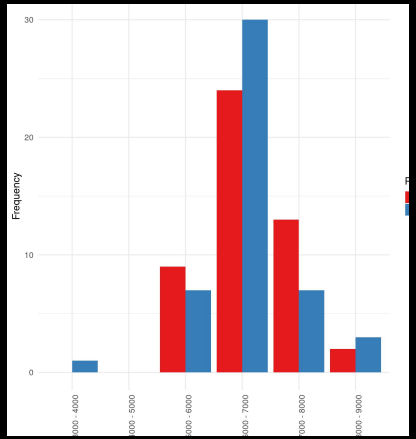
Model fitting and hypothesis testing



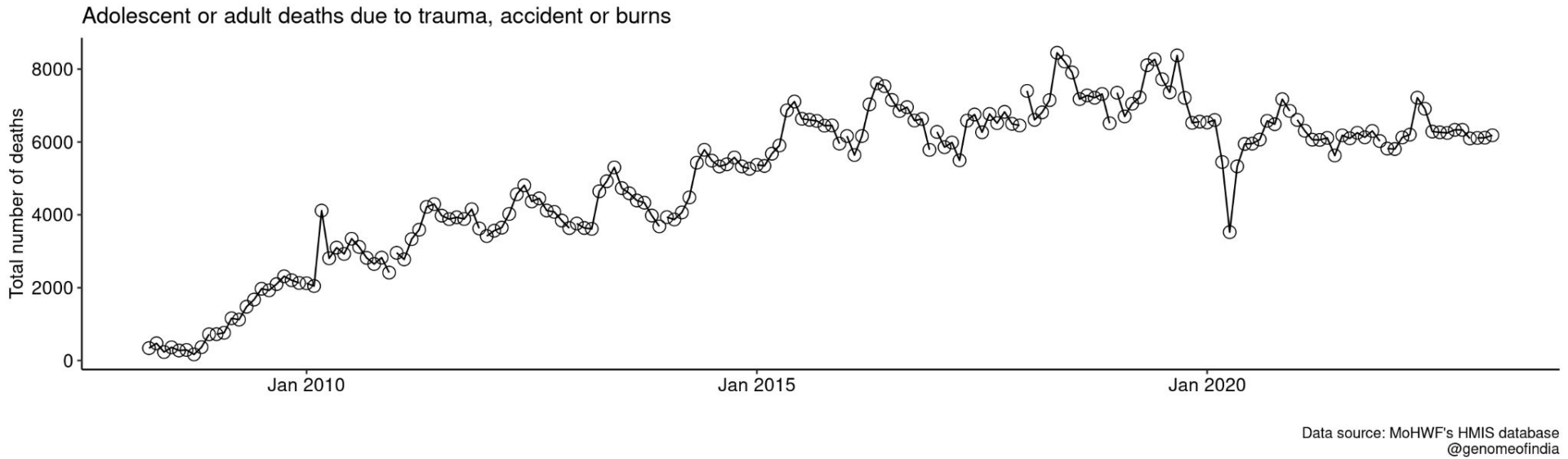
Saket Choudhary
saketc@iitb.ac.in

Introduction to Public Health Informatics
DH 302
Lecture 05 || Wednesday, 22nd January 2025

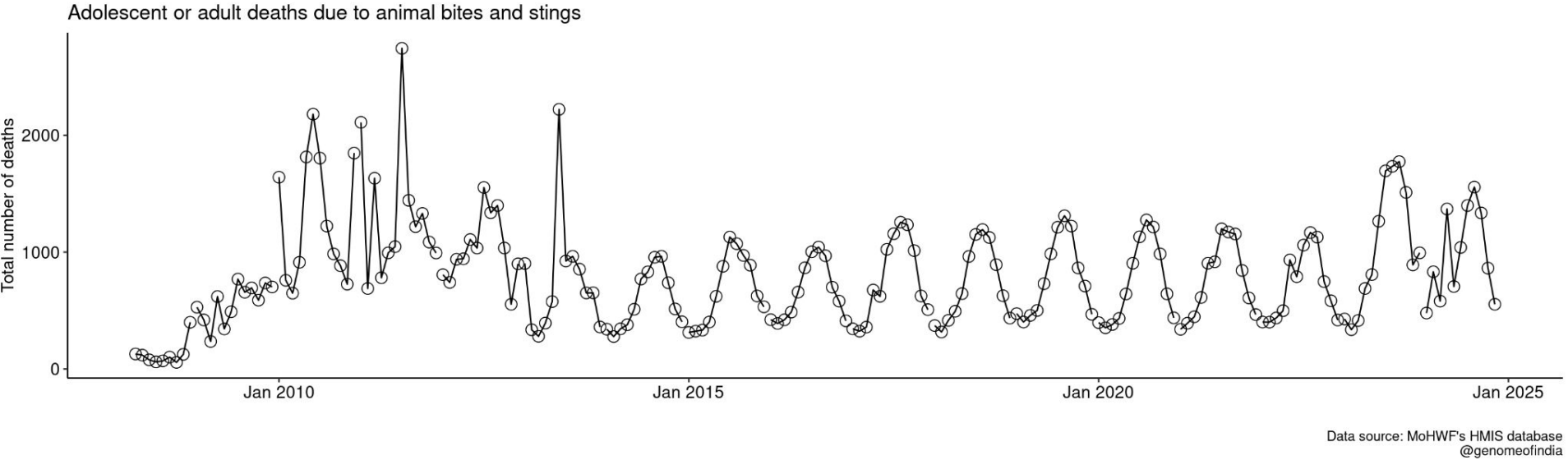
Goals for today

- How to decide what model best explains the observation
- Chi-squared test and G-test
- Expectations, Variances, CLT, Normal approximation

Question: What is going on this plot?



Question: What is going on this plot?



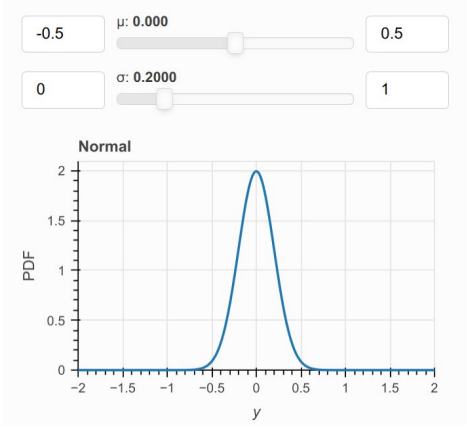
Some potential questions we would like to answer w.r.t the plots

- Are certain points an “**outlier**”?
- Has the trend of deaths **changed “significantly”** with time?

To be able to answer this question, we first need to answer: What is the best model that presents our data?

Even before we answer this question, we should learn about some (important) continuous distributions

Some (important) continuous distributions

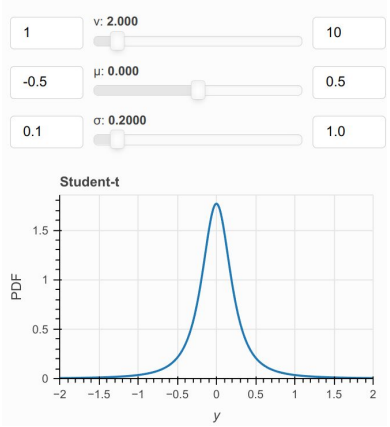


Normal

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

Mean: μ

Variance: σ^2

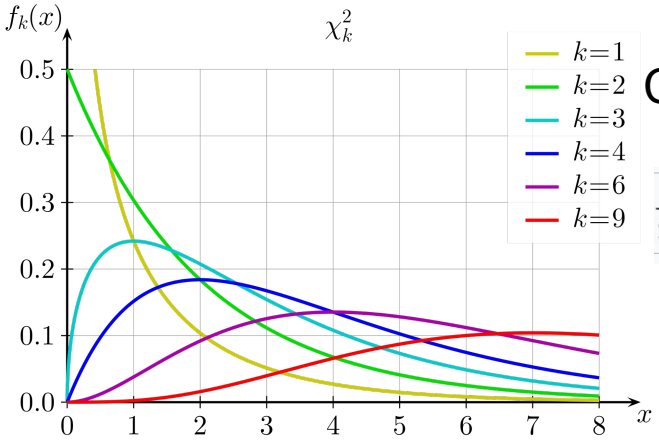


Student's t

$$f(y;\nu,\mu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left(1+\frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

Mean: μ for $\nu > 1$, otherwise undefined.

Variance: $\frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$. If $1 < \nu < 2$, then the variance is infinite. If $\nu \le 1$, the variance is undefined.

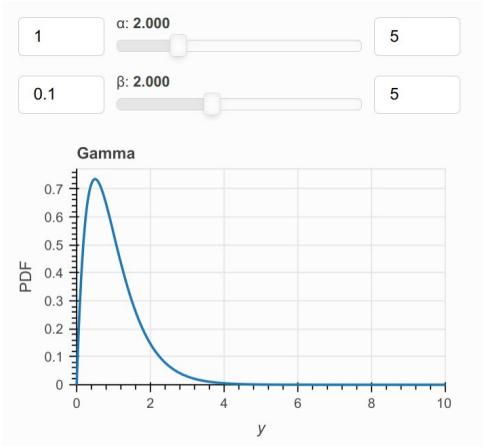


Chi-square

$$\frac{1}{2^{k/2}\Gamma(k/2)}\,x^{k/2-1}e^{-x/2}$$

Mean = k

Variance = 2k



Gamma

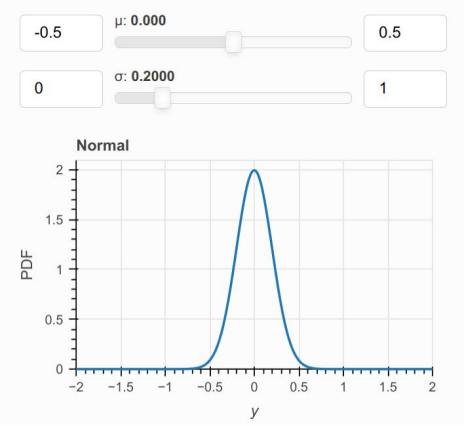
$$f(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\,\frac{(\beta y)^\alpha}{y}\,\mathrm{e}^{-\beta y},$$

Mean: $\frac{\alpha}{\beta}$

Variance: $\frac{\alpha}{\beta^2}$

https://distribution-explorer.github.io/

Gaussian a.k.a Normal distribution



Normal

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

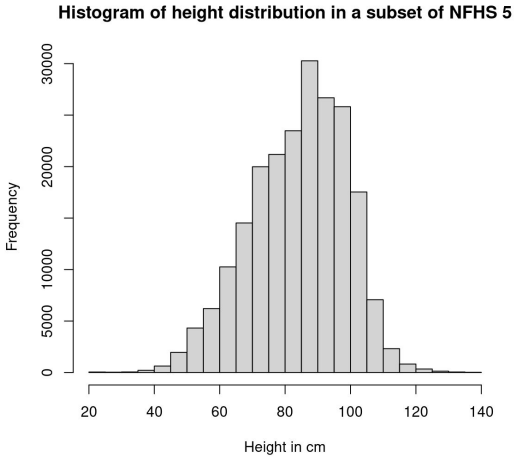
Mean: μ

Variance: σ^2

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

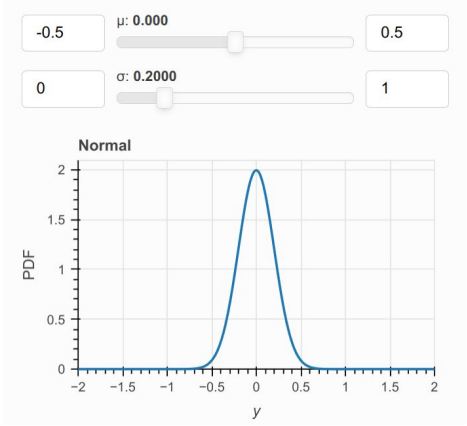
Quantities that are sum of large number of subprocesses tend to be normally distributed

Example: Height/blood pressure distribution of a sample



NFHS5

What is normal about normal?



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

Mean: μ

Variance: σ^2

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

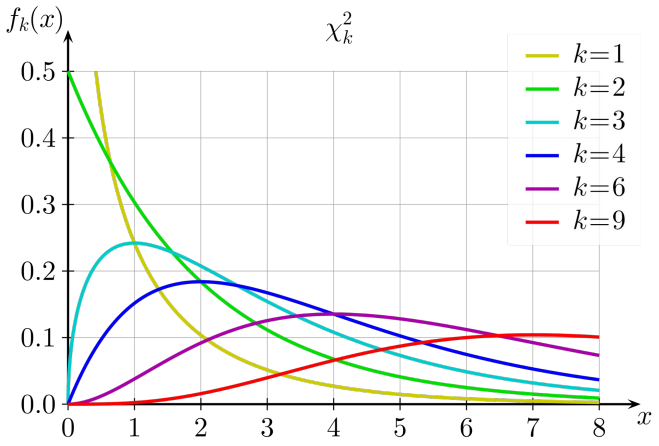
Standard normal has mean 0, variance 1

"The literature gives conflicting evidence about the origin of the term Normal distribution". Karl Pearson (1920) claimed to have introduced it "many years ago", in order to avoid an old dispute over priority between Gauss and Legendre; but he gave no reference. Hilary Seal (1967) attributes it instead to Galton; but again fails to give a reference, so it would require a new historical study to decide this. However, the term had long been associated with the general topic: given a linear model y = Xβ+e where the vector y and the matrix X are known, the vector of parameters and the noise vector e unknown, Gauss (1823) called the system of equations which give the least squares parameter estimates , "the normal equa$X'X\hat{\beta} = X'y$*, ellipsoid of constant probability density was called the "normal surface." It appears that somehow the name got transferred from the equations to the sampling distribution that leads to those equations"*

Source

Chi-square distribution

Given $Z_1, Z_2, \ldots, Z_k$ are independent standard normal distribution , i.e. $Z_i \sim \mathcal{N}(0,1)$, then the sum of their squares follows a χ^2 distribution with k degrees of freedom



Mean = k
Variance = 2k

$$X = \sum_{i=1}^{k} Z_i^2$$

$$X \sim \chi^2_k$$

$$f(x;\, k) = \begin{cases} \dfrac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, \mathrm{d}t, \qquad \Re(z) > 0\,.$$

Example: Estimate the parameters by curve fitting and check how "good" does it explain the observations

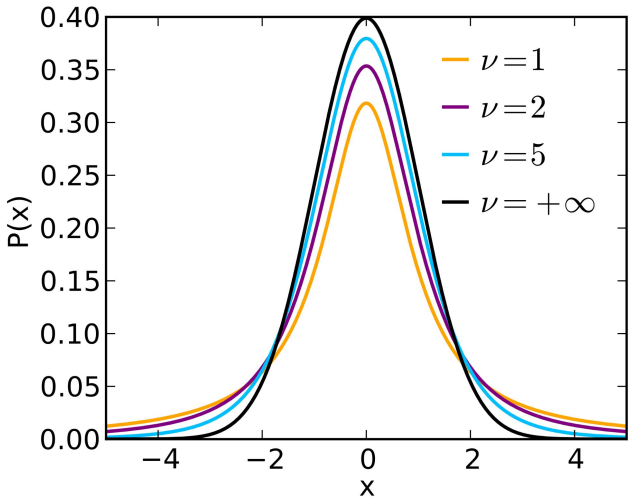
$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$

O_i = an observed count for bin *i*

E_i = an expected count for bin *i*

Student's t

Student's t



$$f(y;\nu,\mu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}}\left(1+\frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

Mean: μ for $\nu > 1$, otherwise undefined.

Variance: $\frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$. If $1 < \nu < 2$, then the variance is infinite. If $\nu \leq 1$, the variance is undefined.

Gaussian like distribution with "heavier tails"
As v → ∞ becomes a "Gaussian"
v = 1 becomes a "cauchy" distribution

Often used as a "statistic" to compare the difference in mean of two populations (samples)

Why the name "Student"?

Student

Why student's t?

William Sealy Gosset

Why the name "Student"?

- Student = William Sealey Gosset used to work for Guinness Brewery in Dublin
- Gossett wrote a paper describing the "t-test" → Used "Student" pseudonym probably following company mandate of not using public names for papers or possibly because Guinness did not want competitors to figure out that they were using t-test to test the quality of barley based on chemical properties of samples

Student 1908

T-test primer

X_1, ..., X_n are independent realizations of the normally-distributed random variable X

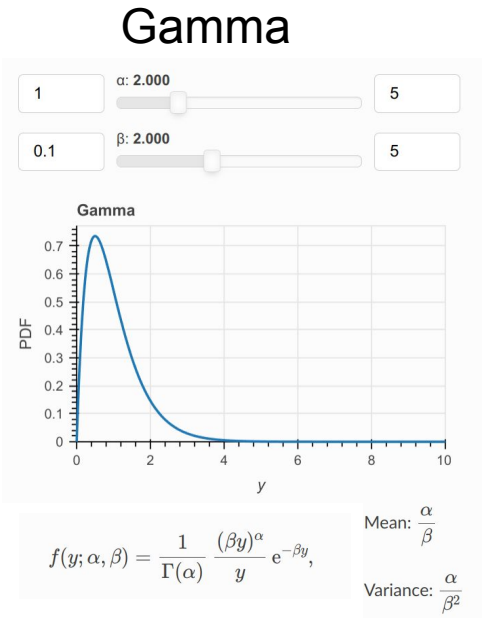
Sample mean $$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

Sample variance $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

T follows student t-distribution $$T \equiv \left(\overline{X}_n - \mu\right)\frac{\sqrt{n}}{s}$$

Student

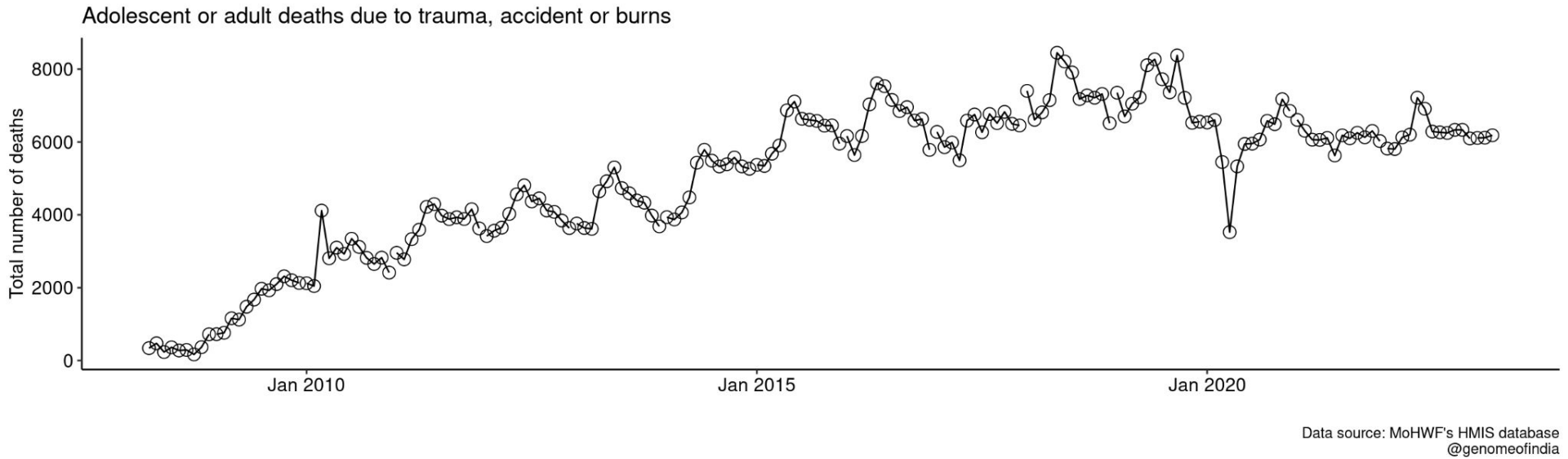
Gamma, the ‘versatile’ distribution



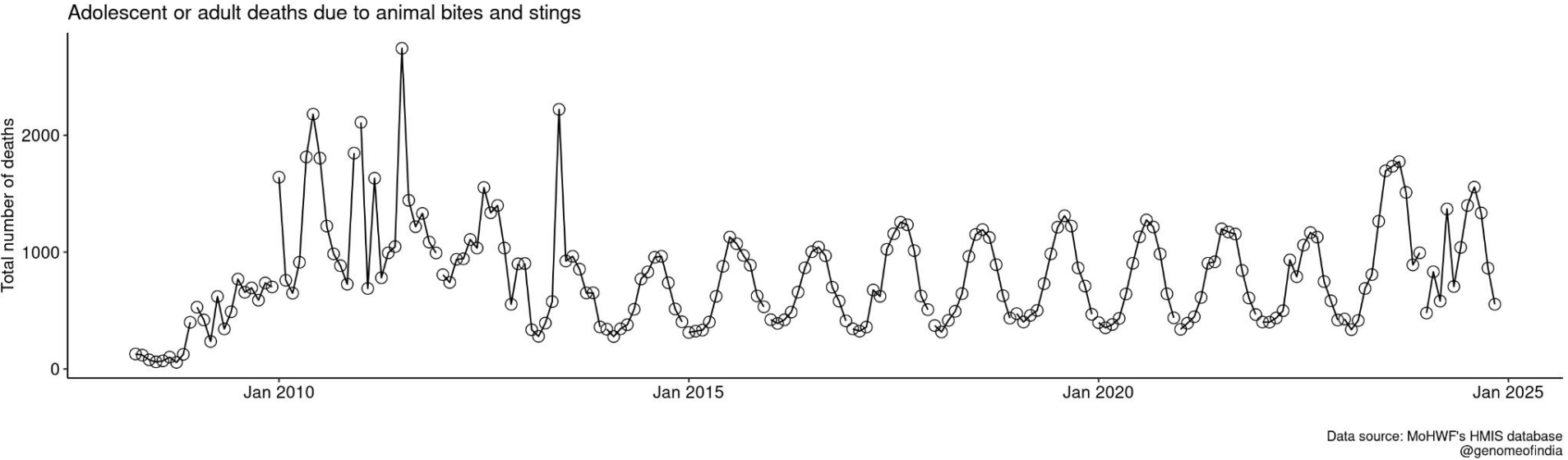
- Versatile two parameter distribution
- Often used to model multistep processes where each step as the same rate
- Example: Waiting time till a system needs to be repaired or cell-division events or to model number of insurance claims, age distribution of cancer events

https://distribution-explorer.github.io/

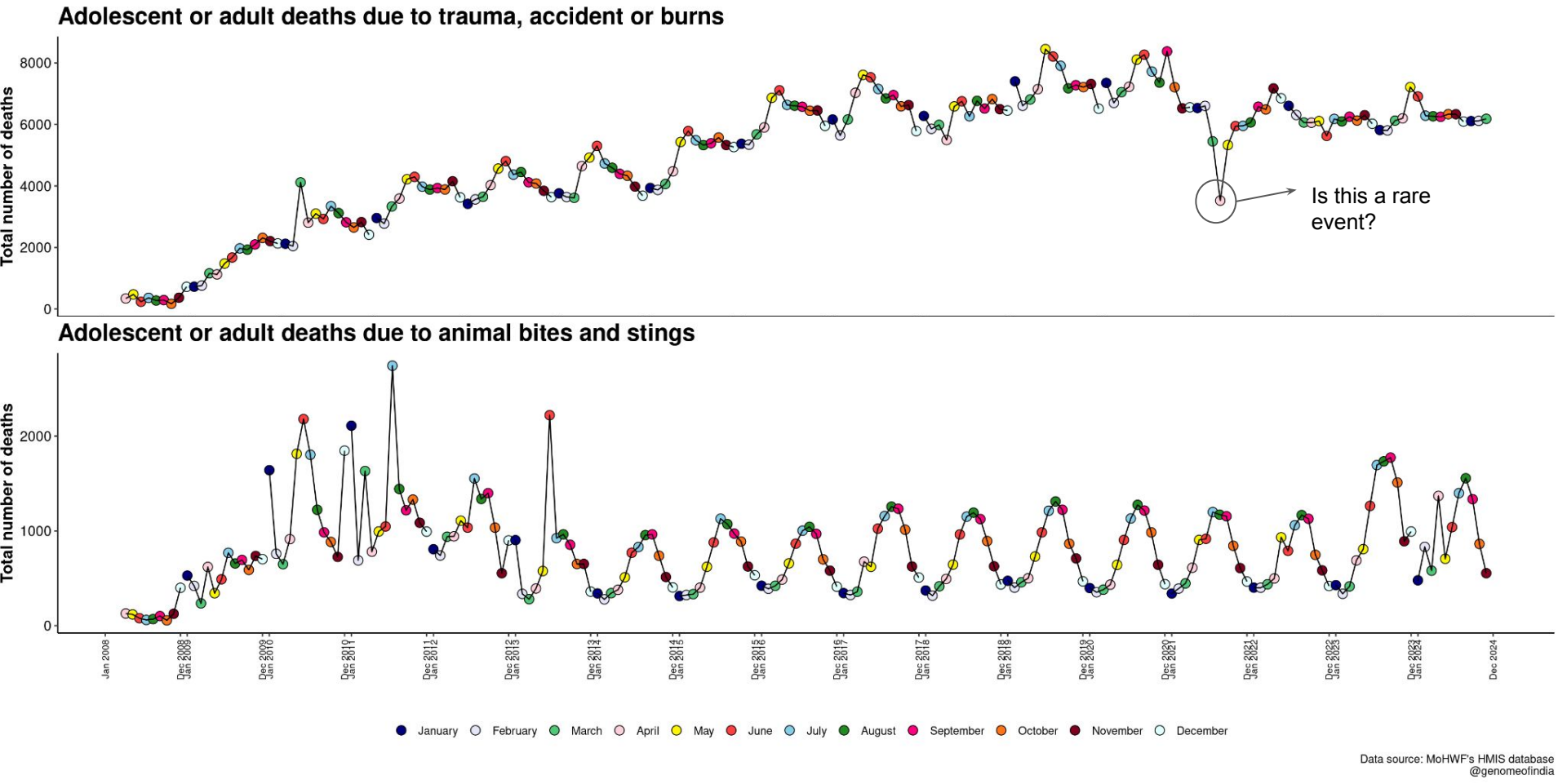
Question: What is going on this plot?



Back to the question: What is going on this plot?

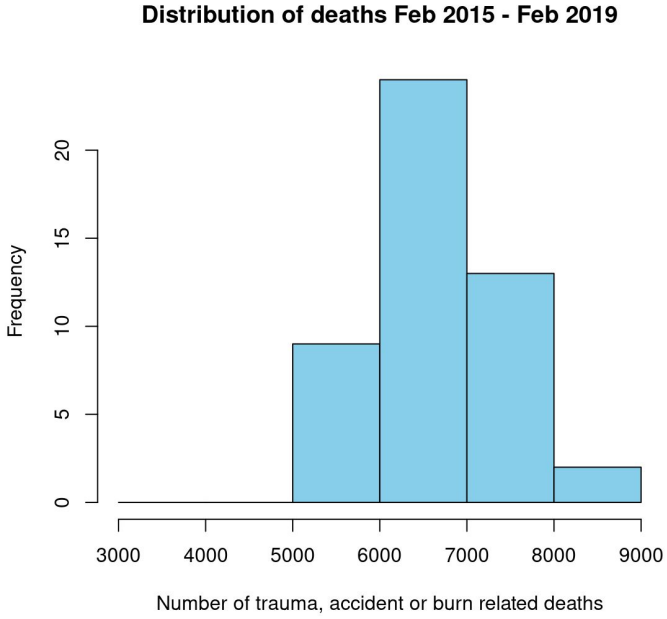


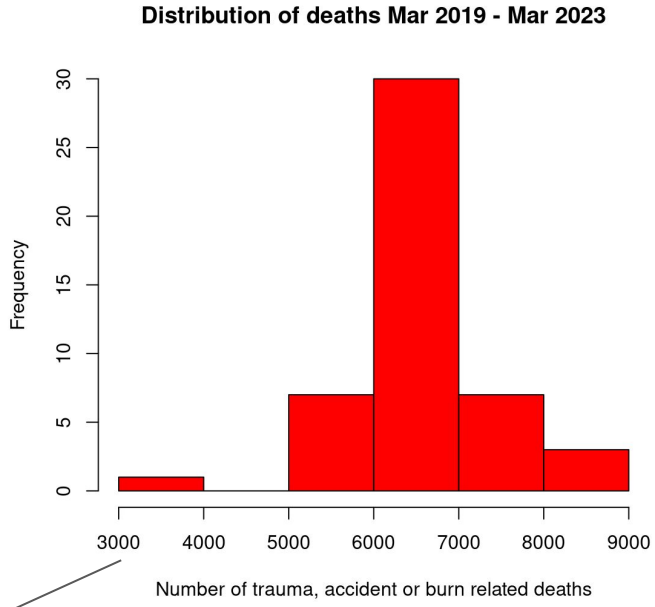
Trauma and bite related deaths are seasonal



Data source: MoHWF's HMIS database
@genomeofindia

Does the historical model fit the latest observations?

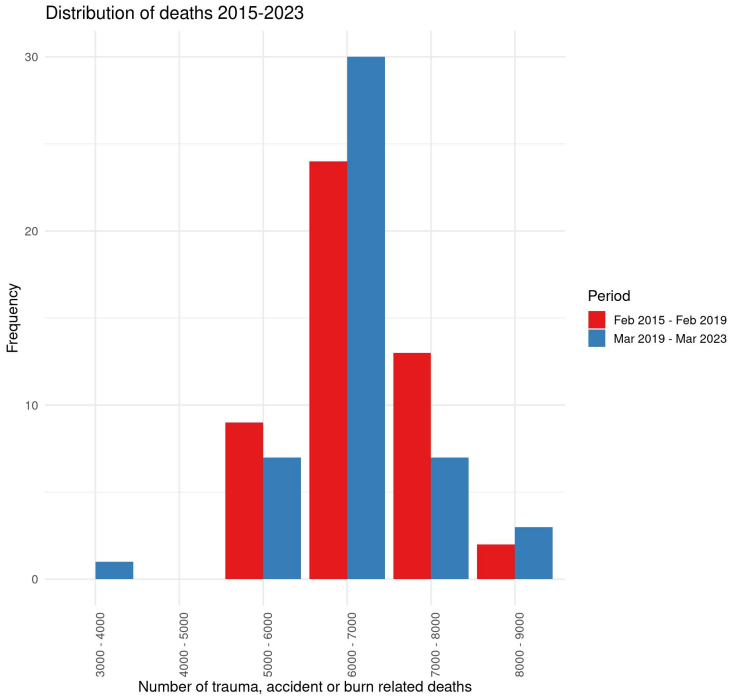




Does the Feb 2015 - Feb 2019 model explain the observations from Mar 2019 - Mar 2023?

Goodness of fit - Chi-squared test

Problem: What distribution should I fit?



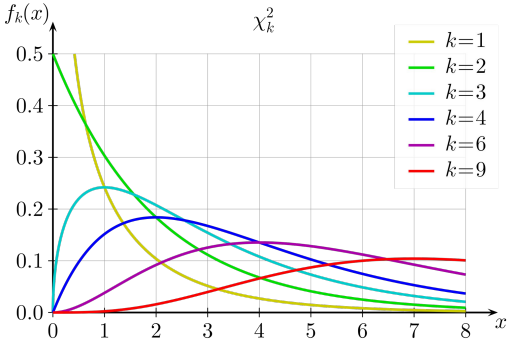
Solution: Quantify how "good" does the expected model (frequencies) explain the observations

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

O_i = an observed count for bin *i*

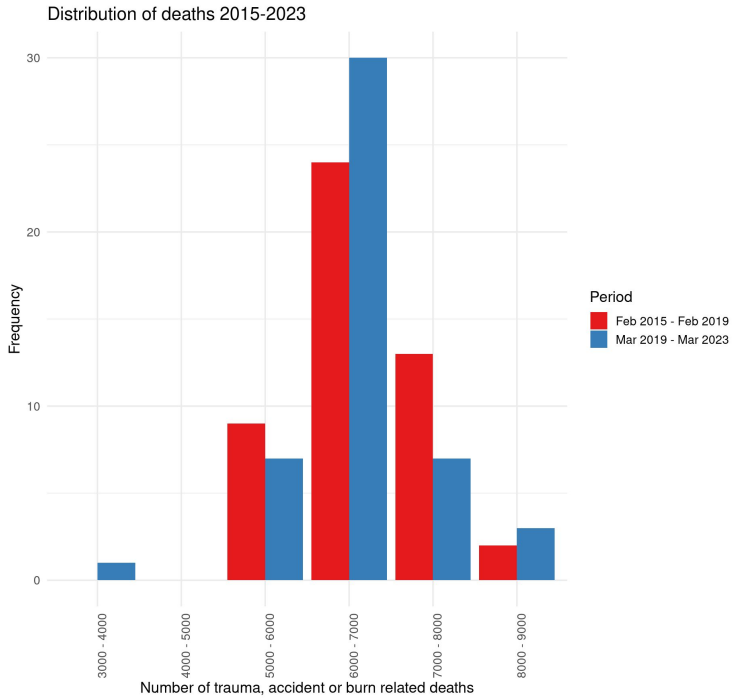
E_i = an expected count for bin *i*, asserted by the null hypothesis

Calculate p-value



Goodness of fit - Chi-squared test

Problem: What distribution should I fit?

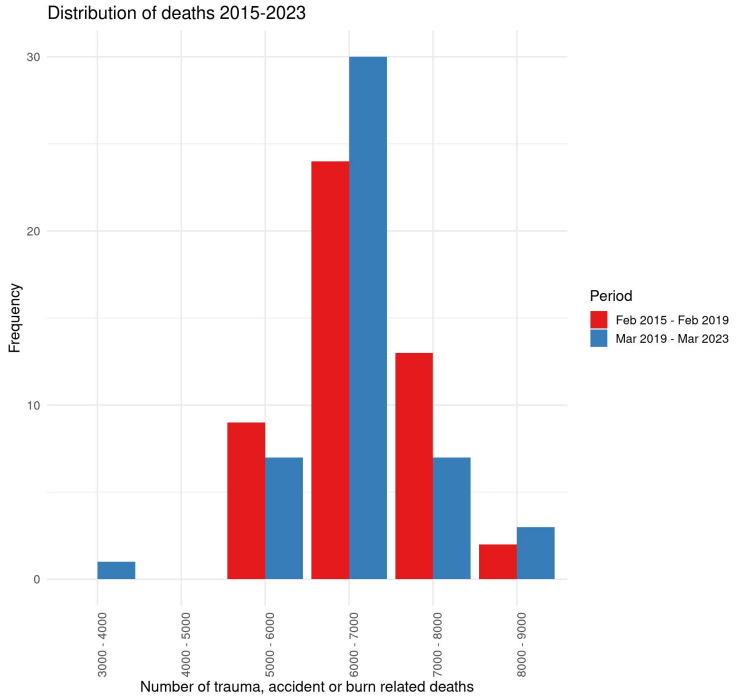


bin	Feb 2015 – Feb 2019	Mar 2019 – Mar 2023
3000 – 4000	0	1
4000 – 5000	0	0
5000 – 6000	9	7
6000 – 7000	24	30
7000 – 8000	13	7
8000 – 9000	2	3

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

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?

Hypothesis testing

- In hypothesis testing, we investigate if the observations could have come by chance or if random processes are sufficient to explain the observations.
- Null hypothesis H_0: Hypothesis that chance alone is responsible for explaining the observations
- Example: The observations in Mar 2019 - Mar 2023 can be explained by the observations in Feb 2015 - Feb 2019 alone.
- Requires constructing a **statistical model** of what the observations would like if chance or random process can alone explain it
- A **“test-statistic”** measures the deviation of observations from the expectations (i.e. the null hypothesis)
- We measure if the test-statistic deviates from an pre-decided threshold. If it does → We “reject” the null hypothesis, if it does not we “fail to reject” the null hypothesis

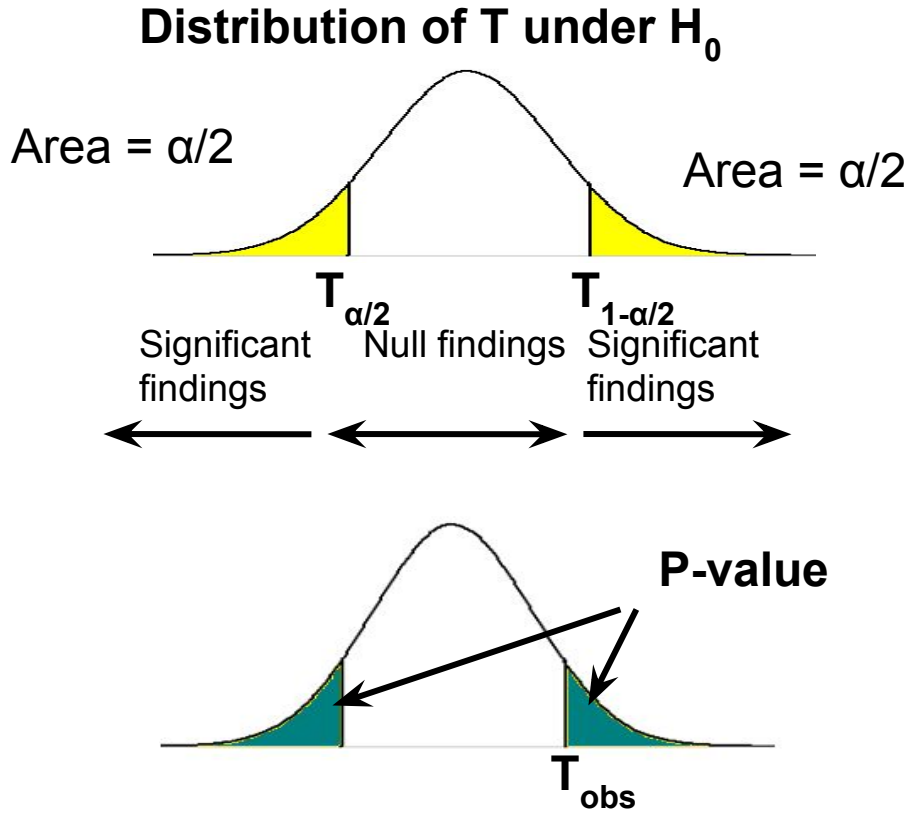
How to evaluate whether the prior model explains the observations?

- We don't know the truth, so we start with a "**null hypothesis**": "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"
- **Summarize the discrepancy** in *expected* to observed values (using a chi-squared test)
- **Select a testing threshold** (α) - the probability threshold below which the null hypothesis can be rejected [this is the extreme ends of your distribution]
- **Compute the test statistic** and reject the null hypothesis if the test-statisic is in the extremes (probability of observing test statistic or extreme is < α)
- We never "accept" the hypothesis: **we find evidence against it probabilistically**

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 (**α**)

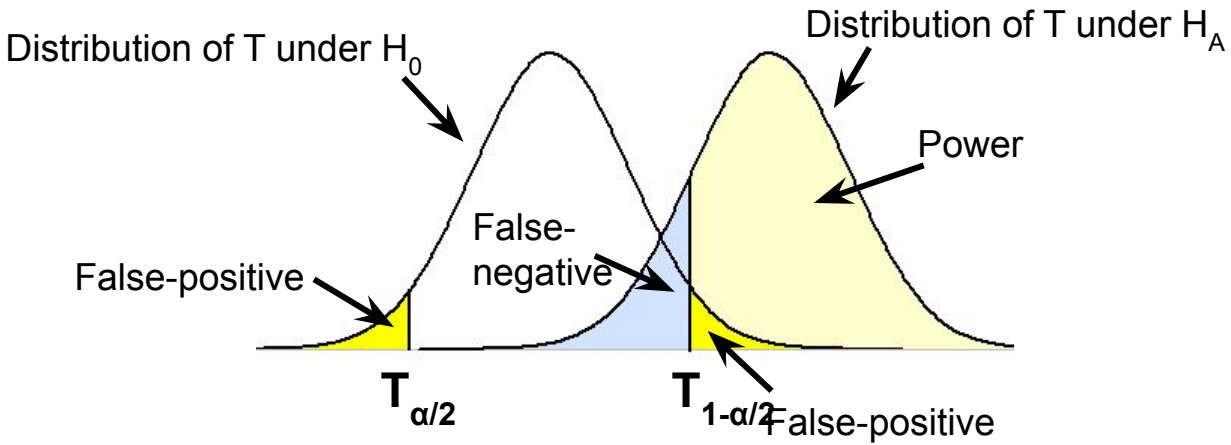


Type I,II errors and Power

- **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.

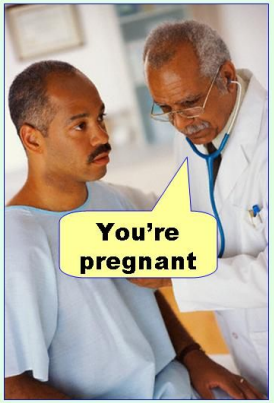


Types of error

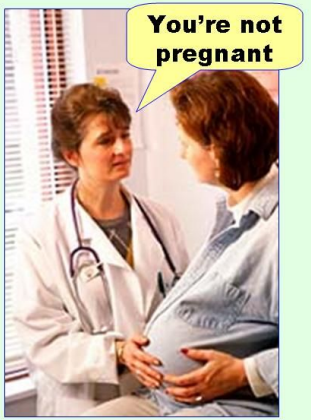
Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference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

Source

Type I error
(false positive)



Type II error
(false negative)



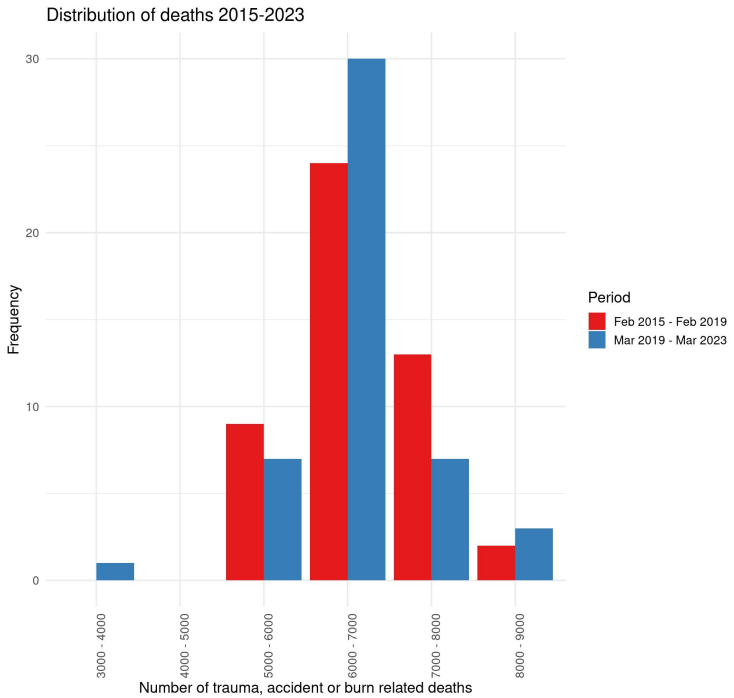
Paul Ellis, 2010

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.

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")

}
```



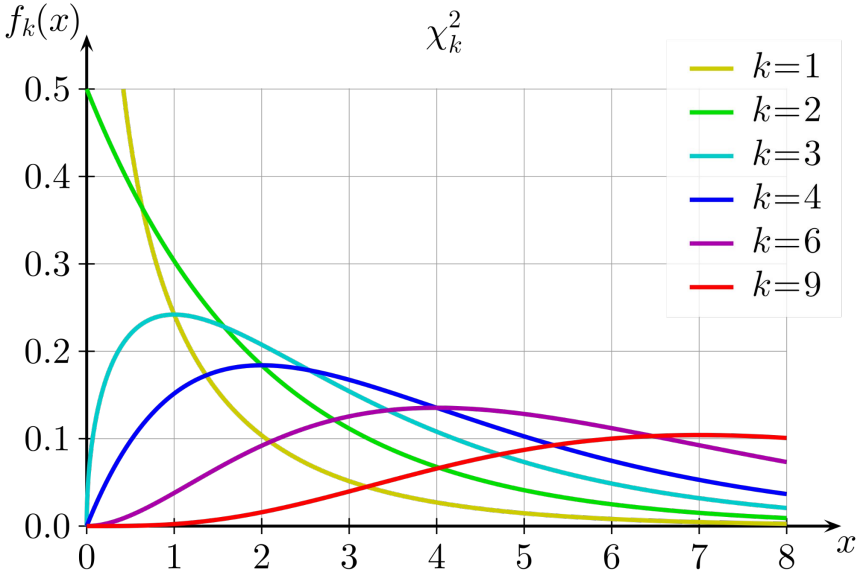
P-value = 0.33 (>0.05)

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)

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

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



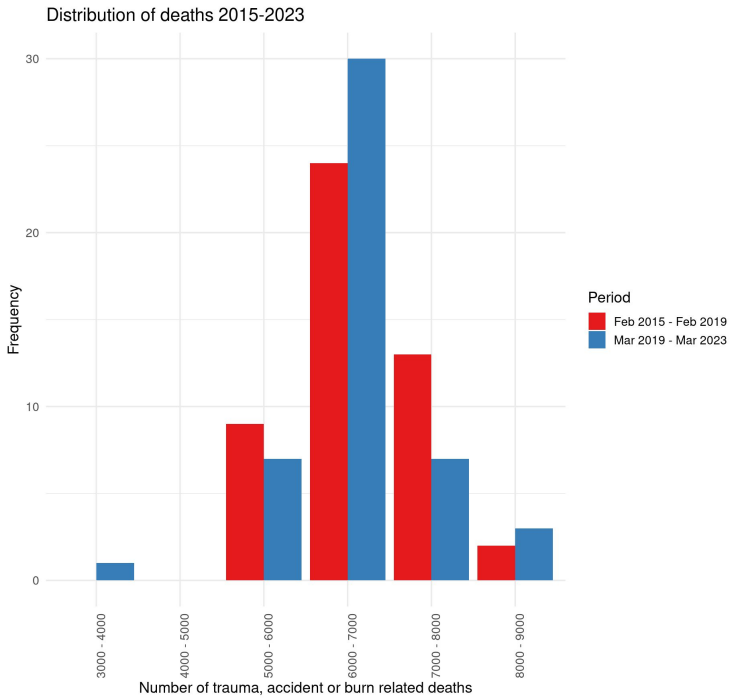
O_i = an observed count for bin i

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”

Questions?