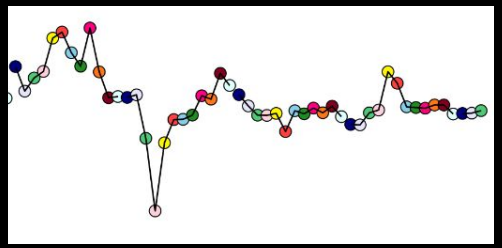
Statistical Modelling - 3



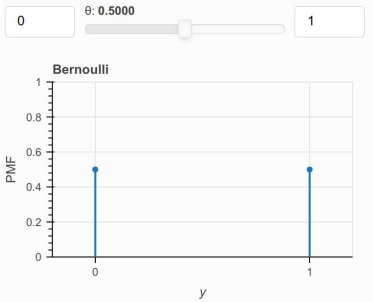
Saket Choudhary
saketc@iitb.ac.in
Introduction to Public Health Informatics
DH 302
Lecture 04 || Friday, 17th January 2025

Goals for today

- How to assess the observations for “extremeness” under a given model
- “Visualizing” likelihood, maximum likelihood and model selection
- How to decide what model best explains the observation

R demo

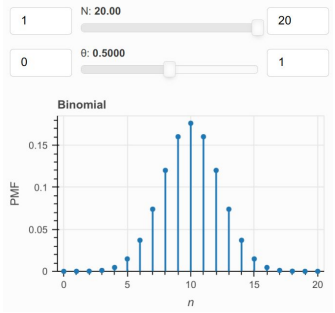
Previously



$$f(y;\theta) = \begin{cases} 1-\theta & y = 0 \\ \theta & y = 1. \end{cases}$$

Mean: θ

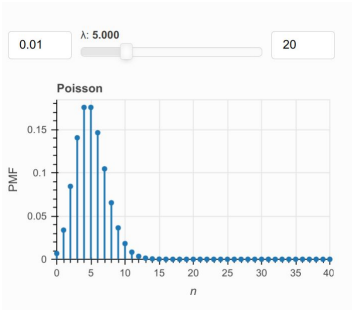
Variance: $\theta(1-\theta)$



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

Mean: $N\theta$

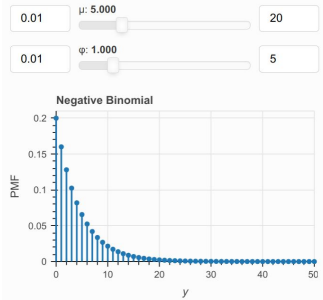
Variance: $N\theta(1-\theta)$



$$f(n;\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda}.$$

Mean: λ

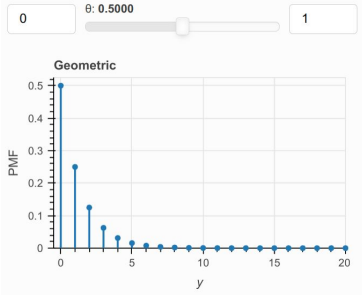
Variance: λ



$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(\phi)\,y!}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}.$$

Mean: μ

Variance: $\mu\left(1+\frac{\mu}{\phi}\right)$.



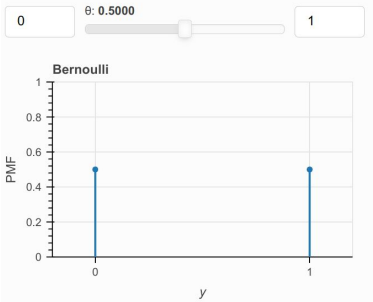
$$f(y;\theta) = (1-\theta)^y\,\theta.$$

Mean: $\frac{1-\theta}{\theta}$

Variance: $\frac{1-\theta}{\theta^2}$

- What are the parameters of a distribution?
  - “Knobs” of a distribution - center and spread
- What is likelihood function?
  - Probability of seeing the observations under a model
- How are the maximum likely parameters learned?
  - Maximum likelihood estimation, we maximize log of likelihood function

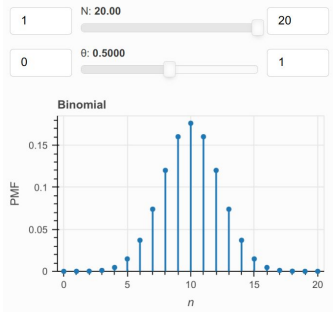
Previously



$$f(y;\theta) = \begin{cases} 1-\theta & y = 0 \\ \theta & y = 1. \end{cases}$$

Mean: θ

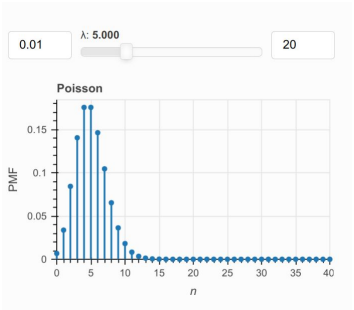
Variance: $\theta(1-\theta)$



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

Mean: $N\theta$

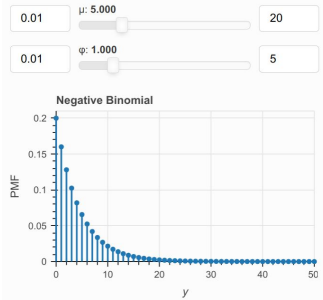
Variance: $N\theta(1-\theta)$



$$f(n;\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda}.$$

Mean: λ

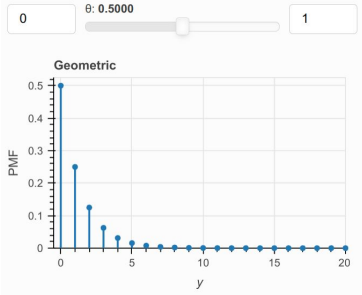
Variance: λ



$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(\phi)\,y!}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}.$$

Mean: μ

Variance: $\mu\left(1+\frac{\mu}{\phi}\right)$.



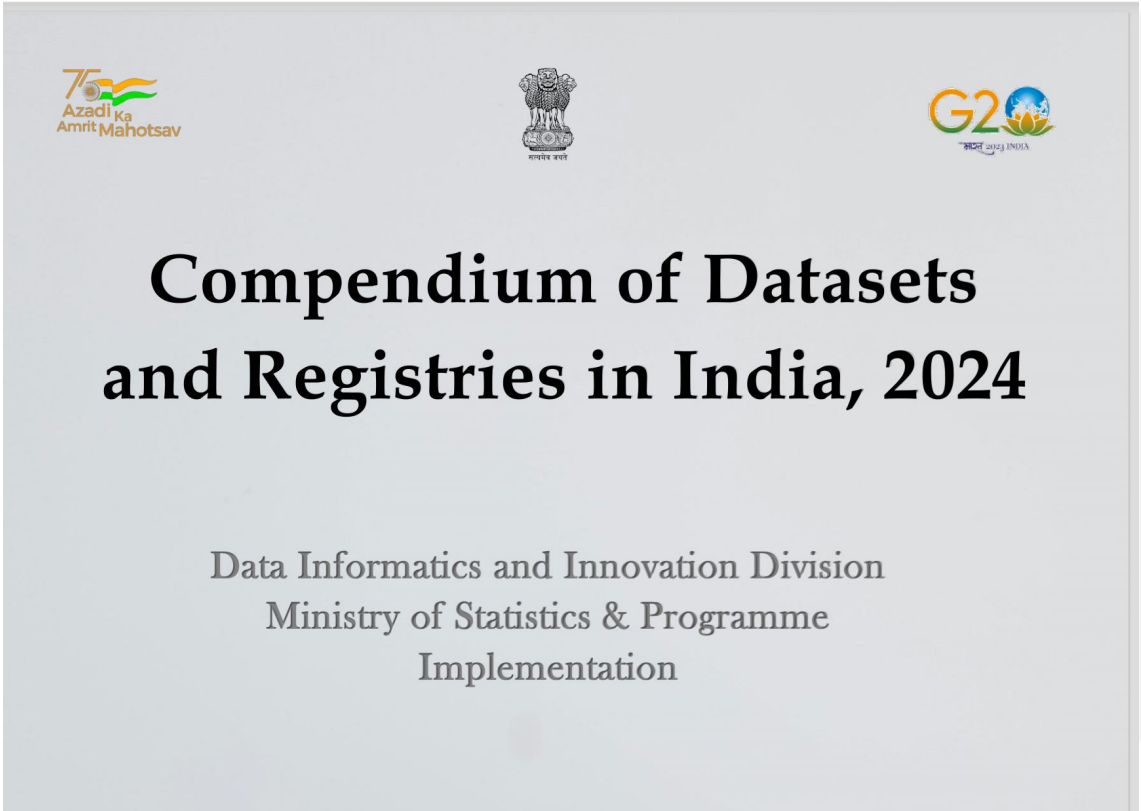
$$f(y;\theta) = (1-\theta)^y\,\theta.$$

Mean: $\frac{1-\theta}{\theta}$

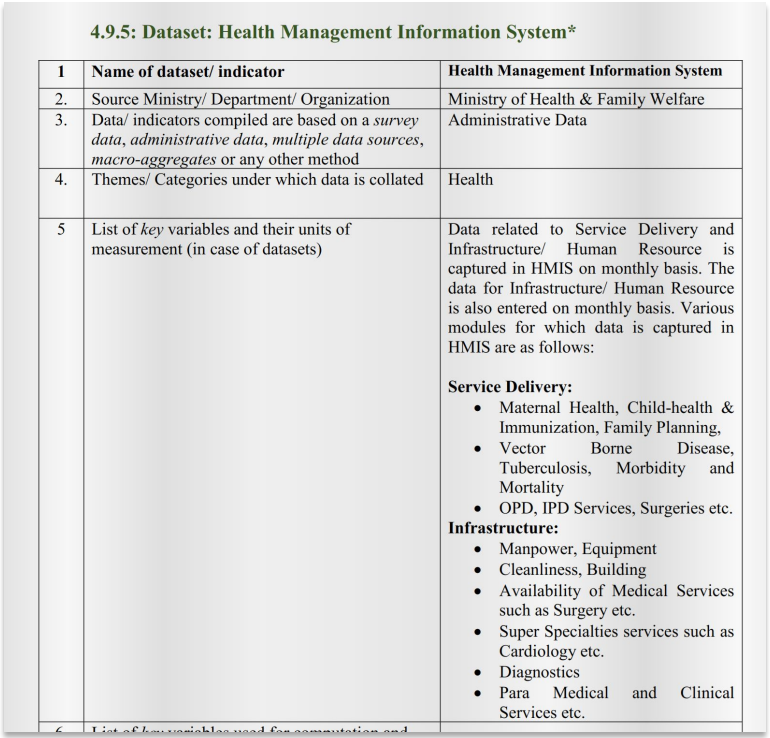
Variance: $\frac{1-\theta}{\theta^2}$

- What are the parameters of a distribution?
  - “Knobs” of a distribution - center and spread
- What is likelihood function?
  - Probability of seeing the observations under a model
- How are the maximum likely parameters learned?
  - Maximum likelihood estimation, we maximize log of likelihood function

Compendium of Indian Data



https://www.mospi.gov.in/sites/default/files/publication_reports/Compendium_of_Datasets_and_Registries_in_India_2024.pdf#page=332.03

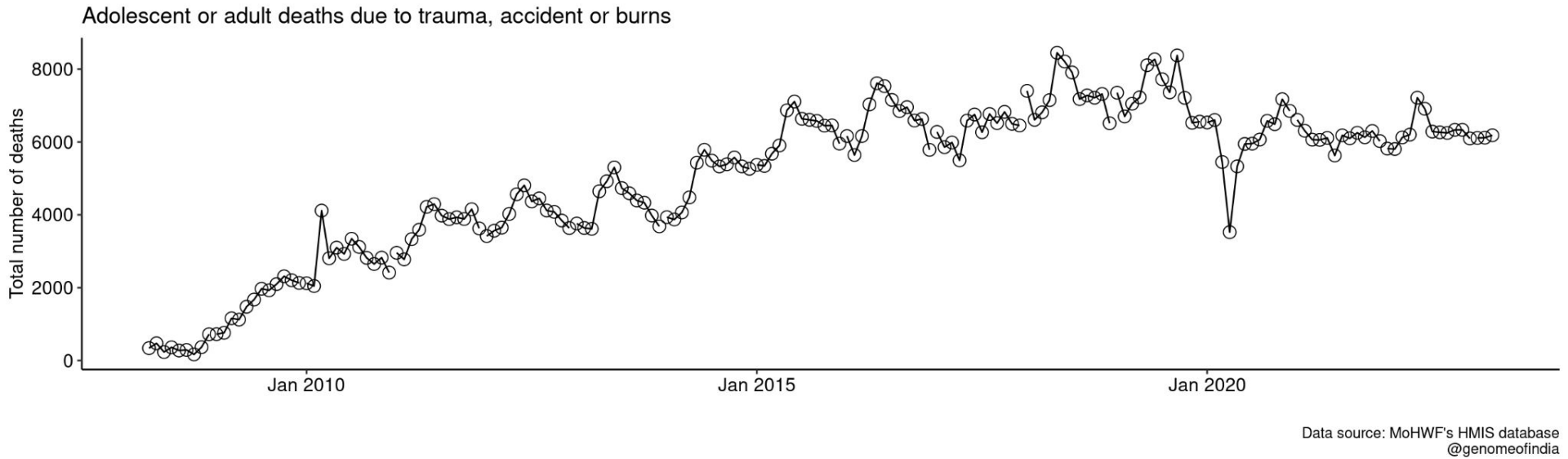


4.9.5: Dataset: Health Management Information System*

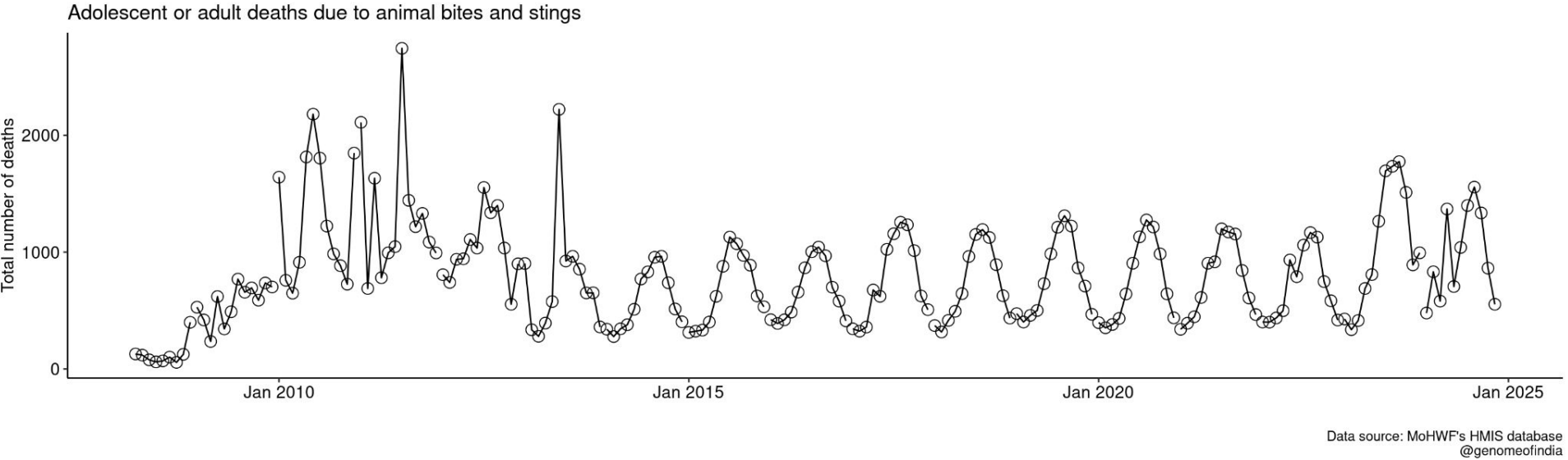
1	Name of dataset/ indicator	Health Management Information System
2.	Source Ministry/ Department/ Organization	Ministry of Health & Family Welfare
3.	Data/ indicators compiled are based on a *survey data*, *administrative data*, *multiple data sources*, *macro-aggregates* or any other method	Administrative Data
4.	Themes/ Categories under which data is collated	Health
5	List of *key* variables and their units of measurement (in case of datasets)	Data related to Service Delivery and Infrastructure/ Human Resource is captured in HMIS on monthly basis. The data for Infrastructure/ Human Resource is also entered on monthly basis. Various modules for which data is captured in HMIS are as follows: **Service Delivery:** • Maternal Health, Child-health & Immunization, Family Planning, • Vector Borne Disease, Tuberculosis, Morbidity and Mortality • OPD, IPD Services, Surgeries etc. **Infrastructure:** • Manpower, Equipment • Cleanliness, Building • Availability of Medical Services such as Surgery etc. • Super Specialties services such as Cardiology etc. • Diagnostics • Para Medical and Clinical Services etc.

https://hmis.mohfw.gov.in/#!/

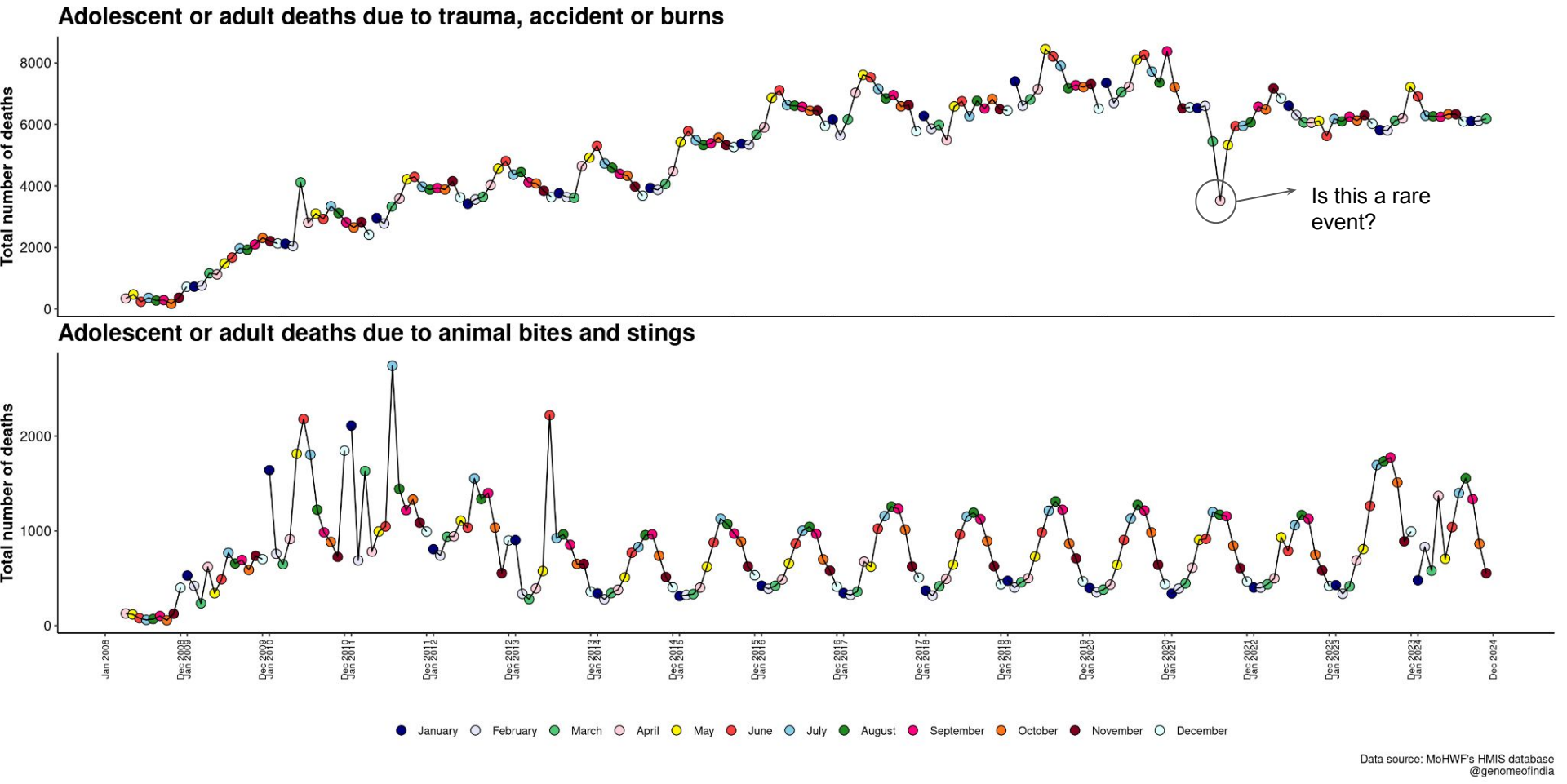
Question: What is going on this plot?



Question: What is going on this plot?



Trauma and bite related deaths are seasonal



Data source: MoHWF's HMIS database
@genomeofindia

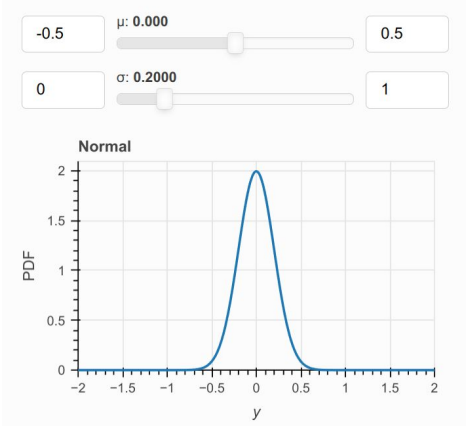
Some potential questions we would like to answer w.r.t the plots

- 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

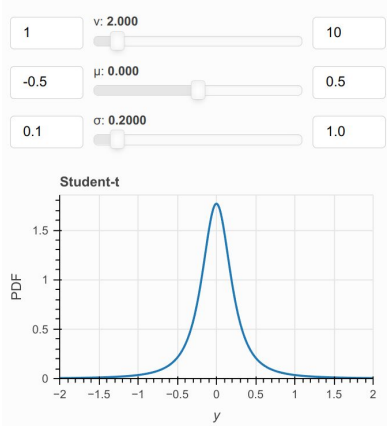


Normal

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

Mean: μ

Variance: σ^2

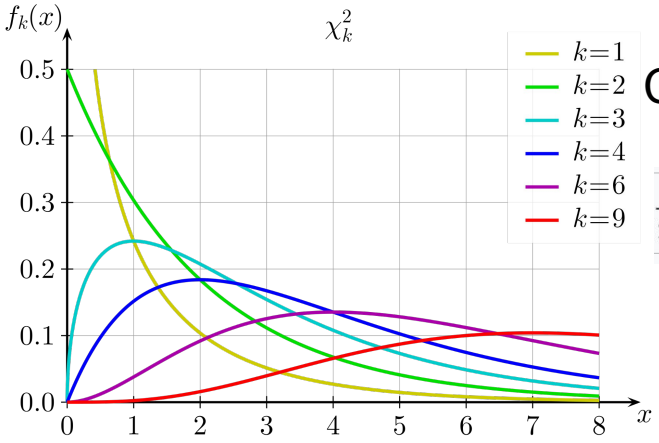


Student's t

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

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.

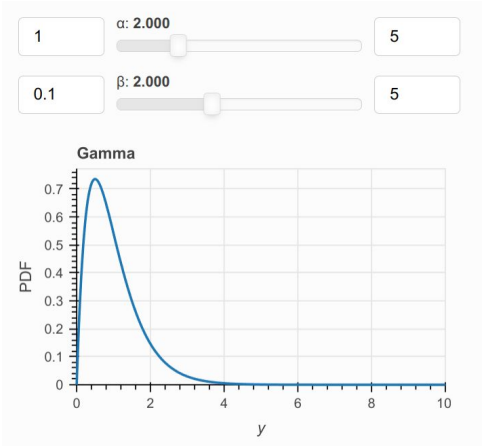


Chi-square

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

Mean = k

Variance = 2k



Gamma

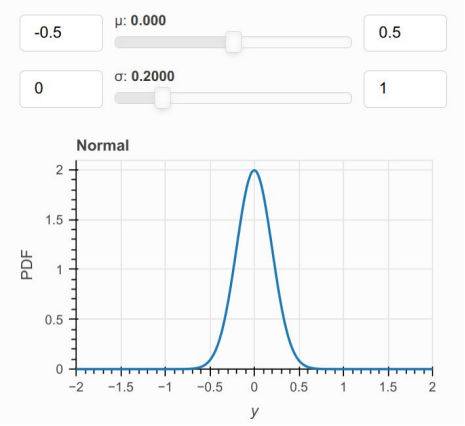
$$f(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\,\frac{(\beta y)^\alpha}{y}\,\mathrm{e}^{-\beta y},$$

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

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

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

Gaussian a.k.a Normal distribution



Normal

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

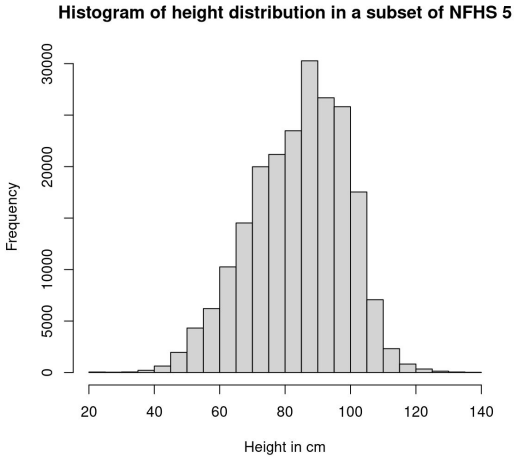
Mean: μ

Variance: σ^2

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

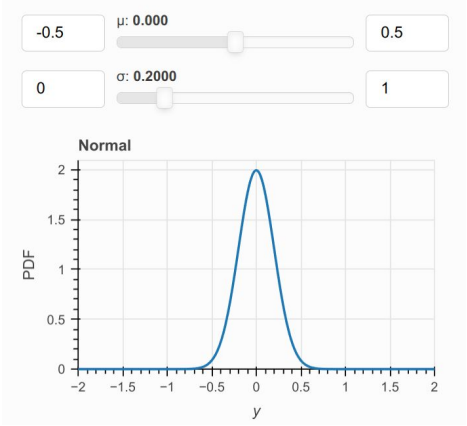
Quantities that are sum of large number of subprocesses tend to be normally distributed

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

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

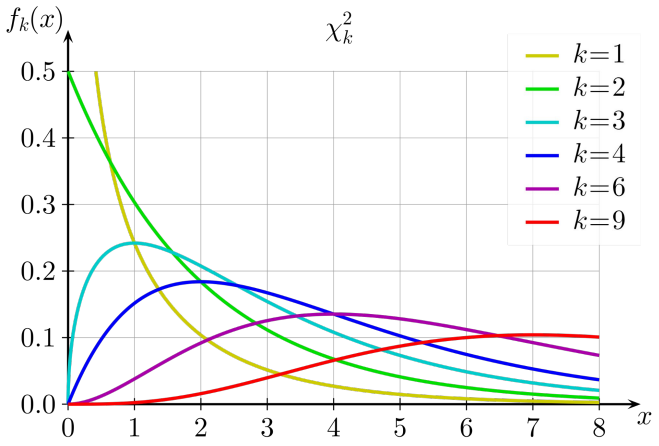
Standard normal has mean 0, variance 1

"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*

Questions?