# Set 2: Descriptive Statistics

Stat 260: May 10, 2024

| Categorical Data or Qualatative Data | Quantitative Data or Numeric Data |
|---|---|
| • **Definition:** observations that can be put in 1 of atleast 2 categories | • **Definition:** Observations that specify amounts, magnitudes or counts.<br>∴ basically numbers<br>∴ discrete or continous |
| • **Examples:**<br>→ eye colour<br>→ occupation<br>→ program of major | • **Examples:**<br>→ Height and weight<br>→ prices<br>→ Distance<br>→ number of subscribed services |

**Univariate data** is data that consists of a single measurement/observation on subjects. We typically denote univariate data on $n$ subjects as $x_1, x_2, \ldots, x_n$ where $x_1$ is the first observation, $x_2$ is the second observation, etc. Univariate data can be either numerical or categorical. sample size = n and it can be qualitative or quantitative

**Example 1:** Suppose that a local clinic wants to study patient appointment times with their doctors. They record the appointment lengths in minutes for all patients visiting one of their doctors on a certain day.

] = includes
) = doesn't include

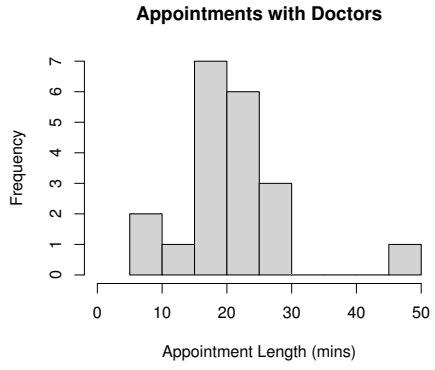
```
Appointment times for patients visiting doctors:
21 9 22 5 21 17 24 26 20 16 25 15 18 15 19 14 25 23 18 48
```

Sample size
n = 20

We can display this appointment time data in both a **frequency table** and a **histogram**.

| Interval | Frequency | Relative Frequency |
|---|---|---|
| $[0, 5)$ | 0 | 0/20 = 0 |
| $[5, 10)$ | 2 | 2/20 = 0.1 |
| $[10, 15)$ | 1 | 1/20 = 0.05 |
| $[15, 20)$ | 7 | 7/20 = 0.35 |
| $[20, 25)$ | 6 | 6/20 = 0.3 |
| $[25, 30)$ | 3 | 3/20 = 0.15 |
| $[30, 35)$ | 0 | 0/20 = 0 |
| $[35, 40)$ | 0 | 0/20 = 0 |
| $[40, 45)$ | 0 | 0/20 = 0 |
| $[45, 50)$ | 1 | 1/20 = 0.05 |

→

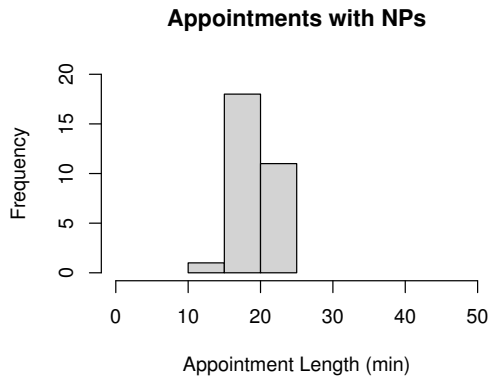


***Example 1 Continued...*** Suppose that the clinic also employs several Nurse Practitioners (NPs). The appointment lengths for patients visiting NPs on the same day is likewise recorded below.

`Appointment times for patients visiting NPs:`

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 20 | 19 | 15 | 19 | 20 | 20 | 19 | 19 | 24 | 21 | 18 | 17 | 19 | 22 | 19 | 23 | 19 | 17 | 15 | 13 | 19 | 17 | 18 | 22 | 19 | 17 | 23 | 22 | 21 |

Sample size
$n = 30$



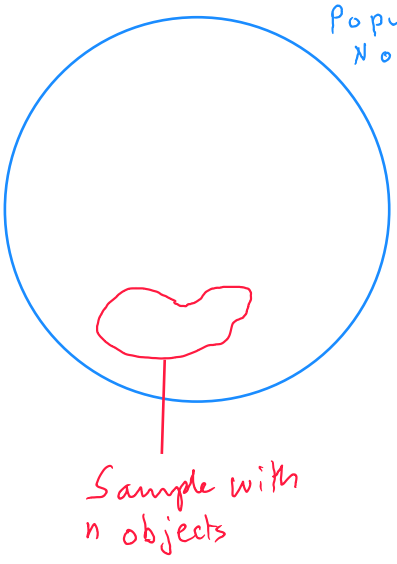

*How can we describe the difference in data spread between the two histograms?*

Two key observation for comparison:
1) how spread out the data is
2) where is the data centered at (most occuring)

---

## Measures of Location of Central Tendency

*How can we find the centre of a dataset?*




Population
- Parameter
  - → mean (average): $\mu$
  - → standard deviation: $\sigma$
  - unattainable typically
    - → median: $\tilde{\mu}$, med, M (any can be used)
    - → variance: $\sigma^2$

Sample
- Statistics
  - → mean (average): $\bar{x}$
  - → standard deviation: $s$
  - → median: $\tilde{x}$
  - → mode
  - → variance: $s^2$

**Mean**: The average of the data

- **Sample Mean:**

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_n}{n}$$

- **Population Mean:**

$$\mu = \frac{x_1 + x_2 + x_3 + x_n}{N}$$

→ typically unattainable, often use $\bar{x}$ to guess $\mu$

**Median**: the value that splits an ordered data sets into an upper and lower half.

- **Sample Median:**

Sort the $n$ observations in increasing order then:

$\tilde{x} =$ the middle value if $n$ is odd, else

the average of two middle values (if $n$ is even) $\frac{m_1 + m_2}{2}$ where $m_1$ and $m_2$ are the two middle values

- **Population Median:**

theoretically the same method as sample median but likely unattainable

**Mode**: the most occurring value

(there can be one more than 1 mode, if that is the case then write both)

(if there are no values occurring more than once e.g. (1,2,3,4,5,6) you write either no mode or all values are the mode)

**Example 2:** Determine the mean, median, and mode of the following samples:

(a) **Sample A**: 0, 0, 2, 3, 6, 7, 10, 11, 20

Mean: $\frac{0+0+2+3+6+7+10+11+20}{9} = 6.56$

Median: 6

Mode: 0

(b) **Sample B**: 1, 1, 2, 8, 15, 15, 25, 100

Mean: $\frac{1+1+2+8+15+15+25+100}{8} = 20.875$

Median: 11.5 → $\frac{15+8}{2}$

Median: 1 and 15

(c) **Sample C**: 5, 0, 3, 28, 3, 7, 10

Mean: $\frac{5+0+3+28+3+7+10}{7} = 8$

Mode: 3

Median: 5

0 3 3 5 7 10 28

Remember to reorder

---

**Question:** Which is better? Mean or median?

Consider **Sample** $B_2$: 1, 1, 2, 8, 15, 15, 25 (Sample B with the **outlier** observations of 100 removed):

B2 → $\overline{x} = 9.57$

→ $\tilde{x} = 8$

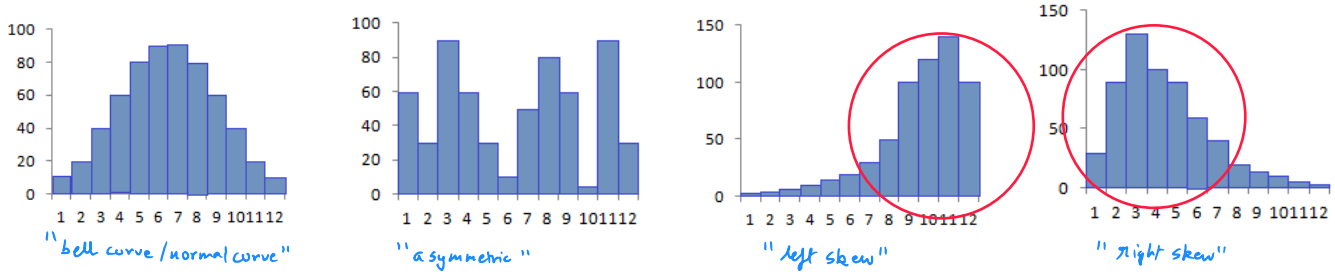
Mean($\overline{x}$) is sensitive to outliers, whereas $\tilde{x}$ is less sensitive. aka median ($\tilde{x}$) is better in this case.

---

## Measures of Variability and Dispersion

*Mean and median tell us where the data is centred, but sometimes we also need to know how spread out the data is.*

The following histograms each represent a sample with $\bar{x} = 6$.
How can we represent the different spreads in the data?



**Variance**: a measurement of how spread out the data is.
→ the larger the variance the greater the spread
→ a variance of 0 implies all observations are the same

- **Population Variance:** "sum of squared differences"

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

- **Sample Variance:**

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

**Standard Deviation**:

the square root of the variance

- **Population Standard Deviation:**

$$\sigma = \sqrt{\sigma^2}$$ ↳ population variance

- **Sample Standard Deviation:**

$$s = \sqrt{s^2}$$ ↳ sample variance

**Example 3:** Determine the variance and the standard deviation of the following samples:

(a) **Sample D**: 10, 48, 49, 50, 51, 52, 90 $\quad n = 7$ , $\bar{x} = 50$

$s^2 = \left(\frac{1}{7-1}\right)\left((10-50)^2 + (48-50)^2 + (49-50^2) + (50-50)^2\ (51-50)^2\ (52-50)^2 + (90-50)^2\right) = \frac{1}{6}(3210) = 535$

$s = \sqrt{535} = 23\sqrt{3}$

(b) **Sample E**: 0, 0, 50, 100, 100

use m+ and mode 1 to add data then RCL to find relevant values written in gray on the calculator

$s = 50$

$s^2 = 2500$

(c) **Sample F**: 48, 49, 50, 51, 52

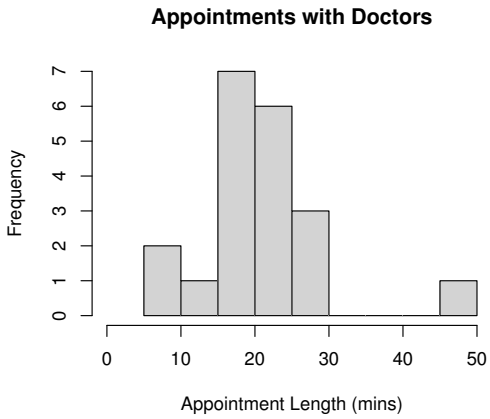
$\bar{x} = 50$

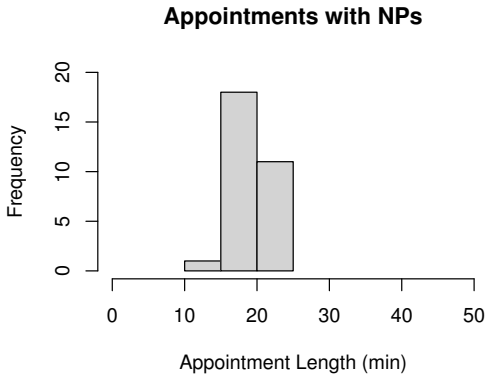
$s = 1.58$

**Sample Variance Short Cut**:

$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$

***Back to Example 1...***




Sample Size: $n = 20$

Mean: $\overline{x} = 20.05$

Median: $\widetilde{x} = 19.5$

Variance: $s^2 = 71.94$

Standard Deviation: $s = 8.48$

Sample Size: $n = 30$

Mean: $\overline{x} = 19.17$

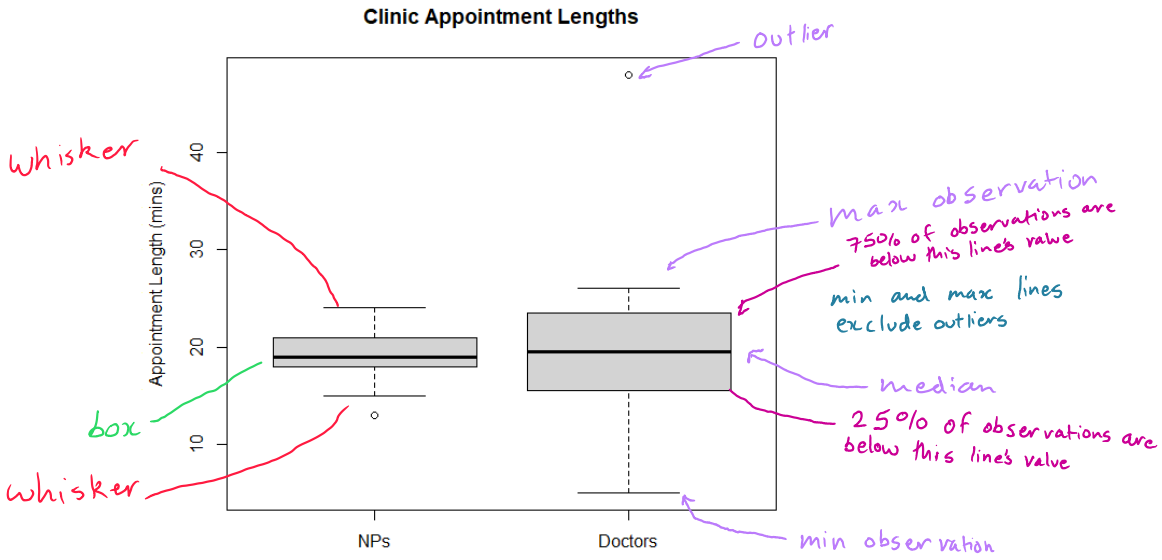
Median: $\widetilde{x} = 19$

Variance: $s^2 = 6.21$

Standard Deviation: $s = 2.49$

* variance is nice in theory, but often doesn't make sense to use in the real world.

*A useful way to represent both samples is using a* ***boxplot*** *(aka* ***box-and-whisker plot****).*



## More on Categorical Data ...

| Ordinal Data | Nominal Data |
|---|---|
| • Definition: categorical data where the categories have some ordering | • Definition: Categorical data where there is no ordering |
| • Examples:<br>– military ranks<br>– letter grades (A, B, C, D, E, F) | • Examples:<br>– eye colour<br>– program / major |

**Example 4:** Ten students are asked to rate their satisfaction with their morning transit commute length on the following scale:

1 = Excellent, 2 = Good, 3 = Adequate, 4 = Poor. ← ordinal data

Their responses are: 1, 1, 2, 2, 2, 2, 3, 3, 3, 4.

$\bar{x} = 2.3$ (this has some meaning but it can be deceptive)
for example, should the difference between "excellent" and "good" be the same as the difference between "adequate" and "poor"

**Example 5:** Suppose we survey 10 people for their eye colour, and encode the results as follows:

$$1 = \text{blue},\ 2 = \text{brown},\ 3 = \text{green}.$$

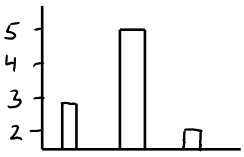
The survey yields the dataset: 1, 1, 1, 2, 2, 2, 2, 2, 3, 3.

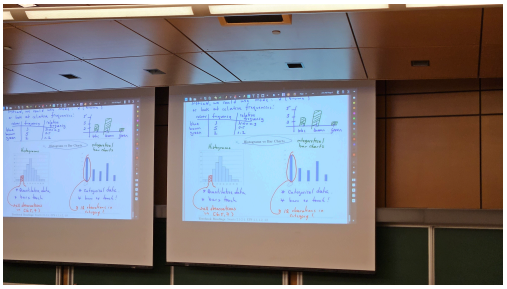
If we take an average ($\bar{x}$) = 1.7 → meaningless in this case

we can't take an average or a median (or standard deviation or variance)

Instead we could use mode = 2 (brown)
or look at relative frequencies =

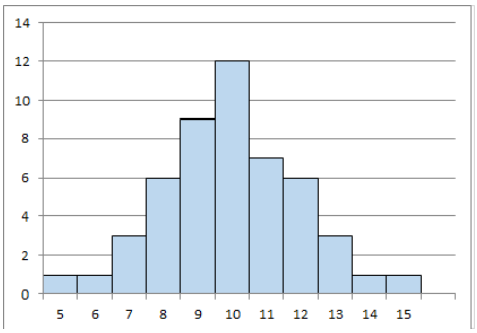
| colour | frequency | relative frequency |
|---|---|---|
| blue | 3 | 0.3 |
| brown | 5 | 0.5 |
| green | 2 | 0.2 |

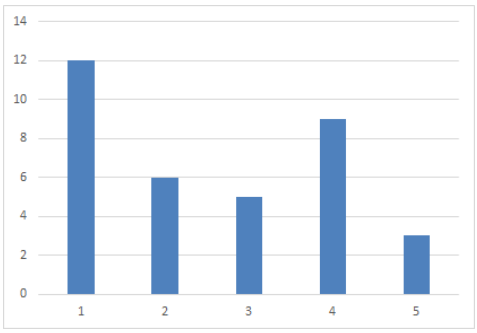


---

## Histograms vs Bar Charts




---

**Textbook Readings**: Swartz 2.1-2.4. EPS 4.1, 4.2, 4.8
**Practice problems**: Swartz 4.3, 4.5, 4.7 (find the variance using the computational formula; check your answer using the statistical functions of your calculator), 4.9