2. DISPLAYING DATA: FPP Ch 3
2.1 TYPES OF DATA VARIABLES

• A variable is a characteristic which may differ from subject to subject in an experiment or study.

• qualitative (not a number): red, green, blue
Math, Physics, Econ, Sociology, Romance languages

• quantitative (a number):
— integer: years of age, answer to "how many?",
counts of qualitative outcomes.
— discrete: can take only certain values
(but not necessarily integer)
— continuous: can take any value – e.g. height
Except in reality, all have some possible range (55" to 85") and limited accuracy (63.17893564335" ???)

• categorical: categories may be
— qualitative (nominal)
—ordinal: ordered categories; sometimes numerical
or first, second, third, ...
— interval : age ranges 20-35, 35-50, 50+

Blank page for your notes

2.2 THE BASIC HISTOGRAM

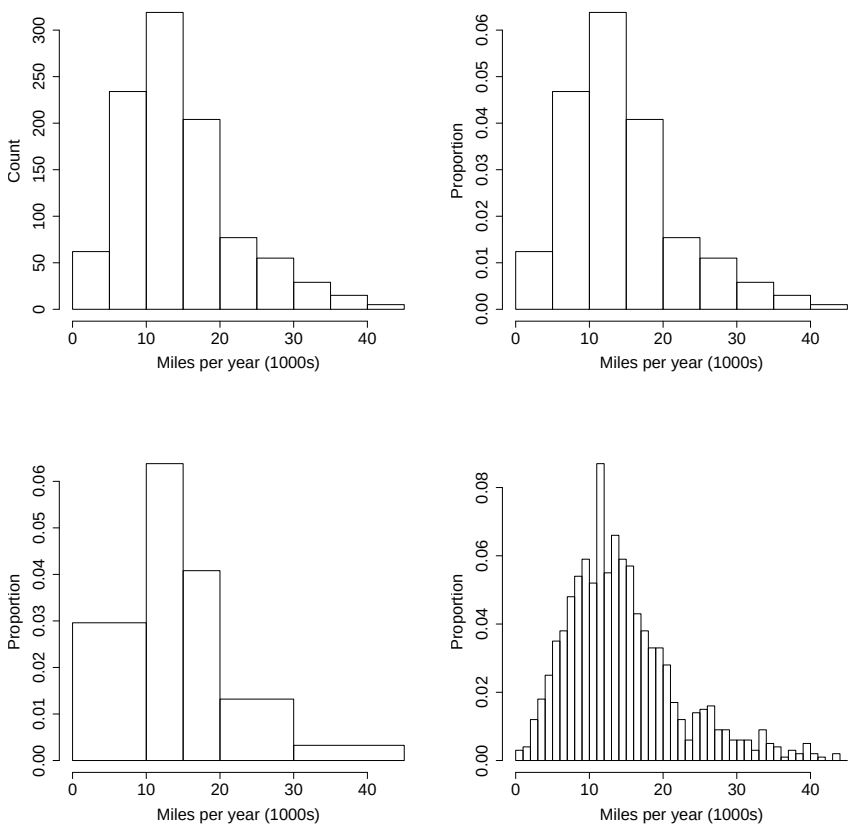
(Almost all data in these sections are "invented")

• Suppose data are counts or proportions (%) of subjects with measurements in specified ranges. For example:

Subjects	Measurements	Ranges
Households	Annual Income (in $1000)	0-20, 21-40, 41-60, 61-80, ...
UW students	Height (in inches)	50-54, 55-59, 60-64, 65-69, 70-74, ...
Days (Seattle?)	Precipitation (inches)	0, <0.1, 0.1-0.5, >0.5 dry, drizzle, wet, really wet
Cars	Gas usage Mpg (city)	<12, 13-19, 20-26, 27-33, >33
Cars	Miles driven (1000s per year)	0-5, 6-10, 11-15, 15-20, 21-30, 31-50, >50.....

- **Normally, the histogram bars should be equal width.**
- **Then, the height represents the count or proportion.**
- **If you do not want to use equal width bars then the AREA should represent the count/proportion.**
- **It is the SHAPE of the histogram that matters.**

HISTOGRAMS FOR MILES PER YEAR FOR 1000 CARS



2.3 ADDING INFORMATION

• Sometimes we want to add more information – for example we may know the make/kind of car/vehicle.

• Or the age of the primary driver. Suppose in sample size 1000, have 500, 200, 300 with drivers 45-59, 60-74, and 75+

• We can look at the separate histograms and compare.

• We had better use same scale on horizontal axis!!

• If we have histogram of counts, and use the same scale on count axis, we can compare total counts, as well as shape of distributions. However, some histograms may have small total counts.

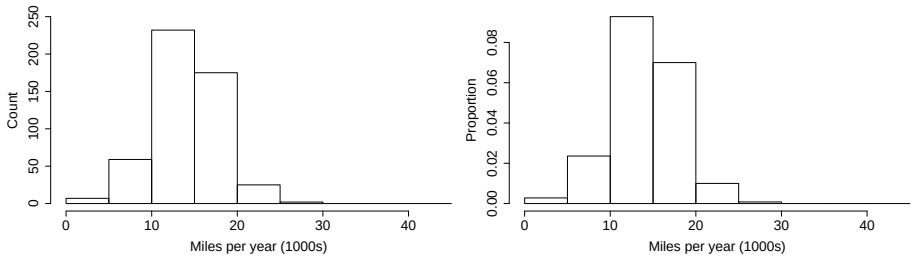
• In this case, AREA is proportional to COUNT

• If we have histograms of proportions, we can compare the different SHAPES (more easily?), but we have lost the information of relative counts.

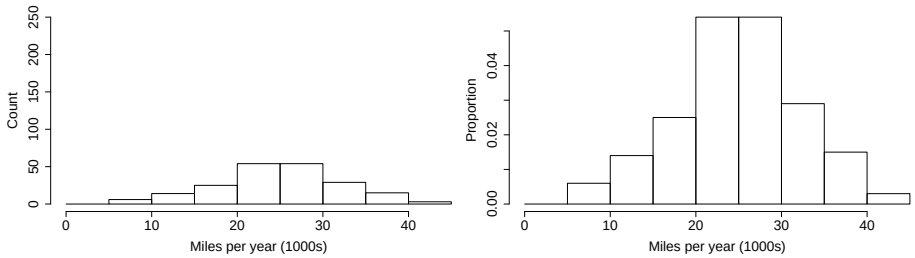
• In this case, AREA represents relative proportions within each group.

• In either case, it is not easy to compare visually a whole page of histograms!!

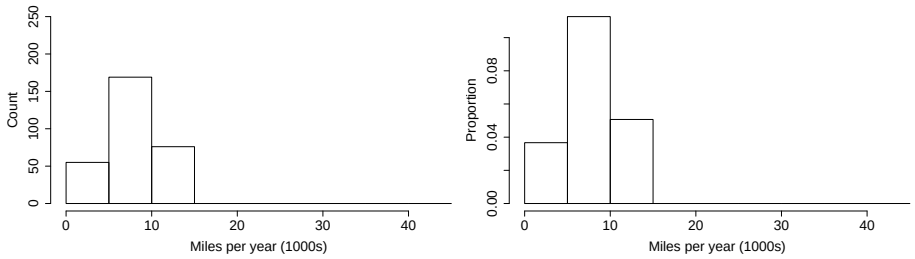
• Primary driver age 45 to 59 (total 500)



• Primary driver age 60 to 74 (total 200)



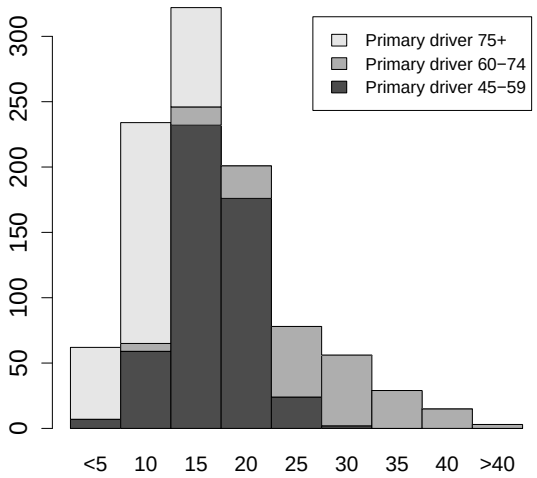
• Primary driver age 75 and over (total 300)



2.4 STACKED HISTOGRAMS

• **Sometimes is is better to to try to represent this information on a single histogram.**

• **We can shade the parts of the histogram bars — a "stacked" histogram**



• **This example is counts, but remember, if doing proportions it is the AREA that represents the proportion of the total sample.**

• **See also stem-and-leaf diagrams, below– that is another way of adding information for the histogram of a small sample.**

Blank page: Draw your histograms here!

2.5 STEM-AND-LEAF DIAGRAMS

• For a small sample, with a discrete, quantitative variable, a stem-and-leaf diagram is a useful form of histogram, that contains more information than a simple histogram.

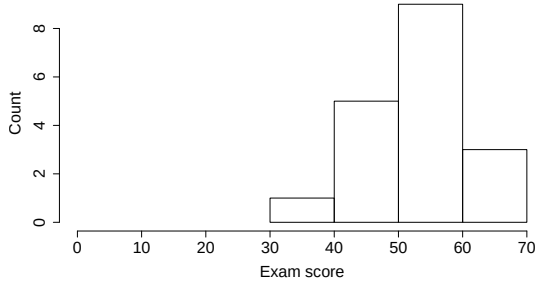
• For example, here are the final-exam scores of a set of 18 students in a small class: 57, 42, 38, 65, 44, 62, 55, 54, 55, 61, 58, 54, 59, 41, 45, 51, 49, 52.

• Find the range: min=38, max=65
(maybe the max possible was 70?)

stem	leaves
30s	8 |
40s	1 2 4 5 9 |
50s	1 2 4 4 5 5 7 8 9 |
60s	1 2 5 |

• Be careful not to overinterpret!
The boundaries are arbitrary, and not necessarily meaningful (38 is close to 41, 49 is close to 51).

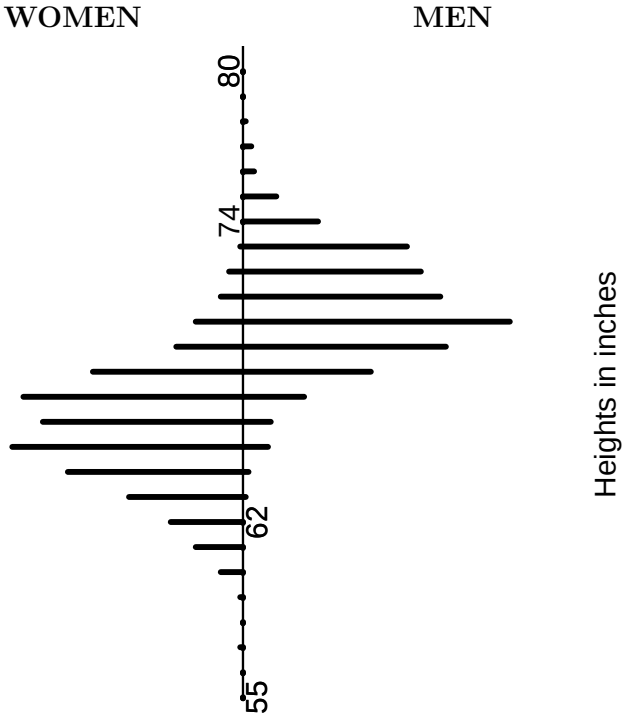
• But at least all the data are there to be seen.



Blank page for your notes

2.6 BACK-TO-BACK HISTOGRAMS

- **If there are just two categories (e.g. male/female), back-to-back histograms can be useful.**
- **For example, heights of 500 women and 500 men (Remember these data are NOT real data)**



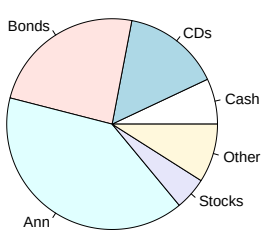
Blank page: Draw your histograms here!

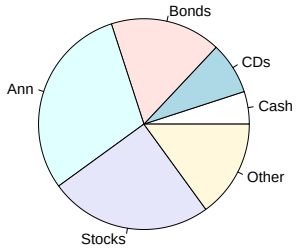
2.7 PIE CHARTS and BAR CHARTS

- **For displaying percentages of a small number of qualitative categories a pie chart is useful.**
- **If categories are ordered, a bar chart works well.**
- **Kinds of assets in a retirement portfolio (not real!)**

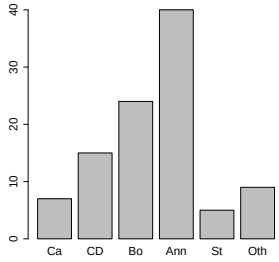
Asset type	Mine	Joe's
Money market/Cash	7%	5%
Certificates (CDs)	15%	8%
Mutual Funds (Bonds)	24%	17%
Annuities (Flex. Returns)	40%	30%
Securities(Stocks)	5%	25%
Other (Comm. Futures)	9%	15%

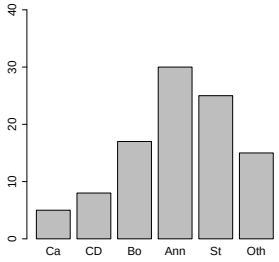
- **The pie charts**





- **Bar chart, ordered by risk.**





Blank page for your notes

2.8 CROSS-TABULATION

• Sometimes, to compare counts across categories, the simplest summary of the data is just a cross-tabulation.

• Back to the example of 1000 cars with miles-per-year and primary-driver-age.

Miles per year (1000s)	Age of primary driver 45-59	60-74	75+	Total
0-9	66	6	224	296
10-19	408	39	76	523
20-29	26	108	0	134
30-39	0	44	0	44
40+	0	3	0	3
Total	500	200	300	1000

• Note we have reduced the number of categories in the miles-per-year.

• Keep the table small enough to be manageable.

Blank page for your notes

2.9 HOMEOPATHY STUDY 1
BACKGROUND

•The Horizon homeopathic dilution experiment.
by Martin Bland, Significance, Vol 2, Pp 106-109 (2005)
Also, BBC2 program transcripts at http://
www.bbc.co.uk/science/horizon/2002/homeopathy.shtml

•A system of medicine with 2 principles:
1. The treatment agent should produce same symptoms as disease. The symptoms are body's defense.
2. Treatment agents are diluted by many orders of magnitude, for example 15C (or 10^{-30}) to 20C (or 10^{-40}).

•The question:
Can a solution that is so diluted it will have no molecules of the agent in it still produce a biological effect?
If so, this effect is known as "water memory".

•Previous evidence:
Well-respected scientists doing careful experiments were surprised to see positive results. (As well as others, not so careful and/or not so skeptical.)

•Horizon, the BBC2 Science Program in UK, decided to do a proper study, of the homeopathic effect of histamine on fresh ("live") red blood cells.

•Each outcome is a quantitative measurement of "percentage activity of basophil cells". A lower percentage activity for the histamine dilutions than the controls (pure water) is what is expected if the histamine is having an effect. (That is what it says?!?)

2.10 HOMEOPATHY STUDY 2
THE EXPERIMENT

•There are 40 "subjects" – test tubes of water, diluted and agitated identically:
20 started as pure water (controls)
other 20, were of histamine, with 5 each of dilutions at 15C, 16C, 17C and 18C (10^{-30} to 10^{-36}).

•They were shuffled, relabeled, and the key to the relabeling locked in a safe.

•Each tube was divided in 2 parts, and half sent to each of 2 Labs.

•Each lab had 5 (human) volunteers who gave fresh blood.

•Each lab, divided each of the 40 samples into 5 parts,

•In each lab, each of the 40 samples was applied to the blood of each of its 5 donors.

•For our analyses, it is enough to consider, for each sample, the average over the five donors.

•So each lab produced 40 quantitative measurements.

•First, for each lab, these were simply divided into lowest 20, and highest 20.

•Then, live, on air, each sample was unblinded as Control (pure water) or Dilution (treatment), and the counts in the top and bottom 20 measurements for each lab were scored.

2.11 HOMEOPATHY STUDY 3 OVERALL CONCLUSIONS

- **Summary results:**

	Lab 1		Lab 2	
	bottom-20	top-20	bottom-20	top-20
Control	11	9	9	11
Dilution	9	11	11	9

- **Martin Bland (Statistician):**
  "There is absolutely no evidence at all to say that there is any difference between the solutions that started as pure water and the solutions that started off with the histamine."

- **John Enderby (Adjudicator, Royal Society VP):**
  "What this has convinced me of is that water does not have a memory."
  (Do you agree with John Enderby ??)

- **We will see more of this study later. For now:**
  Placebo: control and treatment tubes diluted and agitated identically.
  Blinding: The fact of blinding labs, and also (until live on air) data analyzers.
  Randomization: The random shuffling and labeling of 40 original tubes.
  Replication: half of each tube to each lab.

Blank pages for your notes