Basic Biostatistics

B. Burt Gerstman

Statistics for Public Health Practice

Chapter 4: Summary Statistics

Chapter 4 1

In Chapter 4:

The prior chapter used graphs to look at distributional shape, location, and spread.

This chapter uses **numerical summaries** of the same distributional features.

Chapter 4 2

Summary Statistics

- **Central location statistics**
  - Mean
  - Median
  - Mode
- **Spread statistics**
  - Range
  - Interquartile range (IQR)
  - Variance and standard deviation
- **Shape statistics** exist but are seldom used in practice and are *not* covered

Chapter 4 3

Notation

- $n \equiv$ sample size
- $X \equiv$ the variable (e.g., ages of subjects)
- $x_i \equiv$ the value of individual i for variable X
- $\Sigma \equiv$ sum all values (capital sigma)
- Example (n = 10):

21 42 5 11 30 50 28 27 24 52

Let X = AGE

$x_1 = 21$, $x_2 = 42$, …, $x_{10} = 52$

$\Sigma x_i = x_1 + x_2 + \ldots + x_{10} = 21 + 42 + \ldots + 52 = 290$

Chapter 4 4

§4.1: Central Location: Sample Mean

- Arithmetic average
- Traditional measure of central location
- Sum the values and divide by n
- Notation: "x-bar"

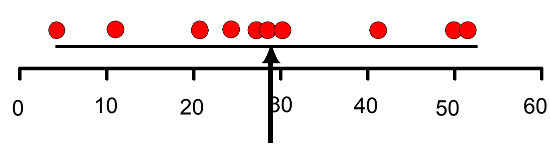
$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Chapter 4 5

Example: Sample Mean

From data: n = 10 with Σx_i = 290

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}(290) = 29.0$$

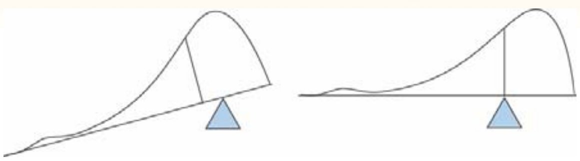


Mean = 29

Gravitational center

6

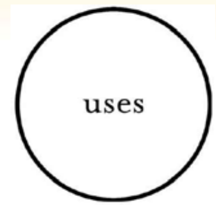
Mean = Gravitational Center



A skew tips the distribution causing the mean to shift toward the tail

Uses of the Sample Mean

- Predicts value of an observation drawn at random from the sample
- Predicts value of an observation drawn at random from the population
- Predicts the population mean



Population Mean

$$\mu = \frac{\sum x_i}{N} = \frac{1}{N}\sum x_i$$

- Same operation as sample mean except based on entire population (N ≡ population size)
- Important conceptually, but not readily available and seldom used

§4.2 Central Location: Median

The median is the value with a *depth* of (n+1)/2

When *n* is even, average the two middle values

For the 10 values below, the median has depth (10+1) / 2 = 5.5, placing it between 27 and 28. Average these two values: median = 27.5

```
05   11   21   24   27   28   30   42   50   52
                         ↑
                       median
        Average the adjacent values: M = 27.5
```


More Examples

- Example A: 2 4 6
  Median = 4
- Example B: 2 4 6 8
  Median = 5 (average of 4 and 6)
- Example C: 6 2 4
  Median ≠ 2
  (Values must be ***ordered*** first)

The Median is Robust

The median is relatively resistant to skews and outlier; it is *robust*.

This data set has x-bar = 1636:

1362 1439 1460 **1614** 1666 1792 1867

Same data set with a data entry error “outlier” ***highlighted***. This data has x-bar = 2743:

1362 1439 1460 **1614** 1666 1792 *9867*

The median is 1614 in both instances,

§4.3: Mode

- Most frequent value in the dataset
- This data set has mode = 7
  {4, 7, 7, 7, 8, 8, 9}
- This data set has no mode
  {4, 6, 7, 8}
  (each point appears once)
- The mode is useful only in large data sets with repeating values

§4.4: Comparison of Mean, Median, Mode

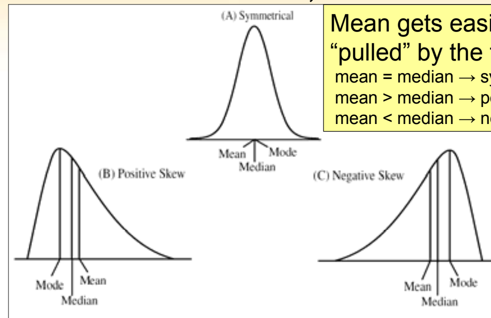


FIGURE 4.4 Effect of a skew on the mean, median, and mode.

Mean gets easily "pulled" by the tail:
mean = median → symmetrical
mean > median → positive skew
mean < median → negative skew

Spread

- Spread = extent to which data vary
- Data at both sites in data to right have x-bar ≈ 36
- Site 1 exhibits greater spread (visual inspection)
- How can we quantify spread?

```
Site 1| |Site 2
---------------
    42|2|
     8|2|
     2|3|234
    86|3|6689
     2|4|0
      |4|
      |5|
      |5|
      |6|
     8|6|
       ×10
```

particulates in air (μg/m³)

Spread: Range

- Range = maximum – minimum
- Illustrative example:
  Site 1 range = 68 – 22 = 46
  Site 2 range = 40 – 32 = 8
- The sample range is ***not*** a good measure of spread: tends to underestimate population range
- Always supplement the range with at least one addition measure of spread

```
Site 1| |Site 2
---------------
    42|2|
     8|2|
     2|3|234
    86|3|6689
     2|4|0
      |4|
      |5|
      |5|
      |6|
     8|6|
       ×10
```


Spread: Interquartile Range

- **Quartile 1 (Q1):** marks bottom quarter of data = middle of the lower half of the data set
- **Quartile 3 (Q3):** marks top quarter of data = middle of the top half of data set
- **Interquartile Range (IQR)** = Q3 – Q1 covers middle 50% of \distribution

```
05  11  21  24  27   28  30  42  50  52
        ↑            ↑           ↑
        Q1         median        Q3
```

Q1 = 21, Q3 = 42, and IQR = 42 – 21 = 21

Quartiles: Tukey's Hinge Method

Data are metabolic rates (cal/day), n = 7

1362 1439 1460 1614 1666 1792 1867
↑
median

When n is odd, include the median in both subsets

Q1 = 1449.5

Bottom half: 1362 1439 1460 1614

Top half: 1614 1666 1792 1867

Q3 = 1729

Five-Point Summary

- Q0 (the minimum)
- Q1 (25th percentile)
- Q2 (median)
- Q3 (75th percentile)
- Q4 (the maximum)

```
05  11  21  24  27   28  30  42  50  52
        ↑            ↑           ↑
        Q1         median        Q3
```

5 point summary = 5, 21, 27.5, 42, 52

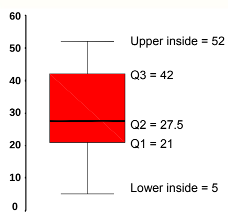
§4.6 Boxplots

1. Draw **box** from Q1 (upper hinge) to Q3 (lower hinge); draw line for **median.**
2. Calculate **fences** as follows:
   Fence_{Lower} = Q1 – 1.5(IQR)
   Fence_{Upper} = Q3 + 1.5(IQR)
   Do **not** draw fences
3. Determine if any values lie outside the fences (**outside values**). If so, plot separately
4. Determine values inside the fences (**inside values**)
   Draw whisker from Q3 to upper inside value.
   Draw whisker from Q1 to lower inside value

Example 1: Boxplot

Data: 05 11 21 24 27 28 30 42 50 52

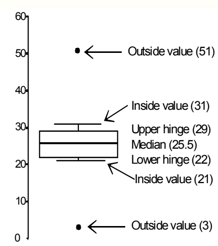
1. 5 pt summary: {5, 21, 27.5, 42, 52};
   Box from 21 to 42 with line @ 27.5
2. IQR = 42 – 21 = 21.
   F_U = Q3 + 1.5(IQR) = 42 + (1.5)(21) = 73.5
   F_L = Q1 – 1.5(IQR) = 21 – (1.5)(21) = –10.5
3. No values above upper fence
   None values below lower fence
4. Upper inside value = 52
   Lower inside value = 5
   Draws whiskers



Example 2: Boxplot

Data: 3 21 22 24 25 26 28 29 31 51

1. 5-point summary: 3, 22, 25.5, 29, 51; *hinges* at 22 and 29
2. IQR = 29 – 22 = 7
   F_U = Q3 + 1.5(IQR) = 29 + (1.5)(7) = 39.5
   F_L = Q1 – 1.5(IQR) = 22 – (1.5)(7) = 11.6
3. One upper outside value (51)
   One lower outside value (3)
4. Upper inside value is 31
   Lower inside value is 21
   Draw whiskers



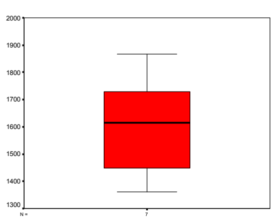
Example 3: Boxplot

Seven metabolic rates (cal / day):
1362 1439 1460 1614 1666 1792 1867

1. 5-point summary: 1362, **1449.5, 1614, 1729**, 1867
2. IQR = 1729 – 1449.5 = 279.5
   F_U = Q3 + 1.5(IQR)
   = 1729 + (1.5)(279.5) = 2148.25

   F_L = Q1 – 1.5(IQR)
   = 1449.5 – (1.5)(279.5) = 1030.25

3. None outside
4. Inside values: **1867** and **1362**

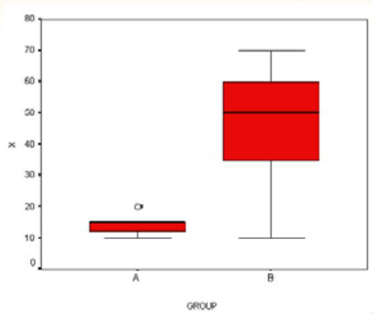


Boxplots: Interpretation

- Location
  - Position of median
  - Position of box
- Spread
  - Hinge-spread (IQR)
  - Whisker-to-whisker spread
- Shape: symmetry? skew? length of tails? outside values?

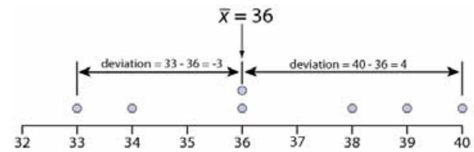
Side-by-side boxplots

Especially useful when comparing groups



Spread: Deviation

- Deviation = distance of a data point from the mean = (x_i – x-bar)
- This data set has x-bar = 36
- Data point 33 has deviation = 33 – 36 = −3
- Data point 40 has deviation = 40 – 36 = 4



Variance and Standard Deviation

Deviation = $x_i - \bar{x}$

Sum of squared deviations = $SS = \sum (x_i - \bar{x})^2$

Sample variance = $s^2 = \dfrac{SS}{n-1}$

Sample standard deviation = $s = \sqrt{s^2}$

Standard deviation (formula)

Sum of Squares

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

Standard Deviation

Observation x_i	Deviations $x_i - \bar{x}$	Sq. deviations $(x_i - \bar{x})^2$
36	36 – 36 = 0	0^2 = 0
38	38 – 36 = 2	2^2 = 4
39	39 – 36 = 3	3^2 = 9
40	40 – 36 = 4	4^2 = 16
36	36 – 36 = 0	0^2 = 0
34	34 – 36 = –2	$–2^2$ = 4
33	33 – 36 = –3	$–3^2$ = 9
32	32 – 36 = –4	$–4^2$ = 16
SUMS ⇒	0 [Always]	SS = 58

Standard Deviation (cont.)

Sample variance (s^2)

Note: squared units

$$s^2 = \frac{SS}{n-1} = \frac{58}{8-1} = 8.286\,(\mu\text{g/m}^3)^2$$

Standard deviation (s)

$$s = \sqrt{s^2} = \sqrt{8.286} = 2.88\ \mu\text{g/m}^3$$

Interpretation of Standard Deviation

- Most common measure spread
- Sample standard deviation *s* is estimator of population standard deviation σ
- 68-95-99.7 rule
- Chebychev's rule

68-95-99.7 Rule
Normal Distributions *Only*

- 68% of data within $\mu \pm \sigma$
- 95% of data within $\mu \pm 2\sigma$
- 99.7% of data within $\mu \pm 3\sigma$
- Example: Normal distribution with $\mu = 30$ and $\sigma = 10$ has:

  68% of values in $30 \pm 10 = 20$ to 40

  95% in $30 \pm (2)(10) = 30 \pm 20 = 10$ to 50

  99.7% in $30 \pm (3)(10) = 30 \pm 30 = 0$ to 60

Chebychev's Rule (All Distributions)

- *At least* 75% of the values within $\mu \pm 2\sigma$
- Example:
  A distribution with $\mu = 30$ and $\sigma = 10$ has *at least* 75% of the values within
  $30 \pm (2)(10)$
  $= 30 \pm 20$
  $= 10$ to 50

Rounding

- There is no single rule for rounding. There are some guidelines.
- The number of significant digits should reflect the precision of the measurement
- You are safe when you carry at least four significant digits during calculations
- Use a reliable reporting guidelines (e.g., APA Publication Manual)
- The round as final step.
- ALWAYS use judgment and "Be kind to your reader"

Choosing Summary Statistics

- Always report a measure of central location, a measure of spread, and the sample size
- Symmetrical distributions ⇒ report the mean and standard deviation
- Asymmetrical distributions ⇒ report the 5-point summaries (or median and IQR)

Software and Calculators

Use 'em