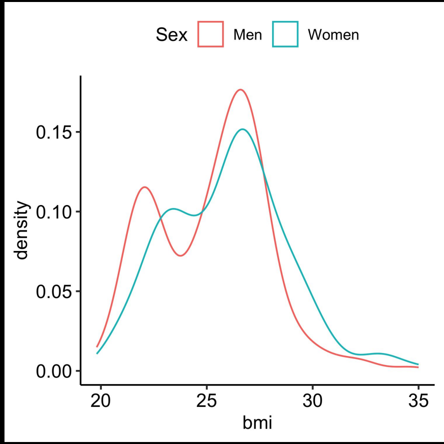
Hypothesis testing - 2



Saket Choudhary
saketc@iitb.ac.in

Introduction to Public Health Informatics

DH 302

Lecture 06 || Friday, 24th January 2025

From last lecture…

In a land of stats, so wild and vast,
Hypothesis testing had students aghast.
"Is it null? Is it not? Should we reject?"
Confusion spread, hard to correct.

Then came the voice, "Here's the key,
State your null as plain as can be.
Assume it's true, don't let it stray,
And let your data have its say."

But oh, the p-values, they played their tricks,
"Below 0.05? It's a statistical fix!"
"Above that line? We must comply—
The null survives, we let it fly."

Coauthored with ChatGPT

From last lecture…

https://xkcd.com/892/

From last lecture…

- Review: Chi-squared test and G-test
- Expectations, Variances, CLT, Normal approximation
- Testing for difference of means
- Dimensionality reduction primer

Why do we select the null as such?

- Purpose of hypothesis testing is largely to impose self-skepticism (“You are innocent unless proven guilty”)
- We usually take the occam’s razor approach, assume the simplest thing that could be true
- "*We cannot conclusively affirm a hypothesis, but we can conclusively negate it*" – Karl Popper
- It is easy to specify the null hypothesis, often we don’t know what the alternate hypothesis explicitly is. For example, there is mean difference between the two populations – but how wide? But easy to say – it is zero (difference is ‘null’).
- Think about this argument: “All swans are white”. What is easier: ‘rejecting it’ or ‘accepting it’?

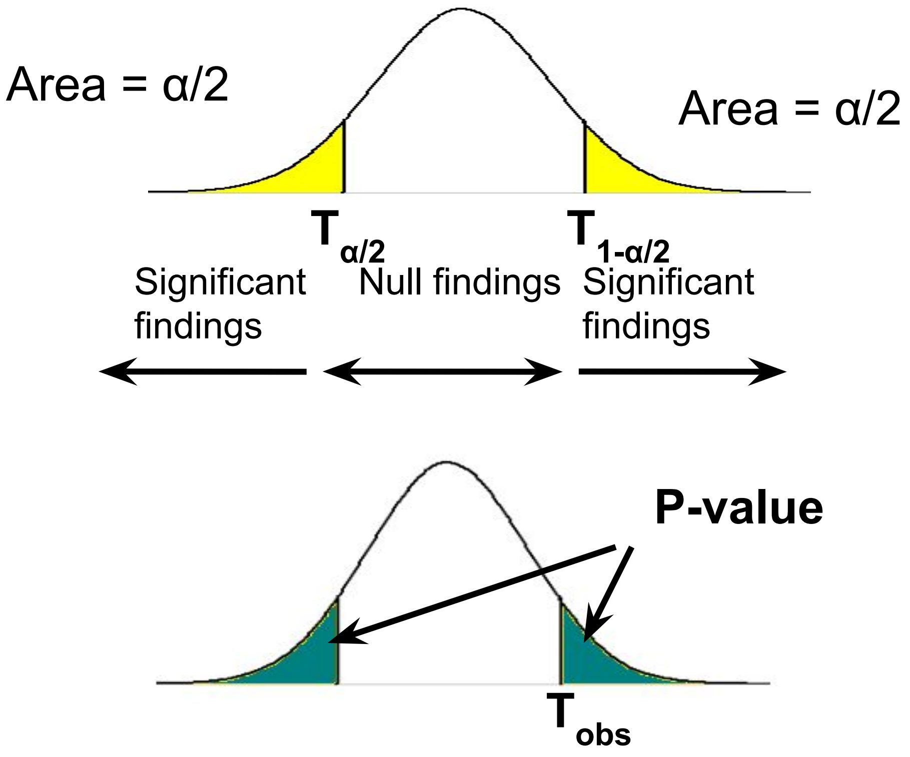
Also see

Visualizing the p-values region

P-value = Probability of sampling a test statistic at least as extreme as the observed test statistic if the null hypothesis is true

We “reject” the null hypothesis (H_0) if the pvalue is below the threshold ($\boldsymbol{\alpha}$)

Distribution of T under H_0

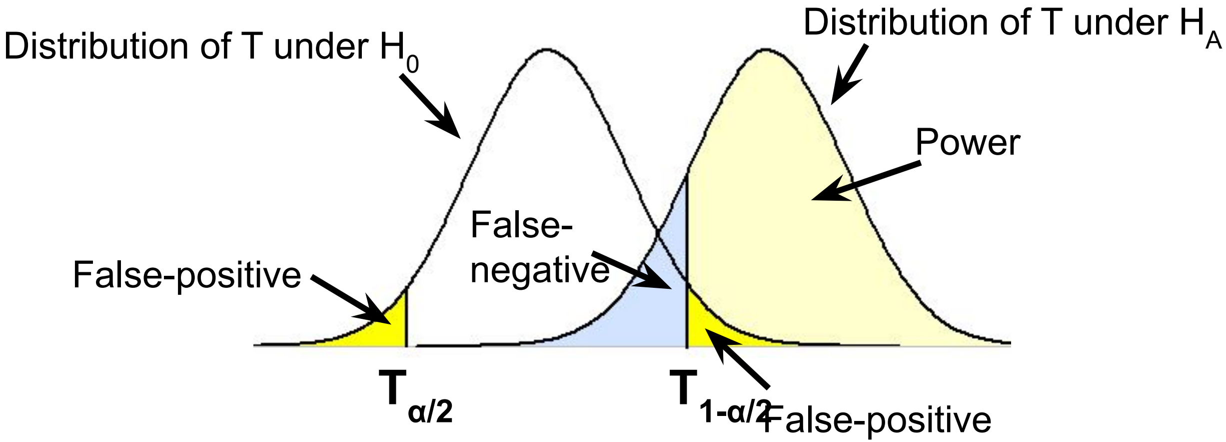


Type I,II errors and Power

- **Type I error:**
  - Probability that the **test incorrectly rejects** the null hypothesis (H_0) when the null H_0 is true
  - Often denoted by $\boldsymbol{\alpha}$
- **Type II error:**
  - Probability that the **test incorrectly fails to reject** the null hypothesis (H_0) when H_0 is false
  - Often denoted by $\boldsymbol{\beta}$
- **Power**:
  - Probability that the **test correctly rejects** the null hypothesis (H_0) when the alternative hypothesis (H_1) is true
  - Commonly denoted by 1- β where β is the probability of making a Type II error by incorrectly failing to reject the null hypothesis.
  - As β increases, the power of a test decreases.

Type I,II errors and Power

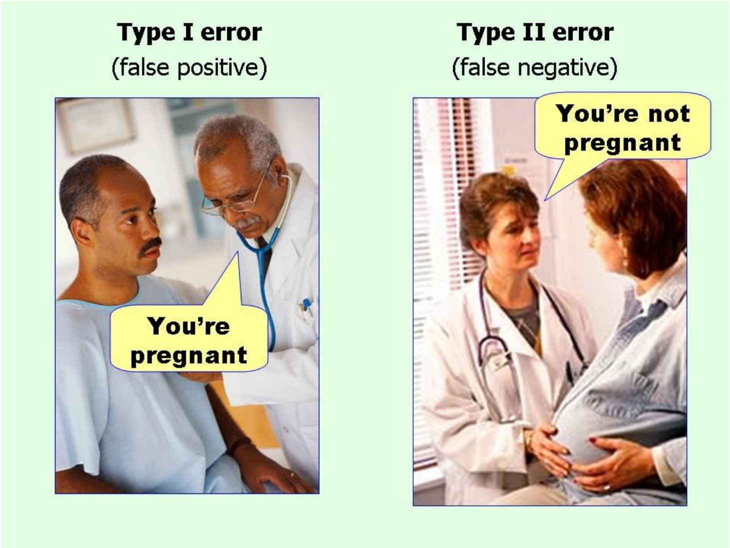
The **false-positive** rate is the probability of **incorrectly *rejecting*** H_0.
The **false-negative** rate is the probability of **incorrectly *accepting*** H_0.
Power = 1 – false-negative rate = probability of **correctly rejecting** H_0.



Types of error

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

Source



Paul Ellis, 2010

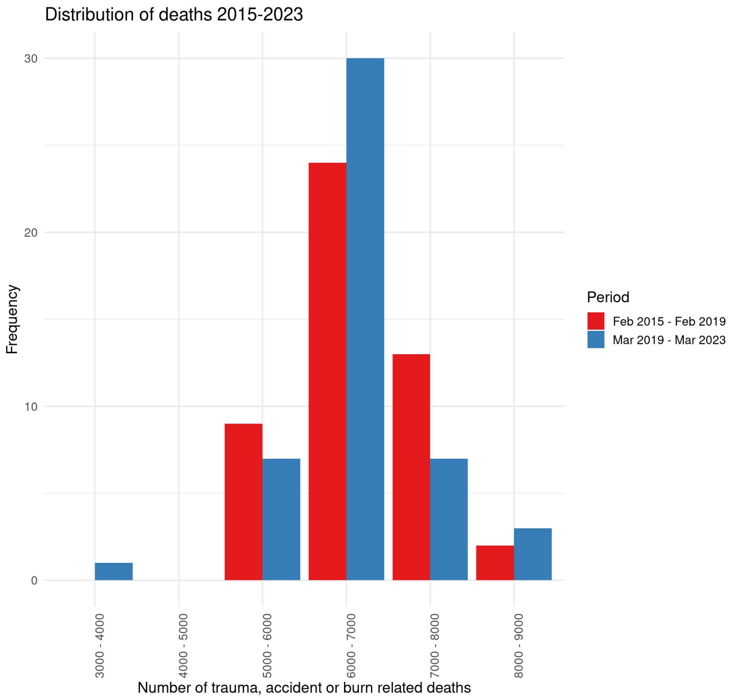
What is p-value?

- P-value is NOT the probability of the alternate hypothesis being correct.
- P-value is NOT the probability of observing the result by chance.
- P-value = Probability of observing a result at least as extreme **if the null hypothesis holds true.**

Goodness of fit - Chi-squared test

Problem: What distribution should I fit?

Use a pseudocount of +1 in frequencies



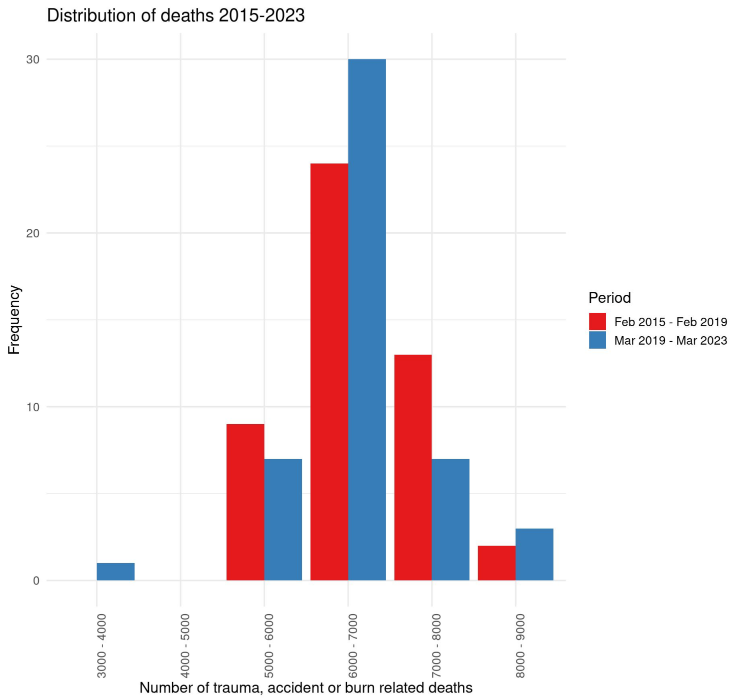
bin	Feb 2015 – Feb 2019	Mar 2019 – Mar 2023	diff	chisq
3000 – 4000	0	1	1	1.0000000
4000 – 5000	0	0	0	0.0000000
5000 – 6000	9	7	−2	0.4000000
6000 – 7000	24	30	6	1.4400000
7000 – 8000	13	7	−6	2.5714286
8000 – 9000	2	3	1	0.3333333

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 5.744762$$

Is 5.7 high/low/medium?

Example of Chi-square in R

```
chi_square_stat <- sum((observed - expected)^2 / expected)

dof <- length(observed) - 1

p_value <- pchisq(chi_square_stat, dof, lower.tail = FALSE)

alpha <- 0.05  # Significance level

if (p_value < alpha) {

  cat("Reject the null hypothesis")

} else {

  cat("Fail to reject the null hypothesis")

}
```



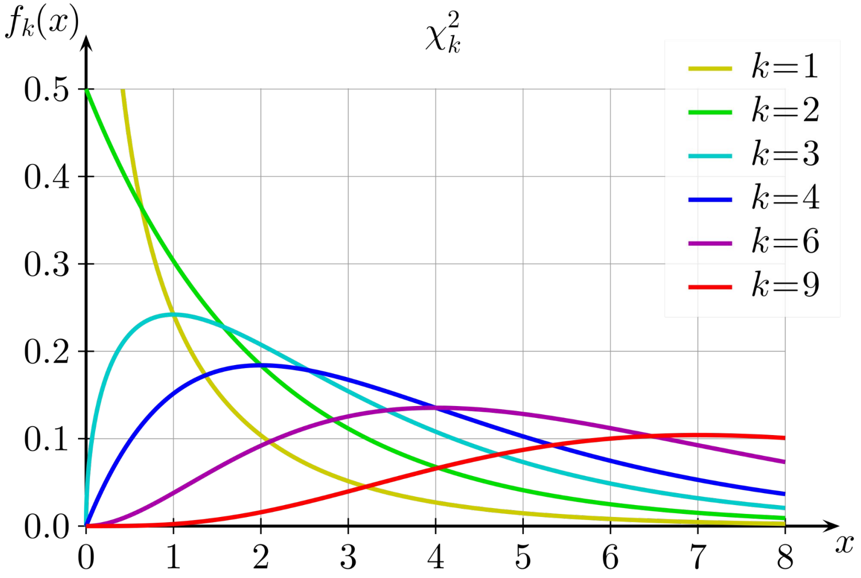
P-value = 0.33 (>0.05)

Thus, we fail to reject the null hypothesis that the there is statistically no significant difference between the frequencies observed in Mar 2019 - Mar 2023 follow the same distribution as the *Feb 2015 - Feb 2019* ones”

Another goodness of fit test - Likelihood ratio test (or G-test)

$$G = 2\sum_i O_i \cdot \ln\left(\frac{O_i}{E_i}\right)$$

$$\sum_i O_i = \sum_i E_i = N$$

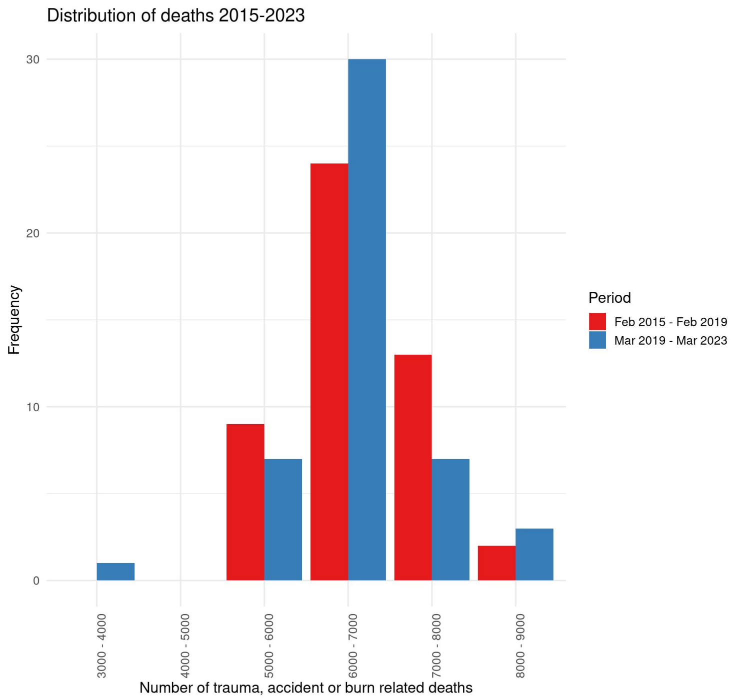


O_i = an observed count for bin i
E_i = an expected count for bin i, asserted by the null hypothesis

G follows a chi-squared distribution with degrees of freedom = (length of observations - 1)

Example of G-test in R

```
G_stat <- 2 * sum(observed * log(observed / expected),
na.rm = TRUE)

dof <- length(observed) - 1

p_value <- pchisq(G_stat, df = dof)

alpha <- 0.05  # Significance level

if (p_value < alpha) {

  cat("Reject the null hypothesis")

} else {

  cat("Fail to reject the null hypothesis")

}
```

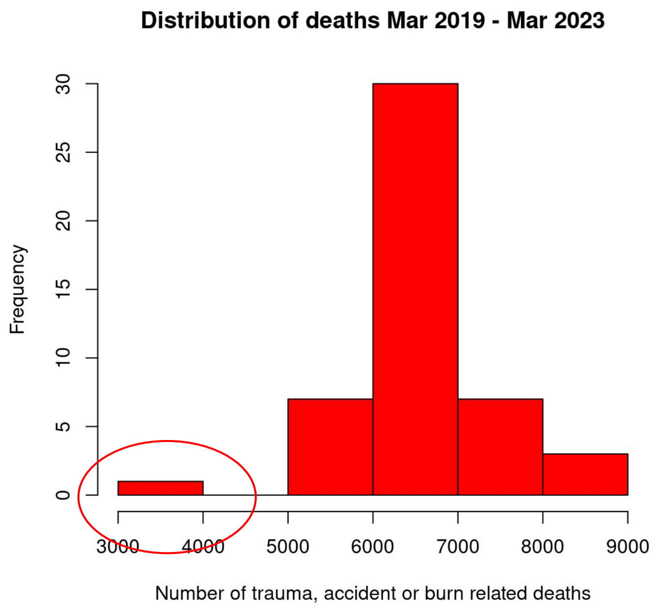


P-value = 0.59 (>0.05)

Thus, we fail to reject the null hypothesis that the there is statistically no significant difference between the frequencies observed in Mar 2019 - Mar 2023 follow the same distribution as the *Feb 2015 - Feb 2019* ones”

Was the rare event statistically different in 4 years?

india	monyear
3524	Apr 2020
5331	May 2020
5450	Mar 2020
5629	Jun 2021
5806	Feb 2022
5818	Jan 2022
5946	Jun 2020
5954	Jul 2020
6025	Dec 2021
6058	Apr 2021



What is the probability of observing something as extreme?
Null hypothesis?

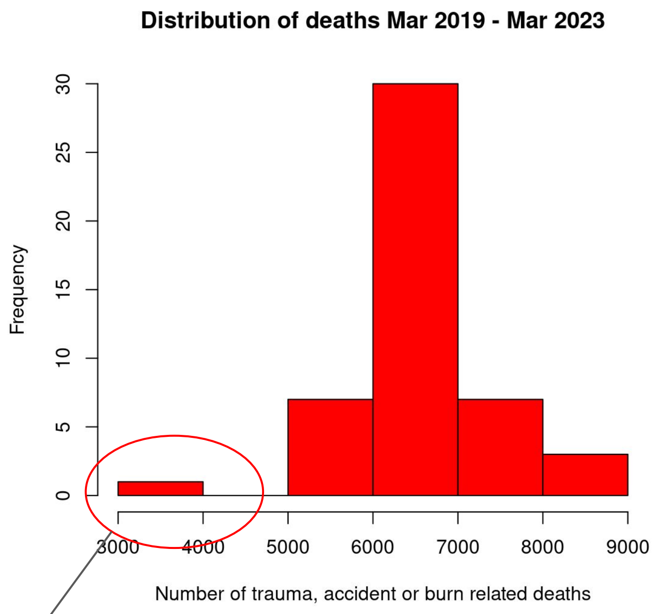
Was the rare event statistically different in 4 years?

What is the probability of observing entries as small as the one in April 2020?

Assume a poisson model

□ = (sum of observations)/length(of observations)

P(X ≤ 3524) = ppois(x = 3524, lambda) < 1e-16 → The rare event is statistically different



Is this event a “rare” event?

A simpler case: Are trauma related deaths in 2020 similarly distributed as 2019?

	month	2019	2020
1	Jan	7353	6535
2	Feb	6701	6604
3	Mar	7050	5450
4	Apr	7225	3524
5	May	8108	5331
6	Jun	8269	5946
7	Jul	7722	5954
8	Aug	7362	6066
9	Sep	8376	6579
10	Oct	7212	6488
11	Nov	6527	7174
12	Dec	6558	6852
	Sum	88463	72503

```
df_wide$diff <- df_wide$`2020`-df_wide$`2019`
df_wide$chisq <- df_wide$diff^2/(df_wide$`2019`)
chi_square_stat <- sum(df_wide$chisq)
dof <- 11
p_value <- pchisq(chi_square_stat, dof, lower.tail = FALSE)

alpha <- 0.05  # Significance level
if (p_value < alpha) {
  cat("Reject the null hypothesis")
} else {
  cat("Fail to reject the null hypothesis")
}
```

Ideally, we should check if (** this was automatically true for the 2015-2019 vs 2019 - 2023 example as we binned the observations)

$$\sum_i O_i = \sum_i E_i = N$$

A simpler case: Are trauma related deaths in 2020 similarly distributed as 2019?

O_i

Probability from 2019

month	2019	2020	p_i	E_i
Jan	7353	6535	0.08311950	6026.413
Feb	6701	6604	0.07574918	5492.043
Mar	7050	5450	0.07969434	5778.078
Apr	7225	3524	0.08167256	5921.506
May	8108	5331	0.09165414	6645.200
Jun	8269	5946	0.09347411	6777.153
Jul	7722	5954	0.08729073	6328.840
Aug	7362	6066	0.08322123	6033.789
Sep	8376	6579	0.09468365	6864.849
Oct	7212	6488	0.08152561	5910.851
Nov	6527	7174	0.07378226	5349.435
Dec	6558	6852	0.07413269	5374.842
Sum	88463	72503	1	72503

```
chisq <- chisq.test(x =
df_wide$`2020`, p =
df_wide$`2019`, rescale.p = T)

> chisq$statistic
X-squared
2738.136
> chisq$p.value
[1] 0
# Method 1
```

```
df_wide$p_i <- df_wide$`2019`/sum(df_wide$`2019`)
df_wide$E_i <- df_wide$p_i * sum(df_wide$`2020`)

chisq_square_stat <-
sum((df_wide$`2020`-df_wide$E_i)^2/df_wide$E_i)
dof <- 11

p_value <- pchisq(chi_square_stat, dof, lower.tail =
FALSE)

> chisq_square_stat
[1] 2738.136
> p_value
[1] 0
# Method 2
```

- Since the assumption of number of deaths in 2020 != number of deaths in 2019, we first calculate the relative probability of deaths in each month 2019 (p_i)
- p_i is then rescaled with total 2020 deaths to give E_i
- Use `chisq.test()` to test 2020 values against p_i or explicitly calculate chisquare

How is G-test related to chi-squared test?

How is G-test (Likelihood ratio test) related to Chi-squared?

Consider $G = 2\sum_i O_i log(\frac{O_i}{E_i})$ and let $O_i = E_i + \delta_i$

$$G = 2\sum_i O_i log(\frac{O_i}{E_i})$$
$$= 2\sum_i (E_i + \delta_i) log(1 + \frac{\delta_i}{E_i})$$

Using taylor expansion for x«1, $log(1+x) = x - x^2 + O(x^3)$
Thus,

$$G = 2\sum_i (E_i + \delta_i) log(1 + \frac{\delta_i}{E_i})$$
$$= 2\sum_i (E_i + \delta_i)(\frac{\delta_i}{E_i} - \frac{1}{2}(\frac{\delta_i}{E_i})^2 + O(\delta^3))$$
$$= 2\sum_i \delta_i + \frac{\delta^2}{E_i} + O(\delta_i^3)$$
$$\approx \frac{\delta^2}{E_i} \qquad \text{as } \sum_i \delta_i = 0$$
$$= \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Central Limit Theorem

Central limit theorem states that the sum or averages of iid random variables is distributed normally. Assuming $X_1, X_2, \ldots X_n$ are iid with mean μ and variance σ^2

$$\bar{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$\mathbb{E}[\bar{X_n}] = \mu$$

$$Var[\bar{X_n}] = \frac{\sigma^2}{n}$$

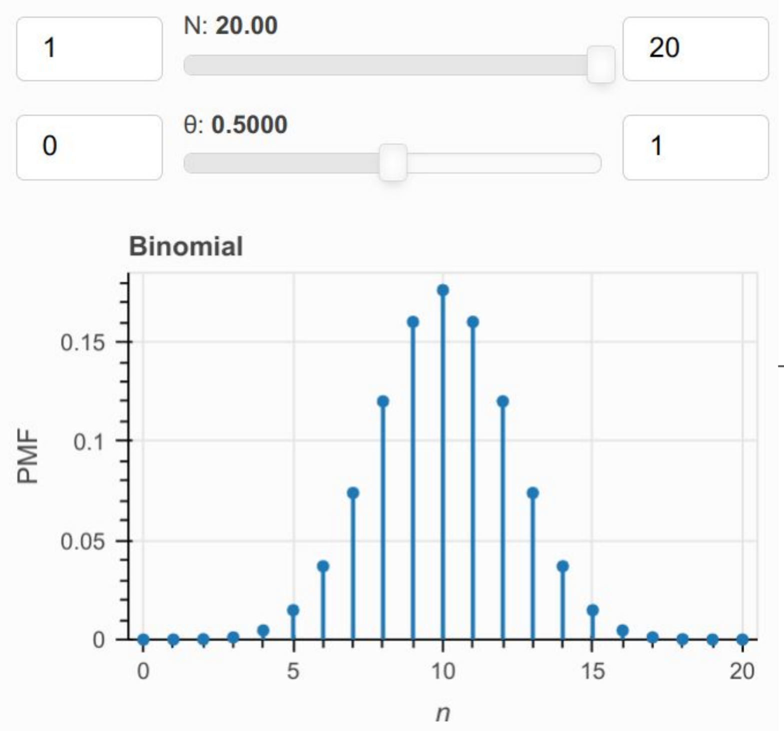
Thus,

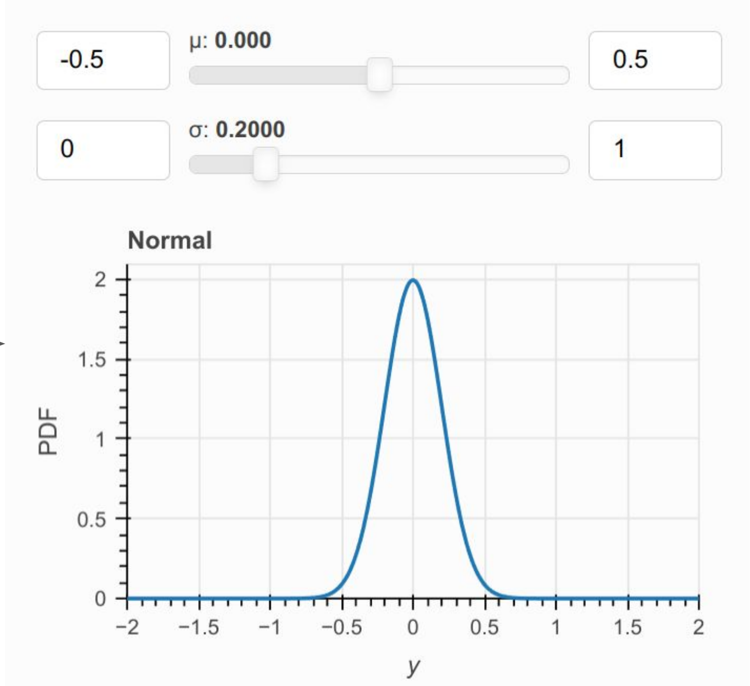
$$\mathbb{E}\left(\frac{\bar{X_n} - \mu}{\sigma/\sqrt{n}}\right) = 0, Var\left(\frac{\bar{X_n} - \mu}{\sigma/\sqrt{n}}\right) = 1$$

Then CLT states (for large n):

$$\frac{\bar{X_n} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Binomial to Normal?





$$f(n; N, \theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}.$$

Mean: $N\theta$

Variance: $N\theta(1-\theta)$

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(y-\mu)^2/2\sigma^2}.$$

Mean: μ

Variance: σ^2

Binomial to Normal

Assume $Y_1, Y_2, \ldots, Y_n$ are IID bernoulli with parameter p. By the CLT,

$$\sqrt{n}\left(\frac{\bar{Y}_n - p}{\sqrt{p(1-p)}}\right) \xrightarrow{\text{in distribution}} \mathcal{N}(0,1) \qquad \text{as } n \to \infty$$

Where $\bar{Y}_n$ is the sample mean and $E[\bar{Y}_n] = p$ and $Var(\bar{Y}_n) = \frac{\sigma_Y^2}{n} = \frac{p(1-p)}{n}$
Now,

$$\sqrt{n}\left(\frac{\bar{Y}_n - p}{\sqrt{p(1-p)}}\right) = \sqrt{n}\left(\frac{\frac{Y_1+\ldots+Y_n}{n} - p}{\sqrt{p(1-p)}}\right) = \sqrt{n}\left(\frac{Y_1 + \ldots + Y_n - np}{n\sqrt{p(1-p)}}\right) = \frac{Y_1 + \ldots + Y_n - np}{\sqrt{np(1-p)}}$$

But $Y_1 + \ldots + Y_n \sim Binomial(n,p)$ and so if we let $X = Y_1 + \ldots + Y_n$ then

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{\text{in distribution}} \mathcal{N}(0,1) \qquad \text{as } n \to \infty$$

And so as $n \to \infty$,

$$\frac{X - np}{\sqrt{np(1-p)}} \approx \mathcal{N}(0,1)$$

$$X \approx Z\sqrt{np(1-p)} + np \sim N(np, np(1-p))$$

So as n gets big, the Binomial RV,

$$X \dot{\sim} \mathcal{N}(np, np(1-p))$$

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,

$$\mathbb{E}[X_1 + X_2 + \cdots + X_N] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_N] \tag{1}$$

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.

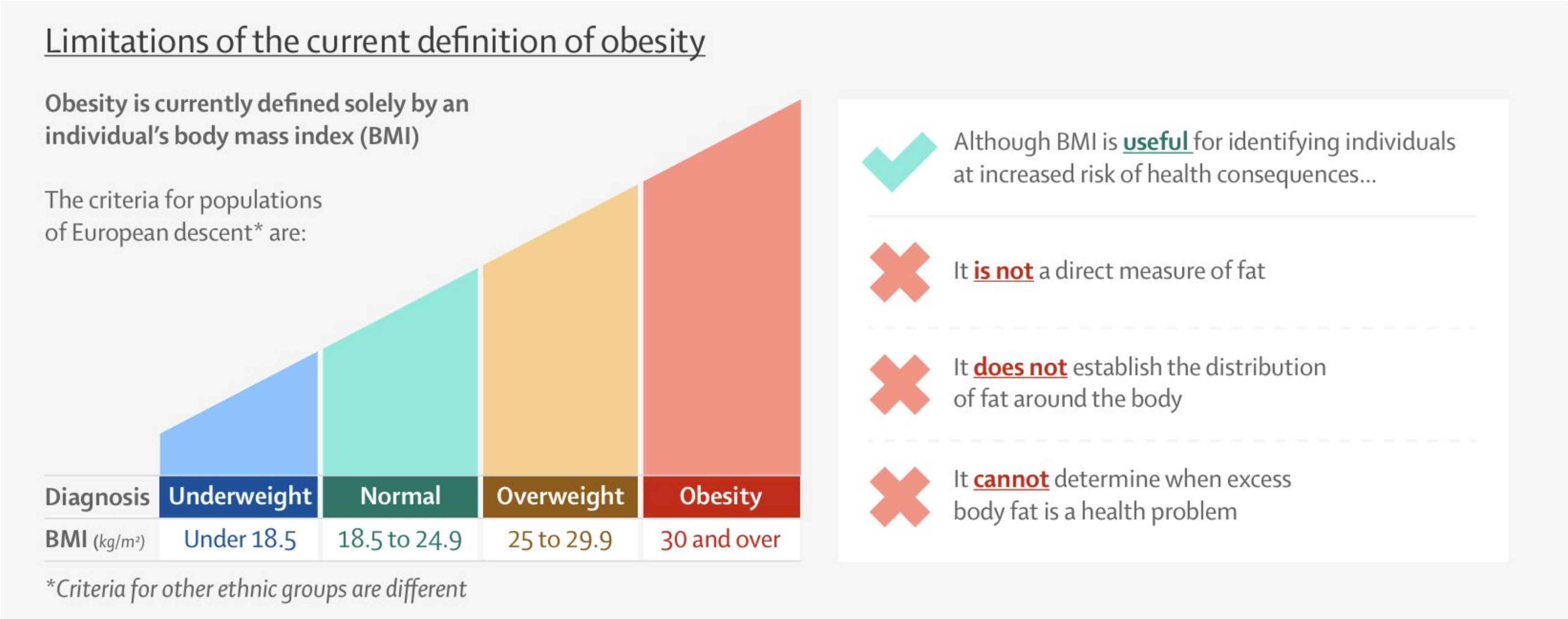
For independent variables, variation of their sum is equal to sum of individual variances, i.e. when X_i are independent,

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

Exercise - Calculate the mean of the binomial random variable

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.

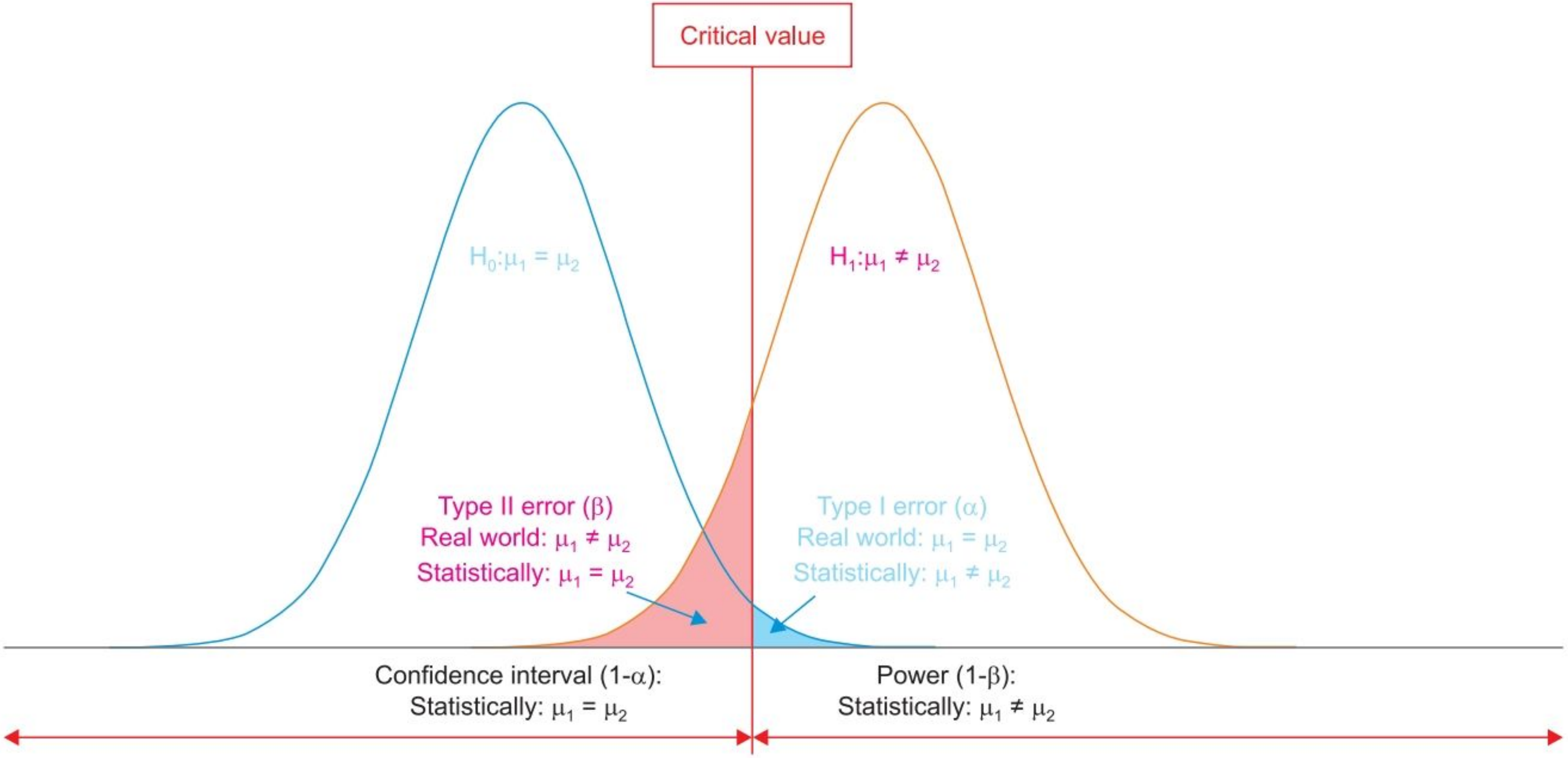
[Lancet]

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

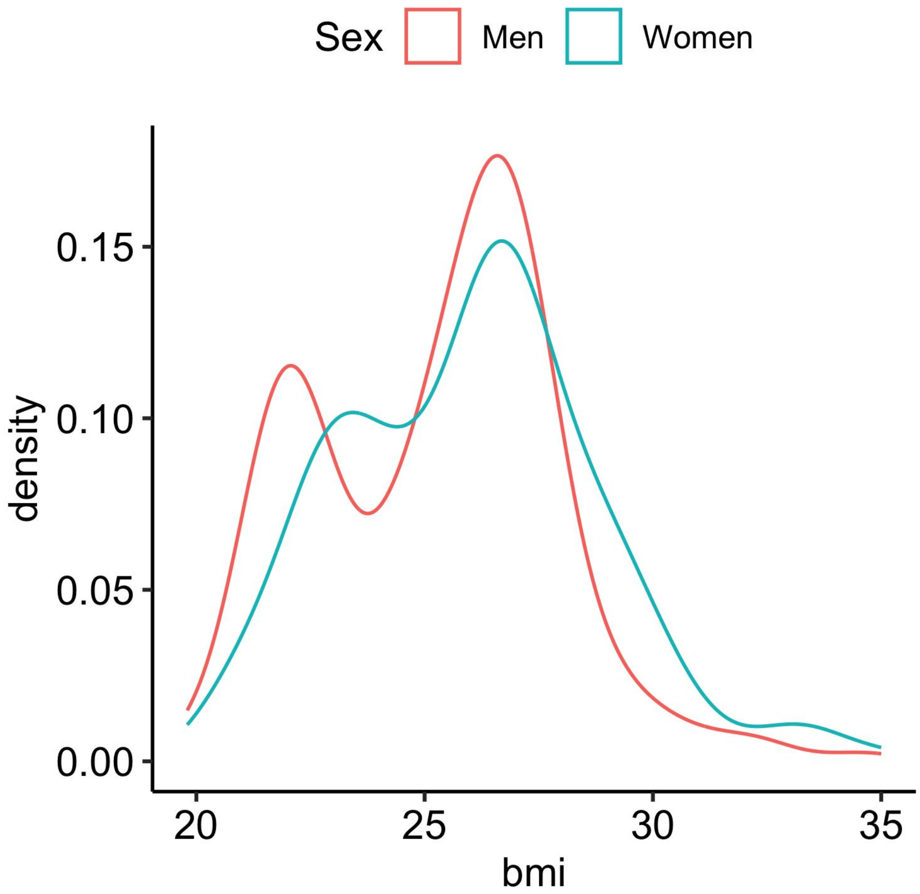
Lancet

Testing for difference in mean (median) of two samples



Source

Next: 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

Questions?