Summarising Data

Mark Lunt

Centre for Epidemiology Versus Arthritis
University of Manchester

18/10/2022

Summarising Data

Today we will consider

- Different types of data
- Appropriate ways to summarise these data
  - Graphical Summary
  - Numerical Summary

Types of Data

Qualitative	Nominal	Outcome is one of several categories
	Ordinal	Outcome is one of several ordered categories
Quantitative	Discrete	Can take one of a fixed set of numerical values
	Continuous	Can take any numerical value

Examples of Types of Data

Nominal	Blood group; Hair colour.
Ordinal	Strongly agree, agree, disagree, strongly disagree.
Discrete	Number of children.
Continuous	Birthweight.

Caveats with Data Types

- Distinction between nominal and ordinal variables can be subjective: e.g. vertebral fracture types: Wedge, Concavity, Biconcavity, Crush.
  Could argue that a crush is worse than a biconcavity which is worse than a concavity . . . , but this is not self-evident.
- Distinction between ordinal and discrete variables can be subjective: e.g. cancer staging I, II, III, IV: sounds discrete, but better treated as ordinal.
- Continuous variables generally measured to a fixed level of precision, which makes them discrete. Not a problem, provide there are enough levels.

Types of Variables

What type of variable are each of the following:

- Number of visits to a G.P. this year
- Marital Status
- Size of tumour in cm
- Pain, rated as minimal/moderate/severe/unbearable
- Blood pressure (mm Hg)

Summarizing Qualitative Data

- Count the number of subjects in each group.
- The count is commonly refered to as the *frequency*
- The proportion in each group is referred to as the *relative frequency*
- Stata command to produce a tabulation is `tabulate` *`varname`*

Numerical Summary of Qualitative Data

```
     region |      Freq.     Percent        Cum.
------------+-----------------------------------
     Canada |        422       22.84       22.84
        USA |        541       29.27       52.11
     Mexico |        223       12.07       64.18
     Europe |        493       26.68       90.85
       Asia |        169        9.15      100.00
------------+-----------------------------------
      Total |      1,848      100.00
```

Graphical Summary of Qualitative Data

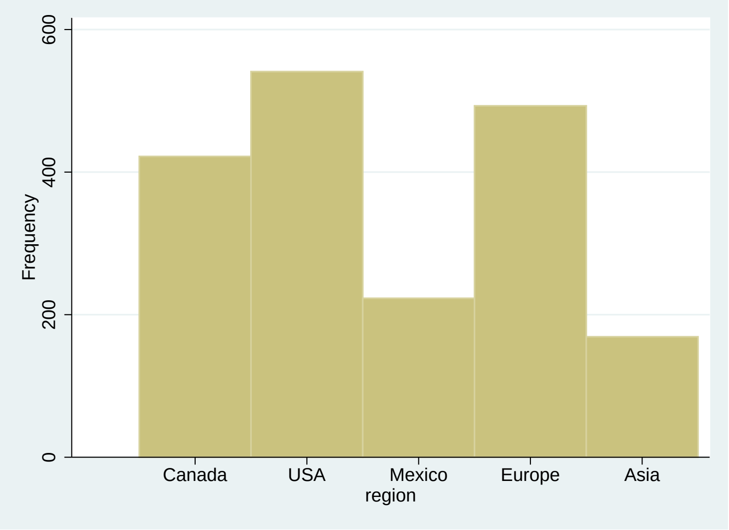
Bar Chart: Data represented as a series of bars, height of bar proportional to frequency.

Pie Chart: Data represented as a circle divided into segments, area of segment proportional to frequency.

Pictograms: Similar to bar chart, but uses a number of pictures to represent each bar.

Bar chart is the easiest to understand.

Bar Chart



Summarizing Quantitative Data

Simplest method: treat as qualitative data.

- Divide observations into groups
  - May be unnecessary for discrete data.
- Look at the frequency distribution of these groups
- Can use table or diagram.

The Histogram

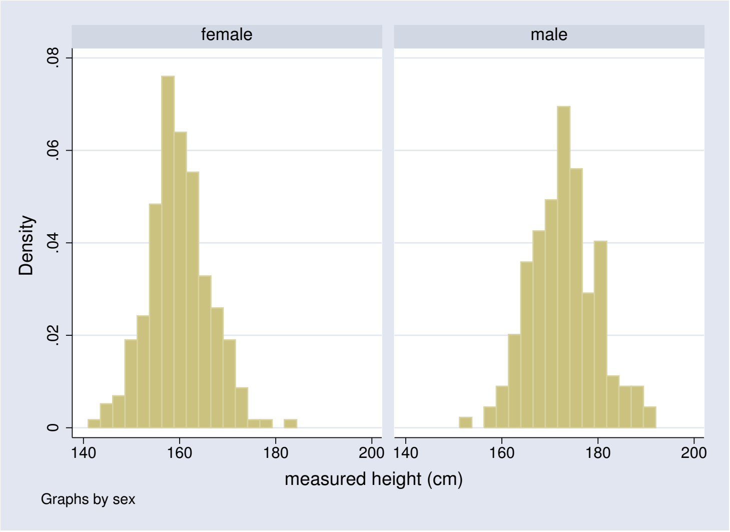
- Similar to a bar chart
- Continuous, not categorical variable
- Area of bars proportional to probability of observation being in that bar
- Axis can be
  - Frequency (heights add up to n)
  - Percentage (heights add up to 100%)
  - Density (*Areas* add up to 1)

How Many Groups ?

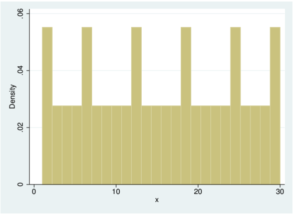
Impossible to say.

- Depends on the number of observations: if individual groups are too small, results are meaningless.
- With discrete variables, exact positions of boundaries may be important.
- Tables need few groups, graphs can have more if sufficient numbers.
- May be decided for you in software.

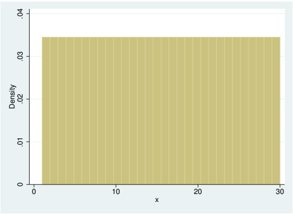
Histograms



Histogram: Effect of Wrong number of bins



24 bins (default)



30 bins (correct)

Bar charts and histograms in Stata

- `histogram` *`varname`* produces a histogram
- Number of bars can by set by option `bin()`
- Width of a bar can be set by option `width()`
- `histogram` *`varname`*`, discrete` produces a bar chart
- What stata calls a bar chart is the mean of second variable subdivided by category, rather than a frequency.

Numerical Summary of Quantitative Data

- Need to know:
  1. What is a typical value ("location")
  2. How much do the values vary ("scale")
- Simplest distribution to summarize is the normal distribution
- Other summary statistics (skewness, kurtosis etc) thought of relative to normal distribution.

Measures of Location

What is the value of a "typical" observation ? May be:

- (Arithmetic) Mean
- Median
- Other forms of mean
  - Rarely used
  - Only if data has been transformed

Arithmetic Mean

"Add them up and divide by how many there are."

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\ &= (\Sigma_{i=1}^{n} x_i)/n \end{aligned}$$

Median

"Arrange in increasing order, pick the middle." If an even number of observations, take mean of middle two.

- Ignores the precise magnitude of most observations
  - Contains less "information" than mean
  - May be useful if there are outliers
- Less easy to use mathematically.

Mean vs. Median

Consider this series of durations of absence from work due to sickness (in days).
1,1,2,2,3,3,4,4,4,4,5,6,6,6,6,7,8,10,10,38,80
Mean = 10
Median = 5
Very few observations are as large as the mean: median is more “typical”.

Percentiles

- The x^{th} percentile is the value than which x% of observations are smaller and $(100 - x)$% are larger.
- The median is the 50th percentile.
- Other centiles can easily be calculated, eg 5th, 25th etc.

Measures of Variation

How close to the “typical” value are other values.

- Range
- Inter-quartile range
- Variance

Simple Measures of Variation

Range
- (Largest measurement) - (smallest measurement)
- Depends on only two measurements
- Can only increase as you add more to the sample

Inter-quartile Range
- (75th centile) - (25th centile).
- Less sensitive to extreme values
- Need fairly large numbers of observations

Standard Deviation

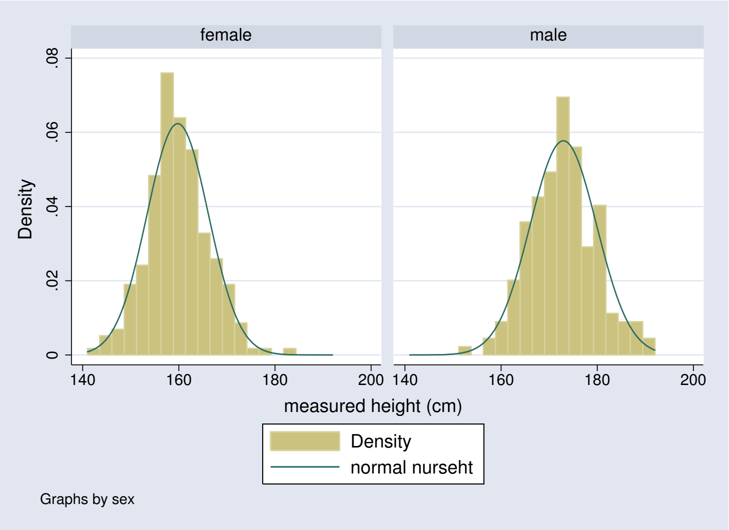
$$\text{Standard Deviation} \quad = \quad \sqrt{\Sigma(x_i - \bar{x})^2/n}$$

- **Nearly** the average difference from the mean
- Uses information from every observation
- Not robust to outliers
- Variance is easy to use mathematically
- Standard deviation is the same units as the observations

The Normal Distribution

- Symmetrical “Bell-shaped” distribution
- Easiest to use mathematically
- Many variables are normally distributed
- Can be described by two numbers
  - Mean (measure of location)
  - Standard Deviation (measure of variation)

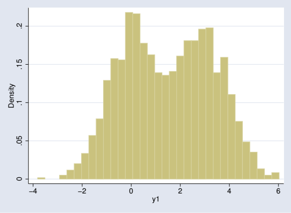
Histogram & Normal Distribution



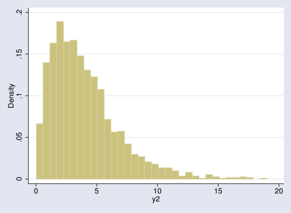
Non-Normal Distributions

- Normal distribution is symmetric.
- Asymmetric distributions are called “skewed”:
  - Positively skewed = some extremely high values (mean > median).
  - Negatively skewed = some extremely low values (mean < median).
- Distribution may have more than one “peak”: bi-modal.
  - Usually formed by mixing two different groups.

Non-Normal Distributions



Bimodal Distribution



Positively Skewed Dist'n

Summary Statistics in Stata

- `summarize` *`varlist`* will give mean, SD, min and max
- `summarize` *`varlist`*`, detail` also gives percentiles
- `tabstat` or `table` can produce tables of summary statistics

Numerical Summary: Table 1

- Quantitative variables
  - Need a measure of location & variation
    - Normal variables: mean and SD
    - Skewed variables: median and IQR
  - Need to give units
- Qualitative variables
  - Number and % in each category

Numerical Summary Example

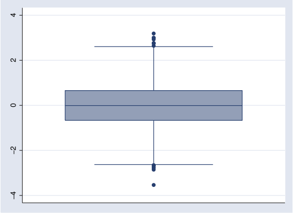
Age in years: Mean (SD)		63 (7.9)
Spine BMD in g/cm^2: Median (IQR)		1.05 (0.78, 1.30)
Gender: n (%)	Male	1537 (44)
	Female	1924 (56)

The Box and Whisker Plot

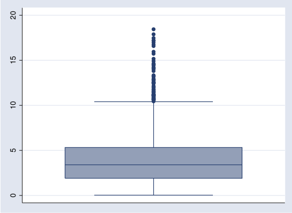
Very efficient summary of distribution:

- Shows median, upper and lower quartiles (25th and 75th percentiles).
- Also shows range of “normal” values and individual “unusual” values.
- Definitions of “normal” and “unusual” differ.
- Will demonstrate skewness, not bimodality.
- Stata command: `graph box` *`varname`*`, [by(`*`groupname`*`)]`

Box and Whisker Plots



Normal Distribution



Positively Skewed Dist’n

Transforming Data

- Skewed distributions may be made symmetric by a transformation.
- Taking logs is the most common.
- Other transformations (e.g. square root, reciprocal) can be used, but can be very difficult to interpret.
- May be better to transform back to original units to present results.
  - Geometric mean is back-transformation of mean of log-transformed data.

Further Reading

- **Edward R. Tufte**, *The Visual Display of Quantitative Information* was the classic text on statistical graphs.
- Huge data visualisation industry now