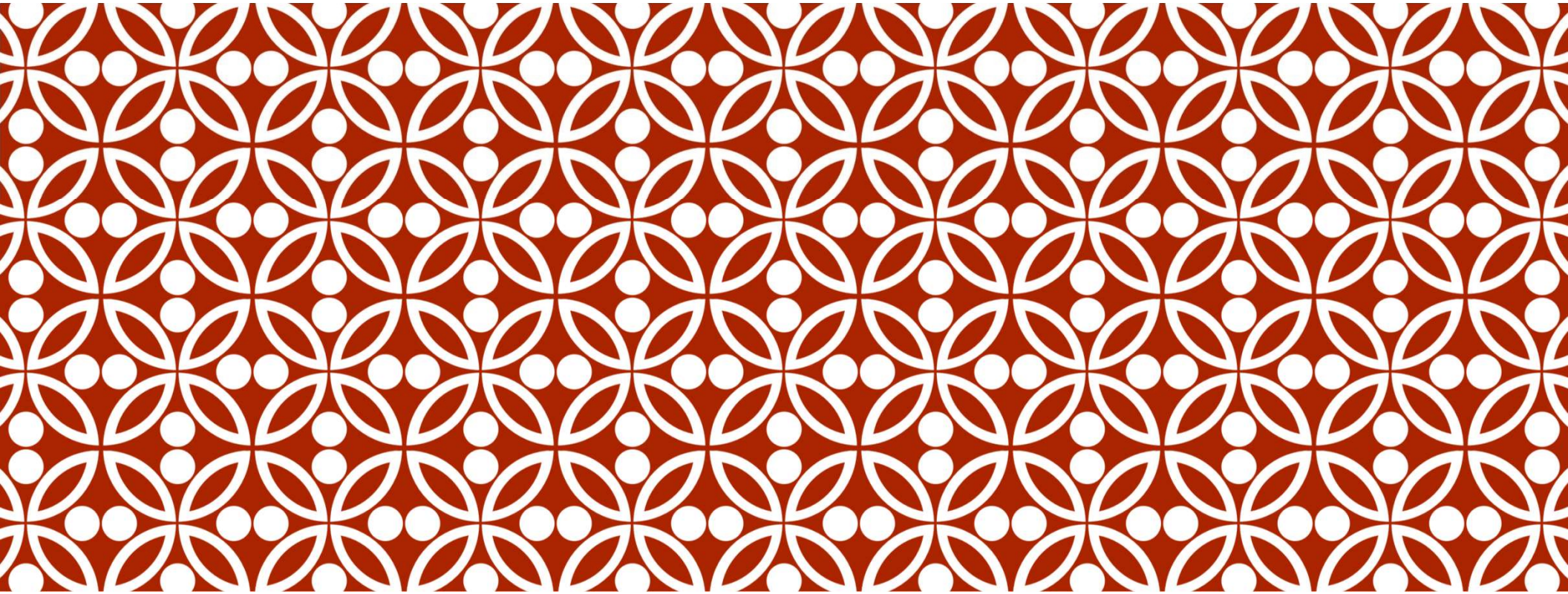


MEASURES OF SPREAD OR DISPERSION

Dr Ebuwa Igho-Osagie

WHY ARE MEASURES OF DISPERSION IMPORTANT?

1. They help you understand the distribution and the behavior of the data

e.g., the mean of -10,-5, 5, 10, 50 is 5; and the mean of 2, 3, 7, 8, 5, is also 5!

2. They help you understand how spread-out data belonging to the same population is e.g., the median of -1000, -400, -702, 3, 400, 700, 1000 is 3. and the median of 1, 2, 3, 4, 5 is also 3

In the first example, the min-max (or range) differentiates the two datasets. In the second example, the IQR gives you a sense of how different both datasets are

3. They help you to calculate statistical significance of the difference between two or more populations. But more on that later

RANGE

- The range is the maximum minus minimum
- Example – the range of $x_i = 1, 2, 2, 1, 4, 10, 8$ is $10 - 1 = 9$
- In practice, we often just report the minimum and the maximum in our papers.
- The range of key variables is important to report in papers as it helps to orient the reader, and helps us compare the variable across samples and populations.

Example
The GHQ-12 questionnaire consists of 12 items, scored 1, 2 or 3 each. The maximum possible GHQ-12 score is 36, and the minimum possible score is 12.
In our study of residents in Lagos, the maximum GHQ-12 score was 34 while the minimum was 18. The range of 34-18=16 is not interpretable.

PERCENTILES

- The p^{th} percentile is the value with p% of the observations at or below it.
- Example, what weight does a child have to be to be in the top 10% of body weights in a Kindergarten class?

- The 25th percentile is the value with 25% of the observations at or below it; or 75% above it. Also called the first quartile, or Q1.
- The 33rd percentile is the value with 33% of the observations at or below it, or 67% above it.
- The 50th percentile is the median.
- The 75th percentile is the third quartile, or Q3.

- Percentiles are called **quantiles** when expressed as a proportion, instead of a percentage.
  - 25th percentile = 0.25 quantile

DATA CLEANING TIP

In most statistics classes, you are provided with clean data for pedagogical reasons.

In your own research, the first task when you collect your data is to clean it.

Data cleaning typically involves looking for errors in your data – that may have arisen from typos or mindless data entry.

Two common ways

- look for extreme patterns that don't make sense – *look at 987 in the table on the right*

- look for patterns that are impossible from your understanding of the world e.g. babies cannot have beards

Statistics

fe_bsl

N	Valid	1484
	Missing	18
Percentiles	1	12.0000
	5	13.0350
	10	14.5600
	20	17.5300
	30	20.9000
	40	25.2100
	50	30.4500
	60	37.4000
	70	44.4000
	80	56.9900
	90	82.4100
	95	105.7875
	99	161.6460
	100	987.0000

INTERQUARTILE RANGE

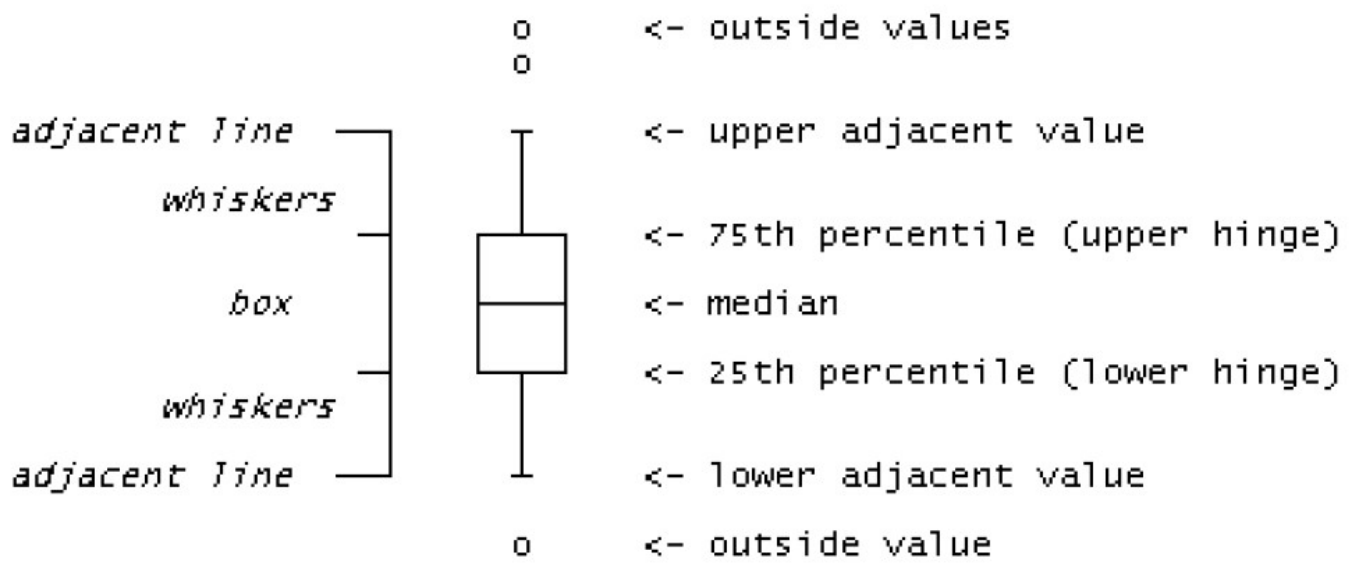
- The interquartile range (IQR) of a distribution is Q3 minus Q1.
- As for range, we often just report IQR as (Q1, Q3).

- **Outliers** or outside values are observations that are more than
  - 1.5 * IQR below Q1, i.e. Q1 – (1.5*IQR)
  - 1.5 * IQR above Q3 , i.e. Q3 + (1.5*IQR)

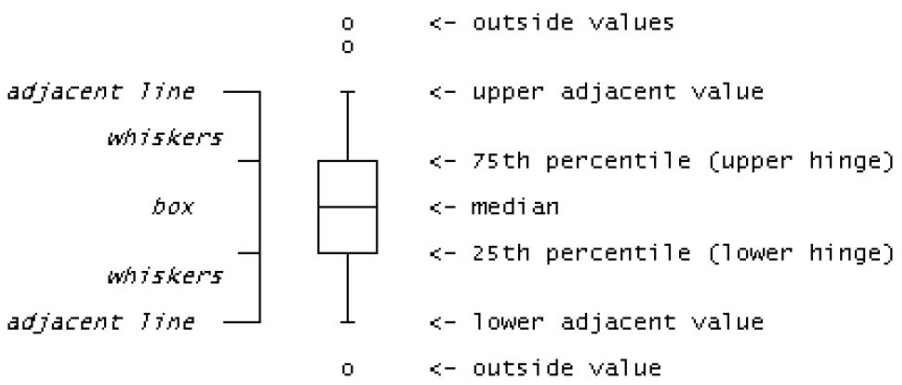
- The adjacent value is the lowest or largest value that is not an outlier.

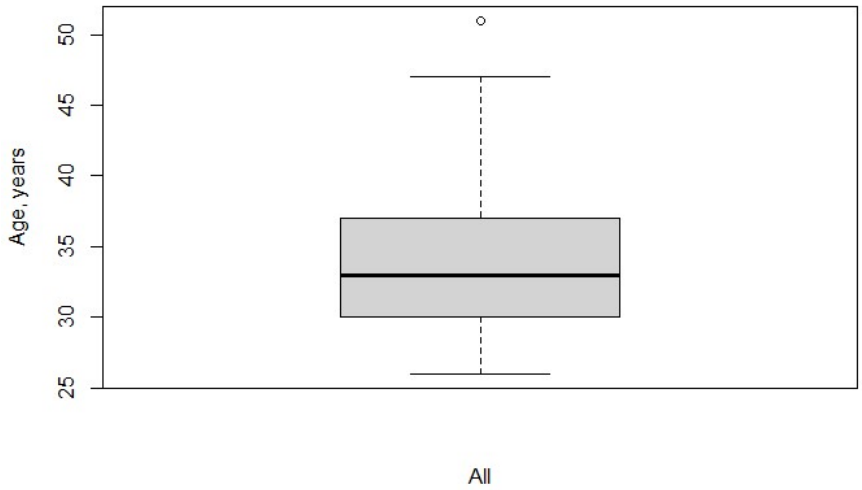
The boxplot is useful to visualize the IQR and outliers.

BOX PLOT


o
o
<- outside values
adjacent line
whiskers
box
whiskers
adjacent line
<- upper adjacent value
<- 75th percentile (upper hinge)
<- median
<- 25th percentile (lower hinge)
<- lower adjacent value
o
<- outside value

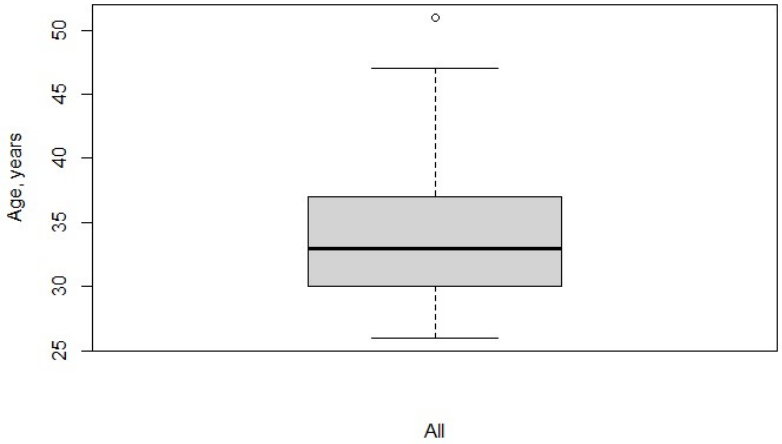
BOX PLOT

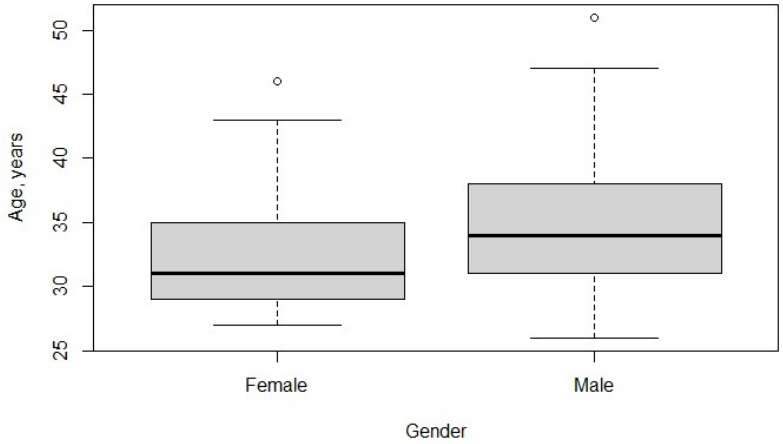

o
o
<- outside values
adjacent line
whiskers
box
whiskers
adjacent line
<- upper adjacent value
<- 75th percentile (upper hinge)
<- median
<- 25th percentile (lower hinge)
<- lower adjacent value
o
<- outside value


Age, years
25
30
35
40
45
50
All

BOX PLOT

Take a look at this example from the age of resident doctors in Lagos Hospital.





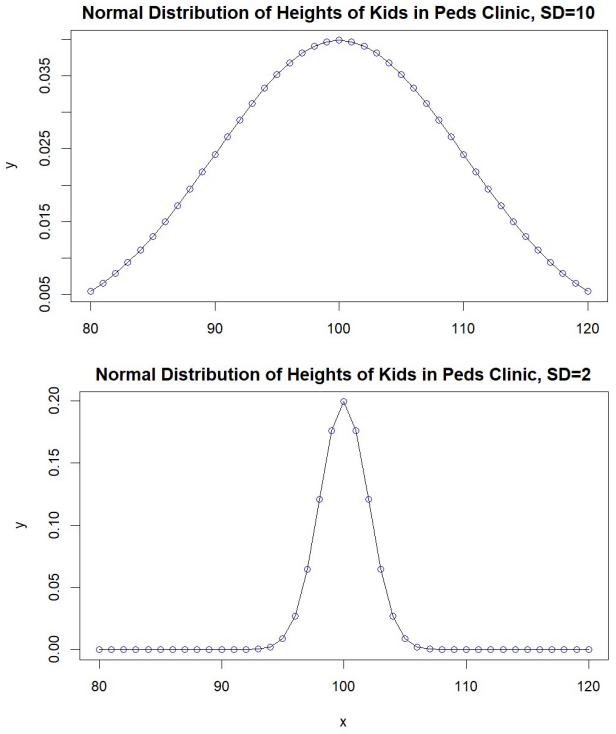
Measure	Value
Mean (SD)	34.1 (4.9)
Median (IQR)	33.0 (30.0, 37.0)

Measure	Female	Male
Mean (SD)	32.8 (4.8)	34.8 (4.9)
Median (IQR)	31.0 (29.0, 35.0)	34.0 (31.0, 37.8)

STANDARD DEVIATION (SD)

- The standard deviation is a measure of the spread of the values of the variable around the mean
  - Small SD suggests the data are clustered around the mean
  - Large SD suggests the data are more spread out
- The SD is an appropriate measure of spread if:
  - The distribution of the variable is symmetric, and the mean is a good measure of the center
  - Outliers are rare

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \qquad \text{Variance} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$



STANDARD DEVIATION (SD)

Take an example dataset for the ages of 10 resident doctors from our dataset.

- 31, 38, 33, 36, 29, 35, 28, 31, 29, 31; Recall the mean, $\bar{x}$= 32.1

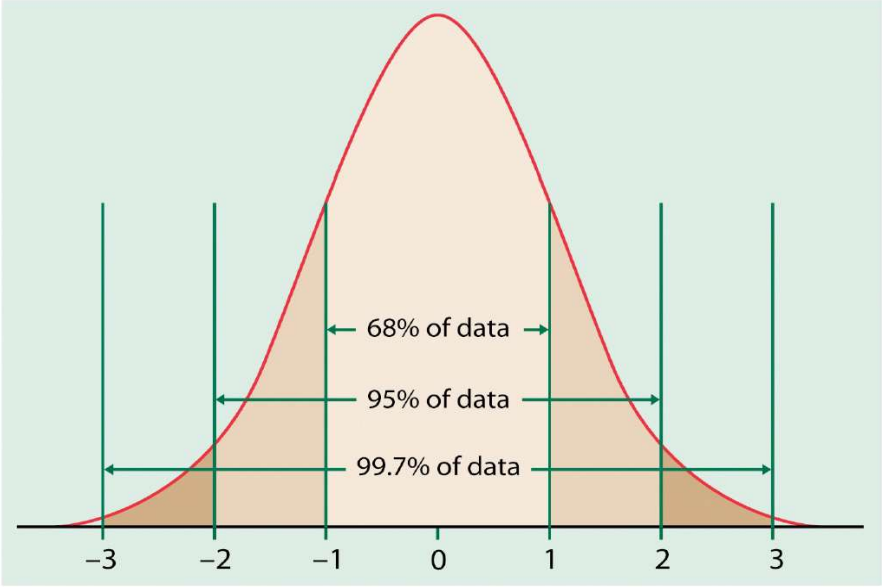
$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{98.9}{9}} = \sqrt{10.99} = 3.31$$

$n-1$ in the formula represents the degrees of freedom.

It is used instead of n because $\sum(x_i - \bar{x}) = 0$

Age_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
31	-4.1	16.81
38	-3.1	9.61
33	-3.1	9.61
36	-1.1	1.21
29	-1.1	1.21
35	-1.1	1.21
28	0.9	0.81
31	2.9	8.41
29	3.9	15.21
31	5.9	34.81
Total	0	98.9

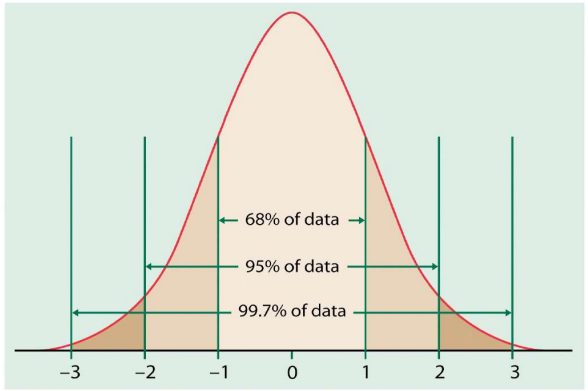
SD and the normal distribution



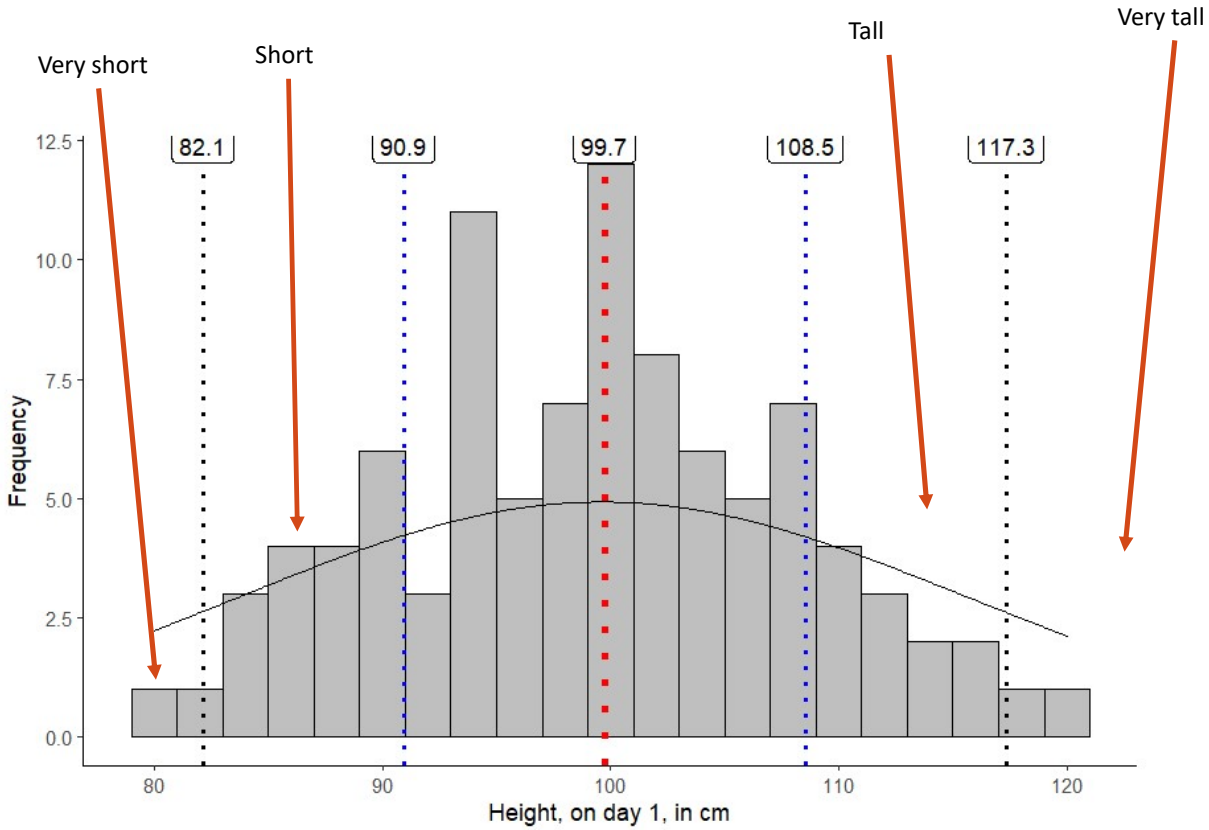
Sometimes, we have a symmetric and unimodal distribution, such as pictured, known as the normal distribution.

- Usually, about ~68% of observations lie within one standard deviation of their mean. That is, mean ±1 SD covers ~68.2% of observations
- Usually ~95% of observations lie within two standard deviations of their mean. That is, mean ±1.96 SD covers approx. 95% of observations
- Mean ±3 SD covers approx. 99.7% of observations

SD and the normal distribution



When we derive reference ranges of a measure, e.g. weight for height, we use this distribution to determine values that are extremely small or extremely large, depending on their distance from the mean.



Converting Means and SD From Unit To Unit

Sometimes, we receive data measured in one unit, e.g. meters, and we want to convert to another, e.g. centimeters.

- If you multiply the mean by a constant, e.g. 100, to convert from meters to centimeters, you multiply the SD by the same constant too.
- Same applies to division.

- If you add or subtract a constant from the mean, the SD stays the same!

POPULATION AND SAMPLE MEASURES

Population

- When we conduct a study, our goal is usually to understand a phenomenon about the population. Say, a **population mean**.
- For instance, we may aim to estimate the mean blood pressure among adults in Freetown.
  - Population mean is denoted as μ
  - Population standard deviation is σ

Sample

- In reality, we conduct a study among a sample of resident of Freetown, and estimate a **sample mean, $\bar{x}$** and a sample standard deviation (sd)
- The sample mean may not necessarily equal the population mean.

STANDARD ERROR (SE)

- Rather than conduct our study of blood pressure among adults in Freetown one time, we can repeat it many times. Each time, we sample a group of adults and measure their blood pressure.
- If we take several random samples (~30 or more times), and calculate the mean of each sample, $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_n$, the mean of those sample means is ~~likely to be~~ the population mean.
- This phenomenon is called the Central Limit Theorem. It is the reason we can conduct studies on small samples and extrapolate to the general population
- We can also calculate the standard deviation for the distribution of those sample means. This is the **standard error**.

STANDARD ERROR (SE)

- In reality, we would conduct our study of blood pressure in Freetown one time, and calculate a sample mean.
- The standard error tells us how close to the population mean our sample mean is likely to be.
- We almost never know the population mean. So, the SE is a good estimate of the precision of our sample mean.
- The SE and the SD are not the same. We must clearly label which we present in our analyses.
- In fact, we can calculate the SE from the SD. $SE = \frac{\sigma}{\sqrt{n}}$

STANDARD ERROR (SE)

Again, using our dataset for the ages of 10 resident doctors.

- 31, 38, 33, 36, 29, 35, 28, 31, 29, 31; Recall the mean, $\bar{x}$= 32.1

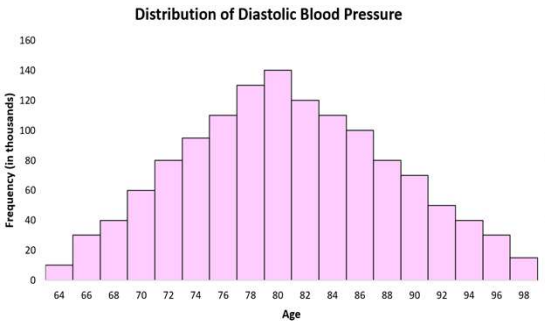
$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{98.9}{9}} = \sqrt{10.99} = 3.31$$

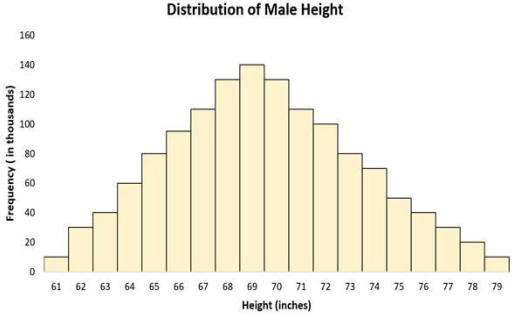
$$SE = \frac{SD}{\sqrt{n}} = \frac{3.31}{\sqrt{10}} = \frac{3.31}{3.16} = 1.05$$

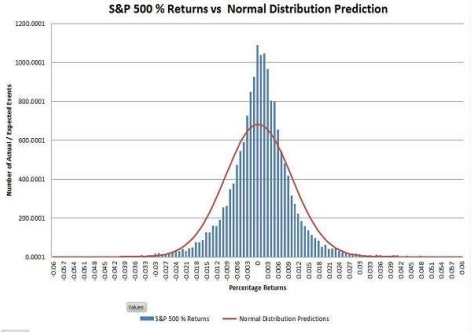
Age_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
31	-4.1	16.81
38	-3.1	9.61
33	-3.1	9.61
36	-1.1	1.21
29	-1.1	1.21
35	-1.1	1.21
28	0.9	0.81
31	2.9	8.41
29	3.9	15.21
31	5.9	34.81
Total	0	98.9

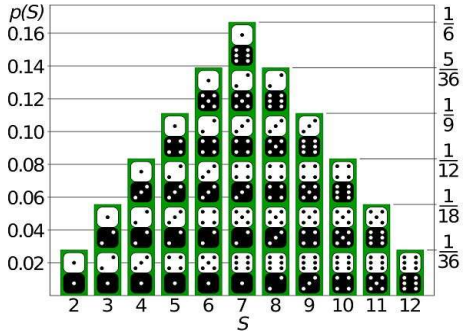
NORMAL DISTRIBUTION

- The normal distribution is the most important distribution in Statistics. It makes a lot of basic and advanced statistics possible.
- Normal distributions are everywhere.

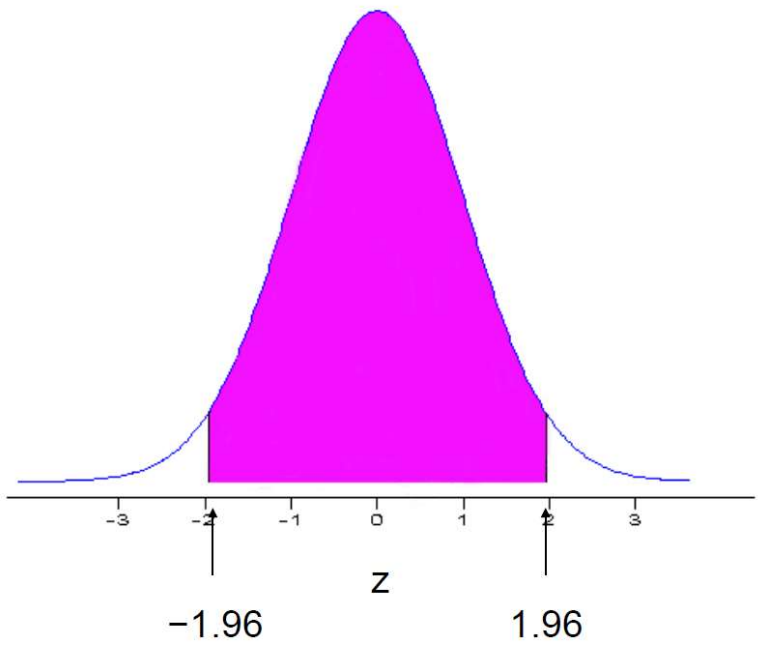








Standardized normal distribution



If the mean = 0 and SD=1, this distribution is referred to as standard normal or bell-shaped.

Any normally distributed variable can be converted to its standard normal distribution by subtracting the mean from each observation and dividing by the standard deviation.

- This is the basis of the diagnosis of malnutrition in children using weight-for-height Z scores, etc

WRAP-UP

What are your key takeaways from today's class?

ACKNOWLEDGEMENT

These slides draw considerably from the following sources:

- Professors' classnotes
  - David Harrington, Harvard
  - Murray Mittleman, Harvard
- Textbook(s)
  - Medical Statistics by Betty Kirkwood and Jonathan Sterne
  - Biostatistics and epidemiology: a primer for health and biomedical professionals by Wassertheil-Smoller S and Smoller J
- Websites
  - Boston University School of Public Health website: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/Menu/index.html
  - Statistical tools for high-throughput data analysis (STHDA): http://www.sthda.com/english/
  - Penn State University's Statistics course websites https://online.stat.psu.edu/stat507/lesson/12/12.3