STAT 311: Lecture 2

Summarizing Data:

- Types of Variables
  - Categorical/ordinal; Quantitative/continuous.
- Graphical summaries of categorical data
  - Pie charts, barplots.
- Graphical summaries of discrete data
  - Dot plots, stem-and-leaf
- The 5-number summary of quantitative data
- Graphical summaries of quantitative data
  - Histograms, Box plots
- Two example graphics with strong visual impact.

Where we are at:

- Canvas: A web course page temporary alternative:
  - Go www.stat.washington.edu/thompson/Stat311
  - It will take you to a page with some links to lecture slides and lab files for Lab1 etc. at bottom of the page.
  - Canvas syllabus page is now public through UW MyPlan.
- Aplia: as of 10:30 p.m. yesterday (Tuesday)
  - 153 people signed up to the course successfully
  - A few people already started on the practice ("Graded:excluded") assignment, and on the graded homeworks!
- Add codes: a few more will be sent out

  If emailing about add codes, PLEASE USE or GIVE UW Id.

Section changes -- not unless you cannot register without, and not if a new lecture add code needed –the few add codes must be saved for genuine adds.

View forward for the week

- Wednesday – Graphical summaries of data. (U/H 2.1-2.4)
- Thurs/ Friday -- Aplia practice and more practice, Also R.
  - “Graded” (but not counting) Math prep assignment “due” Friday
- Friday – Numerical summaries of data (U/H 2.5-2.7)

  – more material needed to complete Hwk 1 and lab 1
- Monday – relationships between two variables (3.1, 3.3)
- Monday 11:00 p.m.: first actually graded homework is due
  - Two parts – 1a: relating to U/H Chapter 1,
  - -- 1b: relates to U/H Chapter 2 (mostly 2.1-2.4)
  - Aplia scores separately, but it will count as single homework grade. (and policy is drop lowest weekly score).
- Tuesday -- quiz section -- more R towards lab 1
- Tues 11:00 p.m.: Lab 1 is due.
- Wednesday: Linear regression (U/H 3.2)
- Thursday -- Quiz section, starting towards Lab 2

Types of Variables

- Categorical: (U/H Chapter 4)

  No logical ordering to the possible values.

  Examples: eye color, nationality, types of investments.

  - Ordinal: Categorical variables for which there is an ordering. Examples: No/Yes (workdays lost to flu)

  Year in college (Fr, Soph, Jun, Sen), T-shirt size (S,M,L,XL)

- Quantitative: (U/H Chapter 3)

  Numerical values for each observation.

  - Discrete: Take only a few (?) possible values.

    Example: Number of cousins, Number of accidents.

  - Continuous: Can, in principle, take any value in a range. Examples: Body temperature, rainfall amount.

  Note the accuracy with which we measure a variable may be limited (e.g. rainfall in 0.01”).

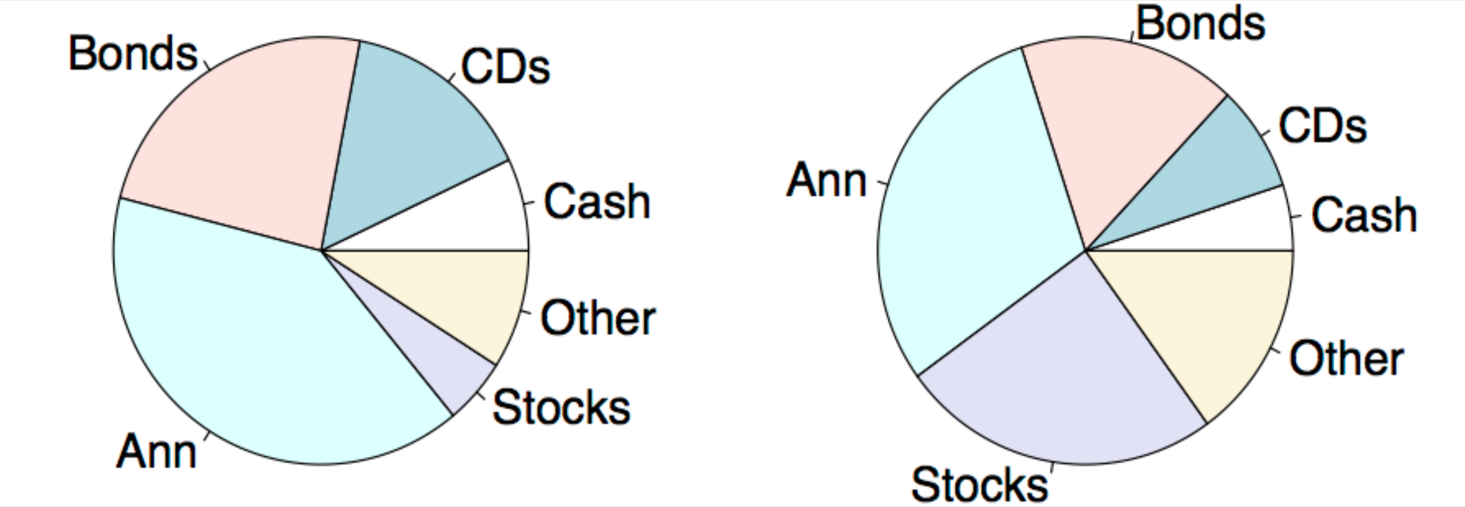
Categorical variables

- The measurement on a case (Observational unit) is a color, or other descriptive type.
- Data are counts or proportions in each category.
- Example: Types of investments in a retirement portfolio:

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

PIE CHARTS

- The easiest way to represent these data is with a pie chart.
- The AREA represents the proportion in each category.

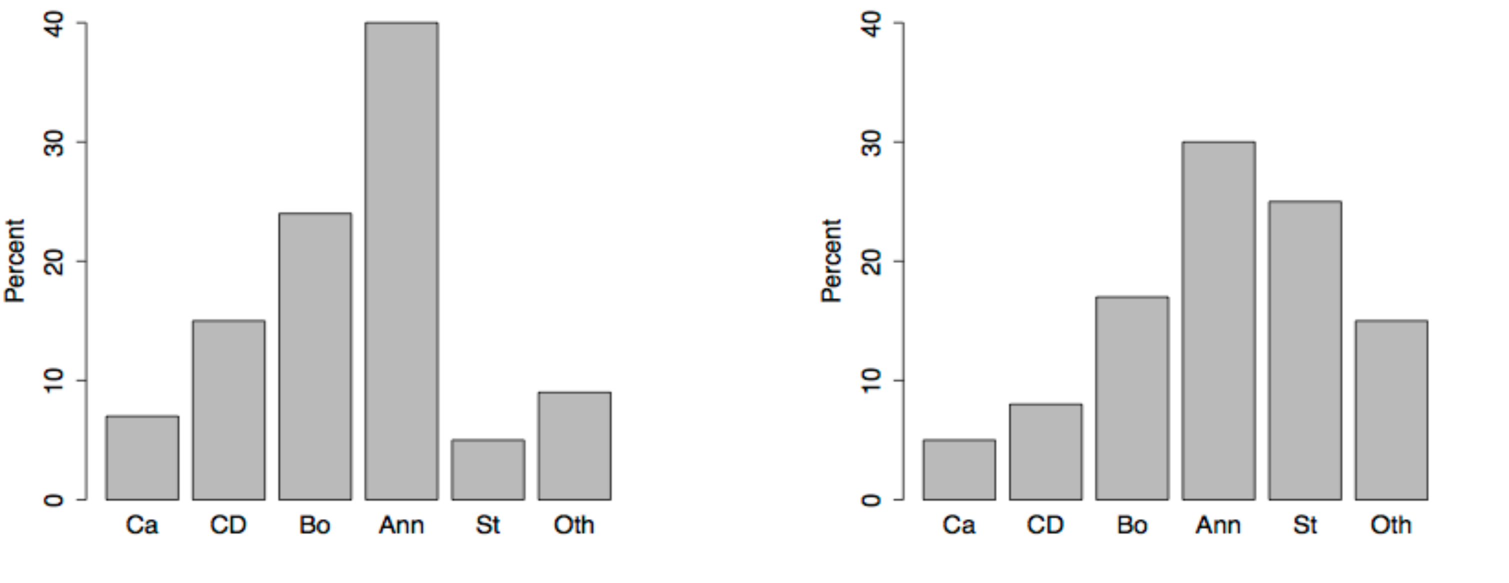


- I have more annuities (as a proportion)
- Joe has more stocks
- But so what?

Ordinal data: The Bar plot

- The investment types are ordered by risk.
- Ordering the categories in a bar plot makes sense.

MINE JOE’s



- Now we see Joe’s portfolio is riskier than mine.
- But also that I am not ultra-risk-averse (not all Cash/CD)
- Or we can color my/Joe’s bars and place alongside – see example 2.4 in U/H textbook.

U/H Ex 2.4: Nightlights and nearsightedness

- Survey of n=479 children
- Response: degree of myopia
    - 3 categories represented by colors.
- Explanatory variable: amount of sleeptime lighting as babies
    - 3 categories across the x-axis.

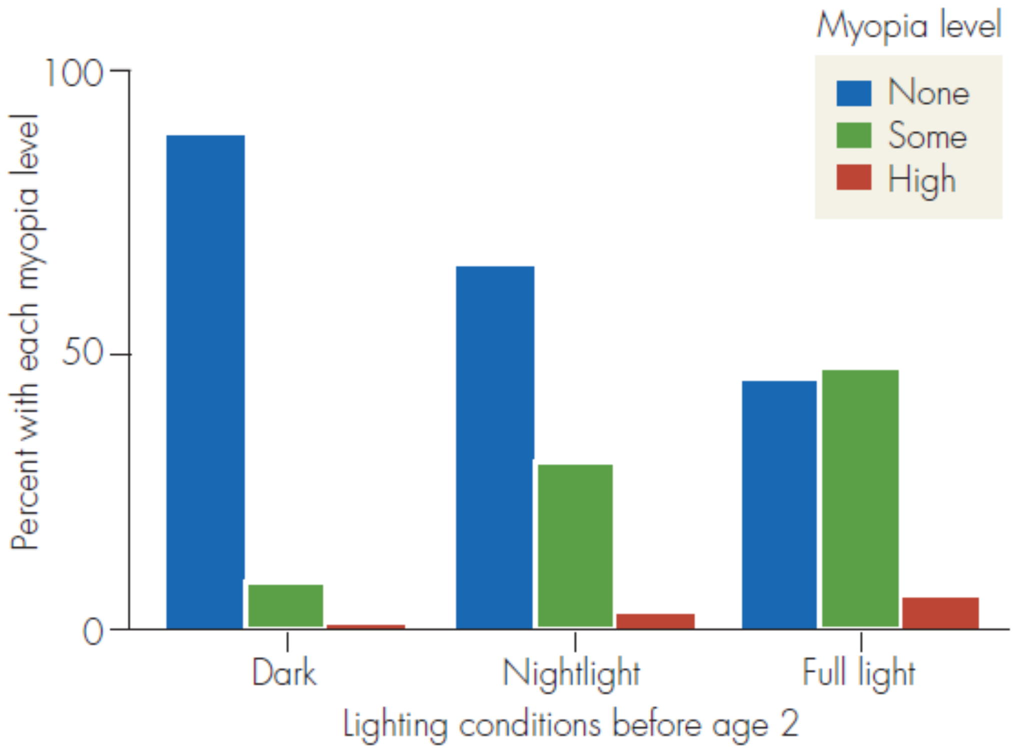


Figure 2.4 Bar chart for myopia and nighttime lighting in infancy

Discrete quantitative data

- Data with only a few values can be represented as a dot plot or stem-and-leaf diagram.
- Example: right hand-spans of 103 female students (2.5 in U/H)

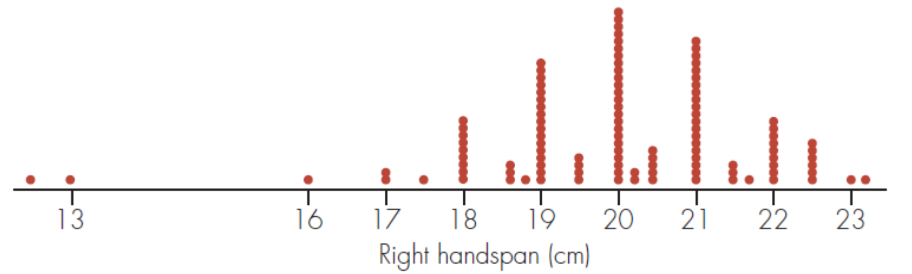
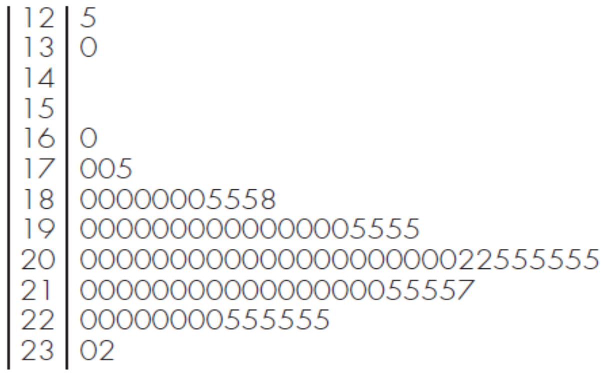


Figure 2.9 Dotplot of females' right handspans

- Values (cm) are on x-axis
- Every observation gets a dot.
- Pile-up the dots for repeated values.



12	5
13	0
14	
15	
16	0
17	005
18	00000005558
19	0000000000000005555
20	00000000000000000000022555555
21	0000000000000000055557
22	00000000555555
23	02

Example: | 12 | 5 = 12.5

Figure 2.8 Stem-and-leaf plot of females' right handspans

- The stem represents coarser scale (here cm)
- Each data point gets a listing as a "leaf" to the right of the stem.
- The digit of the leaf is 0.1 cm (mm).
- Do NOT over-interpret 18.8 is close to 19: the boundaries are arbitrary.

The 5-number summary (or 6)

- How to summarize these data??
  - Most values are around 20 cm.
  - Two values are very low ("outliers")
  - Apart from these: range is 16 to 23 cm.
- More generally??
  - (1) The median: the "middle" of the set of values:
    ~50% are below, ~50% are above
  - The quartiles:
    (2) Q1: Lower quartile: ~25% are below, ~75% are above
    (3) Q3: Upper quartile: ~75% are below, ~25% are above
  - The extremes:
    (4) Minimum: the smallest value
    (5) Maximum: the largest value
- (6) The inter-quartile range: IQR = Q3-Q1

- In the female hand-span example: median=20cm, quartiles = 19cm, 21cm: min=12.5cm, max=23.25cm

Histograms

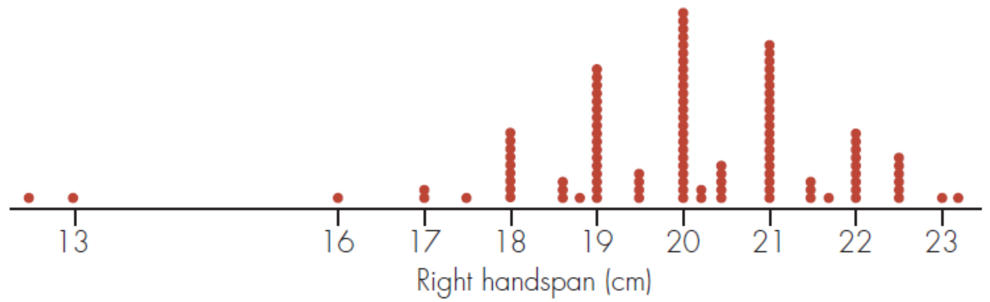


Figure 2.9 Dotplot of females' right handspans

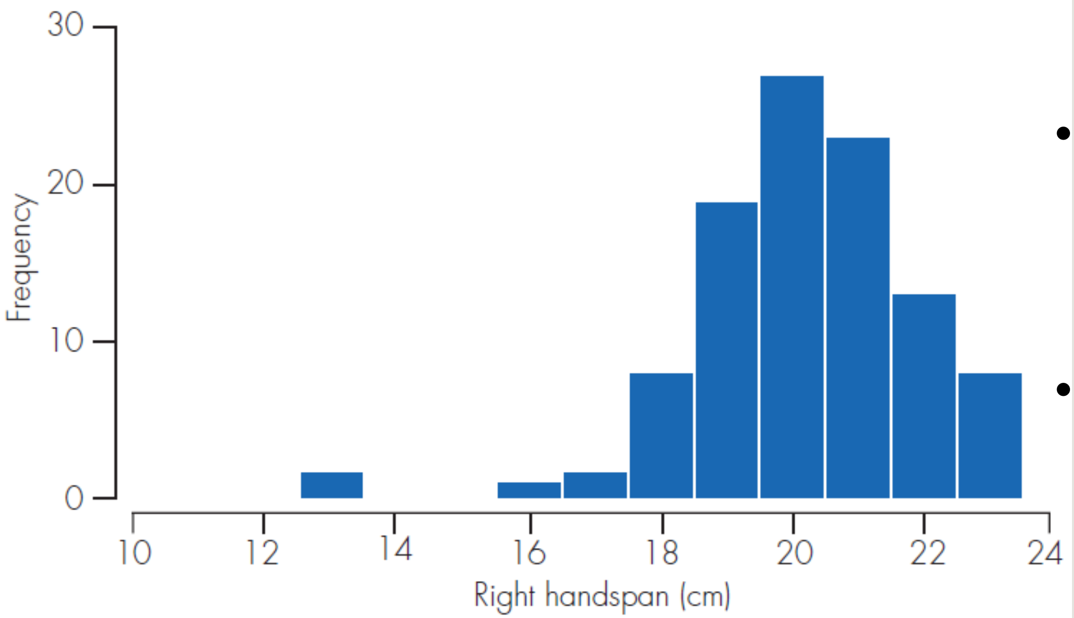
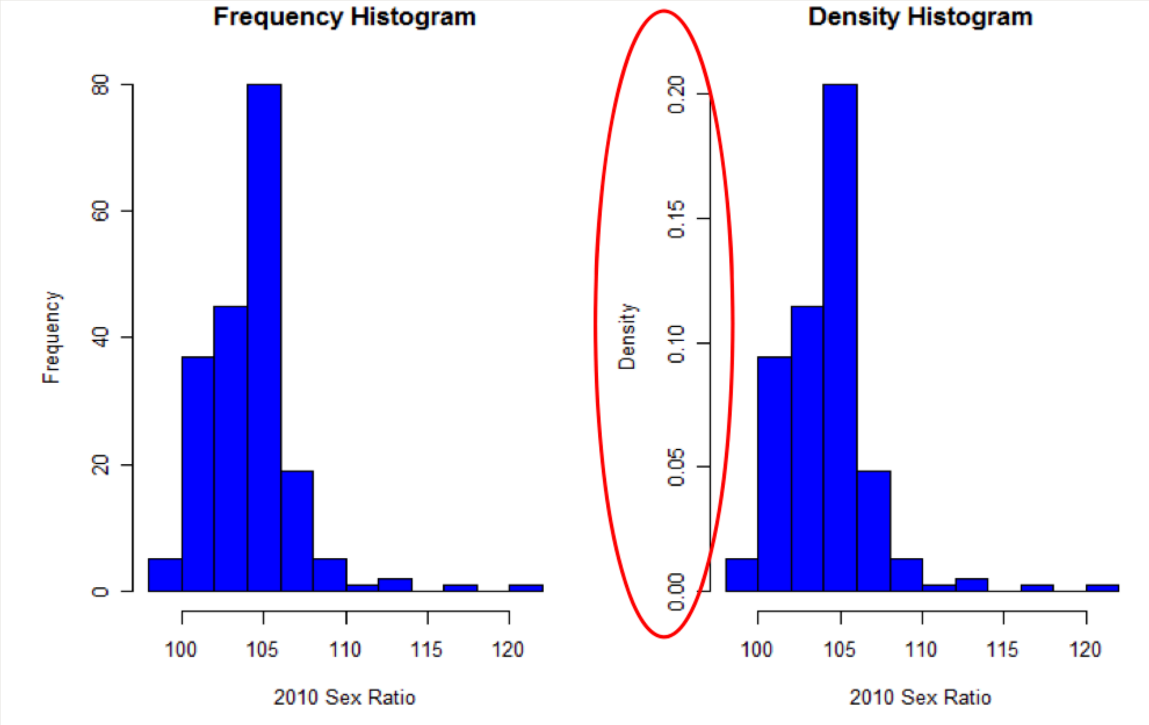


Figure 2.7 Histogram of females' right handspans

- Back to the example of female hand spans
- Above is the dot plot
- Below is the histogram.
- Which do you prefer?

- Note we bin the data: Each bin is width 1 cm, centered on the integer values
- Need to choose the widths and boundaries of bins

Counts or proportions?



- Left:
Vertical axis is count or frequency.
(2 words; same thing).

- Right:
Vertical axis is percentage, proportion, or density (3 words, equivalent things)

- The only difference is what is on the vertical axis.
- Plotting density instead of frequency is useful when comparing samples of different size.

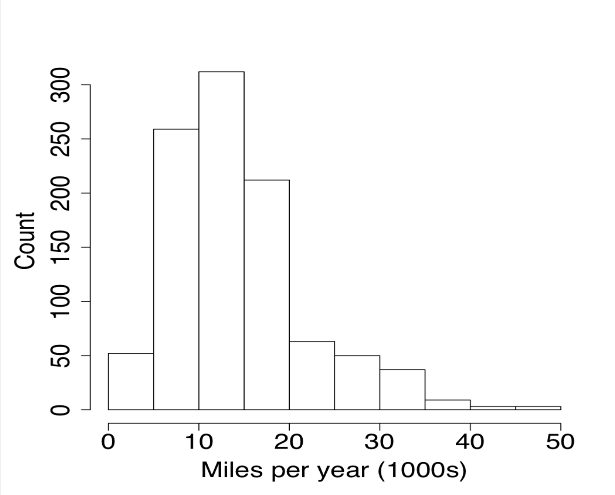
Basic rules for histograms

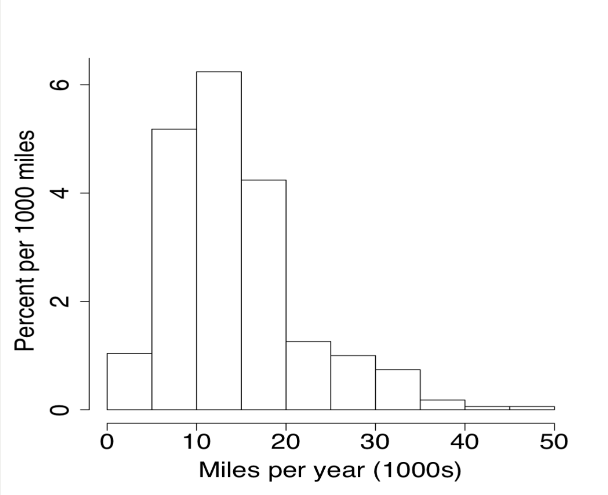
- Can be used for any quantitative data.
- Normally, the histogram bars should be equal width.
- Then, the height represents the count (frequency) or percentage or proportion (density).

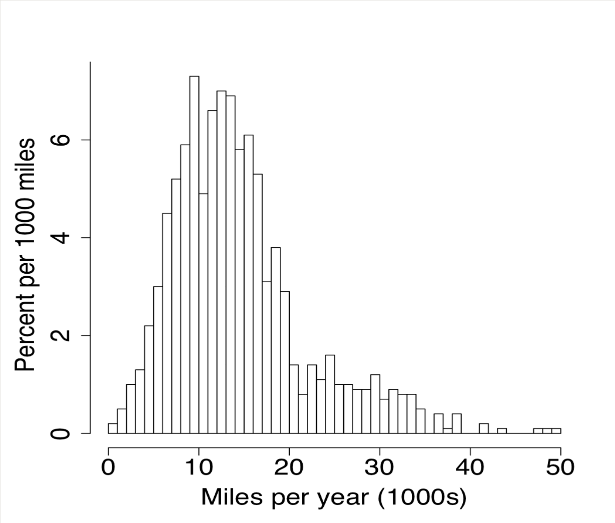
- ALWAYS, the AREA represents the count/ percentage.
- For percentages: the total AREA is 100%.

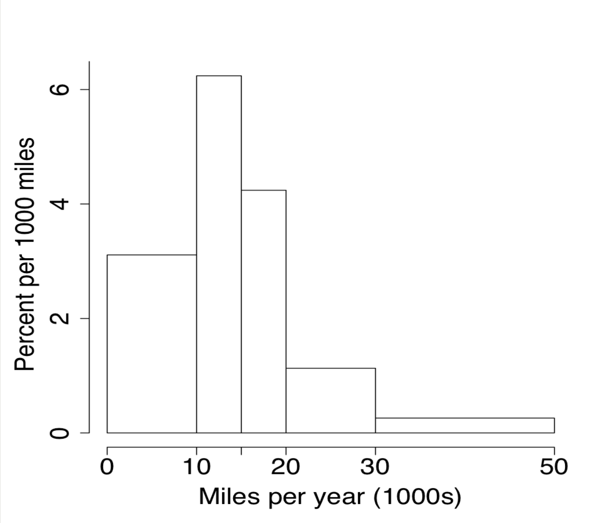
The next page shows four histograms of miles driven per year in a sample of 1000 cars ("main" vehicle in 1000 households).

Four histograms of same data









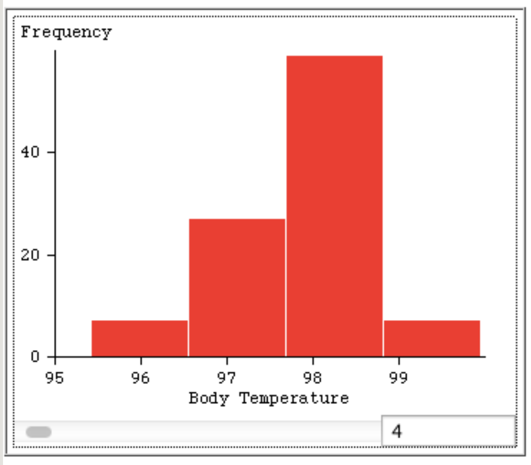
1. Count histogram

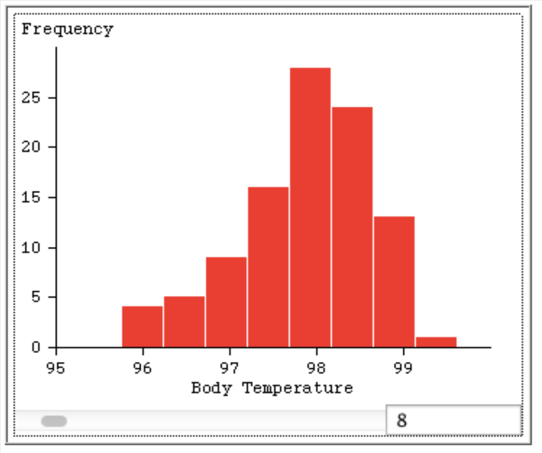
2. Density per unit x-axis
   Unit: 1000 mi.
   Bin: 5000 mi.

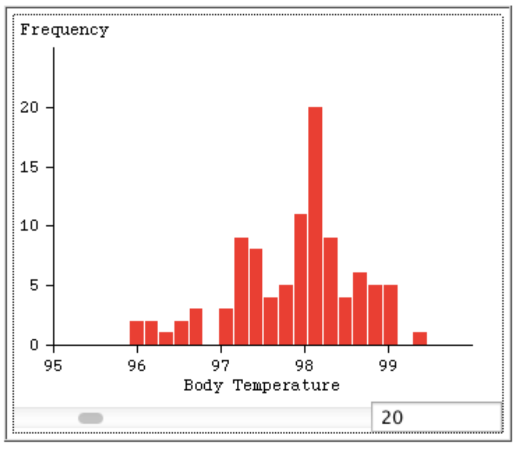
3. Same, but with unit width bins

4. Merging bins unequally.

How many histogram bins?







- Body temperatures (^{0}F) of 100 students.
- Range is 96 to 99.4
- Which histogram do you prefer? And Why?
  - Sample vs population
  - Purpose of measuring the variation
- Does the choice depend on n? (Here n=100)
- It is the SHAPE that matters – which choice gives best idea of shape of distribution in the population?

Histograms vs Barplots

- Barplots/dotplots (Ordinal/Discrete)
  - have a bar for every value in the data
  - Works for discrete variables with a small number of possible values
- Histograms (Quantitative)
  - Divide the data values into ``bins" (normally equal width)
  - Plots the frequency (count) or density (percentage) of data values in each bin
  - Works for both continuous and discrete data

U/H Ex 2.4: Nightlights and nearsightedness

- Survey of n=479 children
- Response: degree of myopia
  - 3 categories represented by colors.
- Explanatory variable: amount of sleeptime lighting
  - 3 categories across the x-axis.

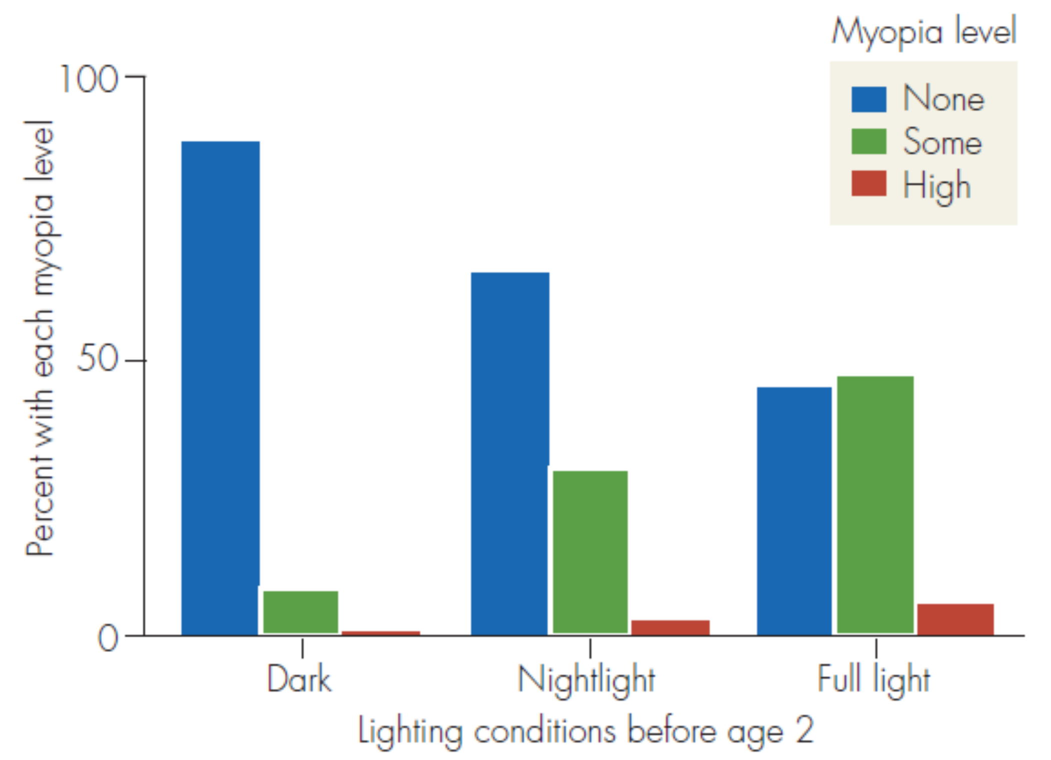
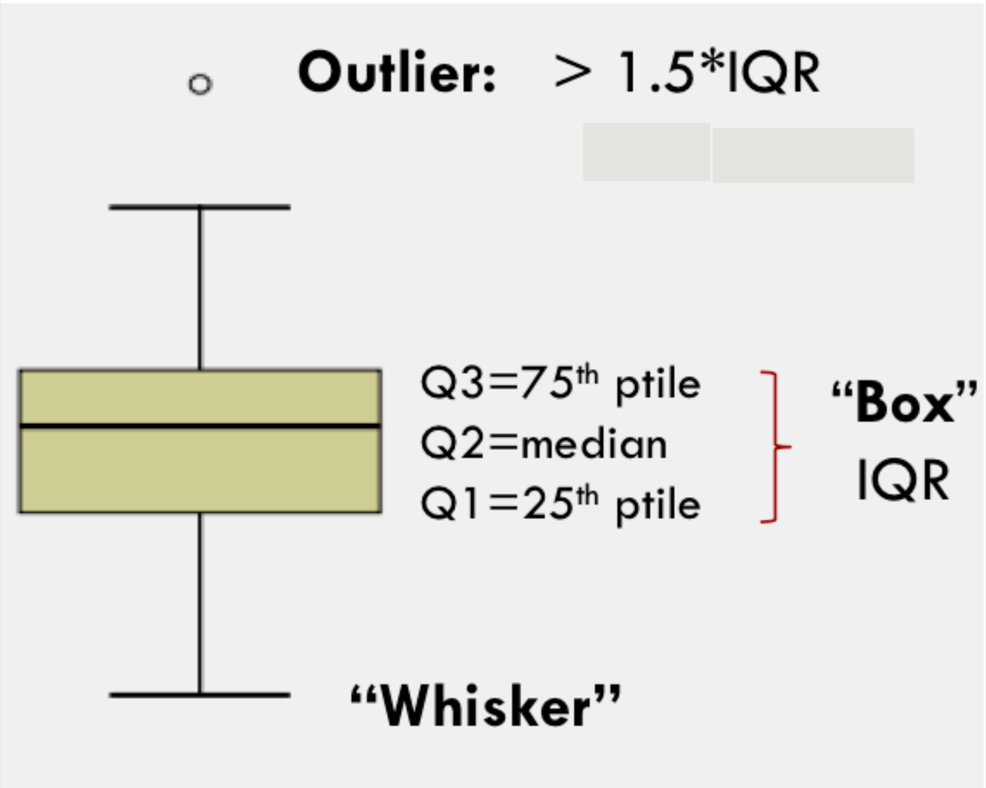


Figure 2.4 Bar chart for myopia and nighttime lighting in infancy

- What is the difference between a bar chart and a histogram
  - Sometimes not much!!

Box plots: defining the box



- The box is defined by the median and lower and upper quartiles.
- The "whisker" extends to either the max/min, or to 1.5 times the IQR below quartile Q1 or above Q3.

- Points beyond 1.5 times the IQR above/below the relevant quartile are often called outliers.
- In the example:

  No low outliers – lower whisker extends to minimum.

  One high outlier, upper whisker extends to Q3 + 1.5 x IQR

Box plots: Example 2.5 of U/H

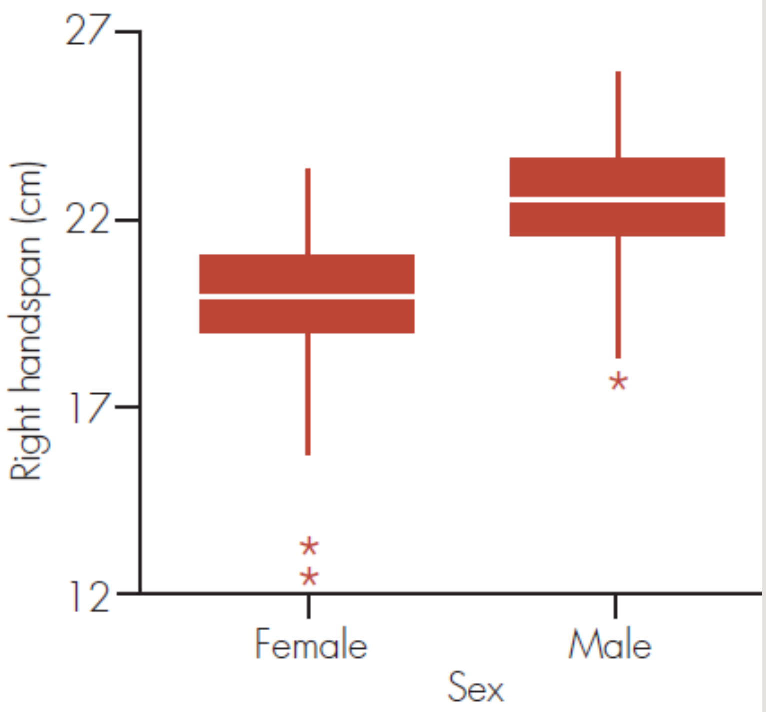


Figure 2.14 Boxplots for right handspans of men and women

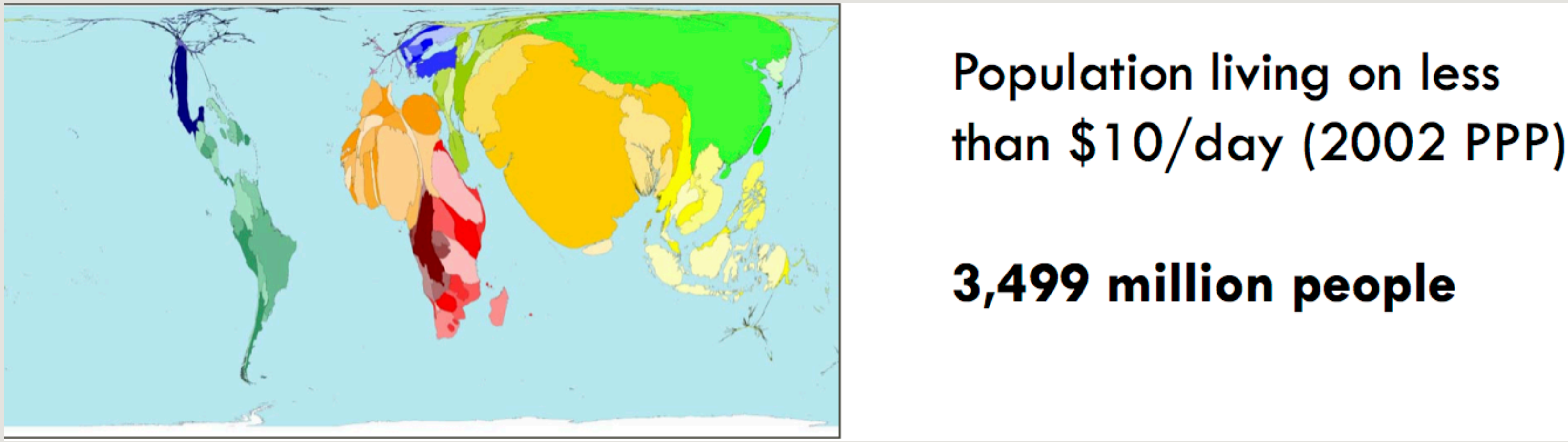
- Right handspans of 190 college students: 103 female, 97 male.
- Females we have seen before: centered at 20cm , most in range 19 to 21cm, range (excluding 2 outliers) is 16cm to 23 cm
- Now we see, male distribution very similar, except 2.5 cm larger.

- With boxplots (as with density histograms) we can compare samples of different sizes.

Which for what variables?

	Pie chart	Barplot	Stem& Leaf or dotplot	Histogram or Boxplot
Categorical	YES	(yes)	NO	NO
Ordinal	(yes)	YES	NO	NO
Quantitative (few discrete values)		(yes) U/H says no	YES	(yes)
Quantitative (continuous)				YES

Graphic with impact: #1



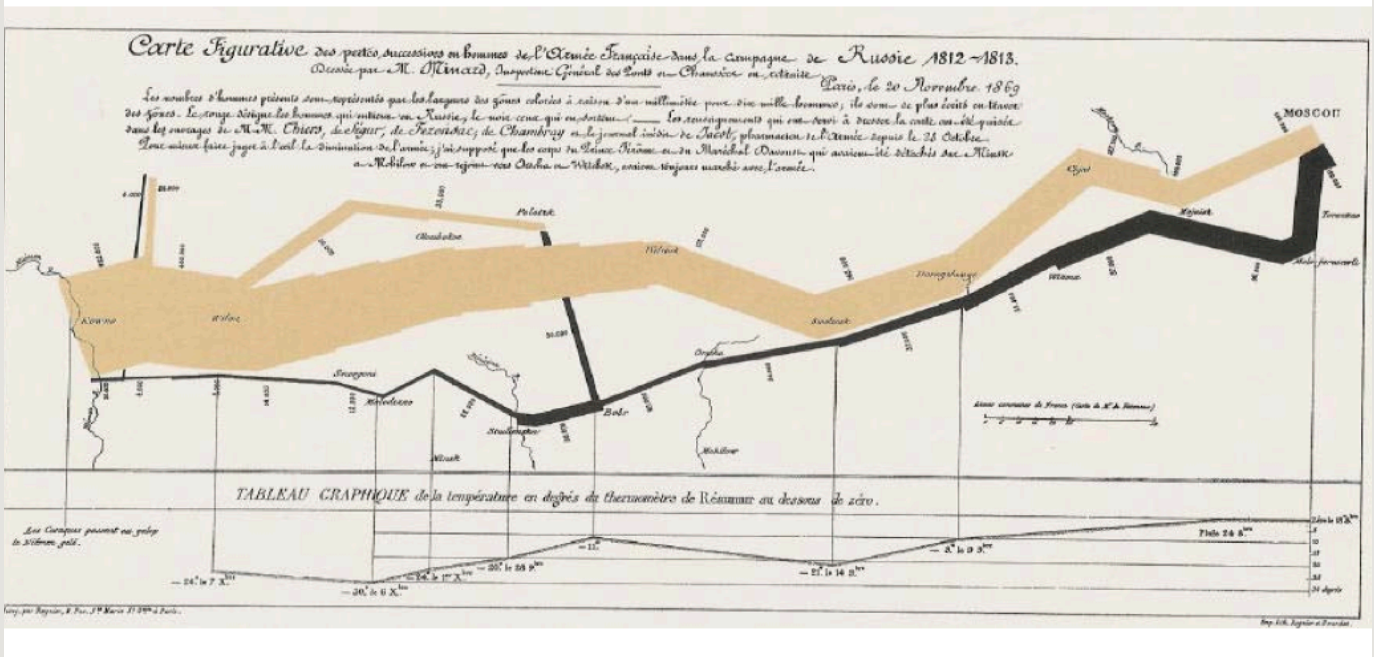
Population living on less than $10/day (2002 PPP)

3,499 million people

- Source: http://www.worldmapper.org/textindex/text_Income.html

- What do we need to know to interpret this graphic?
  - The normal visual world map
    - – North America, West Africa
  - Population densities around the world
    - – what about S. America, Nigeria?

Graphic with impact: #2



- Charles Joseph Minard's 1869
  map of Napoleon's 1812 Russia Campaign
- Source: Wainer, Visual Revelations, p.85
- See also: http://www.datavis.co/gallery/re-minard.php