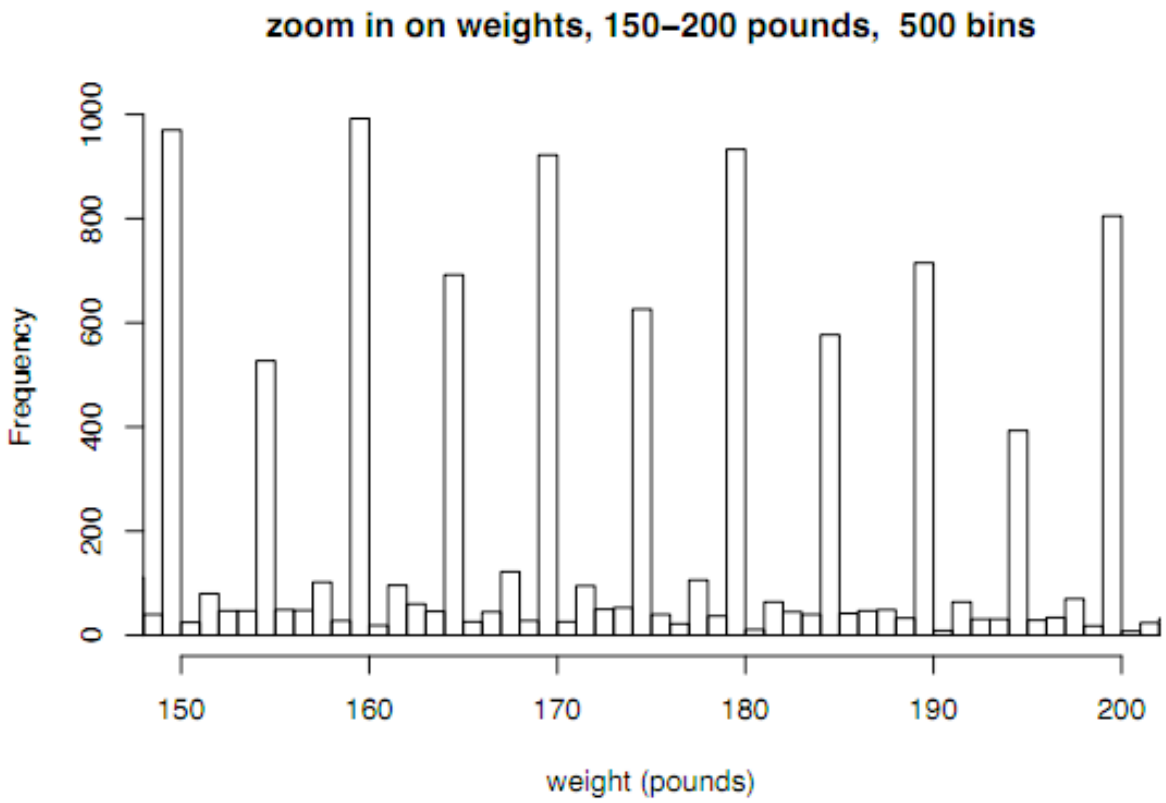
```
> hist(cdc$weight,breaks=50)
> hist(cdc$weight,breaks=500)
```


weights, 50 bins
Frequency
0
500
1000
1500
2000
100
200
300
400
500
weight (pounds)


weights, 500 bins
Frequency
0
200
400
600
800
1000
100
200
300
400
500
weight (pounds)

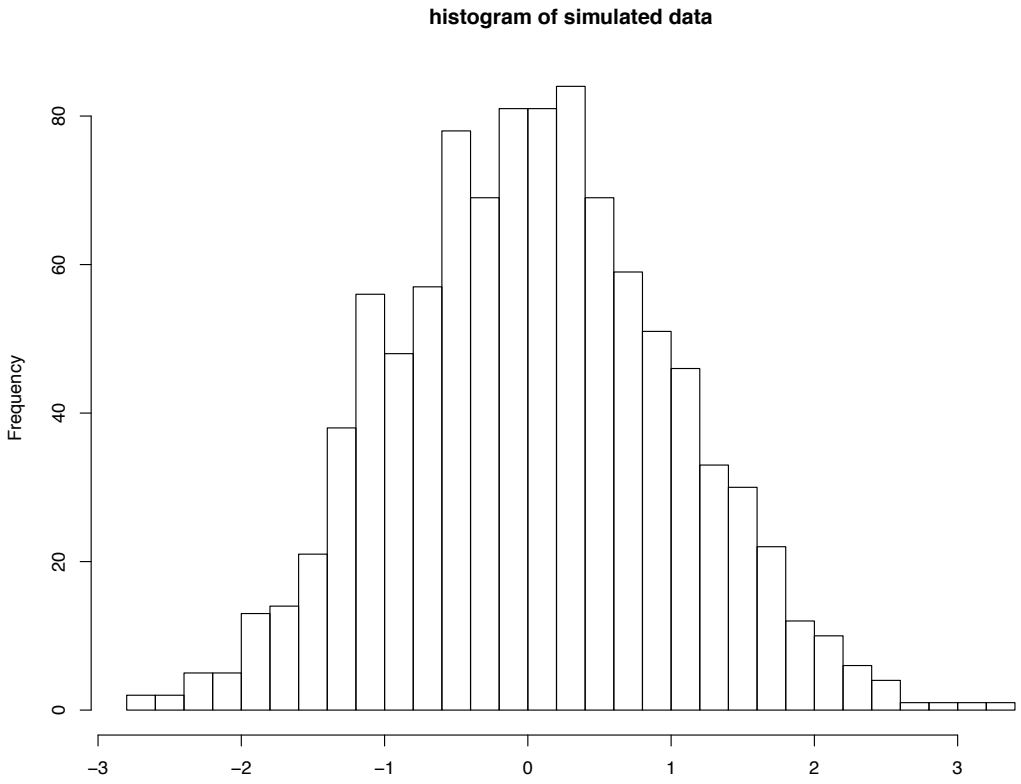
Varying bin sizes

By changing the bin size, we can uncover features in the data; in this case we uncover a basic fact about how people report their weights



A taxonomy of shapes

Over the course of this quarter, you will start to build up a vocabulary describing the shapes of distributions -- We refer to a distribution as symmetric if you can fold the histogram in half and have roughly the same shape on either side



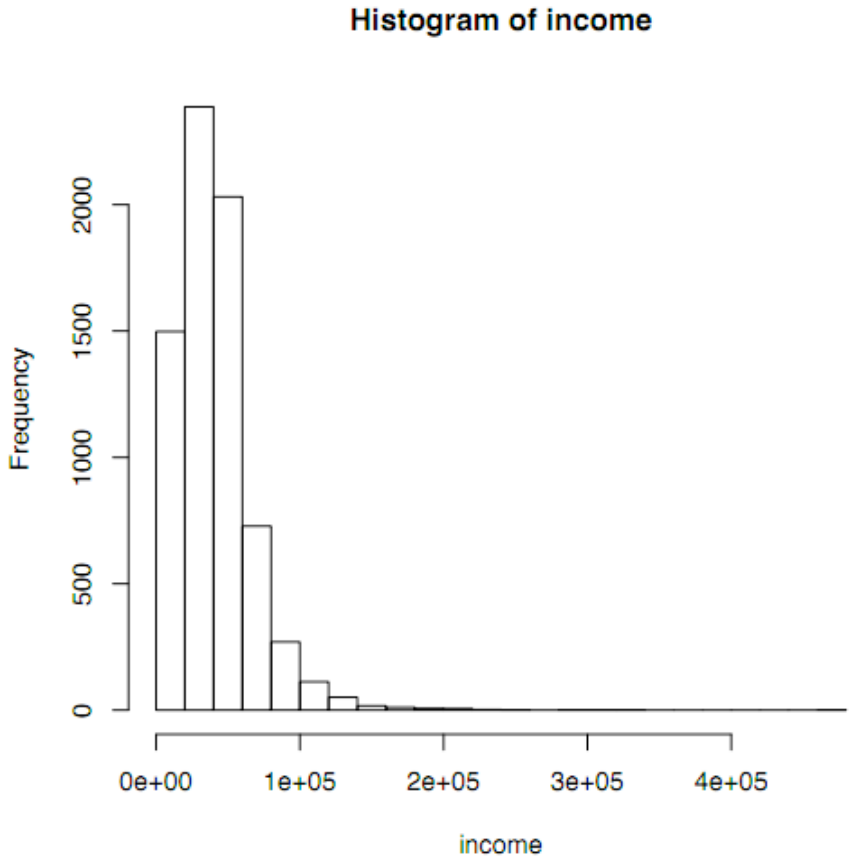
In the case of weight data, the distribution is said to be **skewed to the right** because there are too many large values than we might expect if things were symmetric -- When the distribution is pushed in the other direction we say it is **skewed to the left**

Taxonomy of shapes: Multiple modes

For the 1983 Family Expenditure Survey in the UK, a sample of n=7,125 households reported their annual income*

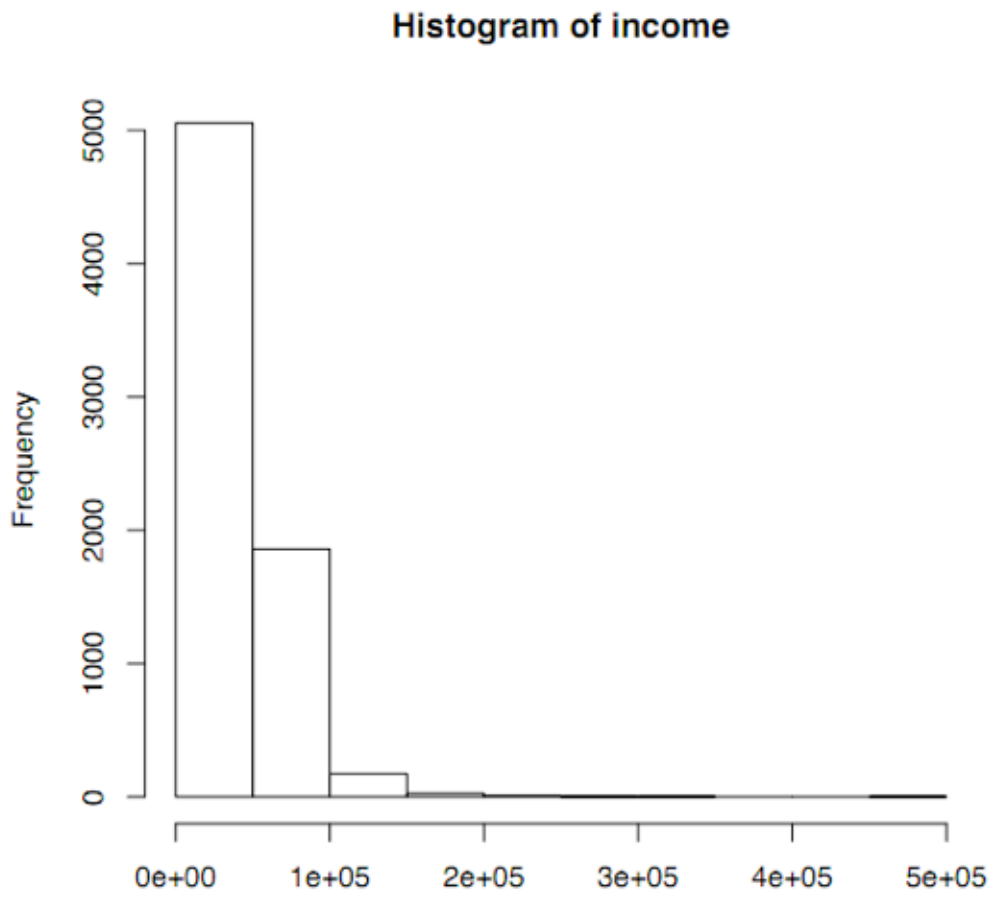
The minimum of these data is 529 (~$850), the maximum 472,821 (~$760K)

The mean of these data is 41,262 (roughly $66K) and the median is 37,520 ($60K)

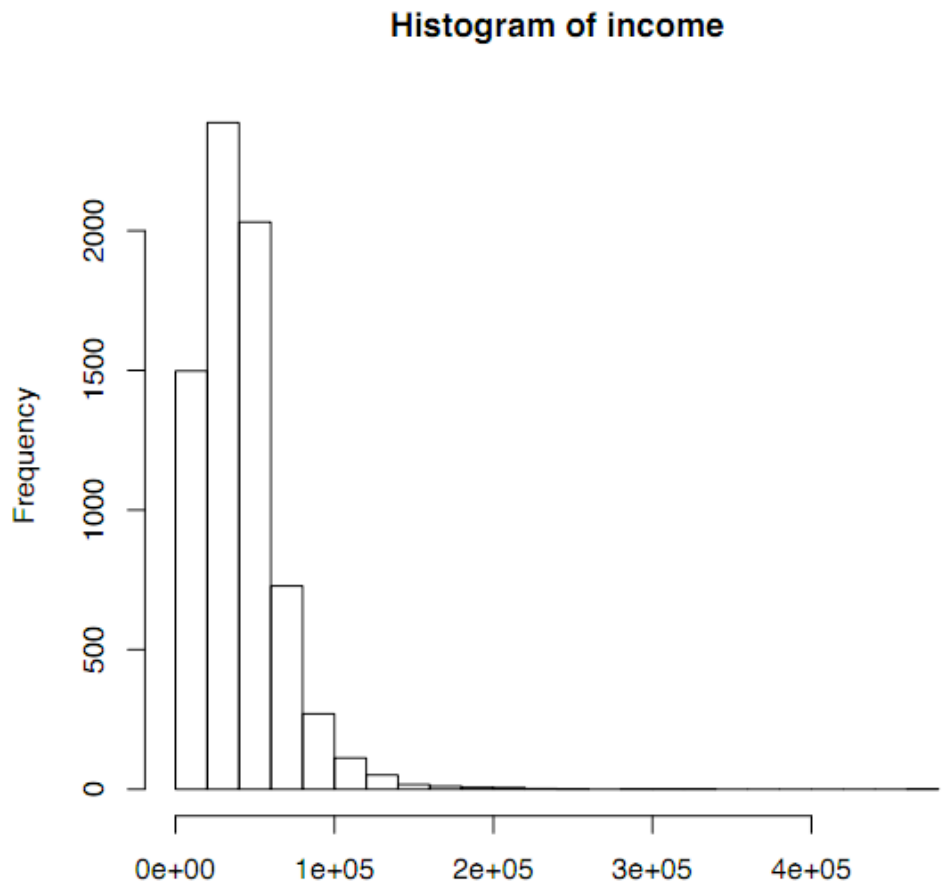


* Yeah, yeah, cheesy example, sorry.

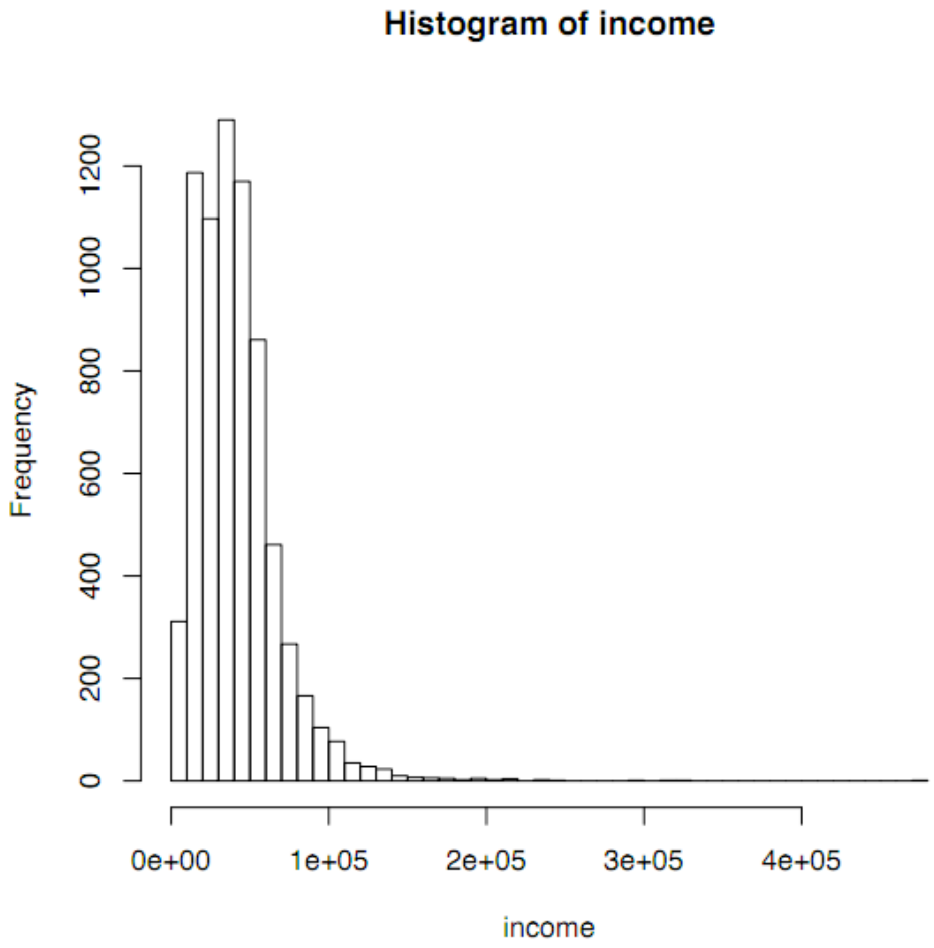
A comment on varying bin size... we see skew



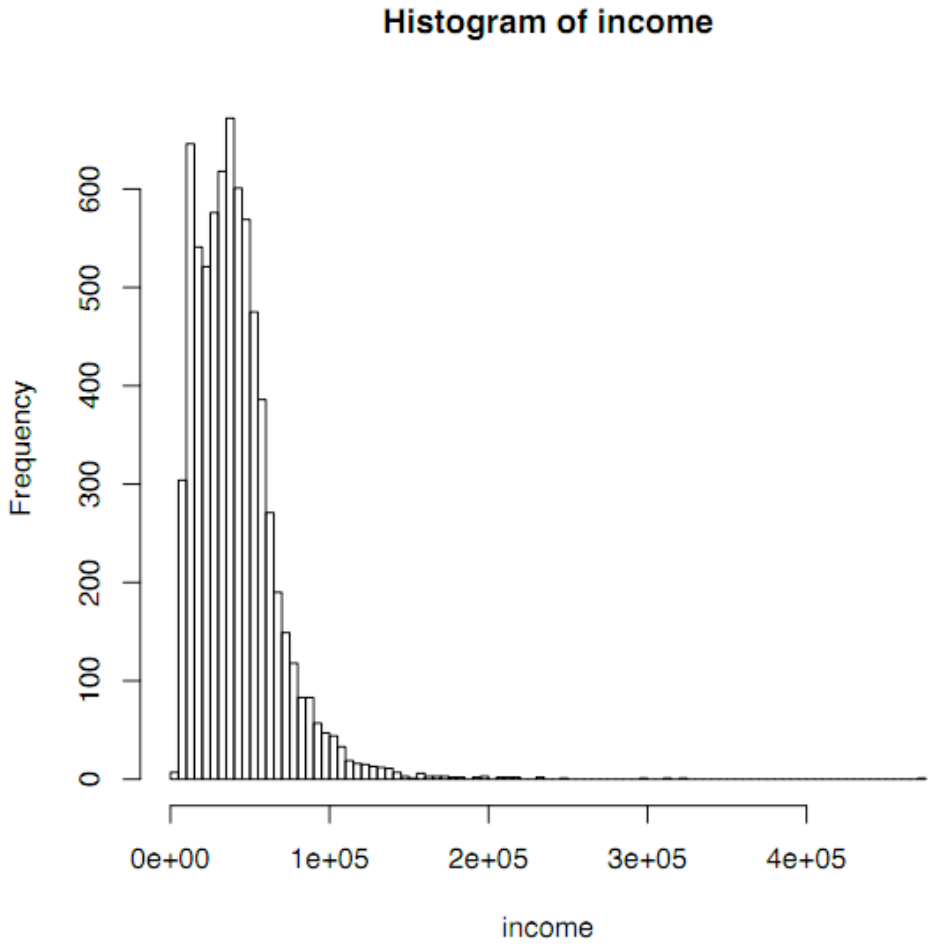
... and more skew



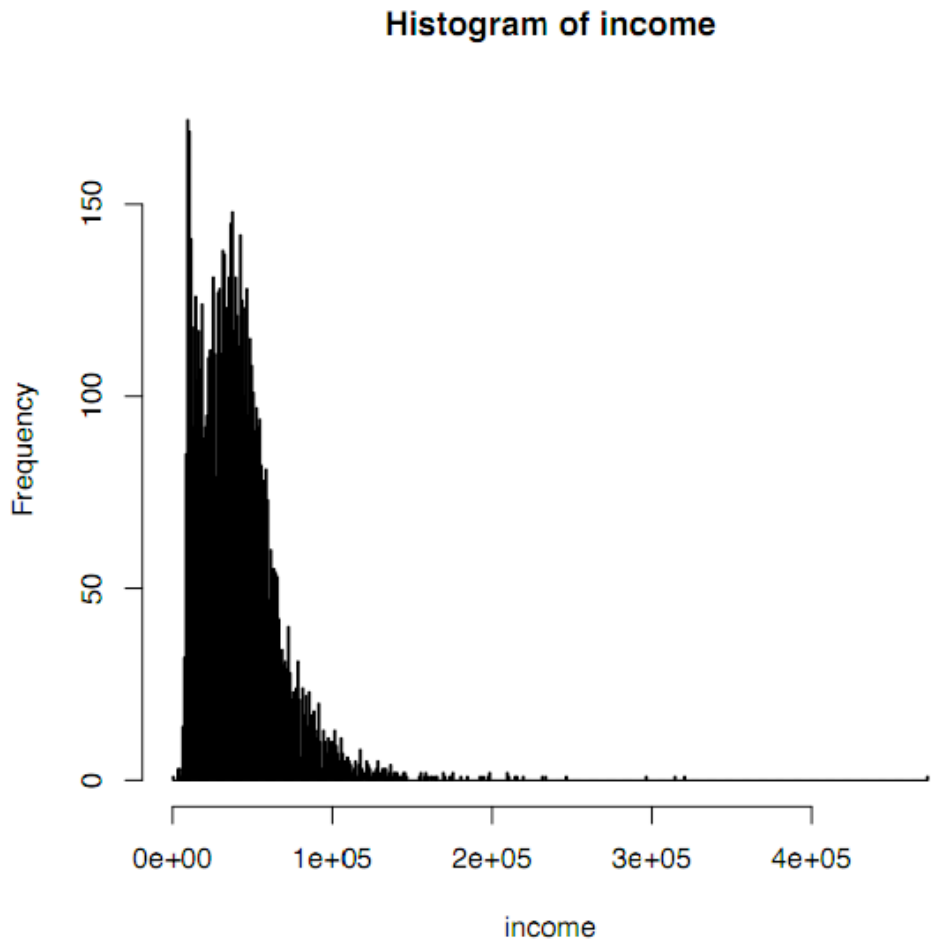
... but something starts to emerge at the left



... and more


Histogram of income
Frequency
0
100
200
300
400
500
600
0e+00
1e+05
2e+05
3e+05
4e+05
income

... and it's quite visible now; what might this be?



Taxonomy of shapes

We say that the income distribution has **two modes or peaks**

In categorical settings, the mode is defined to be the value of the variable receiving the most counts -- By analogy, we call peaks or regions of high concentration in continuous data, modes

Notice however, that the question of the number of modes is **a bit slippery** for continuous data; which bin width do you use?

Default bin size

It is often the case that we don't want to think very hard about how many bins or groups to use when drawing a histogram; the hist() function in R uses a rule of thumb for setting the number of bins based on our sample size

$$\text{number of bins} \approx \log_2(n) + 1$$

Where might a rule like this come from?

Default bin size

At the moment, we are using histograms as a tool to investigate **the structure of data, of a quantitative variable**

We often use histograms, however, as a kind of **estimate of the distribution of the variable in the population**; that is, we hope that aspects of the data we see in the histogram will “hold true” if we were to consider the entire population

In this sense, the default rules attempt to provide us with views of the data that **have features we can expect will exist in the population**; the different rules, then, make different assumptions about what the population data look like

Ultimately, however, these rules are derived mathematically under idealized conditions (maybe the population data have a “bell shape” -- more on that shortly) and typically for large sample sizes, making their use in practice subject to some artful interrogation

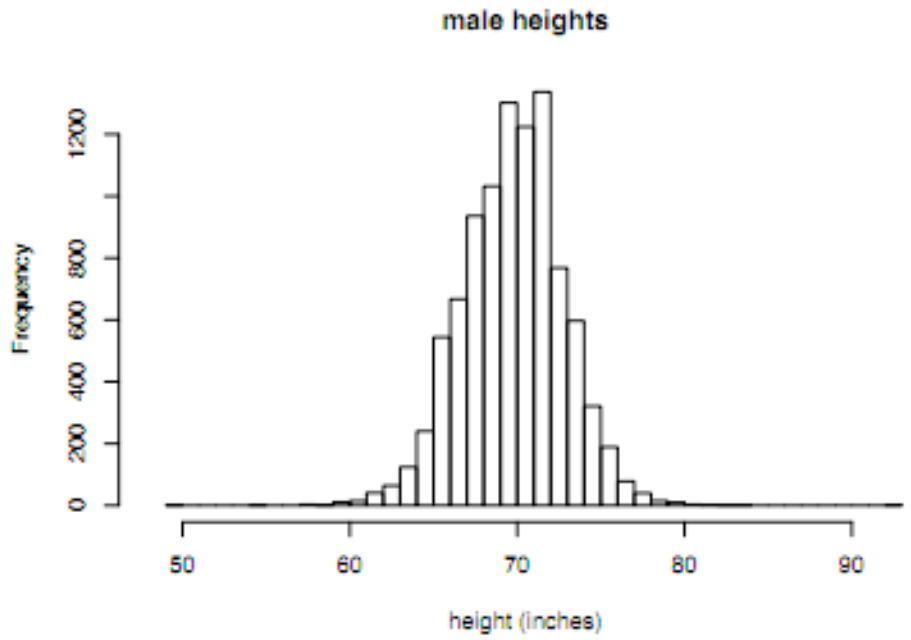
Bottom line: It’s sufficient that you understand there is a choice to be made and that reasonable defaults exist but that **you should question the defaults** if you have the time and examine several bin widths

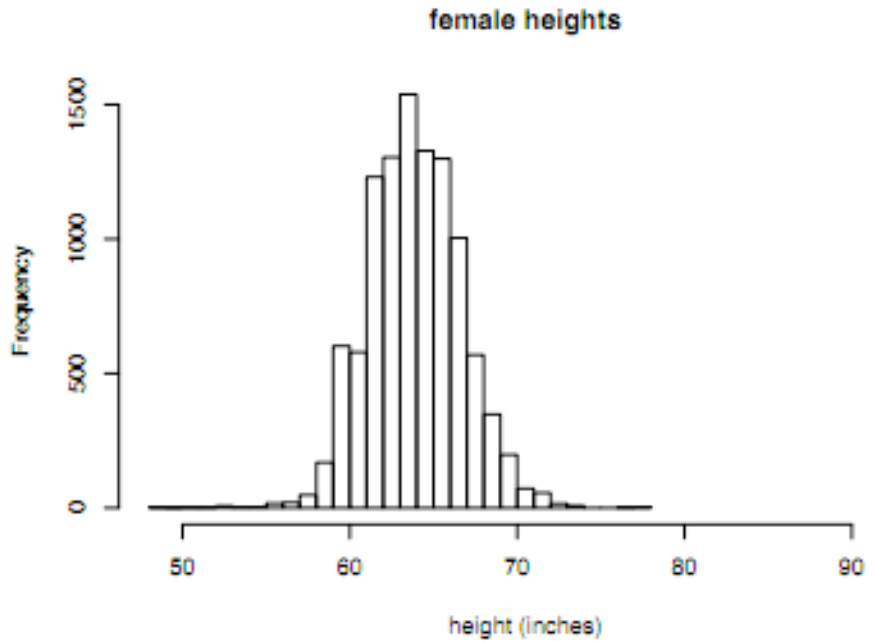
Comparing distributions (I)

We can use these displays to compare distributions

At the left we have separate histograms of the heights of males and females in the sample

What do you see? What terms would you use to describe their shapes?

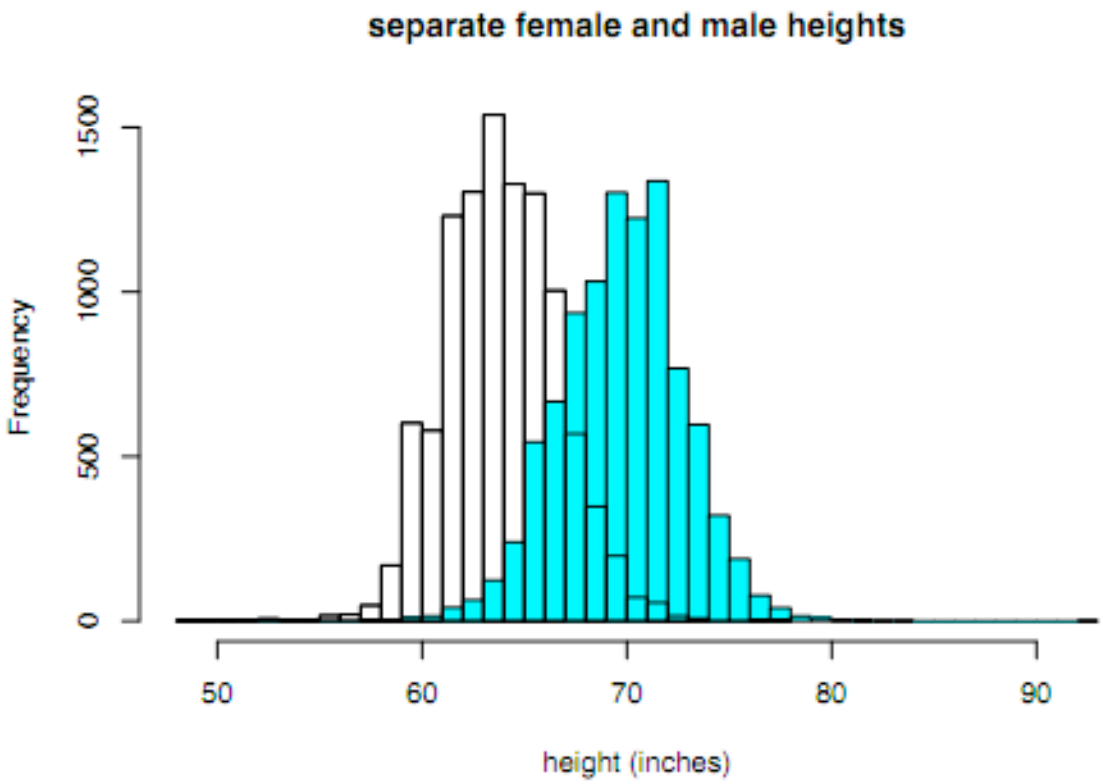




Comparing distributions (I)

A more effective strategy would be to simply overlay one histogram over the other, perhaps adding a snappy color

At this point it should be clear how helpful it is to have a good rule of thumb for picking the number of bins



An aside about skew

Body measurements are very, well, 19th century; it was all the rage at the time, fueled in part by an obsession with criminals, with identifying specific criminals or characterizing criminal "classes"

Lots of body measurements end up having "bell shaped" distributions (like people's heights), or are subject to mild skew (the BRFSS weights); a lot of work went into reasoning around the bell curve, as we'll see

RELEVÉ

DU

SIGNALEMENT ANTHROPOMÉTRIQUE

1. Taille. — 2. Envergure. — 3. Buste. —
4. Longueur de la tête. — 5. Largeur de la tête. — 6. Oreille droite. —
7. Pied gauche. — 8. Médius gauche. — 9. Coudée gauche.


(Réduction photographique 1/7.)
165.670


DESCRIPTIVE


DESCRIPTIVE
43277

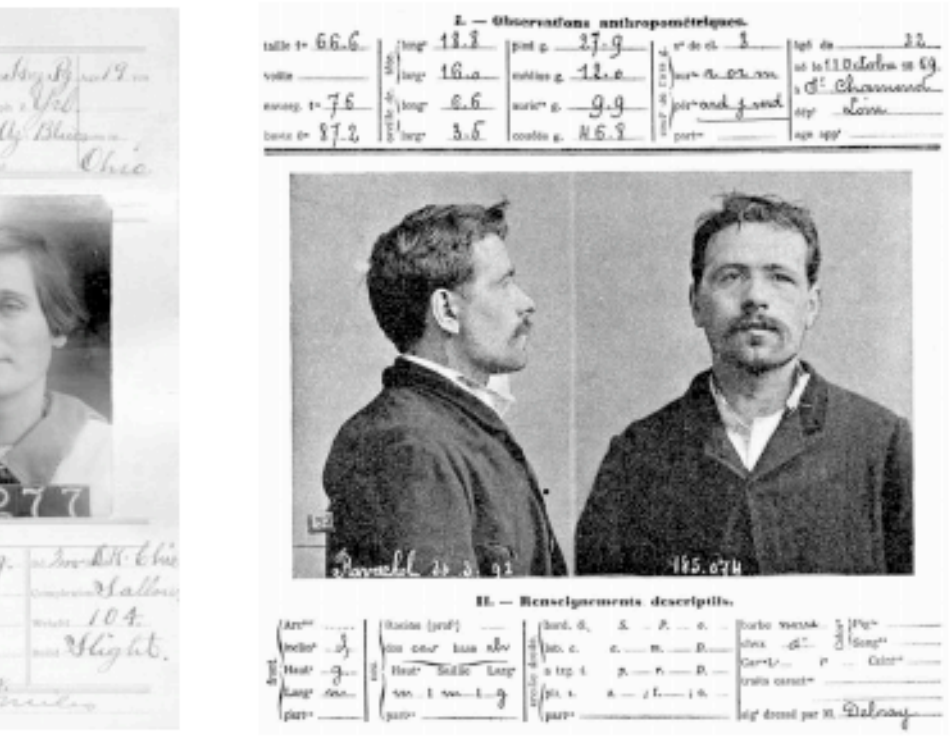

I. — Observations anthropométriques.
II. — Renseignements descriptifs.
165.674

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Table 13 (Special Data).

Relative number of Brothers of various Heights to Men of various Heights, Families of Five Brothers and upwards being excluded.

Heights of the men in inches.	Heights of their brothers in inches.													Total cases.	Medians.
	Below 63	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	Above 74		
74 and above	1	1	...	...	...	...	...	1	1	...	5	3	12	24	
73·5	...	...	...	...	...	1	3	4	8	3	3	2	3	27	
72·5	...	...	...	...	1	1	6	5	9	9	8	3	5	47	71·1
71·5	...	1	...	1	2	8	11	18	14	20	9	4	...	88	70·2
70·5	...	...	1	1	7	19	30	45	36	14	9	8	1	171	69·6
69·5	...	1	2	1	11	20	36	55	44	17	5	4	2	198	69·5
68·5	...	1	5	9	18	38	46	36	30	11	6	3	...	203	68·7
67·5	2	4	8	26	35	38	38	20	18	8	1	1	...	199	67·7
66·5	4	3	10	33	28	35	20	12	7	2	1	...	...	155	67·0
65·5	3	3	15	18	33	36	8	2	1	1	...	...	...	110	66·5
64·5	3	8	12	15	10	8	5	2	1	...	...	...	...	64	65·6
63·5	5	2	8	3	3	4	1	1	...	1	...	...	1	20	
Below 63.....	5	5	3	3	4	2	...	...	...	...	...	...	1	23	
Totals.....	23	29	64	110	152	200	204	201	169	86	47	28	25	1329	

An aside about skew

In the late 1990s, interest was not on bells, but on skew, on extreme skew; so-called heavy tailed distributions and power laws were cropping up everywhere

As an example, let's consider measurements that are personal but not related to the body; say, the number of friends you have on Facebook or the number of visits you get to your web site or the number of bytes you download each day

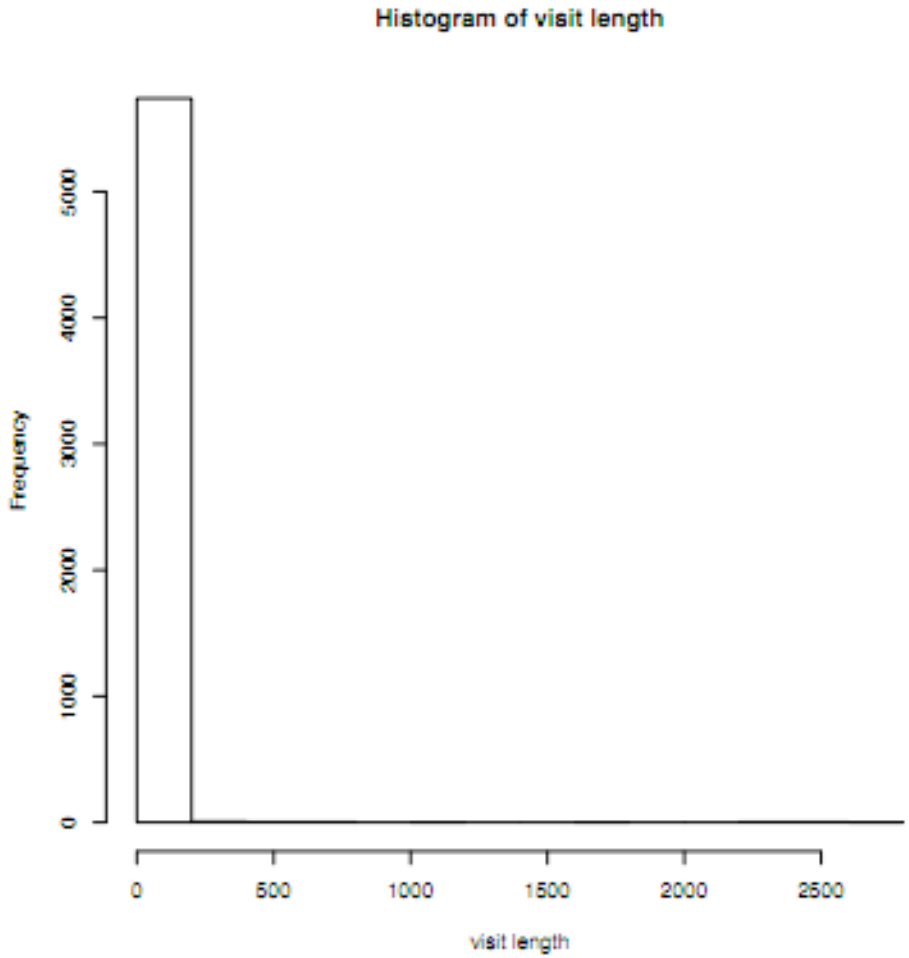
Visits to web sites

Most web sites keep records of who visits and how often; the number of unique pages viewed during a visit is called **the visit length**

For the week between 9/29/10 and 10/02/10 we collected the visit lengths associated with all 5,761 visits to `www.stat.ucla.edu`

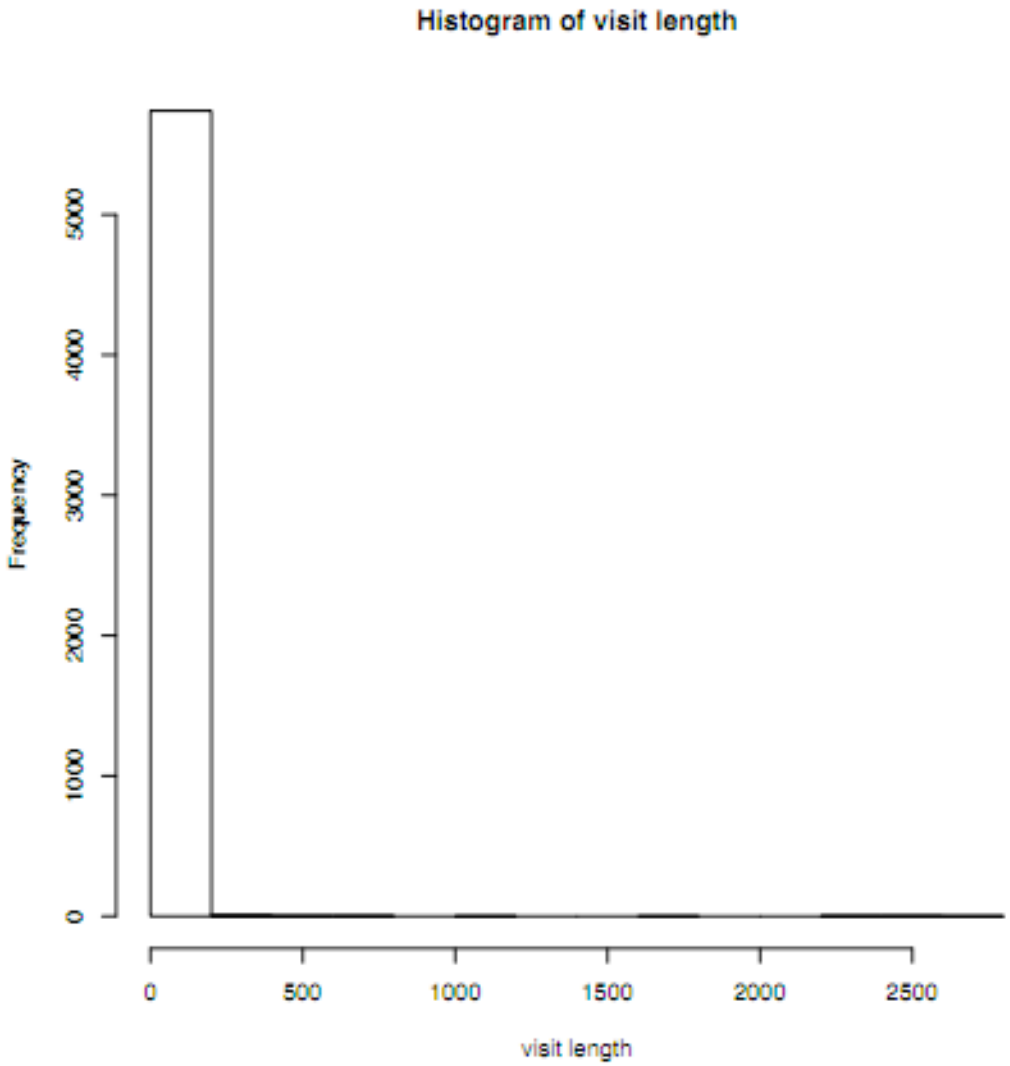
The mean visit length is 11, the median is just 3 (um, I think I just got ahead of myself there)

What does the histogram tell you?



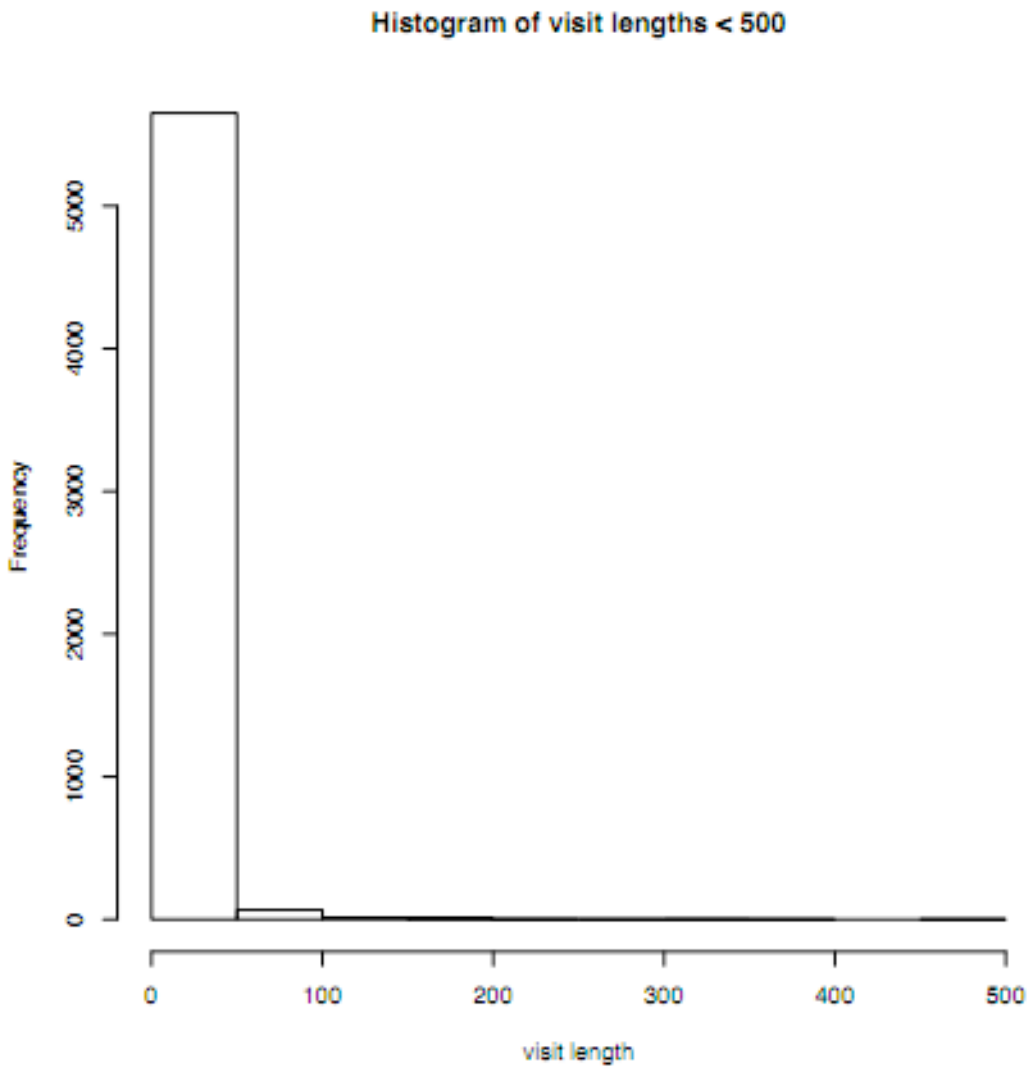
Visits to web sites

The data are badly skewed; the minimum is 1, the maximum is 2,720! We can try to restrict the range of the plot a bit...



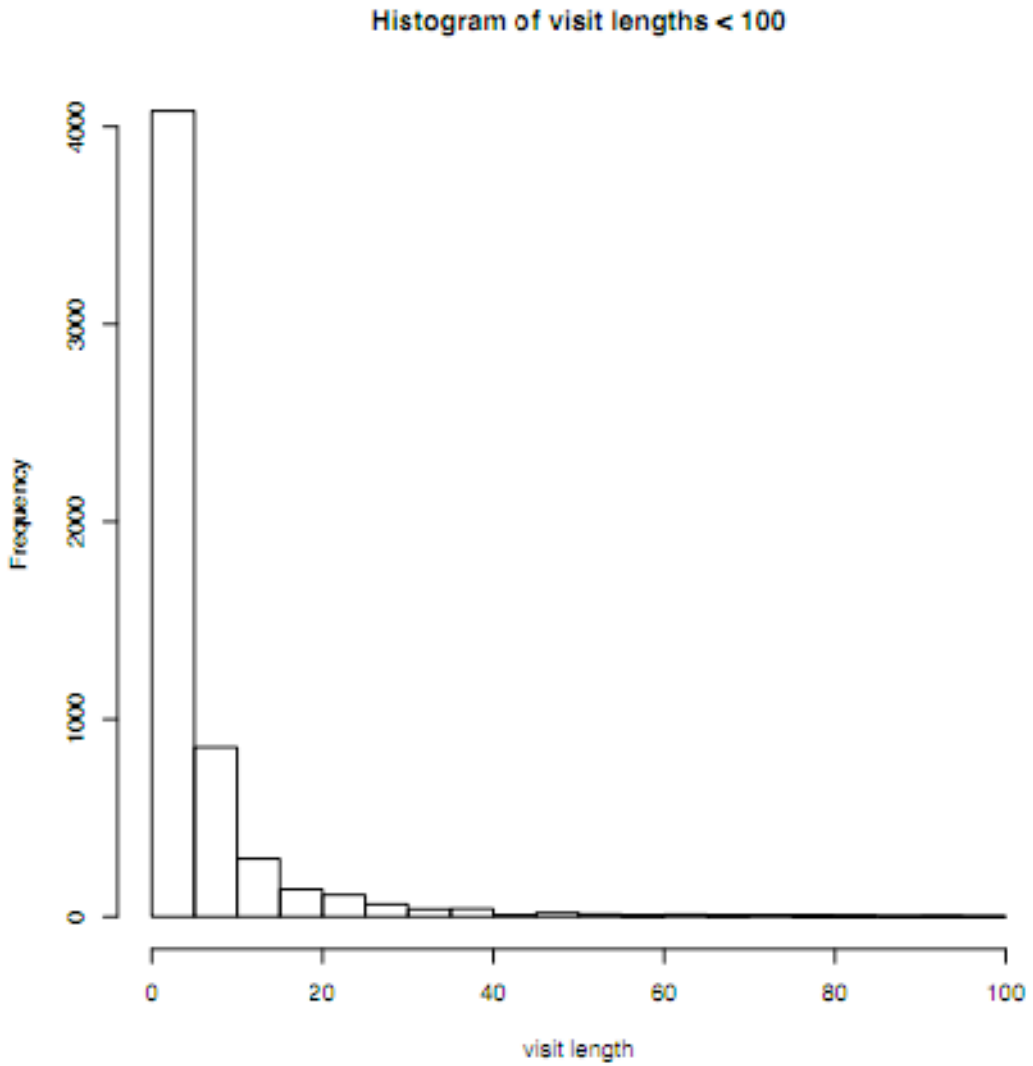
Visits to web sites

The data are badly skewed; the minimum is 1, the maximum is 2,720! We can try to restrict the range of the plot a bit...



Visits to web sites

... and then increase the number of bins; this skew is really dramatic and it will be hard to see a lot in the data this way



An aside about skew

In the late 1990s, John Doyle at CalTech and Jean Carson at UCSB launched a study of complexity and robustness

They provided a theory to explain the heavy tail distributions seen in the sizes of forest fires, in the sizes of files on the web, and a host of other systems with "highly structured, nongeneric, self-dissimilar internal configurations" that are robust yet fragile

While this topic is certainly beyond the reach of this brief introduction, suffice it to say that the world is not all about bell-shaped distributions and that in many cases we are faced with something much more extreme

In these cases, we often prefer to introduce a transformation of the data; typically we would apply a square root or a logarithm, why?

Complexity and robustness

J. M. Carlson*,† and John Doyle‡

+ Author Affiliations

Abstract

Highly optimized tolerance (HOT) was recently framework to study fundamental aspects of cor primarily by systems from biology and engineerin structured, nongeneric, self-dissimilar internal confi fragile external behavior. HOT claims these are the complexity and not accidents of evolution or artifice are inevitably intertwined and mutually reinforcing. I our paper contrasts HOT with alternative perspective real-world examples and also model systems, pa organized criticality.

A vision shared by most researchers in complex syst perhaps even universal, features capture fundamenta manner that transcends specific domains. It is in id sharp differences arise. In disciplines such as biol economics, and ecology, individual complex system of study, but there often appears to be little con models, abstractions, and methods. Highly optimiz one recent attempt, in a long history of efforts, to d for studying complexity. The HOT view is motivated and engineering. Theoretically, it builds on mathen control, communications, and computing. In this pap examples but avoid theories and mathematics th

Judging shape

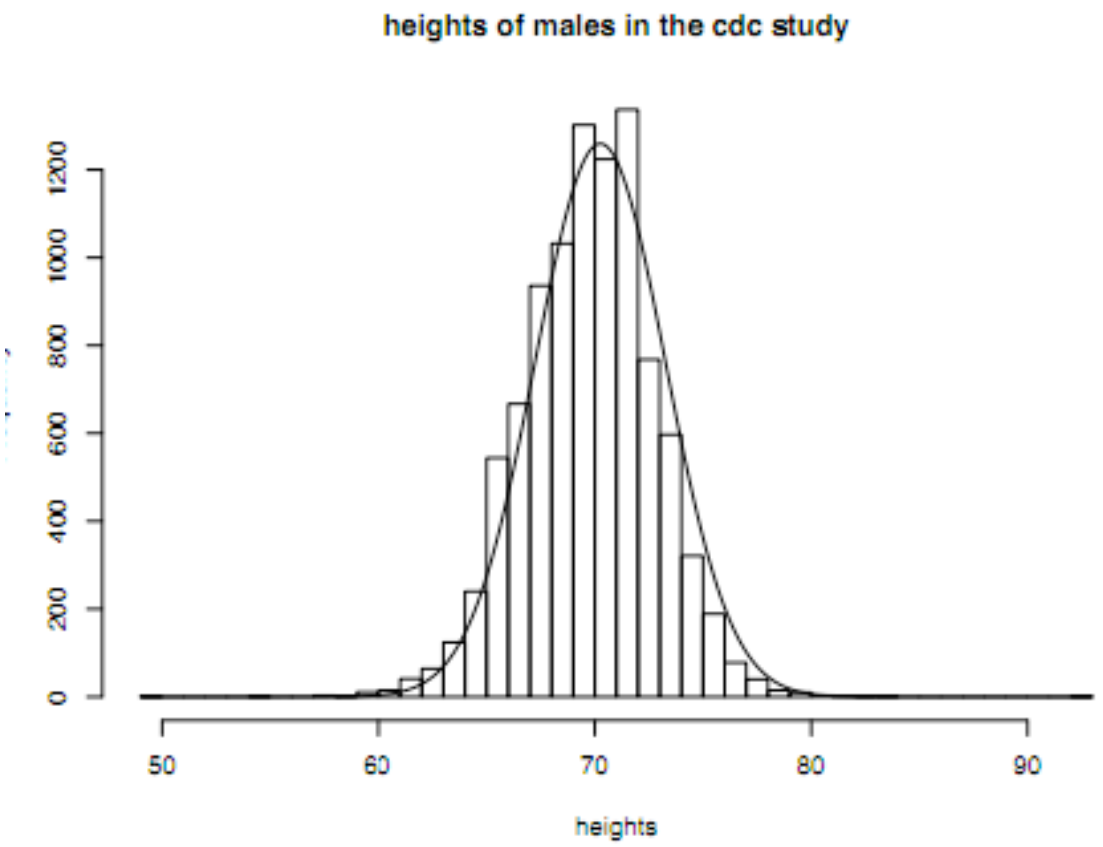
Over the course of the quarter, we're going to want to do more than put data into broad categories (symmetric, skew, etc)

Instead want to judge whether or not it seems reasonable that the data "follow" a particular known distribution

The normal distribution, for example, is often used in this way

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$$

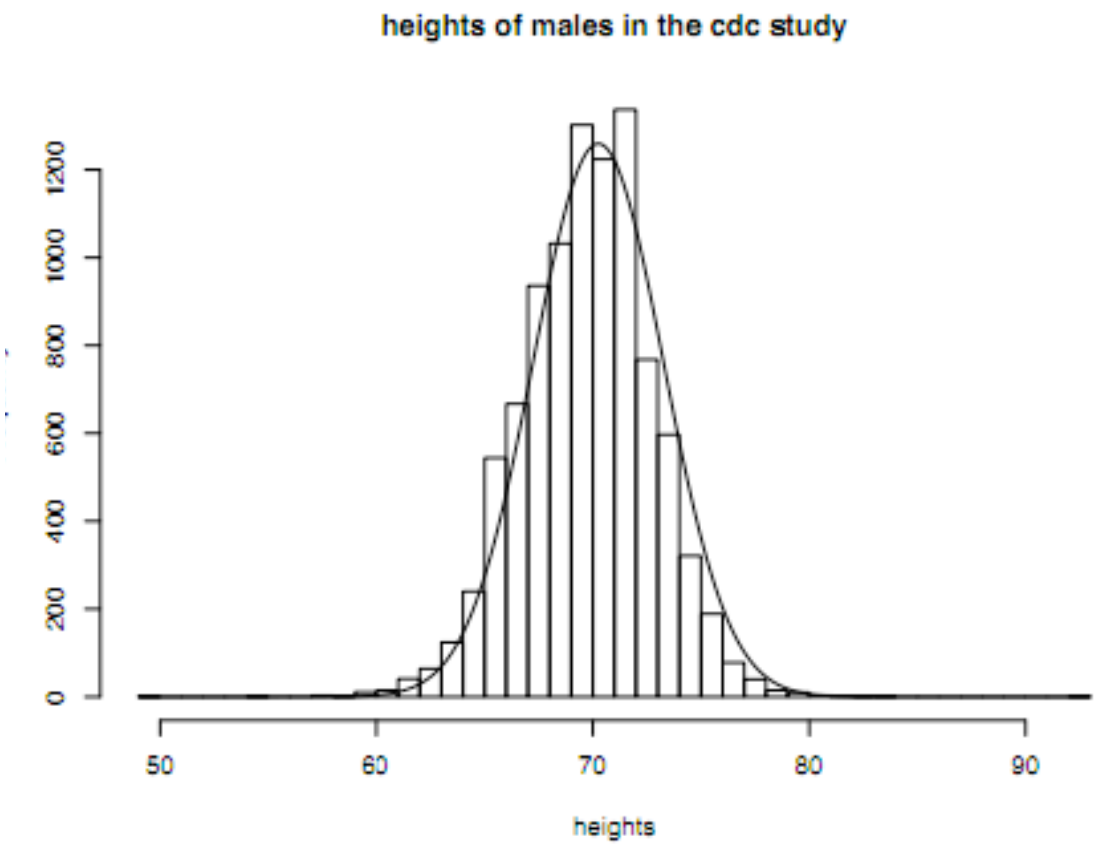
At the right, we have a histogram of men's heights in the BRFSS with an overlay of a normal curve (setting μ to be the sample mean of men's heights in our dataset and σ^2 the sample variance)



Judging shape

What can you see from the display on the right? Do the heights of males in the CDC study look, well, normal?

How easy or hard is this display to read?

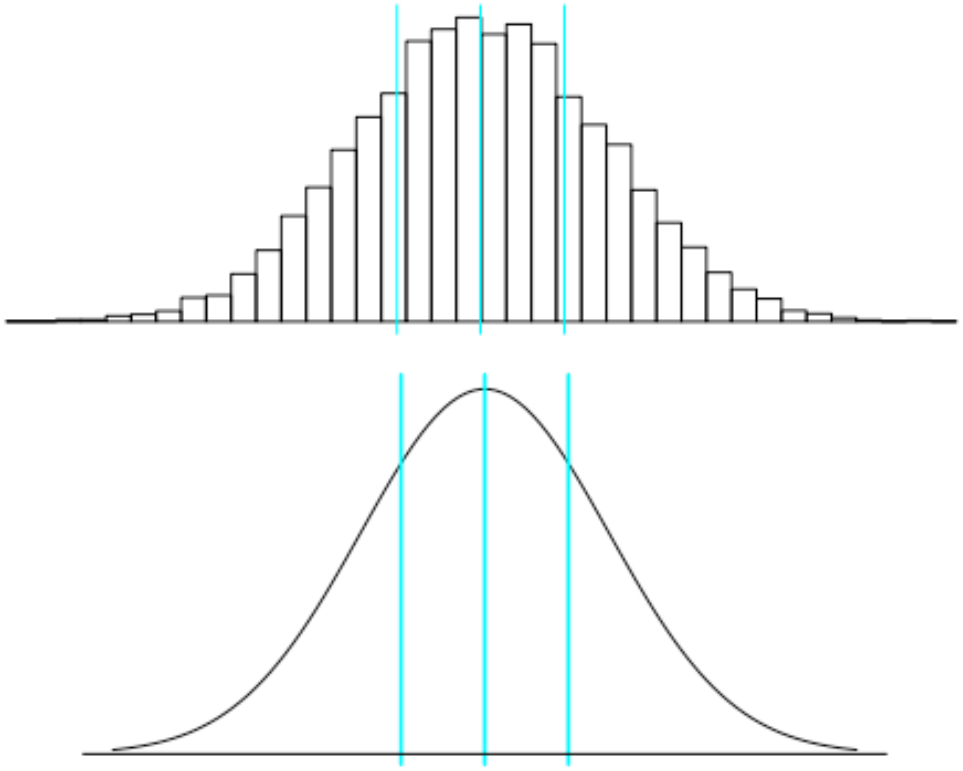


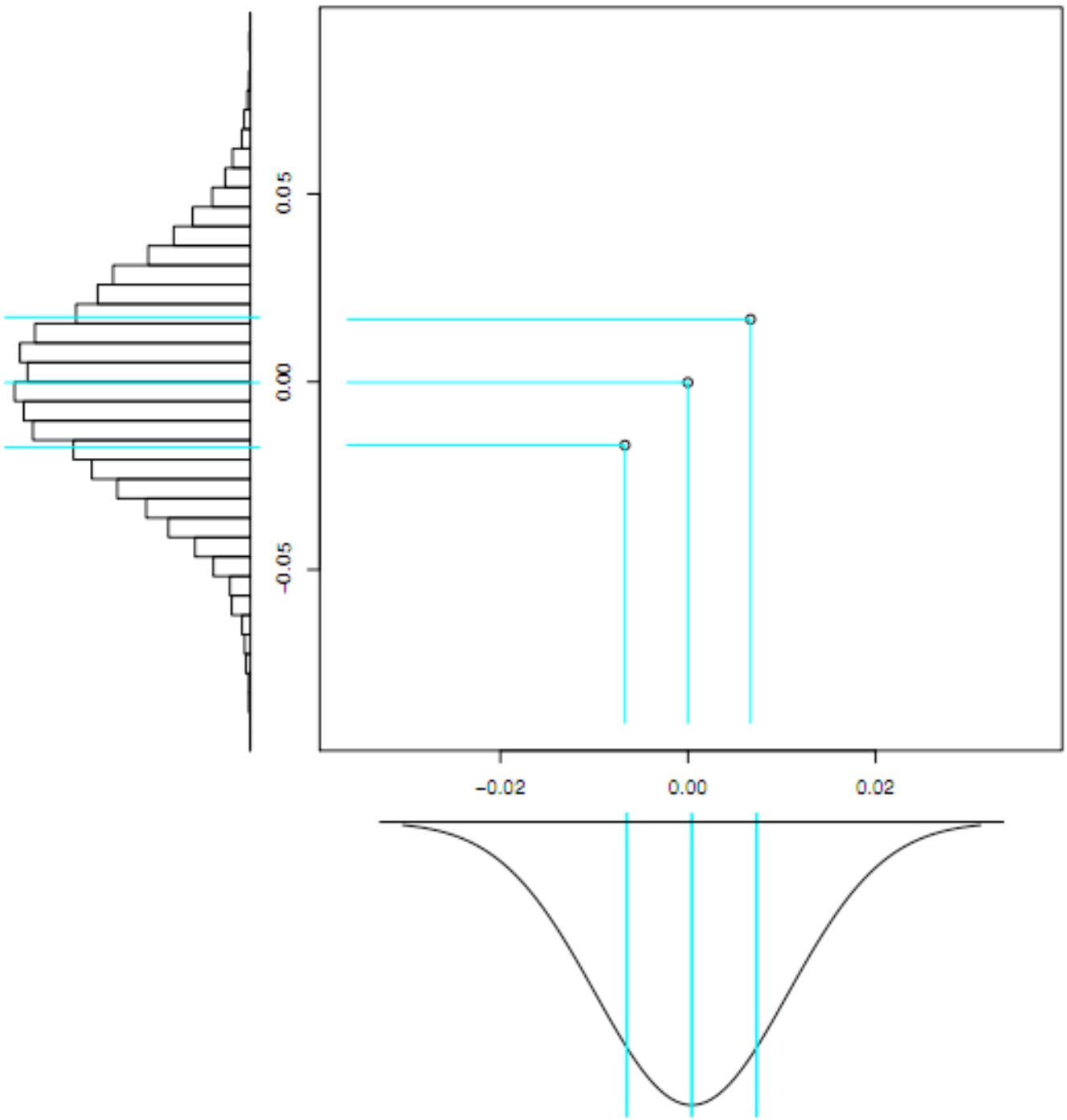
Assessing normality

A normal probability plot compares the way the normal curve distributes probability to the way our sample has arranged its points

Let's start by dividing each into four pieces; for our sample, this means dividing the data using the quartiles we defined for the box plot; for the normal density this means finding regions that divide the total area under the curve into four pieces

To make a more direct comparison, we can try plotting these points against each other...

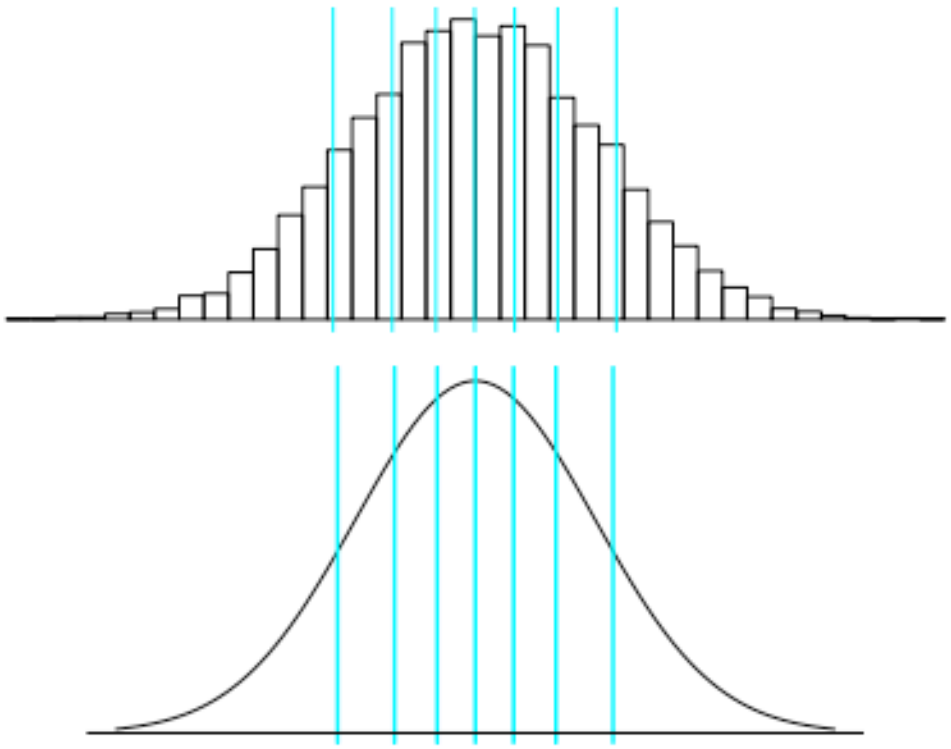


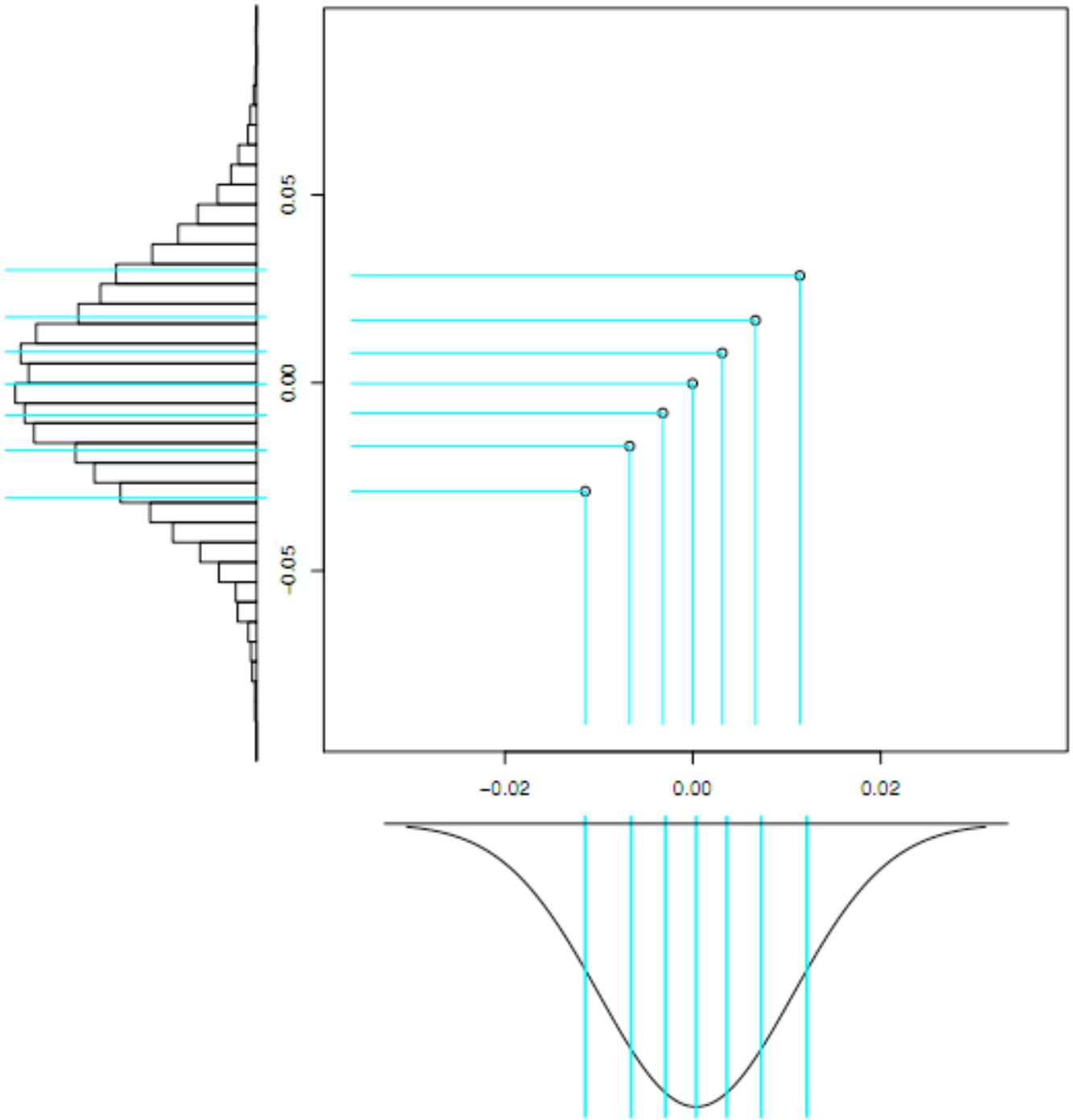

0.05
0.00
−0.05
−0.02
0.00
0.02

Assessing normality

We can continue, this time, dividing the data into 8 pieces (or taking each of the four and dividing them in half)

And again, to make a more direct comparison, we can try plotting these points against each other...

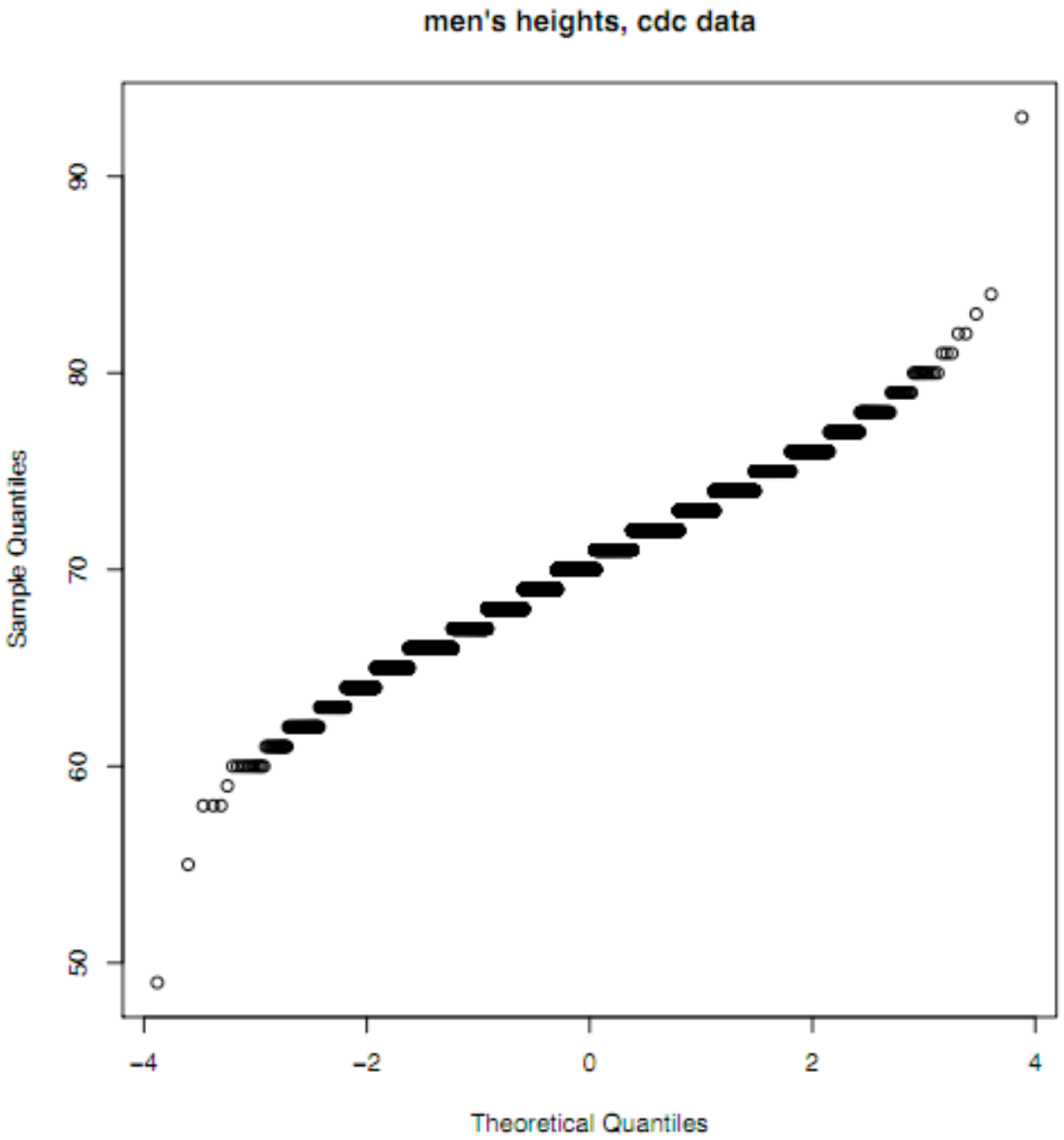


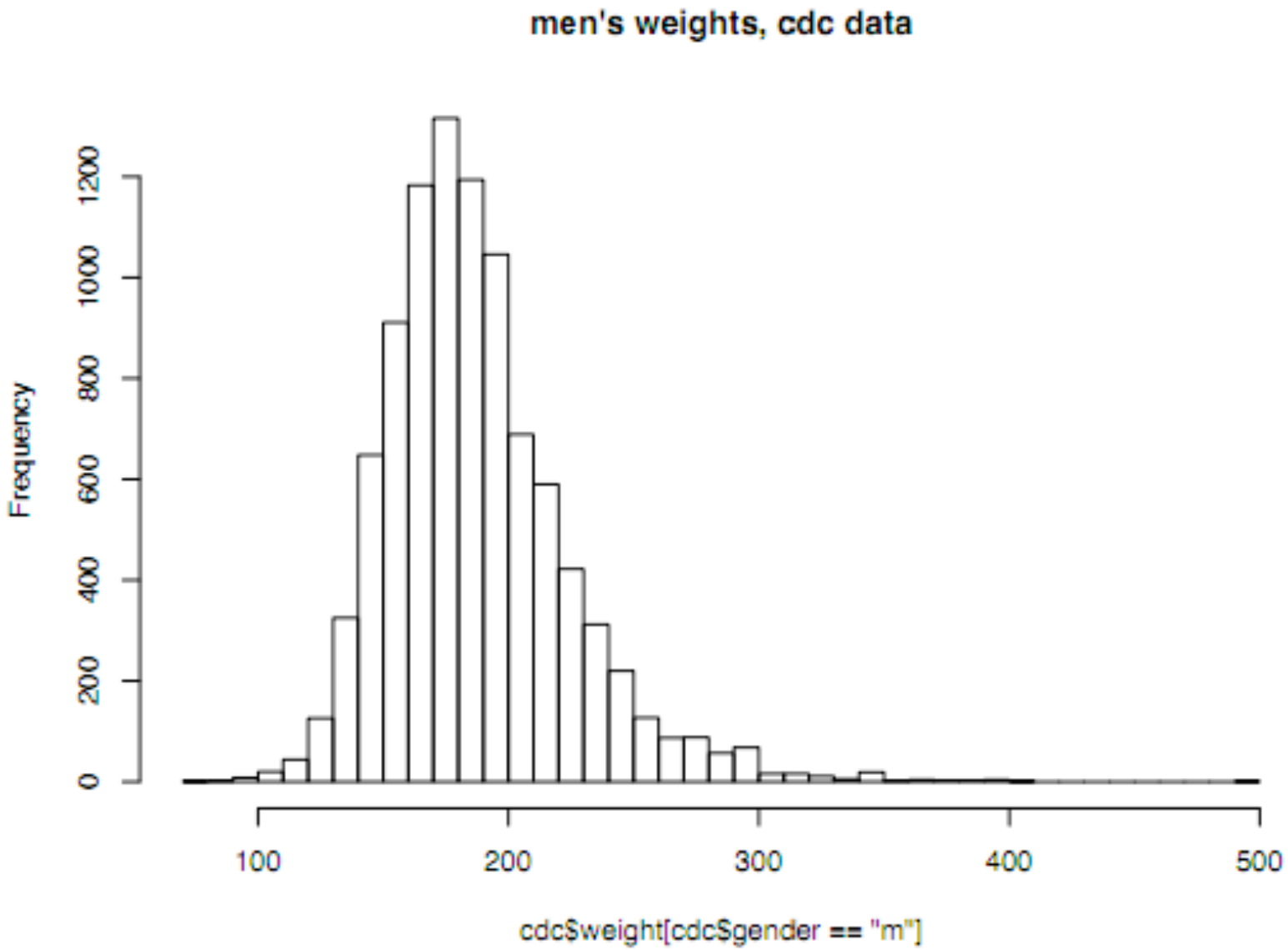

0.05
0.00
−0.05
−0.02
0.00
0.02

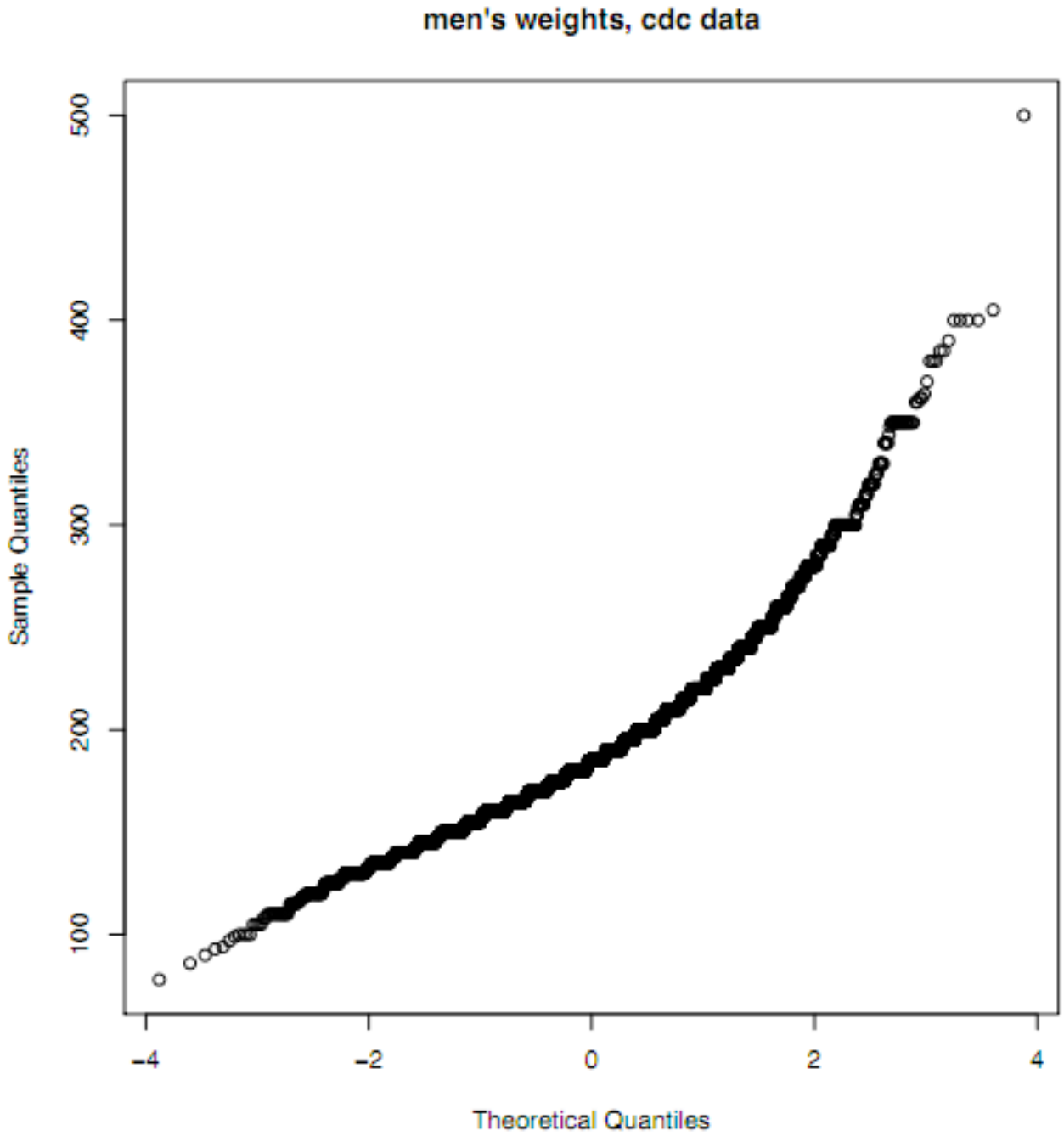
Assessing normality

Continuing this way, we get a plot like the one on the next page for male's heights in the CDC study

The good thing about this kind of plot is that departures from normality are seen as deviations from a straight line -- Visually this is a HUGE advance


men's heights, cdc data
Sample Quantiles
Theoretical Quantiles
50
60
70
80
90
−4
−2
0
2
4


men's weights, cdc data
Frequency
0
200
400
600
800
1000
1200
100
200
300
400
500
cdc$weight[cdc$gender == "m"]


men's weights, cdc data
Sample Quantiles
Theoretical Quantiles
100
200
300
400
500
−4
−2
0
2
4

Comparing distributions (II)

While histograms are powerful, sometimes we want to make very rough comparisons between more than two distributions

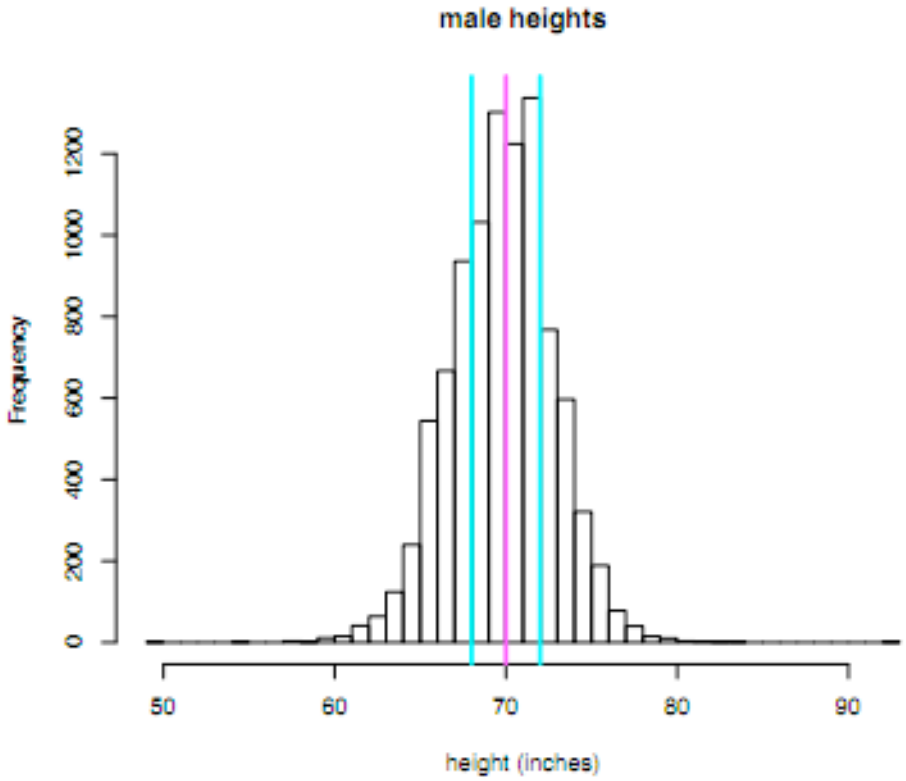
To do this, we will introduce a kind of cartoon representation of a frequency distribution for continuous data

The median and quartiles

To craft this cartoon, we begin by dividing the data in half; that is, we identify a center point

The simplest way to do this is to choose the point for which half of the data lie to the left and half to the right; this is called **the median**

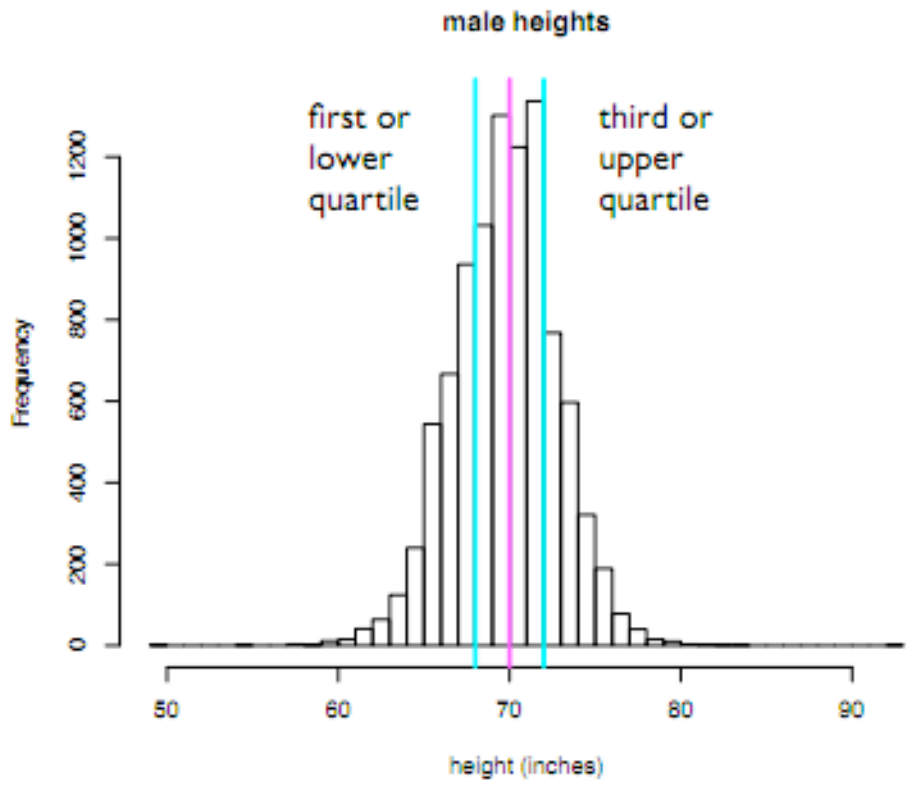
We then take the two halves and in turn divide them in half; in effect, we have split the data into four pieces



The median and quartiles

We often refer to the points marked by the cyan lines as the upper and lower quartiles

They are also called the first and third quartiles; as you might expect, the median is also known as the second quartile



Interquartile range

By design, the interval marked by the upper and lower quartiles contain half of the data; it is known as **the interquartile range**

While the median describes the center of the data, the interquartile range gives us a sense of how **spread out the data are**

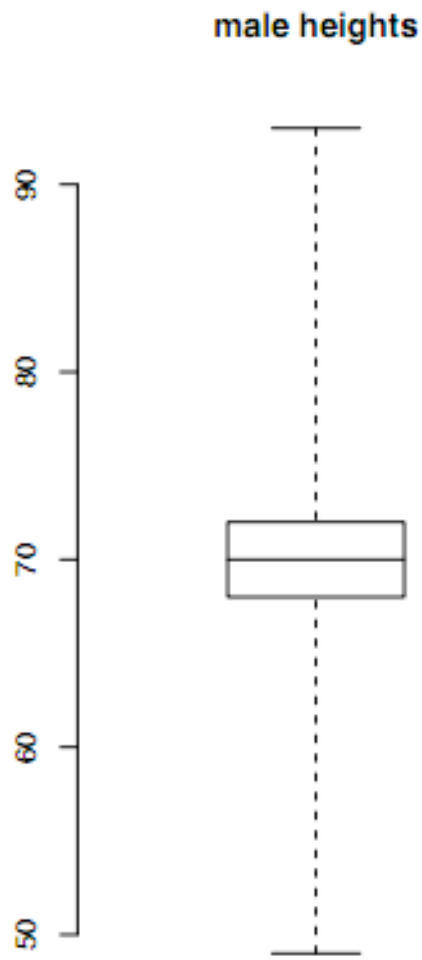
What other measures of spread might we consider?

Boxplots

The boxplot is a graphical representation of a frequency distribution; it is based on five numbers, the minimum data value, the maximum, and the three quartiles

In its standard form*, the whiskers mark the minimum and maximum values; the box is defined by the interquartile range and the median is the horizontal bar in the middle of the box

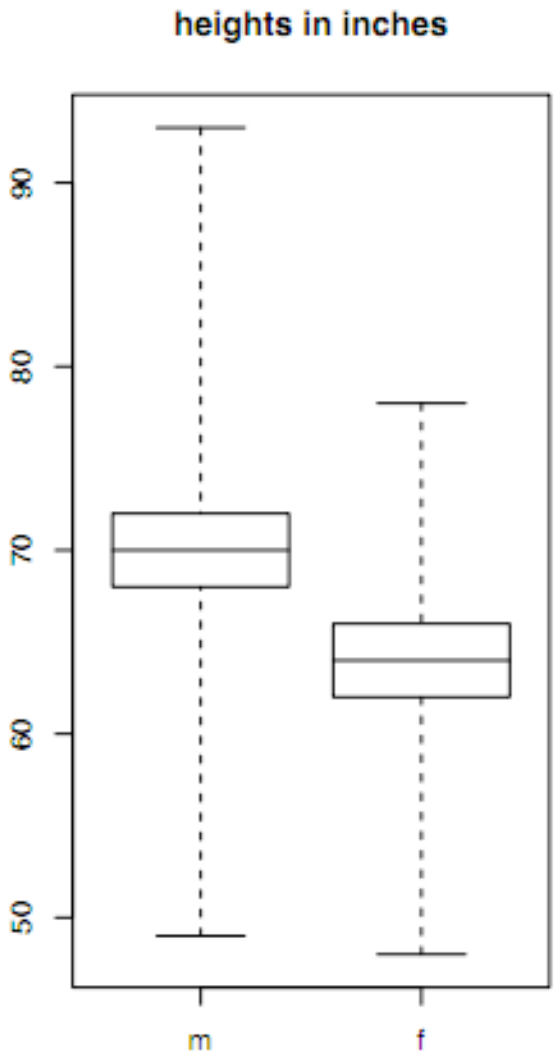
Oh, we should mention that boxplots were developed by John Tukey who we met in the first lecture

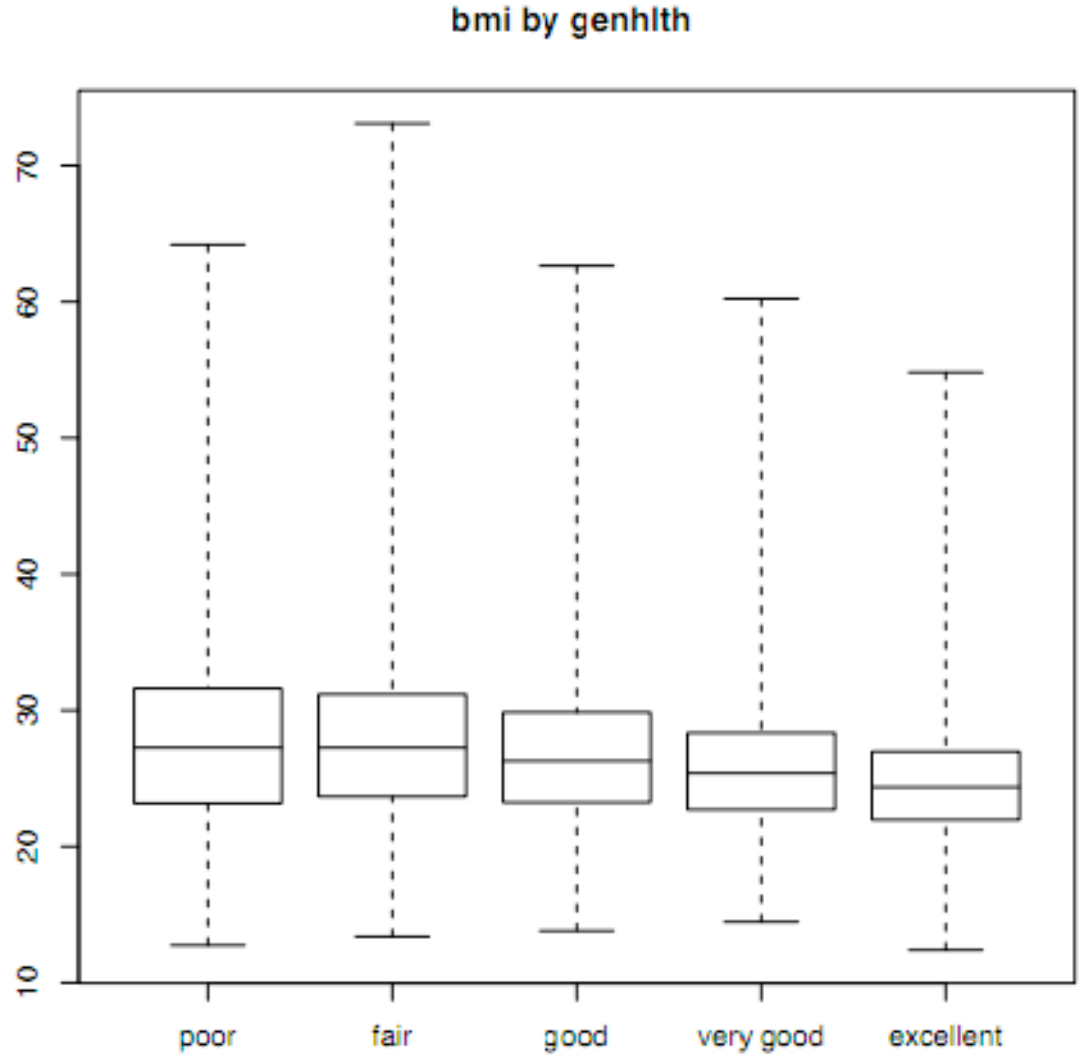


Boxplots

While they are fairly cartoonish, boxplots let us compare a number of distributions at one time

Here we have male and female heights...




bmi by genhlth
70
60
50
40
30
20
10
poor
fair
good
very good
excellent

Boxplots

R implements a modified version of the boxplot, one that incorporates the notion of **outliers**

Outliers are points that stand out from the rest of the data in some way; we typically identify such points on the basis of our prior expectations about how data should behave

Modified boxplots

The boxplots we have seen so far are direct graphical representations of the so-called five number summary

Given a sample of observations on a quantitative variable, the five number summary consists of **the minimum, the lower quartile, the median, the upper quartile and the maximum**

The default boxplot provided by most (if not all) modern statistical software packages is a little more complex; it attempts to highlight values that are **“too extreme”**

Tukey's motivation

*A cautious data analyst often has reason for concern over the distortions that aberrant observations can cause... It is informative, and may be important, to examine samples... for the presence of "outliers" or "exotic values" because their unexpected behavior may indicate failure of a model or point to an unanticipated phenomenon.**

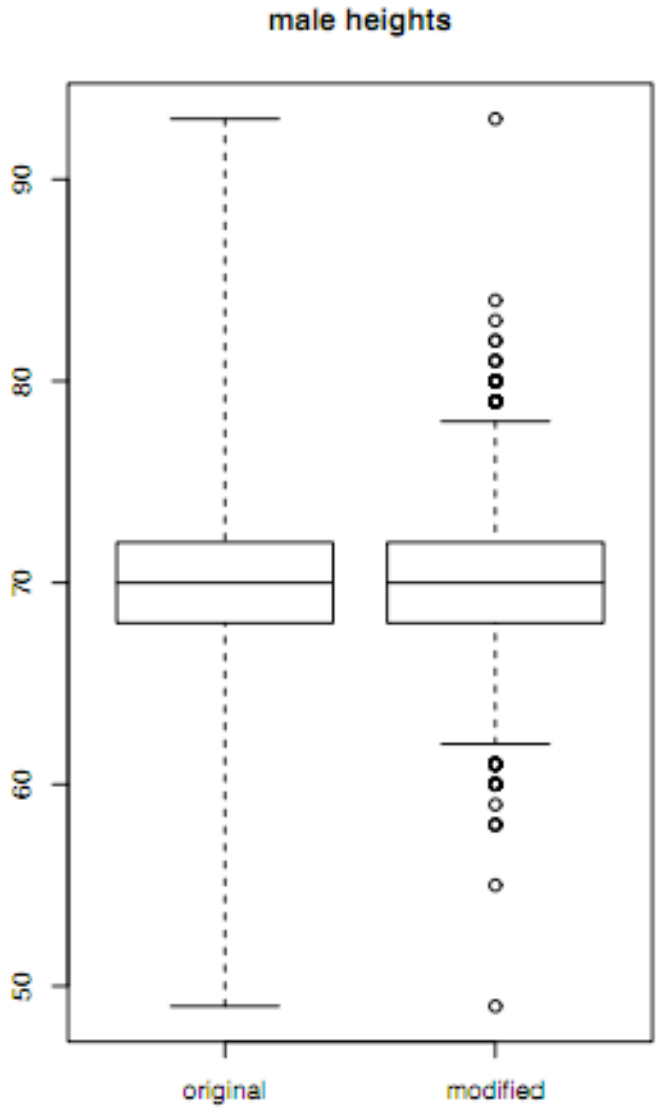
* David C. Hoaglin, Boris Iglewicz, and John W. Tukey (1986), Performance of some resistant rules for outlier labeling, *Journal of the American Statistical Association*, Vol. 81, pp. 991-999.

Modified boxplots

Here we have two boxplots for the heights of males in the CDC study from last lecture; the plot on the left is the original boxplot we created last time, while the one on the right is its "modified" cousin

In this case, the new plot brings the whiskers closer to the middle of the display; by peeling back the whiskers, we expose some of the actual data beyond the "fences"

What does this give us?



Tukey's simple rule

Recall that the distance between the upper and lower quartiles is known as the **interquartile range** or IQR

Define **lower fence** to be $Q_1 - 1.5\,\text{IQR}$; we say that any point below the lower fence is a possible outlier, requiring some investigation

Define the **upper fence** to be $Q_3 + 1.5\,\text{IQR}$; we say that any point above the upper fence is also a possible outlier

Tukey's simple rule

We then highlight potential outliers in our boxplots by adjusting the placement of the upper and lower whiskers

If there are no data points below the lower fence, we leave the lower whisker at the minimum data point; if there are, we place the whisker at the smallest point above the lower fence

If there are no data points above the upper fence, we leave the upper whisker at the maximum data point; if there are, we place the whisker at the largest point below the upper fence

Example

For males, the five number summary of the variable weight is

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
78.0	165.0	185.0	189.3	210.0	500.0

Therefore, the IQR is (210-165) = 45 and our fences are computed to be

lower: 165 - 1.5*45 = 97.5

upper: 210 + 1.5*45 = 277.5

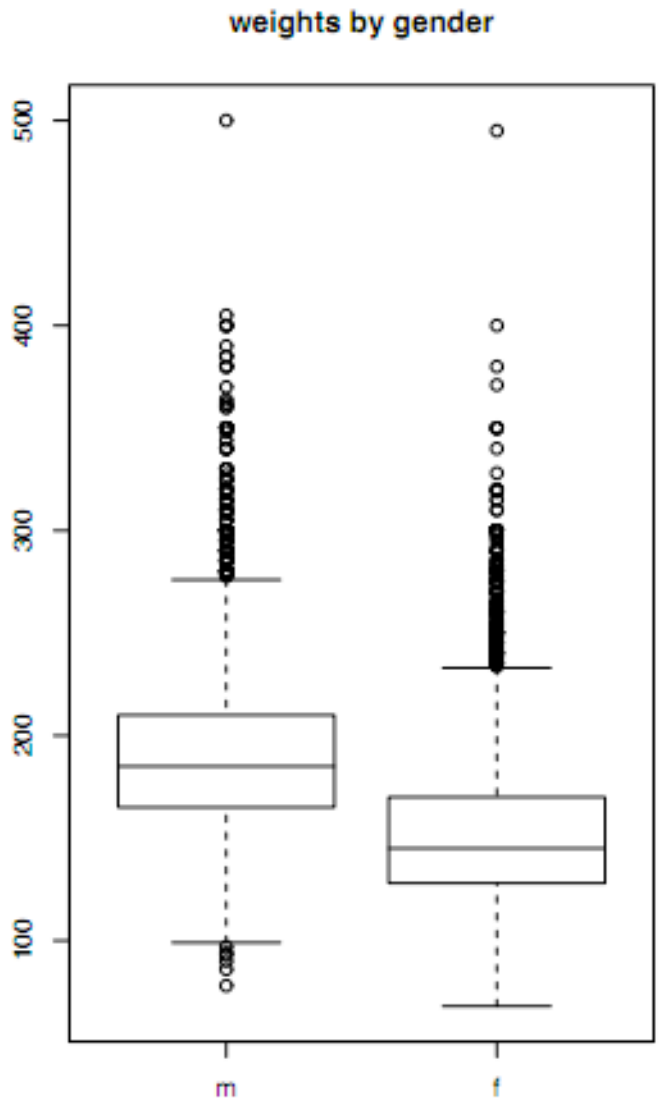
Using this, we identify 6 possible outliers that are extreme and small, and 262 that are extreme and large

Modified boxplots

What do we think of this alteration? What new information does it provide? How might this be useful for us when examining data?

The choice to highlight points falling above or below 1.5 times the IQR, is just that, a choice; why not 1.0 or 3.0 times the IQR, how would those choices change things?

How do you think the value of 1.5 was settled on? What principles were used to guide the choice? Maybe an easier question is, how do we want this



Performance of Some Resistant Rules for Outlier Labeling

DAVID C. HOAGLIN, BORIS IGLEWICZ, and JOHN W. TUKEY*

The techniques of exploratory data analysis include a resistant rule for identifying possible outliers in univariate data. Using the lower and upper fourths, F_L and F_U (approximate quartiles), it labels as "outside" any observations below $F_L - 1.5(F_U - F_L)$ or above $F_U + 1.5(F_U - F_L)$. For example, in the ordered sample $-5, -2, 0, 1, 8$, $F_L = -2$ and $F_U = 1$, so any observation below -6.5 or above 5.5 is outside. Thus the rule labels 8 as outside. Some related rules also use cutoffs of the form $F_L - k(F_U - F_L)$ and $F_U + k(F_U - F_L)$. This approach avoids the need to specify the number of possible outliers in advance; as long as they are not too numerous, any outliers do not affect the location of the cutoffs.

To describe the performance of these rules, we define the some-outside rate per sample as the probability that a sample will contain one or more outside observations. Its complement is the all-inside rate per sample. We also define the outside rate per observation as the average fraction of outside observations. For Gaussian data the population all-inside rate per sample (0) and the population outside rate per observation (.7%) substantially understate the corresponding small-sample values. Simulation studies using Gaussian samples with n between 5 and 300 yield detailed information on the resistant rules. The main resistant rule ($k = 1.5$) has an all-inside rate per sample between 67% and 86% for $5 \le n \le 20$, and corresponding estimates of its outside rate per observation range from 8.6% to 1.7%.

Both characteristics vary with n in ways that lead to good empirical approximations. Because of the way in which the fourths are defined, the sample sizes separate into four classes, according to whether dividing n by 4 leaves a remainder of 0, 1, 2, or 3. Within these four classes the all-inside rate per sample shows a roughly linear decrease with n over the range $9 \le n \le 50$, and the outside rate per observation decreases linearly in $1/n$ for $n \ge 9$.

A more theoretical approximation for the all-inside rate per sample works with the order statistics $X_{(1)} \le \cdots \le X_{(n)}$. In this notation the fourths are $X_{(f)}$ and $X_{(n+1-f)}$ with $f = \frac{1}{2}[(n + 3)/2]$, where $[\cdot]$ is the greatest-integer function. A sample has no observations outside whenever $\{X_{(f)} - X_{(1)}\}/\{X_{(n+1-f)} - X_{(f)}\} \le k$ and $\{X_{(n)} - X_{(n+1-f)}\}/\{X_{(n+1-f)} - X_{(f)}\} \le k$. We first approximate the numerators and denominator in these ratios by constant multiples of chi-squared random variables with the same mean and variance. We then approximate the logarithm of each ratio by a Gaussian random variable, and we calculate the correlation between these variables from the fact that the ratios have the same denominator. Finally, a bivariate Gaussian probability calculation yields the approximate all-inside rate per sample. The error of the result relative to the simulation estimate is typically from 1% to 2% for $5 \le n \le 50$.

To provide an indication of how the two rates behave in alternative "null" situations, the simulation studies included samples from five heavier-tailed members of the family of h-distributions. For a given sample size, the all-inside rate per sample decreases as the tails become heavier, and the outside rate per observation increases.

KEY WORDS: Bivariate normal distribution; Chi-squared distribution; Exploratory data analysis; h-distributions; Masking.

1. INTRODUCTION

A cautious data analyst often has reason for concern over the distortions that aberrant observations can cause. By using summaries that change only slightly in response to an arbitrary change in any small part of the data, robust and resistant methods have made it possible to minimize the effects of such unusual data. It is still informative, however, and may be important, to examine samples and residuals for the presence of "outliers" or "exotic values" because their unexpected behavior may indicate failure of a model or point to an unanticipated phenomenon.

Research on methods of testing for outliers has produced an extensive literature, discussed in books by Barnett and Lewis (1978, 1984) and Hawkins (1980) and in articles by Barnett (1983) and Beckman and Cook (1983). A number of the proposed procedures have difficulty when a sample may contain multiple outliers. The problems include *masking*, in which the presence of other outliers makes each outlier difficult to detect, and *swamping*, in which the procedure tends to declare too many outliers when the null hypothesis of no outliers is rejected. From the point of view of robust/resistant data analysis, masking occurs because the procedure has a too-low breakdown point. [Donoho and Huber (1983) discussed the breakdown point in some detail. They defined it as "roughly, the smallest amount of contamination that may cause an estimator to take on arbitrarily large aberrant values" (p. 157).] Thus, by having higher breakdown points, resistant measures of a sample's location and spread should make it possible to avoid masking in most situations.

As one informal step in this direction, exploratory data analysis (Tukey 1977a) includes a resistant rule of thumb for identifying observations that are extreme and hence are potential outliers. The breakdown point of this rule is roughly 25%. Its practical advantages include simplicity, ability to identify multiple outliers, and routine use in such displays as boxplots. Thus it is valuable to study the probability behavior of the rule in "null" and "alternative" situations.

Small to moderately large samples from the Gaussian distribution constitute the customary null situation. We use simulation to measure the rule's performance in terms of three criteria: (a) the all-inside rate per sample, (b) the probability that (in small samples) as many as three of the observations are "outside," and (c) the outside rate per

* David C. Hoaglin is Research Associate, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138. Boris Iglewicz is Professor, Department of Statistics, Temple University, Philadelphia, PA 19122. John W. Tukey is Senior Research Statistician and Donner Professor of Science Emeritus, Fine Hall, Princeton University, Princeton, NJ 08544. The authors are grateful to Gerry Feit, Jorge Martinez, and Gaston Mendoza for their assistance in carrying out the simulation work, to A. R. DiDonato for supplying the program used to calculate bivariate normal probabilities, and to Michael Dolker, John D. Emerson, Peter J. Kempthorne, Frederick Mosteller, James L. Rosenberger, Roy E. Welsch, Cleo Youtz, an associate editor, and two referees for helpful comments. This work was supported in part by U.S. Army Research Office Contract DAAG29-82-K-0085 with Harvard University, by National Science Foundation Grant SOC75-15702 to Harvard University, by National Cancer Institute Grant CA-31247 to Harvard University, and by U.S. Army Research Office Contract DAAG29-82-K-0178 with Princeton University. An earlier version of this article appeared in the *1980 Proceedings of the Statistical Computing Section*.

But, it's just a graphical tool...

We could examine how far out to draw the fences by **trying out different choices in situations where we know the answer** (well behaved data with no anomalous points, data with a few odd points) and see what happens

This is precisely what Tukey does in his paper; **he simulates data from a model** where there are no strange points (a bell-shaped distribution, actually, but we'll get to that later) and tries to make sure that in those "normal" cases, the plot doesn't flag too many points as possibly anomalous

Tukey and company comment that...

> *Any observations that fall below the [lower fence] or above [the upper fence] are termed outside. We would inquire into the circumstances surrounding any outside data values in an attempt to learn the reasons for their unusual behavior, and we are likely to level the corresponding points with appropriate identifiers in any graphical display.*

But keep in mind that **there is a difference between a point being "outside," or an "outlier" as is it is commonly called, and there being a real problem with the data represented by that point**; the modified box plot is meant to call your attention to data that might be strange, **it is not a rule defining what is strange or a problem**

The upshot

This is now the second time you've seen **a kind of tuning parameter** (the bin width of a histogram and the fence distance of a boxplot) that you can easily vary thanks to software

No matter whether you use the default choices (which you will typically do for boxplots, whereas for histograms, experimentation is encouraged) or not, it's important to remember that they are guides; **they are nothing more than devices that tend to make the graphics we're considering informative** in a large number of cases

So, while a graphic may indicate some points are outliers, it is up to you, the data analyst, to have a look and make a determination; at this point, all we've presented are guides that help you see things, **they are not reasoning for you!**

Extensions (I)

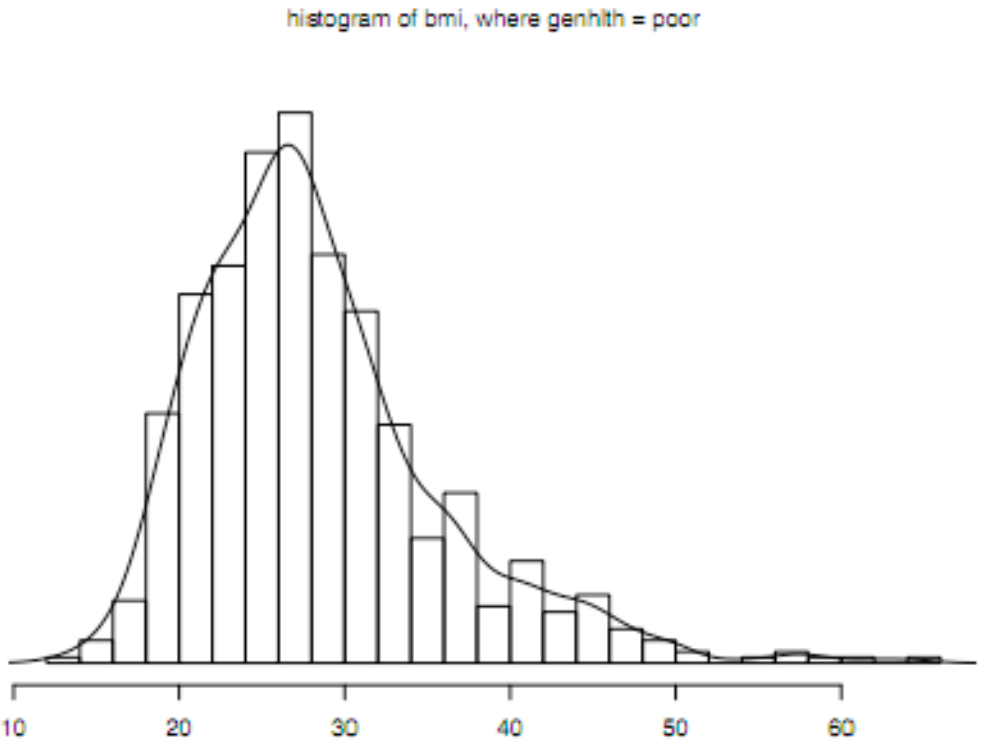
Boxplots represent a rather extreme compression of the frequency distribution; after all, only 5 numbers are being shown

They were also developed at a time when EDA might be done by hand (Tukey advocates extensive use of tracing paper, for example)

An extension of the boxplot attempts to make more of the distribution visible; it starts with a "smoothed" histogram

On the right we have a histogram of BMI for people who reported being in poor health; the curve is the smoothed histogram

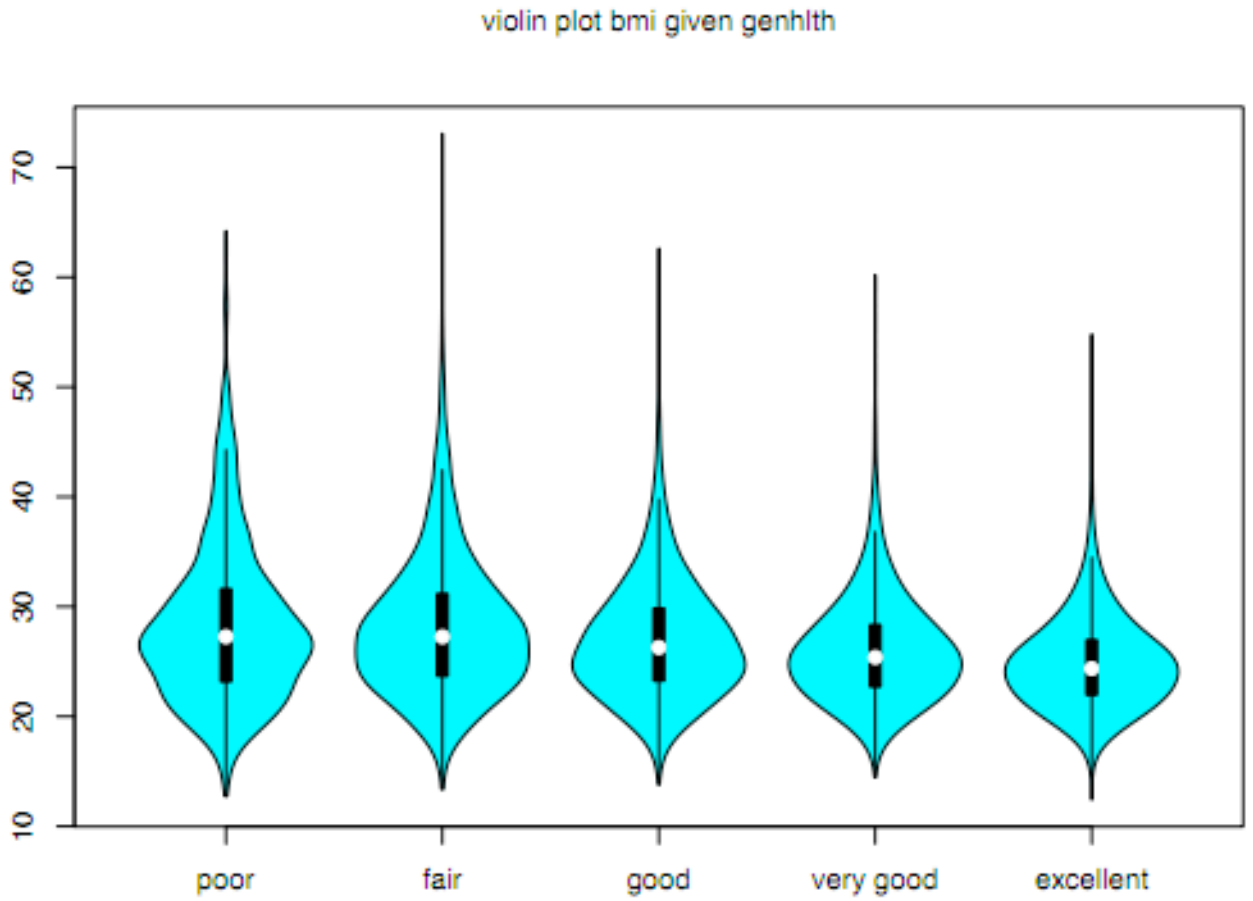
How can we use this to make the boxplot a bit more expressive?



Violin plots

The so-called violin plot might be more artistry than data analysis; but it uses the smoothed histogram tipped on its side and mirrored left and right in place of a box

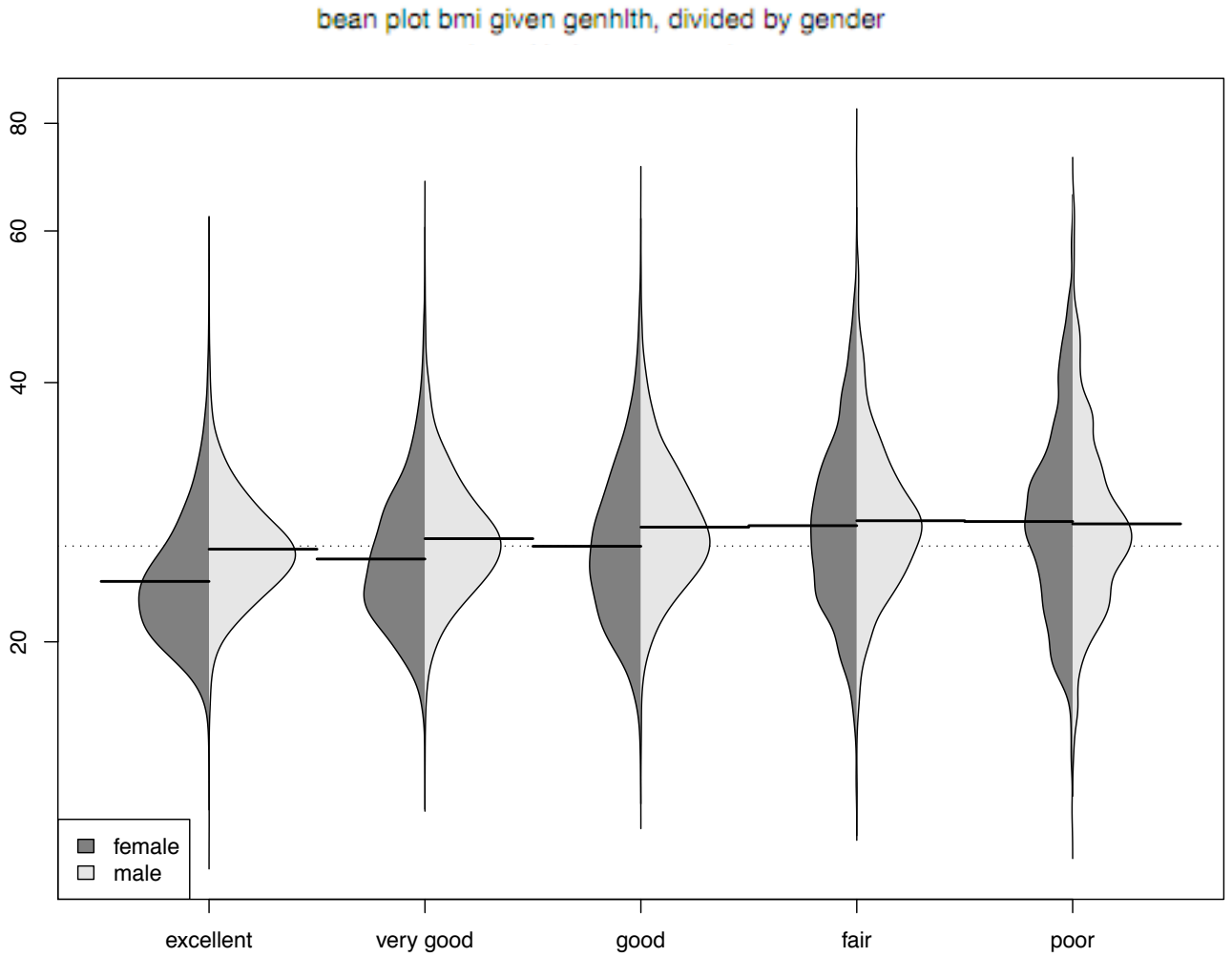
Compare this plot to the boxplot three slides back; what do you think?



Bean plots

This might be getting a bit farther out there, but with a relative of the violin plot (the so-called bean plot, named because the plots look like, um, beans) you can go farther and condition on, say, gender, producing a side-by-side plot

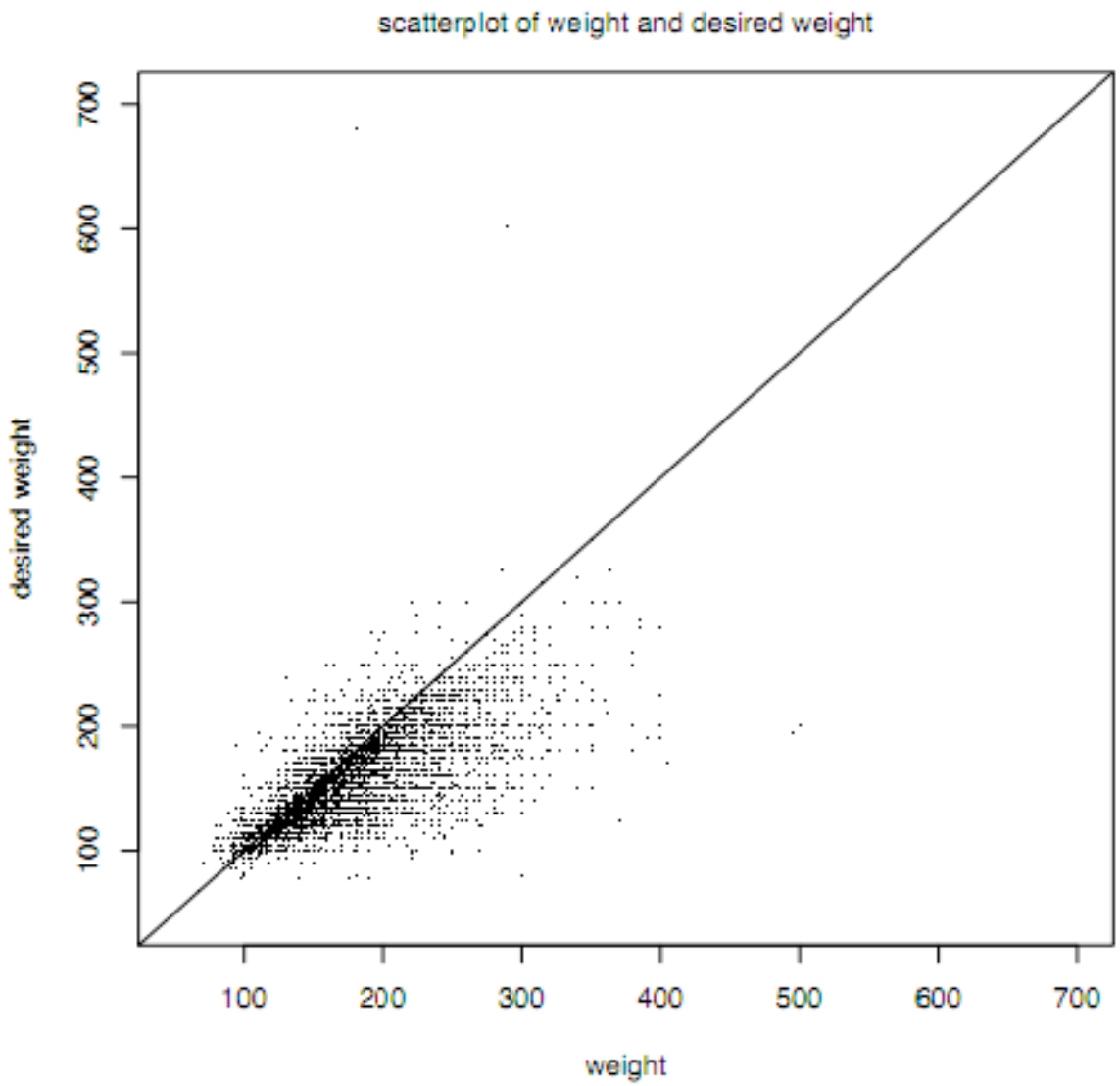
OK, now what do you think?



Extensions (II)

Now, let's consider another kind of extension to the boxplot; suppose we have two variables that we would like to describe

Take, for example, `weight` and `wtdesire`; we can create a simple **scatterplot** to look at how these two variables relate to each other


scatterplot of weight and desired weight
desired weight
weight
100
200
300
400
500
600
700

Scatterplots

We've added a line with unit slope to the plot; Why? What do you notice? What strikes you as expected? Unexpected?

Now, suppose we want to create something like a boxplot for these data; what concepts do we have to extend?