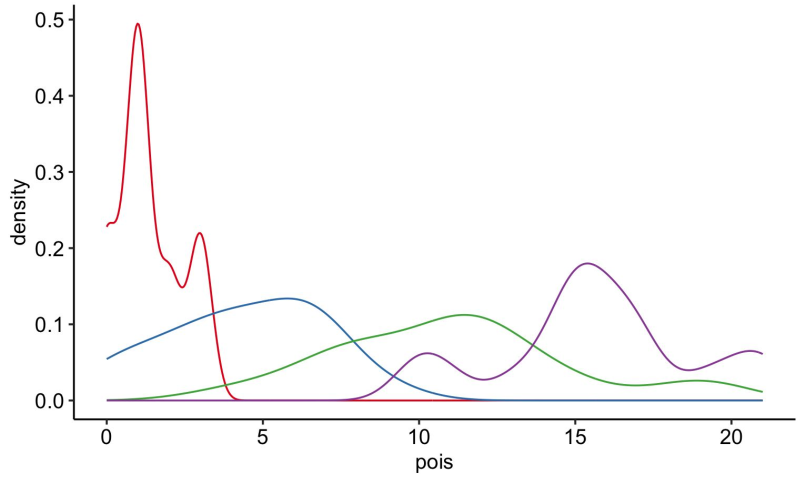
Statistical models for discrete health* data



Saket Choudhary
saketc@iitb.ac.in

Introduction to Public Health Informatics

DH 302

Lecture 03 || Wednesday, 15th January 2025

Class website

https://saket-choudhary.me/DH302/

DH607 - Introduction to Public Health Informatics   Home   Syllabus   Programming resources

Syllabus

Week	Date	Topic	Slides	Assignment	Resources
01	01-08 (Wed)	1. Introduction to the course and history of public health informatics			
01	01-10 (Fri)	2. Statistical models for health data - 1			
02	01-15 (Wed)	3. Statistical models for health data - 2 and Hands On			
02	01-17 (Fri)	4. Exploratory data analysis		• Assignment 01 released	
03	01-22 (Wed)	5. Dimensionality reduction for healthcare data			

All discussions on Piazza: https://piazza.com/iit_bombay/spring2025/dh302

Pooja Sankar Talk

TAs and office hours

Anisha Karmakar
23D1622@iitb.ac.in
Friday, 3-4 pm, BSBE (Lab 605)

Chetan Patil
20b030012@iitb.ac.in
Wednesday, 2-3 PM, KCDH Lab

Devendra Singh
devendrasb@iitb.ac.in
Friday, 5-6 PM KCDH Lab

Kriti A
210100083@iitb.ac.in
Tuesdays, 5-6 PM, ME Department

Shobhit Aggarwal
20d100026@iitb.ac.in
Wednesdays,4-5PM, KCDH Lab

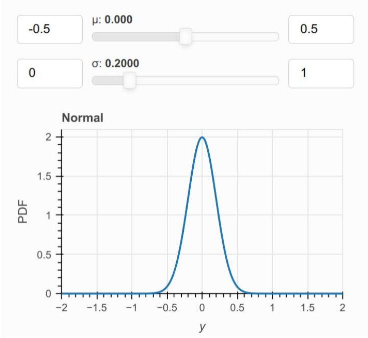
Sunny Gupta
sunnygupta@iitb.ac.in
Friday, 4-5 pm, Medal EE

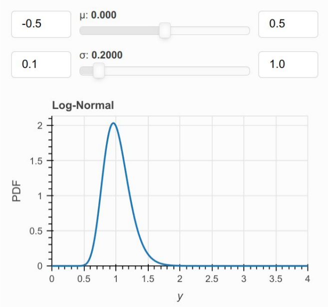
Probability models for health data

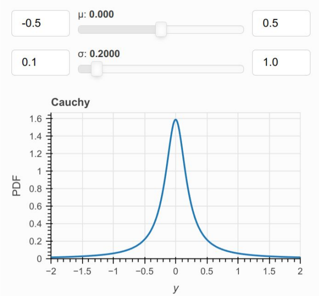
- How should the number of Covid-19 cases be modeled?
- What is the correct statistical model for representing deaths as a function of time?
- What is the distribution of height of males in a village? What about children in village? What about children in a village known to be suffering from stunting?

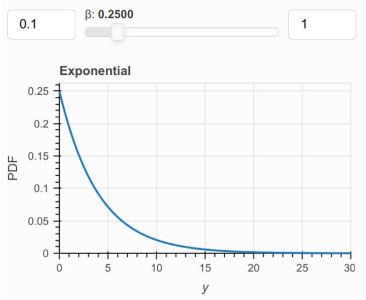
How to think about distributions? The most important ones..

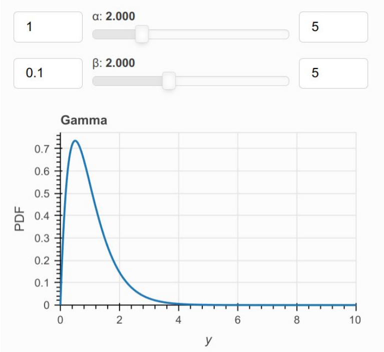
Continuous



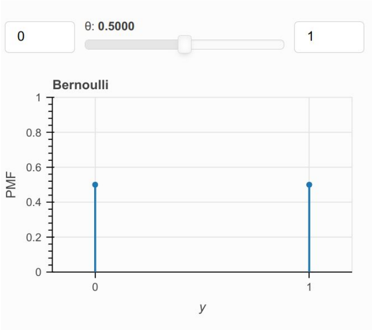


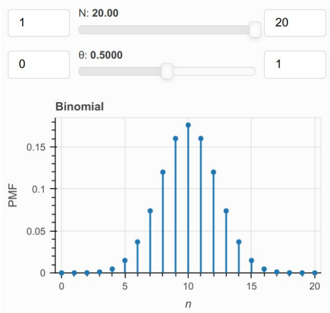


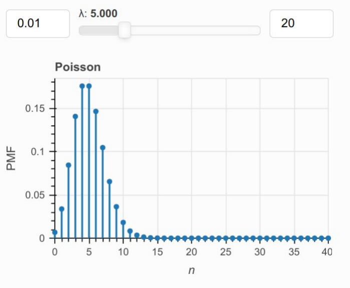


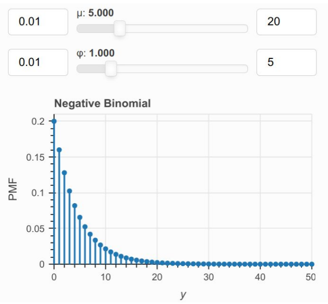


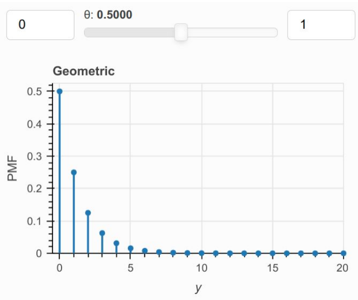
Discrete







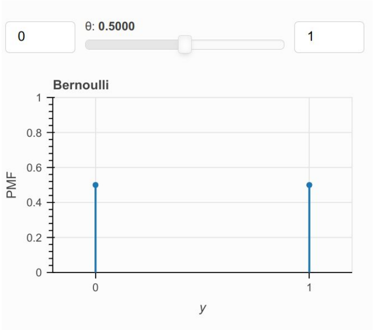


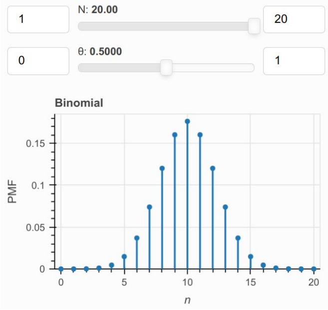


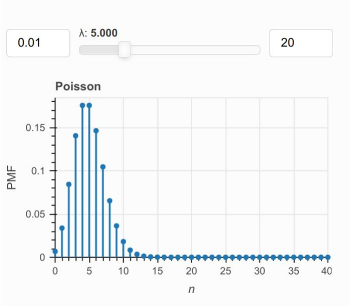
https://distribution-explorer.github.io/

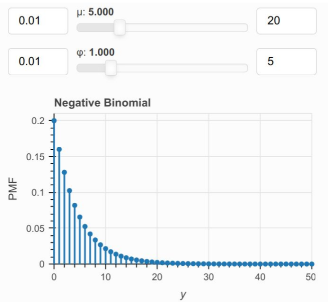
How to think about distributions? The most important ones..

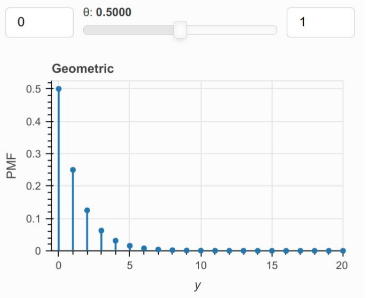
Discrete











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

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

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

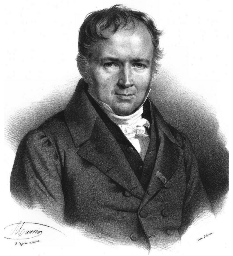
Mean: μ

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

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

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

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



Simeon Poisson,

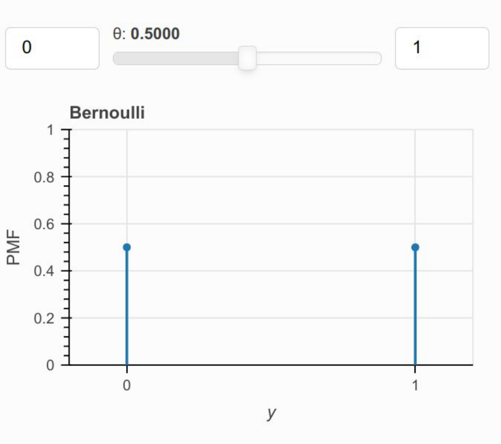
https://distribution-explorer.github.io/

What are the parameters of a distribution?

Parameters are the “knobs” of a distribution: Controls the “center” and the “spread” of the distribution

Distribution	Parameters
Bernoulli	θ
Binomial	θ
Poisson	λ
Negative Binomial	μ, ϕ
Geometric	θ

Bernoulli



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

Mean: θ

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

A *Bernoulli trial* is an experiment that has two outcomes that can be encoded as success (x=1) or (x=0). The result x of a Bernoulli trial is Bernoulli distributed.

Example: Outcomes of a drug on a patient (success or failure);

Parameter: θ = Probability of success of the trial

```
R: dbern(x, prob, log = FALSE)
```

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

Why care about “fitting” a distribution?

If you know the “knobs” of your distribution, the underlying probability distribution is the “hardware” of your machine

Different machines are good at different tasks

Our goal is to find the best hardware for a given set of observations (tasks)

Likelihood of a bernoulli distribution

We observe $x_1, x_2, \ldots, x_n$ outcomes from a Bernoulli trial and are interested in estimating the θ parameter.

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

It is often easier to work with log likelihood

$$\ell(\theta) = \log\theta \sum_{i=1}^{n} x_i + \log(1-\theta) \sum_{i=1}^{n} (1-x_i)$$

Maximum likelihood estimation (MLE)

In maximum likelihood estimation (MLE), our goal is to estimate the value of θ such that the value of our likelihood function is maximized. More formally, in MLE we estimate $\hat{\theta}$ such that

$$\ell(\theta) = \log\theta \sum_{i=1}^{n} x_i + \log(1-\theta) \sum_{i=1}^{n} (1-x_i)$$

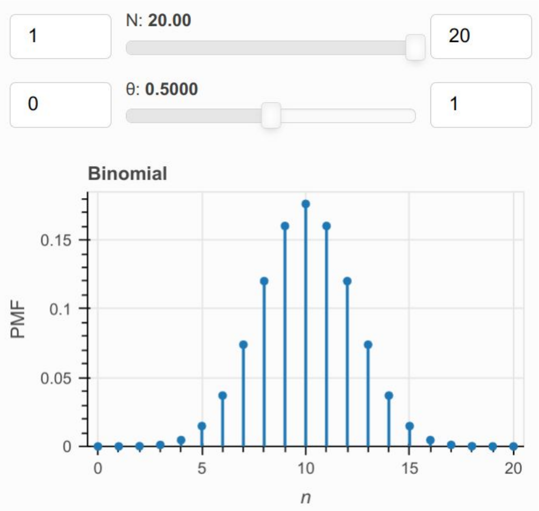
$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-\theta} \overset{\text{set}}{=} 0$$

$$\sum_{i=1}^{n} x_i - \theta \sum_{i=1}^{n} x_i = \theta \sum_{i=1}^{n} (1-x_i)$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = \frac{-\sum_{i=1}^{n} x_i}{\theta^2} - \frac{\sum_{i=1}^{n}(1-x_i)}{(1-\theta)^2} < 0$$

Binomial



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

Mean: $N\theta$

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

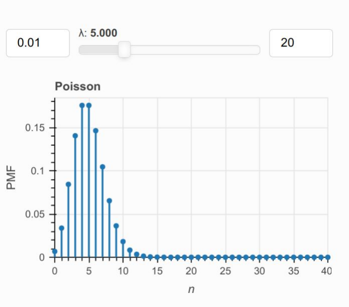
Perform N Bernoulli trials, each with probability θ of success. The number of successes, n, is Binomially distributed.

Parameters: N, θ

Example: Number of people getting infected from a virus giving the probability of infection is θ

```
R: dbinom(k,n,p)
```

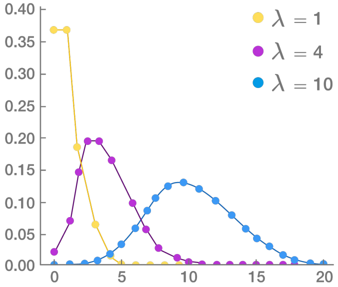
Poisson

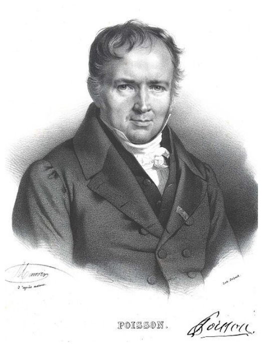


$$f(n;\lambda) = \frac{\lambda^n}{n!}\,\mathrm{e}^{-\lambda}.$$

Mean: λ

Variance: λ





Probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate. Usually used to model “rare” events

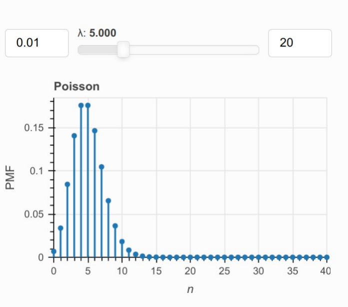
$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Example: Number of mutations in a strand of DNA; Number of maternal deaths during labor

Parameter: λ

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

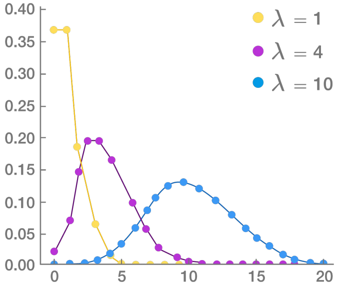
Poisson

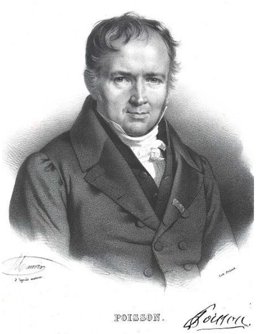


$$f(n;\lambda) = \frac{\lambda^n}{n!}\,e^{-\lambda}.$$

Mean: λ

Variance: λ





Poisson distribution is often used to

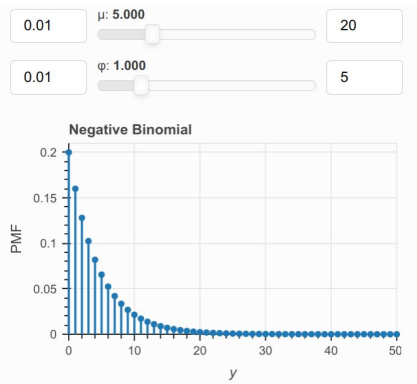
$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Example: Number of mutations in a strand of DNA; Number of maternal deaths during labor

Parameter: λ

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

Negative Binomial



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

Mean: μ

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

Parameters: ϕ, μ

Perform a series of Bernoulli trials with probability $\beta/(1+\beta)$

The number of failures, *y*, before we get *α* successes is Negative Binomially distributed. (Usually employed when a simple poisson model fails)

$$f(y;\alpha,\beta) = \binom{y+\alpha-1}{\alpha-1}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{y}.$$

Generally speaking, *α* need not be an integer, so we may write the PMF as

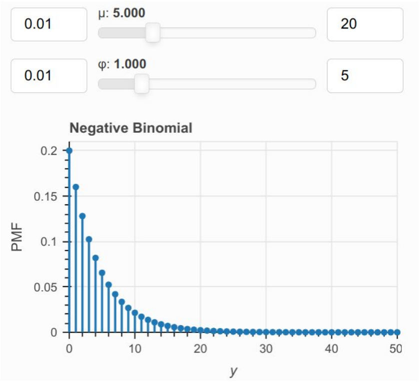
$$f(y;\alpha,\beta) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\,y!}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{y}.$$

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

```
R: dnbinom(x, size, prob, mu, log = FALSE)
```

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

Why is Negative Binomial called so?



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

Mean: μ

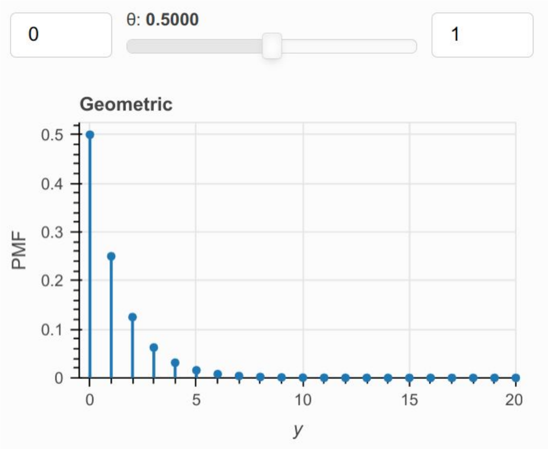
Variance: $\mu\left(1+\frac{\mu}{\phi}\right)$.

Parameters: ϕ, μ

$$f(y;\alpha,\beta) = \binom{y+\alpha-1}{\alpha-1}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{y}.$$

$$\begin{aligned}\binom{y+\alpha-1}{\alpha-1} &= \frac{(y+\alpha-1)!}{(\alpha-1)!y!} \\ &= \frac{(y+\alpha-1)(y+\alpha-2)\dots\alpha}{y!} \\ &= (-1)^{\alpha}\frac{(-\alpha-1)(-\alpha-2)\dots(-\alpha-y+1)}{y!} \\ &= (-1)^{\alpha}\binom{-\alpha}{y}\end{aligned}$$

Geometric



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

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

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

Perform a series of Bernoulli trials with probability of success θ until we get a success. The number of **failures** before the success is geometrically distributed

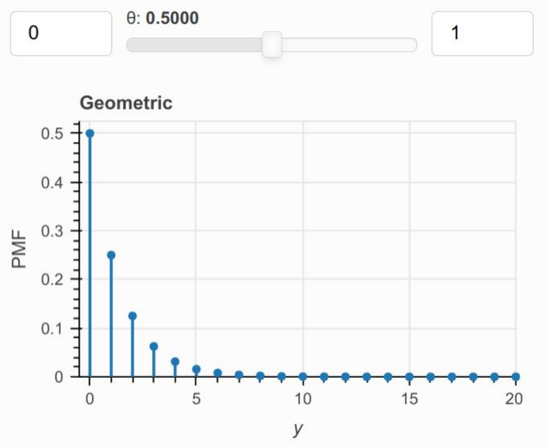
Parameters: θ

Examples: Number of visits to a primary health care center before the doctor actually sees you

```
R:dgeom(x, prob, log = FALSE)
```

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

Geometric



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

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

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

Perform a series of Bernoulli trials with probability of success θ until we get a success. The number of **failures** before the success is geometrically distributed

Parameters: θ

Examples: Number of visits to a primary health care center before the doctor actually sees you

```
R:dgeom(x, prob, log = FALSE)
```

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

R demo

Questions?