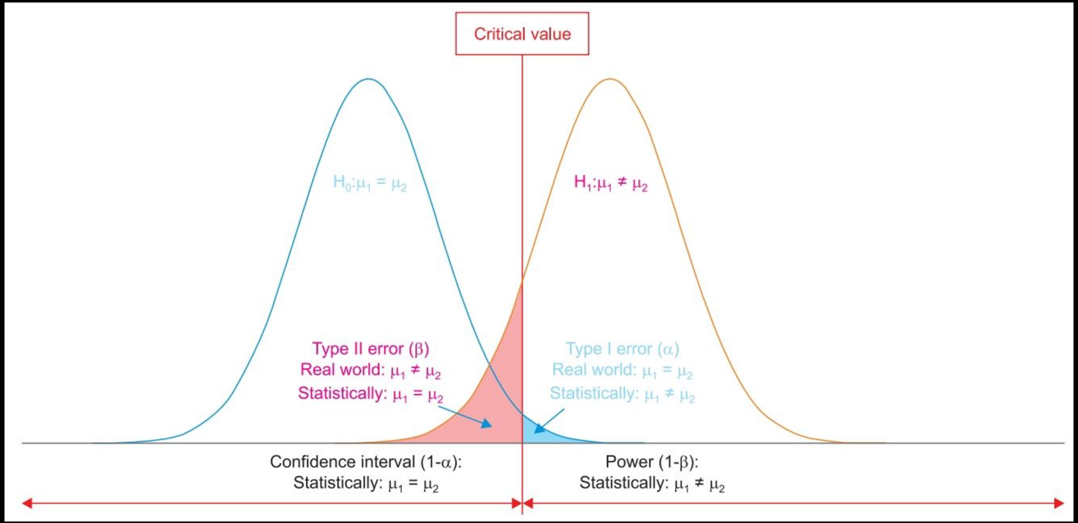
Testing difference of means



Saket Choudhary
saketc@iitb.ac.in

Introduction to Public Health Informatics

DH 302

Lecture 07 || Wednesday, 29th January 2025

From last lecture…

- Review: Expectations, Variances, CLT, Normal approximation
- Testing for difference of means
- Dimensionality reduction primer

Expectations and Variances

The expected value of a random variable is the average value of all expected outcomes. More formally,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{j=1}^{J} x_j P(X = x_j) \\ &= x_1 p_1 + x_2 p_2 + \cdots + x_J p_J.\end{aligned}$$

The expectation of the sum = Sum of the expectations. More formally,

$$\mathbb{E}[X_1 + X_2 + \cdots + X_N] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_N] \quad (1)$$

Expectations and Variances

When we observe (realize) different values of $\mathcal{X}$, one natural question to ask is how scattered these values tend to be from the expected value. This is quantified by "variance". More formally,

$$
\begin{aligned}
Var[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] && (2) \\
&= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] && (3) \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && (4)
\end{aligned}
$$

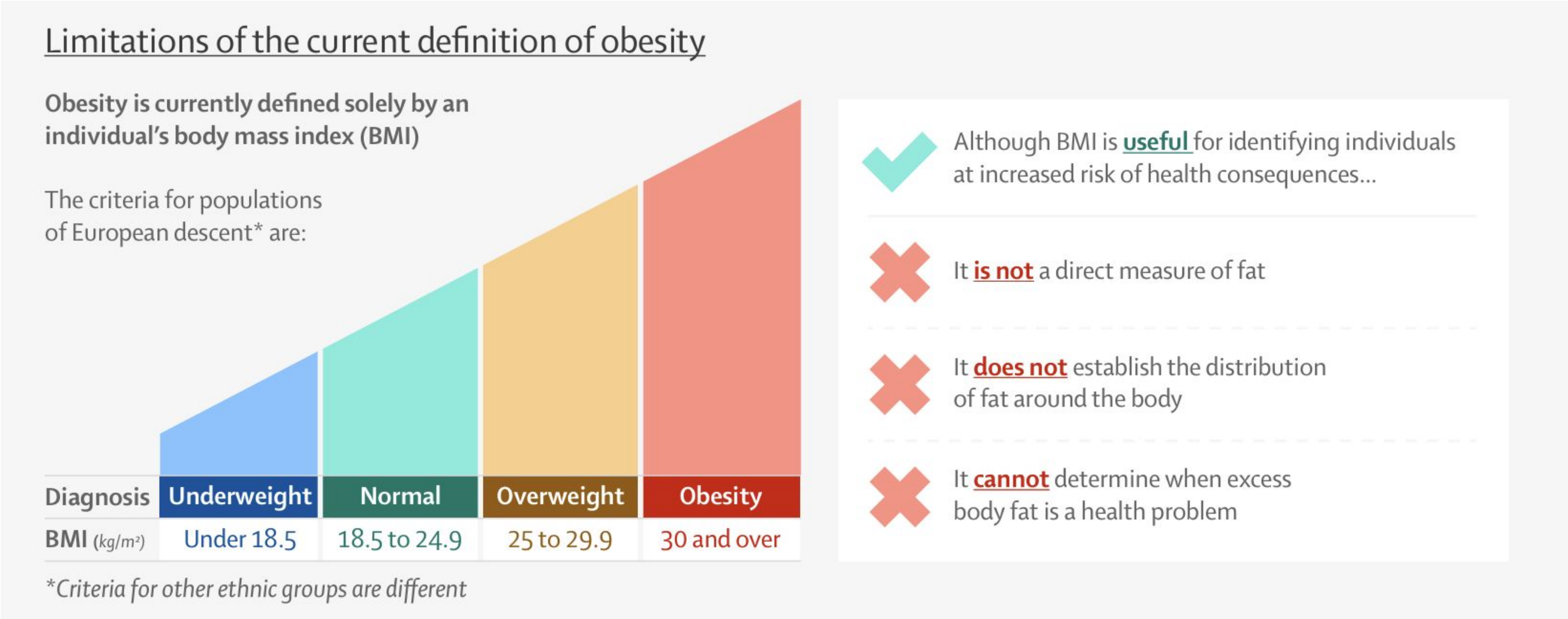
The positive square root of $Var[X]$ is called standard deviation of X.

For independent variables, variation of their sum is equal to sum of individual variances, i.e. when X_i are independent,

$$Var(X_1 + X_2 + \dots X_N) = Var(X_1) + Var(X_2) + \dots Var(X_N)$$

Some digression

Obesity and BMI - The old paradim



Lancet

Obesity: requirement of the new definition

#	1	2	3	4
BMI (kg/m²)	28.2	30.1	36.2	36.2
Diagnosis	Overweight	Obesity	Obesity	Obesity
Excess body fat?	Yes	No	Yes	Yes
Signs and symptoms?†	No	No	No	Yes
Notes	Under-diagnosis of obesity	Over-diagnosis of obesity	Obesity with preserved health	Obesity with ongoing illness

Limitations of BMI-based diagnosis

!

People with excess body fat do not always have a BMI above 30, meaning that their health risk can go unnoticed.

Individuals with high muscle mass (eg, athletes) tend to have high BMIs despite normal fat mass. Diagnosing such people as having obesity or a disease is inappropriate.

Some people with excess body fat (and high BMI) can nevertheless maintain normal organ function and an unhindered ability to conduct daily activities (hence, they have no illness); others instead manifest objective evidence of ongoing illness. Current definition and measures of obesity do not reflect health/illness at individual level and are therefore inadequate for disease diagnosis.

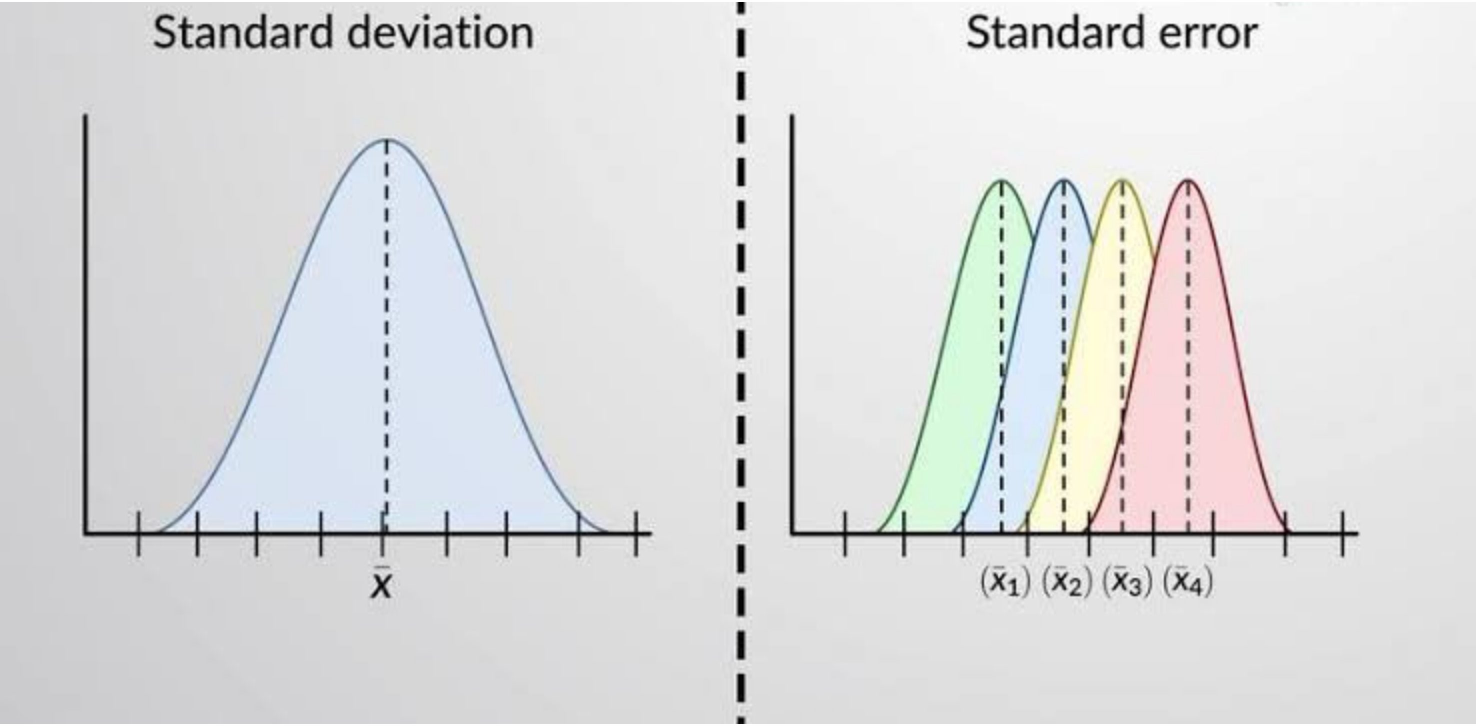
Lancet

Obesity the new definition

#	1	2	3	4	5	6
BMI (kg/m²)	23.7	28.8	28.8	32.4	39.2	39.2
Excess body fat?	No	No	Yes	No	Yes	Yes
Muscle mass	Normal / High	Normal	Normal / Low	High	Normal / Low	Normal / Low
Signs and symptoms?*	No	No	No	No	No	Yes
Old diagnosis	No obesity	Overweight	Overweight	Obesity	Obesity	Obesity
New diagnosis	No obesity	No obesity	Preclinical obesity	No obesity	Preclinical obesity	Clinical obesity

Testing for mean differences

Primer: Standard error of the mean



Primer: Standard error of the mean

Standard error of the mean:

- $\bar{X}$ is an **estimate of the population mean μ**, but how good of an estimate is it?
- How "good" (close to the population mean) the sample mean x It depends on the sample and is often referred to as the sampling error.
- If you could sample the entire population this sampling error will be zero.
- The standard deviation of the sampling distribution of the sample mean is given by: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ where n is the number of times sample mean was drawn
- But, we do not always have σ available, so we use its estimate, the sample standard deviation s. The standard error of the mean is then defined as: $\mathrm{SE}_{\bar{X}} = \frac{s}{\sqrt{n}}$

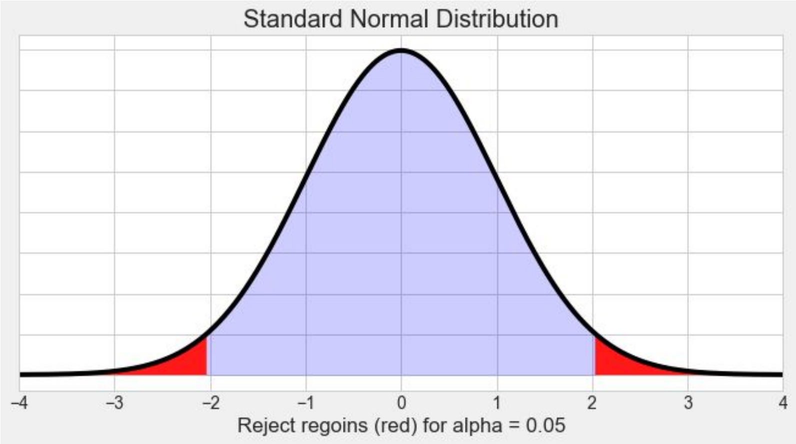
Standard Error vs Standard Deviation:

- Standard deviation (SD) **describes the dispersion of the data**
- **Standard error (SE) describe the unreliability of the mean of the sample** due to sampling error
- As the sample size increases, sample mean and the standard deviation both approach the population mean and standard deviation respectively
- The standard error tends to decrease as the sample size n increases → sample mean becomes more and more precise estimate of the population mean.

Testing for the mean of a distribution: Z-test

Null hypothesis: Is the mean value of the sample a given quantity

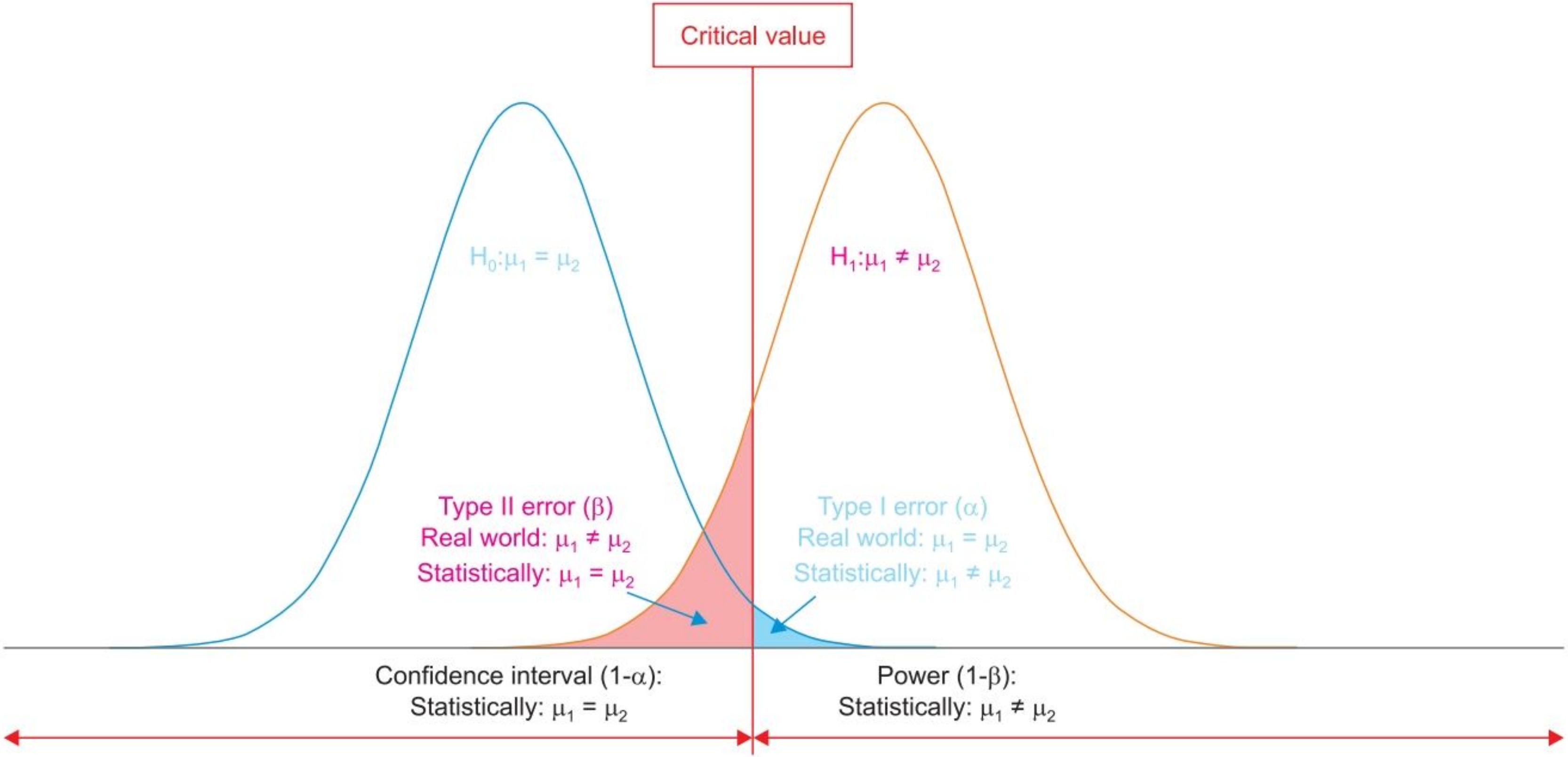
$$Z = \frac{(\bar{X} - \mu_0)}{s}$$



Only valid if the population variance is known

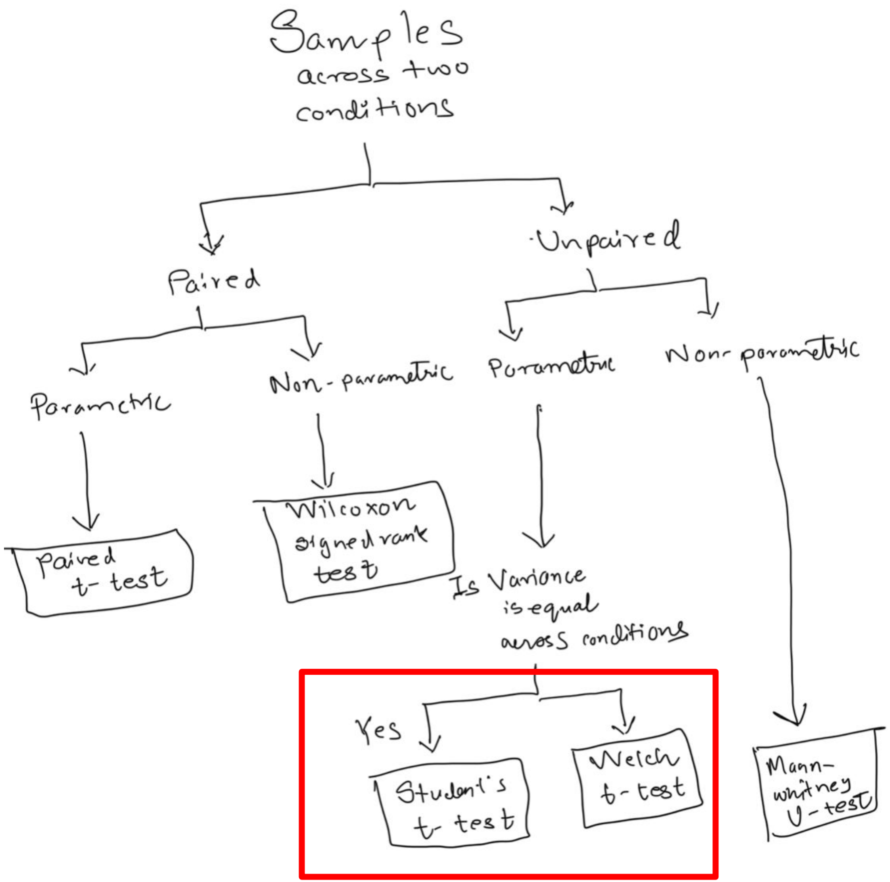
[Source]

Testing for difference in mean (median) of two samples



Source

Testing for difference in mean (median) of two samples



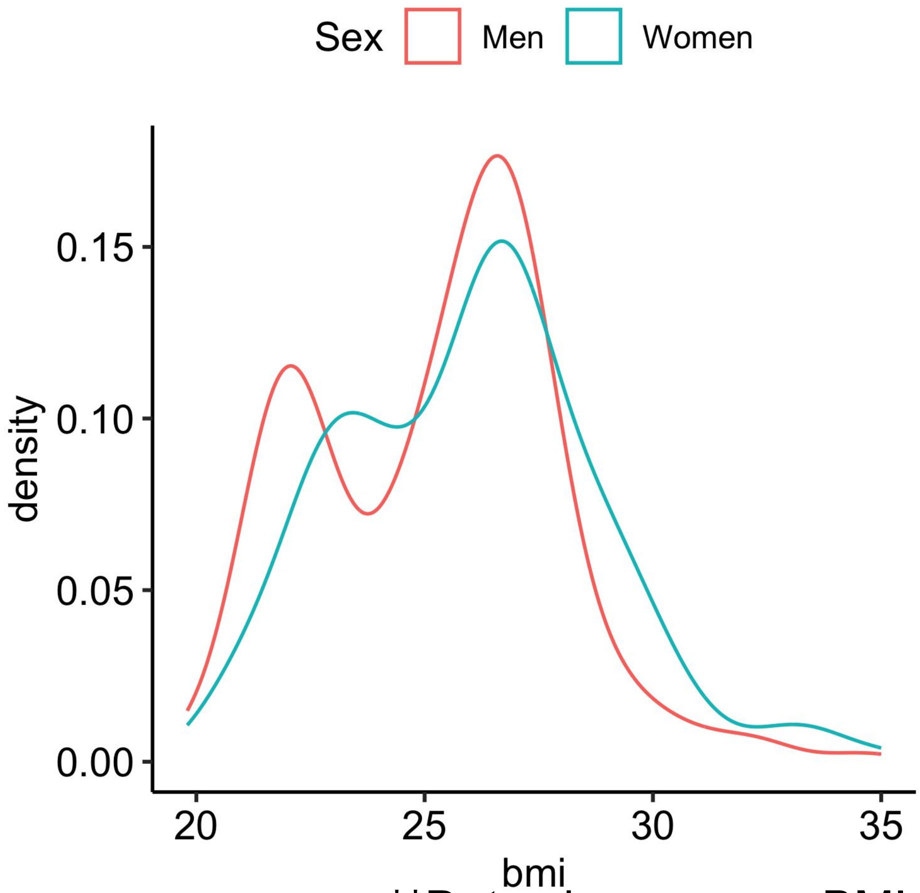
Parameteric tests = Have a parameter → used when the distribution is known

Non-parameteric tests → used when the distribution is unknown

Paired tests: Used when observation is made on related or same pair of objects (before and after)

Unpaired tests: For non-related observations across two groups

Testing for difference of means



<u>Question</u>: Is there statistically significant difference in mean between men and women BMI?

What is the null hypothesis?

Null Hypothesis: The mean bmi is same for men and womean

[Data source]

**Data shows mean BMI distribution across countries in 2017

T-test paradigm

$$H_0 : \mu_1 = \mu_2$$

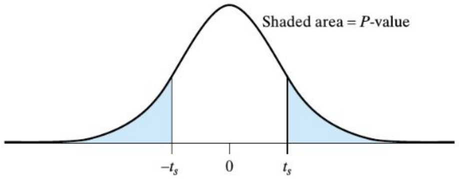
The t-statistic t_s measures how far are the mean difference of the groups $\bar{y}_1 - \bar{y}_2$ from the expectation of zero, if the null were true:

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}}$$

The 's' in t_s indicates that this t-statistic is calculated from the samples. If H_0 is true, then t_s follows a Student's t-distribution approximately with degrees of freedom given by

$$df = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}}$$

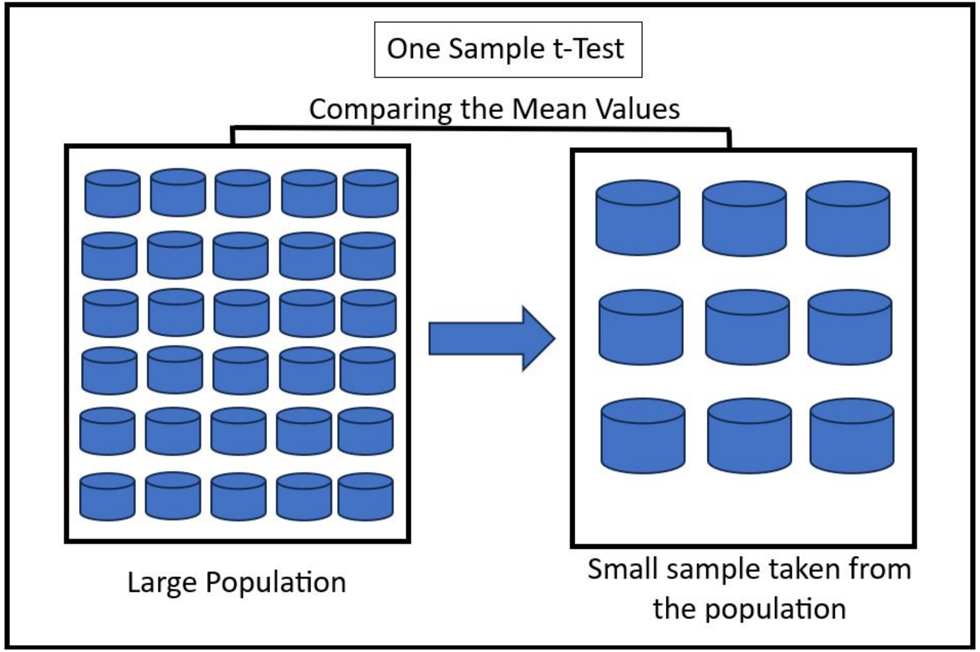
t_s Follows a t-distribution with df degrees of freedom



T-test assumptions

- Observations from both the groups are approximately normally distributed
- The difference of means $\bar{y}_1 - \bar{y}_2$ is independent of the standard error of the difference $SE_{\bar{y}_1 - \bar{y}_2}$

T-test one sample



Question: Is the mean of the sample equal to a pre-specified value?

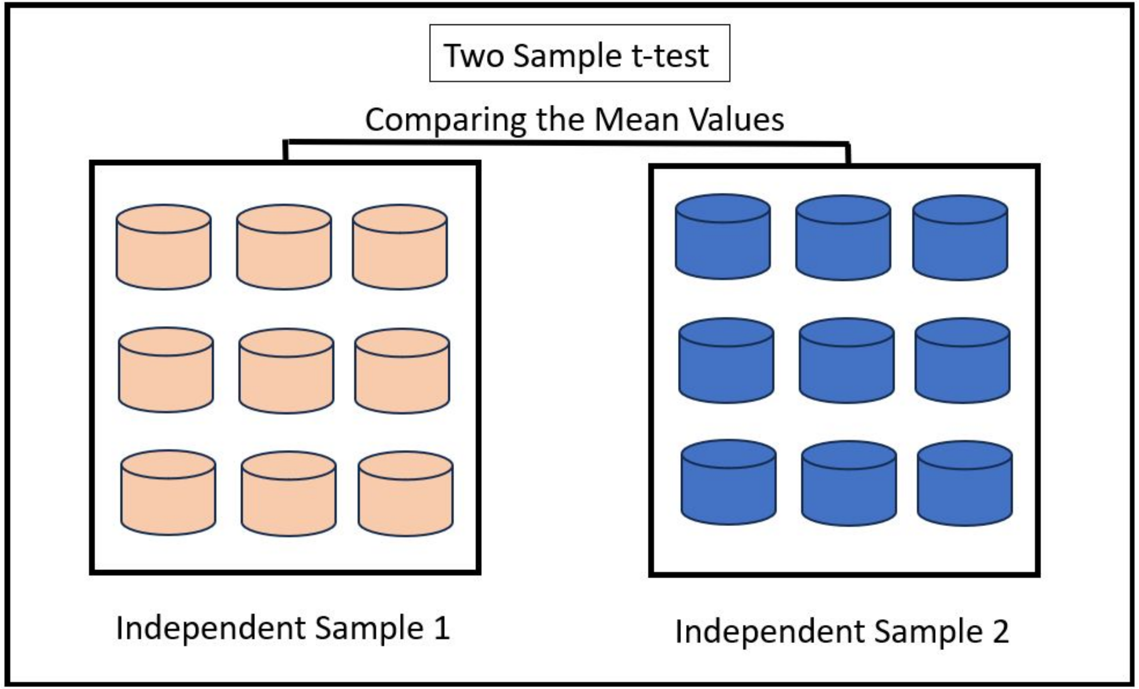
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

μ_0: Specified value of the mean

s: standard deviation of the sample

Source

T-test two sample (Equal size and equal variance)

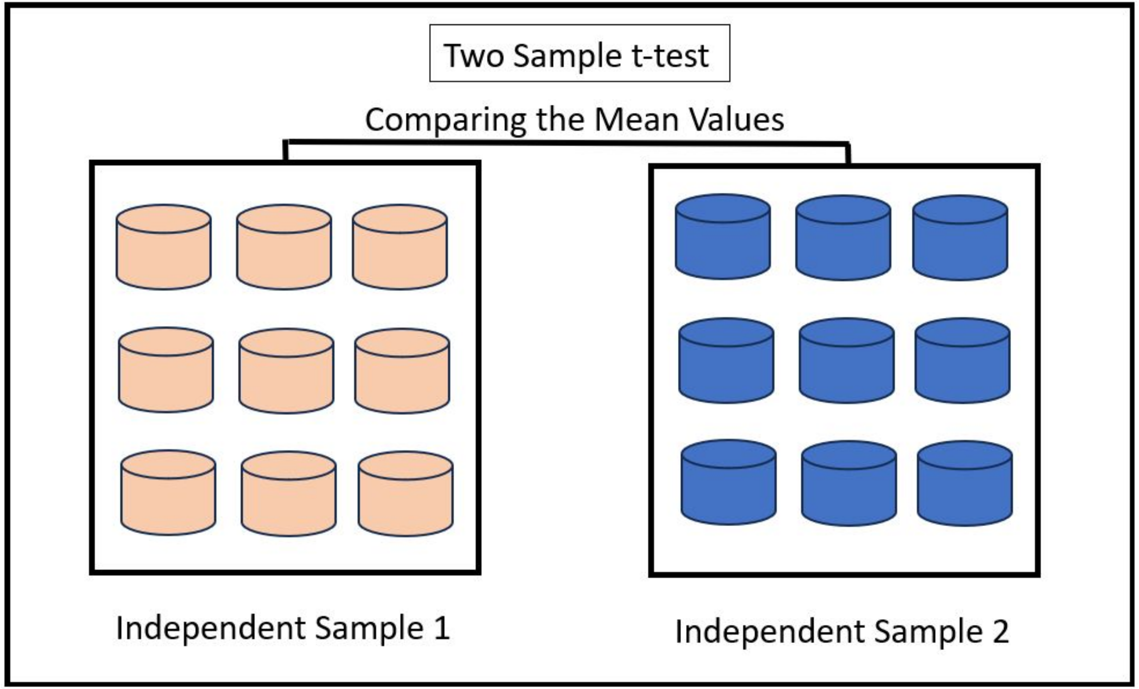


Question: Are the mean of two samples equal?

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}}$$

We will assume (conservatively) the variance of the two samples are different (this version of the t-test is called the "Welch" t-test)

T-test two sample



Question: Are the mean of two samples equal?

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}}$$

We will assume (conservatively) the variance of the two samples are different (this version of the t-test is called the "Welch" t-test)

Sample sizes need not be equal (unlike the picture)

T-test: The scenarios

Equal sizes, same variance

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$$

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$$

d.f. = 2n-2

Equal or unequal sizes, similar variance

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

$$\frac{1}{2} < \frac{s_{X_1}}{s_{X_2}} < 2$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

d.f. = n1+n2-2

Equal or unequal sizes, unequal variance (Welch test)

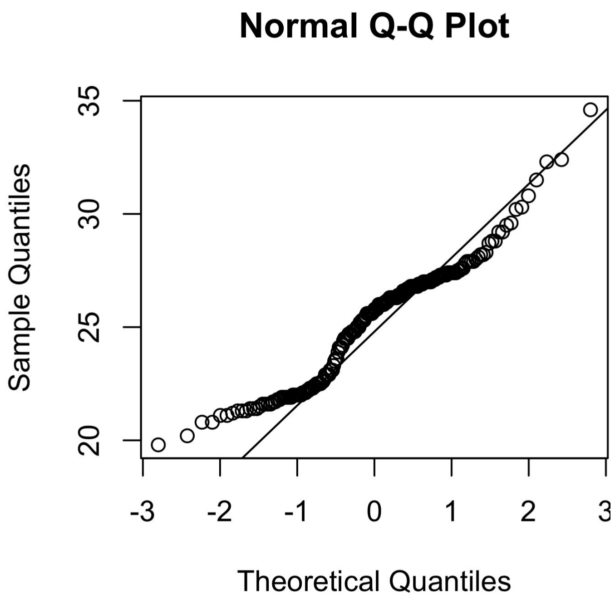
$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

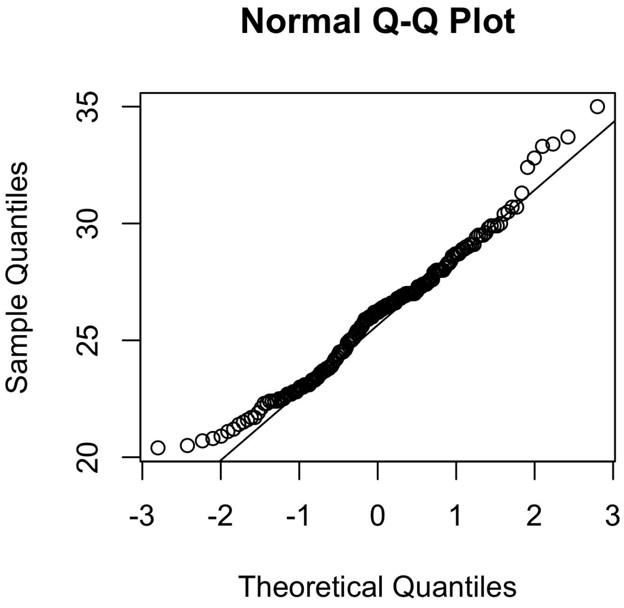
Verifying the ‘normality’ assumption using QQplot

```
qqnorm(men_rural_2017$bmi)
qqline(men_rural_2017$bmi)
```



Men

```
qqnorm(women_rural_2017$bmi)
qqline(women_rural_2017$bmi)
```



Women

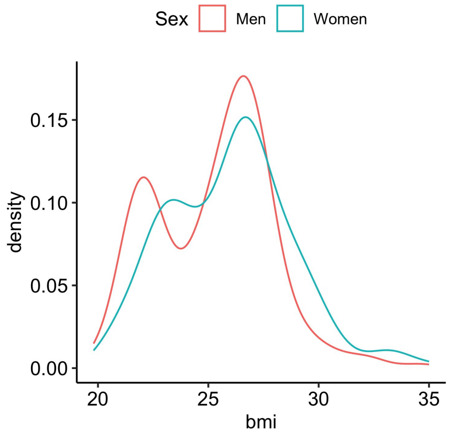
- We can test the normality assumption by plotting the sample quantiles against the expected “theoretical” quantiles of a standard normal in a QQ plot
- QQ plot is the plot of the quantiles of the first dataset against the quantiles on the second dataset
- What is a Quantile? Fraction of points below the given value
- For example, 30% quantile value represents that 30% of the values in dataset lie below this point (and 70% lie above)

T-test: BMI differences between males and females are statistically significant?

```
> t.test(x = men_rural_2017$bmi, y = women_rural_2017$bmi)

    Welch Two Sample t-test

data:  men_rural_2017$bmi and women_rural_2017$bmi
t = -2.6774, df = 388.7, p-value = 0.007734
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -1.2741967 -0.1951911
sample estimates:
mean of x mean of y
 25.24490  25.97959
```



We reject the null hypothesis that the mean bmi of men and women is equal

Problem

Twenty-two volunteers at a cold research institute caught a cold after having been exposed to various cold viruses. A random selection of 10 of these volunteers was given tablets containing 1 gram of vitamin C. These tablets were taken four times a day. The control group consisting of the other 12 volunteers was given placebo tablets that looked and tasted exactly the same as the vitamin C tablets. This was continued for each volunteer until a doctor, who did not know if the volunteer was receiving the vitamin C or the placebo tablets, decided that the volunteer was no longer suffering from the cold. The length of time the cold lasted was then recorded.

Treated with Vitamin C	Treated with Placebo
5.5	6.5
6.0	6.0
7.0	8.5
6.0	7.0
7.5	6.5
6.0	8.0
7.5	7.5
5.5	6.5
7	7.5
6.5	6.0
	8.5
	7.0

This can further be summarized as follows:

	Treated with Vitamin C	Treated with Placebo
$\bar{y}$	6.45	7.125
n	10	12
SD	0.761	0.882
SE	0.240	0.254

In the context of this study, state the null and alternative hypotheses.

Test the hypothesis.

Questions?