

# Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapters 2 & 3: Data presentations and summaries

Jinko Graham

2019-09-07

## Data Presentation (Chapter 2)

# Types of Variables

- ▶ Categorical versus quantitative variables
- ▶ Categorical variables: variable has categories or classes
  - ▶ Nominal data: unordered categories or classes (e.g., gender, cancer type).
  - ▶ Ordinal data: categories with a natural order (e.g., cancer stage, injury severity score)
- ▶ Quantitative variables: numbers represent measurable quantities
  - ▶ Discrete variables: Restricted values, e.g. integers. Some examples included number of days in hospital, parity.
  - ▶ Continuous data: No restriction on values (though in practice, the measuring device may impose restrictions). Some examples include blood pressure, height, weight.

# Tables

- ▶ Tables can be used to display the frequency distribution of a categorical variable
- ▶ Example: Frequency distribution of gender among 21,737 bladder cancer patients. Data from Mungan et al. (2000)

```
## Gender
## Female    Male
##    5536  16201
```

## Tables, cont.

- ▶ Joint frequency distribution of two categorical variables:

##		Cancer.Stage			
##	Gender	I	II	III	IV
##	Female	3926	402	356	852
##	Male	12418	995	883	1905

- ▶ More on analysis of two-way tables in Chapter 15.

# Tabulating Quantitative Variables

- ▶ Can tabulate quantitative variables after “binning”
- ▶ Divide the range of possible values into bins
  - ▶ For example, with age data ranging from 15 to 30 years, could create three 5-year bins for age
- ▶ Count the number of subjects in each bin
- ▶ Histograms are a graphical display of quantitative data that makes use of binning.
  - ▶ We will see an example histogram in two slides.

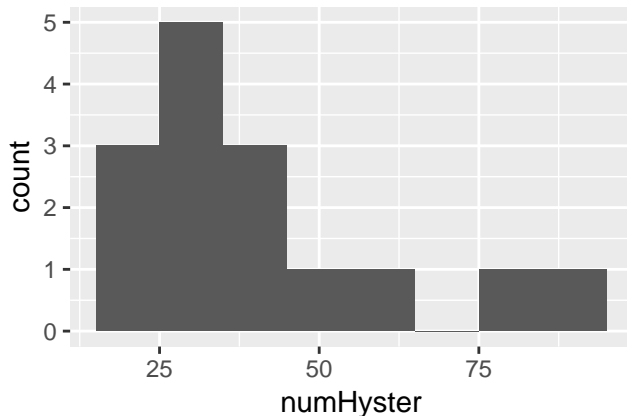
# Graphs

- ▶ Bar plots to display categorical variables or discrete variables with a limited number of possible values
- ▶ Histograms to display discrete variables with many possible values and for continuous variables
- ▶ Bars above categories (or bins) indicate number of observations in that category (or bin)

## Example Histogram

- Data on the numbers of hysterectomies performed by 15 male Swiss doctors:

20 25 25 27 28 31 33 34 36 37 44 50 59 85 86



- Observations far from the others are called outliers



# Notes on Histograms

- ▶ The purpose is a graphical representation of a distribution.
  - ▶ The details of how the picture are drawn are not that important.
  - ▶ But you may find yourself wondering . . .
- ▶ The bins in the previous example were of width 10.
  - ▶ It looks like they were set at 15 to 25, 25 to 35, etc.
- ▶ Which bin do the 25's go in?
  - ▶ The 15-25 bin. The bins in R's `histogram()` function don't include their left end-point, but do include their right end-point

## Summary Statistics (Chapter 3)

# Summary Statistics Overview

- ▶ For quantitative variables, numerical summaries are used to measure different aspects of a data distribution.
  - ▶ Mean and median measure centre (central tendency) of the distribution
  - ▶ Inter-quartile range and standard deviation measure spread (dispersion)
- ▶ *Five-number summary* to summarize a distribution: min, max, median, 1st and 3rd quartiles. Graphed with a boxplot (more later).

## Centre: The mean

- ▶ The population mean,  $\mu$ , is the ordinary arithmetic average of a variable in the population.
- ▶ The corresponding statistic is the sample mean,  $\bar{x}$ .
- ▶ The sample mean,  $\bar{x}$ , is the ordinary arithmetic average of the observations in a random sample from the population.
  - ▶ Notation: Let  $x_1, x_2, \dots, x_n$  denote the observed values of a variable measured on  $n$  individuals. The sample mean,  $\bar{x}$  is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▶ Hysterectomy example data:

20 25 25 27 28 31 33 34 36 37 44 50 59 85 86

has sample mean

$$\bar{x} = \frac{20 + 25 + \dots + 86}{n} = 41.3$$

## Centre: The Median

- ▶ The population median is the “middle value” of the variable in the population.
- ▶ The corresponding statistic is the sample median,  $M$ .
- ▶ The sample median is the “middle value” of the variable in a random sample from the population.
- ▶ The sample median of the hysterectomy data is:

20, 25, 25, 27, 28, 31, 33, **34**, 36, 37, 44, 50, 59, 85, 86

- ▶ The centre observation is  $M = 34$ .
- ▶ The median represents a “typical” observation.

## Spread: The Standard Deviation (SD) and Variance

- ▶ The variance,  $\sigma^2$ , is the average of squared deviations from the mean in the population.
- ▶ The SD,  $\sigma$ , measures spread about the mean.
- ▶ The corresponding statistics are:
  - ▶ the sample variance,  $s^2$ , an average (almost) of the squared deviations about the sample mean in a random sample from the population
  - ▶ the sample SD,  $s$ , the square root of the sample variance.
- ▶ Notation:  $x_1, x_2, \dots, x_n$  are observed measurements on  $n$  individuals with sample mean  $\bar{x}$ .

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

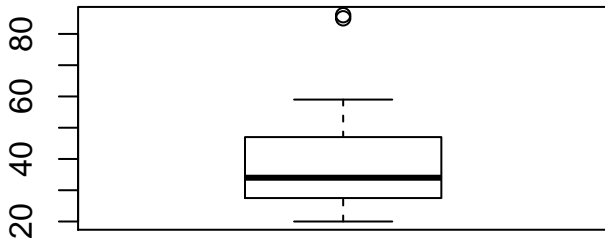
- ▶ Hysterectomy example:  $s = 20.6$

## Spread: The Inter-Quartile Range (IQR)

- ▶ The first and third quartiles mark the first and third quarters of the observations, whether in a population or in a random sample from the population.
  - ▶ These are also called the 25th and 75th percentiles, respectively.
- ▶ For the hysterectomy data, R calculates these quartiles to be 27.5 and 47, respectively.
  - ▶ Details of how R calculates quartiles are unimportant to us.
- ▶ The middle half of the data lies between.
- ▶ The range of the middle half, or IQR, is  $47 - 27.5 = 19.5$ .

# Boxplots

- ▶ The five-number summary is the minimum, maximum, median, 1st and 3rd quartiles; graphed with a boxplot:



- ▶ Box represents the IQR, thick horizontal line the median.
- ▶ Whiskers extend out to the min/max if these are within a certain distance from the box,
- ▶ However, **outliers** that are far from the box are plotted separately (as above)
- ▶ There are different ways to draw boxplots; details unimportant



## Further Examples

- ▶ Summary statistics and graphics can be very powerful.
- ▶ See, for example, the late Hans Rosling's work

[http://www.ted.com/talks/hans\\_rosling\\_shows\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen.html](http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html)