WEEK 2

Discrete random variables

Felipe Campelo

Adapted from Montgomery and Runger's
Applied Statistics and Probability for Engineers
lecture notes (Ch 3)

About this slide deck

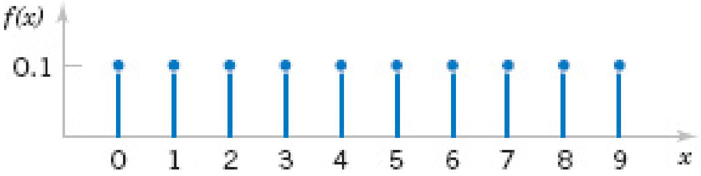
These slides summarise some discrete probability distributions, which are used to model different sorts of processes.

These slides should be considered in the context of **Chapter 3** our main textbook (*DC Montgomery, GC Runger, [Applied statistics and probability for engineers, 5th ed.]*), more specifically sections 3.5 - 3.9.

You're encouraged to read that chapter and try your hand at the worked examples.

Discrete uniform distribution

A random variable X has a **discrete uniform distribution** if each of the k values in its range $\{x_1, \dots, x_k\}$ has equal probability. Then,



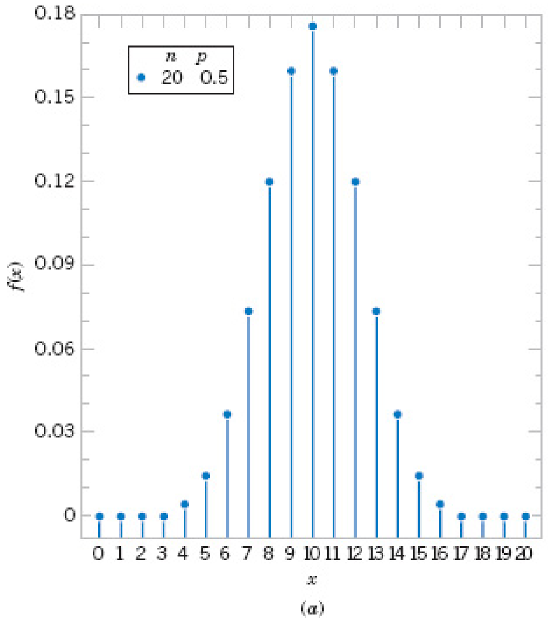
$f(x_i) = 1/k, \text{for } i = 1, \dots, k$

This distribution emerges in process such as dice rolls, last digits of student ID numbers, attribution of calls to operators in customer support centres, etc.

Suppose the range of X is the consecutive integers $a, a+1, \dots, b, \text{for } a \leq b$, which contains $\mathrm{k} = \mathrm{b} - \mathrm{a} + 1$ possible values, each with a probability $1/k$. Then, it can be shown that:

$\mu = E[X] = \frac{a+b}{2}$ and $\sigma^2 = V(X) = E[(X-\mu)^2] = \frac{(b-a+1)^2 - 1}{12}$

Binomial distribution



Remember that a **Bernoulli trial** is a random experiment which can have only two outcomes, success (1) or failure (0), each with a given probability.

Consider a random experiment that consists of n independent and identically distributed (*iid*) Bernoulli trials, each with a probability of success given by p (e.g., $n = 20$ tosses of a fair coin, $p = 0.5$).

The random variable X that equals the number of trials out of n that result in a success is a **binomial random variable,** with parameters $p \in [0,1]$ and $n \in \mathbb{Z}_{\geq 0}$. This variable has a pmf given by:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{for } x = 0,1,\dots,n$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Binomial distribution

Binomial distributions emerge in several relevant situations, such as:

- A machine produces 1% defective parts. The number of defective parts in the next 25 parts produced is a Binomial RV, $X \sim Binom(p = 0.01, n = 25)$.
- When sampling for hidden dangerous chemicals, assume that each sample of air from a military complex has a 0.5% chance of containing a particular chemical signature if it was present in the last 6 months. If an inspection team collects 30 samples from the complex, the number of air samples that will detect the chemical can be modelled as $X \sim Binom(p = 0.005, n = 30)$.

The mean and variance of a binomial RV X are given by:

$\mu = E[X] = np$ and $\sigma^2 = V(X) = np(1 - p)$

Example

Assume the chemical detection example from the previous slide, modelled as $X \sim Binom(p = 0.005, n = 30)$. If a facility actually had the material in it, what is the total probability that the inspection team will be able to detect the material if they take 30 air samples? Assume that the team can only declare that the material was present if 2 or more samples test positive.

Example

Assume the chemical detection example from the previous slide, modelled as $X \sim Binom(p = 0.005, n = 30)$. If a facility had the material in it, what is the total probability that the inspection team will be able to detect the material if they take 30 air samples? Assume that the team can only declare that the material was present if 2 or more samples test positive.

In this case, we want to know the sum of probabilities for the cases $x \geq 2$. Given that the total probability needs to add to one, it is easier to compute this as:

$$P(x \geq 2) = 1 - P(x \leq 1) = 1 - f(0) - f(1)$$

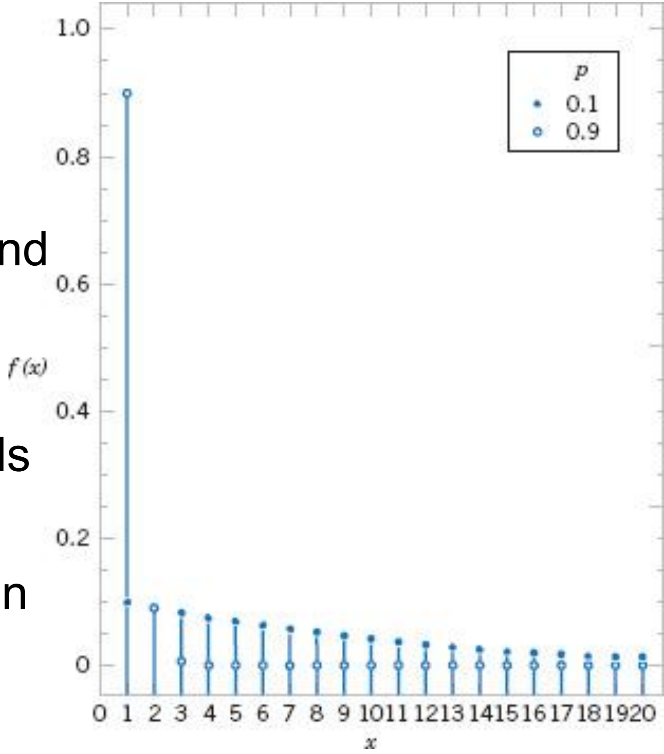
$$f(0) = \binom{30}{0} 0.005^0 . 0.995^{30} = 0.8604$$

$$f(1) = \binom{30}{1} 0.005^1 . 0.995^{29} = 0.1297$$

$$P(x \geq 2) = 1 - f(0) - f(1) = 0.0099$$

```
R 4.4.1 · ~/
>
> # Using the binomial CDF function in R:
> 1 - pbinom(1, 30, .005)
[1] 0.009909649
>
```

Geometric distribution



The Binomial distribution has a fixed number of trials and a random number of successes. The **Geometric distribution** has this reversed: a random number of trials, and the number of successes fixed at 1.

This distribution models the number of *iid* Bernoulli trials (with probability parameter $p \in [0,1]$) required until the first success is observed. If X is the random variable denoting the number of trials until the first success, then X is a **Geometric random variable**, with pmf:

$$f(x) = p(1-p)^{x-1}, \text{for } x = 1,2, \ldots$$

and

$$\mu = E[X] = \frac{1}{p}, \qquad \sigma^2 = V(X) = (1-p)/p^2$$

Example

Going back to the chemical detection example $(p = 0.005)$, suppose that samples are collected and analysed sequentially. What is the probability of the first detection being observed at any point in the first 10 samples?

Example

Going back to the chemical detection example ($p = 0.005$), suppose that samples are collected and analysed sequentially. What is the probability of the first detection being observed at any point in the first 10 samples?

The pmf in this case is given by $f(x) = 0.005 \times 0.995^{x-1}$. We want the sum of probabilities for all cases where $x \leq 10$, i.e., the value of the cumulative distribution function at $x = 10$,

$$F(x = 10) = P(x \leq 10) = 0.005 \sum_{x=1}^{10} 0.995^{x-1} = 0.0489$$

R 4.4.1 · ~/

```
>
> # Using the geometric CDF function in R:
> pgeom(9, 0.005)
[1] 0.04888987
>
```

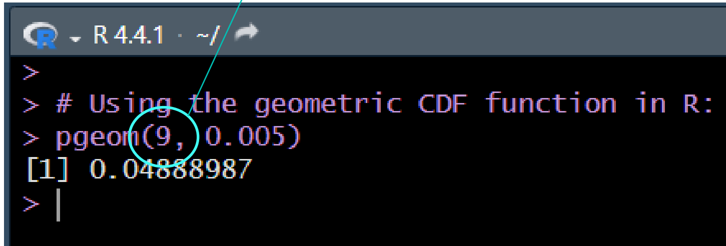
Example

Going back to the chemical detection example ($p = 0.005$), suppose that samples are collected and analysed sequentially. What is the probability of the first detection being observed at any point in the first 10 samples?

The pmf in this case is given by $f(x) = 0.005 \times 0.995^{x-1}$. We want the sum of probabilities for all cases where $x \leq 10$, i.e., the value of the cumulative distribution function at $x = 10$,

$$F(x = 10) = P(x \leq 10) = 0.005 \sum_{x=1}^{10} 0.995^{x-1} = 0.0489$$

In R, the binomial cdf is parameterised by "number of failures before a success is observed"



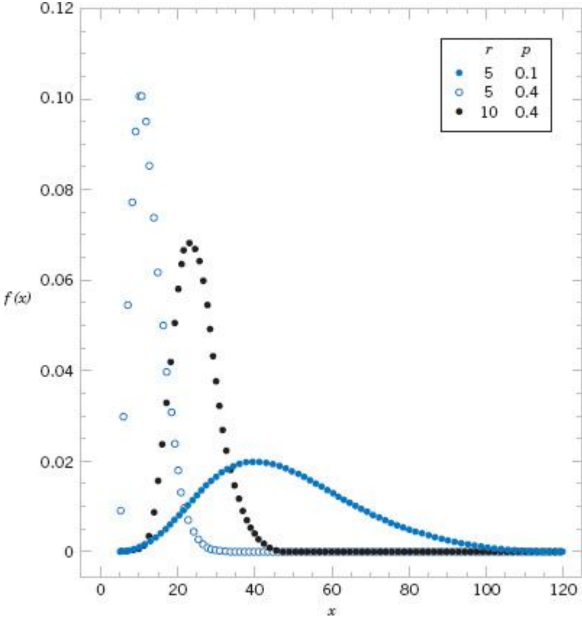
Lack of memory

For a Geometric RV, the trials are **independent**: the “count of number of trials until a success is observed” can be started at any trial, without changing the probability distribution of the random variable. Mathematically,

$$P(x > a + b | x > a) = P(x > b), \text{for } a, b \geq 0$$

The implication is that the system is unaffected by its history – hence, the Geometric distribution is said to lack any memory.

Negative binomial distribution

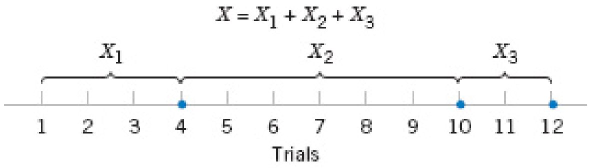


The **Negative Binomial** distribution is a generalisation of the geometric distribution, in which the RV represents the number of Bernoulli trials with parameter $p \in [0,1]$ required to obtain $r \in \mathbb{Z}_{>0}$ successes.

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

and

$$\mu = E[X] = r/p, \quad \sigma^2 = V(X) = r(1-p)/p^2$$



*Note: the **lack of memory** property of the Geometric distribution makes it possible to express a Negative Binomial distribution as a sum of iid Geometric RVs.*

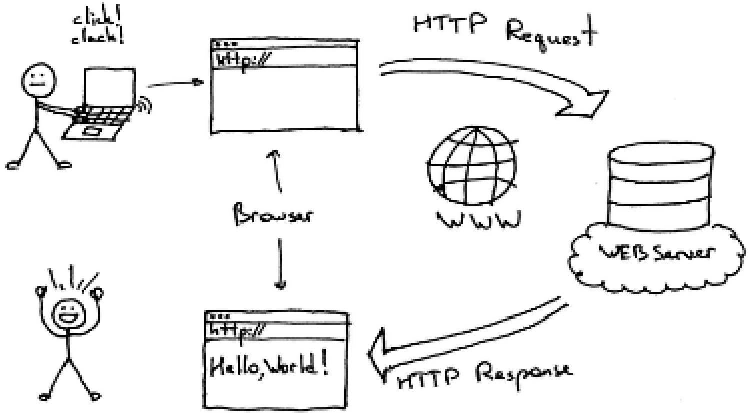
Poisson distribution

Imagine that you have a system monitoring the (random) arrival of requests on a web service.

Let X be the RV that counts the number of requests in a time interval Δt, and call the average number of requests per Δt interval as λ.

Suppose now that we partition Δt into n sub-intervals of infinitesimally small length. The probability p that a sub-interval will contain a request decreases rapidly, and the probability that two or more requests arrive in the same sub-interval tends to zero.

We can model the distribution of X as approximately a Binomial random variable, and since $\mathrm{E}[X] = \lambda = np$, we have $p = \lambda/n$.



Poisson distribution

More generally, consider an interval T of real numbers partitioned into subintervals of small length δt, and assume that as δt tends to zero:

1. The probability of more than one event in any subinterval tends to zero.
2. The probability of one event in a subinterval tends to $\lambda\delta t/T$.
3. The event in each subinterval is independent of other subintervals.

A random experiment with these properties is called a **Poisson process**. The random variable X that counts the number of events in a Poisson process is a **Poisson random variable** with parameter $\lambda \in \mathbb{R}_{>0}$ and a pmf given by:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{for } x \in \mathbb{Z}_{>0}$$

Both the **mean** and the **variance** of a Poisson variable are equal to λ**.**