Basic Probability Topics

Moshe Zukerman

Electronic Engineering Department
City University of Hong Kong
Hong Kong SAR, PRC

Text/Reference Books

Moshe Zukerman, Introduction to Queueing Theory and Stochastic Teletraffic Models (Chapter 1)
http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf

D. Bertsekas and J. N. Tsitsiklis, Introduction to Probability, Athena Scientific, Belmont, Massachusetts 2002.

S. M. Ross, A first course in probability, Macmillan, New York, 1976.

Events, Sample Space, and Random Variables

- Consider an <u>experiment</u> (e.g. tossing a coin, or rolling a die).
- <u>Sample space</u> - set of all possible outcomes.
- <u>Event</u> - a subset of the sample space.
- <u>example: experiment consisting of rolling a die once.</u>

Sample space = {1, 2, 3, 4, 5, 6}

Possible events:

- {2, 3},
- {6},
- empty set {} (often denoted by Φ)
- the entire sample space {1, 2, 3, 4, 5, 6}

Question

Consider the experiment to be tossing a coin. What is the Sample Space? What are the events associated with this Sample Space?

Question

Consider the experiment to be tossing a coin. What is the Sample Space? What are the events associated with this Sample Space?

Answer

Notice that although the sample space includes only the outcome of the experiments which are Head (H) and Tail (T), the events associated with this samples space includes all subsets of the state space which include also the empty set which in this case is the event $\{H \cap T\}$ and the entire sample space which in this case is the event $\{H \cup T\}$.

Events are called <u>mutually exclusive</u> if their intersection is the empty set.

A set of events is <u>exhaustive</u> if its union is equal to the sample space.

<u>Example 1:</u> tossing a coin only once

The events {H} (Head) and {T} (Tail) are both mutually exclusive and exhaustive.
What is the state space (the set) of all possible events in this case?

<u>Example 2:</u> rolling a die only once
The events {1}, {2}, {3}, {4}, {5}, and {6} are both mutually exclusive and exhaustive.
The events {4}, {5}, and {6} are mutually exclusive but are not exhaustive.

A random variable is a real valued function defined on the sample space.

This function $X = X(\omega)$ assigns a number to each outcome ω of the experiment.

Example: tossing a coin experiment
X = 1 for Head {H}
X = 0 for Tail {T}

Note that the function X is deterministic (not random), but the ω is unknown before the experiment is performed. Therefore $X(\omega)$ is called a random variable.

If X is a random variable then $Y = g(X)$ for some function g is also a random variable.

Examples:

$Y = cX$ for some scalar c is a random variable.

$Y = X^n$ for some integer n is a random variable.

If $X_1, \ X_2, \ X_3, \ \ldots, \ X_n$ is a sequence of random variables, then
$Y = \sum_{i=1}^{n} X_i$ is also a random variable.

Probability, Conditional Probability and Independence

Consider a sample space *S*. Let *A* be a subset of *S*. The probability of *A* is the function on *S* and all its subsets, denoted *P(A)* that satisfies the following three axioms:

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. The probability of the union of mutually exclusive events is equal to the sum of the probabilities of these events.

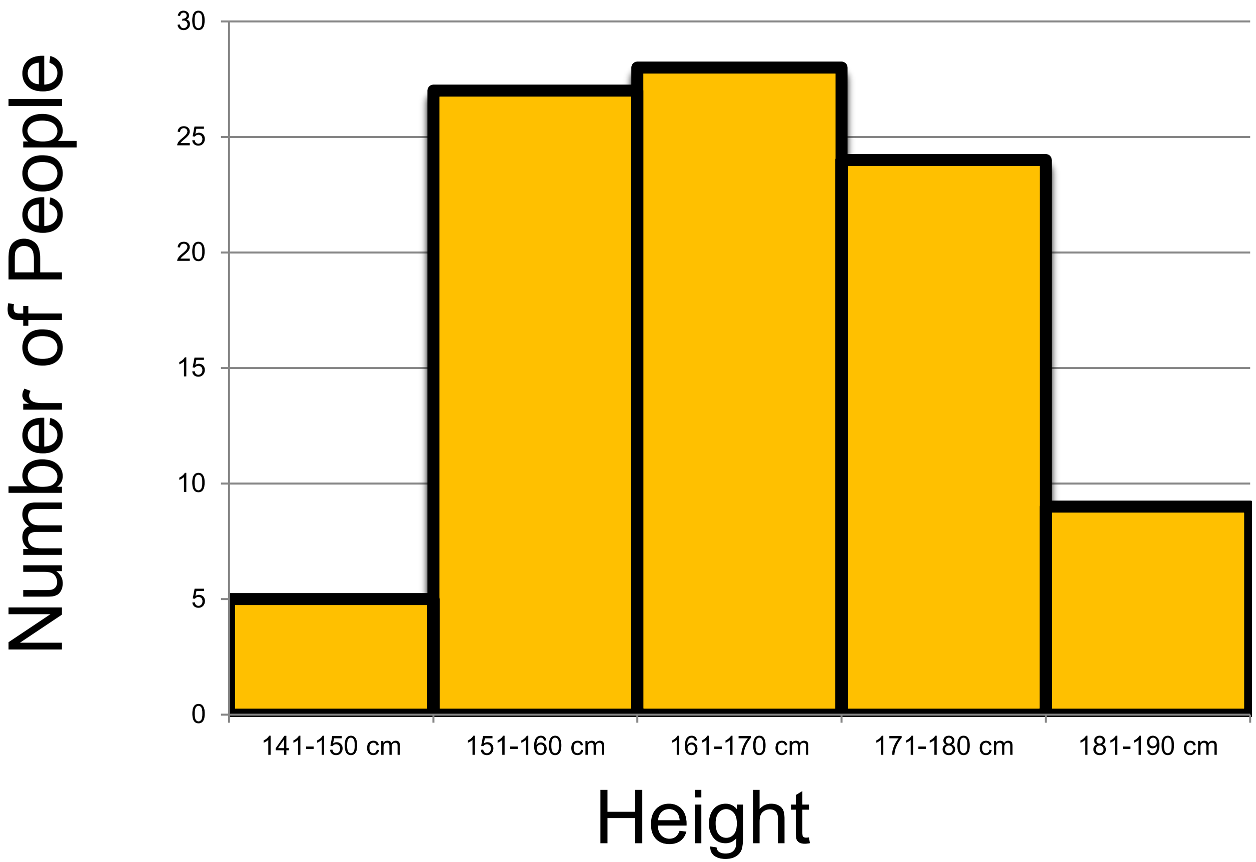
Questions

Question 1

Consider again the experiment to be tossing a coin. Assume that $P(H) = P(T) = 0.5$. Illustrate each of the Probability Axioms for this case.

One intuitive interpretation of probability of an event is its **limiting relative frequency**

Let an outcome of an experiment be a person height. Let p_i be the probability that a person's height is i cm. Consider a sample of N people. Each one of them reports his/her height, rounded to the nearest cm. Let n_i be the number of people reported a height of i cm. These n_i values can be graphically presented as what is known as a **histogram**. An example of a histogram is shown in the following graph.



An example of a Histogram with 5 ranges (bins) and each range is 10 cm.

In every range (bin), 10 n_i values are added up.

In this case, N = 93 people.

Limiting Relative Frequency (continued)

The relative frequency n_i/N approximates p_i.

This approximation becomes more and more accurate as N increases.

This approximation is consistent with the requirement

$$\sum_i p_i = 1.$$

If we set $p_i = n_i/N$, then since $\sum_i n_i = N$, we obtain

$$\sum_i p_i = 1.$$

The Average Height

$$\text{The average height} = \frac{\sum_i i n_i}{N}.$$

For large N, we set $p_i = n_i/N$, then,

$$\text{The average height} = \sum_i i p_i.$$

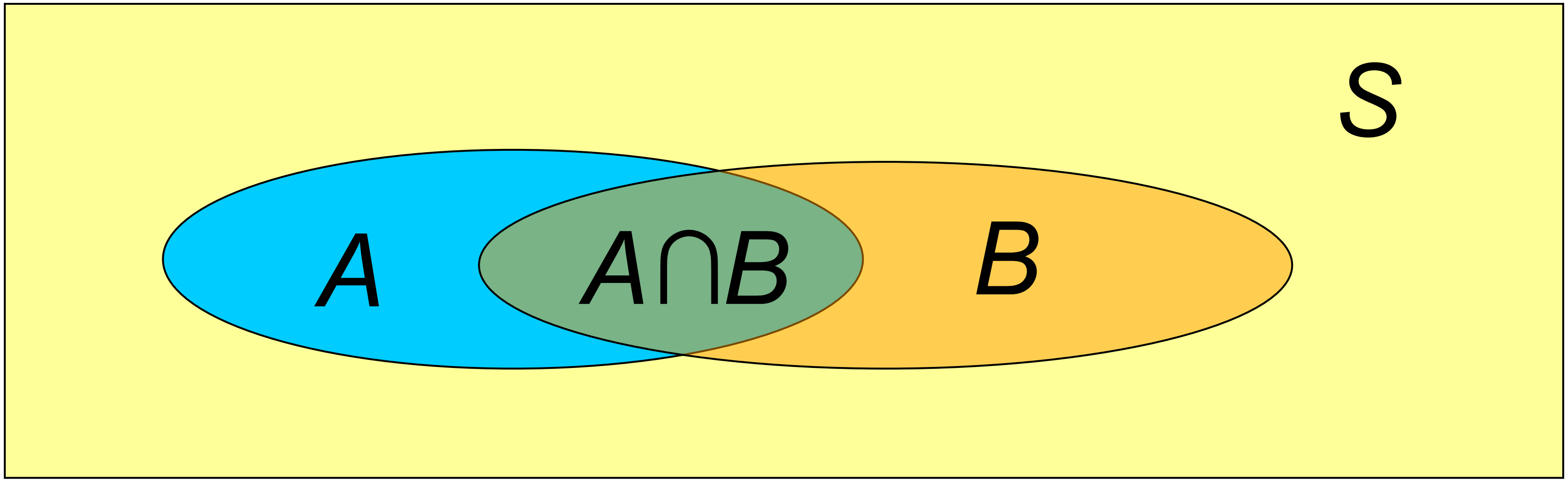
This is related to the well-known

Law of Large Numbers

Conditional Probability

The conditional probability of event A given event B is denoted by

$$P(A \mid B)$$



"Given event *B*" is equivalent to "*B* becomes the sample space".

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example: consider rolling a die and *B*={1,2,3} (*B* = outcome is 1 or 2 or 3), and *A*={1}, then

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Now, since

$$P(A \cap B) = P(B \cap A) = P(B \mid A)P(A)$$

we obtain

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Events *A* and *B* are said to be **independent** if and only if

$$P(A \mid B) = P(A).$$

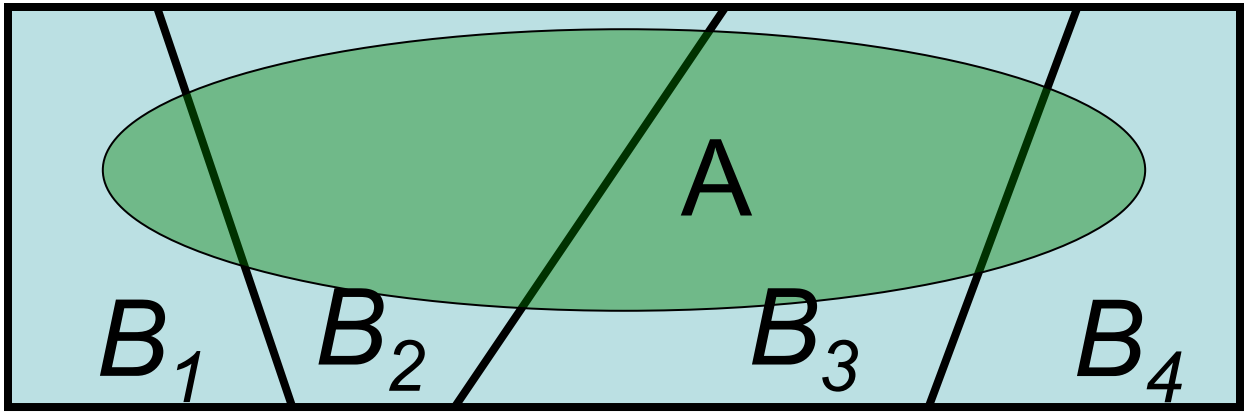
Equivalent definitions are:

$$P(A \cap B) = P(A)P(B)$$

$$P(B \mid A) = P(B).$$

Independence between two events means that if one of them occurs, the probability of the other to occur is not affected.

Homework: Show the equivalence between these three relationships.



$B_1, B_2, B_3, \ldots, B_n$ are mutually exclusive and exhaustive events in S. A is another event in S. Then, $A = \bigcup_{i=1}^{n} (A \cap B_i)$

Notice that $(A \cap B_1), (A \cap B_2), \cdots, (A \cap B_n)$ are also mutually exclusive (but not exhaustive).

Considering the 3rd probability axiom, we obtain:

$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$,

Recall, $P(A \cap B) = P(A \mid B)P(B)$,

law of total probability:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \times P(B_i)$$

Bayes' theorem:

$$P(B_1 \mid A) = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{n} P(A|B_i)\times P(B_i)}.$$

$P(B_1 \mid A)$ is the posterior probability of B_1.

$P(B_1)$ is the prior probability of B_1.

Other names for Bayes’ Theorem: Bayes' law and Bayes' rule

Homework: Make sure you know how to derive the law of total probability and Bayes’ theorem.

Question 2

Now consider an experiment involving three coin tosses. The outcome of the experiment is now a 3-long string of Heads and Tails. Assume that all coin tosses have probability 0.5, and that the coin tosses are independent events.

1. Write the sample space where each outcome of the experiment is an ordered 3-long string of Heads and Tails.

2. What is the probability of each outcome?

3. Consider the event
$A = \{Exactly\ one\ head\ occurs\}$.
Find $P(A)$ using the additivity axiom.

Partial Answer: $P(A) = 1/8 + 1/8 + 1/8 = 3/8$.

Question 3

Now consider again three coin tosses. Find the probability $P(A \mid B)$ where A and B are the events:

$A =$ more than one head came up

$B =$ 1st toss is a head.

Guide:

$P(B) = 4/8$; $P(A \cap B) = 3/8$;

$P(A \mid B) = (3/8)/(4/8) = 3/4.$

Question 4

Consider a medical test for a certain disease. The medical test detects the disease with probability 0.99 and fails to detect the disease with probability 0.01. If the disease is not present, the test indicates that it is present with probability 0.02 and that it is not present with probability 0.98. Consider two cases:

Case a: The test is done on a randomly chosen person from the population where the occurrence of the disease is 1/10000.

Case b: The test is done on patients that are referred by a doctor that have a prior probability (before they do the test) of 0.3 to have the disease.

Find the probability of a person to have the disease if the test shows positive outcome in each of these cases.

Guide:

A = person has the disease.
B = test is positive.
$\bar{A}$ = person does not have the disease.
$\bar{B}$ = test is negative.
We need to find $P(A \mid B)$.

Case a:

We know: $P(A) = 0.0001$.
$P(\bar{A}) = 0.9999$.
$P(B \mid A) = 0.99$.
$P(B \mid \bar{A}) = 0.02$. By the law of total probability:
$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})$.
$P(B) = 0.99 \times 0.0001 + 0.02 \times 0.9999 = 0.020097$.
Now put it all together and use Bayes' Theorem to obtain:
$P(A \mid B) = 0.004926108$.

Case b:

$P(A) = 0.3$.
Repeat the previous derivations to show that for this case $P(A \mid B) = 0.954983923$.

Question 5

In a multiple choice exam, there are 4 answers to a question. A student knows the right answer with probability 0.8 (Case 1), with probability 0.2 (Case 2), and with probability 0.5 (Case 3). If the student knows the answer s/he answers correctly with probability 1. However, if the student does not know the answer s/he always guesses with probability of success being 0.25. Given that the student marked the right answer, what is the probability he/she knows the answer.

Guide:

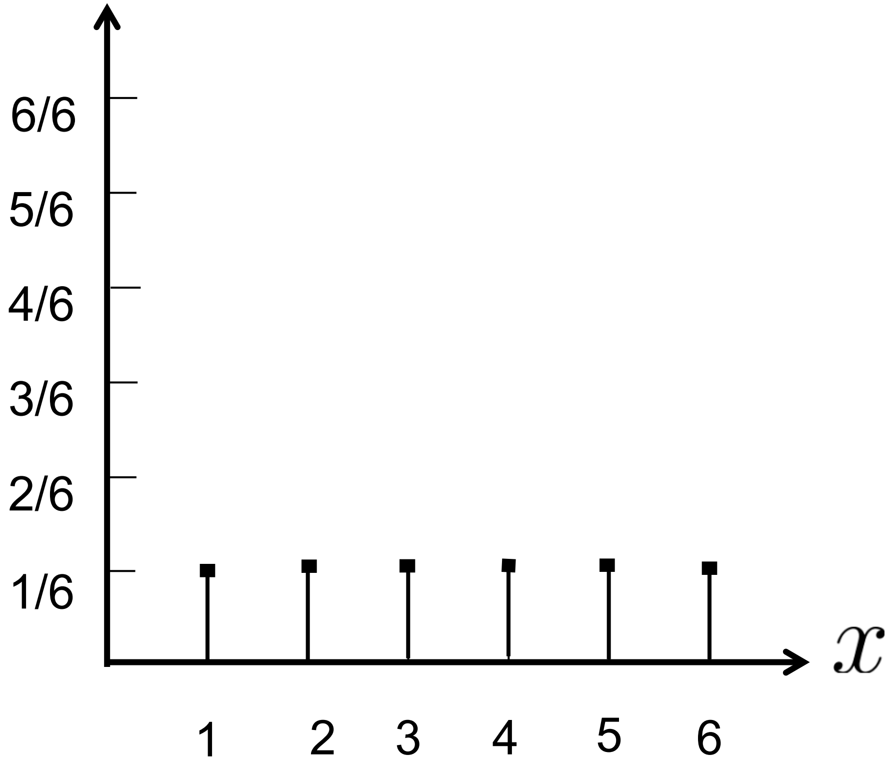
A = Student knows the answer.
B = Student marks correctly.
$\bar{A}$ = Student does not know the answer.
$\bar{B}$ = Student marks incorrectly.
We need to find $P(A \mid B)$.
Case 1: We know: $P(A) = 0.8$. $P(\bar{A}) = 0.2$. $P(B \mid A) = 1$. $P(B \mid \bar{A}) = 0.25$. By the law of total probability:
$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})$.
$P(B) = 1 \times 0.8 + 0.25 \times 0.2 = 0.85$.
Now put it all together and by Bayes' Theorem obtain:
$P(A \mid B) = 0.941176471$.
Case 2: Repeat the previous derivations to obtain:
$P(A) = 0.2$. $P(B) = 0.4$. $P(A \mid B) = 0.5$.
Case 3: Repeat the previous derivations to obtain:
$P(A) = 0.5$. $P(B) = 0.625$. $P(A \mid B) = 0.8$.

Probability and Distribution Functions

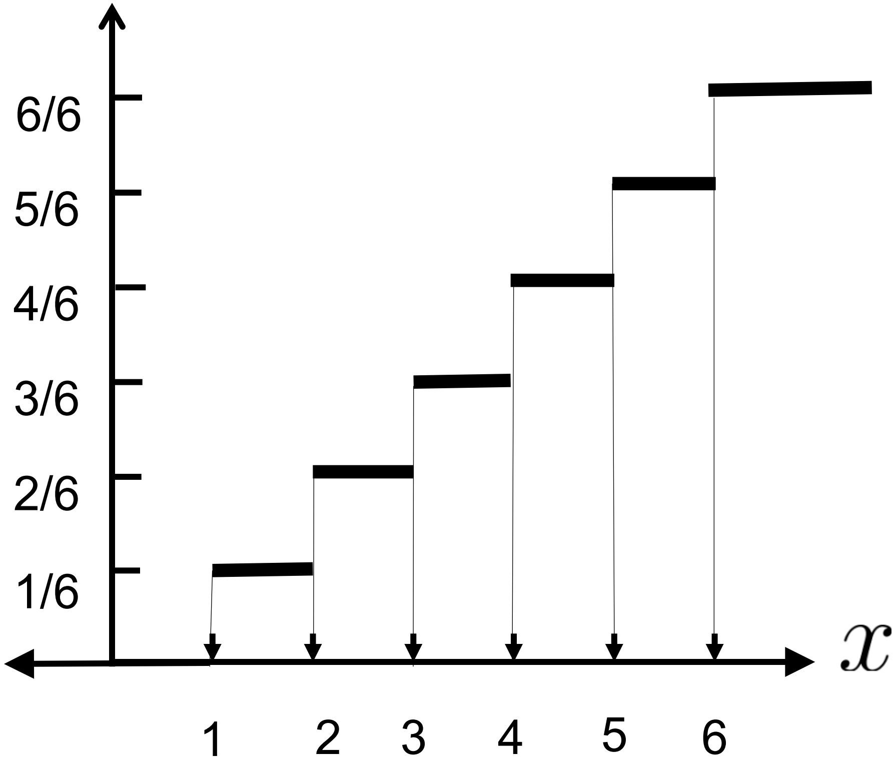
X is a random variable (r.v.).
x is a number that represents an outcome of an experiment.
$\{X = x\}$ is an event.
$P_X(x) = P(X = x)$ is a probability function.
Other names: probability distribution function, probability mass function.
$F_X(x) = P(X \leq x)$ is the cumulative distribution function (CDF) of r.v. X.

Rolling a die

$$P_X(x) = P(X = x)$$



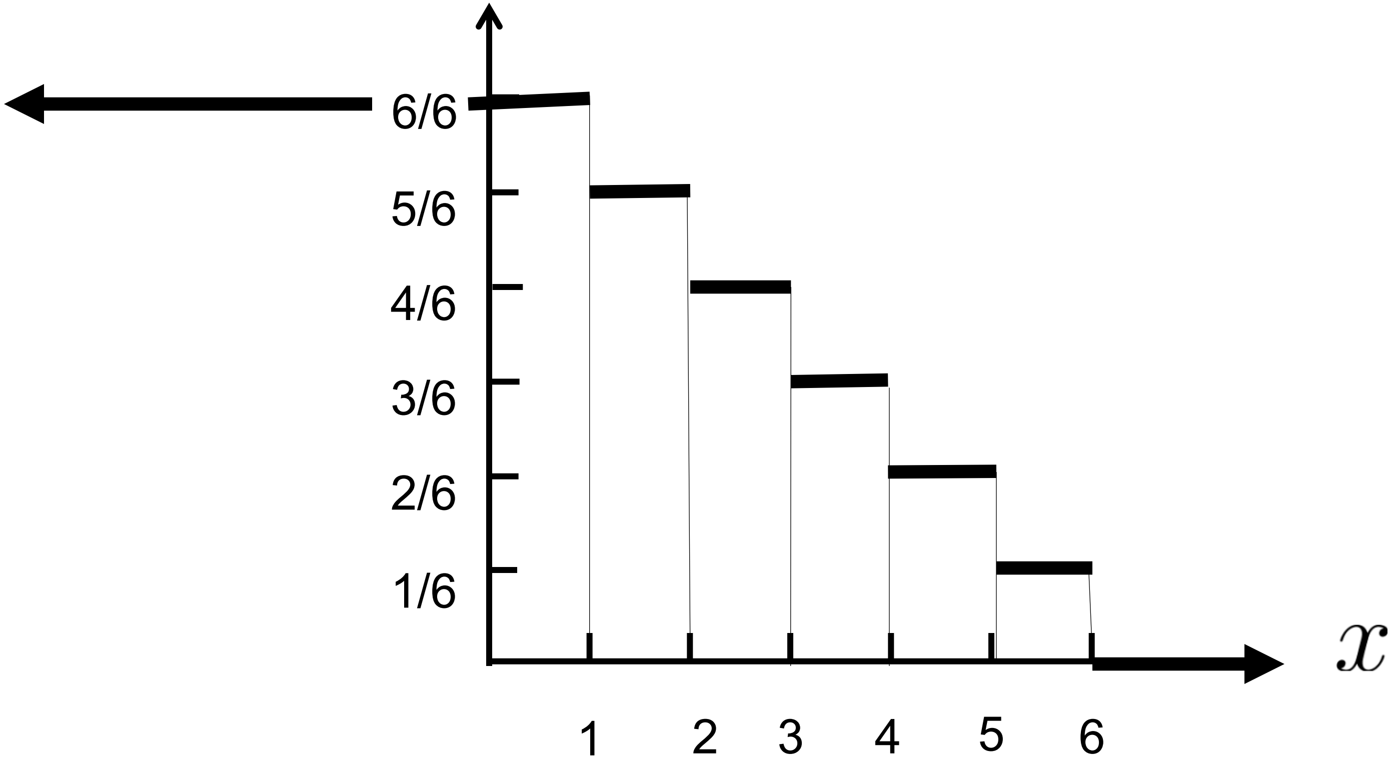
$$F_X(x) = P(X \le x)$$



$\bar{F}_X(x) = P(X > x)$ is the **complementary distribution function** of random variable X

For the case of Rolling a die

$$\bar{F}_X(x) = P(X > x)$$



For every r.v. X and any $x \in R$,

$$F_X(x) + \bar{F}_X(x) = 1.$$

Notation: $A, B = A \cap B$

$$F_{X_1,\ ...,X_n}(x_1,\ ..., x_n) = P(X_1 \leq x_1,\ \ldots,\ X_n \leq x_n).$$

is the joint distribution function of the random variables $X_1,\ X_2,\ \ldots,\ X_n$.

Then

$$F_{X_1}(x_1) = F_{X_1,\ X_2,\ ...,\ X_n}(x_1,\ \infty,\ \ldots,\ \infty).$$

A random variable is called ***discrete*** if it takes at most a countable number of possible values.

A ***continuous*** random variable takes an uncountable number of possible values.

For discrete random variables $X_1,\ X_2,\ ..., X_n$ the joint probability function is:

$$P_{X_1,\ X_2,\ ...,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n) = P(X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n)$$

and the probability function of a single discrete random variable is:

$$P_{X_1}(x_1) = \textstyle\sum_{x_2} \cdots \sum_{x_n} P_{X_1,\ X_2,\ ...,\ X_n}(x_1,\ x_2,\ \ldots,\ x_n).$$

Conditional Probability for Discrete Random Variables

$$P_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \tfrac{P(X=x,Y=y)}{P(Y=y)} = \tfrac{P_{X,Y}(x,y)}{P_Y(y)}.$$

Because of the above and since $P_Y(y) = \sum_x P_{X,Y}(x, y),$

we obtain $\sum_x P_{X|Y}(x \mid y) = 1.$

The implication is that the event {*Y*=*y*} is the new sample space and *X* has a legitimate distribution function in this new sample space.

$$P_Y(y) = \sum_x P_{X,Y}(x, y) = \sum_x P_{Y|X}(y \mid x)P_X(x)$$

is another version of the law of total probability.

Independence between Random Variables

"random variables U and V are independent" is equivalent to:
"the events $U = u$ and $V = v$ are independent for every u and v."
Accordingly, random variables U and V are independent if and only if:
$P_{U,V}(u, v) = P_U(u)P_V(v) \quad$ for all u, v.

An equivalent definition of independence between U and V is $P_{U|V}(u \mid v) = P_U(u)$ for all u, v.

Example

You roll a fair 6-side die twice. X is a result of the first roll and Y is the result of the second roll. Define U = max(X,Y) and V = min(X,Y).

Find: P(U=5|V=3)

Convolution

Consider independent random variables V_1 and V_2 with probability functions $P_{V_1}(v_1)$ and $P_{V_2}(v_2)$, respectively.

Let $V = V_1 + V_2$.

The **<u>convolution</u>** of $P_{V_1}(v_1)$ and $P_{V_2}(v_2)$ is

$$\begin{aligned} P_V(v) &= P(V_1 + V_2 = v) \\ &= \sum_{v_1} P(V_1 = v_1, V_2 = v - v_1) \\ &= \sum_{v_1} P_{V_1}(v_1) P_{V_2}(v - v_1). \end{aligned}$$

Question

Explain the last equation of convolution using the Law of Total Probability.

Guide

The factor $P_{V_2}(v - v_1)$ represents $P(V = v \mid V_1 = v_1)$ (conditioning) and the second is $P_{V_1}(v_1)$ (unconditioning).

Now consider *k* random variables $X_i,\ i = 1, 2, 3,\ \ldots,\ k.$

Let $P_{X_i}(x_i)$ be the probability function of X_i

$$Y = \sum_{i=1}^{k} X_i$$

The convolution of the *k* probability functions is:

$$P_Y(y) = \sum_{x_2,\ x_3,\ \ldots,\ x_k:\ x_2+x_3+\ \ldots\ +x_k \leq y} \left(P_{X_1}(y - \Sigma_{i=2}^{k} x_i) \prod_{i=2}^{k} P_{X_i}(x_i) \right).$$

If all the X_i are independent and identically distributed (IID) random variables, with probability function $P_{X_1}(x)$, then $P_Y(y)$ is called the *k-fold convolution* of $P_{X_1}(x)$.

Some discrete random variables

1. Bernoulli (with parameter p)

$$\begin{aligned} P(X=1) &= p && \text{“success”} \\ P(X=0) &= 1-p && \text{“failure”} \end{aligned}$$

2. Geometric (with parameter p)

The number of independent Bernoulli trials until the first success

$$P(X=i)=(1-p)^{i-1}p \quad \text{for} \;\; i=1,2,3, \;\; \ldots$$

$P(X>i)=(1-p)^i$ for $i=0,1,2, \;\ldots\;,$ and $P(X>i)=1$ for $i<0$.

Geometric random variable is memoryless.

$$P(X>m+n \mid X>m)=P(X>n), \;\; m=0,1,2,\ldots, \;\; n=0,1,2,\ldots$$

It is the ONLY memoryless discrete random variable.

3. Binomial (with parameters p and n)

The number of successes in *n* independent Bernoulli trials

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \qquad i = 0,\ 1,\ 2,\ \ldots,\ n.$$

Can be used to model users activity.

A user is active with probability *p* and non-active with probability 1-*p*.

$X = i$ is the event where i users are active.

4. Poisson (with parameter λ)

$$P(X=i)=e^{-\lambda}\frac{\lambda^i}{i!} \qquad i=0,\ 1,\ 2,\ 3,\ \ldots$$

How to compute these values?

Use Recursion and start from values around λ.

Set arbitrary initial value then normalize.

Poisson-Binomial Relationship

Consider a sequence of binomial random variables $X_n, \quad n = 1, 2, \ldots$ with parameters (n, p) where $\lambda = np$, or $p = \lambda/n$. Then the probability function

$$\lim_{n\to\infty} P(X_n = k)$$

is a Poisson probability function with parameter λ.

$\Rightarrow$ Poisson can be used to model traffic from a large number of sources.

To prove this we write:

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k}.$$

Substituting $p = \lambda/n$, we obtain

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

or

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \frac{n!}{(n-k)!n^k} \left(\frac{\lambda^k}{k!}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.$$

Now notice that

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda},$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}=1,$$

and

$$\lim_{n\to\infty}\frac{n!}{(n-k)!n^k}=1.$$

Therefore,

$$\lim_{n\to\infty}P(X_n=k)=\frac{\lambda^k e^{-\lambda}}{k!}.$$

QED.

Sum of two Poisson Random variables

Let $Y = X_1 + X_2$ where X_1 and X_2 are two independent Poisson random variables with parameters λ_1 and λ_2, respectively.

Use convolution to show that Y follows a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

$$\begin{aligned}
P_Y(k) &= P(X_1 + X_2 = k) \\
&= \sum_{i=0}^{k} P(\{X_1 = i\} \cap \{X_2 = k - i\}) \\
&= \sum_{i=0}^{k} P_{X_1}(i) P_{X_2}(k - i) \\
&= \sum_{i=0}^{k} \frac{\lambda_1^i}{i!} e^{-\lambda_1} \frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_2} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{k} \frac{\lambda_1^i \lambda_2^{k-i}}{i!(k-i)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{i=0}^{k} \frac{k! \lambda_1^i \lambda_2^{k-i}}{i!(k-i)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda_1^i \lambda_2^{k-i} \\
&= \frac{e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^k}{k!}. \quad \textbf{QED}
\end{aligned}$$

5. Pascal $(\text{with parameters } k \geq 1, p \in (0,1])$

The number of independent Bernoulli trials until the kth success, or equivalently, sum of k geometric random variables.

For Pascal random variable X, the event $\{X = i\}$ requires a "success" in the ith trial, and $k-1$ "successes" in $i - 1$ trials. These two events are independent. The first is Bernoulli and the second is Binomial. Therefore,

$$P(X = i) = \binom{i-1}{k-1} p^k (1-p)^{i-k} \qquad i = k,\ k+1,\ k+2,\ \ldots\ .$$

6. Discrete Uniform

(with parameters a and b with $b > a$)

The discrete uniform probability function with integer parameters a and b has equal non-zero values for $x = a, a+1, a+2, \ldots, b$. It is given by

$$P_X(x) = \begin{cases} \frac{1}{b-a+1} & \text{if } x = a, a+1, a+2, \ldots, b \\ 0 & \text{otherwise.} \end{cases}.$$

Rolling a fair die is one example that govern by this probability function with $a = 1$ and $b = 6$.

Continuous Random Variables and Distributions

Now the set of possible outcomes is uncountable.

A continuous random variable X which assigns a real number to any outcome of an experiment, is characterized by the existence of a function called the *probability density function* (or simply the *density*) of X defined for all $x \in R$, which has the property that for any set $A \subset R$,

$$P(X \in A) = \int_A f(x)dx.$$

To guarantee that all the relevant probabilities are nonnegative (recall the first probability axiom), we only consider nonnegative density functions.

Let $A = [a, b]$, we obtain,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

Notice that the probability of a continuous random variable taking a particular value is equal to zero. If we set $a = b$ in the above, we obtain

$$P(X = a) = \int_a^a f(x)dx = 0.$$

Therefore, for continuous random variable X,

$F_X(x) = P(X \leq x) = P(X < x)$,

and $P(X \geq x) = P(X > x)$.

We obtain

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)ds.$$

Hence, the probability density function $f(x)$ is the derivative of the distribution function $F_X(x)$.

Notation: $F_X(x) = F(x)$ and $f_X(x) = f(x)$

Let X and Y be two continuous random variables. The joint density of X and Y denoted $f_{X,Y}(x,y)$ is a nonnegative function that satisfies

$$P(\{X,Y\} \in A) = \iint_{\{X,Y\} \in A} f_{X,Y}(x,y)dxdy.$$

for any set $A \subset R^2$.

Equivalently to the discrete case:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

Let X and Y be two continuous random variables with joint density $f_{X,Y}(x,y)$. The conditional density of X given Y is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

For every given fixed y, it is a legitimate density because

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1.$$

By the definition of conditional density,

$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x \mid y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x \mid y)dy.$$

$$P(A) = P(X \in A) = \int_A f_X(x)dx = \int_A \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x \mid y)dydx.$$

$$P(A) = \int_{-\infty}^{\infty} f_Y(y) \int_A f_{X|Y}(x \mid y)dxdy$$

and therefore

$$P(A) = \int_{-\infty}^{\infty} f_Y(y) P(A \mid Y = y)dy$$

which is the continuous equivalence of the Law of Total Probability.

Example

Consider the following joint density:

$$f_{X,Y}(x,y) = \begin{cases} 2 & 0 \leq x+y \leq 1, \quad x \geq 0, \quad y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

1. Show that this is a legitimate density by showing first that all relevant probabilities are nonnegative and that the 2-dimensional integral of this joint density over the entire state space is equal to 1.
2. Derive the marginal density $f_Y(y)$.
3. Derive the conditional density $f_{X|Y}(x \mid y)$.

Please complete all steps in the following.

To show that this is a legitimate density observe that the joint density is nonnegative and also

$$\int_0^1 \int_0^{1-x} 2dydx = 1.$$

$$f_Y(y) = \begin{cases} \int_0^{1-y} f_{X,Y}(x,y)dx = 2 - 2y & 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$f_{X|Y}(x \mid y) = \begin{cases} \frac{2}{2-2y} = \frac{1}{1-y} & 0 \leq x \leq 1 - y \\ 0 & \text{otherwise.} \end{cases}$$

Convolution of continuous random variables

Consider independent random variables U and V. Let $X = U + V$.

$$f_X(x) = \int_u f_U(u) f_V(x-u).$$

The latter is the *convolution* of the densities $f_U(u)$ and $f_V(v)$.

Convolution of *k* continuous random variables

As in the discrete case the convolution $f_Y(y)$, of k densities $f_{X_i}(x_i)$, $i = 1, 2, 3, \ \ldots, \ k$, of random variables X_i, $i = 1, 2, 3, \ \ldots, \ k$, respectively, is given by

$$f_Y(y) = \iint_{x_2, \ ..., \ x_k: \ x_2+ \ ...,+x_k \leq y} \left(f_{X_1}(y - \Sigma_{i=2}^k x_i) \prod_{i=2}^k f_{X_i}(x_i) \right) .$$

And again, in the special case where all the random variable X_i, $i = 1, 2, 3, \ \ldots, \ k$, are IID, the density f_Y is the k-fold convolution of f_{X_1}.

Equivalence between discrete and continuous random variables and their probability functions/densities and distributions

Discrete	**Continuous**
$P_Y(y) = \sum_x P_{X,Y}(x,y)$	$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$
$P_{X\|Y}(x \mid y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}$	$f_{X\|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \times P(B_i)$	$P(A) = \int_{-\infty}^{\infty} f_Y(y)P(A \mid Y = y)dy$
If X and Y are independent $P_{X\|Y}(x \mid y) = P_X(x)$ $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ $P(V_1 + V_2 = v) = \sum_{v_1} P_{V_1}(v_1)P_{V_2}(v - v_1)$	**If X and Y are independent** $f_{X\|Y}(x \mid y) = f_X(x)$ $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ $P(U + V = x) = \int_u f_U(u)f_V(x - u)du$

Some Continuous Random Variables

1. Uniform (with parameters *a,b*)

Its probability density function is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

A special case - uniform (0,1).
Its probability density function is

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Inverse transform sampling

Using uniform (0,1) deviates to generate sequence of random deviates of any distribution

For any uniform (0,1) deviate $U(0,1)$ and any CDF $F(x)$, set $U(0,1) = F(x*)$, so $x^* = F^{-1}(U(0,1))$ is the corresponding random deviate from $F(x)$.

Why does it work? Let U be a uniform (0,1) random variable. Let $F(x)$ be an arbitrary CDF. Let $Y = F^{-1}(U)$. That is, $U = F(Y)$. Now,
$P(Y \leq x) = P[F^{-1}(U) \leq x] = P[U \leq F(x)]$.
Because U is a uniform (0,1) random variable, then
$P[U \leq F(x)] = F(x)$. Thus, $P(Y \leq x) = F(x)$. **QED**

Derive the convolution of two independent uniform (0,1) random variables.

$$f_X(x) = \int_u f_U(u) f_V(x-u) du$$

Since U and V are uniform(0,1) random variables, for $f_U(u)f_V(x-u)$ to be non-zero, u and x must satisfy:

$0 \leq u \leq 1$ and $0 \leq x - u \leq 1$,

or

$\max(0, x-1) \leq u \leq \min(1, x)$

and

$0 \leq x \leq 2$.

Therefore,

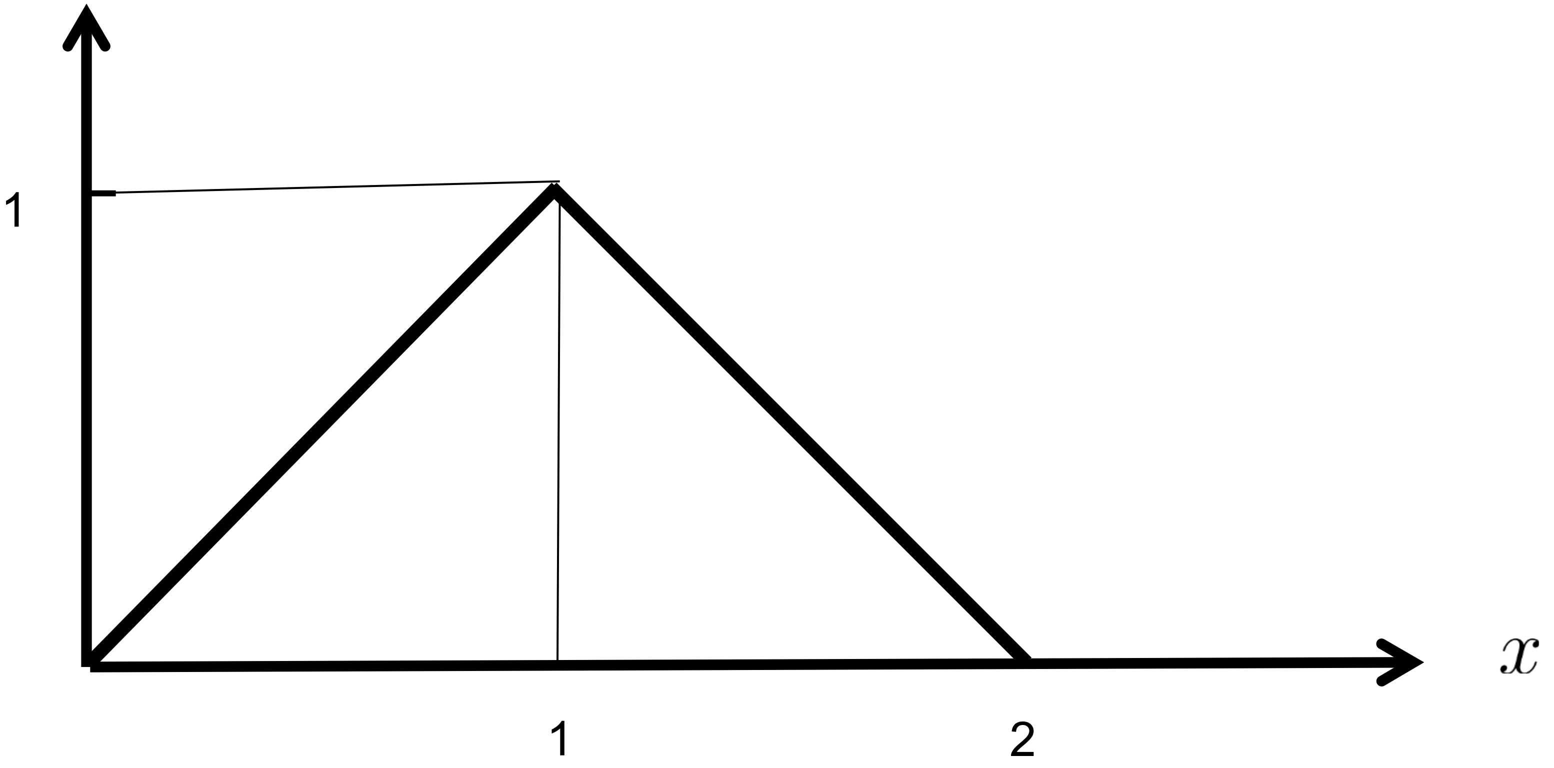
$$f_X(x) = \begin{cases} u\Big|_{\max(0,x-1)}^{\min(1,x)} & 0 \le x \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

or

$$f_X(x) = \begin{cases} \min(1,x) - \max(0,x-1) & 0 \le x \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

Convolution of two independent uniform (0,1) random variables

$$f_X(x)$$



2. Exponential (with parameter *μ*)

Density function:

$$f(x) = \begin{cases} \mu e^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Cumulative Distribution Function (CDF):

$$F(x) = \begin{cases} \int_0^x \mu e^{-\mu s} ds = 1 - e^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Complementary Distribution Function:

$$\bar{F}(x) = 1 - F(x) = \begin{cases} e^{-\mu x} & \text{if } x \geq 0 \\ 1 & \text{otherwise.} \end{cases}$$

Example

Show how to apply the Inverse transform sampling to generate exponential deviates.

Guide

Due to symmetry, we can use the complementary distribution function instead of the cumulative distribution function. Set $U(0,1) = \bar{F}(x*) = e^{-\lambda x^*}$, and obtain

$$\ln U(0,1) = -\lambda x^*$$

or

$$x^* = -\frac{\ln U(0,1)}{\lambda}.$$

Memorylessness

A continuous random variable is called memoryless if for any $t \geq 0$ and $s \geq 0$, $P(X > s+t \mid X > t) = P(X > s)$. The following proves that the exponential random variable is memoryless.

$$\begin{aligned}
P(X > s+t \mid X > t) &= \frac{P(X > s+t \cap X > t)}{P(X > t)} \\
&= \frac{P(X > s+t)}{P(X > t)} \\
&= \frac{e^{-\mu(s+t)}}{e^{-\mu t}} \\
&= e^{-\mu s} = P(X > s).
\end{aligned}$$

The exponential random variable is the ONLY memoryless continuous random variable. (Recall that the Geometric is the ONLY memoryless discrete random variable.)

The minimum of two independent exponential random variables

Let X_1 and X_2 be independent and exponentially distributed random variables with parameters λ_1 and λ_2. Let $X = \min[X_1, X_2]$. Then

$$\begin{aligned} P(X > t) &= P(\min[X_1, X_2] > t) \\ &= P(X_1 > t, X_2 > t) \\ &= e^{-\lambda_1 t} e^{-\lambda_2 t} \\ &= e^{-(\lambda_1 + \lambda_2)t}. \end{aligned}$$

Thus, the distribution of X is exponential with parameter $\lambda_1 + \lambda_2$.

Competition between two independent exponential random Variables

Let X_1 and X_2 be independent and exponentially distributed random variables with parameters λ_1 and λ_2. What is the probability of $X_1 < X_2$?

By the continuous version of the law of total probability,

$$\begin{aligned} P(X_1 < X_2) &= \int_0^\infty (1 - e^{-\lambda_1 t})\lambda_2 e^{-\lambda_2 t} dt \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2}. \end{aligned}$$

To understand the latter, note that X_2 can obtain many values: $t_1, t_2, t_3, \ldots$, infinitely many values ...

All these values, that X_2 may take, lead to the events $X_2 = t_1, X_2 = t_2, X_2 = t_3, \ldots$ that are mutually exclusive and exhaustive.

Then, using the continuous version of the Law of Total Probability, namely, integration of the product $P(X_1 < t)$ times the density of X_2 at t, will give us the probability of $X_1 < X_2$.
By integrating over all t we "add up" the probabilities of infinitely many mutually exclusive and exhaustive events that make up the event $X_1 < X_2$. And this is exactly what the Law of Total Probability does!

Relationship between the memoryless random variables: Exponential and Geometric

Let X_{exp} and X_{geo} be exponential and geometric random variables with parameters λ and p, respectively. Let δ be an "interval" size used to discretize the continuous values that X_{exp} takes, and we are interested to find δ such that

$$F_{X_{exp}}(n\delta) = F_{X_{geo}}(n), \quad n = 1, 2, 3, \ldots.$$

To this end we consider the complementary distributions and aim to find δ that satisfies $P(X_{exp} > n\delta) = P(X_{geo} > n)$, or $e^{-\lambda n\delta} = (1-p)^n$, or $e^{-\lambda\delta} = 1 - p$, or $p = 1 - e^{-\lambda\delta}$.

We can observe that as the interval size δ approaches zero the probability of success p also approaches zero, and under these conditions the two distributions approach each other.

3. Hyper-Exponential

Let X_i for $i = 1, 2, 3, \ \ldots, \ k$ be k independent exponential random variables with parameters λ_i, $i = 1, 2, 3, \ \ldots, \ k$, respectively. Let p_i for $i = 1, 2, 3, \ \ldots, \ k$ be k nonnegative real numbers such that $\sum_{i=1}^{k} p_i = 1$. A random variable X that is equal to X_i with probability p_i is called Hyper-exponential. By the Law of total probability, its density is

$$f_X(x) = \sum_{i=1}^{k} p_i f_{X_i}(x).$$

4. Erlang (with parameters k and λ)

A random variable X has Erlang distribution with parameters λ (positive real) and k (positive integer) if its density is given by

$$f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}. \tag{1}$$

Let X_i, $i = 1, 2, \ldots, k$, be k independent exponentially distributed random variables each with parameter λ, prove by induction that the random variable X defined by the sum $X = \sum_{i=1}^{k} X_i$ has Erlang distribution with parameters k and λ.
Complete the other homework problems in the textbook on Erlang random variable.

5. Hypo-Exponential

Let X_i, $i = 1,\ 2,\ \ldots,\ k$ be k independent exponentially distributed random variables each with parameters λ_i, respectively.

The random variable X defined by the sum $X = \sum_{i=1}^{k} X_i$ is called hypo-exponential. In other words, the density of X is a convolution of the k densities $\lambda_i e^{-\lambda_i x}$, $i = 1,\ 2,\ \ldots,\ k$.

The Erlang distribution is a special case of hypo-exponential when all the k random variables are identically distributed.

6. Gaussian (with parameters m and σ^2)

The Gaussian random variable X with parameters m and σ^2 has the following density.

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2} \qquad -\infty < x < \infty.$$

This density is symmetric and bell shaped.

Very widely used due to the **The central limit theorem** which says that the sum of a large number of independent random variables (not necessarily of the same distribution, but each has a finite variance) has Gaussian (normal) distribution.

7. Pareto (with parameters γ and δ)

It is useful in modelling lengths of data bursts in data and multimedia networks.

Its complementary distribution function is defined by

$$P(X > x) = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta \\ 1, & \text{otherwise.} \end{cases}$$

Here $\delta > 0$ is the scale parameter representing a minimum value for the random variable, and $\gamma > 0$ is the shape parameter of the Pareto distribution.

Mean

The **mean**, or the **expectation**, of a discrete random variable is defined by

$$E[X] = \sum_{\{n: P_X(n) > 0\}} n P_X(n).$$

This is related to the term **average** discussed before in the context of limiting relative frequency.

Equivalently, the mean of a continuous random variable is defined by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Example

Consider the following probability function:

$$P\left(X = x\right) = \begin{cases} 0.2 & x = 1 \\ 0.3 & x = 2 \\ 0.3 & x = 3 \\ 0.2 & x = 4 \end{cases}$$

Find the mean of X.

Solution:

$$E[X] = 1 \times 0.2 + 2 \times 0.3 + 3 \times 0.3 + 4 \times 0.2 = 2.5$$

As mentioned before, a function of a random variable is also a random variable.

Mean of a function of discrete random variables:

$$E[g(X)] = \sum_{\{k:P_X(k)>0\}} g(k)P_X(k)$$

Mean of a function of continuous random variables:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

If a and b are constants then for a random variable X (either discrete or continuous) we have:

$$E[aX] = aE[X],$$

$$E[X - b] = E[X] - b,$$

and

$$E[aX - b] = aE[X] - b.$$

For random variables $X_1, X_2, \ldots X_n$ (not necessarily independent) we have:

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i].$$

The n**th moment** of the random variable X is defined by $E[X^n]$. Substituting $g(X) = X^n$ in the definitions of $E[g(X)]$, the nth moment of X is given by:

$$E[X^n] = \sum_{\{k:P_X(k)>0\}} k^n P_X(k)$$

for a discrete random variable and

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx$$

for a continuous random variable.

The n**th central moment** of random variable X is defined by $E[(X-E[X])^n]$. Substituting $g(X) = (X-E[X])^n$ in the definitions of $E[g(X)]$, the nth central moment of X is given by:

$$E[(X-E[X])^n] = \sum_{\{k:P(k)>0\}} (k-E[X])^n P_X(k)$$

for a discrete random variable and

$$E[(X-E[X])^n] = \int_{-\infty}^{\infty} (x-E[X])^n f_X(x)dx$$

for a continuous random variable.

The mean is the first moment.

The variance is the second central moment defined by:

$$Var[X] = E[(X - E[X])^2].$$

The variance of a discrete random variable X is

$$Var[X] = \sum_{\{k:P(k)>0\}} (k - E[X])^2 P_X(k)$$

The variance of a continuous random variable X is

$$Var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx.$$

Another equation for the variance can be obtained by

$$Var[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] = E[X^2] - (E[X])^2.$$

Example

Consider again the following probability function:

$$P\left(X=x\right)=\begin{cases}0.2 & x=1\\0.3 & x=2\\0.3 & x=3\\0.2 & x=4\end{cases}$$

Find the variance of X.

Solution:

We already know that $E[X]=2.5$,
so $Var[X]=(1-2.5)^2\times 0.2+(2-2.5)^2\times 0.3+$
$(3-2.5)^2\times 0.3+(4-2.5)^2\times 0.2=1.05$

An Alternative Solution:

Now we solve the same problem using the formula
$Var[X] = E[X^2] - (E[X])^2$

We already know that
$(E[X])^2 = (2.5)^2 = 6.25,$
and
$E[X^2] = 1^2 \times 0.2 + 2^2 \times 0.3 + 3^2 \times 0.3 + 4^2 \times 0.2 = 7.3.$
Then,
$Var[X] = E[X^2] - (E[X]^2) = 7.3 - 6.25 = 1.05.$

If a is a constant then for a random variable X (either discrete or continuous) we have:

$$Var[aX] = a^2 Var[X],$$

If the random variables $X_1, X_2, X_3, \ldots, X_n$ are **independent**, then

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var[X_i].$$

The **standard deviation** of r.v. X is:

$$\sigma_X = \sqrt{Var[X]}.$$

When X is obvious, we use σ for the standard deviation. Hence the variance is often denoted by σ^2.

Covariance and Correlation

The **covariance** of two random variables X_1 and X_2, denoted by $cov(X_1, X_2)$, is defined by

$$cov(X_1, X_2) = E[(X_1 - E[X_1])(X_2 - E[X_2])].$$

The larger the dependency, the larger the covariance.

The variance of the sum of two random variables X_1 and X_2 is given by

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2] + 2cov(X_1, X_2).$$

If X_1 and X_2 are independent, then $cov(X_1, X_2) = 0$, so

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2].$$

Problem

Prove that $cov(X_1, X_2) = 0$ does not necessarily imply that X_1 and X_2 are independent.

Problem

Prove that $cov(X_1, X_2) = 0$ does not necessarily imply that X_1 and X_2 are independent.

Guide

The proof is by a counter example.

Consider two random variables X and Y and assume that both have Bernoulli distribution with parameter p.

Consider random variable X_1 defined by $X_1 = X + Y$ and another random variable X_2 defined by $X_2 = X - Y$. Show that $cov(X_1, X_2) = 0$ and that X_1 and X_2 are not independent.

The covariance can take any value between $-\infty$ and $+\infty$, and in some cases, it is convenient to have a normalized dependence measure - a measure that takes values between -1 and 1. Such measure is the **correlation**. Notice that the covariance is bounded by

$$cov(X_1, X_2) \leq \sqrt{Var[X_1]Var[X_2]},$$

the **correlation** of two random variables X and Y denoted by $corr(X, Y)$ is defined by

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

assuming $Var[X] \neq 0$ and $Var[Y] \neq 0$.

Homework

Consider an experiment of tossing a die with 6 sides. Assume that the die is fair, i.e., each side has the same probability (1/6) to occur. Consider a random variable X that takes the value i if the outcome of the toss is i, for $i = 1, 2, 3, \cdots, 6$.

1. Find $E[X]$, $Var[X]$ and σ_X.
2. Plot the probability function, cumulative distribution function and the complementary distribution function of X.

Some answers

$E[X] = 3.5$; $E[X^2] = 15.16666667$; $Var[X] = 2.916666667$; $\sigma_X = 1.707825128$.

Homework

Consider an exponential random variable with parameter λ. Derive its mean and Variance.

Guide

Find the mean by

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x} dx.$$

Use integration by parts to show that:

$$E[X] = -xe^{-\lambda x}\Big]_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Guide (continued)

Then use integration by parts to derive the second moment. Understand and verify the following derivations:

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\ &= -x^2 e^{-\lambda x}\Big]_0^\infty + 2\int_0^\infty x e^{-\lambda x} dx \\ &= \left(-x^2 e^{-\lambda x} - \frac{2}{\lambda} x e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right)\Bigg]_0^\infty \\ &= \frac{2}{\lambda^2}. \end{aligned}$$

$$Var[X] = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Sample Mean and Sample Variance

Consider a sample of n realizations of a random variable X, denoted $X(1), X(2), \ldots, X(n)$

The **Sample Mean**

$$S_m = \frac{\sum_{i=1}^{n} X(i)}{n}$$

is an estimator for the mean of X.

The **Sample Variance**

$$S_v = \frac{\sum_{i=1}^{n} [X(i) - S_m]^2}{n-1} \qquad n > 1$$

is an estimator for the variance of X.

The sample standard deviation is $\sqrt{S_v}$.

Homework

Consider the following data for students heights (in cm) taken from a sample of 10 students: 172, 178, 162, 167, 168, 175, 182, 161, 171, 170.

Compute the sample mean, the sample variance and the sample standard deviation.

Answers

Sample Mean = 170.6
Sample Variance = 43.6
Sample Standard Deviation = 6.603029608

Homework

Generate 10 deviates from an exponential distribution of a given mean and compute the Sample Mean and Sample Variance. Compare them with the real mean and variance. Then increase the sample to 100, 1000, . . . , 1,000,000. Observe the difference between the real mean and variance and the sample mean and variance. Repeat the experiment for a Pareto deviates of the same mean. Discuss differences.

Use your generated data of random deviates to compare between the exponential and Pareto densities and the equivalent histograms. Use small enough ranges to achieve good fit.

Conditional Expectation (or Mean)

$E[X \mid Y]$ denotes the conditional expectation of random variable X given the event $\{Y = y\}$ for each relevant value of y.

The conditional expectation of two discrete random variables is defined by

$$E[X \mid Y = j] = \sum_{i} iP(X = i \mid Y = j).$$

If X and Y are continuous, their conditional expectation is defined as

$$E[X \mid Y = y] = \int_{x=-\infty}^{\infty} x f_{X|Y}(x \mid y) dx.$$

$E[X \mid Y]$ is itself a random variable which is a function of the random variable Y. Therefore, for **discrete** random variables:

$$\begin{aligned}
E_Y[E[X \mid Y]] &= \sum_j E[X \mid Y = j] P(Y = j) \\
&= \sum_j \sum_i i P(X = i \mid Y = j) P(Y = j) \\
&= \sum_i i \sum_j P(X = i \mid Y = j) P(Y = j) \\
&= \sum_i i P(X = i) = E[X].
\end{aligned}$$

Thus,

$$E[X] = E_Y[E[X \mid Y]].$$

For **continuous** random variables:

$$
\begin{aligned}
E_Y[E[X \mid Y]] &= \int_{y=-\infty}^{\infty} E[X \mid Y = y] f_Y(y) dy \\
&= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} x f_{X|Y}(x \mid y) dx f_Y(y) dy \\
&= \int_{x=-\infty}^{\infty} x \int_{y=-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy dx \\
&= \int_{x=-\infty}^{\infty} x f_X(x) dx = E[X].
\end{aligned}
$$

Thus,

$$E[X] = E_Y[E[X \mid Y]].$$

Homework: Show that the latter holds also for the case where X is discrete and Y is continuous and vice versa.

Question

Let X be a geometric random variable. Use the concept of conditional expectation to derive $E[X]$.

Guide

Condition on the result of the first Bernoulli trial and obtain $E[X] = p(1) + (1-p)(1+E[X])$ and then solve it to obtain $E[X] = 1/p$.

Note that $P(X = x \mid Y = y)$ is itself a random variable that is a function of the values y taken by random variable Y.

Therefore, by definition

$$E_Y[P(X = x \mid Y = y)] = \sum_y P(X = x \mid Y = y)P(Y = y)$$

which lead to another way to express the Law of Total Probability:

$$P_X(x) = E_Y[P(X = x \mid Y = y)].$$

Henceforth, we normally ommit the subsript X or Y for mean, variance and probability notations.

Define the **Conditional Variance** as

$$Var[X \mid Y] = E[(X - E[X \mid Y])^2 \mid Y].$$

This gives rise to the following useful formula for the variance of a random variable known as EVVE:

$$Var[X] = E[Var[X \mid Y]] + Var[E[X \mid Y]].$$

Homework: Try to prove it, and if you are not successful, see my book or other sources.

Example

The number of Internet flows that arrive at a router per second is ϕ which has mean ϕ_e and variance ϕ_v. The number of packets in each flow is ς which has mean ς_e and variance ς_v. Assume that ϕ and ς are independent. The total number of packets arriving at the router per second is W which has mean W_e and variance W_v. Assume $W = \varsigma\phi$. To meet certain quality of service requirements, it is required that the router has the capacity to serves the arriving packets at the rate of $s_r = W_e + 4\sqrt{W_v}$ per second. Find s_r.

Solution

To compute s_r one needs to have the values of W_e and W_v. Because ϕ and ς are independent $E[W|\phi] = \phi\varsigma_e$ and therefore

$$W_e = E[W] = E[E[W|\phi]] = E[\phi]E[\varsigma] = \phi_e\varsigma_e.$$

By EVVE,

$$Var[W] = E[Var[W|\phi]] + Var[E[W|\phi]] = \varsigma_v E[\phi^2] + (\varsigma_e)^2 Var[\phi].$$

Therefore

$$W_v = \phi_v\varsigma_v + \varsigma_v\phi_e^2 + \phi_v\varsigma_e^2.$$

The following table provides the mean and the variance of some of the above-mentioned random variables.

r.v.	parameters	mean	variance
Bernoulli	$0 \leq p \leq 1$	p	$p(1-p)$
geometric	$0 \leq p \leq 1$	$1/p$	$(1-p)/p^2$
binomial	n and $0 \leq p \leq 1$	np	$np(1-p)$
Poisson	$\lambda > 0$	λ	λ
discrete uniform	a and b	$(a+b)/2$	$[(b-a+1)^2-1]/12$
uniform	a and b	$(a+b)/2$	$(b-a)^2/12$
exponential	$\mu > 0$	$1/\mu$	$1/\mu^2$
Gaussian	m and σ	m	σ^2
Pareto	$\delta > 0$ and $1 < \gamma \leq 2$	$\delta\gamma/(\gamma-1)$	∞

The Central Limit Theorem

Let $X_1, X_2, X_3, \ \ldots, \ X_k$ be k independent and identically distributed (IID) random variables with common mean λ and variance σ^2. Define random variable Y_k as

$$Y_k = \frac{X_1 + X_2 + X_3 + \ \ \ldots \ \ + X_k - k\lambda}{\sigma\sqrt{k}}.$$

Then,

$$\lim_{k \to \infty} P(Y_k \le y) = \Phi(y)$$

where $\Phi(\cdot)$ is the distribution function of a standard Gaussian random variable given by

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt.$$

The central limit theorem is considered the most important result in probability.

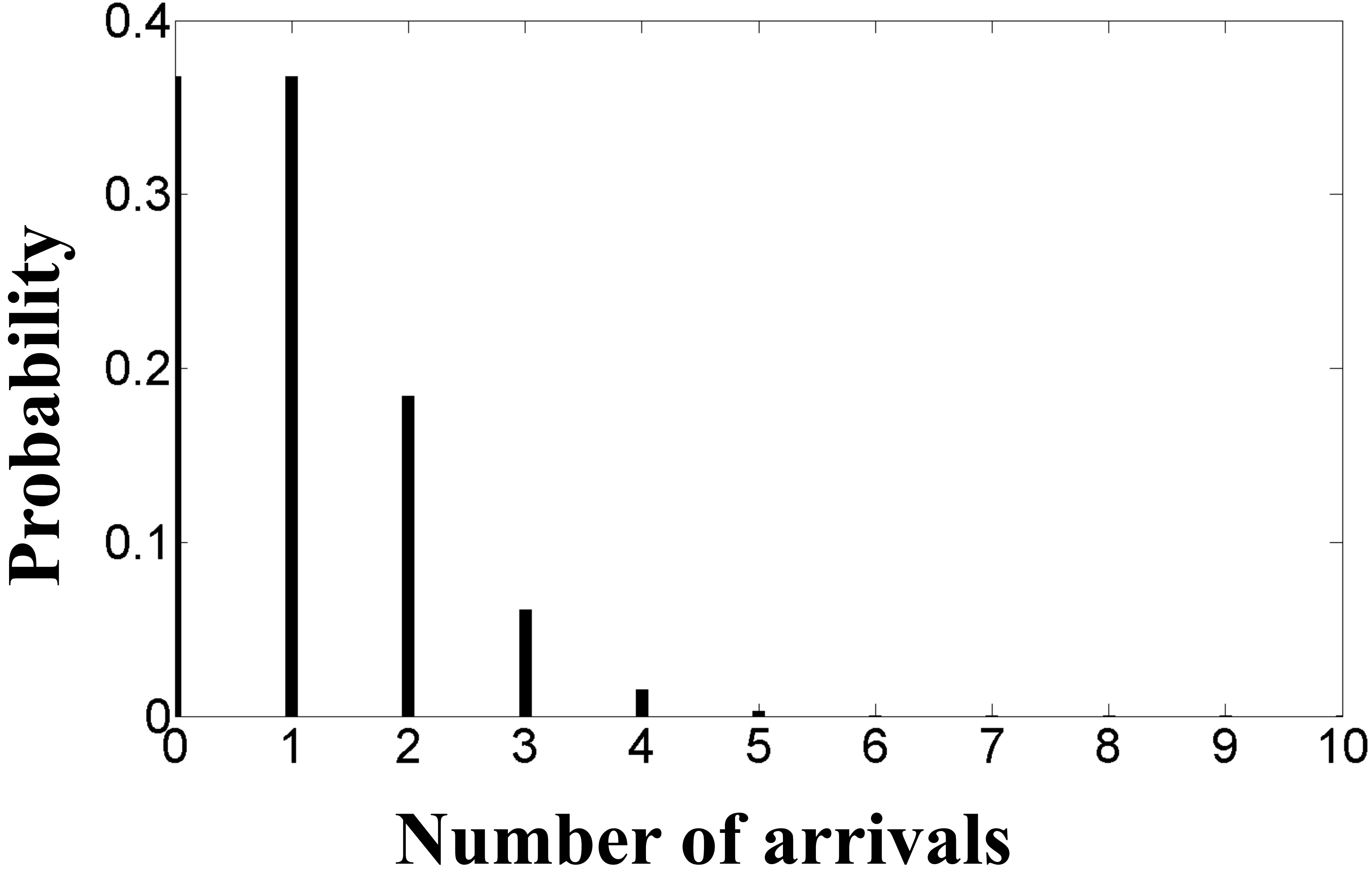
It implies that for large k, the sum of k IID random variable with common mean λ and variance σ^2 is approximately Guassian with mean $k\lambda$ and variance $k\sigma^2$ *regardless* of the distribution of these variables.

Furthermore, under certain conditions, the central limit theorem also applies in the case of sequences that are not identically distributed.

Homework:

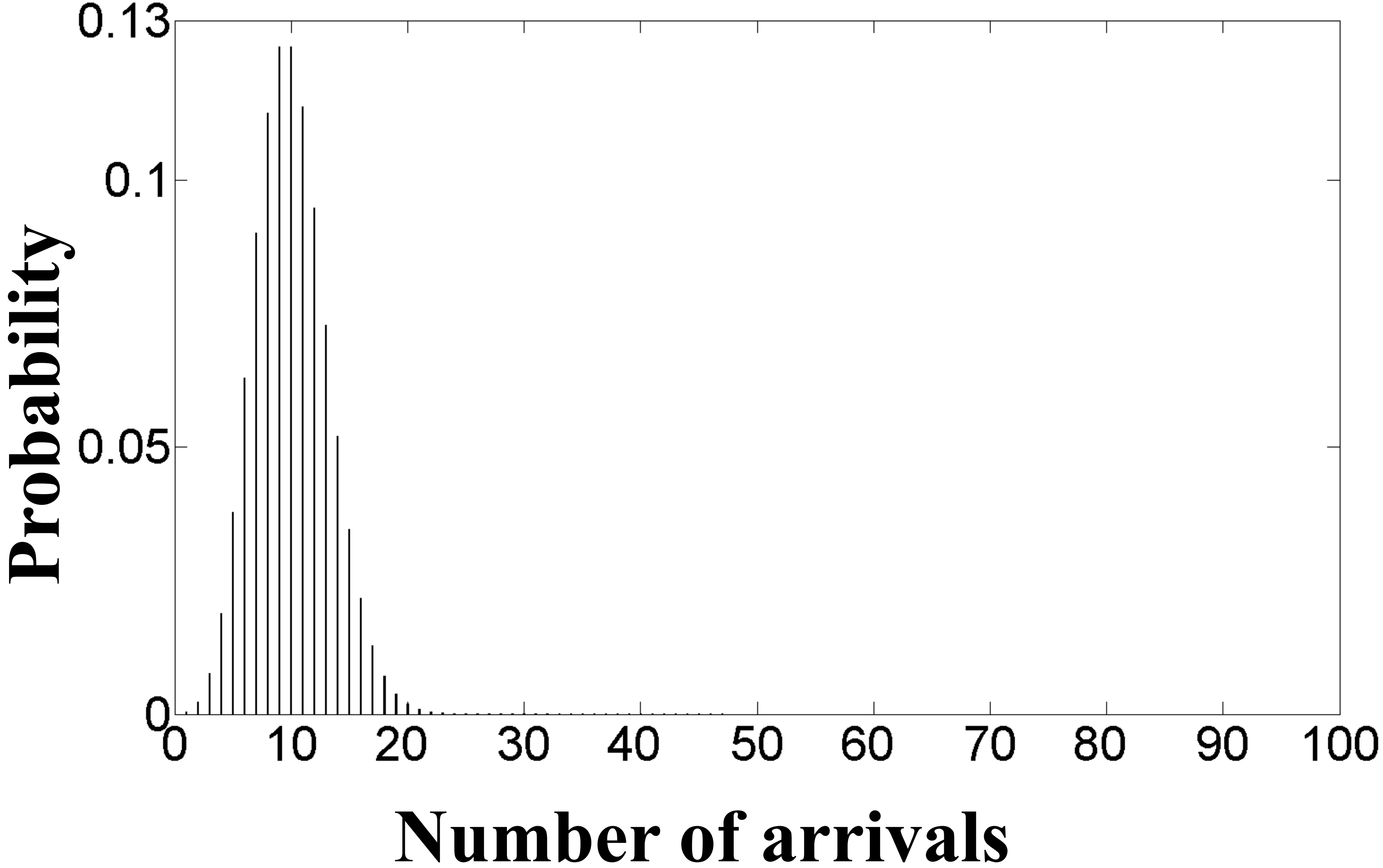
Observe the following behavior of the Poisson probability function and provide explanation.

Poisson Probability function with λ= 1



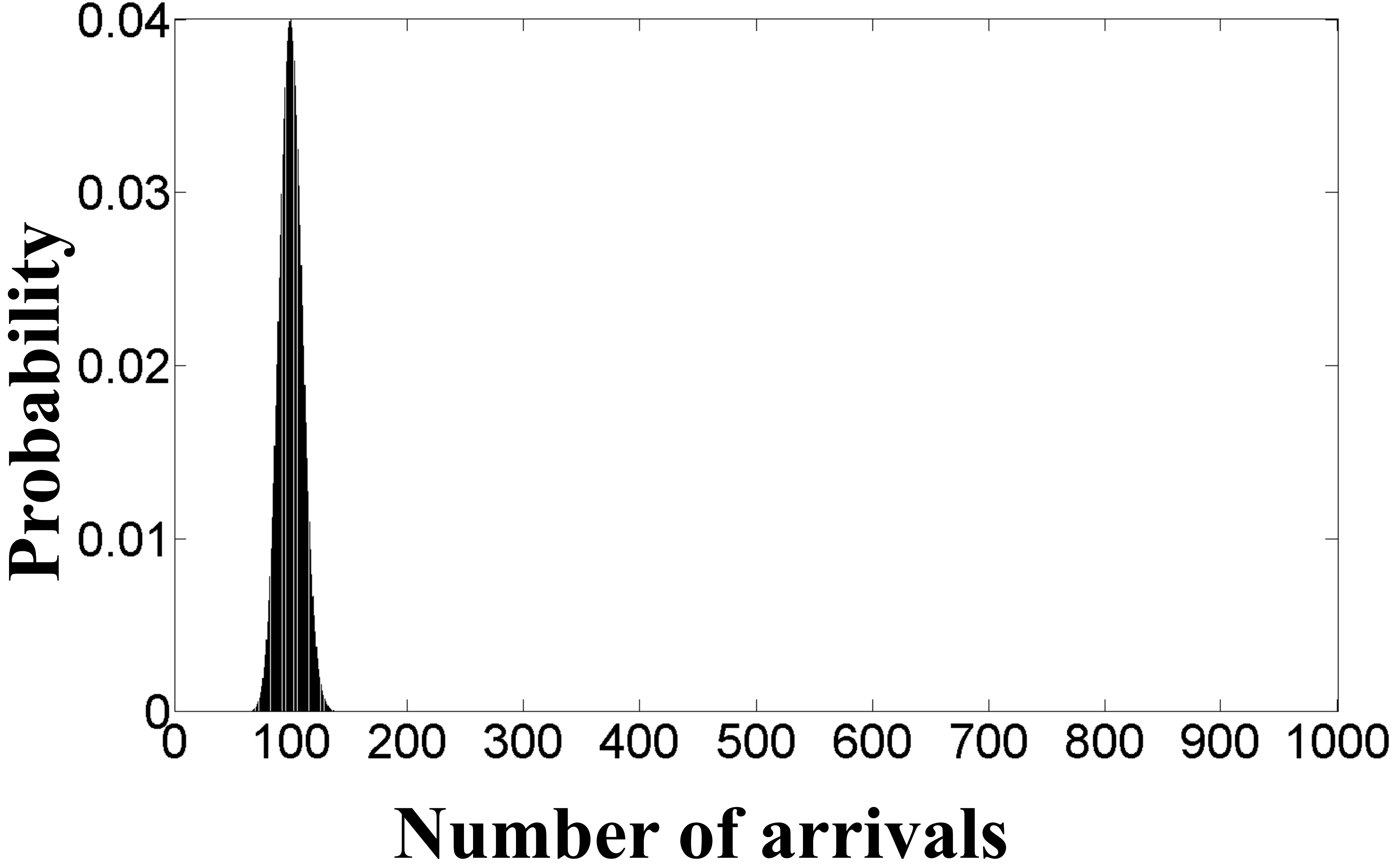
Credit: Li Fan

Poisson Probability function with λ= 10



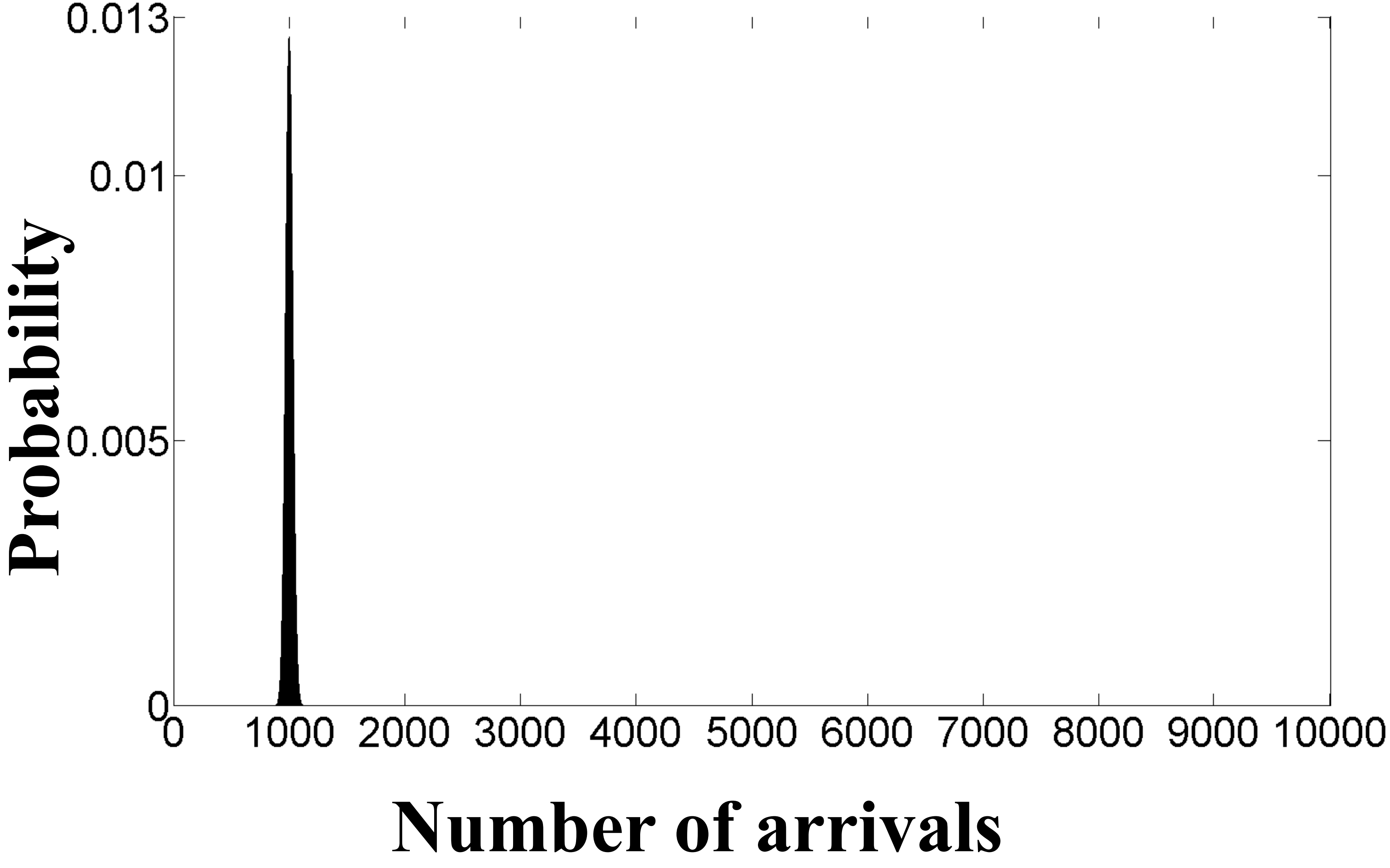
Credit: Li Fan

Poisson Probability function with λ= 100



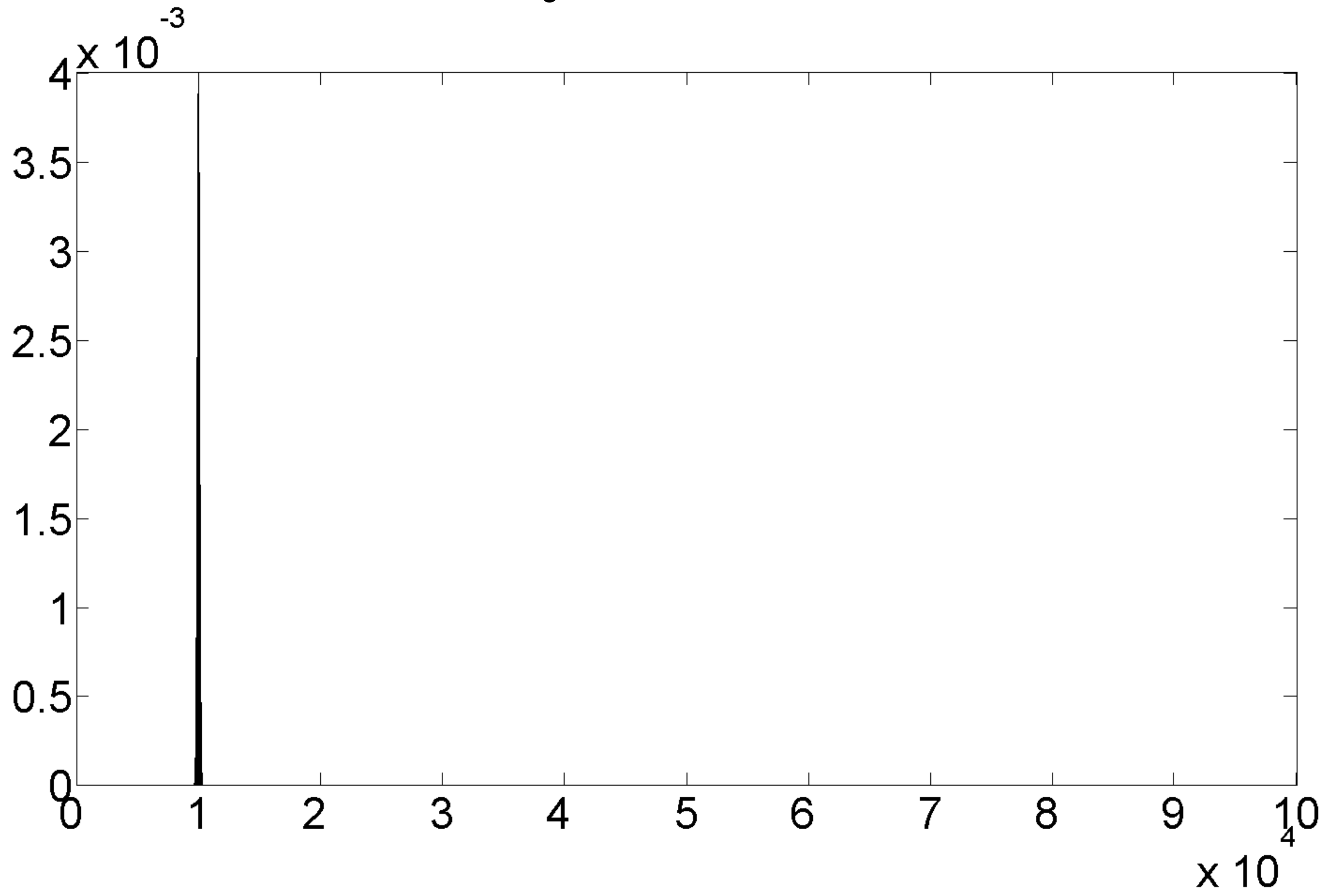
Credit: Li Fan

Poisson Probability function with λ= 1000



Credit: Li Fan

Poisson Probability function with λ= 10000



Probability

Number of arrivals

Credit: Li Fan

Link Dimensioning

We will consider several scenarios of sources (individuals or families) sharing a communication link. Each of the sources has certain requirements for capacity and the common link must be dimensioned in such a way that minimizes the cost for the telecommunications provider, but still meets the individual QoS requirements. The link dimensioning procedures that we consider apply to user requirements for capacity either upload or download.

Case 1: Homogeneous Individual Sources

Consider N independent sources (end-terminals), sharing a transmission link of capacity C [Mb/s]. Any of the sources transmits data in accordance with an on-off process. That is, a source alternates between two states: 1) the on state during which the source transmits at a rate R [Mb/s], and 2) the off state during which the source is idle. Assume that the proportion of time the source is in the on-state is p, so it is in the off-state $1 - p$ of the time. The question is how much capacity should the link have so it can serve all N sources such that the probability that the demand exceeds the total link capacity is no higher than α.

Without loss of generality, let us normalize the traffic generated by a source during on period by setting $R = 1$.

The demand generated by a single source is Bernoulli distributed with parameter p, so the demand generated by all N sources has Binomial distribution with parameters p and N.

Accordingly, finding the desired capacity is reduced to finding the smallest C such that

$$\sum_{i=C+1}^{N} \binom{N}{i} p^i (1-p)^{N-i} \leq \alpha.$$

If N is large we can use the central limit theorem and approximate the Binomial distribution by a Gaussian distribution.

Accordingly, the demand can be approximated by a Gaussian random variable with mean Np and variance $Np(1-p)$ and finding C_G such that the probability of our Gaussian random variable to exceed C_G is α.

It is well known that Gaussian random variables obey the so-called 68-95-99.7% Rule which means that the following apply to a random variable X with mean m and standard deviation σ.

$$\begin{aligned} P(m-\sigma \leq X \leq m+\sigma) &= 0.68 \\ P(m-2\sigma \leq X \leq m+2\sigma) &= 0.95 \\ P(m-3\sigma \leq X \leq m+3\sigma) &= 0.997. \end{aligned}$$

Therefore, if $\alpha = 0.0015$ then C_G should be three standard deviations above the mean, namely,

$$C_G = Np + 3\sqrt{Np(1-p)}.$$

Recall that for our original problem, before we introduced the Gaussian approximation, $C = N$ guarantees that there is sufficient capacity to serve all arriving traffic without losses. Therefore, we set our dimensioning rule for the optimal C value as follows:

$$C_{opt} = \min\left[N, Np + 3\sqrt{Np(1-p)}\right].$$

Case 2: Non-homogeneous Individual Sources

Now we generalize the above scenario to the case where the traffic and the peak rates of different sources can be different.

In this case where the sources are non-homogeneous, we must invoke a generalization of the central limit theorem that allows for non IID random variables (i.e., the so-called "Lyapunov's central limit theorem").

Consider N sources where the ith source transmits at rate R_i with probability p_i, and at rate 0 with probability $1 - p_i$.

Let $R_X(i)$ be a random variable representing the rate transmitted by source i. We obtain:

$$E[R_X(i)] = p_i R_i.$$

and

$$Var[R_X(i)] = R_i^2 p_i - (R_i p_i)^2 = R_i^2 p_i (1 - p_i).$$

The latter is consistent with the fact that $R_X(i)$ is equal to R_i times a Bernoulli random variable.

We now assume that the random variable

$$\Sigma_R = \sum_{i=1}^{N} R_X(i)$$

has a Gaussian distribution with mean

$$E[\Sigma_R] = \sum_{i=1}^{N} E[R_X(i)] = \sum_{i=1}^{N} p_i R_i$$

and variance

$$Var[\Sigma_R] = \sum_{i=1}^{N} Var[R_X(i)] = \sum_{i=1}^{N} R_i^2 p_i (1 - p_i).$$

Notice that the allocated capacity should not be more than the total sum of the peak rates of the individual sources. Therefore, in this more general case, for the QoS requirement $\alpha = 0.0015$, our optimal C value is set to:

$$C_{opt} = \min \left[\sum_{i=1}^{N} R_i, E[\Sigma_R] + 3\sqrt{Var[\Sigma_R]} \right].$$

For lower α value, mean + 4 or even 5 standard deviations may be required

Homework

There are 20 sources each transmits at a peak-rate of 10 Mb/s with probability 0.1 and is idle with probability 0.9, and there are other 80 sources each transmits at a peak-rate of 1 Mb/s with probability 0.05 and is idle with probability 0.95.

A service provider aims to allocate the minimal capacity C_{opt} such that no more than 0.0015 of the time, the demand of all these 100 sources exceeds the available capacity. Find an appropriate C_{opt}.

Answer: $C_{opt} = 64.67186$ Mb/s.

Notice the difference in contributions to the total variance of sources from the first group versus such contributions of sources from the second group.

Consider a range of examples where the variance is the dominant part of C_{opt} versus examples where the variance is not the dominant part of C_{opt}.