Statistics for Data Science

Lecture 5

Distributions

Ken Cameron

Admin

- Background maths tutorials
- Wednesdays 10.15am 3W 3.9
- We’ll start with differentation

Content

- Statistical Distributions
  - Modelling
  - Discrete
  - Continuous

Modelling

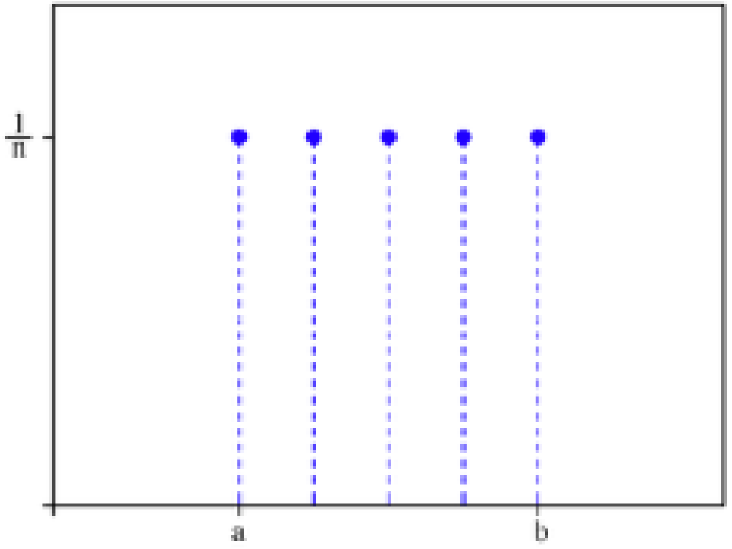
- Statistical distributions can be used to model populations.
  - We work with a family of distributions.
  - The family is defined by one or more parameters
- Example: The normal distribution
  - With μ, the mean as a parameter, $-\infty < \mu < \infty$

Discrete Distributions

- Uniform
- Hypergeometric
- Binomial
- Poisson
- Negative Binomial
- Geometric

Uniform

- $P(X = x|N) = 1/N$, $x = 1, 2, 3, ..., N$.
- Where N is an integer.
- Equal chance of each outcome.



- Mean

$$EX = \frac{(N+1)}{2}$$

- Variance

$$EX^2 = \sum_{x=1}^{N} x^2 \frac{1}{N} = \frac{(N+1)(2N+1)}{6}$$

$$VarX = EX^2 - (EX)^2 = \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{(N+1)(N-1)}{12}$$

Hypergeometric

- Example: Bag containing N balls, M red, N-M green. Select K balls.
  - What is the probability that x are red?

$$P(X \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, x = 0,1,\ldots,K.$$

M >= x and N-M >= K – x
M – (N – K) <= x <= M

- Mean

$$EX = \sum_{x=1}^{K} x \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} = \frac{KM}{N}$$

- Variance

$$VarX = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right)$$

Binomial

- Based on Bernoulli trial

  - X = 1 with probability P
  - X = 0 with probability 1 – P
  - 0 <= 0 <= 1

  - EX = 1p + 0(1 - p) = p

  - Var X = $(1 - p)^2 p + (0 – p)^2 (1 - p) = p(1 - p)$

n Bernoulli trials

- A_i = { X = 1 on the i^{th} trial}, i = 1, 2, ... ,n.
- Assume A_1, ... A_n are independent events.
- Random Variable Y = sum of X_i
- Has the binomial distribution with two parameters: n, p.

$$P(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \text{ y= 0,1,…,n}$$

- Mean

$$EX = np$$

- Variance

$$\mathrm{Var}\, X = np(1 - p)$$

Example

- What are the chances of one 6 in four rolls of a fair die?
- Model as four Bernoulli trials p=1/6.
- Binomial(4, 1/6)
- X = total number of 6s in four rolls.

Example (cont.)

- P(at least one 6) = P(X > 0) = 1 – P(X = 0)

$$1-\binom{4}{0}\left(\frac{1}{6}\right)^{0}\left(\frac{5}{6}\right)^{4}=1-\left(\frac{5}{6}\right)^{4}=0.518$$

Poisson

- Often used to model waiting for an event.
  - E.g. Bus arriving.

- Single parameter λ
  - Referred to as the intensity.

$$P(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,\ldots$$

- Mean

$$EX = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^{x}}{x!} = \lambda$$

- Variance

Var X = λ

Example

- Website accessed on average 5 times every 3 minutes.
  - What is the probability of no accesses in the next minute?
- Random variable X number of accesses in a minute.
  - Poisson distribution with $\lambda = 5/3$.

Example (cont.)

- P(no accesses in the next minute) = P(X = 0)

$$\frac{e^{-5/3}(5/3)^0}{0!} = e^{-5/3} = 0.189$$

- Quick test: P(at least two accesses in the next minute) ?

Negative Binomial

- The binomial distribution counts the number of successes for a fixed number of Bernoulli trials.

- Suppose we count the number of trials required to get a fixed number of successes.

- In a sequence of Bernoulli(p) trials, let X denote the trial at which r^{th} success occurs.

Negative Binomial

$$P(X = x \mid r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, x = r, r+1, \ldots$$

It is often easier to consider it in terms of the number of failures before the r^{th} success. Y = X - r

$$P(Y = y) = \binom{x+y-1}{y} p^r (1-p)^y, y = 0,1,..$$

Negative Binomial

- Mean

$$EY = \frac{r(1-p)}{p} \to \lambda$$

- Variance

$$VarY = \frac{r(1-p)}{p^2} \to \lambda$$

- Includes the poisson distribution as the limiting case.

Geometric

- The simplest of the waiting time distributions.

- A special case of the negative binomial distribution.

- Set $r = 1$

$$P(X = x|p) = p(1 - p)^{x-1}, \; x=1,2,....$$

Geometric

- Recall $Y = X - r$.
  - So $X = Y + 1$

- Mean

$$EX = EY + 1 = 1/p$$

- Variance

$$\text{Var } X = (1 - p)/p^2$$

Geometric

- Useful property : memoryless
- For integers $s > t$: $P(X>s|X>t) = P(X > s-t)$
- It forgets what has occurred.
  - The chance of getting an additional s-t failures having already observed t failures is the same as observing s-t failures at the start of the sequence.
- <u>The chance of getting a run of failures depends only on the length of the run not its position.</u>

Continuous Distributions

- Uniform
- Normal
- Lognormal
- Double Exponential

Uniform

- Defined by spreading mass uniformly over an interval [a, b]
- pdf:

$$f(x \mid a,b) = \begin{cases} \dfrac{1}{a-b}, ifx \in [a,b] \\ 0, otherwise \end{cases}$$

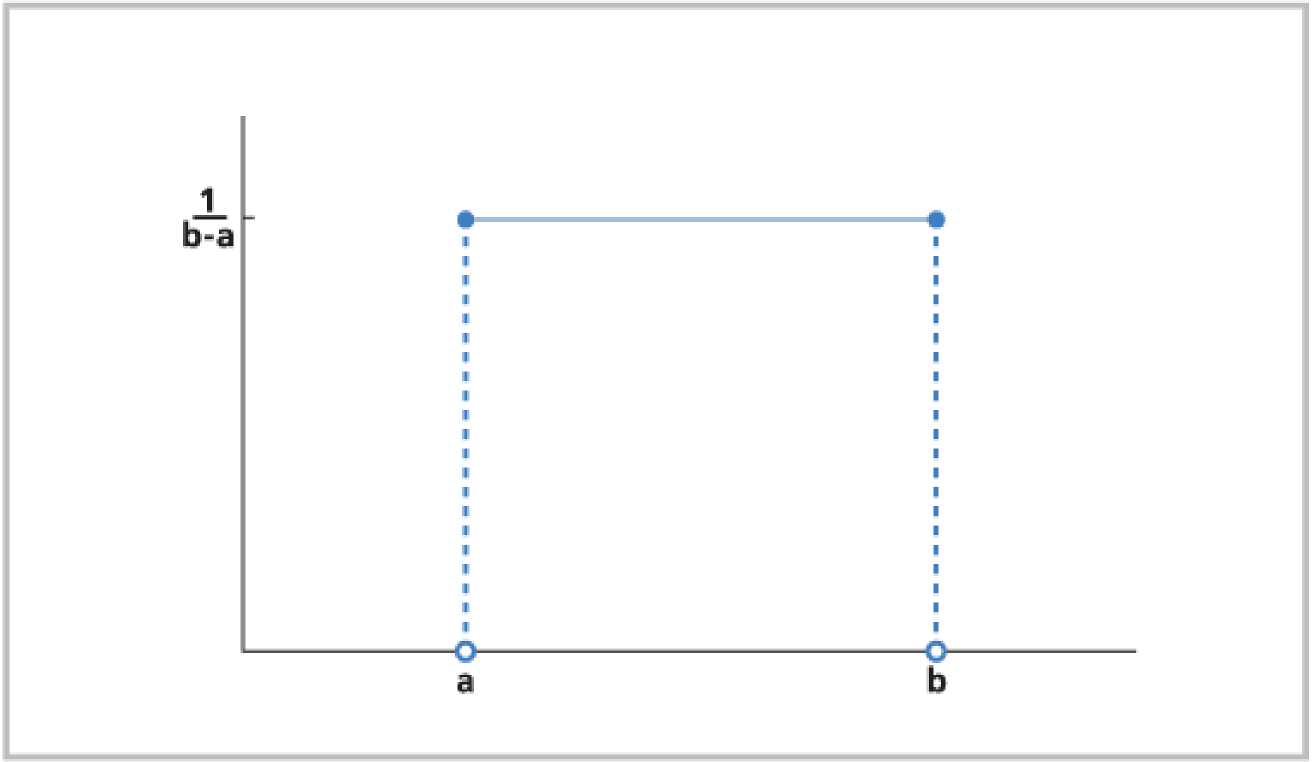
- Mean

$$EX = \int_a^b \frac{x}{b-a} dx = \frac{b+a}{2}$$

- Variance

$$VarX = \int_a^b \frac{(x - \frac{b+a}{2})^2}{b-a} dx = \frac{(b-a)^2}{12}$$

Uniform

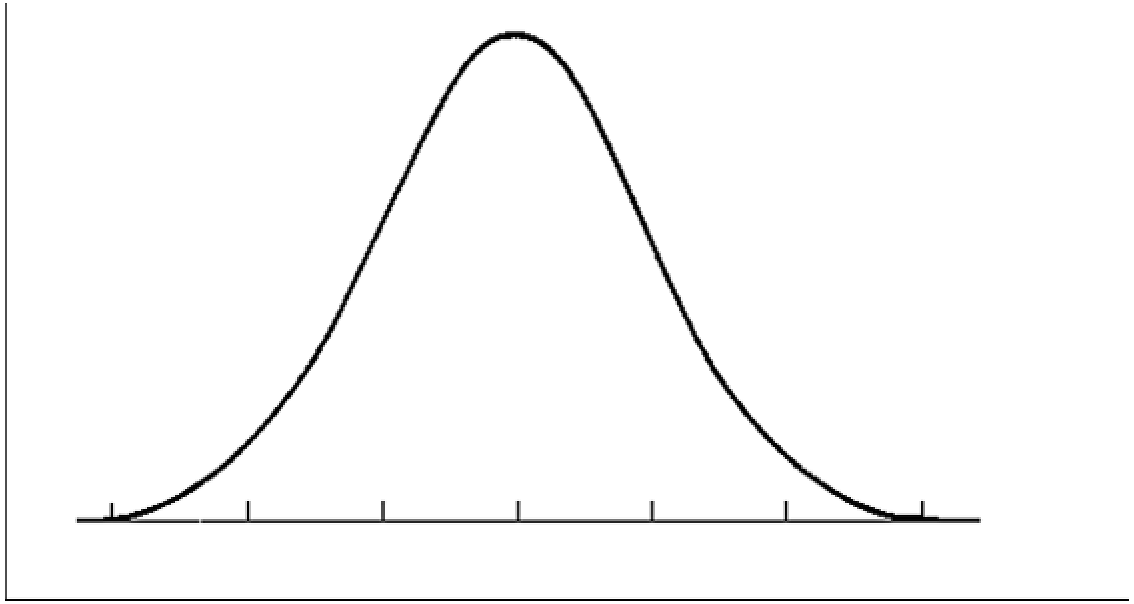


Normal (Gaussian)

- Popular model
  - Very tractable analytically
  - Symmetric (bell curve)
  - Central Limit Theorem (a later lecture) shows that normal distribution can be used to approximate a large range of distribution, in large samples.

Parameters

- Two parameters:
  - μ The mean
  - σ^2 The variance



- pdf:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/(2\sigma^2)}, -\infty < x < \infty$$

Lognormal

- If X is a random variable whose logarithm is normally distributed.
  - Log X ~= n(μ,σ^2)
- Then X has a lognormal distribution.

$$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(\log x-\mu)^2/(2\sigma^2)}, 0 < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

- Mean

$$EX = e^{\mu + (\sigma^2/2)}$$

- Variance

$$VarX = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$$

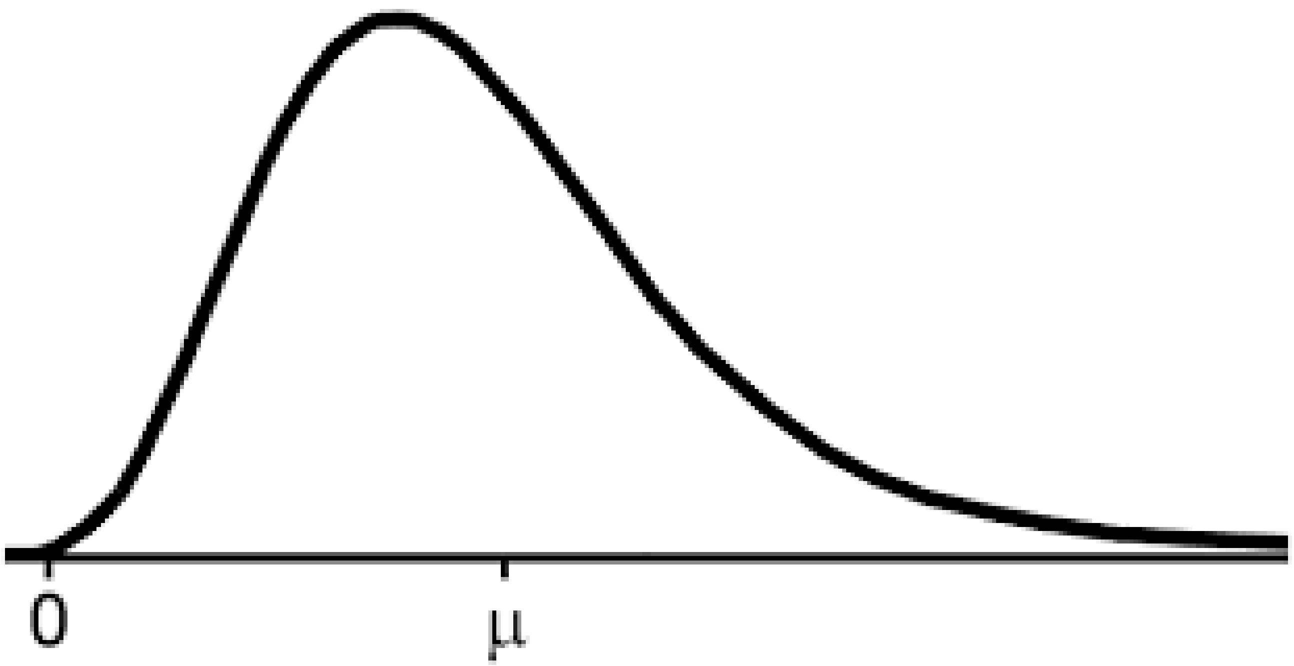
Uses

- Popular for modelling where the variable of interest is skewed to the right.

- For example, incomes

  necessarily skewed to the right.

  - We can use normal-theory statistics on log(income).



Double Exponential

- Formed by reflecting the exponential distribution around its mean.

- pdf

$$f(x\,|\,\mu,\sigma) = \frac{1}{2\sigma} e^{-|x-\mu]/\sigma}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

- EX = μ
- Var X = $2\sigma^2$

Other distributions are available.

- This is only a small set of the available distributions.
  - Hopefully the most useful.
- Others may crop up in the lectures.
- For a more complete list see:
  - Multi-volume work Distributions in Statistics
  - By Johnson, Kotz and Balakarishnan (1994,1995) and Johnson, Kotz and Kept (1992).

Challenge

- Assume you take delivery of a consignment of 25 disc drives.
  - As part of acceptance testing, you run the self test on 10 of them.
  - What is the probability of all 10 passing if 6 of the 25 are faulty?
- Start by picking the right distribution to model the problem.