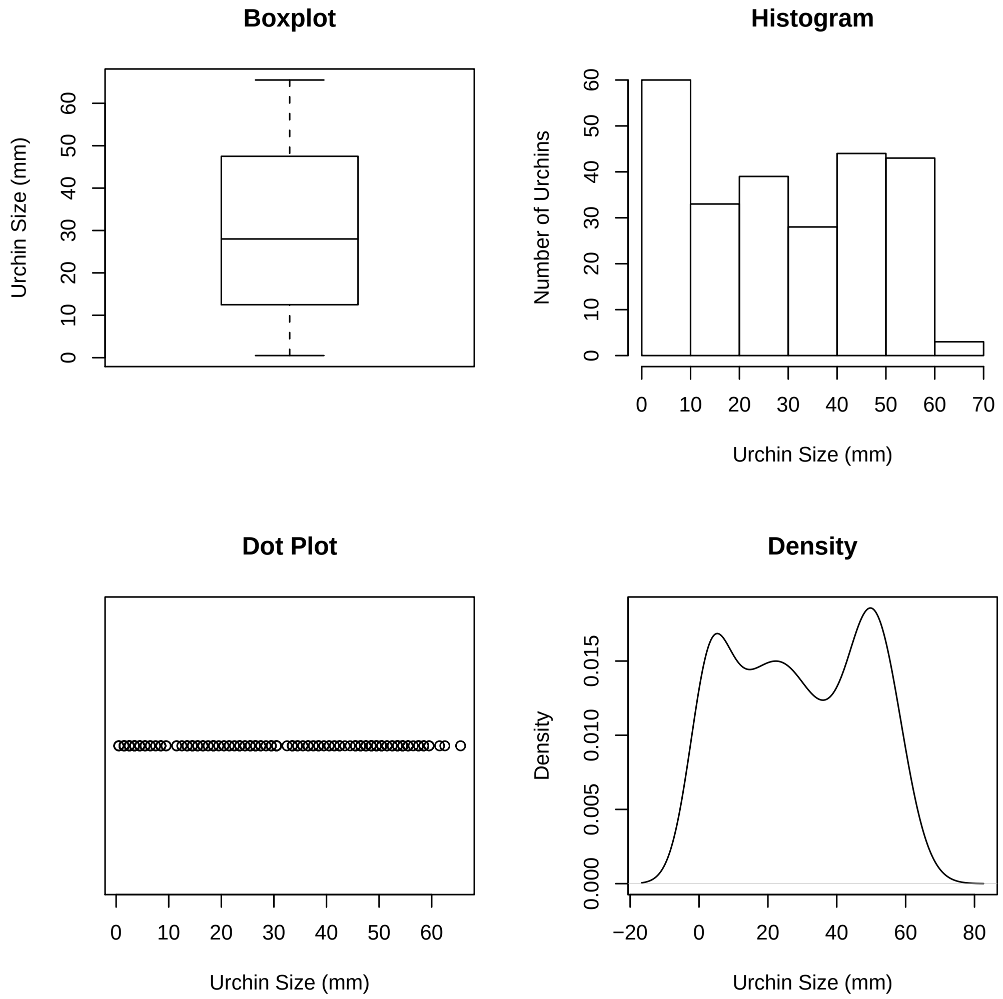
Univariate Descriptive Statistics

Displays: pie charts, bar graphs, box plots, histograms, density estimates, dot plots, stem-leaf plots, tables, lists.

Example: sea urchin sizes



Points:

1) Useful for quantitative variables.

2) Boxplot shows five point summary: minimum, first quartile, median, third quartile, maximum.

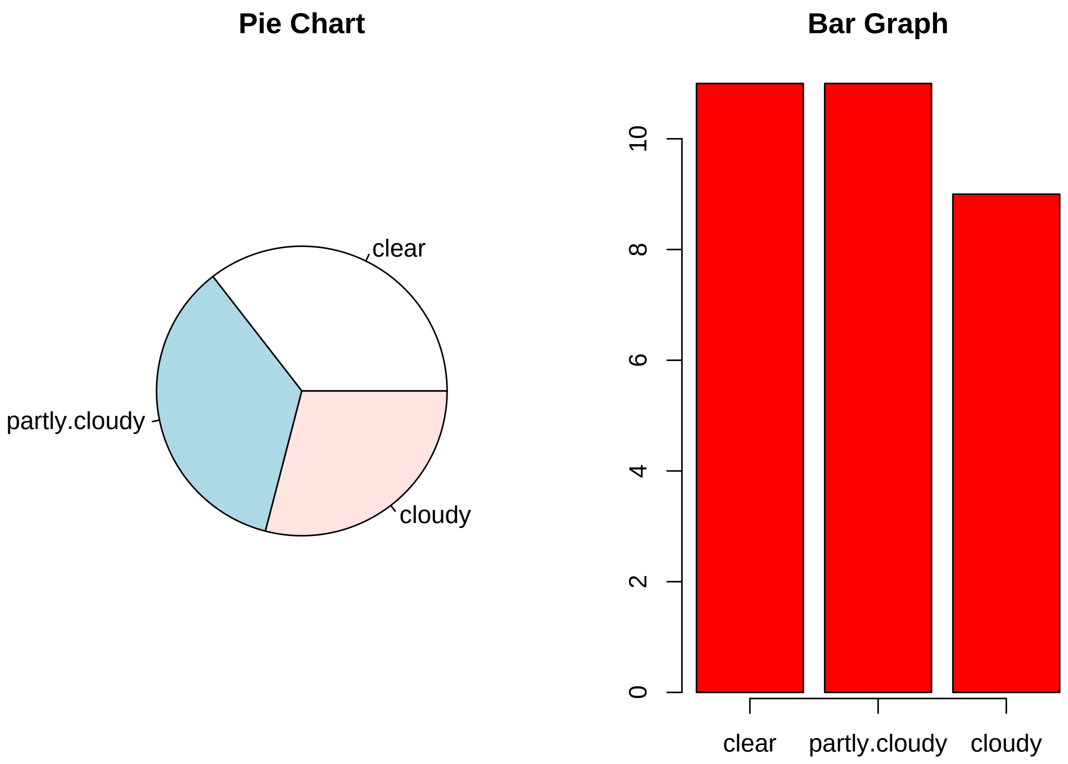
3) Dot Plot illegible with 250 data points. (1 dot for each size plotted on line.)

4) Histogram, density plot serve similar purposes.

5) Density goes below 0: bad.

6) Histogram doesn't show clustering density plot shows.

Example: Categorical: Weather in Central Park



Pie chart harder to read.

General summary: Pie Charts are bad.

More useful with more categories.

Ordering of categories important for nominal variables.

Cloudiness is ordinal.

Pie charts: wedge has area proportional to # of individuals in category.

Bar chart: bar has height equal to # of individuals in category.

Density estimates not discussed in this course.

Histogram:

1) divide range of values into intervals.

2) Count numbers of individuals in each interval.

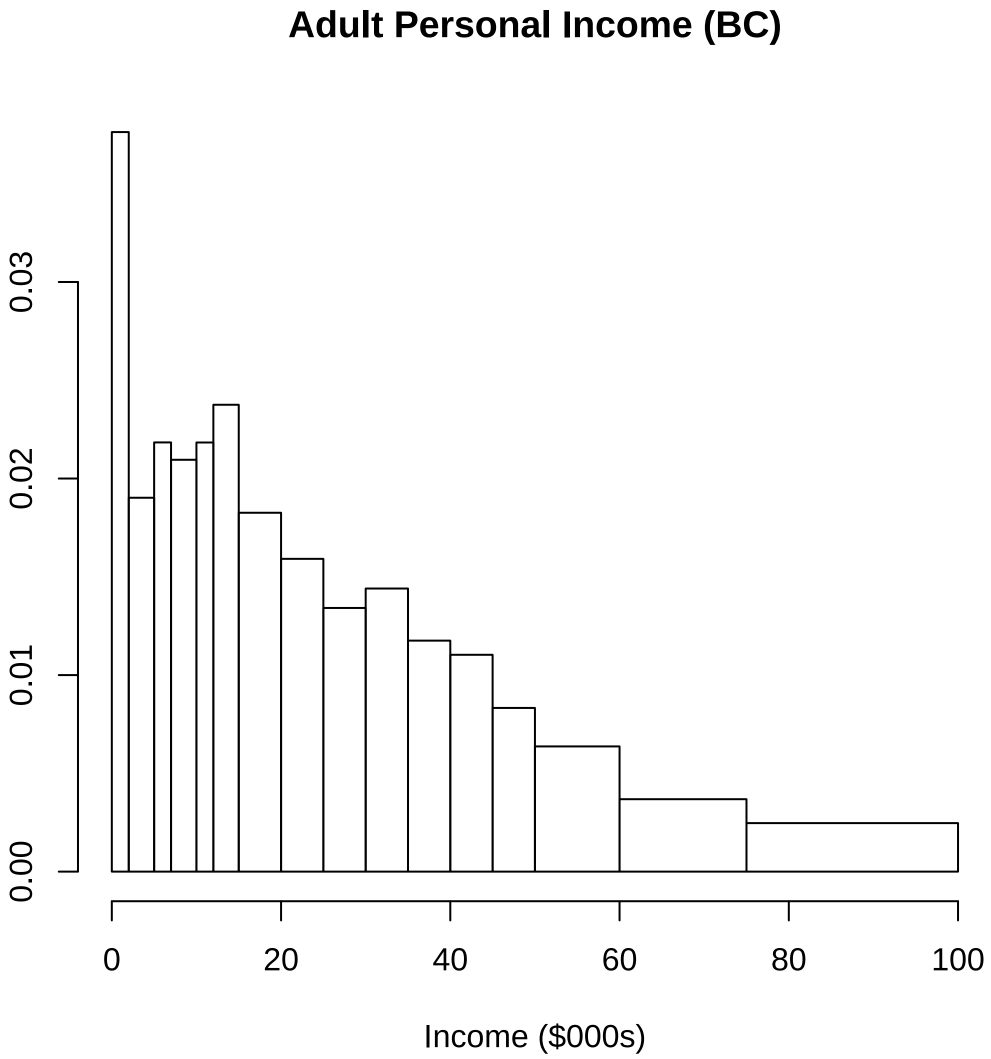
3) bar AREA is proportional to # of individuals in interval; width is length of interval.

4) equal width bars best – then height proportional to # of individuals.

5) label x-axis; include units.

6) label y-axis.

Example: Personal Income for BC (ages 15+). (For those with income.) Source: 2001 Census.



Points

1) Bar widths unequal – census tables given that way.

2) So take width times height to get area = fraction of population in that income group.

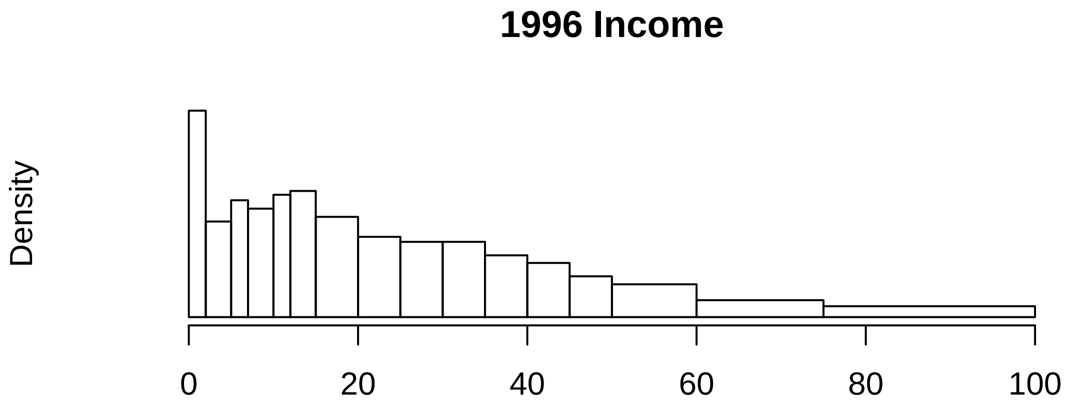
3) Last group on right open ended – artificially cut off at $100,000 by me.

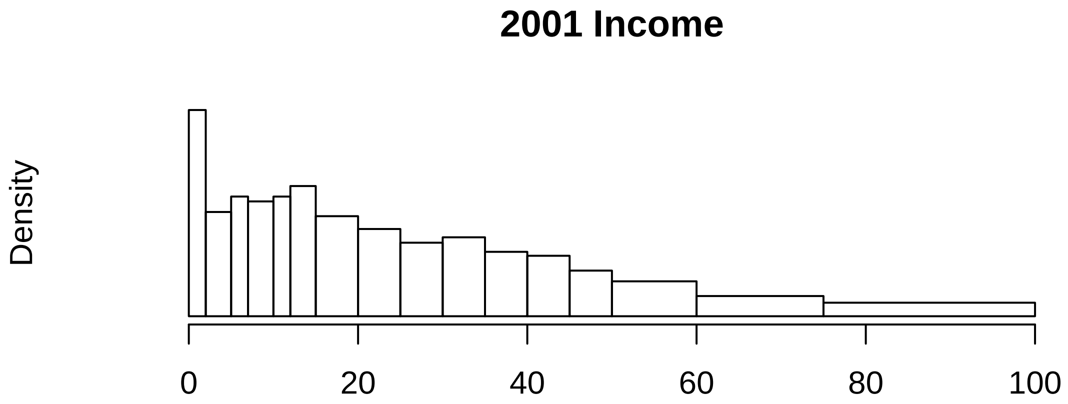
4) Plot is “long-tailed to the right” or “skewed to the right”.

5) Based on 20% sample of 1,523,720 people aged 15 + in BC on census day, 2001.

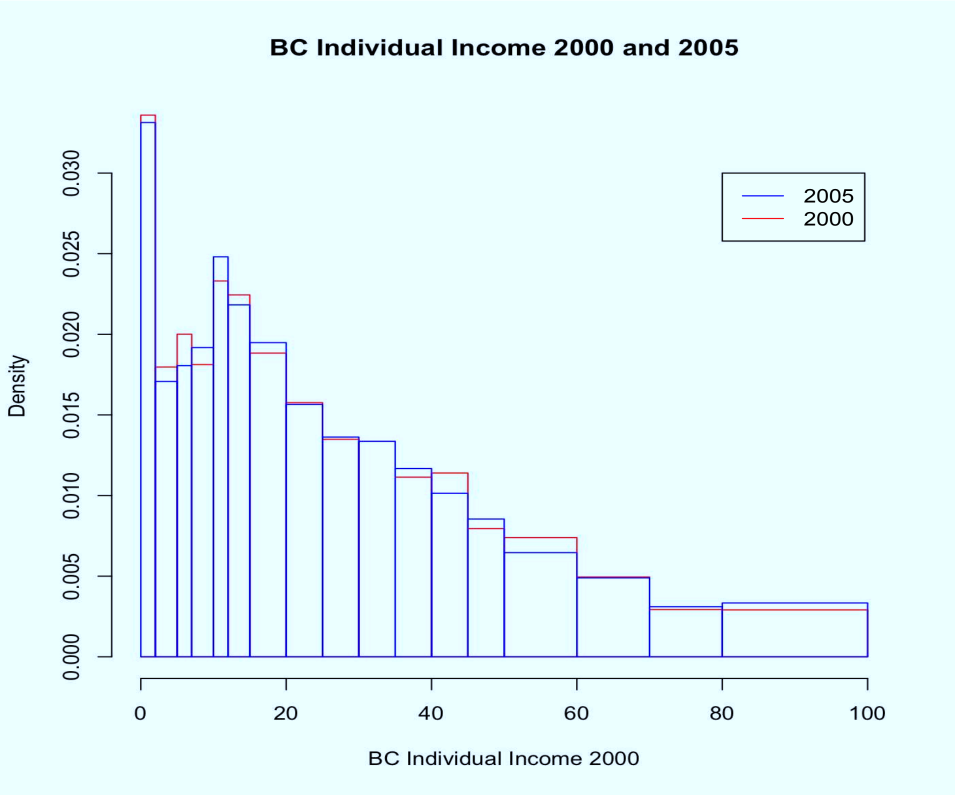
6) Income is for previous year – 2000.

Comparison of 1995, 2005.





Comparison of 2000, 2005.



Summarizing the pictures.

Purposes: less space in text than a graph; precise numerical comparison between groups.

Summarizing a histogram:

Where is centre of the x-axis values? Jargon: **location** or **centre**.

How far do the x values extend on either side? Jargon: **spread**, **variation**, **width**.

Is the picture symmetric or does it extend farther to right than left?

Location and number of bumps.

Measures of location:

Mean, **Arithmetic Mean**, **Average**, **Arithmetic Average**: total of x-values divided by number of x values.

Histogram balances at mean. (**First Moment** in physics.) Think of See-Saw: small kid far from centre balances big kid close to centre.

Formula: data $X_1, \ldots, X_n$.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Utility of summation notation in this course: NIL. But $\bar{X}$ is standard notation for average of X.

Median: number such that 1/2 of X values at least that large, and 1/2 of X values at least that small.

Sort list: if n is odd median is middle of sorted list. If n is even take average of two middle values.

Numerical examples: ages in my family:

$$50, 50, 20, 15, 8, 8.$$

$$\bar{A} = \frac{50 + 50 + 20 + 15 + 8 + 8}{6} = \frac{151}{6} \approx 25.2$$

Median age: middle numbers are 15, 20.

Halfway between is median = 17.5.

Mode: most common value. Not useful concept in most cases. Location of tallest bar in histogram (affected by definition of classes).

Mode of ages is not unique: 50 or 8. Not useful summary of centre.

Comparison:

Advantages of mean:

1) if your average weekly income is $100 you know how you will do in the long run; not so if median weekly income is $100.

2) Same point: average and sample size tells you total.

3) Has simpler mathematical behaviour than median.

Advantages of median:

Not influenced by extreme members of list.

Median income, for instance, gives more information about typical person.

Measures of spread:

Standard Deviation

Interquartile Range

Mean Absolute Deviation.

Deviations from the mean: subtract mean from each number in list: $X_i - \bar{X}$. For my family deviations are

$$24.8, 24.8, -5.2, -10.2, -17.2, -17.2.$$

Summarize size of deviations:

Average is 0. Not useful as measure of size since pluses cancel minuses.

Mean absolute deviation: take absolute values (ignore $-$ signs) and average

$$\frac{24.8+24.8+5.2+10.2+17.2+17.2}{6}$$
$$= 16.6 \text{ years}$$

Standard deviation: square deviations, average, take square root:

$$s = \sqrt{\frac{(24.8)^2+\cdots+(-17.2)^2}{5}}$$
$$= 19.8 \text{ years.}$$

WARNING: notice the 5 not 6. This is Traditional. Not important in large data sets.

Jargon: **variance** is s^2:

$$s^2 = \frac{(24.8)^2+\cdots+(-17.2)^2}{5}$$
$$= 390.6 \text{ years}^2$$

Interquartile Range:

First define quartiles, quintiles, etc.

First, second and third quartiles split list into 4 equal pieces.

One quarter of list below first quartile, two quarters below second, three quarters below third.

Second quartile is median.

Interquartile range is third quartile minus first quartile.

Book gives method to find quartiles.

Quintiles split list into 5 equal parts.

Percentiles split list into 100 equal parts.

Comparison:

Advantages of IQR: like median not influenced by extremes.

Easily related to proportions of population.

But: rather than use 2 number summary (median, IQR) typically use 3 number summary (quartiles) or 5 number summary (min, max, quartiles).

Boxplot is graph of 5 number summary.

Advantages of Mean Absolute Deviation.

Seems intuitive.

Less influenced by extremes than Standard Deviation.

But: poor mathematical properties.

We mostly use Standard Deviation.

Why the Standard Deviation?

Usual explanation: squares nicer mathematically than absolute values.

Real explanation (WARNING: personal view): ONLY the SD works in normal approximations for sums.

Normal approximations? A common summary for curves.

Rule of thumb: in many lists of data about 2/3 of the observations are within 1 SD of the mean, about 95% within 2 SDs of the mean and almost all within 3 SDs of the mean.

NEXT TOPIC: the normal curve. (bell curve, Gaussian)