Applied Epidemiology I: Summary statistics

Enoch Yi-Tung Chen

Department of Medical Epidemiology and Biostatistics, Karolinska Insitutet

February 1, 2021

Acknowledgements

This course material is based on my learning from Anastasia Lam's teachings in last year's Applied Epidemiology I lab sessions, and readings from *Epidemiology* by Gordis [1], *A First Course in Probability and Statistics* by Goldsman and Goldsman [2], *Principles of Biostatistics* by Pagano and Gauvreau [3], and *Biostatistics I* by Gabriel and Frumento [4]. I especially want to thank Marlene Stratmann for reviewing the slides and Prof. Paul Dickman for providing me with suggestions to improving the teaching.

Outline

Summary statistics:
Measures of Central Tendency: mean, median, mode

- Mean: the sum of the values of a variable and dividing by number of the observations

Summary statistics:
Measures of Central Tendency: mean, median, mode

- Mean: the sum of the values of a variable and dividing by number of the observations
- Median: the middle (the 50th centile) observation

Summary statistics:
Measures of Central Tendency: mean, median, mode

- Mean: the sum of the values of a variable and dividing by number of the observations
- Median: the middle (the 50th centile) observation
- Mode: the value that occurs most frequently

Summary statistics:
Measures of Central Tendency: mean, median, mode

- Mean: the sum of the values of a variable and dividing by number of the observations
- Median: the middle (the 50th centile) observation
- Mode: the value that occurs most frequently

```
. tabstat age // will only give you mean

    variable |      mean
-------------+----------
         age |  56.41176
----------------------

. tabstat age, s(count mean median) // s stands for statistics
    variable |         N      mean       p50
-------------+------------------------------
         age |        34  56.41176        56
--------------------------------------------
```

Summary statistics: Measures of Dispersion: range, IQR, variance, standard deviation

- Range: the difference between the maximum and the minimum

Summary statistics: Measures of Dispersion: range, IQR, variance, standard deviation

- Range: the difference between the maximum and the minimum
- Interquartile range: the absolute difference between the 25th percentile of the observations and the 75th.

Summary statistics: Measures of Dispersion: range, IQR, variance, standard deviation

- Range: the difference between the maximum and the minimum
- Interquartile range: the absolute difference between the 25th percentile of the observations and the 75th.
- Variance, standard deviation (sd): a measure of spread of the data

Summary statistics: Measures of Dispersion: range, IQR, variance, standard deviation

- Range: the difference between the maximum and the minimum
- Interquartile range: the absolute difference between the 25th percentile of the observations and the 75th.
- Variance, standard deviation (sd): a measure of spread of the data

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

```
. tabstat age, s(count range min max iqr var sd)

    variable |         N     range       min       max
-------------+--------------------------------------------
         age |        34        20        47        67
----------------------------------------------------------

    variable |       iqr  variance        sd
-------------+------------------------------
         age |        10  36.12834  6.010686
-----------------------------------------
```

References

1. Gordis L. *Epidemiology*. Philadelphia, PA: Elsevier/Saunders, 2014. ISBN 9781455737338.
2. David Goldsman PG. *A First Course in Probability and Statistics*. Georgia Institute of Technology, 2020.
3. Marcello Pagano KG. *Principles of Biostatistics*. Cengage Learning, Inc, 2000. ISBN 0534229026.
4. Erin Gabriel PF. Epidemiology PhD program, Karolinska Institutet, 2020.