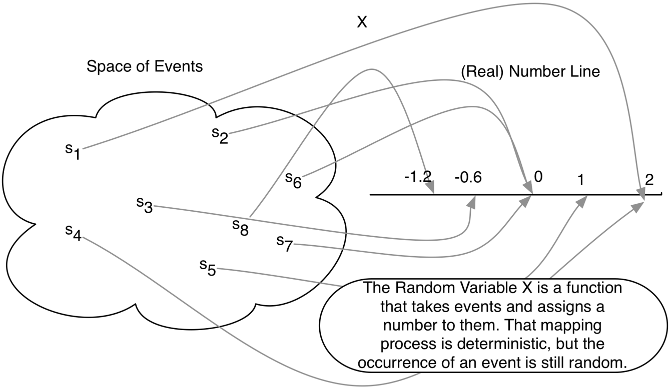
Probability and Statistics

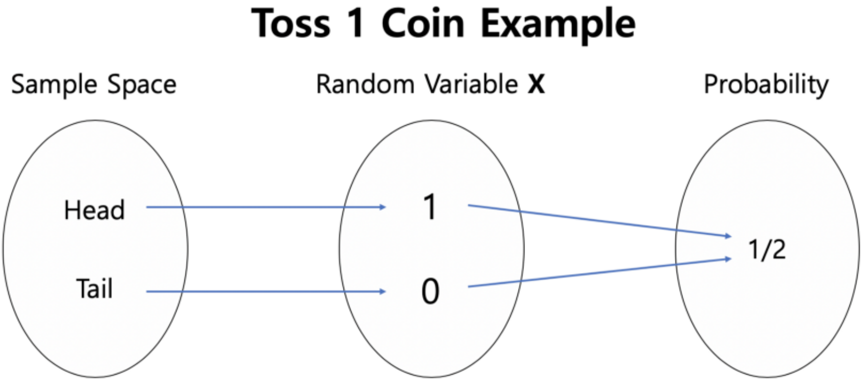
Overview

1. Random variables
   - Main concepts
   - Distribution of a random variable
   - Types of random variables
2. Expectation and variance
   - Expectation
   - Variance and standard deviation
   - Chebyshev's inequality
3. Families of discrete distributions
   - Bernoulli distribution
   - Binomial distribution
   - Geometric distribution
   - Negative binomial distribution
   - Poisson distribution

Random variables I



Random variables II



Random variables III

Definition

A random variable is a function which assigns a numerical value to the outcomes of an experiment,

$$X = f(\omega),\ \omega \in \Omega$$

where $f : \Omega \to \mathbb{R}$. In other words it is a quantity that depends on chance.

- The domain of a random variable is the sample space Ω.
- Its range can be $\mathbb{R}$, $\mathbb{Z}$, $(0, 1)$, etc., depending on what possible values the random variable can potentially take.

Random variables IV

Example 1a.

Consider an experiment of tossing 3 fair coins and counting the number of heads.

•The same model suits the number of girls in a family with 3 children, the number of 1's in a random binary code consisting of 3 characters, etc.

•The sample space is $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

•Let X be the number of heads, therefore X takes the values 0 to 3.

• Since each value of X is an event, we can compute probabilities,

$$P(X=0) = P(TTT) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8},$$

$$P(X=1) = P(HTT) + P(THT) + P(TTH) = \frac{3}{8},$$

$$P(X=2) = P(HHT) + P(HTH) + P(THH) = \frac{3}{8},$$

Random variables V

$$P(X = 3) = P(HHH) = \frac{1}{8}.$$

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

This table contains everything that is known about random variable X prior to the experiment. Before we know the outcome ω, we cannot tell what X equals to. However, we can list all the possible values of X and determine the corresponding probabilities.

This is called the **distribution** of the random variable X.

Example 1b.

Consider the random variable X which represents the sum of the numbers which appear when we roll two dice. Determine the distribution of X.

Distribution of a random variable I

Definition

The **distribution** of X is the collection of all probabilities related to X.
The function $f : \mathbb{R} \to [0, 1]$,

$$f(x) = P(X = x)$$

is the **probability mass function** (**PMF**).
The **cumulative distribution function** (**CDF**) is defined as
$F : \mathbb{R} \to [0, 1]$,

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y).$$

The CDF F(x) is a non-decreasing function of x, always between 0 and 1.
We also have that

$$P(a < X \leq b) = F(b) - F(a).$$

Exercise. Compute the PMF and CDF of X in Example 1.

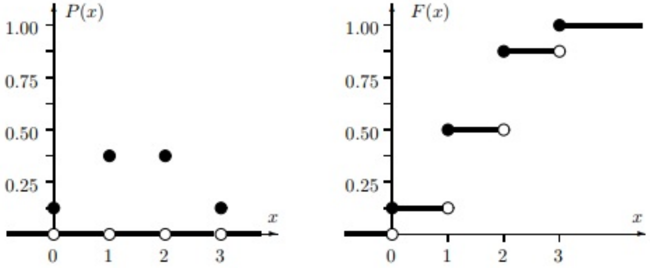
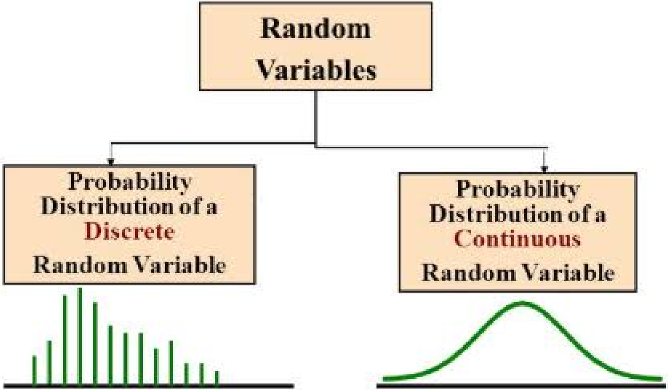


FIGURE .1: *The probability mass function $P(x)$ and the cumulative distribution function $F(x)$ for Example 1. White circles denote excluded points.*

Types of random variables I



Types of random variables II

So far, we are dealing with **discrete random variables**. These are variables whose **range** is **finite or countable** (the values can be listed in a sequence).

Examples: the number of jobs submitted to a printer, the number of errors, the number of error-free modules, the number of failed components, etc.

On the contrary, **continuous random variables** assume a whole interval of values. This could be a bounded interval (a, b), or an unbounded interval such as (a, ∞), $(-\infty, b)$ or $(-\infty, \infty)$.

Examples: various times (software installation time, code execution time, connection time, waiting time, lifetime), or physical variables like weight, height, voltage, temperature, distance, the number of miles per gallon, etc.

Expectation and variance

The detailed information of a random variable X can be summarized in a few vital **characteristics** describing the **average value**, the most likely value of a random variable, its **spread, variability**, etc.

The most commonly used are the **expectation, variance, standard deviation**, introduced in this section.

Also rather popular and useful are **the mode, moments, quantiles, and interquartile range**.

Expectation I

Definition

Expectation or expected value of a random variable X is its mean, the average value.

The expected value of a discrete random variable X with the following PMF is:

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

$$E[X] = \sum_{k=1}^{n} x_k \cdot p_k.$$

Expectation II

Properties of the expected value:

- $E(X+Y) = E(X) + E(Y)$
- $E(aX) = aE(X)$
- $E(c) = c$
- $E(X \cdot Y) = E(X) \cdot E(Y)$, if X, Y are independent.

Example 2.

Determine the expected value of the following random variable:

$$X : \begin{pmatrix} 5.20 & 5.30 & 5.80 & 6.00 \\ 0.4 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

Variance and standard deviation I

Expectation shows where the average value of a random variable is located, or where the variable is expected to be, plus or minus some error. How large could this "error" be, and how much can a variable vary around its expectation? Let us introduce some measures of variability.

Definition

Variance of a random variable is defined as the expected squared deviation from the mean. For discrete random variables, variance is

$$V[X] = E[(X - E[X])^2] = E[X^2] - E^2[X]$$

Definition

Standard deviation is a square root of variance,

$$Std[X] = \sqrt{V[X]}.$$

Variance and standard deviation II

Properties of the variance:

- $V(aX) = a^2 V(X)$
- $V(c) = 0$
- $V(X+Y) = V(X) + V(Y)$, if X, Y are independent.

Example 3.

Determine the expected value of the following random variable:

$$X : \begin{pmatrix} 5.20 & 5.30 & 5.80 & 6.00 \\ 0.4 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

Chebyshev's inequality

Theorem

For all $L > 0$ we have the following inequality:

$$P[|X - E(X)| < L] \geq 1 - \frac{\sigma^2(X)}{L^2}.$$

Property

If $L = k \cdot \sigma$ then:

$$P[|X - E(X)| < k \cdot \sigma] \geq 1 - \frac{1}{k^2}.$$

If $k = 3$ we have

$$P[|X - E(X)| < 3 \cdot \sigma] \geq \frac{8}{9}.$$

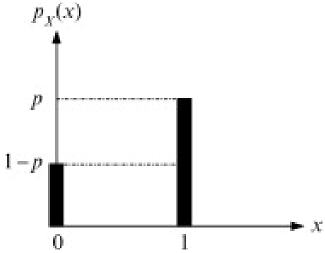
This means that 89% ($\frac{8}{9}$) of the absolute deviations of the variable X do not exceed 3σ (**the "three sigma" rule**).

Families of discrete distributions

Next, we introduce the most commonly used families of **discrete distributions**. Amazingly, absolutely different phenomena can be adequately described by the same mathematical model, or a family of distributions.

For example, as we shall see below, the number of virus attacks, received e-mails, error messages, network blackouts, telephone calls, traffic accidents, earthquakes, and so on can all be modeled by the same **Poisson family of distributions**.

Bernoulli distribution I



Definition

A random variable with two possible values, 0 and 1, is called a **Bernoulli variable**, its distribution is **Bernoulli distribution**, and any experiment with a binary outcome is called a **Bernoulli trial**.

Bernoulli distribution II

Examples of Bernoulli trials: good or defective components, parts that pass or fail tests, transmitted or lost signals, working or malfunctioning hardware, benign or malicious attachments, sites that contain or do not contain a keyword, girls and boys, heads and tails, etc.

For all these cases we shall use generic names for the two outcomes: **"successes"** and **"failures"**.

The probability of *success* is $P(X=1)=p$, then $P(X=0)=1-p=q$ is the probability of *failure*.

$$X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

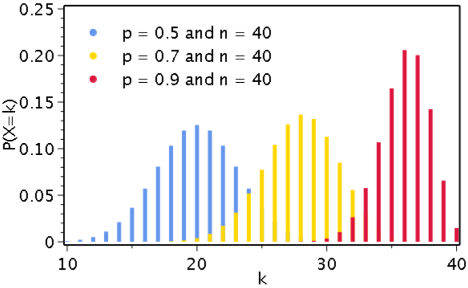
Bernoulli distribution III

Compute the expected value and the variance of the Bernoulli random variable X.

$$E[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

$$V[X] = E[X^2] - E[X]^2 = p - p^2 = p(1-p) = pq$$

Binomial distribution I



Now consider a **sequence of independent Bernoulli trials** and count the number of successes in it.
This may be the number of defective computers in a shipment, the number of updated files in a folder, the number of girls in a family, the number of e-mails with attachments, etc.

Binomial distribution II

Definition

A variable described as the number of successes in a sequence of independent Bernoulli trials has **Binomial distribution**. Its parameters are n, the number of trials, and p, the probability of success.

The probability mass function is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, \ldots, n$$

Binomial distribution III

Any binomial variable X can be represented as a sum of independent Bernoulli variables, $X \sim Bin(n, p)$,

$$X = X_1 + \cdots + X_n$$

so it is easy to compute

$$E[X] = E[X_1] + \cdots + E[X_n] = np$$

and

$$V[X] = V[X_1] + \cdots + V[X_n] = npq$$

Binomial distribution IV

Example 4

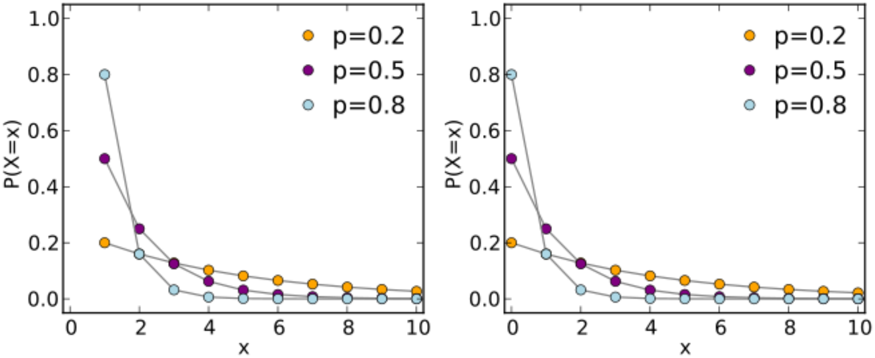
As part of a business strategy, randomly selected 20% of new internet service subscribers receive a special promotion from the provider. A group of 10 neighbors signs for the service. What is the probability that at least 4 of them get a special promotion?

Example 5

An exciting computer game is released. Sixty percent of players complete all the levels. Thirty percent of them will then buy an advanced version of the game. Among 15 users, what is the expected number of people who will buy the advanced version? What is the probability that at least two people will buy it?

Definition

The number of Bernoulli trials needed to get the first success has **Geometric distribution**.



Geometric distribution II

Example: 1) A search engine goes through a list of sites looking for a given key phrase. Suppose the search terminates as soon as the key phrase is found. The number of sites visited is Geometric.

2) A hiring manager interviews candidates, one by one, to fill a vacancy. The number of candidates interviewed until one candidate receives an offer has Geometric distribution.

Geometric distribution III

If X has a geometric distribution with parameter p (the probability of succes), then the values of X are: $1, 2, \ldots, n, \ldots$ (there is an infinity number of values).

The PMF is

$$P(X = k) = q^{k-1}p, k = 1, 2, \ldots$$

The expected value and variance:

$$E[X] = \sum_{k=1}^{\infty} kq^{k-1}p = \frac{1}{p}$$

$$V[X] = \frac{1-p}{p^2}.$$

Geometric distribution IV

Example 6a.

Products produced by a machine has a 3% defective rate. What is the probability that the first defective occurs in the fifth item inspected?

Example 6b.

Terminals on an on-line computer system are attached to a communication line to the central computer system. The probability that any terminal is ready to transmit is 0.95. What is the probability that exactly 3 terminals need to be polled until the first ready one is found? What is the probability that we need to poll at least 3 terminals until the first ready one is located? In average, how many terminals need to be polled until the first ready terminal is located?

Geometric distribution V

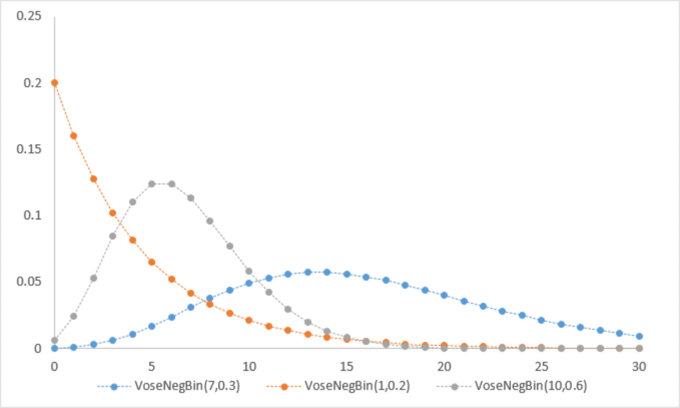
Example 6c.

A representative from the National Football League's Marketing Division randomly selects people on a random street in Kansas City, Kansas until he finds a person who attended the last home football game. Let p, the probability that he succeeds in finding such a person, equal 0.20. And, let X denote the number of people he selects until he finds his first success. What is the probability that the marketing representative must select 4 people before he finds one who attended the last home football game? What is the probability that the marketing representative must select more than 6 people before he finds one who attended the last home football game?

Negative binomial distribution I

Definition

In a sequence of independent Bernoulli trials, the number of trials needed to obtain k successes has **Negative Binomial distribution**.



Negative binomial distribution II

In some sense, Negative Binomial distribution is opposite to Binomial distribution. Binomial variables count the number of successes in a fixed number of trials whereas Negative Binomial variables count the number of trials needed to see a fixed number of successes. Other than this, there is nothing "negative" about this distribution.

The PMF is

$$P(X = x) = P(\text{the x-th trial results in the k-th success}) =$$

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

Negative binomial distribution III

A Negative Binomial variable $X \sim NB(k, p)$ can be represented as a sum of independent Geometric variables,

$$X = X_1 + \cdots + X_k,$$

withe the same probability of success p.
The expected value and variance of the geometric distribution:

$$E[X] = \frac{k}{p}$$

$$V[X] = \frac{k(1-p)}{p^2}$$

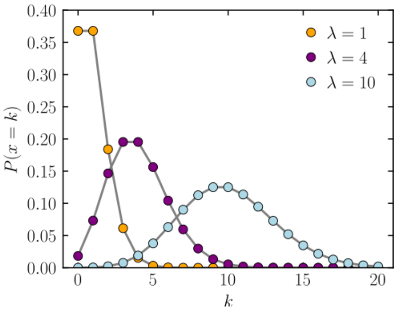
Negative binomial distribution IV

Example 7.

An archer hits a bull's-eye with the probability of 0.09 and the results of different attempts can be taken as independent of each other. (a) If the archer shoots a series of arrows, what is the probability that the first bull's-eye is scored with the fourth arrow? (b) What is the probability that the third bull's-eye is scored with the tenth arrow? (c) What is the expected number of the arrows shot before the first bull's-eye is scored? (d) What is the expected number of the arrows shot before the third bull's-eye is scored?

Poisson distribution I

The next distribution is related to a concept of **rare events**, or **Poissonian events**. Essentially it means that two such events are extremely unlikely to occur simultaneously or within a very short period of time. Arrivals of jobs, telephone calls, e-mail messages, traffic accidents, network blackouts, virus attacks, errors in software, floods, and earthquakes are examples of rare events.



Poisson distribution II

Definition

The number of rare events occurring within a fixed period of time has Poisson distribution.

The PMF of $X \sim Po(\lambda)$ is:

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

The expected value and variance:

$$E[X] = V[X] = \lambda$$

Poisson distribution III

Example 8.

(New accounts). Customers of an internet service provider initiate new accounts at the average rate of 10 accounts per day. (a) What is the probability that more than 8 new accounts will be initiated today? (b) What is the probability that more than 16 accounts will be initiated within 2 days?

Poisson approximation of Binomial distribution I

Poisson distribution can be effectively used to approximate Binomial probabilities when the number of trials n is large, and the probability of success p is small. Such an approximation is adequate, say, for $n \geq 30$ and $p \leq 0.05$, and it becomes more accurate for larger n.

$$Binomial(n, p) \approx Poisson(\lambda),$$

where $n \geq 30, p \leq 0.05, np = \lambda$, also the PMF of both distributions are close

$$\lim_{n\to\infty, p\to 0, np\to\lambda} \binom{x}{n} p^x (1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!}$$

Poisson approximation of Binomial distribution II

When p is large ($p \geq 0.95$), the Poisson approximation is applicable too. The probability of a failure $q = 1 - p$ is small in this case. Then, we can approximate the number of failures, which is also Binomial, by a Poisson distribution.

Example 9.

Ninety-seven percent of electronic messages are transmitted with no error. What is the probability that out of 200 messages, at least 195 will be transmitted correctly?

The End