Famnit

Probability

The written Corona lectures

Published in the year of our Lord 2021,
when we were all suffering from the Plague
Michaelus Permanus

The Corona lecture notes

PROBABILITY

Mihael Perman

Note:

The following are the Corona lecture notes. They cover most of the material we covered by Zoom except Chapter 2.2 on continuous distributions.

1. Outcomes, events, probabilities

1.1. Outcomes and events

Example: Italian gamblers in the 17th century liked the game where they placed a bet on the outcome of rolling three dice. Popular bets were 9 and 10. The gamblers had a „theory" that the two popular bets are equivalent in the sense that the probability of winning is the same for both bets. They wrote down two lists:

Sum 9	Sum 10
1 2 6	1 3 6
1 3 5	1 4 5
1 4 4	2 2 6
2 2 5	2 3 5
2 3 4	2 4 4
3 3 3	3 3 4

Based on these two lists the two games were deemed equivalent. However, gambling experience suggested they were not.

The problem was solved by Galileo Galilei (1564 – 1642).

He wrote down all possible outcomes.

111	112	113	114	115	116
121	122	123	124	125	126
⋮	⋮	⋮	⋮	⋮	⋮
661	662	663	664	665	666

There are $6^3 = 216$ possible triplets. Galileo found that 25 sum to 9 and 27 to 10.

Assuming all triplets have the same chance of appearing the problem is solved.

The moral of the story is that we have to write down all possibilities when dealing with an experiment involving chance.

In mathematical language we will talk about the set of all possible outcomes and denote it by Ω.

Examples:

(i) In Galileo's example we have

$$\Omega = \{(i,j,k) : 1 \le i,j,k \le 6\}$$
$$= \{1,2,3,\dots,6\}^3$$

(ii) If we toss a coin n times we get a sequence of n heads and tails. In this case

$$\Omega = \{H,T\}^n.$$

(iii) Suppose we arrange n objects in random order. This means that we choose a random permutation or

$$\Omega = S_n = \text{set of all permutations.}$$

(iv) We can think of tossing a coin infinitely many times. In this case

$$\Omega = \{H, T\}^{\mathbb{N}}$$

which is the set of all countably infinite sequences of symbols H and T.

The next concept is the event. If we roll three dice an event is, say, that the sum is 9. An event can either happen or not. But what is an event mathematically? All the triplets that give a sum of 9 are a subset of $\Omega = \{1, 2, 3, 4, 5, 6\}^3$.

It is plausible to understand events as subsets of Ω.

We will denote events by $A, B, C, \dots$

For mathematical reasons denote the family of all events by $\mathcal{F}$.

We will require the following.

(i) $\Omega \in \mathcal{F}$.

(ii) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(iii) if $A_1, A_2, \dots \in \mathcal{F}$, then

$$\bigcup_i A_i \in \mathcal{F}.$$

The union is either finite or countable.

Remark: In mathematics a family of subsets with the above properties is called a σ-algebra.

Remark: In cases of infinite sets Ω not all subsets are necessarily events. For finite Ω we will usually assume that all subsets are events.

In Galileo's example we assumed that all outcomes in Ω are equally likely. The probability of

$A = \{\text{sum is } 9\}$ is then $25/216$.

If $B = \{\text{sum is } 10\}$ then $P(B) = 27/216$.

We have $A \cap B = \emptyset$ and

$$P(A \cup B) = \frac{25 + 27}{216} = P(A) + P(B).$$

For mathematical reasons it turns out to be better to assign probabilities to events rather than outcomes. The example shows that for disjoint A and B we should have $P(A \cup B) = P(A) + P(B)$.

Any assignment of probabilities should have this property. The mathematical definition is more general.

<u>Definition</u> : Probability is an assignment to every event $A \in \mathcal{F}$ of a real number in such a way that

(i) $0 \leq P(A) \leq 1$, $P(\Omega) = 1$.

(ii) if $A_1, A_2, \ldots$ are disjoint we have

$$P(\bigcup_i A_i) = \sum_i P(A_i)$$

Remark: The sum in (ii) can be finite or infinite.

Remark: The property (ii) is called σ-additivity.

Let us look at some simple consequences of the above definition.

(i) we have $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. By additivity

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

so $\boxed{P(A^c) = 1 - P(A)}$

(ii) Let A, B be events. We can write

$$A \cup B = \underbrace{(A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)}_{\text{disjoint}}.$$

So

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$$

But $\underbrace{(A \cap B^c) \cup (A \cap B)}_{\text{disjoint}} = A$ so

$$P(A \cap B^c) + P(A \cap B) = P(A) \Rightarrow$$

$$P(A \cap B^c) = P(A) - P(A \cap B)$$

and similarly

$$P(A^c \cap B) = P(B) - P(A \cap B)$$

Using this in the above expression gives

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

If we have three events we get

$$P(A \cup B \cup C) = P((A \cup B) \cup C)$$

$$= P(A \cup B) + P(C)$$

$$- P((A \cup B) \cap C)$$

$$= P(A) + P(B) - P(A \cap B) + P(C)$$

$$- P((A \cap C) \cup (B \cap C))$$

$$= P(A) + P(B) + P(C) - P(A \cap B)$$

$$- P(A \cap C) - P(B \cap C)$$

$$+ P(A \cap B \cap C)$$

From this we generalize to

<u>Theorem 1.1</u> (inclusion-exclusion formula)

Let $A_1, A_2, \dots, A_n$ be events.

We have

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j)$$

$$+ \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k)$$

$$- \dots$$

$$+ (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n)$$

Proof: We know that the formula is valid for $n = 2$. Suppose it is valid for n. We write

$$\bigcup_{i=1}^{n+1} A_i = \bigcup_{i=1}^{n} A_i \cup A_{n+1} \quad \text{so}$$

$$P\left(\bigcup_{i=1}^{n+1} A_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) - \underbrace{P\left(\bigcup_{i=1}^{n} A_i \cap A_{n+1}\right)}_{= P\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right)}.$$

By the induction assumption the formula is valid for unions of n sets. This means

$$P\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right) = \sum_{i=1}^{n} P(A_i \cap A_{n+1}) - \sum_{1 \le i < j \le n} P(A_i \cap A_j \cap A_{n+1}) + \dots + (-1)^{n-1} P(A_1 \cap \dots \cap A_{n+1})$$

Using this gives the inclusion exclusion formula for $n+1$ sets, and the induction step is completed.

<u>Example</u>: n couples go dancing. When they are about to leave the power goes out and each woman grabs a man at random. What is the probability that no woman will grab her man?

In probability language we are talking about choosing a random permutation of n numbers. All permutations have the same probability $\frac{1}{n!}$.

<u>Figure</u>:

Women	1	2	3	. i .	n
Men	3	5	4	$\sigma(i)$	1

Define $A_i = \{$ woman i grabs her man $\}$.

and $A = \{$ no woman grabs her man $\}$

We have

$$A^c = \bigcup_{i=1}^{n} A_i$$

To use the exclusion-exclusion formula we need the following probabilities:

\# of permutations sending $i \to i$.

$$P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

\# of all probabilities

$$P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$

$\vdots$

$$P(A_{i_1} \cap \cdots \cap A_{i_r}) = \frac{(n-r)!}{n!}$$

The inclusion - exclusion formula gives

$$P(A^c) = \binom{n}{1} \cdot \frac{1}{n} - \binom{n}{2} \cdot \frac{(n-2)!}{n!}$$

$$+ \binom{n}{3} \frac{(n-3)!}{n!} -$$

$$\vdots$$

$$\vdots$$

$$+ (-1)^n \quad \frac{0!}{n!}$$

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1} \cdot \frac{1}{n!}$$

The symbol $\binom{n}{r}$ counts the number of different intersections.

Finally

$$P(A) = 1 - P(A^c)$$

$$= \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \cdot \frac{1}{n!}$$

From Analysis we know

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots$$

Take $x = -1$ to get

$$e^{-1} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots$$

The probability $P(A)$ is a partial sum of the above series which converges fast. We can approximate

$$P(A) \approx e^{-1} = 0.3697.$$

1.2. Conditional probabilities and independence

<u>Example</u>: Let us return to Galileo's example. We have $\Omega = \{1,2,3,4,5,6\}^3$ and all triplets are equally likely. Suppose you know that the first component is 1 but not the other two components. What is your opinion about the probability that the sum is 9? There are 36 triplets of the form $(1,j,k)$. Of these the triplets

$(1,2,6)$
$(1,3,5)$
$(1,4,4)$
$(1,5,3)$
$(1,6,2)$

give a sum of 9.

Is it reasonable to assum that given the information that the first component is 1 all the 36 triplets are equally likely? Yes. So the updated probability is $5/36$ which is different from $25/216$. We rewrite

$$5/36 = \frac{5/216}{36/216}$$

and denote $A = \{\text{sum is } 9\}$ and $B = \{\text{first component is } 1\}$.

We have

$$5/36 = \frac{P(A \cap B)}{P(B)}$$

Definition: The conditional probability of A given B is

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

Remark: If we have additional information about an outcome this usually means that the outcome is in a restricted subset of Ω. In the above example this restricted subset is B.

Rewriting the definition we get

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

If $A_1, A_2, \ldots, A_n$ are events we can write

$$P(A_1 \cap \cdots \cap A_{n-1} \cap A_n) =$$

$$P(A_n \mid A_1 \cap \cdots \cap A_{n-1}) \, P(A_1 \cap \cdots \cap A_{n-1})$$

Iterating the rule gives

$$\boxed{\begin{aligned} &P(A_1 \cap \dots \cap A_n) \\ &= P(A_n \mid A_1 \cap \dots \cap A_{n-1}) \cdot \\ &\quad P(A_{n-1} \mid A_1 \cap \dots \cap A_{n-2}) \\ &\quad \ddots \\ &\quad P(A_2 \mid A_1) \cdot P(A_1) \end{aligned}}$$

<u>Definition</u>: A collection $\{H_1, H_2, \dots, H_n\}$ is a partition of Ω if $H_i \cap H_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^{n} H_i = \Omega$.

<u>Theorem 1.2</u> (law of total probabilities)

Let $\{H_1, H_2, \dots, H_n\}$ be a partition and A an event. We have

$$P(A) = \sum_{i=1}^{n} P(A \mid H_i) \cdot P(H_i)$$

<u>Proof</u>: We write

$$
\begin{aligned}
A &= A \cap \Omega \\
&= A \cap \bigcup_{i=1}^{n} H_i \\
&= \underbrace{\bigcup_{i=1}^{n} (A \cap H_i)}_{\text{disjoint events.}}
\end{aligned}
$$

It follows that

$$
\begin{aligned}
P(A) &= \sum_{i=1}^{n} P(A \cap H_i) \\
&= \sum_{i=1}^{n} \frac{P(A \cap H_i)}{P(H_i)} \cdot P(H_i) \\
&= \sum_{i=1}^{n} P(A \mid H_i) \cdot P(H_i)
\end{aligned}
$$

<u>Remark</u>: This formula is useful to compute probabilities. We can often guess conditional probabilities.

Example: In an Internet game of chance you have 12 tickets.

$\boxed{1}$ $\boxed{1}$ $\boxed{1}$ $\boxed{1}$ $\boxed{2}$ $\boxed{2}$ $\boxed{3}$ $\boxed{D}$ $\boxed{S}$ $\boxed{3}$ $\boxed{S}$ $\boxed{S}$

The tickets are randomly permuted and turned around so that the player sees

The player then turns around the tickets until the first $\boxed{S}$ = STOP. One example is

$\boxed{1}$ $\boxed{2}$ $\boxed{D}$ $\boxed{1}$ $\boxed{S}$

The payoff is the sum of digits multiplied by 2 if $\boxed{D}$ = DOUBLE is among the visible tickets.
In the above example the payoff is 8.

What is the probability that the player will see the ticket $\boxed{D}$?

Define $H_i = \{$ first $\boxed{S}$ is in position $i\}$ for $i = 1, 2, \dots, 9$. The collection $\{H_1, H_2, \dots, H_9\}$ is a partition.

Let $A = \{$ we see $\boxed{D}\}$.

First we compute

$$P(H_i) = \frac{8}{12} \cdot \frac{7}{11} \cdot \; \cdot \frac{8-i+2}{12-i+2} \cdot \frac{4}{12-i+1}$$

What about $P(A \mid H_i)$? Idea?

If $\boxed{S}$ appears in position i then the first $i-1$ tickets are randomly chosen from

$\boxed{A}$ $\boxed{1}$ $\boxed{1}$ $\boxed{1}$ $\boxed{2}$ $\boxed{2}$ $\boxed{3}$ $\boxed{D}$. So we choose $i-1$ tickets out of 8 and ask for the (conditional) probability that $\boxed{D}$ is among the tickets chosen. We have

$$P(A \mid H_i) = \frac{\binom{7}{i-2}}{\binom{8}{i-1}}$$

$\binom{7}{i-2}$ ← # of samples of size $i-1$ containing $\boxed{D}$.

$\binom{8}{i-1}$ ↑ # of possible samples

Cancelling we get

$$P(A \mid H_i) = \frac{i-1}{8}$$

<u>Check</u>: For $i = 9$ we should get 1.

The rest is adding fractions.

$$P(A) = \sum_{i=1}^{9} \frac{8!\,(12-i)!}{12!\,(8-i+1)!} \cdot 4 \cdot \frac{(i-1)}{8}$$

$$= \frac{8!}{12!} \cdot \frac{1}{2} \sum_{i=1}^{9} \frac{(12-i)!\,(i-1)}{(8-i+1)!}$$

$$= 1/5$$

Example : Prisoner's Paradox

Three prisoners are in jail in a dark country. They are all sentenced to death but the ruler will choose one of them at random and pardon him. Here is a conversation between the guard in jail and prisoner A:

A: Guard, you already know who will be pardoned. If you tell me who of the other two will not be pardoned you do not give me any information.

G: If I tell you there will be only two of you left. Your probability of survival is then 1/2. I do give you some information.

Who is right? To talk about conditional probabilities we need a space of all possible outcomes. Here is a suggestion:

$$\Omega = \left\{ \begin{array}{ll} \boxed{AB\,|\,B} & 1/3 \\ \boxed{AC\,|\,C} & 1/3 \\ \boxed{BC\,|\,B} & \left.\begin{array}{l} \\ \end{array}\right\} \ 1/3 \ \begin{array}{l} \to x/3 \\ \to (1-x)/3 \end{array} \\ \boxed{BC\,|\,C} & \end{array} \right.$$

Last letter is what the guard says

First two letters are the wretched prisoners who will be hanged

There is no indication how the last probability of $1/3$ is distributed between the last two outcomes. Let us say $\frac{x}{3}$ and $\frac{1-x}{3}$ for $x \in [0,1]$.

We compute

$$P(\text{A survives} \mid \text{Guard says B})$$

$$= \frac{P(\{\text{A survives}\} \cap \{\text{Guard says B}\})}{P(\text{Guard says B})}$$

$$= \frac{x/3}{1/3 + x/3}$$

$$= \frac{x}{1+x}$$

This function has values from 0 to $1/2$ on $[0,1]$.

Two cases:

(i) if $x = 1/2$ the guard chooses at random when he has the choice. In this case the conditional probability is

$$\frac{1/2}{1+1/2} = 1/3$$ as before.

(ii) If $x = 0$ then the conditional probability is 0!

Why?

What if $P(A|B) = P(A)$? Then B does not "tell us anything" about the probability of A. The word we choose is independence. The above equality can be written as

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

by definition. If we have events A, B and C and they are "independent" then $A \cap B$ ought to be independent of C. This leads to the following definition.

<u>Definition</u>:

(i) The events A and B are independent if

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

(ii) Events $A_1, A_2, \dots, A_n$ are independent if for all collections of indeces $1 \le i_1 < i_2 < \dots < i_m \le n$ we have

$$P(A_{i_1} \cap \dots \cap A_{i_m})$$

$$= P(A_{i_1}) P(A_{i_2}) \dots P(A_{i_m}).$$

<u>Remark</u>: Typically independence is associated with physically different objects, like several dice, differen coins.

Example (Paradox of Chevalier de Méré, Antoine Gombaud, 1607 – 1684).

Chevalier de Méré considered the following two games of chance.

(i) You roll a die 4 times. You win if you see at least one ace (ace = ⚀)

(ii) You roll two dice 24 times. You win if you see at least one double ace. i.e. ⚀⚀

Which of the two games has a higher probability of winning?

Let us look at the first game.

Define $A_i = \{$ the i-th roll is not ⚀ $\}$ and $A = \{$ we win $\}$.

We have

$$A^c = A_1 \cap A_2 \cap A_3 \cap A_4$$

It is reasonable to assume that subsequent rolls are independent which means that A_1, A_2, A_3, A_4 are independent. It follows

$$P(A^c) = P(A_1 \cap A_2 \cap A_3 \cap A_4)$$

$$= P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot P(A_4)$$

$$= 5/6 \cdot 5/6 \cdot 5/6 \cdot 5/6$$

$$= \left(5/6\right)^4$$

Finally we get

$$P(A) = 1 - P(A^c) = 1 - \left(5/6\right)^4 \doteq 0.5177$$

For the second game we similarly define

$$A_i = \{\text{not a double ace on roll } i\},$$

$$i = 1, 2, \ldots, 24.$$

$A = \{ \text{we win} \}$

We have

$$A^c = A_1 \cap A_2 \cap \dots \cap A_{24}$$

We assume independence and get

$$P(A^c) = P(A_1) P(A_2) \cdots P(A_{24})$$

$$= 35/36 \cdot 35/36 \cdot \dots \cdot 35/36$$

$$= (35/36)^{24},$$

and finally

$$P(A) = 1 - P(A^c) = 1 - (35/36)^{24} \doteq 0.4914.$$

Comment: The difference is small but nevertheless important.

Example (Gambler's ruin).

Two gamblers A and B start out with m and n sequins (gold coins) respectively. In each round of the game they toss a coin. If it is heads A gets a coin from B; if it is tails B gets a coin from A. They play until one of them is left with no coins. What is the probability that A will get all the coins? We assume that tosses are independent and the probability of heads is $p \in (0,1)$.

Let $A = \{\text{gambler } A \text{ wins}\}$ and denote $p_{m,n} = P(A \text{ wins})$.

Let $H = \{\text{first toss is heads}\}$.

Then

$$P(A) = \underbrace{P(A \mid H)}_{= p_{m+1,n-1}} \cdot P(H) + \underbrace{P(A \mid H^c)}_{= p_{m-1,n+1}} P(H^c)$$

We get the recursion

$$p_{m,n} = p \cdot p_{m+1,n-1} + (1-p)\, p_{m-1,n+1}.$$

Obviously $p_{0,m+n} = 0$ and $p_{m+n,0} = 1$.

Denote

$$\pi_m = p_{m,n}.$$

We can rewrite the recursion as

$$\pi_m = p \cdot \pi_{m+1} + (1-p)\, \pi_{m-1}.$$

with $\pi_{m+n} = 1$ and $\pi_0 = 0$.

Define $q := 1-p$. We rewrite

$$\underbrace{(p+q)}_{=1} \pi_m = p\,\pi_{m+1} + q\,\pi_{m-1}, \text{ or}$$

$$p(\pi_{m+1} - \pi_m) = q(\pi_m - \pi_{m-1}), \text{ or}$$

$$\pi_{m+1} - \pi_m = q/p\,(\pi_m - \pi_{m-1})$$

Write

$$\pi_2 - \pi_1 = q/p\,(\pi_1 - \pi_0)$$

$$\pi_3 - \pi_2 = q/p\,(\pi_2 - \pi_1)$$

$$= (q/p)^2 (\pi_1 - \pi_0)$$

$$\vdots$$

$$\pi_{m+n} - \pi_{m+n-1} = (q/p)^{m+n-1} (\pi_1 - \pi_0)$$

From this we get by adding

$$\pi_1 \left(1 + q/p + \dots + (q/p)^{m+n-1}\right)$$

$$= \pi_{m+n}$$

$$= 1$$

It follows that

$$\pi_1 = \frac{1}{1 + q/p + \dots + q/p^{\,m+n-1}}.$$

As a consequence we have

$$\pi_m = p_{m,n} = \frac{1 + (q/p) + \dots + (q/p)^{m-1}}{1 + q/p + \dots + (q/p)^{m+n-1}}.$$

How do we know that the game will end?

Let B_k be defined as

$B_k = \{$ tosses $(m+n)k+1, (m+n)k+2, \dots$

$\dots (m+n)(k+1)-1$

produce heads $\}$

By independence

$$P(B_k) = p^{m+n}.$$

But B_k depend on disjoint blocks of events so they are independent. But

$$\{\text{game ends}\} \supseteq \bigcup_{k=1}^{\infty} B_k$$

We compute

$$P\left(\bigcup_{k=1}^{\infty} B_k\right) = \lim_{k \to \infty} P\left(\bigcup_{k=1}^{r} B_k\right)$$

$$= \lim_{r \to \infty} \left(1 - P\left(\bigcap_{k=1}^{n} B_k^c\right)\right)$$

$$= 1 - \lim_{r \to \infty} P(B_k^c)^r$$

$$= 1 - \lim_{r \to \infty} (1 - p^{m+n})^r$$

$$= \underline{1}.$$

Lemma 1.3: Let $A_1, A_2, \dots$ be events.

(i) If $A_1 \subseteq A_2 \subseteq \dots$ then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} P(A_n)$$

(ii) If $A_1 \supseteq A_2 \supseteq \dots$ then

$$P\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} P(A_n)$$

Proof: We only prove (i). The second assertion follows by de Morgan rules. Write

$$\bigcup_{k=1}^{\infty} A_k = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_1 \cup A_2) \cup \dots$$

The events in the union on the right are disjoint. It follows that

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = P(A_1) + P(A_2 \setminus A_1) + \dots$$

The infinite series on the right converges and its partial sum is

$$P(A_1) + P(A_2 \setminus A_1) + \dots + P(A_n \setminus A_1 \cup \dots \cup A_{n-1})$$

$$= P\left(\bigcup_{k=1}^{n} A_k\right) = P(A_n).$$

The assertion follows.

2. Random variables

2.1. Discrete random variables

In Galileo's example we had $\Omega = \{1, 2, 3, 4, 5, 6\}^3$. The outcomes are of the form (i, j, k). The gamblers, however, were not so interested in the inner structure of the outcome but in the random number that was the sum of dots. We can imagine that random numbers are created through some process that involves chance. In probability we call such random numbers random variables. We denote them by letters X, Y, Z.

Technical note: Formally we understand random variables as functions from Ω to the real numbers $\mathbb{R}$. We imagine that some invisible „hand" chooses the outcome ω and the random variable X gives the random number $X(\omega)$.

Definition: A random variable X is a function $X : \Omega \to \mathbb{R}$ such that $X^{-1}((a, b])$ is an event for all $a < b$, $a, b \in \mathbb{R}$.

Note: The choice of intervals of the form $(a, b]$ is arbitrary. Intervals of the form (a, b), $[a, b]$ do the same.

Definition: A random variable X is discrete if it can only take values in a finite or countable set $\{x_1, x_2, \dots\}$.

We can see that for discrete random variables we can require $X^{-1}(\{x_k\})$ to be an event for all possible values. This definition is equivalent to the more general one.

To continue with Galileo's example gamblers were interested in the probabilities that the sum is 9 or 10.
We can ask the question for any $k \in \{3, 4, \dots, 18\}$.

A note on notation: We will write $\{X = x_k\}$ for the event $X^{-1}(\{x_k\})$. When we write probabilities $P(\{X = x_k\})$ we will drop the curly brackets and write $P(X = x_k)$.

We obviously have $\bigcup_{k=3}^{18} \{X = k\} = \Omega$. Since the events are disjoint we have

$$\sum_{k=3}^{18} P(X = k) = P(\Omega) = 1.$$

The total probability 1 is "distributed" among all possible values of X.

We say that these probabilities determine the distribution of X.

<u>Definition</u>: Let X be a discrete random variable. The distribution of X is given by the probabilities $P(X = x_k)$ for all possible values of X.

There is a number of standard distributions in probability.

<u>Binomial distribution</u>

Suppose we toss a coin n times. Suppose the tosses are independent and the probability of heads is $p \in (0,1)$. Let $X = \#$ of heads in n tosses. This random variable has values $k = 0, 1, 2, \ldots, n$.

To describe the distribution we need to compute $P(X = k)$ for all k.

We have $\Omega = \{H, T\}^n$ and the event $\{X = k\}$ consists of all sequences HTHHT . . . . . H that contain exactly k heads. Every such outcome has the probability $p^k (1-p)^{n-k}$ because of independence. So we only need to compute how many such outcomes there are. But this is given by $\binom{n}{k}$ because we need to choose k positions for heads among n positions. We have

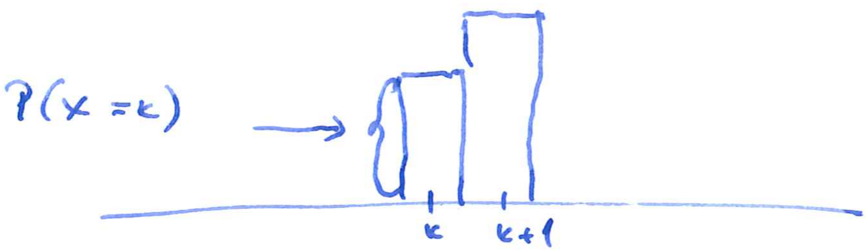
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$. We say that X has binomial distribution with parameters n and p.

Notation: $X \sim Bin(n, p)$.

The way to visualize a distribution is to draw a histogram. If X is a random variable with integer values we draw a column over a possible value k of X with base 1 and height $P(X=k)$ centered on k.

Figure:



Let us consider $X \sim Bin(n,p)$.

For $k \geq 1$ we can compute

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

$$= \frac{n-k+1}{k} \cdot \frac{p}{1-p}$$

If $\dfrac{P(X=k)}{P(X=k-1)} > 1$, then we have $P(X=k) > P(X=k-1)$, i.e. the column over k is taller than the column over $k-1$. This happens if

$$\frac{n-k+1}{k} \cdot \frac{p}{1-p} > 1 \Rightarrow$$

$$(n-k+1)p > k(1-p) \Rightarrow$$

$$(n+1)p > k$$

We have two cases:

1. If $(n+1)p$ is not an integer then the tallest column is over $k = \lfloor (n+1)p \rfloor$.

2. If $(n+1)p$ is an integer then for $k = (n+1)p$ we have

$$\frac{P(X=k)}{P(X=k-1)} = 1$$

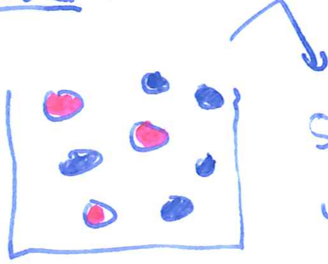
This means that the two columns over $k = (n+1)p$ and $(n+1)p - 1$ are the tallest and equal.

Coin tosses are methaphors for conting „successes" in identical and independent repetitions of the same experiment.

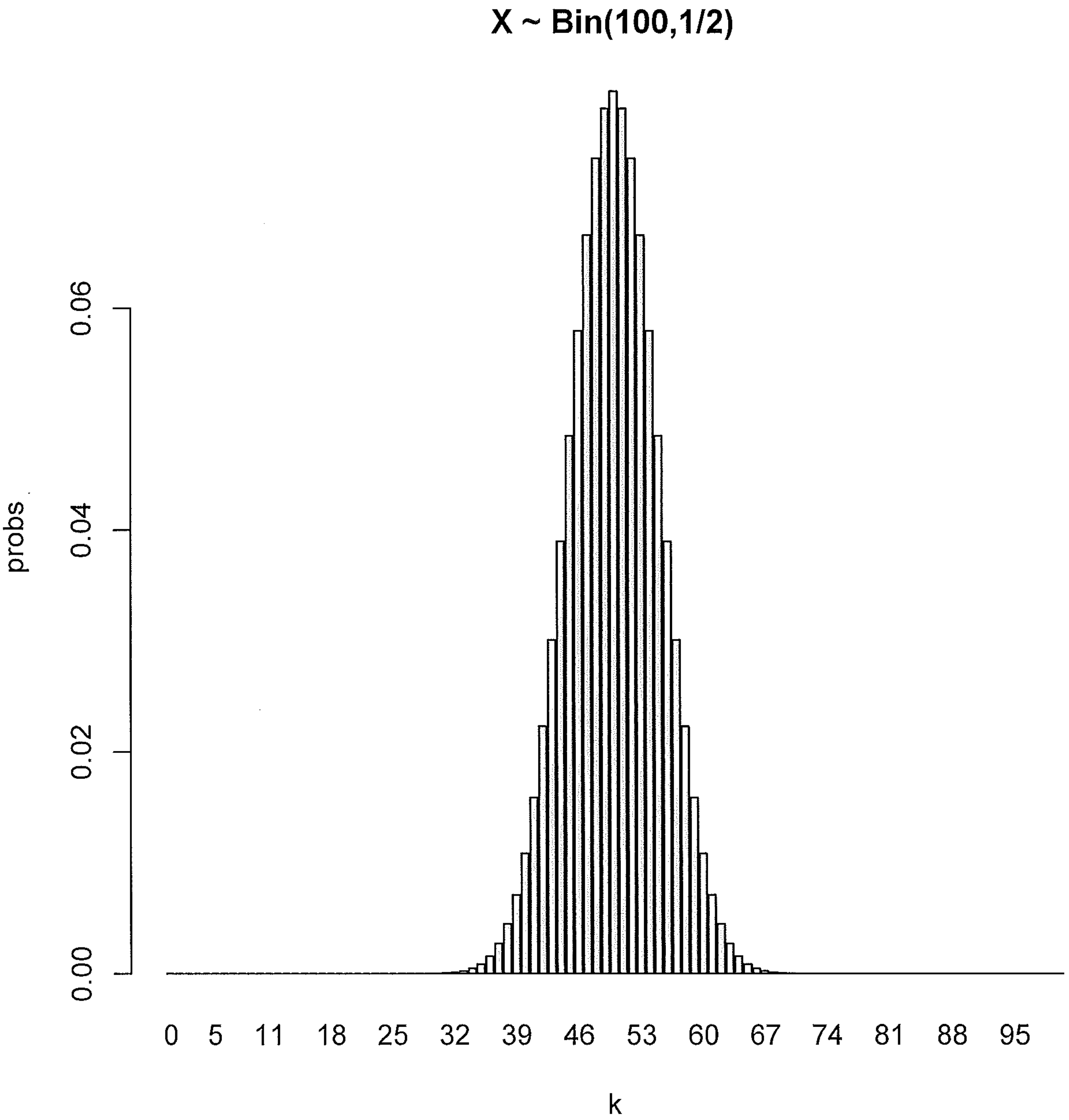
Hyper-geometric distribution

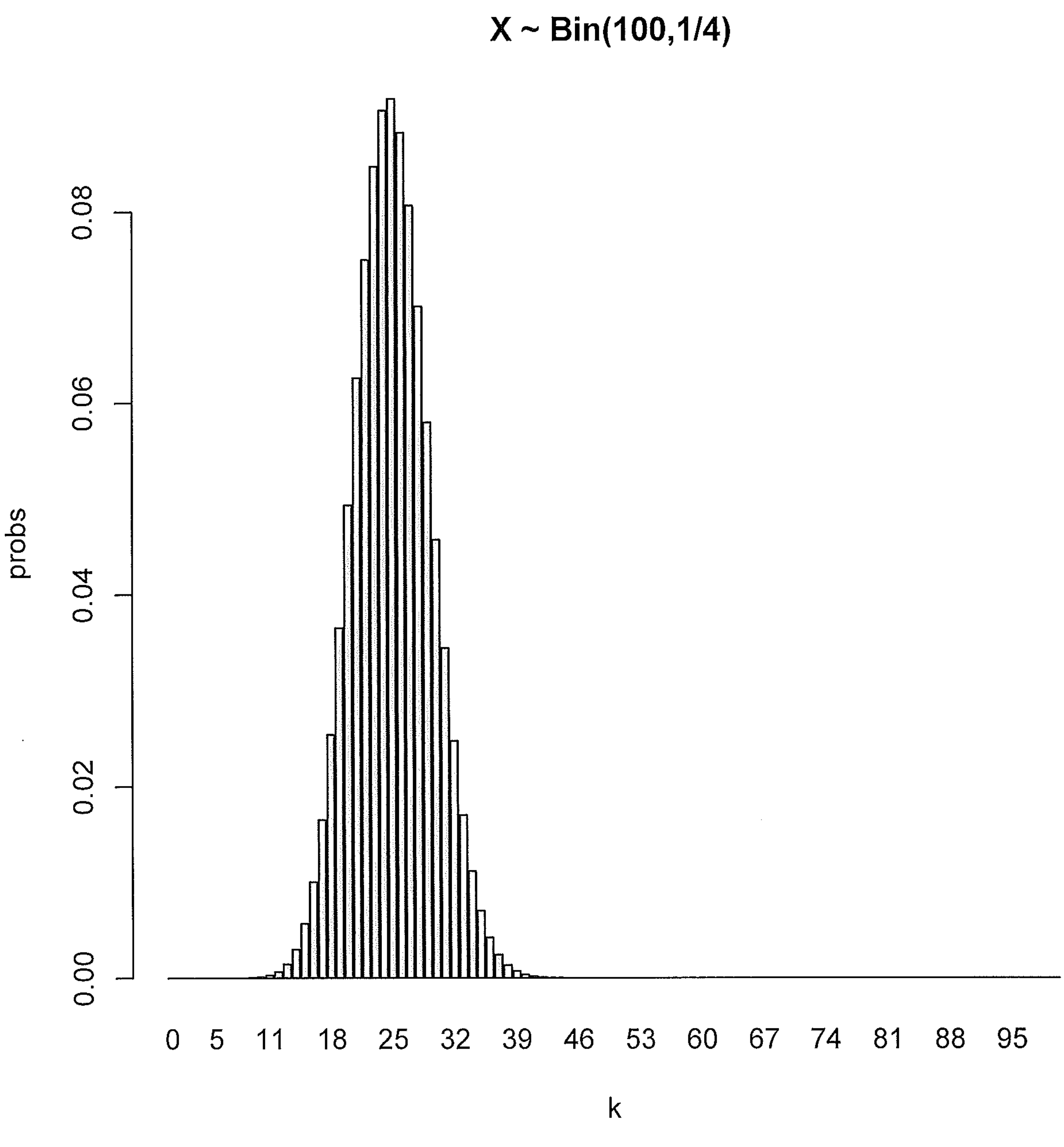
Suppose we have an urn with B black and R red balls. Denote $N = B + R$. Suppose we select $n \leq N$ balls at random from the urn. In mathematical terms this means that all $\binom{N}{n}$ possible selections of n balls are equally likely.

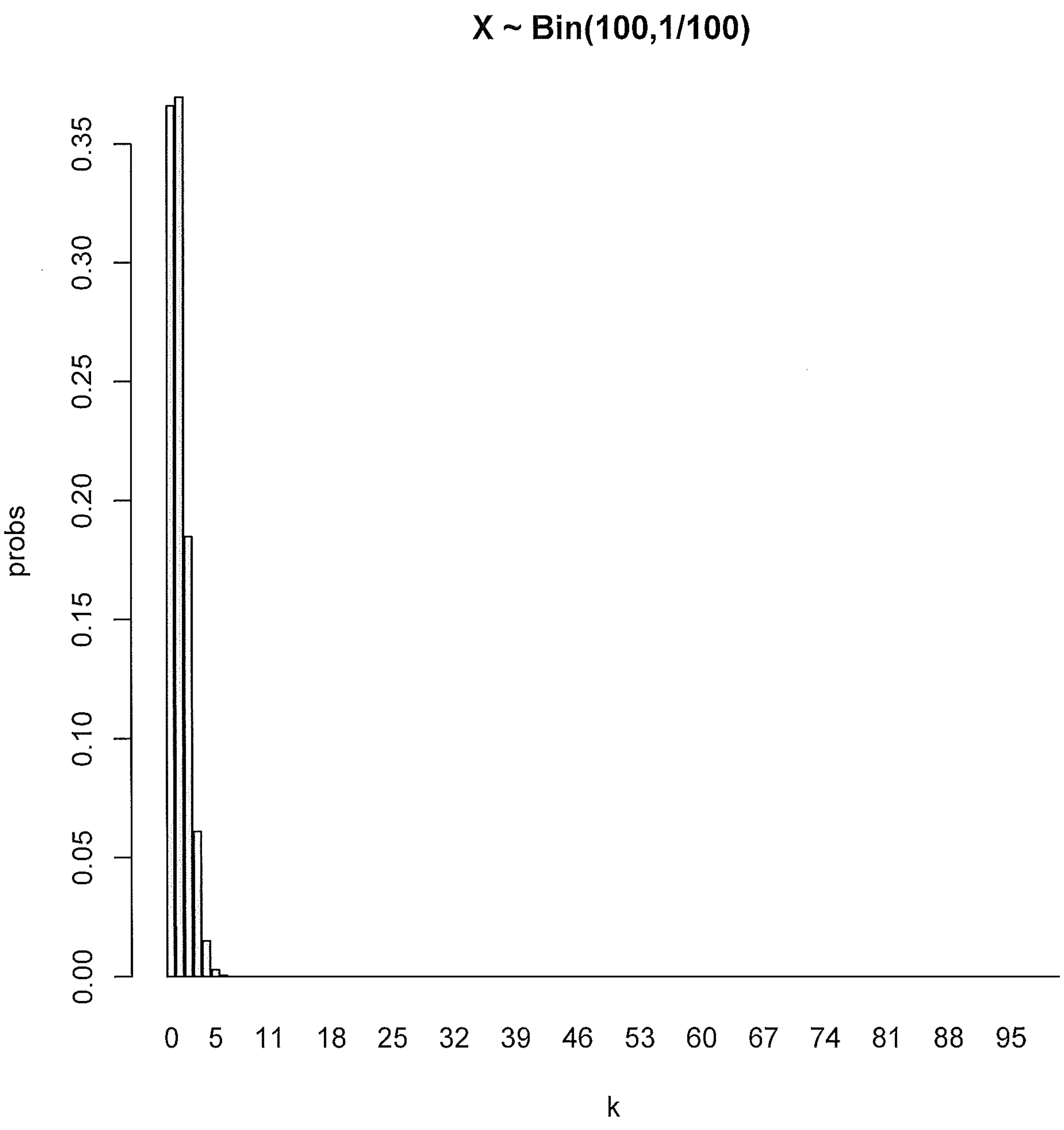
Figure:



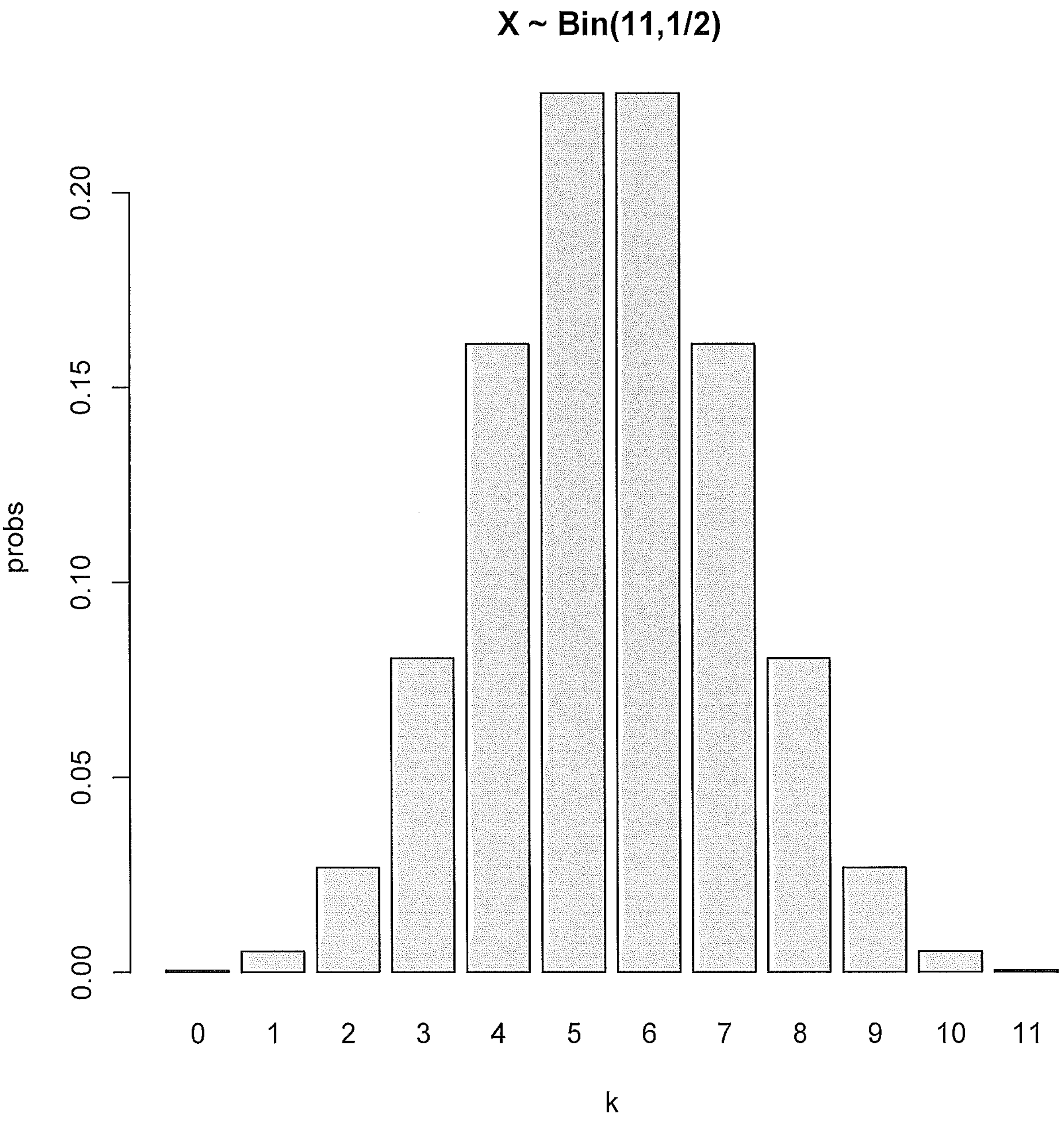
Select n balls at random without replacement.


X ~ Bin(100,1/2)
probs
0.00
0.02
0.04
0.06
0
5
11
18
25
32
39
46
53
60
67
74
81
88
95
k


X ~ Bin(100,1/4)
probs
0.00
0.02
0.04
0.06
0.08
0 5 11 18 25 32 39 46 53 60 67 74 81 88 95
k



X ~ Bin(100,1/100)
probs
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
5
11
18
25
32
39
46
53
60
67
74
81
88
95
k


X ~ Bin(11,1/2)
probs
0.00
0.05
0.10
0.15
0.20
0
1
2
3
4
5
6
7
8
9
10
11
k

<u>Definition</u> : If

$$P(X=k) = \frac{\binom{B}{k}\binom{R}{n-k}}{\binom{N}{n}} \quad \text{for}$$

$$\max(0, n-R) \le k \le \min(n, B)$$

we say that X has the hyper-geometric distribution with parameters n, B and $N = B+R$.

<u>Shorthand</u> :

$$X \sim \text{HiperGeom}(n, B, N).$$

Let X = number of black balls among the n selected. X is a random variable with values k that must satisfy

$$\max(0, n-R) \leq k \leq \min(n, B).$$

We have

$$P(X=k) = \frac{\binom{B}{k}\binom{R}{n-k}}{\binom{N}{n}}$$

The denominator is the number of all possible selection and the numerator is the number of selections with <u>exactly</u> k black and $n-k$ white balls. As with the binomial distribution we can calculate

$$\frac{P(X=k)}{P(X=k-1)} = \frac{(B-k+1)}{k} \cdot \frac{(n-k+1)}{R-n+k}$$

After some calculation we find that

$$\frac{P(X=k)}{P(X=k-1)} > 1 \quad \text{if} \quad k < \frac{(B+1)(n+1)}{N+2}$$

Again we have two cases:

1. $\frac{(B+1)(n+1)}{(N+2)}$ is not an integer.

   Then $k = \left\lfloor \frac{(B+1)(n+1)}{N+2} \right\rfloor$ is the tallest column.

2. $\frac{(B+1)(n+1)}{N+2}$ is an integer.

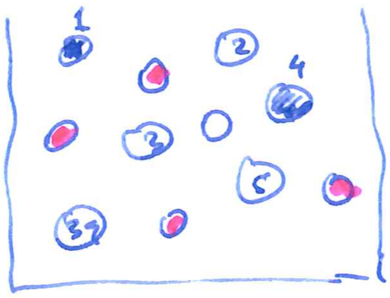
   Then $k = \frac{(B+1)(n+1)}{N+2}$ is still the largest probability but is equal to $P(X=k-1)$.

<u>Example</u>: Lottery.

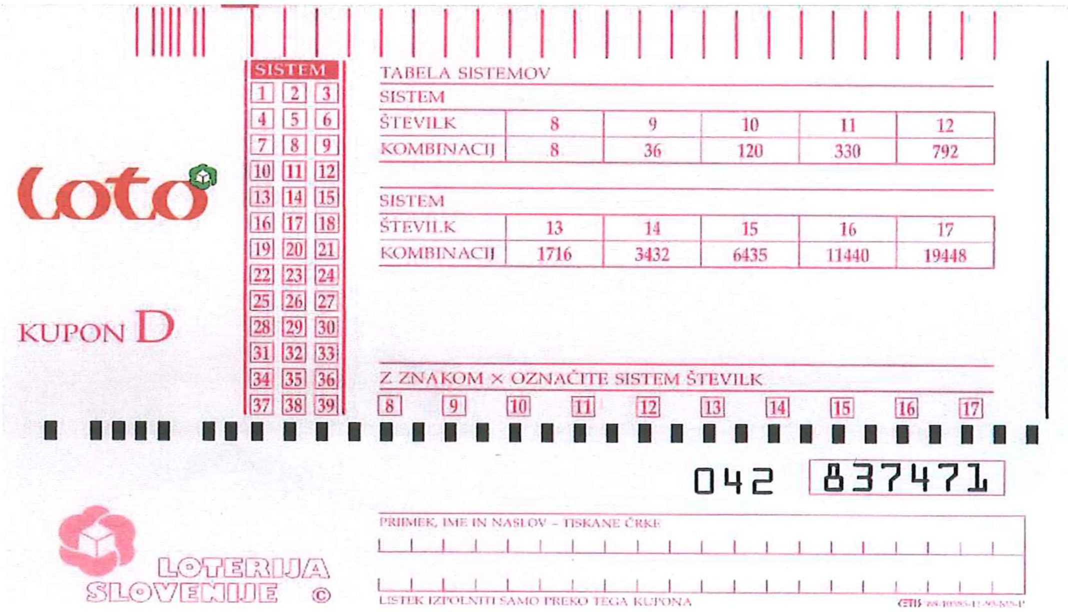
A lottery ticket has 39 numbers. We cross out m numbers where $m = 8, 9, \ldots, 17$;

The winnings depend on the draw. Each week 7 numbers are drawn. If all the numbers are among the ones we crossed out we win a large amount of money. We can translate this problem into a problem involving the hyper-geometric distribution. Imagine you have balls numbered 1-39. You put them into an urn. When we cross the numbers on the Lottery ticket we paint those numbers black. The others we paint red. When 7 balls are drawn the number of correct guesses is $X \sim HiperGeom(7, m, 39)$ where m is the number of balls we painted black.

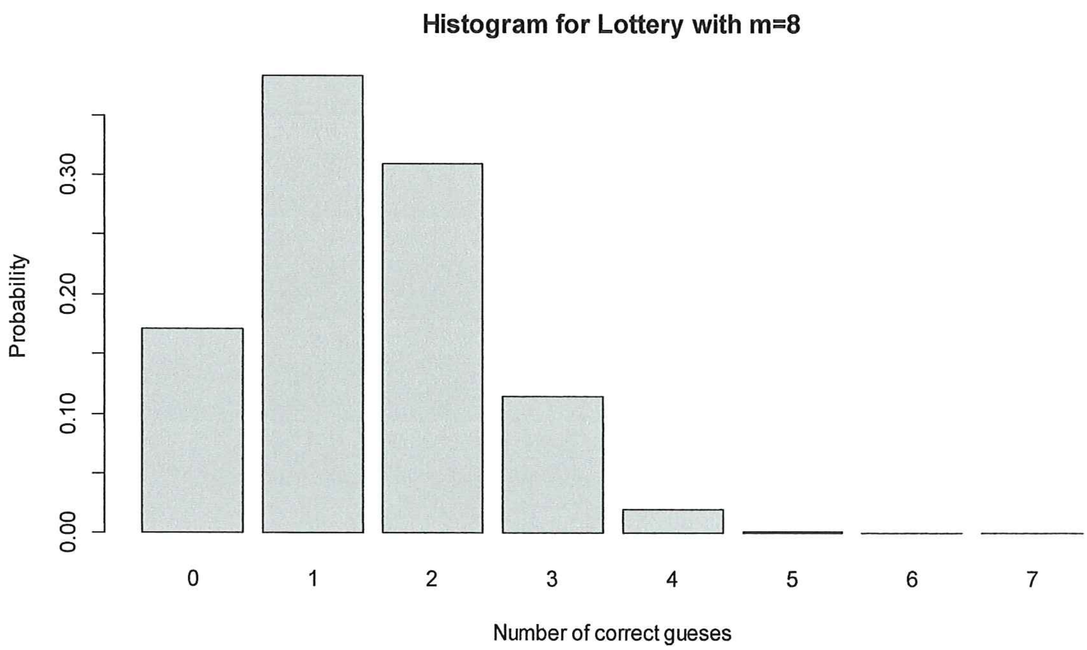
<u>Figure</u>:



Below are distributions for correct number of guesses in Lottery for m=8,13,17. We translated the winning odds in Lottery to a question about the hyper-geometric distribution. The Lottery ticket looks like

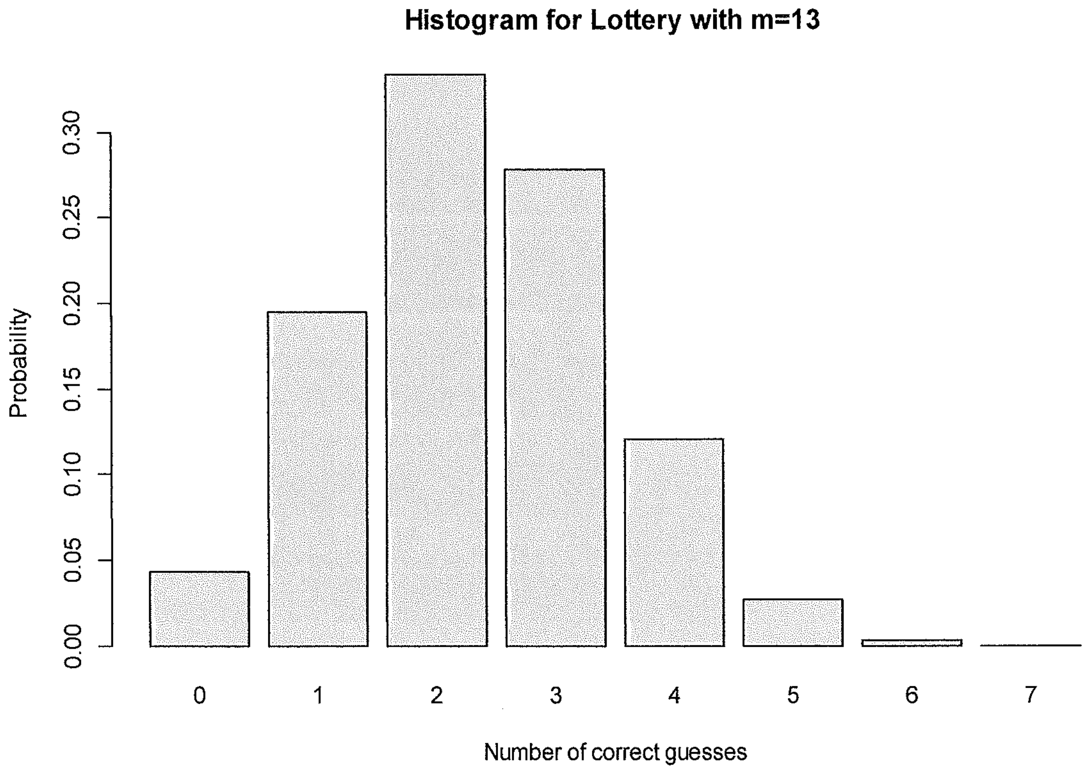


On the ticket the player can cross from m=8 to m=17 numbers. The number of correct guesses is the basis for determining the winnings. The correct guesses are a random variable X. The probability $P(X = 7)$ is the most interesting as it is the probability of jack-pot.



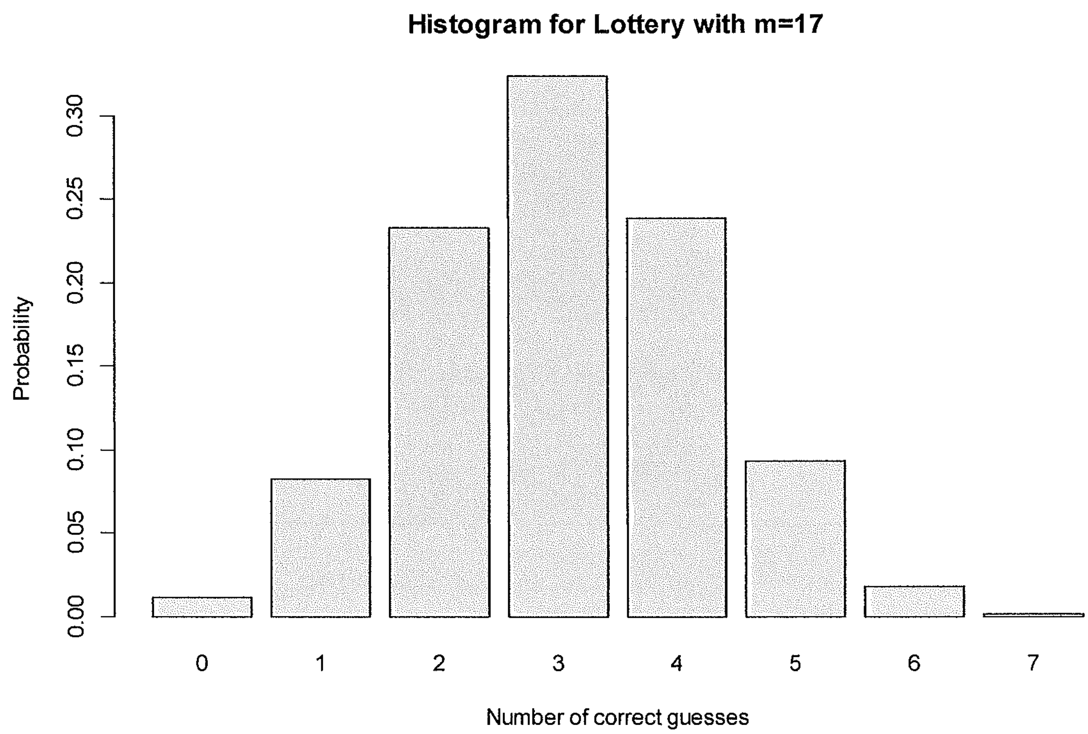
The numerical values of the above probabilities are:

1.709633e-01 3.829577e-01 3.093120e-01 1.145600e-01 2.045714e-02 1.693005e-03 5.643349e-05 5.201244e-07



The numerical values are:

0.0427672254 0.1945908757 0.3335843584 0.2779869653 0.1208638980 0.0271943770 0.0029007336 0.0001115667



Numerical values:

0.011088011 0.082467082 0.232848233 0.323400323 0.238294975 0.092935040 0.017701912 0.001264422

Geometric and negative binomial distribution

Suppose we toss a coin. The tosses are independent and the probability of heads is p.

Let X be the number of tosses until the first heads.

X is a random variable with values in $k = 1, 2, \ldots$

<u>Remark</u>: The values of X can be arbitrarily large. The event $\{X = k\}$ happens if we get $\underbrace{TT\cdots T}_{(k-1)\text{ times}} H$. By independence this implies

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

Definition: If

$$P(X=k) = (1-p)^{k-1} \cdot p$$

we say that X has geometric distribution with parameter p.

Shorthand: $X \sim Geom(p)$.

Example: We play roulette and wait for the number 17 to appear. What is the probability that 17 will not appear in the first n tosses?

$$P(X>n) = P(\underbrace{TT\cdots T}_{n\text{-times}}) = (1-p)^n$$

Because $X \sim Geom(1/37)$ this means

$$P(X>n) = \left(\frac{36}{37}\right)^n$$

If instead of waiting for heads we wait for the appearance of m heads then X is a random variable with values $k = m, m+1, \ldots$. We have

$$P(X = k) = P(\{\text{exactly } m-1 \text{ heads in first } k-1 \text{ tosses}\} \cap \{\text{heads of } k\text{-th toss}\})$$

(independence)

$$= P(\{\text{exactly } m-1 \text{ heads in first } (k-1) \text{ tosses}\}) \times P(\text{heads on toss } k)$$

$$= \underbrace{\binom{k-1}{m-1} p^{m-1} \cdot (1-p)^{(k-1)-(m-1)}}_{\text{by binomial distribution}} \times p$$

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

$$k = m, m+1, \ldots .$$

Definition: we say that X has negative binomial distribution with parameters m and p if

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

$$k = m, m+1, \dots .$$

Shorthand: $X \sim NegBin(m, p)$.

<u>Example</u>: The Polish mathematician Stefan Banach (1897 - 1945) was a chain smoker. He always carried two boxes of matches in his pockets. Assume Banach starts with two boxes of n matches. Then he randomly reaches into the left or right pocket at random with probability $1/2$. At some stage Banach will take the last match from a box but will not notice it. The first time Banach pulls an empty matchbox from his pockets, the number of matches in the other box is random. Call it X. Possible values for X are $k = 0, 1, \dots, n$. We would like to compute the distribution of X.

Let us define

$A = \{X = k\} \cap \{$Banach pulls the empty box from left pocket$\}$

By symmetry and the law of total probabilities we have

$$P(X=k) = 2 \cdot P(A).$$

Let us picture Banach cigarettes.

The event A happens when Banach lights his $(n + (n-k) + 1)$-st cigarette and at that point has pulled a box from his left pocket exactly $(n+1)$-st time.

Declare: reach into the left pocket = „success".

Poisson distribution.

Let us look at the binomial distribution $Bin(n, \frac{\lambda}{n})$ for a given $\lambda > 0$. What happens if n is „large"? Let k be fixed. For $n \geq k$ we have for $X_n \sim Bin(n, \frac{\lambda}{n})$ that

$$P(X = k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

What happens when $n \to \infty$?

From Analysis we know that

$$\left(1 + \frac{x}{n}\right)^n \to e^x \quad \text{for } x \in \mathbb{R}.$$

Rewrite

$$P(X_n = k) = \frac{\lambda^k}{k!} \cdot \underbrace{\frac{n(n-1)\cdots(n-k+1)}{n^k}}_{\to 1} \times$$

$$\times \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \times \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1}$$

We find that

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

This motivates the following definition

<u>Definition</u>: If $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$

for $\lambda > 0$ and $k = 0, 1, 2, \ldots$

we say that X has the Poisson distribution with parameter $\lambda > 0$.

Shorthand: $X \sim Po(\lambda)$.

<u>Remark</u>: From analysis we know that

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}.$$

If $X \sim Po(\lambda)$ we do get

$$\sum_{k=0}^{\infty} P(X=k) = 1.$$

2.2. Continuous distributions

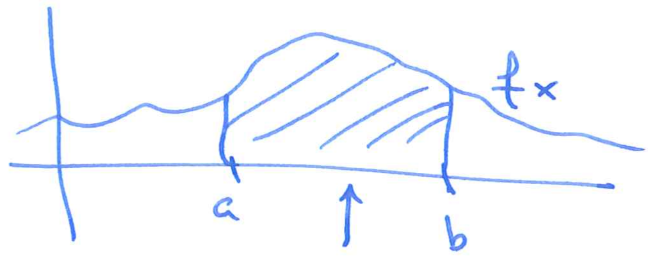
We can imagine „random numbers" that can take any real number as a value. Examples are lifetimes of components, a randomly chosen point on the interval $[0,1]$, ...

Technically, X is still a function on Ω and we require $X^{-1}((a,b])$ to be an event for all $a < b$.

<u>Definition:</u> The distribution of a random variable is given by the probabilities $P(X \in (a,b])$ for all $a < b$.

The idea of continuous random variables is to describe probabilities $P(X \in (a,b])$ by integrals of a single function.

<u>Figure</u>:



Area $= P(a < X \leq b)$

<u>Definition</u>: The random variable X has continuous distribution if there is a non-negative function $f_X(x)$ called the density such that

$$P(a < X \leq b) = \int_a^b f_X(x)\,dx$$

for all $a < b$.

Continuous distributions are used in financial modelling and statistics because of their practicality.

There are standard distributions that are often used.

Normal distribution

The random variable X has normal distribution with parameters μ a $\sigma > 0$ if the density is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Remark: For densities we must have $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

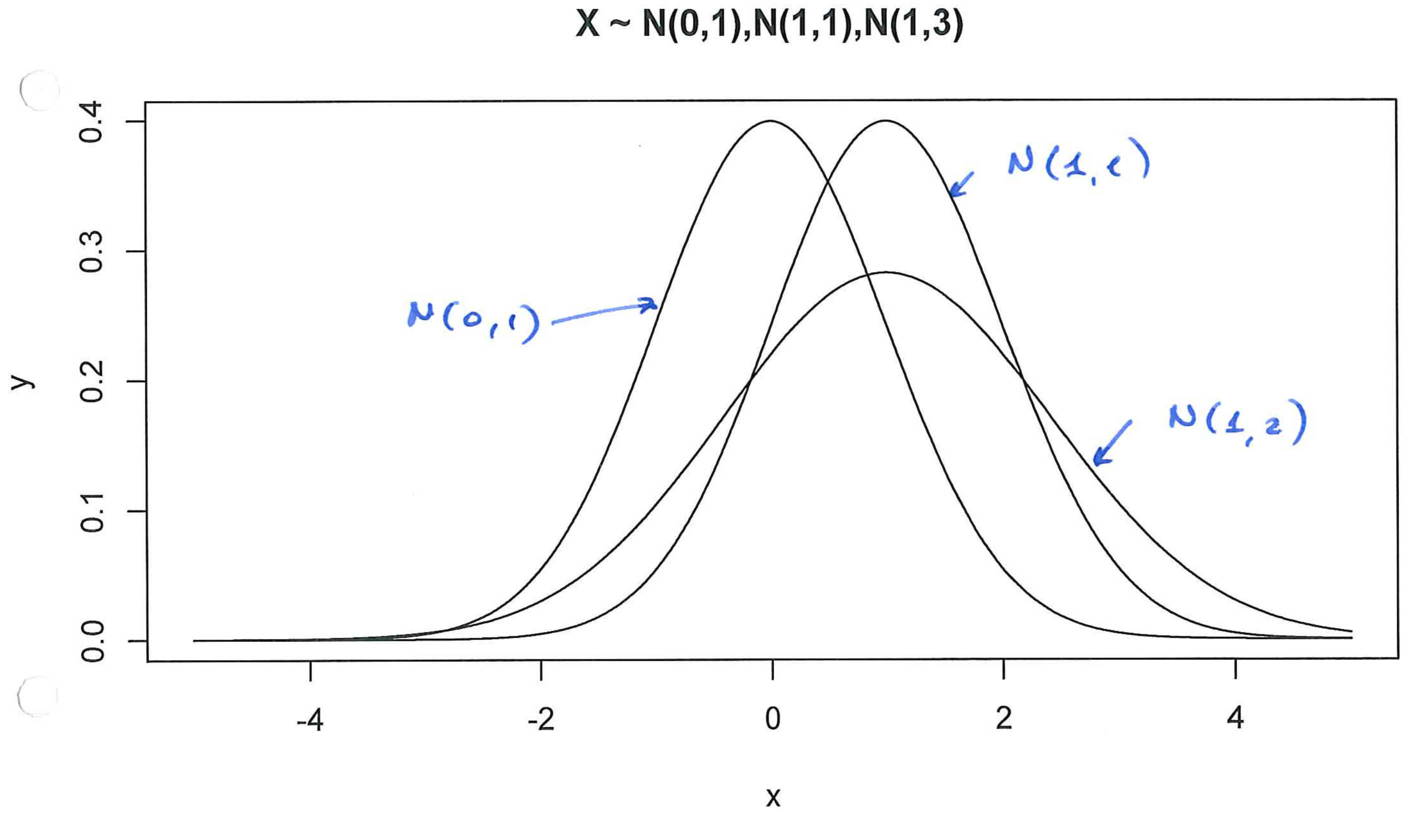
We believe mathematics that f_X is a density.

We will use the notation:

$$X \sim N(\mu, \sigma^2).$$

<u>Remark</u>: The name normal distribution was chosen by the Belgian statistician Adolphe Quetelet because the histograms of human characteristics such as IQ, height and others are close to normal.

<u>Remark</u>: We will give the right interpretation to parameters μ in σ^2 later.



X ~ N(0,1),N(1,1),N(1,3)
N(0,1)
N(1,1)
N(1,2)
y
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

Exponential and gamma distribution

The density of the exponential distribution is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{else.} \end{cases}$$

We say that X has the exponential distribution with parameter λ.

<u>Notation</u> : $X \sim \exp(\lambda)$

The exponential distribution is used to mode lifetimes of electronic components.

To define the gamma distribution recal the definition of the gamma function.

$$\Gamma(x) = \int_0^\infty u^{x-1} \cdot e^{-u} \, du.$$

The most important properties are:

(i) $\Gamma(x+1) = x \cdot \Gamma(x)$

(ii) $\Gamma(n) = (n-1)!$

(iii) $\Gamma(1/2) = \sqrt{\pi}$.

The density

$$f_X(x) = \begin{cases} \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}, & x > 0 \\ 0 & \text{else.} \end{cases}$$

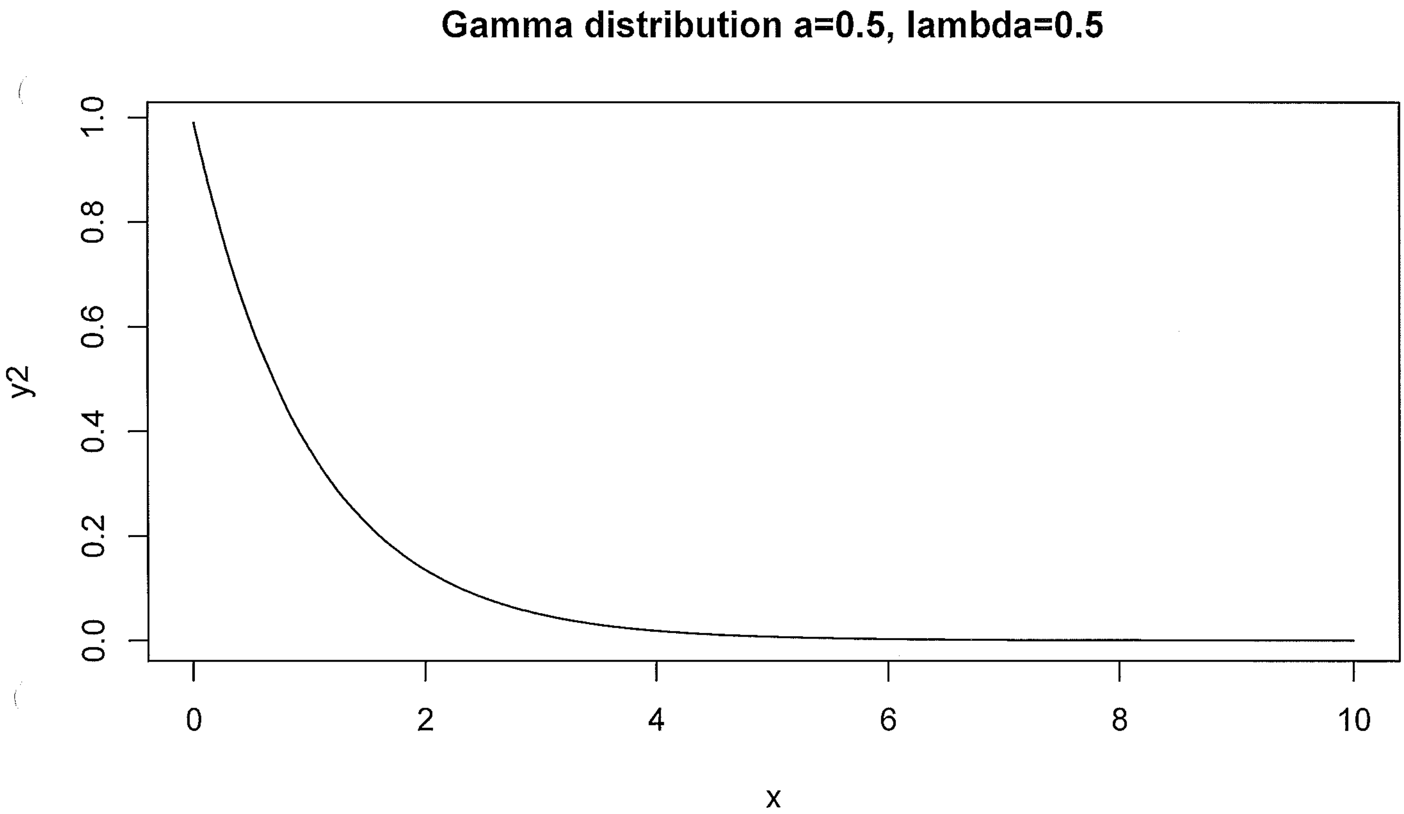
is called the gamma density with parameters a and λ.

a : shape parameter

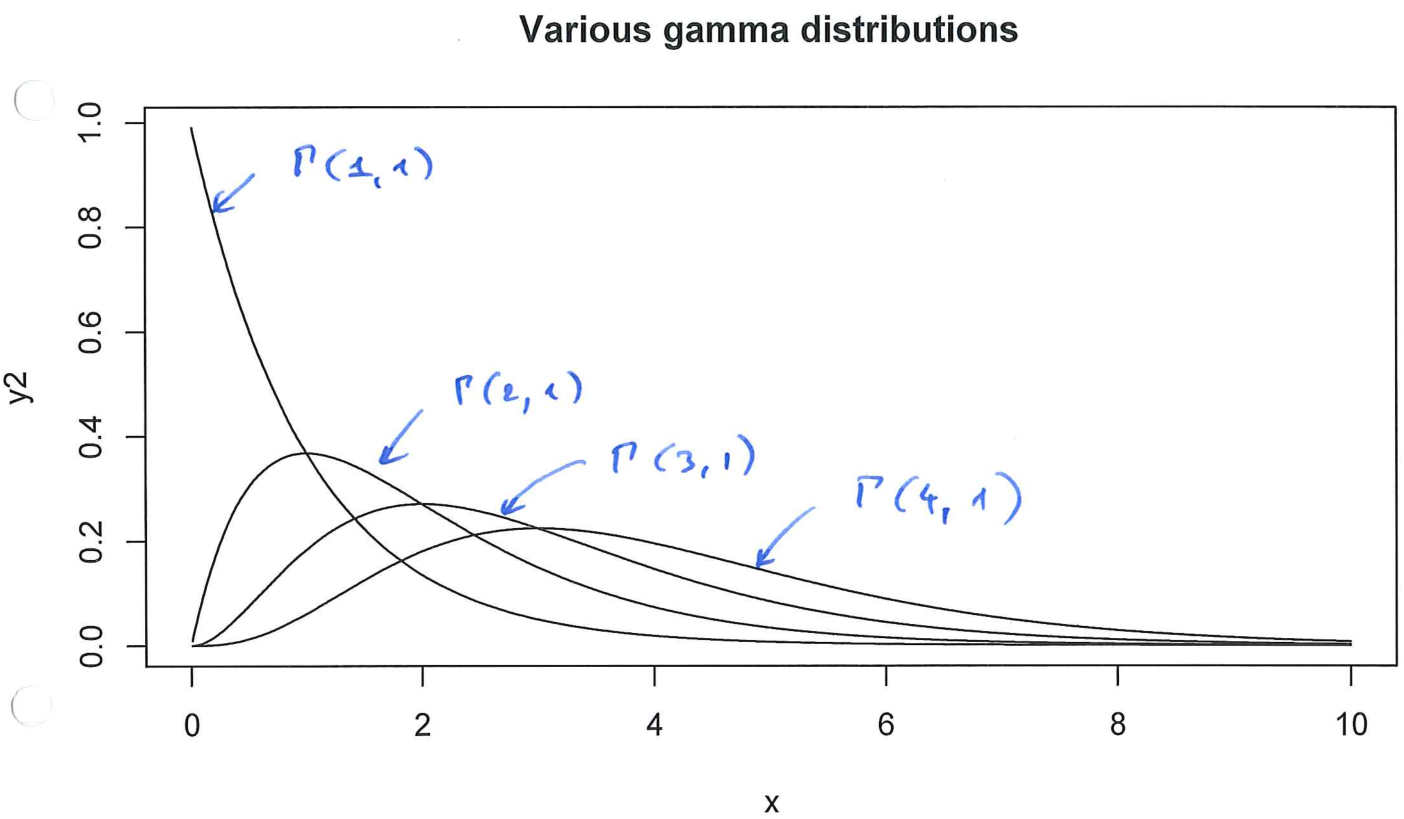
λ : scale parameter.

Remark: Sometimes $1/\lambda$ is used instead of λ.

Notation: $X \sim \Gamma(a, \lambda)$



Gamma distribution a=0.5, lambda=0.5
y2
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x



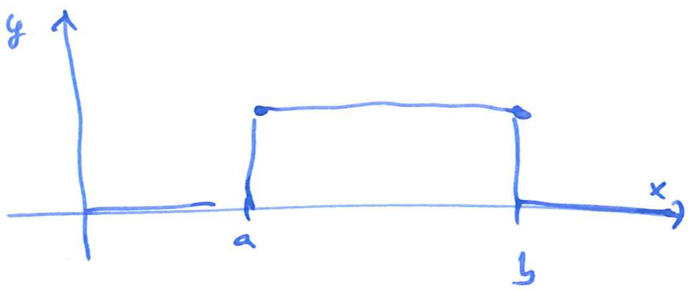
Various gamma distributions
Γ(1,1)
Γ(2,1)
Γ(3,1)
Γ(4,1)
y2
x
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

Uniform distribution

The uniform distribution models the choice of a point at random uniformly on the interval (a, b).
The density is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b. \\ 0 & \text{, else.} \end{cases}$$

Figure:



We use the notation:

$$X \sim U(a, b)$$

Most computer generated random numbers are uniform on $[0, 1]$.

2.3. Function of random variables

Let X be a random variable.

We have that $\{X \le x\}$ is an event. So the probability is defined.

<u>Definition</u>: The distribution function of X is defined as the function

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

<u>Theorem 2.1</u>: Let X be a random variable with distribution function F_X.

(i) F_X is nondecreasing.

(ii) $\lim\limits_{x \to \infty} F_X(x) = 1$, $\lim\limits_{x \to -\infty} F_X(x) = 0$

(iii) F_X is right continuous.

Proof :

(i) For $x < y$ we have $\{X \leq x\} \subseteq \{X \leq y\}$.

It follows that $P(X \leq x) \leq P(X \leq y)$.

(ii) We have $\Omega = \bigcup_{n=1}^{\infty} \{X \leq n\}$.

The sets in the union are increasing so

$$1 = P(\Omega) = \lim_{n \to \infty} P(X \leq n)$$

$$= \lim_{n \to \infty} F_X(n)$$

The conclusion follows because F_X is nondecreasing. The other limit is proved similarly.

(iii) Fix $x \in \mathbb{R}$. Let $x_n \downarrow x$.

We have $\{X \leq x\} = \bigcap_{n=1}^{\infty} \{X \leq x_n\}$.

The sets $\{X \leq x_n\}$ are decreasing. It follows

$$P(X \leq x) = \lim_{n \to \infty} P(X \leq x_n) \quad \text{or}$$

$$F_X(x) = \lim_{n \to \infty} F_X(x_n)$$

The last statement is equivalent to right continuity.

If X has density $f_X(x)$ then

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du.$$

Inversely, if

$$F_X(x) = \int_{-\infty}^{x} g(u)\,du$$

for all $x \in \mathbb{R}$ and a nonnegative g then g is (a version of) the density. If g is continuous at x then

$$F_X'(x) = g(x) = f_X(x).$$

Example : Let $X \sim N(0,1)$ i.e.

the density of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Denote $F_X(x)$ by $\Phi(x)$, i.e.

$$\Phi(x) = \int_{-\infty}^{x} f_X(u)\, du.$$

Let $Y = aX + b$ for $a > 0$. What is the density of Y? We have

$$\begin{aligned} P(Y \le y) &= P(aX + b \le y) \\ &= P(aX \le y - b) \\ &= P\left(X \le \frac{y-b}{a}\right) \\ &= \int_{-\infty}^{\frac{y-b}{a}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \end{aligned}$$

New variable :

$$\frac{v-b}{a} = u \qquad \frac{dv}{a} = du$$

<u>Example</u>: Let $X \sim N(0,1)$ and $Y = X^2$. Density of Y? For $y > 0$

$$P(Y \le y) = P(X^2 \le y)$$

$$= P(-\sqrt{y} \le X \le \sqrt{y})$$

$$= \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$$

<u>Comment</u>: In general

$$\boxed{P(a < X \le b) = F_X(b) - F_X(a)}.$$

For continuous random variables the probabilities $P(X = x) = 0$ for all x and it is not relevant whether we write $<$ or $\le$. We have

$$P(Y \le y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}).$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{\sqrt{y}} e^{-u^2/2} \, du$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{\sqrt{y}} e^{-u^2/2} \, du$$

New variable: $\quad u^2 = v \Rightarrow$

$$2u\,du = dv$$

$$du = \frac{dv}{2\sqrt{v}}$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^y \frac{1}{2\sqrt{v}} e^{-v/2} dv$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^y \frac{1}{\sqrt{v}} e^{-v/2} dv$$

Since $P(Y \geq 0) = 1$ we have

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi} \cdot \sqrt{y}} e^{-y/2}, & y > 0 \\ 0, & \text{else.} \end{cases}$$

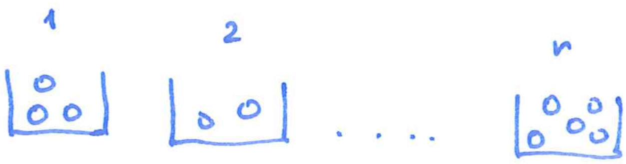
We recognize: $\quad Y \sim \Gamma(1/2, 1/2)$.

3. Multivariate distributions

3.1. Discrete multivariate distributions

Example: Suppose we have r boxes. We are dropping balls into these boxes at random. The probabilities that we hit box k is p_k for $k = 1, 2, \dots, r$; Assume the subsequent drops are independent. There are n balls.

Figure:



We end up with random numbers of balls in boxes. Denote these random numbers by $X_1, X_2, \dots, X_r$. These random numbers will „in the collective" equal to $k_1, k_2, \dots, k_r$ where $k_i \geq 0$ and $\sum_{i=1}^{r} k_i = n$. All the random variables $X_1, X_2, \dots, X_r$ simultaneously take a collection of values. The mathematical objects with several components are vectors.

By analogy we will say that

$\underline{X} = (X_1, X_2, \dots, X_r)$ is a random vector.

The possible values of this random vector are vectors $(k_1, k_2, \dots, k_r)$ with $k_i \geq 0$ and $\sum_{i=1}^{r} k_i = n$.

For discrete random variables we had that the distribution was given by $P(X = x)$ for all possible x. By analogy the distribution of the random vector $\underline{X}$ will be given by probabilities $P(\underline{X} = \underline{x})$ where $\underline{x}$ are possible collections/vectors of values. In the above example we need to compute

$$P(\underline{X} = (k_1, k_2, \dots, k_r)) = P(\underbrace{X_1 = k_1, X_2 = k_2, \dots, X_r = k_r}_{})$$

This notation means

$$\bigcap_{i=1}^{r} \{X_i = k_i\}$$

If we want to hit box 1, k_1 times, box 2 k_2 times, ... the possible disjoint ways for this to happen is to get a sequence of hits

$$n_1, n_2, \ldots\ldots, n_n$$

where k_1 of the $n_1, n_2, \ldots, n_n$ are equal to 1, k_2 are equal to k_2, ...
The probability of such a sequence of hits is by independence

$$p_1^{k_1} \cdot p_2^{k_2} \cdots p_r^{k_r}$$

How many sequences of this type are there? we have n positions

$$\overset{1}{\sqcup}\,\overset{2}{\sqcup}\,\sqcup \,\ldots\ldots\ldots\, \overset{n}{\sqcup}$$ ← Positions

We first choose k_1 positions for 1s.
We can do this in $\binom{n}{k_1}$.

Among the $n-k_1$ positions left we choose k_2 positions for 2s. We can do this in $\binom{n-k_1}{k_2}$ wayse.

By the fundamental theorem of combinatorics the total number of possibilities is

$$\binom{n}{k_1}\binom{n-k_1}{k_2}\cdots\binom{n-k_1-\ \ -k_{r-1}}{k_r}$$

$$= \frac{n!}{k_1!\,(n-\cancel{k_1})!} \cdot \frac{(n-\cancel{k_1})!}{k_2!\,(n-\cancel{k_1}-k_2)!} \cdots \frac{(n-\cancel{k_1}\ -k_r)!}{k_r!\cdot 0!}$$

$$= \frac{n!}{k_1!\cdot k_2!\cdots k_r!}$$

All the sequences are disjoint events with the same probabilities. It follows

$$P(X_1=k_1,\ldots,X_r=k_r) = \frac{n!}{k_1!\cdots k_r!}\, p_1^{k_1}\cdots p_r^{k_r}$$

for $k_i \geq 0$ for $i=1,2,\ldots,r$ and $\sum_{i=1}^{r} k_i = n$.

Definition: For a vector with the above distribution we say that it has the multinomial distribution with parameters n and $\underline{p} = (p_1, p_2, \dots, p_r)$.

Shorthand: $\underline{X} \sim \text{Multinom}(n, \underline{p})$.

Definition: A discrete random vector $\underline{X} = (X_1, X_2, \dots, X_r)$ is a function $\underline{X} : \Omega \to \{\underline{x}_1, \underline{x}_2, \dots, \dots\}$ where $\{\underline{x}_1, \underline{x}_2, \dots\}$ is a finite or countable set of possible values, and such that all components $X_1, X_2, \dots, X_r$ are random variables.

Definition: The distribution of a random vector $\underline{X}$ with values in $\{\underline{x}_1, \underline{x}_2 \dots\}$ is given by probabilities $P(\underline{X} = \underline{x}_k)$ for all $k = 1, 2, \dots$

<u>Remark</u>: Typically we will write

$P(X_1 = x_1, \dots, X_r = x_r)$. When the number of components is small we often write $P(X = x, Y = y)$ or $P(X = x, Y = y, Z = z)$.

<u>Example</u>: Let $N \geq 3$. Choose three numers at random ~~pr~~ from $\{1, 2, \dots, N\}$ without replacement so that all subsets of three numbers are equally likely. Let X be the smallest of the three numbers, Z the largest and Y the remaining one.

<u>Example</u>: If we choose $5, 3, 7$ ~~we~~ we have $X = 3$, $Y = 5$, $Z = 7$.

What is the distribution of (X, Y, Z)?

The possible values are triplets (i, j, k) with $1 \leq i < j < k \leq N$.

We have

$$P(X=i,\ Y=j,\ Z=k)$$

$$= P(\text{ we select the subset } \{i,j,k\})$$

$$= \frac{1}{\binom{N}{3}}$$

What is the distribution of X ?

It has possible values $1, 2, \dots, N-2$.

We notice $\{X=i\} = \bigcup_{i<j<k\le N} \{X=i, Y=j, Z=k\}$

$$P(X=i) = \sum_{i<j<k\le N} P(X=i, Y=j, Z=k)$$

$$= \frac{\binom{N-i}{2}}{\binom{N}{3}}$$

$$= \frac{\frac{(N-i)(N-i-1)}{2}}{\frac{N(N-1)(N-2)}{6}}$$

$$= \frac{3(N-i)(N-i-1)}{N(N-1)(N-2)}$$

Definitions:

(i) The distributions of components of a random vector $\underline{X} = (X_1, X_2, \dots, X_r)$ are called univariate marginal distributions.

(ii) The distribution of subvectors like $(X_1, X_2, \dots, X_s)$ for $s < r$ are called multivariate marginal distributions.

Example (continuation): What is the distribution of (X, Y)? We write

$$\{X = i, Y = j\} = \underbrace{\bigcup_{k=j+1}^{N} \{X = i, Y = j, Z = k\}}_{\text{disjoint events}}$$

We have

$$P(X = i, Y = j) = \sum_{k=j+1}^{N} P(X = i, Y = j, Z = k)$$

$$= \frac{N-j}{\binom{N}{3}}$$

for $1 \leq i < j < N$.

If $\underline{X} = (X_1, \ldots, X_r)$ is a random vector let us write

$$\underline{X}^1 = (X_1, \ldots, X_s) \quad \text{and} \quad \underline{X}^2 = (X_{s+1}, \ldots, X_r).$$

Theorem 3.1 : Let $R = \{\underline{x}_1, \underline{x}_2 \ldots\}$ be the set of possible values of $\underline{X}$.

The marginal distribution of $\underline{X}^1$ is given by

$$P(\underline{X} = \underline{x}^1) = \sum_{(\underline{x}^1, \underline{x}^2) \in R} P(\underline{X} = (\underline{x}^1, \underline{x}^2))$$

$$= \sum_{(\underline{x}^1, \underline{x}^2) \in R} P(\underline{X}^1 = \underline{x}^1, \underline{X}^2 = \underline{x}^2)$$

Proof : We write

$$\{\underline{X}^1 = \underline{x}^1\} = \underbrace{\bigcup_{(\underline{x}^1, \underline{x}^2) \in R} \{\underline{X}^1 = \underline{x}^1, \underline{X}^2 = \underline{x}^2\}}_{\text{disjoint union}}.$$

It follows that

$$P(\underline{X}^1 = \underline{x}^1) = \sum_{(\underline{x}^1, \underline{x}^2) \in R} P(\underline{X}^1 = \underline{x}^1, \underline{X}^2 = \underline{x}^2).$$

Independence

For two events A and B we say that they are independent if $P(A \cap B) = P(A) \cdot P(B)$. We would like to define independence for random variables. If X and Y are to be independent we expect the events $\{X = x\}$ and $\{Y = y\}$ to be independent. So we need $P(X = x, Y = y) = P(X = x) P(Y = y)$

This is the right intuition. For the formal definition we generalize to

$$
\begin{aligned}
P(X \in A, Y \in B) &= \sum_{(x,y) \in A \times B} P(X = x, Y = y) \\
&= \sum_{(x,y) \in A \times B} P(X = x) P(Y = y) \\
&= \Big(\sum_{x \in A} P(X = x) \Big) \Big(\sum_{y \in B} P(Y = y) \Big) \\
&= P(X \in A) \cdot P(Y \in B).
\end{aligned}
$$

Definitions:

(i) Discrete random variables X and Y are independent if

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B)$$

for any two sets A and B.

(ii) Random variables $X_1, X_2, \ldots, X_r$ are independent if

$$P(X_1 \in A_1, \ldots, X_r \in A_r) = P(X_1 \in A_1) \cdots P(X_r = A_r)$$

for any sets $A_1, A_2, \ldots, A_r$.

Remark: The second definition is equivalent to saying that

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r) =$$

$$= P(X_1 = x_1)\, P(X_2 = x_2) \cdots P(X_r = x_r)$$

for all possible values $(x_1, \ldots, x_r)$ of $\underline{X} = (X_1, \ldots, X_r)$.

<u>Example</u>: Let $\underline{X} \sim \text{Multinom}(n, \underline{p})$.

We can easily guess that

$$X_k \sim \text{Bin}(n, p_k) \quad \text{for } k = 1, 2, \dots, r.$$

So

$$P(X_1 = k_1) \dots P(X_r = k_r)$$

$$= \binom{n}{k_1} p_1^{k_1} (1-p_1)^{n-k_1} \cdot \dots \binom{n}{k_r} p_r^{k_r} (1-p_r)^{n-k_r}$$

and

$$P(X_1 = k_1, \dots, X_r = k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r}.$$

Since $P(X_1 = k_1, \dots, X_r = k_r) \neq P(X_1 = k_1) \dots P(X_r = k_r)$ there is no independente.

<u>Example</u>: Suppose the number ~~to~~ of children in a family is Poisson with parameter $\lambda > 0$. Suppose all children are boys or girls with equal probability independently of.

each other. Let X be the number of boys and Y the number of girls.

We compute with $N = X + Y$

$$P(X = k, Y = \ell) = P(X = k, Y = \ell, N = k + \ell)$$

$$= P(X = k, Y = \ell \mid N = k + \ell)\, P(N = k + \ell)$$

$$= \binom{k+\ell}{k} \left(\frac{1}{2}\right)^{k+\ell} \cdot e^{-\lambda} \cdot \frac{\lambda^{k+\ell}}{(k+\ell)!}$$

$$= e^{-\lambda/2} \cdot \frac{(\lambda/2)^k}{k!} \cdot e^{-\lambda/2} \cdot \frac{(\lambda/2)^\ell}{\ell!}$$

On the other hand

$$P(X = k) = \sum_{n=k}^{\infty} P(X = k, N = n)$$

$$= \sum_{n=k}^{\infty} P(X = k \mid N = n)\, P(N = n)$$

$$= \sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{1}{2}\right)^n \cdot e^{-\lambda} \cdot \frac{\lambda^n}{n!}$$

$$= \sum_{n=k}^{\infty} \frac{n!}{k!\,(n-k)!} \left(\frac{1}{2}\right)^n \cdot e^{-\lambda} \frac{\lambda^n}{n!}$$

$$= \frac{e^{-\lambda}\lambda^k (1/2)^k}{k!} \underbrace{\sum_{n=k}^{\infty} \frac{(\lambda/2)^{n-k}}{(n-k)!}}_{e^{\lambda/2}}$$

$$= \frac{e^{-\lambda/2}(\lambda/2)^k}{k!}$$

We have (the same calculation is valid for girls)

$$P(X=k, Y=\ell) = P(X=k)\,P(Y=\ell)$$

so X, Y are independent.

<u>Theorem 3.2</u>: Suppose X, Y are discrete random variables with values in $\{x_1, x_2, \dots\}$ and $\{y_1, y_2, \dots\}$. Suppose we have

$$P(X=x, Y=y) = f(x)\,g(y)$$

for all pairs $(x,y) \in \{x_1, x_2 \dots\} \times \{y_1, y_2 \dots\}$ for some functions $f : \{x_1, x_2 \dots\} \to \mathbb{R}$ and $g : \{y_1, y_2, \dots\} \to \mathbb{R}$. Then X and Y are independent.

<u>Proof</u>: By Theorem 3.1 the marginal distribution are

$$P(X=x) = \sum_y P(X=x, Y=y)$$

$$= \sum_y f(x)\, g(y)$$

$$= f(x) \cdot \underbrace{\sum_y g(y)}_{= c_1}$$

Similarly

$$P(Y=y) = c_2 \cdot g(y).$$

It follows that

$$P(X=x, Y=y) = \cancel{P(X=x)}$$

$$= f(x)\, g(y)$$

$$= \frac{P(X=x)}{c_1} \cdot \frac{P(Y=y)}{c_2}$$

To finish the proof we need $c_1 c_2 = 1$.

But

$$\sum_{x,y} P(X=x, Y=y) = 1 \qquad \text{and}$$

$$\sum_{x,y} P(X=x)P(Y=y)$$

$$= \Big(\sum_x P(X=x)\Big)\Big(\sum_y P(Y=y)\Big)$$

$$= 1 \cdot 1 .$$

Summing up we get

$$\sum_{x,y} P(X=x, Y=y) = \frac{1}{c_1 c_2} \sum_{x,y} P(X=x)P(Y=y)$$

or $\quad 1 = \frac{1}{c_1 \cdot c_2} \cdot 1 \implies c_1 \cdot c_2 = 1 .$

<u>Definition</u>: Random vectors $\underline{X}$ and $\underline{Y}$ are independent if

$$P(\underline{X} \in A, \underline{Y} \in B) = P(\underline{X} \in A) \cdot P(\underline{Y} \in B).$$

for all sets A, B.

Remark: The definition is equivalent to

$$P(\underline{X} = x, \underline{Y} = y) = P(\underline{X} = x)\, P(\underline{Y} = y)$$

for all pairs of possible values.

Theorem 3.2 is valid in the following form:

If $P(\underline{X} = x, \underline{Y} = y) = f(x)\, g(y)$ for some functions f, g then $\underline{X}, \underline{Y}$ are independent.

3.2. Expected value

Example: In one of on-line games you have 12 tickets

$$\boxed{1}\ \boxed{1}\ \boxed{1}\ \boxed{1}\ \boxed{2}\ \boxed{2}\ \boxed{3}\ \boxed{D}\ \boxed{S}\ \boxed{S}\ \boxed{S}\ \boxed{S}$$

The tickets are turned around and randomly permuted. The player sees

The player then turns tickets from left to right until the ticket

$\boxed{S}$ = STOP appears. Example:

$\boxed{1}$ $\boxed{2}$ $\boxed{D}$ $\boxed{1}$ $\boxed{S}$

The payout is the sum of all numbers, multiplied by 2 if $\boxed{D}$ = double appears among the tickets. In the above example the payout is 8.

What is the fair price for this game?

Suppose we played this game many times. We can interpret the payout as a random variable, X say. Possible values of X are $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 16, 18, 20, 22\}$.

We have denoting possible values of X by $\{x_1, x_2, \ldots\ldots, x_{17}\}$:

$$\frac{v_1 + \cdots + v_n}{n}$$

$$= \sum_{k=1}^{17} x_k \cdot \underbrace{\frac{\text{\# of occurrencies of } x_k}{n}}_{\approx P(X = x_k)}$$

So the „long term" average will be

$$\sum_{k=1}^{17} x_k P(X = x_k)$$

We will call this average the expected value of a random variable.

<u>Definition</u> : Let X be a discrete random variable with values $\{x_1, x_2, \ldots\}$. The <u>expected value</u> $E(X)$ is defined as

$$\boxed{E(X) = \sum_{x_k} x_k P(X = x_k)}$$

<u>Technical note</u>: We say that X exist if the sum

$$\sum_{x_k} |x_k| \cdot P(X = x_k) \text{ converges.}$$

If f is a function then $Y = f(X)$ is again a discrete random variable. If we „repeat" X we also „repeat" Y. The expectation $E(Y)$ will be approximately

$$\frac{f(v_1) + \dots + f(v_n)}{n} \approx \sum_{x_k} f(x_k) P(X = x_k)$$

by exactly the same argument as before. Formally, we state:

<u>Theorem 3.3</u>: If X is a discrete random variable with values in $\{x_1, x_2, \dots\}$. Let $f : \{x_1, \dots, \dots\} \to \mathbb{R}$.

We have

$$E[f(X)] = \sum_{x_k} f(x_k) P(X = x_k)$$

Proof : Denote $Y = f(X)$. Possible values are $\{y_1, y_2, \dots\}$. By definition

$$
\begin{aligned}
E[f(X)] &= E(Y) \\
&= \sum_{y_\ell} y_\ell \, P(Y = y_\ell) \\
&= \sum_{y_\ell} y_\ell \sum_{\{x_k : f(x_k) = y_\ell\}} P(X = x_k) \\
&= \sum_{y_\ell} \sum_{\{x_k : f(x_k) = y_\ell\}} f(x_k) \, P(X = x_k) \\
&= \sum_{x_k} f(x_k) \, P(X = x_k)
\end{aligned}
$$

Technical note : We say that $E(f(X))$ exists if the sum

$$\sum_{x_k} |f(x_k)| \, P(X = x_k)$$

exists.

Examples:

(i) Let $X \sim \mathrm{Bin}(n,p)$. We compute

$$E(X) = \sum_{k=0}^{n} k \cdot P(X=k)$$

$$= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \qquad q := 1-p$$

$$= \sum_{k=1}^{n} n \cdot p \cdot \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)}$$

$$= n \cdot p \underbrace{\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)}}_{= (p+q)^{n-1} = 1}$$

$$= n \cdot p$$

Similarly

$$E(X^2) = \sum_{k=0}^{n} k^2 \cdot P(X=k)$$

$$= \sum_{k=1}^{n} [k(k-1)+k] \binom{n}{k} p^k q^{n-k}$$

$$= \sum_{k=2}^{n} k(k-1) \binom{n}{k} p^k q^{n-k}$$

$$+ \underbrace{\sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k}}_{= n \cdot p}$$

$$= \sum_{k=2}^{n} n(n-1)\binom{n-2}{k-2}\cdot p^2\, p^{k-2}\, q^{(n-2)-(k-2)}$$

$$+ np$$

$$= n(n-1)p^2 \underbrace{\sum_{k=2}^{n}\binom{n-2}{k-2} p^{k-2} q^{(n-2)-(k-2)}}_{=(p+q)^{n-2} = 1}$$

$$+ np$$

$$= n(n-1)p^2 + np$$

$$= n^2p^2 + npq$$

(ii) Let $X \sim \text{NegBin}(m,p)$.

We have

$$P(X=k) = \binom{k-1}{m-1} p^m q^{k-m}$$

for $k = m, m+1, \ldots$. We compute

$$E(X) = \sum_{k=m}^{\infty} k\cdot\binom{k-1}{m-1} p^m\cdot q^{k-m}$$

$$= \sum_{k=m}^{\infty} \binom{(k+1)-1}{(m+1)-1}\cdot m\cdot p^m q^{k-m}$$

$$= \sum_{k=m}^{\infty} \frac{m}{p}\cdot\binom{(k+1)-1}{(m+1)-1} p^{m+1} q^{(k+1)-(m+1)}$$

$$= \frac{m}{p} \cdot \underbrace{\sum_{k=m}^{\infty} \binom{(k+1)-1}{(m+1)-1} p^{m+1} q^{(k+1)-(m+1)}}$$

= 1, because this is the sum of all probabilities in the NegBin$(m+1, p)$ distribution

$$= \frac{m}{p}.$$

In a similar way we find that

$$E(X^2) = \sum_{k=m}^{\infty} k^2 \binom{k-1}{m-1} p^m \cdot q^{k-m}$$

$$= \sum_{k=m}^{\infty} [k(k+1) - k] \binom{k-1}{m-1} p^m q^{k-m}$$

$$= \sum_{k=m}^{\infty} \binom{(k+2)-1}{(m+2)-1} \frac{m(m+1)}{p^2} p^{m+2} q^{k-m}$$

$$- \frac{m}{p}$$

$$= \frac{m(m+1)}{p^2} - \frac{m}{p}$$

$$= \frac{m^2}{p^2} + \frac{m}{p}\left(\frac{1}{p} - 1\right)$$

$$= \frac{m^2}{p^2} + \frac{m \cdot q}{p^2}$$

(iii) Let $X \sim \text{HiperGeom}(n, B, N)$.

Let us agree that $\binom{a}{b} = 0$ if $b > a$ or $b < 0$. We compute

$$E(X) = \sum_k k \cdot \frac{\binom{B}{k}\binom{R}{n-k}}{\binom{N}{n}}$$

$$= \sum_k \frac{B\binom{B-1}{k-1} \cdot \binom{R}{(n-1)-(k-1)} \cdot n}{\binom{N-1}{n-1} \cdot N}$$

$$= n \cdot \frac{B}{N} \cdot \underbrace{\sum_k \frac{\binom{B-1}{k-1}\binom{R}{(n-1)-(k-1)}}{\binom{N-1}{n-1}}}_{= 1}$$

$= 1$, because this is the sum of all probs. in HiperGeom$(n-1, B-1, N-1)$ distribution.

$$= n \cdot \frac{B}{N}$$

The most important theoretical property of expectation is linearity.

<u>Theorem 3.4</u> : Let X, Y be discrete random variables. We have

$$\boxed{E(aX+bY) = aE(X) + bE(Y)}$$

<u>Proof</u> : Denote $Z = aX + bY$. Z is a discrete random variable with values $\{z_1, z_2, \dots\}$. We have

$$E(Z) = \sum_{z_m} z_m \cdot P(Z = z_m)$$

$$= \sum_{z_m} z_m \cdot \sum_{\{(x_k, y_\ell) : a x_k + b y_\ell = z_m\}} P(X = x_k, Y = y_\ell)$$

$$= \sum_{z_m} \sum_{-''-} (a x_k + b y_\ell) \, P(-''-)$$

$$= \sum_{x_k, y_\ell} (a x_k + b y_\ell) P(X = x_k, Y = y_\ell)$$

$$= a \cdot \sum_{x_k, y_\ell} x_k P(X = x_k, Y = y_\ell)$$

$$+ b \sum_{x_k, y_\ell} y_\ell P(X = x_k, Y = y_\ell)$$

$$= a \cdot \sum_{x_k} x_k P(X = x_k)$$

$$+ b \cdot \sum_{y_\ell} y_\ell P(Y = y_\ell)$$

$$= E(X) + E(Y)$$

Technical note : We assume that $E(X)$ and $E(Y)$ exist. In this case $E(aX + bY)$ exist as well.

Remark : We have derived that

$$E(X) = \sum_{x_k, y_\ell} x_k P(X = x_k, Y = y_\ell)$$

A consequence of Theorem 3.4 is that linearity is valid for more general linear combinations.

If $X_1, X_2, \ldots, X_r$ are random variables such that $E(X_k)$ exists then

$$\boxed{E\left[\sum_{k=1}^{r} a_k X_k\right] = \sum_{k=1}^{r} a_k E(X_k).}$$

Finally, we state

<u>Theorem 3.5</u>: Let $\underline{X}$ be a discrete random vector in $\mathbb{R}^n$ and let $f: \mathbb{R}^n \to \mathbb{R}$ be a function. We have

$$E[f(\underline{X})] = \sum_{\underline{x}_k} f(\underline{x}_k) P(\underline{X} = \underline{x}_k)$$

<u>Proof</u>: The proof is identical to the proof of Theorem 3.3.

<u>Example</u>: Let $\underline{X} \sim \text{Multinomial}(n, \underline{p})$. What is $E(X_k \cdot X_\ell)$? We know that

$X_k + X_\ell \sim \text{Bin}(n, p_k + p_\ell)$ so

$$E[(X_k + X_\ell)^2] = n(p_k + p_\ell)(1 - p_k - p_\ell) + n^2(p_k + p_\ell)^2$$

$$\parallel$$

$$E[X_k^2 + 2X_k X_\ell + X_\ell^2]$$

$$\parallel$$

$$E(X_k^2) + 2E(X_k X_\ell) + E(X_\ell^2)$$

$$= n p_k (1 - p_k) + n^2 p_k^2$$

$$+ 2 E(X_k X_l)$$

$$+ n p_l (1 - p_l) + n^2 p_l^2$$

This is an equation for $E(X_k X_l)$ from which we compute

$$E(X_k X_l) = - n p_k p_l + n^2 p_k p_l$$

<u>Definition</u> : A random variable X with values in $\{0,1\}$ is called an indicator or a Bernoulli random variable. We denote $p = P(X=1)$

<u>Shorthand</u> : $X \sim \text{Bernoulli}(p)$.

By definition

$$E(X) = 0 \cdot P(X=0) + 1 \cdot P(X=1) = p$$

<u>Remark</u> : Since $X : \Omega \to \{0,1\}$ we can denote $A = \{X = 1\}$ which is an event.

Every indicator is associated with an event A. We will write I_A or 1_A for the indicator of A i.e. the random variable X, for which $X(\omega) = 1$ if $\omega \in A$ and 0 else.

In many cases complicated random variables can be written as linear combinations of ~~more~~ ~~complicate~~ simpler random variables. Expectctions can then be computed in simpler ways using linearity.

<u>Example</u>: Let us return to the first example.

$$\overset{1}{\boxed{1}}\ \overset{2}{\boxed{1}}\ \overset{3}{\boxed{1}}\ \overset{4}{\boxed{1}}\ \overset{1}{\boxed{2}}\ \overset{2}{\boxed{2}}\ \boxed{3}\ \boxed{0}\ \boxed{5}\ \boxed{5}\ \boxed{5}\ \boxed{5}$$

Label the tickets with 1 from 1 to 4, and the tickets with 2 from 1 to 2.

If we know whether a ticket has contributed to the final payout and whether $\boxed{D}$ ~~appeared~~ appeared we can reconstruct the payout.

Example:

	1	2	3	4	1	2						
	$\boxed{1}$	$\boxed{1}$	$\boxed{1}$	$\boxed{1}$	$\boxed{2}$	$\boxed{2}$	$\boxed{3}$	$\boxed{D}$	$\boxed{S}$	$\boxed{S}$	$\boxed{S}$	$\boxed{S}$
	↑	↑	↑	↑	↑	↑	↑	↑	.			
Contr.:	Y	N	N	Y	N	Y	N	Y				

Payout $= 4 \times 2 = 8$

Define events

$A_{1,i} = \{$ ticket $\boxed{1}^{\,i}$ contributed, but $\boxed{D}$ did not $\}$

$A_{2,i} = \{$ -"- , and $\boxed{D}$ did $\}$

$B_{1,i} = \{$ ticket $\boxed{2}^{\,i}$ contributed, $\boxed{D}$ did not $\}$

$B_{2,i} = \{$ -"- , $\boxed{D}$ did $\}$

$C_1 = \{$ $\boxed{3}$ contributed, $\boxed{D}$ did not $\}$

$C_2 = \{$ -"- , $\boxed{D}$ did $\}$.

We have

$$X = \sum_{i=1}^{4} 1_{A_{1,i}} + 2 \sum_{i=1}^{4} 1_{A_{2,i}}$$

$$+ 2 \sum_{i=1}^{2} 1_{B_{1,i}} + 4 \sum_{i=1}^{2} 1_{B_{2,i}}$$

$$+ 3\, 1_{C_1} + 6\, 1_{C_2}$$

By symmetry

$$P(A_{1,i}) = P(B_{1,i}) = P(C_1)$$

and

$$P(A_{2,i}) = P(B_{2,i}) = P(C_2).$$

This means that

$$E(X) = 11 \cdot P(A_{1,1}) + 22 \cdot P(A_{2,1})$$

We compute $P(A_{1,1})$ and $P(A_{2,1})$ by noticing that if we only look at tickets

$\boxed{1}$ $\boxed{D}$ $\boxed{S}$ $\boxed{S}$ $\boxed{S}$ $\boxed{S}$ among the 12 permuted tickets they too are randomly permuted. We say that the induced permutation is random. It follows that $A_{1,1}$ happens if we see

$$\boxed{1}\ \boxed{3}\ *\ *\ *\ *$$

The probability is

$$\frac{1}{6} \times \frac{4}{5} = \frac{2}{15}$$

The event $A_{2,1}$ happens if we see

$$\boxed{1}\ \boxed{D}\ *\ *\ *\ * \quad \text{or} \quad \boxed{D}\ \boxed{1}\ *\ *\ *\ *.$$

The probability is

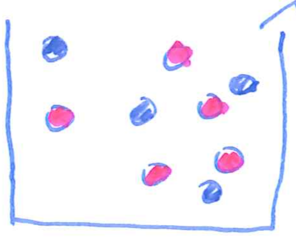
$$2 \cdot \frac{1}{6} \cdot \frac{1}{5} = \frac{1}{15}$$

Finally,

$$E(X) = 11 \cdot \frac{2}{15} + 22 \cdot \frac{1}{15} = \frac{44}{15}$$

$$\doteq 2.93$$

<u>Example</u>: The hyper-geometric distribution is created by selecting balls out of a box.



Select n balls at random

$X = \#$ of black balls

We can imagine that balls are selected one by one at random until we have n balls. Define

$$I_k = \begin{cases} 1, & \text{if the } k\text{-th ball is black.} \\ 0, & \text{else,} \end{cases}$$

for $k = 1, 2, \ldots, n$. We have

$$E(X) = E(\underbrace{I_1 + \cdots + I_n}_{= X})$$

$$= E(I_1) + \cdots + E(I_n)$$

$$= P(I_1 = 1) + P(I_2 = 1) + \cdots + P(I_n = 1)$$

But the k-th ball is equally likely to be any of the N balls. We are asking this question before the selection process begins. This means that

$$P(I_1 = 1) = P(I_2 = 1) = \cdots = P(I_n = 1).$$

But

$$P(I_1 = 1) = P(\text{first ball selected black})$$

$$= \frac{B}{N}.$$

It follows that

$$E(X) = n \cdot \frac{B}{N}.$$

<u>Comment</u>: The idea to write X as a linear combination of indicators is called the method of indicators.

3.3. Joint continuous distributions

For a continuous random variable X with density f_X we have

$$P(a \le X \le b) = \int_a^b f_X(x)\,dx.$$

More generally, for a set A we can say

$$P(X \in A) = \int_A f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} f_X(x)\,\chi_A(x)\,dx,$$

where χ_A is the characteristic function of the set A. This last form has an easy extension to $\mathbb{R}^2$, $\mathbb{R}^3$ or $\mathbb{R}^4$.

For $\mathbb{R}^2$ we can say that

$$P(\underline{X} \in A) = \iint_A f_{\underline{X}}(x, y)\,dx\,dy$$

for an appropriate non-negative function. In $\mathbb{R}^3$

we have for $\underline{X} = (X_1, X_2, X_3)$

$$P(\underline{X} \in A) = \iiint_A f_{\underline{X}}(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3 .$$

In probability we will write single integrals even in higher dimensions.
If $A \subseteq \mathbb{R}^n$ we will write

$$\iint_A \cdots \int f(x_1, x_2 \dots x_n)\, dx_1\, dx_2 \dots dx_n$$

$$= \int_A f(\underline{x})\, d\underline{x}$$

<u>Definition</u>: If for a random vector $\underline{X}$ we have

$$P(\underline{X} \in A) = \int_A f_{\underline{X}}(\underline{x})\, d\underline{x}$$

for a non-negative function $f_{\underline{X}} : \mathbb{R}^n \to \mathbb{R}$ and all (reasonable) sets A we say that $\underline{X}$ has continuous distribution with density $f_{\underline{X}}$.

Technical note: In more dimensions in general we say that the distribution of $\underline{X}$ is described by probabilities $P(\underline{X} \in A)$ for all reasonable sets $A \subseteq \mathbb{R}^n$.

„Reasonable" means all sets that are formed from open sets by complements, countable unions and countable intersections. Such sets are called Borel sets.

Example: Let (X, Y) be a random vector with density $f_{X,Y}(x,y)$ given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}}$$

for $\rho \in (-1,1)$. Let us check that $f_{X,Y}$ is a density. This means that it is non-negative and integrates to 1. We know that

$$\frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1,$$

because the latter is the integral of the normal density. We integrate

$$\int_{\mathbb{R}^2} f_{X,Y}(x,y)\, dx\, dy =$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{\mathbb{R}^2} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}} dx\, dy$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \underbrace{\int_{-\infty}^{\infty} dx \cdot \int_{-\infty}^{\infty} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}} dy}_{\text{This is called Fubini's theorem.}}$$

$$= (*)$$

We note

$$(x^2 - 2\rho x y + y^2)/(1-\rho^2)$$

$$= [(y-\rho x)^2 + (1-\rho^2)x^2]/(1-\rho^2)$$

$$= \frac{(y-\rho x)^2}{1-\rho^2} + x^2$$

$$(*) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \cdot \underbrace{\int_{-\infty}^{\infty} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}} dy}_{=\sqrt{2\pi}\cdot\sqrt{1-\rho^2}}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx$$

$$= 1.$$

Let us continue with this example and compute $P(X \geq 0, Y \geq 0)$.

In other words

$$P(X \geq 0, Y \geq 0) = P\left((X,Y) \in [0,\infty)^2\right).$$

By definition

$$P\left((X,Y) \in [0,\infty)^2\right)$$

$$= \int_{[0,\infty)^2} f_{X,Y}(x,y)\, dx\, dy$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{[0,\infty)^2} e^{-\frac{x^2 - 2\rho x y + y^2}{2(1-\rho^2)}}\, dx\, dy$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^\infty e^{-x^2/2}\, dx \cdot \underbrace{\int_0^\infty e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}\, dy}_{\text{New variable: } \frac{y-\rho x}{\sqrt{1-\rho^2}} = u}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^\infty e^{-x^2/2}\, dx \cdot \int_{-\frac{\rho x}{\sqrt{1-\rho^2}}}^{\infty} e^{-u^2/2}\, du \cdot \sqrt{1-\rho^2}$$

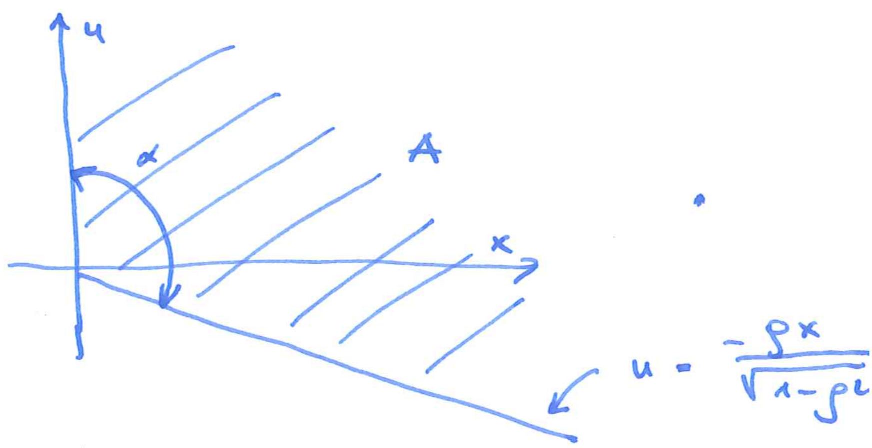
This last integral is the integral of the function

$$f(x,u) = \frac{1}{2\pi} e^{-\frac{x^2+u^2}{2}}$$ over the set

$$A = \{(x,u) : x \geq 0,\ u \geq -\frac{\rho x}{\sqrt{1-\rho^2}}\}$$

by Fubini.

Figure :



We observe :

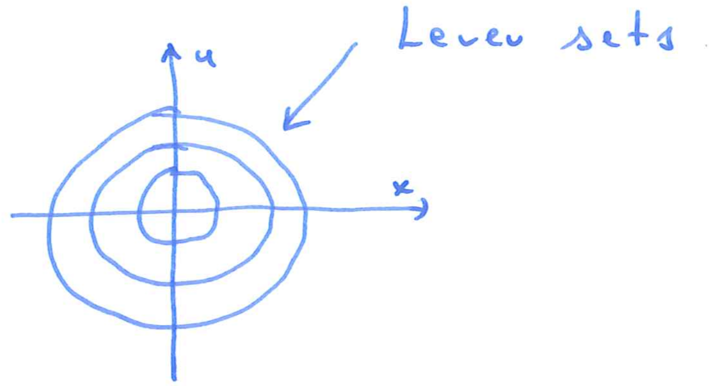
(i) $f(x,u)$ integrates to 1.
We get this by taking $\rho = 0$ in the previous example.

(ii) The function

$f(x,u)$ is rotationally

symmetric.

Figure :

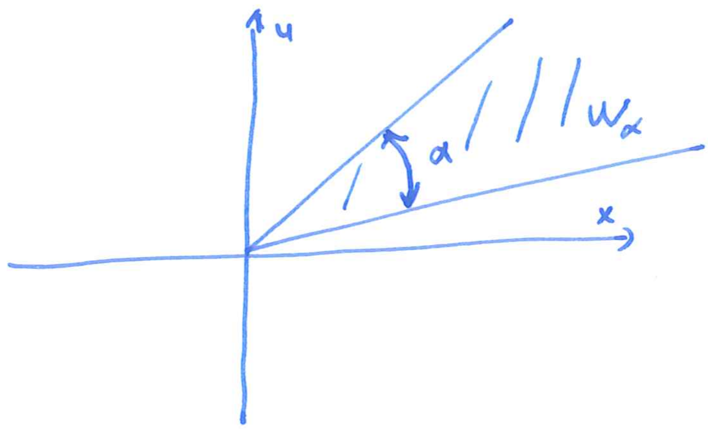


The integral over a wedge

of angle α is proportional

to the angle

Figure :



$$\int_{W_\alpha} f(x,u)\,dx\,du = \frac{\alpha}{2\pi}$$

In our case the angle α equals

$$\alpha = \frac{\pi}{2} + \operatorname{arctg}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right).$$

Finally we have

$$P(X \geq 0, Y \geq 0) =$$

$$= \frac{1}{4} + \frac{1}{2\pi} \operatorname{arctg}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

Marginal distributions

Suppose (X,Y) has density $f_{X,Y}(x,y)$.
What is the density of X ?

We know from the ~~1st~~ 2nd Chapter that if for any $a < b$

$$P(a \leq X \leq b) = \int_a^b g(x)\,dx \quad \text{this}$$

implies that $g(x) = f_X(x)$.

we compute

$$P(a \le X \le b) = P(a \le X \le b, Y \in \mathbb{R})$$

$$= P((X,Y) \in [a,b] \times \mathbb{R})$$

$$= \int_{[a,b]\times\mathbb{R}} f_{X,Y}(x,y)\,dx\,dy$$

$$= \int_a^b dx \cdot \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy}_{\text{This is a function of } x\text{, say } g(x)}$$

$$= \int_a^b g(x)\,dx.$$

Theorem 3.6: Let (X,Y) have the density $f_{X,Y}(x,y)$. We have

$$\boxed{\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \end{aligned}}$$

<u>Proof</u>: Done already.

<u>Comments</u>

(i) The two formulae in Theorem 3.6 are called formulae for <u>marginal densities</u>.

(ii) A rigorous statement must include the assumptions that $x \mapsto f_{X,Y}(x,y)$ and $y \mapsto f_{X,Y}(x,y)$ are Riemann integrable, and that f_X and f_Y are Riemann integrable.

Example :

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2-2xy+y^2}{2(1-\rho^2)}}$$

We have

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

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-x^2/2} \cdot e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}} dy$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}} dy}$$

= 1 , because it is the integral of the $N(\rho x, 1-\rho^2)$ dist

$$= \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}.$$

Conclusion : $X \sim N(0,1)$.

Theorem 3.6 has a more general version.

<u>Theorem 3.6 a</u>: Let $(\underline{X}, \underline{Y})$ be a random vector with $\underline{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with density $f_{\underline{X},\underline{Y}}(\underline{x}, y)$. Then

$$f_{\underline{X}}(\underline{x}) = \int_{\mathbb{R}^q} f_{\underline{X},\underline{Y}}(\underline{x}, y)\, dy$$

$$f_{\underline{Y}}(y) = \int_{\mathbb{R}^p} f_{\underline{X},\underline{Y}}(\underline{x}, y)\, d\underline{x}$$

<u>Proof</u>: Same as before.

<u>Example</u>: If (X, Y, Z) has density $f_{X,Y,Z}(x,y,z)$ then

$$f_{X,Y}(x,y) = \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\, dz$$

and

$$f_X(x) = \int_{\mathbb{R}^2} f_{X,Y,Z}(x,y,z)\, dy\, dz.$$

Independence

In general we say that X, Y are independent if

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

If (X, Y) has density $f_{X,Y}(x,y)$ this means that for $A = [a,b]$ and $B = [c,d]$ we have

$$\underbrace{\int_{[a,b]\times[c,d]} f_{X,Y}(x,y)\, dx\, dy}$$

$$= P(X \in [a,b], Y \in [c,d])$$

$$= \underbrace{\left(\int_a^b f_X(x)\, dx \right)}_{P(a \le X \le b)} \cdot \underbrace{\left(\int_c^d f_Y(y)\, dy \right)}_{P(c \le Y \le d)}$$

Fubini:

$$= \int_{[a,b]\times[c,d]} f_X(x) \cdot f_Y(y)\, dx\, dy$$

We borrow a statement from Analysis 2: if for functions $f(x,y)$ and $g(x,y)$ we have:

(i) For all rectangles $Q = [a,b] \times [c,d]$ the functions are Riemann integrable.

(ii)

$$\int_Q f(x,y)\,dx\,dy = \int_Q g(x,y)\,dx\,dy$$

for all Q

then $f(x,y) = g(x,y)$.

<u>Technical note</u>: in fact f, g can differ but only on a set of measure 0.

If X, Y are independent we have

$$\int_Q f_{X,Y}(x,y)\,dx\,dy = \int_Q f_X(x) f_Y(y)\,dx\,dy$$

Theorem 3.7: Let (X,Y) have density $f_{X,Y}(x,y)$. The random variables X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

Proof: If $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$

then

$$\underbrace{P((X,Y) \in A \times B)}_{\int_{A\times B} f_{X,Y}(x,y)\,dx\,dy} = P(X \in A, Y \in B)$$

$$\overset{\text{Fubini}}{=} \underbrace{\int_A f_X(x)\,dx}_{P(X \in A)} \cdot \underbrace{\int_B f_Y(y)\,dy}_{P(Y \in B)}$$

Independence follows.

If X, Y are independent we proved above that $f_{X,Y}(x,y) =$

$= f_X(x) \cdot f_Y(y)$.

Theorem 3.7 has a more general version.

<u>Theorem 3.7a</u>: Let $\underline{X}, \underline{Y}$ be continuous random vectors with density $f_{\underline{X},\underline{Y}}(\underline{x},\underline{y})$. The vectors $\underline{X}$ and $\underline{Y}$ are independent if and only if $f_{\underline{X},\underline{Y}}(\underline{x},\underline{y}) = f_{\underline{X}}(\underline{x}) \cdot f_{\underline{Y}}(\underline{y})$.

<u>Proof</u>: Same as above.

<u>Theorem 3.8</u>: Let (X,Y) have density $f_{X,Y}(x,y)$. If

$$f_{X,Y}(x,y) = g(x)\,h(y)$$

for nonnegative functions g and h then X, Y are independent.

Proof: By the formulae for marginal density we have

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

$$= \int_{-\infty}^{\infty} g(x)\,h(y)\,dy$$

$$= g(x) \cdot \underbrace{\int_{-\infty}^{\infty} h(y)\,dy}_{= c_1}$$

Similarly

$$f_Y(y) = h(y) \cdot \underbrace{\int_{-\infty}^{\infty} g(x)\,dx}_{c_2}$$

It follows

$$f_{X,Y}(x,y) = \frac{f_X(x)}{c_1} \cdot \frac{f_Y(y)}{c_2}$$

We need to prove that $c_1 \cdot c_2 = 1$.

Integrate both sides over $\mathbb{R}^2$.

We get

$$1 = \int_{\mathbb{R}^2} f_{X,Y}(x,y)\,dx\,dy$$

$$= \frac{1}{c_1 c_2} \int_{\mathbb{R}^2} f_X(x) f_Y(y)\,dx\,dy.$$

Fubini

$$= \frac{1}{c_1 c_2} \int_{-\infty}^{\infty} f_X(x)\,dx \cdot \int_{-\infty}^{\infty} f_Y(y)\,dy$$

$$= \frac{1}{c_1 \cdot c_2} \cdot 1 \cdot 1$$

It follows $c_1 \cdot c_2 = 1$.

<u>Example</u>: Let

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{x^2 - 2\rho x y + y^2}{2(1-\rho^2)}}$$

We computed

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

We see that

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

only when $\rho = 0$.

3.4. Functions of random vectors

Discrete case

In the discrete case we will only look at integrer valued random variables. If X, Y are two such variables then $Z = X + Y$ is an integer valued random variable. We have

$$\{Z = n\} = \underbrace{\bigcup_{k \in \mathbb{Z}} \{X = k, Y = n-k\}}_{\text{disjoint union}}$$

We have

$$\boxed{P(Z = n) = \sum_{k \in \mathbb{Z}} P(X = k, Y = n-k)}$$

Special cases:

(i) if X, Y are non-negative we have

$$P(Z = n) = \sum_{k=0}^{n} P(X = k, Y = n-k)$$

(ii) if X, Y are independent then

$$P(Z = n) = \sum_{k \in \mathbb{Z}} P(X = k) P(Y = n-k)$$

<u>Examples</u>: (i) Let X, Y be independent and $X \sim Po(\mu)$, $Y \sim Po(\lambda)$. Let $Z = X + Y$. By the above formula

$$P(Z=n) = \sum_{k=0}^{n} P(X=k) \cdot P(Y=n-k)$$

$$= \sum_{k=0}^{n} \frac{e^{-\mu}\mu^k}{k!} \cdot \frac{e^{-\lambda} \cdot \lambda^{n-k}}{(n-k)!}$$

$$= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^{n} \underbrace{\frac{n!}{k!\,(n-k)!}}_{=\binom{n}{k}} \mu^k \lambda^{n-k}$$

$$= \frac{e^{-(\lambda+\mu)}}{n!} (\lambda+\mu)^n.$$

<u>Conclusion</u>: $Z = X + Y \sim Po(\lambda+\mu)$.

(ii) Let X, Y be independent and have the Pólya distribution.

This means that

$$P(X=k) = \frac{\beta^a (a)_k}{k! \, (1+\beta)^{a+k}} \qquad k = 0, 1, \dots$$

$$P(Y=\ell) = \frac{\beta^b (b)_\ell}{\ell! \, (1+\beta)^{b+\ell}} \qquad \ell = 0, 1, \dots$$

Here $(a)_0 = 1$ and

$$(a)_k = a(a+1)\dots(a+k-1)$$

is the Pochhammer symbol.

Let $Z = X+Y$. We are looking for the distribution of Z. By the formula we have

$$P(Z=n) = \sum_{k=0}^{n} P(X=k)P(Y=n-k)$$

$$= \frac{\beta^{a+b}}{(1+\beta)^{a+b+n}} \sum_{k=0}^{n} \frac{(a)_k (b)_{n-k}}{k! \, (n-k)!}$$

$$= \frac{\beta^{a+b}}{(1+\beta)^{a+b+n} \cdot n!} \sum_{k=0}^{n} \binom{n}{k} (a)_k (b)_{n-k}.$$

The last formula is similar to the binomial formula.

To prove it we will use a few facts from Analysis:

(i) The gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} \cdot e^{-u} \, du , \quad x > 0$$

Integration by parts gives

$$\Gamma(x+1) = x\,\Gamma(x)$$ and

as a consequence

$$\Gamma(a+n) = (a+n-1)(a+n-2) \cdots a \cdot \Gamma(a)$$

We can write

$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}$$

(ii) The Beta function is defined as

$$B(p,q) = \int_0^1 u^{p-1}(1-u)^{q-1}\,du \quad , \; p,q>0$$

The connection between Γ and B functions is given by Euler:

$$\boxed{B(p,q) = \frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}}$$

We compute

$$\sum_{k=0}^{n} \binom{n}{k} (a)_k (b)_{n-k}$$

$$= \sum_{k=0}^{n} \binom{n}{k} \frac{\Gamma(a+k)}{\Gamma(a)} \cdot \frac{\Gamma(b+n-k)}{\Gamma(b)}$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{n} \binom{n}{k} \frac{\Gamma(a+k)\,\Gamma(b+n-k)}{\Gamma(a+b+n)}$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{n} \binom{n}{k} B(a+k, b+n-k)$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{n} \binom{n}{k} \int_0^1 u^{a+k-1} (1-u)^{b+n-k-1} \, du$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a)\Gamma(b)} \int_0^1 u^{a-1}(1-u)^{b-1} \cdot \underbrace{\sum_{k=0}^{n} \binom{n}{k} u^k (1-u)^{n-k}}_{=1} \, du$$

$$\overset{\text{def.}}{=} \frac{\Gamma(a+b+n)}{\Gamma(a)\Gamma(b)} B(a,b)$$

$$\overset{\text{Euler}}{=} \frac{\Gamma(a+b+n)}{\cancel{\Gamma(a)}\cancel{\Gamma(b)}} \cdot \frac{\cancel{\Gamma(a)}\cancel{\Gamma(b)}}{\Gamma(a+b)}$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a+b)}$$

$$= (a+b)_n$$

Finally we have

$$P(Z=n) = \frac{\beta^{a+b} (a+b)_n}{n! \, (1+\beta)^{a+b+n}} \, , \quad n = 0, 1, \ldots$$

<u>Example</u>: Suppose X, Y are independent and $X \sim \mathrm{Bin}(m, p)$, $Y \sim \mathrm{Bin}(n, p)$. We expect $Z = X + Y \sim \mathrm{Bin}(m+n, p)$. The formal proof:

$$P(Z = \ell) = \sum_{k=0}^{\ell} P(X = k, Y = \ell - k)$$

$$= \sum_{k = \max(0, n-\ell)}^{\min(\ell, m)} \binom{m}{k} p^k q^{m-k} \binom{n}{\ell - k} p^{\ell - k} q^{n - \ell + k}$$

$$= \sum_{k = \max(0, n-\ell)}^{\min(\ell, m)} \binom{m}{k} \binom{n}{\ell - k} \underbrace{p^{\ell} q^{m+n-\ell}}_{\text{does not depend on } k}$$

The sum is computed by the following combinatorial argument:

Suppose we need to choose l elements from the union of sets with m and n elements. This can be done in $\binom{m+n}{l}$ ways. We can count in another way: we choose k elements from the first set and $l-k$ from the other. This is possible for $k \geq \max(0, n-l)$ and $l \leq \min(l, m)$. This splits all the choices in disjoint subsets so

$$\binom{m+n}{l} = \sum_{k=\max(0,n-l)}^{\min(l,m)} \binom{m}{k}\binom{n}{l-k}.$$

Finally

$$P(Z=l) = \binom{m+n}{l} p^l q^{m+n-l}$$

Continuous case

The most important formula is the transformation formula.

Suppose the vector (X,Y) has density $f_{X,Y}(x,y)$. We form a new vector (U,V) by

$$\Phi(X,Y) = (\Phi_1(X,Y), \Phi_2(X,Y)).$$

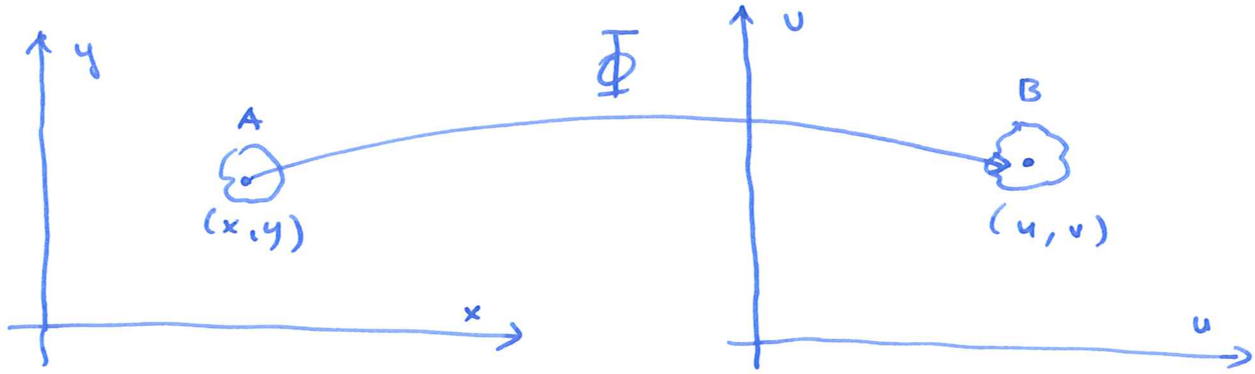
Example:

$$\Phi(x,y) = \left(\frac{x}{x+y}, x+y\right)$$

$$(U,V) = \left(\frac{X}{X+Y}, X+Y\right).$$

Question: What is the density $f_{U,V}(u,v)$ of (U,V)?

Idea :



By definition for „small" A and B

$$P((x,y) \in A) \approx f_{X,Y}(x,y)\,|A|$$

$$P((U,V) \in B) \approx f_{U,V}(u,v)\cdot|B|$$

If Φ is bijective then

$$P((X,Y) \in A) = P((U,V) \in B)$$

if $B = \Phi(A)$. So

$$f_{X,Y}(x,y)\,|A| \approx f_{U,V}(u,v)\cdot|B|$$

or

$$f_{U,V}(u,v) \approx f_{X,Y}(x,y)\cdot\frac{|A|}{|B|}$$

But from Analysis 2 we know

$$\frac{|A|}{|B|} \approx |J_{\Phi^{-1}}(u,v)|.$$

<u>Theorem 3.9</u> (transformation formula). Let (X,Y) be a vector with density $f_{X,Y}(x,y)$. Suppose $P((X,Y) \in \mathcal{U}) = 1$ for an open set $\mathcal{U}$. Let $\Phi : \mathcal{U} \to \mathcal{S}$ be a bijective map which is continuously partially differentiable. Let

$$(U,V) = \Phi(X,Y).$$

The the density $f_{U,V}(u,v)$ is

$$f_{U,V}(u,v) = f_{X,Y}(\Phi^{-1}(u,v)) \cdot |J_{\Phi^{-1}}(u,v)|$$

where $J_{\Phi^{-1}}$ is the Jacobian determinant of Φ^{-1}.

Proof: Let $B \subseteq S$. We compute

$$P((U,V) \in B)$$

$$= P((X,Y) \in \Phi^{-1}(B))$$

$$= \int_{\Phi^{-1}(B)} f_{X,Y}(x,y)\,dx\,dy$$

$$= (*)$$

New variable: $(x,y) = \Phi^{-1}(u,v)$

$$dx\,dy = |J_{\Phi^{-1}}(u,v)|\,du\,dv$$

$$(*) = \int_B f_{X,Y}(\Phi^{-1}(u,v))\,|J_{\Phi^{-1}}(u,v)|\,du\,dv.$$

Comment: We used the formula for a new variable in double integrals.

Example: Let X, Y be independent with $X \sim \Gamma(a, \lambda)$ and $Y \sim \Gamma(b, \lambda)$. This means

$$f_X(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} \cdot e^{-\lambda x}, \quad x > 0$$

$$f_Y(y) = \frac{\lambda^b}{\Gamma(b)} y^{b-1} e^{-\lambda y}, \quad y > 0$$

By independence

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

Let

$$\Phi(x,y) = \left(\frac{x}{x+y}, x+y \right)$$

for $x, y > 0$. We can take

$\mathcal{U} = (0,\infty)^2$ and $\mathcal{S} = (0,1) \times (0,\infty)$.

Φ is bijective and continuously differentiable. To find Φ^{-1} we need to solve equations

$$\frac{x}{x+y} = u \,, \qquad x + y = v .$$

We get

$$\begin{aligned} x &= u \cdot v \\ y &= v - x = v - u \cdot v \\ &= v(1-u) \end{aligned}$$

This means

$$\Phi^{-1}(u,v) = (uv, v(1-u)).$$

We compute

$$J\Phi^{-1}(u,v) = \det\begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix}$$

$$= v$$

The density $f_{U,V}(u,v)$ is given by

$$f_{U,V}(u,v) = f_{X,Y}(uv, v(1-u)) \cdot |J\Phi^{-1}(u,v)| =$$

$$= f_X(uv) \cdot f_Y(v(1-u)) \cdot v$$

$$= \frac{\lambda^a}{\Gamma(a)} (uv)^{a-1} \cdot e^{-\lambda uv}$$

$$\cdot \frac{\lambda^b}{\Gamma(b)} [v(1-u)]^{b-1} \cdot e^{-\lambda v(1-u)}$$

$$\cdot v$$

$$= \frac{\lambda^{a+b}}{\Gamma(a)\,\Gamma(b)}\, u^{a-1} (1-u)^{b-1}$$

$$\cdot\, v^{a+b-1} \cdot e^{-\lambda v}$$

for $(u,v) \in (0,1) \times (0,\infty)$.

We note:

(i) U, V are independent

(ii) $f_U(u) = \text{const.}\; u^{a-1} (1-u)^{b-1}$

$f_V(v) = \text{const.}\; v^{a+b-1} \cdot e^{-\lambda v}$.

It follows $U = \frac{X}{X+Y} \sim \text{Beta}(a,b)$

and $V = X+Y \sim \Gamma(a+b, \lambda)$.

<u>Example</u> : Suppose (X,Y) has density $f_{X,Y}(x,y)$. Let

$$\Phi(x,y) = (x, x+y) = (x,z)$$

What is the density $f_{X,Z}(x,z)$?

By the transformation formula

$$f_{X,Z}(x,z) = f_{X,Y}(x, z-x) \cdot |J_{\Phi^{-1}}(x,z)|.$$

But $\Phi^{-1}(x,z) = (x, z-x) \Rightarrow$

$$J_{\Phi^{-1}}(x,z) = 1.$$

We have

$$\boxed{f_{X,Z}(x,z) = f_{X,Y}(x, z-x)}$$

The density of Z is the marginal density of $f_{X,Z}(x,z)$.

We have

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx}$$

If X, Y are independent we get

$$\boxed{f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\,dx}$$

<u>Comment</u>: The above formula is known as <u>convolution</u> in Analysis.

Example: Let X, Y be independent with $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\nu, \tau^2)$. What is the density of $Z = X + Y$. Assume first $\mu = \nu = 0$ and $\sigma^2 + \tau^2 = 1$.

In this case

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx$$

$$= \frac{1}{2\pi \cdot \sigma \tau} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \cdot e^{-\frac{(z-x)^2}{2\tau^2}}\, dx$$

$$= \frac{1}{2\pi \sigma \cdot \tau} \int_{-\infty}^{\infty} e^{-x^2\left[\frac{1}{2\sigma^2} + \frac{1}{2\tau^2}\right]} \cdot e^{\frac{xz}{\tau^2}} \cdot e^{-\frac{z^2}{2\tau^2}}\, dx$$

$$= \frac{1}{2\pi \sigma \cdot \tau} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2\tau^2} + \frac{xz}{\tau^2}} \cdot e^{-\frac{z^2}{2\tau^2}}\, dx$$

$$= \frac{1}{2\pi \sigma \cdot \tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2\tau^2}(x - z\sigma^2)^2} \cdot e^{\frac{z^2\sigma^2}{2\tau^2}} \cdot e^{-\frac{z^2}{2\tau^2}} \, dx$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}\,\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{(x - z\sigma^2)^2}{2\sigma^2 \cdot \tau^2}} \, dx}_{=1} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}\underbrace{\left[\frac{1}{\tau^2} - \frac{\sigma^2}{\tau^2}\right]}_{=1}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

<u>Conclusion</u>: $Z \sim N(0,1)$.

We know: if $X \sim N(\mu, \sigma)$ then

$aX + b \sim N(a\mu + b, a^2\sigma^2)$.

In general: $X \sim N(\mu, \sigma^2)$, $Y \sim N(\nu, \tau^2)$

$$X + Y = \sqrt{\sigma^2 + \tau^2} \times \left(\underbrace{\frac{X - \mu}{\sqrt{\sigma^2 + \tau^2}}}_{X'} + \underbrace{\frac{Y - \nu}{\sqrt{\sigma^2 + \tau^2}}}_{Y'} \right) + \mu + \nu$$

We have $X' \sim N(0, \frac{\sigma^2}{\sigma^2+\tau^2})$ in

$Y' \sim N(0, \frac{\tau^2}{\sigma^2+\tau^2})$. The

expression $X' + Y' \sim N(0,1)$.

It follows that

$$\boxed{X + Y \sim N(\mu+\nu, \sigma^2+\tau^2)}$$

Example : Let X, Y be

independent standar normal.

Let $Z = \frac{Y}{X}$. Density of Z ?

Define

$$\Phi(x,y) = (x, y/x)$$

$$\Phi^{-1}(x,z) = (x, xz) \implies$$

$$J\Phi^{-1}(x,z) = \det\begin{pmatrix} 1 & 0 \\ z & x \end{pmatrix} = x$$

The density of (X, Z) is

$$f_{X,Z}(x,z) = f_X(x)\, f_Y(xz) \cdot |x|$$

$$= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(xz)^2}{2}} \cdot |x|$$

We get the density of Z as the marginal density

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{2\pi} \cdot e^{-\frac{x^2(1+z^2)}{2}} \cdot |x|\, dx$$

$$= \frac{1}{\pi} \int_0^{\infty} e^{-\frac{x^2(1+z^2)}{2}}\, x \cdot dx$$

$$= \frac{1}{\pi(1+z^2)} \left(- e^{-\frac{x^2(1+z^2)}{2}} \right) \Bigg|_0^{\infty}$$

$$= \frac{1}{\pi(1+z^2)}$$

<u>Example</u>: Let X, Y be independent with $X \sim \Gamma(a, \lambda)$ and $Y \sim \Gamma(b, \lambda)$. Let $Z = X + Y$. We established that $Z \sim \Gamma(a+b, \lambda)$ but will ~~do~~ it again using convolution.

$$f_Z(z) = \int_0^z f_X(x)\, f_Y(z-x)\, dx$$

$$= \int_0^z \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \cdot \frac{\lambda^b}{\Gamma(b)} (z-x)^{b-1} e^{-\lambda(z-x)}\, dx$$

$$= \frac{\lambda^{a+b}}{\Gamma(a)\,\Gamma(b)} \cdot e^{-\lambda z} \cdot \int_0^z x^{a-1} (z-x)^{b-1}\, dx$$

New variable: $x = z \cdot u$

$dx = z \cdot du$

$$= \frac{\lambda^{a+b}}{\Gamma(a)\,\Gamma(b)} \cdot e^{-z} \cdot \int_0^1 u^{a-1} \cdot (1-u)^{b-1} \cdot z^{a+b-1}\, du$$

$$= \frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)} e^{-z} \cdot z^{a+b-1} \quad B(a,b)$$

The result is a density which means that it integrates to 1.

But we know that

$$\frac{\lambda^{a+b}}{\Gamma(a+b)} \int_0^\infty z^{a+b-1} \cdot e^{-\lambda z} dz = 1.$$

This means that

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot B(a,b) = 1$$

which is Euler identity!

We have used probability to derive Euler's identity.

Theorem 3.6 has a more general version.

<u>Theorem 3.6 a</u>: Let $\underline{X}$ be a random vector with density $f_{\underline{X}}(\underline{x})$. Assume $P(\underline{X} \in \mathcal{U}) = 1$ for some open set $\mathcal{U} \subseteq \mathbb{R}^n$ and let $\Phi : \mathcal{U} \to \mathcal{S} \subseteq \mathbb{R}^n$ be a bijective map between $\mathcal{U}$ and $\mathcal{S}$ such that Φ and Φ^{-1} are continuously partially differentiable. Let $\underline{Y} = \Phi(\underline{X})$. Then $\underline{Y}$ has the density

$$\boxed{f_{\underline{Y}}(y) = f_{\underline{X}}(\Phi^{-1}(y)) \cdot |J_{\Phi^{-1}}(y)|.}$$

<u>Proof</u>: Same as before.

<u>Example</u>: Let $\underline{X} = (X_1, X_2, \dots, X_r)$ such that $X_1, X_2, \dots, X_r$ are independent and $X_k \sim N(0,1)$ for all $k = 1, 2, \dots, r$; Let $\underline{A}$ be an invertible matrix. Define

$$\underline{\Phi}(\underline{x}) = \underline{A}\,\underline{x} + \underline{\mu} \quad \text{for} \quad \underline{\mu} \in \mathbb{R}^n.$$

We have $\underline{\Phi}^{-1}(\underline{y}) = \underline{A}^{-1}(\underline{x} - \underline{\mu})$

and $J\underline{\Phi}^{-1}(\underline{y}) = \det(\underline{A}^{-1}) = \dfrac{1}{\det(A)}$.

The transformation formula gives for $\underline{Y} = \underline{A}\,\underline{X} + \underline{\mu}$

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{\Phi}^{-1}(\underline{y})) \cdot |J\underline{\Phi}^{-1}(\underline{y})|.$$

We have

$$f_{\underline{X}}(\underline{x}) = \prod_{k=1}^{r} f_{X_k}(x_k)$$

$$= \frac{1}{(2\pi)^{n/2}} \cdot e^{-\frac{1}{2}\sum_{k=1}^{n} x_k^2}$$

$$= \frac{1}{(2\pi)^{n/2}} \cdot e^{-\frac{1}{2}\underline{x}^T\cdot\underline{x}}$$

It follows

$$f_{\underline{Y}}(y) = \frac{1}{(2\pi)^{n/2}\,|\det(\underline{A})|}$$

$$\times \quad e^{-1/2\,[\underline{A}^{-1}(\underline{x}-\underline{\mu})]^T[\underline{A}^{-1}(\underline{x}-\underline{\mu})]}$$

$$= \frac{1}{(2\pi)^{n/2}\,|\det(\underline{A})|}$$

$$\times \quad e^{-1/2\,(\underline{x}-\underline{\mu})^T(\underline{A}^{-1})^T\underline{A}^{-1}(\underline{x}-\underline{\mu})}$$

Denote $\underline{\Sigma} = \underline{A}\cdot\underline{A}^T$. We have

$$\underline{\Sigma}^{-1} = (\underline{A}^T)^{-1}\cdot A^{-1} = (\underline{A}^{-1})^T(\underline{A}^{-1})$$

We have

$$\frac{1}{|\det(\underline{A})|} = \frac{1}{\sqrt{\det(\underline{\Sigma})}}$$

and

$$f_{\underline{Y}}(y) = \frac{1}{(2\pi)^{r/2}\sqrt{\det(\underline{\Sigma})}}$$

$$\times \quad e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x}-\underline{\mu})}.$$

<u>Comment</u>: The above density is called the multivariate normal density with parameters $\underline{\mu} \in \mathbb{R}^r$ and $\underline{\Sigma}$ $(r \times r)$.

Example: Let $\underline{X}$ have density

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \underline{\Sigma}}}$$

$$\times \; e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x}-\underline{\mu})}$$

If $\underline{X} = (X_1, X_2, \dots, X_n)$ what is the distribution of $\underline{X}^{(1)} = (X_1, X_2, \dots, X_p)$, $p < n$? Denote $\underline{X} = (\underline{X}^{(1)}, \underline{X}^{(2)})$,

$$\underline{\mu} = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \quad \text{and} \quad \underline{\Sigma} = \begin{pmatrix} \underline{\Sigma}_{11} & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} \end{pmatrix}.$$

$\underline{\mu}^{(1)}$ is a p-dimensional vector, $\underline{\Sigma}_{11}$ $(p \times p)$, $\underline{\Sigma}_{21}$ $(p \times q)$, $\underline{\Sigma}_{21}$ $(q \times p)$, $\underline{\Sigma}_{22}$ $(q \times q)$. Define $\underline{\Phi} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\underline{\Phi}(\underline{x}) = \begin{pmatrix} x^{(1)} \\ \underline{x}^{(2)} - \underline{\Sigma}_{21}\underline{\Sigma}_{11}^{-1}\underline{x}^{(1)} \end{pmatrix}.$$

$\underline{\Phi}(\underline{x})$ is a linear map

$$\underline{\Phi}(\underline{x}) = \underbrace{\begin{pmatrix} \underline{I}_r & 0 \\ -\underline{\Sigma}_{21}\underline{\Sigma}_{11}^{-1} & \underline{I}_\ell \end{pmatrix}}_{\underline{A}} \underline{x} .$$

Since the matrix is lower triangular we have

$$D\underline{\Phi} = A \Rightarrow J_{\underline{\Phi}}(\underline{x}) = 1$$

$$\Rightarrow J_{\underline{\Phi}^{-1}}(\underline{y}) = 1 .$$

We have

$$\underline{\Phi}^{-1}(\underline{y}) = \begin{pmatrix} \underline{y}^{(1)} \\ \underline{y}^{(2)} + \underline{\Sigma}_{21}\underline{\Sigma}_{11}^{-1}\underline{y}^{(1)} \end{pmatrix}$$

It follows that

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}(\underline{\Phi}^{-1}(\underline{y})) \cdot 1$$

We need some linear algebra. Suppose $\underline{A}, \underline{B}$ are invertible matrices. Write

$$\underline{A} = \begin{pmatrix} \underline{A}_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{pmatrix} \quad \text{and} \quad \underline{B} = \begin{pmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{pmatrix}$$

where $\underline{A}_{ij}$ and $\underline{B}_{ij}$ are of the same dimension. If $\underline{A} \cdot \underline{B} = \underline{I}$ we have

$$\underline{A}_{11} \underline{B}_{11} + \underline{A}_{12} \underline{B}_{21} = \underline{I}$$
$$\underline{A}_{11} \underline{B}_{12} + \underline{A}_{12} \cdot \underline{B}_{22} = 0$$

For simplicity we assume $\underline{\mu} = 0$. We need to compute

$$[\underline{\Phi}^{-1}(y)]^T \cdot \underline{\Sigma}^{-1} [\underline{\Phi}^{-1}(y)].$$

In matrix form this means

$$\mathbf{y}^T \begin{pmatrix} \mathbf{I}_p & \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{21} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \mathbf{\Sigma}^{-1} \begin{pmatrix} \mathbf{I}_p & 0 \\ \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1} & \mathbf{I}_q \end{pmatrix}$$

Denote $\mathbf{A} = \mathbf{\Sigma}^{-1}$. From

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} = \mathbf{I}_n$$

we have

$$\mathbf{A}_{11}\mathbf{\Sigma}_{11} + \mathbf{A}_{21}\mathbf{\Sigma}_{21} = \mathbf{I}_p$$

$$\mathbf{A}_{11}\mathbf{\Sigma}_{12} + \mathbf{A}_{12}\mathbf{\Sigma}_{22} = \mathbf{0}$$

$$\mathbf{A}_{21}\mathbf{\Sigma}_{11} + \mathbf{A}_{22}\mathbf{\Sigma}_{21} = \mathbf{0}$$

We compute

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & 0 \\ \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1} & \mathbf{I}_q \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}, & \mathbf{A}_{12} \\ \mathbf{A}_{21} + \mathbf{A}_{22}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}, & \mathbf{A}_{22} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{\Sigma}_{11}^{-1} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}$$

Continue to get

$$\begin{pmatrix} \underline{I}_p & \underline{\Sigma}_{11}^{-1}\underline{\Sigma}_{12} \\ 0 & \underline{I}_q \end{pmatrix} \begin{pmatrix} \underline{\Sigma}_{11}^{-1} & \underline{A}_{12} \\ \underline{0} & \underline{A}_{22} \end{pmatrix}$$

$$= \begin{pmatrix} \underline{\Sigma}_{11}^{-1} , & \underline{A}_{12} + \underline{\Sigma}_{11}^{-1}\underline{\Sigma}_{12}\underline{A}_{22} \\ -''- , & \underline{A}_{22} \end{pmatrix}$$

But

$$\begin{pmatrix} \underline{\Sigma}_{11} & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \underline{A}_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{pmatrix} = \underline{I}_n$$

gives

$$\underline{\Sigma}_{11}\underline{A}_{12} + \underline{\Sigma}_{12}\underline{\Sigma}_{22} = 0 \text{ , so}$$

$$\underline{\Sigma}_{11}\left(\underline{A}_{12} + \underline{\Sigma}_{11}^{-1}\underline{\Sigma}_{12}\underline{A}_{22}\right) = 0$$

The linear equations give

$$\underline{A}_{22} = \left(\underline{\Sigma}_{22} - \underline{\Sigma}_{12}\underline{\Sigma}_{11}^{-1}\underline{\Sigma}_{21}\right)$$

(see Appendix)

So we have

$$[\Phi^{-1}(y)]^T \Sigma^{-1} [\Phi^{-1}(y)]$$

$$= y^T \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix} y$$

$$= [y^{(1)}]^T \Sigma_{11}^{-1} y^{(1)}$$

$$+ [y^{(2)}]^T (\Sigma_{22} - \Sigma_{21}\Sigma_{11}\Sigma_{12})^{-1} y^{(2)}$$

Comment: in general replace y by $y - \mu$. So we have

$$f_{Y}(y) = f(y^{(1)}) \cdot g(y^{(2)}).$$

This means that

$$Y^{(1)} = (X_1, \dots, X_r)$$

$$Y^{(2)} = X^{(2)} - \Sigma_{21} \Sigma_{11}^{-1} X^{(1)}$$

are independent vectors.

Appendix: if we have

$$\begin{pmatrix} \underline{\Sigma}_{11} & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \underline{A}_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{pmatrix} = \underline{I}_n$$

then

$$\underline{\Sigma}_{11}\,\underline{A}_{11} + \underline{\Sigma}_{12}\,\underline{A}_{21} = \underline{I}_p$$

$$\underline{\Sigma}_{11}\,\underline{A}_{12} + \underline{\Sigma}_{12}\,\underline{A}_{22} = \underline{0}$$

$$\underline{\Sigma}_{21}\,\underline{A}_{11} + \underline{\Sigma}_{22}\,\underline{A}_{21} = \underline{0}$$

$$\underline{\Sigma}_{21}\,\underline{A}_{12} + \underline{\Sigma}_{22}\,\underline{A}_{22} = \underline{I}_q$$

We have a system of 4 linear equations with 4 unknowns.

Multiply the second equation with $\underline{\Sigma}_{11}^{-1}$ from the left to get

$$\underline{A}_{12} + \underline{\Sigma}_{11}^{-1}\,\underline{\Sigma}_{12}\,\underline{A}_{22} = \underline{0}$$

Insert this into the last equation to get

$$- \underline{\Sigma}_{21} \underline{\Sigma}_{11}^{-1} \underline{\Sigma}_{12} \underline{A}_{22} + \underline{\Sigma}_{22} \underline{A}_{22} = \underline{I}_{\ell}$$

we have

$$\underline{A}_{22} = (\underline{\Sigma}_{22} - \underline{\Sigma}_{12} \underline{\Sigma}_{11}^{-1} \underline{\Sigma}_{21})^{-1}$$

This result is known as the inversion lemma.

<u>Remark</u>: Invertibility follows from the fact that the product is $\underline{I}_{\ell}$.

3.5. Conditional distributions

In elementary probability we have that $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

If X is a discrete random variable with values $\{x_1, x_2, \dots\}$ the distribution is given by the probabilities $P(X = x_k)$.

If we have additional information in the sense that the event B has happened our opinion about the probabilities of event $\{X = x_k\}$ change to the conditional probabilities

$$P(\{X = x_k\} | B) = \frac{P(\{X = x_k\} \cap B)}{P(B)}.$$

We can verify easily that

$$\sum_{x_k} P(X = x_k \mid B) = 1 .$$

This observation motivates the definition of conditional probabilities and distributions.

Definition : Let X be a discrete random variable with values in $\{x_1, x_2, \dots\}$. The conditional distribution of X given B with $P(B) > 0$ is given by

$$P(X = x_k \mid B) = \frac{P(\{X = x_k\} \cap B)}{P(B)}.$$

Comment : In most cases the event B is of the form $B = \{Y = y_\ell\}$ for some random variable Y.

<u>Example</u>: Let X, Y be independent with $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$. Let $Z = X + Y$. We know that $Z \sim \text{Bin}(m+n, p)$.

The conditional distribution of X given $\{Z = r\}$ is given by

$$P(X = k \mid Z = r) = \frac{P(X = k, Z = r)}{P(Z = r)}$$

$$= \frac{P(X = k, Y = r-k)}{P(Z = r)}$$

$$= \frac{P(X = k)\, P(Y = r-k)}{P(Z = r)} \qquad \underline{\text{indep}}$$

$$= \frac{\binom{m}{k} p^k q^{m-k} \cdot \binom{n}{r-k} p^{r-k} \cdot q^{n-r+k}}{\binom{m+n}{r} p^r q^{m+n-r}}$$

$$= \frac{\binom{m}{k}\binom{n}{r-k}}{\binom{m+n}{r}}$$

for $k \leq \min(m, r)$ and

$k \geq \max(0, r-n)$.

We recognize the hypergeometric distribution. We write

$$X \mid_{Z=r} \sim \text{HiperGeom}(r, m, m+n).$$

<u>Definition</u>: Let $\underline{X}$ be a discrete random vector with values $\{\underline{x}_1, \underline{x}_2, \ldots\}$. Let B be an event. The conditional distribution of $\underline{X}$ given B with $P(B) > 0$ is given by conditional probabilities

$$P(\underline{X} = \underline{x}_k \mid B) = \frac{P(\{X = x_k\} \cap B)}{P(B)}.$$

As before in most cases B is of the form $B = \{\underline{Y} = y\}$ for some random vector $\underline{Y}$.

Example: Let $\underline{X} = (X_1, \dots, X_r)$ be multinomial with parameters n and $\underline{p} = (p_1, p_2, \dots, p_r)$. Let $s < r$. What is the conditional distribution of $(X_1, X_2, \dots, X_s)$ given $Y = X_1 + X_2 + \dots + X_s = m$.

Denote $Z = X_{s+1} + \dots + X_r$. We know that $Y \sim \text{Bin}(n, p_1 + \dots + p_s)$. We compute for $k_1 + \dots + k_s = m$

$$P(X_1 = k_1, \dots, X_s = k_s \mid Y = m)$$

$$= \frac{P(X_1 = k_1, \dots, X_s = k_s, Z = n-m)}{P(Z = n-m)}$$

$$= \frac{n!}{k_1! \cdots k_s! \, (n-m)!} \times p_1^{k_1} \cdots p_s^{k_s} (1 - p_1 - \cdots - p_s)^{n-m} \Big/$$

$$\Big/ \binom{n}{m} (p_1 + \cdots + p_s)^m (1 - p_1 - \cdots - p_s)^{n-m}$$

$$= \frac{m!}{k_1! \cdots k_s!} \times \frac{p_1^{k_1} \cdots p_s^{k_s}}{(p_1 + \cdots + p_s)^m}$$

$$= (*)$$

We denote: $\hat{p}_k = \dfrac{p_k}{(p_1 + \cdots + p_s)}$

for $k = 1, 2, \ldots, s$. We have:

$$* = \frac{m!}{k_1! \cdots k_s!} \hat{p}_1^{k_1} \cdots \hat{p}_s^{k_s}.$$

Conclusion: $(X_1, X_2, \ldots, X_s)$

has the multi~~nomial~~ distribution

with parameters $n-m$ and $\tilde{p} = (\tilde{p}_1, \dots, \tilde{p}_s)$. We write

$$\underline{X}' = (X_1, \dots, X_s) \quad \text{and}$$

$$\underline{X}' \,|_{X_1 + \dots + X_s = m} \sim \text{Multinom}(m, \tilde{p}).$$

For the continuous case the intuitive idea is that we will define conditional densities. If (X, Y) has density $f_{X,Y}(x,y)$ then the conditional density of Y given $\{X = x\}$ should be proportional to the function

$$y \longmapsto f_{X,Y}(x,y)$$

$$\boxed{f_{Y|X=x}(x,y) = \frac{f_{X,Y}(x,y)}{f_X(x)}}$$

<u>Example</u>: Let

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \times$$

$$\times\ e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}$$

for $|\rho|<1$. We know that $X \sim N(0,1)$ i.e.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

We write

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-x^2/2} \times$$

$$\times\ e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}$$

It follows that

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}$$

We notice

$$Y \mid X=x \sim N(\rho x, 1-\rho^2)$$

The definition has a vector version.

<u>Definition</u>: Let $(\underline{X}, \underline{Y})$ have density $f_{\underline{X},\underline{Y}}(\underline{x},\underline{y})$. Assume $f_{\underline{X}}(\underline{x}) > 0$. The conditional density of $\underline{Y}$ given $\{\underline{X}=\underline{x}\}$ is given by

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

Example : Let $\underline{X} = (\underline{X}^{(1)}, \underline{X}^{(2)}) \sim N(\underline{\mu}, \underline{\Sigma})$.

What is $f_{\underline{X}^{(2)} | \underline{X}^{(1)} = \underline{x}^{(1)}} (\underline{x}^{(2)})$?

Direct calculation is difficult but we found out that $\underline{X}^{(1)}$ and

$\underline{X}^{(2)} - \underline{\Sigma}_{21} \underline{\Sigma}_{11}^{-1} \underline{X}^{(1)}$ are independent

vectors. If we write $\underline{Y} = \underline{X}^{(2)} - \underline{\Sigma}_{21} \underline{\Sigma}_{11} \underline{X}^{(1)}$

we know that

$$f_{\underline{Y}}(y) = \frac{1}{(2\pi)^{q/2} \sqrt{\det (\underline{\Sigma}_{22} - \underline{\Sigma}_{21} \underline{\Sigma}_{11}^{-1} \underline{\Sigma}_{12})}}$$

$$\times\ e^{-\frac{1}{2} (y - \mu^{(2)} + \underline{\Sigma}_{21} \underline{\Sigma}_{11}^{-1} \underline{\mu}^{(1)})^T (\underline{\Sigma}_{22} - \underline{\Sigma}_{21} \underline{\Sigma}_{11}^{-1} \underline{\Sigma}_{12})^{-1} (- '' -)}.$$

But

$$\begin{pmatrix} \underline{X}^{(1)} \\ \underline{X}^{(2)} \end{pmatrix} = \begin{pmatrix} \underline{I}_p & 0 \\ \underline{\Sigma}_{21} \underline{\Sigma}_{11} & \underline{I}_q \end{pmatrix} \begin{pmatrix} \underline{X}^{(1)} \\ \underline{Y} \end{pmatrix}$$

The Jacobian of this is 1 so we can write

$$f_{\underline{X}}(\underline{x}) = f_{\underline{X}^{(1)}}(\underline{x}^{(1)})$$

$$\times \; f_{\underline{Y}}\left(\underline{x}^{(2)} - \underline{\Sigma}_{21}\underline{\Sigma}_{11}^{-1}\underline{x}^{(1)}\right)$$

Now it is easy to divide by $f_{\underline{X}^{(1)}}(\underline{x}^{(1)})$. We get

$$f_{\underline{X}^{(2)} \mid \underline{X}^{(1)} = \underline{x}^{(1)}}(\underline{x}^{(2)})$$

$$= f_{\underline{Y}}\left(\underline{x}^{(2)} - \underline{\Sigma}_{12}\underline{\Sigma}_{11}^{-1}\underline{x}^{(1)}\right).$$

Using the form of $f_{\underline{Y}}$ we find:

$$\underline{X}^{(2)} \mid \underline{X}^{(1)} = \underline{x}^{(1)} \sim N\left(\underline{\mu}^{(2)} + \underline{\Sigma}_{21}\underline{\Sigma}_{11}^{-1}(\underline{x}^{(1)} - \underline{\mu}^{(1)}), \; \underline{\Sigma}_{22} - \underline{\Sigma}_{21}\underline{\Sigma}_{11}^{-1}\underline{\Sigma}_{12}\right)$$

4. Expectation and variance

4.1. Expectation in general

For a discrete random variable we defined

$$\boxed{\begin{aligned} E(X) &= \sum_{x_k} x_k P(X = x_k) \\ E[f(X)] &= \sum_{x_k} f(x_k) P(X = x_k) \end{aligned}}$$

For a discrete random vector we have

$$\boxed{E[f(\underline{X})] = \sum_{\underline{x}_k} f(\underline{x}_k) P(\underline{X} = \underline{x}_k)}$$

We need to extend the notion of expectation to continuous random variables and vectors.

Definitions:

(i) Let X have density $f_X(\underline{x})$.

We define

$$\boxed{\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx \\ E[f(X)] &= \int_{-\infty}^{\infty} f(x)\, f_X(x)\, dx \end{aligned}}$$

Technical note: we say that $E(X)$ exists if the integral $\int_{-\infty}^{\infty} |x|\, f_X(x)\, dx$ converges and similarly for $E[f(X)]$.

(ii) Let the random vector $\underline{X}$ have density $f_{\underline{X}}(\underline{x})$. We define

$$\boxed{E[f(\underline{X})] = \int_{\mathbb{R}^n} f(\underline{x})\, f_{\underline{X}}(\underline{x})\, d\underline{x}}$$

Examples:

(i) $X \sim N(\mu, \sigma^2)$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$$

$= (*)$ New variable: $\frac{x-\mu}{\sigma} = u$

$$= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (\sigma u + \mu)\, e^{-u^2/2}\, du$$

$$= \mu \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du$$

$$= \mu$$

because $\int_{-\infty}^{\infty} u \cdot e^{-u^2/2}\, du = 0$

(odd function).

We continue

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,dx$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \int_{-\infty}^{\infty} x^2 \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma u+\mu)^2\, e^{-u^2/2}\,du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma^2 u^2+\mu^2) \cdot e^{-u^2/2}\,du$$

(the middle term = 0)

$$= \sigma^2 \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 \cdot e^{-u^2/2}\,du + \mu^2 = (*)$$

We integrate by parts

$$\int_{-\infty}^{\infty} u^2 \cdot e^{-u^2/2}\,du =$$

$$= \int_{-\infty}^{\infty} u \cdot u \cdot e^{-u^2/2}\,du$$

$$= -u \cdot e^{-u^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-u^2/2}\,du$$

$$= \sqrt{2\pi}$$

$$(*) = \sigma^2 + \mu^2$$

Definition: Let X be a random variable. Let $\mu = E(X)$.

We call $E(X^m)$ the m-th **moment** of X and $E[(X-\mu)^m]$ the m-th **central moment** of X.

Example: Let $X \sim \Gamma(a, \lambda)$.

We compute the m-th moment of X as

$$E(X^m) = \int_0^\infty x^m \cdot f_X(x)\, dx$$

$$= \int_0^\infty \frac{\lambda^a}{\Gamma(a)} \cdot x^m \cdot x^{a-1} \cdot e^{-\lambda x}\, dx$$

$$= \frac{\lambda^a}{\Gamma(a)} \int_0^\infty x^{m+a-1} e^{-\lambda x}\, dx$$

$$= \frac{\lambda^a}{\Gamma(a)} \cdot \frac{\Gamma(m+a)}{\lambda^{m+a}}$$

$$= \frac{\Gamma(m+a)}{\Gamma(a)\ \lambda^{m}}$$

$$= \frac{(a)_m}{\lambda^m}$$

where $(a)_m = a(a+1)\cdots(a+m-1)$.

Example: Let (X,Y) have the density $f_{X,Y}(x,y) =$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}}\ e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}$$

with $|\rho|<1$. We have

$$E(XY) = \int_{\mathbb{R}^2} xy \cdot f_{X,Y}(x,y)\,dx\,dy$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{\mathbb{R}^2} xy \cdot e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}\,dx\,dy$$

Fubini

$$= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x \cdot e^{-x^2/2}\,dx \ \times$$

$$\times \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} y \cdot e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}\,dy$$

The last integral is ρx.
We computed it in the first example. We have

$$E(XY) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \rho x^2 \cdot e^{-x^2/2} \, dx$$

$$= \rho \cdot E(X^2)$$

$$= \rho$$

If $X \sim N(0,1)$ then $E(X^2) = \sigma^2 + \mu^2 = \sigma^2$.

Theorem 4.1: Let (X,Y) have density $f_{X,Y}(x,y)$. We have

$$E[\alpha X + \beta Y] = \alpha E(X) + \beta E(Y)$$

Proof: We have

$E[\alpha X + \beta Y]$

$$= \int_{\mathbb{R}^2} (\alpha x + \beta y)\, f_{X,Y}(x,y)\, dx\, dy$$

$$= \alpha \cdot \int_{\mathbb{R}^2} x\, f_{X,Y}(x,y)\, dx\, dy$$

$$+ \beta \int_{\mathbb{R}^2} y\, f_{X,Y}(x,y)\, dx\, dy$$

$$= \alpha \cdot E(X) + \beta E(Y)$$

Remarks:

(i) The same proof works with $\alpha f(X,Y) + \beta g(X,Y)$ i.e

$$E[\alpha f(X,Y) + \beta g(X,Y)]$$

$$= \alpha\, E[f(X,Y)] + \beta E[g(X,Y)].$$

(ii) The theorem has a vector version:

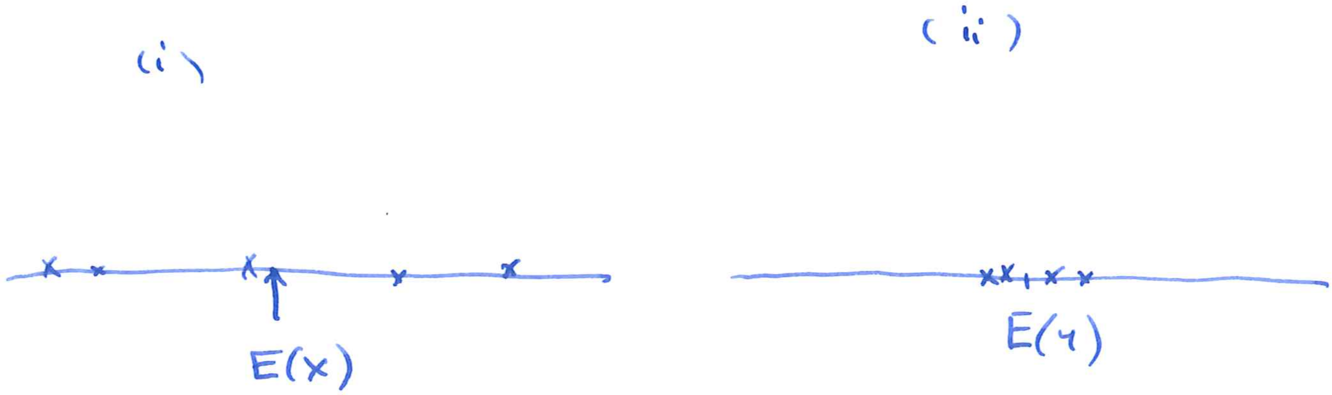
$$\boxed{E\left[\sum_{k=1}^{r} \alpha_k X_k\right] = \sum_{k=1}^{r} \alpha_k E(X_k).}$$

The expectation is always linear.

4.2 Variance and covariance

We motivated the expectation as a „long term average".
We would like to devise a measure of dispersion of a random variable X.

Figure:



In the figure we have „repeated" values of X and Y. The values of X are more „dispersed". Why do we say this? On average the values of X are further away from $E(x)$. Repetitions to the right and to the left contribute an equal amount to dispersion so we take absolute distances. However, Gauss chose the square. His choice was motivated by mathematical considerations. Denote $v_1, v_2, \ldots, v_n$ the repetitions of X.

The dispersion according to Gauss is

$$\frac{(v_1 - E(x))^2 + \cdots + (v_n - E(x_n))^2}{n}.$$

If we take $f(x) = (x - E(x))^2$

we see that the above average

$$\approx E[(x - E(x))^2].$$

<u>Definition</u>: The variance of the random variable X is given by

$$\boxed{\operatorname{var}(x) = E[(X - E(x))^2]}$$

We compute

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

$$= E[X^2 - 2E(x) \cdot X + E(x)^2]$$

lin.

$$= E(x^2) - 2E(x)\cdot E(x) + E(x)^2$$

$$= E(x^2) - [E(x)]^2$$

Alternative form:

$$\boxed{var(x) = E(x^2) - [E(x)]^2}$$

Examples:

(i) $X \sim Bin(n, p)$

We know

$$E(x^2) = npq + n^2p^2 \text{ and}$$

$$E(x) = n\cdot p$$

$$var(x) = E(x^2) - E(x)^2$$

$$= npq$$

(ii) $X \sim NegBin(m, p)$

We know

$$E(X) = \frac{m}{p}$$

$$E(X^2) = \frac{m \cdot q}{p^2} + \frac{m^2}{p^2}$$

$$var(X) = \frac{m \cdot q}{p^2}$$

(iii) $X \sim N(\mu, \sigma^2)$

We know

$$E(X) = \mu$$

$$E(X^2) = \sigma^2 + \mu^2$$

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

$$= \sigma^2$$

<u>Comment</u>: If $X \sim N(\mu, \sigma^2)$ the two parameters have a nice

interpretation. They are the expectation and the variance.

What about the variance of sums? We compute

$$\mathrm{var}(X+Y) = E[(X+Y)^2] - [E(X+Y)]^2$$

$$= E(X^2+2XY+Y^2) - (E(X)+E(Y))^2$$

lin.

$$= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2$$

$$= \mathrm{var}(X) + \mathrm{var}(Y) + 2[E(XY) - E(X)E(Y)]$$

There is no reason for the term in square brackets to be 0.

Definition: Let X, Y be random variables. The quantity

$$E(XY) - E(X) \cdot E(Y)$$

is called the covariance of X and Y and denoted by $\operatorname{cov}(X, Y)$.

Remark: An application of linearity gives that

$$\operatorname{cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

Theorem 4.2: Let $X_1 \dots X_r$ and $Y_1, Y_2, \dots, Y_s$ be random variables. We have

$$\operatorname{cov}\left(\sum_{k=1}^{r} \alpha_k X_k, \sum_{\ell=1}^{s} \beta_\ell Y_\ell\right) = \sum_{k=1}^{r} \sum_{\ell=1}^{s} \alpha_k \beta_\ell \operatorname{cov}(X_k, Y_\ell)$$

<u>Proof</u>: We compute

$$E\left[\left(\sum_{k=1}^{r}\alpha_k X_k\right)\left(\sum_{\ell=1}^{s}\beta_\ell Y_\ell\right)\right]$$

$$= E\left[\sum_{k=1}^{r}\sum_{\ell=1}^{s}\alpha_k\beta_\ell X_k Y_\ell\right]$$

$$\overset{\text{lin.}}{=} \sum_{k=1}^{r}\sum_{\ell=1}^{s}\alpha_k\beta_\ell E(X_k Y_\ell)$$

On the other hand

$$E\left(\sum_{k=1}^{r}\alpha_k X_k\right)\cdot E\left(\sum_{\ell=1}^{s}\beta_\ell Y_\ell\right)$$

$$\overset{\text{lin}}{=} \sum_{k=1}^{r}\sum_{\ell=1}^{s}\alpha_k\beta_\ell E(X_k)E(Y_\ell)$$

We subtract and get the result.

<u>Remark</u>: The property is called bilinearity.

The definitions give us further properties of covariances that follow from definitions:

(i) $\operatorname{var}(aX) = a^2 \operatorname{var}(X)$

(ii) $\operatorname{cov}(X,X) = \operatorname{var}(X)$

(iii) $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$

(iv) $\operatorname{cov}(\alpha X, \beta Y) = \alpha\beta \operatorname{cov}(X,Y)$.

<u>Theorem 4.3</u>: Let $X_1, \dots X_r$ be random variables. We have

$$\operatorname{var}\left(\sum_{k=1}^{r} \alpha_k X_k\right)$$

$$= \sum_{k=1}^{r} \alpha_k^2 \operatorname{var}(X_k) + \sum_{\substack{k,\ell=1 \\ k\neq \ell}}^{r} \alpha_k \alpha_\ell \operatorname{cov}(X_k, X_\ell).$$

<u>Proof</u>: Follows directly from Theorem 4.2.

<u>Special case</u>: Let X, Y be discrete and independent.

Then

$$E(X \cdot Y) = \sum_{x,y} x \cdot y \, P(X = x, Y = y)$$

$$= \sum_{x,y} x \cdot y \, P(X = x) P(Y = y)$$

$$= \left(\sum_{x} x P(X = x) \right) \cdot \left(\sum_{y} y P(Y = y) \right)$$

$$= E(X) E(Y).$$

This means $\operatorname{cov}(X, Y) = 0$.

A similar calculation is valid for continuous X, Y.

<u>Remark</u>: For independent X, Y and functions f, g we have

$$E[f(X) g(Y)] = E(f(X)) E(g(Y))$$

with the same proof.

As a consequence for <u>independent</u> $X_1, X_2, \dots, X_n$ we have

$$\operatorname{var}(X_1 + \dots + X_n) = \operatorname{var}(X_1) + \dots + \operatorname{var}(X_n)$$

Examples: (i) If $\underline{X} = (X_1, \dots X_r)$ is multinomial we have

$$E(X_k X_l) = -np_k p_l + n^2 p_k p_l$$

and $E(X_k) = np_k$ and $E(X_l) = np_l$.

We have

$$\operatorname{cov}(X_k, X_l) = -np_k p_l$$

(ii) If

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}}$$

we have

$$E(X \cdot Y) = \rho \quad \text{and}$$

$$E(X) = E(Y) = 0$$

so

$$\operatorname{cov}(X,Y) = \rho .$$

Method of indicators:

If we can write a random variable X as a sum of indicators we can in many cases compute variances by Theorem 4.3.

If $I \sim \text{Bernoulli}(p)$ then

$$E(I^2) = E(I) = p \quad \text{so}$$

$$\text{var}(I) = E(I^2) - E(I)^2$$
$$= p - p^2$$
$$= p(1-p)$$

If I, J are indicators then

$$\boxed{\begin{aligned} \text{cov}(I,J) &= E(I \cdot J) - E(I)E(J) \\ &= P(I=1, J=1) \\ &\quad - P(I=1)P(J=1) \end{aligned}}$$

<u>Example</u>: Let $X \sim \text{HyperGeom}(n, B, N)$.

We wrote

$$X = I_1 + \dots + I_n .$$

We have

$$\operatorname{var}(X) = \sum_{k=1}^{n} \operatorname{var}(I_k)$$

$$+ \sum_{\substack{k,l=1 \\ k \neq l}}^{n} \operatorname{cov}(I_k, I_l)$$

We know: $I_k \sim \text{Bernoulli}\left(\frac{B}{N}\right)$

so $\operatorname{var}(I_k) = \frac{B}{N} \cdot \left(1 - \frac{B}{N}\right)$.

We have

$$P(I_1 = 1, I_2 = 1) = P(\text{1st \& 2nd ball black})$$

$$= \frac{B}{N} \cdot \frac{B-1}{N-1}.$$

It follows

$$\operatorname{cov}(I_1, I_2) = \frac{B}{N} \cdot \frac{B-1}{N-1} - \left(\frac{B}{N}\right)^2$$

$$= \frac{B}{N}\left[\frac{(B-1)N - B(N-1)}{N(N-1)}\right]$$

$$= \frac{B}{N}\left[\frac{-N+B}{N(N-1)}\right]$$

$$= -\frac{B}{N}\left(1-\frac{B}{N}\right) \cdot \frac{1}{N-1}$$

But by symmetry (I_k, I_l) has the same distribution as (I_1, I_2).

So all covariances are the same.

We have

$$\operatorname{var}(X) = n\frac{B}{N}\left(1-\frac{B}{N}\right) + n(n-1) \times (-1) \times \frac{B}{N}\left(1-\frac{B}{N}\right) \cdot \frac{1}{N-1}$$

$$= n\frac{B}{N}\left(1-\frac{B}{N}\right)\left(1-\frac{n-1}{N-1}\right)$$

$$= n \cdot \frac{B}{N}\left(1-\frac{B}{N}\right)\frac{N-n}{N-1}$$

4.3. Conditional expectation

Idea: If X is a discrete random variable then

$$E[f(X)] = \sum_{x_k} f(x_k) P(X = x_k).$$

The expectation is computed using the distribution. We can use the same idea with the conditional distribution $P(X = x_k \mid B)$ for $P(B) > 0$.

Definition: Let X be a discrete random variable. The conditional expectation of X given B is given by

$$E(X \mid B) = \sum_{x_k} x_k \cdot P(X = x_k \mid B)$$

and

$$E(f(X) \mid B) = \sum_{x_k} f(x_k) P(X = x_k \mid B).$$

Technical note: We understand the existence of $E(X \mid B)$ the same way as for usual expectations.

In most cases B will be of the form $B = \{Y = y_e\}$ for some random variable Y.

<u>Example</u>: Players A and B get 5 cards each from a well shuffled deck of cards. Let X be the number of aces of A and Y the number of aces of B. We have

$Y|_{X=k} \sim \text{HiperGeom}(5, 4-k, 47)$. It follows that

$$E(Y \mid X = k) = 5 \cdot \frac{4-k}{47}$$

We know that for $Z \sim \text{HiperGeom}(n, B, N)$ we have $\text{var}(Z) = n \cdot \frac{B}{N}\left(1 - \frac{B}{N}\right)\frac{N-n}{N-1}$.

so

$$E(Z^2) = \text{var}(Z) + n^2 \cdot \frac{B^2}{N^2}$$

We have

$$E(Y^2 \mid X = k)$$

$$= 5 \cdot \frac{4-k}{47}\left(1 - \frac{4-k}{47}\right) \cdot \frac{47-5}{47-1}$$

$$+ 5^2 \cdot \frac{(4-k)^2}{47^2}$$

There is an alternative way to write the conditional expectation.

If X is discrete and B is an event we compute

$$E[\underbrace{X \cdot \mathbb{1}_B}] = \sum_{x_k} x_k \cdot P(\{X = x_k\} \cap B)$$

$$= (*)$$

This random variable has value x_k with probability $P(\{X = x_k\} \cap B)$ and possibly 0 with probability

$P(B \cup \{X = 0\})$

$$(*) = \sum_{x_k} x_k \frac{P(\{X = x_k\} \cap B)}{P(B)} \cdot P(B)$$

$$= P(B) \cdot E(X \mid B)$$

We have

$$E(X \mid B) = \frac{E(X \cdot 1_B)}{P(B)}$$

$$E[f(X) \mid B] = \frac{E[f(X) \cdot 1_B]}{P(B)}.$$

!

<u>Theorem 4.4</u>: Let $\{H_1, H_2, \dots\}$ be a partition of $(\Omega, \mathcal{F}, P)$. Let X be a discrete random variable with $E(|X|) < \infty$. We have

$$E(X) = \sum_k E(X \mid H_k) \cdot P(H_k)$$

Proof: We compute

$$\sum_k E(X \mid H_k) \cdot P(H_k)$$

$$= \sum_k \left(\sum_{x_\ell} x_\ell \, P(X = x_\ell \mid H_k) \right) P(H_k)$$

$$= \sum_{x_\ell} x_\ell \underbrace{\sum_k P(X = x_\ell \mid H_k) P(H_k)}_{= P(X = x_\ell)}$$

$$= \sum_{x_\ell} x_\ell \, P(X = x_\ell)$$

$$= E(X)$$

We used the law of total probabilities. The statement is the law of total expectations.

Example: We toss a coin until we get r consecutive heads. Tosses are independent and the probability of heads is p. Let X be the number of tosses needed.

Example : if $r = 4$ and we get

$$\underbrace{H\,T\,T\,H\,H\,T\,H\,T\,T\,T\,H\,T\,H\,H\,H\,H}_{X = 15}$$

We want $E(X)$. Let

$H_k = \{$the first T appears in position $k\}$.

We have

$$E(X \mid H_k) = r \quad \text{if} \quad k = r+1, \dots$$

and

$$E(X \mid H_k) = k + E(X) \quad \text{if} \quad k = 1, \dots, r \quad ;$$

The law of total expectation gives

$$E(X) = \sum_{k=1}^{\infty} E(X \mid H_k)\, P(H_k)$$

$$= \sum_{k=1}^{r} (k + E(X))\, P(H_k)$$

$$+ \sum_{k=r+1}^{\infty} r \cdot P(H_k)$$

This last expression is a linear equation for $E(x)$. We compute

(i) $$\sum_{k=r+1}^{\infty} r\, P(H_k) =$$

$$= r \cdot \sum_{k=r+1}^{\infty} p^{k-1} \cdot q$$

$$= r \cdot p^r \cdot q \cdot \sum_{k=0}^{\infty} p^k$$

$$= r \cdot p^r \cdot q \cdot \frac{1}{1-p}$$

$$= r \cdot p^r$$

(ii) $$\sum_{k=1}^{r} P(H_k) = q \cdot \sum_{k=1}^{r} p^{k-1}$$

$$= q \cdot \frac{1-p^r}{1-p}$$

$$= 1 - p^r$$

(iii) $$\sum_{k=1}^{r} k \cdot P(H_k) = \sum_{k=1}^{r} k \cdot p^{k-1} \cdot q$$

$$= \frac{d}{dp}\left(p + p^2 + \quad + p^r\right) \cdot q$$

$$= (*)$$

$$(*) = \frac{d}{dp}\left(\frac{p(1-p^r)}{1-p}\right)\cdot q$$

$$= \frac{[(1-p^r) - rpp^{r-1}](1-p) + p(1-p^r)}{(1-p)^2}\cdot q$$

$$= \frac{1-p^r - rp^r + rp^{r+1}}{(1-p)^2}\cdot q$$

$$= \frac{1-p^r - rp^r + rp^{r+1}}{q}$$

Rewrite

$$E(x) = \frac{1-p^r - rp^r + rp^{r+1}}{q}$$

$$+ E(x)\cdot(1-p^r)$$

$$+ r\cdot p^r$$

We have

$$E(x) = r + \frac{1-p^r - rp^r + r\cdot p^{r+1}}{q\cdot p^r}$$

$$= r + \frac{1-p^r + rp^r(p-1)}{q\cdot p^r}$$

$$= \frac{1-p^r}{q\cdot p^r}$$

$$E[f(\underline{X})] = \sum_k E[f(\underline{X}) \mid H_k]\, P(H_k)$$

The proof is identical to the proof we had before.

<u>Definition</u>: We define

$$\operatorname{var}(X \mid B) = E(X^2 \mid B) - [E(X \mid B)]^2$$

and

$$\operatorname{cov}(X, Y \mid B) = E(XY \mid B) - E(X \mid B)\, E(Y \mid B).$$

For continuous random variables we use conditional densities to compute conditional expectations. We have:

<u>Definition</u>: Let Y have conditional density $f_{Y \mid X = x}(y)$ given $X = x$.

We define

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y \, f_{Y \mid X = x}(y) \, dy$$

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

<u>Technical note</u>: We define existence the way existence is defined for usual expectations.

<u>Comment</u>: The same definition holds for vectors. We define

$$E[f(\underline{Y}) \mid \underline{X} = \underline{x}] = \int_{\mathbb{R}^2} f(y) \, f_{\underline{Y} \mid \underline{X} = x}(y) \, dy.$$

<u>Definition</u>: We define

$$\operatorname{var}(Y \mid \underline{X} = \underline{x}) = E(Y^2 \mid \underline{X} = \underline{x}) - [E(Y \mid \underline{X} = \underline{x})]^2$$

and

$$\operatorname{cov}(Y_1, Y_2 \mid \underline{X} = \underline{x})$$

$$= E(Y_1 \cdot Y_2 \mid \underline{X} = \underline{x})$$

$$- E(Y_1 \mid \underline{X} = \underline{x}) E(Y_2 \mid \underline{X} = \underline{x})$$

<u>Example</u>: Let

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho x y + y^2}{2(1-\rho^2)}}$$

We have computed that

$$f_{Y \mid X = x}(y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}$$

or $Y \mid X = x \sim N(\rho x, 1-\rho^2)$. From this we have

$$E(Y \mid X = x) = \rho x \quad \text{and}$$

$$\operatorname{var}(Y \mid X = x) = 1 - \rho^2.$$

5. Generating functions

5.1. Definition and basic properties

The idea of generating functions comes from analysis and combinatorics. If $c_0, c_1, \dots$ is a sequence of complex numbers then we can define the power series

$$G(s) = \sum_{k=0}^{\infty} c_k \cdot s^k \quad \text{for } s \in \mathbb{C}.$$

We know from analysis that such power series converge for $|s| < R$ where R is the radius of convergence. Analysis further gives that

$$\frac{1}{R} = \limsup_{n\to\infty} \sqrt[n]{|c_n|}.$$

If $|c_n| \leq 1$ for all n then

$\frac{1}{R} \leq 1 \Rightarrow R \geq 1.$

In this chapter we will only look at non-negative integer valued random variables.

<u>Definition</u>: Let X be a random variable with values $0, 1, 2, \ldots$. We define the generating function of X, denoted by $G_X(s)$ as the power series

$$G_X(s) = \sum_{k=0}^{\infty} P(X=k) \cdot s^k$$

<u>Comments</u>:

(i) The idea is to "pack" up the distribution in a function.

(ii) Since $\sum_{k=0}^{\infty} P(X=k) = 1.$

the power series is dominated by $P(X=k)$ for $|s| \leq 1$ and converges uniformly to a continuous function.

Examples :

(i) if $X \sim Bin(n,p)$ we have

$$G_X(s) = \sum_{k=0}^{n} P(X=k) \cdot s^k$$

$$= \sum_{k=0}^{n} \binom{n}{k} p^k \cdot q^{n-k} \cdot s^k$$

$$= \sum_{k=0}^{n} \binom{n}{k} (ps)^k q^{n-k}$$

$$= (ps+q)^n .$$

(ii) if $X \sim Po(\lambda)$ we have

$$G_X(s) = \sum_{k=0}^{\infty} P(X=k)\, s^k$$

$$= \sum_{k=0}^{\infty} \frac{e^{-\lambda} \cdot \lambda^k}{k!} \cdot s^k$$

$$= (*)$$

$$(*) = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!}$$

$$= e^{-\lambda} \cdot e^{\lambda s}$$

$$= e^{-\lambda(1-s)}$$

(iii) Let $X \sim NegBin(m, p)$.

From analysis we have that for $|x| < 1$

$$(1+x)^a = \sum_{k=0}^{\infty} \binom{a}{k} x^k \quad \text{where}$$

$$\binom{a}{k} = \frac{a(a-1)\cdots(a-k+1)}{k!}$$

The above formula is known as the Newton formula. Replace x by $-x$ and let $a = -r$ for some integer $r > 0$.

We get

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \binom{-r}{k} \cdot (-x)^k$$

$$= \sum_{k=0}^{\infty} \frac{(-r)(-r-1)\cdots(-r-k+1)}{k!} (-1)^k \cdot x^k$$

$$= \sum_{k=0}^{\infty} \frac{r(r+1)\cdots(r+k-1)}{k!} \cdot x^k$$

$$= \sum_{k=0}^{\infty} \frac{(r-1)!\, r(r+1)\cdots(r+k-1)}{(r-1)! \cdot k!} x^k$$

$$= \sum_{k=0}^{\infty} \frac{(r+k-1)!}{(r-1)! \cdot k!} x^k$$

$$= \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} x^k.$$

We compute

$$G_X(s) = \sum_{k=m}^{\infty} P(X=k) \cdot s^k$$

$$= \sum_{k=m}^{\infty} \binom{k-1}{m-1} \cdot p^m \cdot q^{k-m} \cdot s^k$$

$$= (*)$$

$$(*) = \sum_{\ell=0}^{\infty} \binom{m+\ell-1}{m-1} p^m \cdot q^\ell s^{m+\ell}$$

$$= p^m \cdot s^m \cdot \sum_{\ell=0}^{\infty} \binom{m+\ell-1}{m-1} s^\ell \cdot q^\ell$$

$$= \frac{p^m \cdot s^m}{(1-qs)^m}$$

$$= \left(\frac{ps}{1-qs} \right)^m$$

(iv) The computation in previous example gives

$$(1-x)^{-a} = \sum_{k=0}^{\infty} \frac{(a)_k}{k!} \cdot x^k \qquad |x| < 1.$$

Let X have the Pólya distribution

$$P(X=k) = \frac{\beta^a \, (a)_k}{k! \, (1+\beta)^{a+k}}$$

We have

$$G_X(s) = \sum_{k=0}^{\infty} P(X=k) \cdot s^k$$

$$= \sum_{k=0}^{\infty} \frac{\beta^a (a)_k}{k!\,(1+\beta)^{a+k}} \cdot s^k$$

$$= \sum_{k=0}^{\infty} \frac{\beta^a (a)_k}{k!\,(1+\beta)^a} \left(\frac{s}{1+\beta}\right)^k$$

$$= \frac{\beta^a}{(1+\beta)^a} \cdot \left(1 - \frac{s}{1+\beta}\right)^{-a}$$

$$= \left(\frac{\beta}{1+\beta-s}\right)^a$$

<u>Theorem 5.1</u>: Let X be a nonnegative integer valued random variable and let $G_X(s)$ be its generating function. Then $G_X(s)$ uniquely determines the distribution of X.

Proof: Since $G_X(s)$ converges for $|s| < 1$ we have

$$G_X^{(n)}(0) = n!\, P(X = n).$$

Theorem 5.2: Let X be an integer valued random variable with generating function $G_X(s)$.

(i)

$$E(X) = \lim_{s \uparrow 1} G_X'(s)$$

(ii)

$$E[X(X-1) \cdots (X-m+1)]$$

$$= \lim_{s \uparrow 1} G_X^{(m)}(s)$$

Proof: Let $\varepsilon > 0$ and assume first that $E(X) < \infty$.

There is a N_ε such that for $n \geq N_\varepsilon$ we have $\sum_{k=n}^{\infty} k P(X=k) < \varepsilon$.

This means that

$$E(X) - \sum_{k=0}^{N_\varepsilon - 1} k P(X=k) < \varepsilon.$$

Since all the coefficients in the power series are non-negative we have that for $s \in (0,1)$

$$\sum_{k=1}^{N_\varepsilon - 1} k P(X=k) s^{k-1} \leq G_X'(s) \leq E(X).$$

As $s \uparrow 1$ we have

$$\sum_{k=1}^{N_\varepsilon - 1} k P(X=k) \leq \lim_{s \uparrow 1} G_X'(s) \leq E(X)$$

The limit exists because $G_X(s)$ is nondecreasing on $(0,1)$. But the above means that

$$E(X) - \varepsilon \leq \lim_{s \uparrow 1} G_X(s) \leq E(X)$$

for all $\varepsilon > 0$.

If $E(X) = \infty$ we have that

$$\lim_{s \uparrow 1} G_X'(s) \geq \sum_{k=1}^{N} k P(X=k)$$

for any finite N. This implies that $\lim_{s \uparrow 1} G_X'(s) = \infty$.

(ii) The proof is similar.

<u>Theorem 5.3</u>: Let X, Y be independent. Then

$$\boxed{G_{X+Y}(s) = G_X(s) \cdot G_Y(s)}$$

<u>Proof</u>: We can write

$$E(s^X) = \sum_{k=0}^{\infty} s^k \cdot P(X=k) = G_X(s)$$

Then

$$E(s^{X+Y}) = E(s^X \cdot s^Y) \overset{\text{indp}}{=} E(s^X) \cdot E(s^Y)$$

$= \quad G_X(s) \cdot G_Y(s)$

<u>Comment</u>: This is the most important property of generating functions.

By extension we have for independent $X_1, X_2, \ldots, X_r$

$$\boxed{G_{X_1+X_2+\cdots+X_r}(s) = G_{X_1}(s) \cdot G_{X_2}(s) \cdots G_{X_r}(s)}$$

<u>Examples</u>:

(i) X, Y independent $X \sim Bin(m,p)$ and $Y \sim Bin(n,p)$. We have

$$\begin{aligned} G_{X+Y}(s) &= G_X(s) \cdot G_Y(s) \\ &= (ps+q)^m \cdot (ps+q)^n \\ &= (ps+q)^{m+n} \end{aligned}$$

This last function is the generating function of the $\mathrm{Bin}(m+n, p)$ distribution. It follows, by uniqueness, $X+Y \sim \mathrm{Bin}(m+n, p)$.

(ii) Let X, Y be independent and

$$P(X=k) = \frac{\beta^a (a)_k}{k!\,(1+\beta)^{a+k}}, \quad k = 0, 1, \dots$$

$$P(Y=l) = \frac{\beta^b (b)_l}{l!\,(1+\beta)^{b+l}}, \quad l = 0, 1, \dots$$

We have

$$G_{X+Y}(s) = G_X(s) \cdot G_Y(s)$$

$$= \left(\frac{\beta}{1+\beta-s}\right)^a \cdot \left(\frac{\beta}{1+\beta-s}\right)^b$$

$$= \left(\frac{\beta}{1+\beta-s}\right)^{a+b}$$

Conclusion:

$$P(X+Y=k) = \frac{\beta^{a+b}\,(a+b)_k}{k!\,(1+\beta)^{a+b}}, \quad k=0,1,\ldots$$

<u>Comment</u>: We have computed the distribution of $X+Y$ before but the above is much more elegant.

(iii) Suppose $X_1, X_2, \ldots, X_r$ are independent and $X_k \sim \text{Geom}(p)$. We have

$$G_{X_k}(s) = \sum_{i=1}^{\infty} s^i \cdot q^{i-1} \cdot p$$

$$= ps \sum_{i=0}^{\infty} (qs)^i$$

$$= \frac{ps}{1-qs}$$

It follows

$$G_{X_1+\ldots X_r}(s) = \left(\frac{ps}{1-qs}\right)^r$$

Conclusion: $X_1+\cdots+X_r \sim \text{NegBin}(r,p)$.

5.2. Branching processes

In applications of probability we often calculate sums of a random number of random variables. Let $X_1, X_2, \ldots$ be random variables and N an integer valued nonnegative random variables. We need to define $X_1 + X_2 + \cdots + X_N$. Formally we define

$$\boxed{X = \sum_{k=1}^{\infty} X_k \cdot 1(N \geq k)}$$

<u>Comment</u>: For a fixed $\omega \in \Omega$ we have $N(\omega) < \infty$ and so the sum is finite because only a finitely many terms are $\neq 0$.

We will write

$$X = X_1 + X_2 + \cdots + X_N.$$

Theorem 5.4: Let $N, X_1, X_2, \ldots$ be independent, $X_1, X_2, \ldots$ equally distributed non-negative integer valued random variables and N non-negative integer valued. Let $X = X_1 + X_2 + \cdots + X_N$. Then

$$\boxed{G_X(s) = G_N(G_{X_1}(s)).}$$

Proof: We use the formula for the total expectation.

$$\begin{aligned}
G_X(s) &= E(s^X) \\
&= \sum_{k=0}^{\infty} E(s^X \mid N = k) P(N = k) \\
&= \sum_{k=0}^{\infty} E(s^{X_1 + \cdots + X_k} \mid N = k) P(N = k) \\
&\overset{\text{indep}}{=} \sum_{k=0}^{\infty} E(s^{X_1 + \cdots + X_k}) P(N = k) \\
&= (*)
\end{aligned}$$

$$(*) = \sum_{k=0}^{\infty} [G_{X_1}(s)]^k \cdot P(N=k)$$

$$= G_N(G_{X_1}(s)).$$

Example: A hen lays N eggs. A chick hatches from each egg with probability p independent of all other eggs. Suppose $N \sim Po(\lambda)$. What is the distribution of the number of chicks? In mathematical notation we are asking about the distribution of $I_1 + I_2 + \dots + I_N$ where $I_1, I_2, \dots$ are independent with $I_k \sim \text{Bernoulli}(p)$ and independent of N. We have

$$G_X(s) = G_N(G_{X_1}(s)).$$

$$G_{I_1}(s) = s^0 P(I_1 = 0) + s^1 \cdot P(I_1 = 1)$$
$$= q + ps$$

We have

$$G_X(s) = G_N(q + ps)$$

$$= e^{-\lambda(1-q-ps)}$$

$$= e^{-\lambda(p-ps)}$$

$$= e^{-\lambda p(1-s)}$$

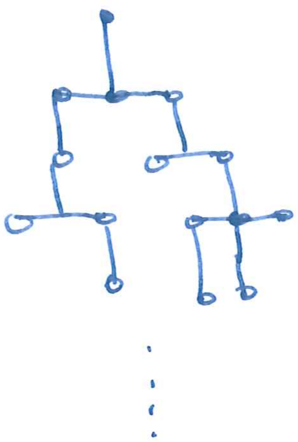
This last function is the generating function of the $P(\lambda p)$ distribution so $X \sim Po(\lambda p)$.

Branching processes

In 1874 Sir Francis Galton (1822-1911) asked the following question: suppose you take an English aristocrat. He will have a random number of sons. His sons will have a random number of sons, ...

The problem is to determine the probability that the family tree will die out.

<u>Figure</u>: A possible family tree



$z_0 = 1$

$z_1 = 3$

$z_2 = 3$

$z_3 = 5$

$z_4 = 3$

The problem was solved by Galton and Watson in 1875 (F. Galton, H.W. Watson, [illegible] (1875) On the probability of the extinction of families, Journal of the Royal Anthropological Institute 4, 138 – 144) using generating functions.

To solve the problem mathematically we need a few additional assumptions:

(i) Generations are simultaneous.

(ii) Each individual has sons independently of all the others.

(iii) The random number of sons has the same distribution for all individuals.

The above assumptions imply the following mathematical formulation:

Let $\{\xi_{n,k}\}_{n\geq 1, k\geq 1}$ be independent, equally distributed non-negative integer valued random variables with generating function G.

We define

$Z_0 = 1$ and recursively

$Z_{n+1} = \xi_{n+1,1} + \xi_{n+1,2} + \cdots + \xi_{n+1,Z_n}$

The sequence $Z_0, Z_1, \ldots$ of random variables is called the <u>branching process</u>.

The above means that Z_n individuals in the n-th generation have randomly many offspring.

The random variable Z_n depends on $\xi_{m,k}$ for $m \le n$ so it is independent of

$$\xi_{n+1,1}, \ \xi_{n+1,2}, \ \ldots$$

Denote $G_n(s) = G_{Z_n}(s)$. By Theorem 5.4 we have

$$G_{n+1}(s) = G_n(G(s)).$$

By definition $G_1(s) = G(s)$ and by the above recursion

$$G_2(s) = G_1(G(s)) = (G \circ G)(s)$$

$$G_3(s) = G_2(G(s)) = (G \circ G \circ G)(s)$$

$$\vdots$$

$$G_n(s) = (G \circ G \circ \cdots \circ G)(s).$$

Since composition is associative we have

$$\boxed{G_{n+1}(s) = G(G_n(s))}$$

Let $A = \{$the family tree dies out$\}$.

The family tree dies out if one of the generations is empty so

$$A = \bigcup_{n=1}^{\infty} \{Z_n = 0\}.$$

But $\{Z_1 = 0\} \subseteq \{Z_2 = 0\} \subseteq \dots\dots$

In the first chapter we proved that for $A_1 \subseteq A_2 \subseteq \dots$ we have

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n).$$

Denote $\eta = P(A)$. We have

$$\eta = P(A) = \lim_{n\to\infty} P(Z_n = 0).$$

But $P(Z_n = 0) = G_n(0)$.

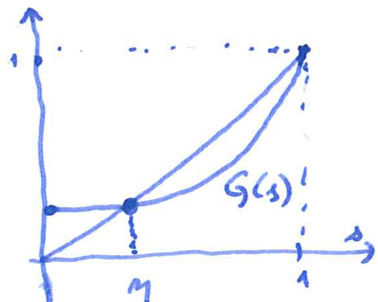
Theorem 5.5: The probability η satisfies the equation $\eta = G(\eta)$ and is the smallest solution of the above equation on $[0,1]$.

Comments:

(i) If $\eta = G(\eta)$ we say that η is a fixed point of G.

(ii) Since $G(1) = 1$ there is always at least one fixed point on $[0,1]$. The set of fixed points is compact so it contains a smallest point.

Figure:



Proof: $G(s)$ is continuous on $[0,1]$.

So we have

$$\eta = \lim_{n\to\infty} G_{n+1}(0) = \lim_{n\to\infty} G(G_n(0))$$

$$= G\left(\lim_{n\to\infty} G_n(0)\right) = G(\eta).$$

So η is a fixed point. To prove that η is the smallest fixed point let $\overline{\eta}$ be a fixed point on $[0,1]$. We have

$$0 \leq \overline{\eta}$$

Since G is non decreasing on $[0,1]$ it follows

$$G(0) \leq G(\overline{\eta}) = \overline{\eta}$$

$$G(G(0)) \leq G(\overline{\eta}) = \overline{\eta}$$

$$\vdots$$

$$(G \circ \cdots \circ G)(0) = G_n(0) \leq \overline{\eta}$$

So

$$\lim_{n\to\infty} G_n(0) = \eta \leq \overline{\eta}.$$

This means that any fixed point $\bar{\eta}$ is $\geq \eta$ which proves the theorem.

<u>Example</u>: Suppose every individual has 0, 1, 2, 3 sons with probability 1/4 each. This means that

$$G(s) = \frac{1+s+s^2+s^3}{4}.$$

We need all solutions of

$$G(s) = s \iff 1 - 3s + s^2 + s^3 = 0$$

We know that $s = 1$ is a solution so we can factor

$$1 - 3s + s^2 + s^3 = (s-1)(s^2+2s-1)$$

The solutions are

$s = 1$

$s = -1+\sqrt{2}$

$s = -1-\sqrt{2}$

The smallest fixed point on $[0,1]$ is $-1+\sqrt{2}$ $\doteq 0.4142$.

Example: Suppose $G(s) = \frac{p}{1-qs}$.

What is $G_n(s)$?

$$G_2(s) = G(G(s))$$

$$= \frac{p}{1 - q \cdot \frac{p}{1-qs}}$$

$$= \frac{p(1-qs)}{1 - pq - qs}$$

$$G_3(s) = \frac{p}{1 - q \cdot \frac{p(1-qs)}{1-pq-qs}}$$

$$= \frac{p(1-pq-qs)}{1-pq-qs-pq(1-qs)}$$

$$= \frac{p(1-pq-qs)}{1-2pq-q(1-pqs)}$$

We see that

$$G_n(s) = \frac{a_n - b_n s}{c_n - d_n s}$$

$$G_{n+1}(s) = \frac{a_n - b_n \cdot \frac{p}{1-qs}}{c_n - d_n \cdot \frac{p}{1-qs}}$$

Multiplying out we get

$$a_{n+1} = a_n - p \cdot b_n$$
$$b_{n+1} = q a_n$$

We have $g_0(s) = s \Rightarrow a_0 = 0, b_0 = -1$

Write in matrix form

$$\begin{pmatrix} a_{n+1} \\ b_{n+1} \end{pmatrix} = \begin{pmatrix} 1, & -p \\ q, & 0 \end{pmatrix} \begin{pmatrix} a_n \\ b_n \end{pmatrix}.$$

Iteration gives

$$\begin{pmatrix} a_n \\ b_n \end{pmatrix} = \begin{pmatrix} 1, & -p \\ q & 0 \end{pmatrix}^n \begin{pmatrix} a_0 \\ b_0 \end{pmatrix}$$
$$= \begin{pmatrix} 1, & -p \\ q & 0 \end{pmatrix}^n \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

We need to find the power of the matrix. Suppose $p \neq q$.

We can check by multiplication that

$$\underbrace{\begin{pmatrix} p & 1 \\ q & 1 \end{pmatrix}}_{A} \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \underbrace{\begin{pmatrix} \frac{1}{p-q}, & -\frac{1}{p-q} \\ -\frac{q}{p-q} & \frac{p}{p-q} \end{pmatrix}}_{A^{-1}}$$
$$= \begin{pmatrix} 1, & -p \\ q, & 0 \end{pmatrix}$$

We have diagonalized the matrix $\begin{pmatrix} 1, & -p \\ q & 0 \end{pmatrix}$. This means

$$\begin{pmatrix} 1, & -p \\ q, & 0 \end{pmatrix}^n = \begin{pmatrix} p & 1 \\ q & 1 \end{pmatrix}\begin{pmatrix} p^n & 0 \\ 0 & q^n \end{pmatrix}\begin{pmatrix} 1, & -1 \\ -q & p \end{pmatrix} \cdot \frac{1}{p-q}$$

$$= \begin{pmatrix} p^{n+1} - q^{n+1}, & -p^{n+1} + pq^n \\ qp^n - q^{n+1}, & -qp^n + pq^n \end{pmatrix} \cdot \frac{1}{p-q}$$

We find:

$$\begin{pmatrix} 1, & -p \\ q & 0 \end{pmatrix}^n \begin{pmatrix} 0 \\ -1 \end{pmatrix} =$$

$$= \frac{1}{p-q} \begin{pmatrix} p(p^n - q^n) \\ pq(p^{n-1} - q^{n-1}) \end{pmatrix} = \begin{pmatrix} a_n \\ b_n \end{pmatrix}$$

For c_n, d_n the procedure is the same except that $\begin{pmatrix} c_0 \\ d_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and

$$\begin{pmatrix} c_n \\ d_n \end{pmatrix} = \begin{pmatrix} p^{n+1} - q^{n+1} \\ q(p^n - q^n) \end{pmatrix} \cdot \frac{1}{p-q}$$

Finally,

$$G_n(s) = \frac{p\left(p^n - q^n - qs\left(p^{n-1} - q^{n-1}\right)\right)}{p^{n+1} - q^{n+1} - qs\left(p^n - q^n\right)}$$

We find

$$P(Z_n = 0) = G_n(0)$$

$$= \frac{p\left(p^n - q^n\right)}{p^{n+1} - q^{n+1}}$$

We get: if $p > q$, then

$$\lim_{n \to \infty} P(Z_n = 0) = 1$$

if $p < q$

$$\lim_{n \to \infty} P(Z_n = 0) = \frac{p}{q} < 1.$$

<u>Comment</u>: The fixed points satisfy $\frac{p}{1 - qs} = s \Rightarrow$

$$qs^2 - s + p = (s-1)(qs-p) = 0 \Rightarrow$$

There is a fixed point in $[0,1]$ other than 1 if $p < q$.

<u>Comment</u> : For $p = q = 1/2$ we get

$$G_n(s) = \frac{n - (n-1)s}{n+1 - ns}$$

and

$$G_n(0) = \frac{n}{n+1} \to 1, \quad \text{as } n \to \infty.$$

Theorem 5.6: Let $Z_0, Z_1, Z_2, \dots$ be a branching process. Let $\mu = E(Z_1)$.

(i) If $\mu < 1$ then $\eta = 1$.

(ii) If $\mu > 1$ then $\eta \in [0,1)$.

(iii) If $\mu = 1$ and $G(s) \not\equiv s$ then $\eta = 1$.

Comment: The case $G(s) = s$ is uninteresting.

The proof of the theorem has more to do with analysis than probability.

Proof: Note that $\mu = \lim_{s \uparrow 1} G'(s)$.

The function $G'(s)$ is nondecreasing on $(0,1)$.

(i) if $\mu < 1$, then $G'(s) \leq \mu < 1$ for all $s \in (0,1)$. If $G(\bar{\eta}) = \bar{\eta}$ for $\bar{\eta} \in (0,1)$ then by Lagrange theorem

$$G(1) - \underbrace{G(\bar{\eta})}_{} = G'(\xi)(1-\bar{\eta}) \quad \text{for } \xi \in (\bar{\eta}, 1).$$

$$\overset{\shortparallel}{1 - \bar{\eta}}$$

But $G'(\xi) < 1$. So we have a contradiction.

(ii) If $\mu > 1$ there is a $\delta > 0$ such that $G'(s) > 1$ for $s \in (1-\delta, 1)$. By Lagrange for $s \in (1-\delta, 1)$ we have

$$G(1) - G(s) = G'(\xi)(1-s) > (1-s)$$

for $\xi \in (s,1)$. This implies

$$1 - G(s) > 1 - s \Rightarrow G(s) < s.$$

On the other hand $G(0) \geq 0$.

For $s \in (1-\delta, 0)$ we have that

$$F(0) = G(0) - 0 \geq 0$$

$$F(s) = G(s) - s < 0.$$

There must be a zero of F on $(0, s) \subset (0,1) \Rightarrow \eta \in (0,1)$.

(iii) If $G(s) \neq s$ the either $G(s) = 1$ in which case $\eta = 1$ or G is strictly convex on $(0;1)$, or $G'(s)$ is strictly increasing on $(0,1)$.

If $G(\bar{\eta}) = \bar{\eta}$ for some $\bar{\eta} \in (0,1)$ this would imply

$$G(1) - G(\bar{\eta}) = 1 - \bar{\eta}$$
$$= G'(\xi)(1-\bar{\eta})$$

for some $\xi \in (\bar{\eta}, 1)$. This means $G'(\xi) = 1$. But $G'(s)$ is strictly increasing meaning $\lim_{s \uparrow 1} G'(s) > 1$.

A contradiction.

Panjer recursion

If $X = X_1 + X_2 + \cdots + X_N$ we know that

$$G_X(s) = G_N(G_{X_1}(s)).$$

In principle we get $P(X=k)$ by expanding the right side into power series. But this is often difficult and recursive formulae are needed. This problem is often dealt with in insurance.

Definition: The random variable N is of Panjer class if

$$P(N=n) = \left(a + \frac{b}{n}\right) P(N=n-1)$$

for $n = 1, 2, \ldots\ldots$

Examples: (i) Take $a = 0$ and $b > 0$.

We get $P(N=n) = \frac{b}{n} P(N=1) \Rightarrow$

$P(N=n) = e^{-b} \cdot \frac{b^n}{n!} \Rightarrow N \sim Po(b)$.

(ii) Suppose $N \sim Bin(M, p)$. We have computed

$$\frac{P(N=n)}{P(N=n-1)} = \frac{M-n+1}{n} \cdot \frac{p}{q}$$

$$= \left(-\frac{p}{q} + \frac{(M+1)p}{q \cdot n} \right)$$

We take $a = -\frac{p}{q}$, $b = \frac{(M+1)p}{q}$

We see that $P(N = M+1) = 0$.

We compute

$$P(N=1) = \left(-p/q + \frac{(M+1)p}{q} \right) P(N=0)$$

$$= \frac{p}{q} M \cdot P(N=0)$$

$$P(N=2) = \left(-p/q + \frac{(M+1)p}{2q} \right) P(N=1)$$

$$= \frac{p}{q} \left(-1 + \frac{M+1}{2q} \right) P(N=1)$$

$$= \frac{p}{q} \cdot \frac{M-1}{2} \cdot \frac{p}{q} \cdot \frac{M}{1}$$

Continuing we get

$$P(N=n) = \left(\frac{p}{q}\right)^n \cdot \frac{M(M-1)\cdots(M-n+1)}{n!} \; P(N=0)$$

$$= \left(\frac{p}{q}\right)^n \binom{M}{n} \cdot P(N=0)$$

Because all probabilities must add to 1 we have

$$\left(1+\frac{p}{q}\right)^M P(N=0) = 1 \Rightarrow$$

$$P(N=0) = q^M$$

or

$$P(N=n) = \binom{M}{n} p^n q^{M-n}.$$

Conclusion: N is in the Panjer class.

Theorem 5.7: For $|s| < 1$ the generating function of N satisfies

$$(1 - as)\, G_N'(s) = (a+b)\, G_N(s).$$

Proof: From the recursion equation we have

$$P(N=n) \cdot s^n = \left(a + \frac{b}{n}\right) P(N=n-1) \cdot s^n.$$

Sum both sides over $n = 1, 2, \dots$. We get

$$G_N(s) - G_N(0)$$

$$= a \sum_{n=1}^{\infty} P(N=n-1)\, s^n$$

$$+ b \cdot \sum_{n=1}^{\infty} \frac{s^n}{n} P(N=n-1)$$

$$= as\, G_N(s) + b \sum_{n=0}^{\infty} \left(\int_0^s u^n\, du\right) P(N=n)$$

$$= as\, G_N(s) + b \cdot \int_0^s \left(\sum_{n=0}^{\infty} u^n P(N=n)\right) du$$

$$= as\, G_N(s) + b \int_0^s G_N(u)\, du.$$

It is legitimate to interchange summation and integration because the sum converges uniformly on $[0,1]$.

Take derivatives to get

$$G_N'(s) = a\,G_N(s) + as\,G_N'(s) + b\,G_N(s).$$

Rearranging gives the equation.

If $X = X_1 + X_2 + \cdots + X_N$ where X_1, X_2 are independen equally distributed we get

$$G_X(s) = G_N(G_{X_1}(s))$$

Taking derivatives we get

$$G_X'(s) = G_N'(G_{X_1}(s)) \cdot G_{X_1}'(s)$$

Multipy both sides by

$$1 - aG_{X_1}(G_{X_1}(s))$$

and use Theorem 5.7. We get

$$\boxed{G_X'(s)(1 - a G_{X_1}(s)) = (a+b) G_X(s) G_{X_1}'(s)}$$

Denote $P(N = n) = p_n$ for $n = 0, 1, \ldots$.

Denote $P(X = r) = g_r$ and

$P(X_1 = k) = f_k$ for $k = 0, 1, \ldots$

We have that $X = 0$ if either

$N = 0$ or $N > 0$ and $X_1 + \cdots + X_N = 0$.

It follows

$$P(X = 0) = P(N = 0) + \sum_{n=1}^{\infty} P(N = n) \cdot f_0^n$$

$$= G_N(f_0)$$

In our notation

$$P(X = 0) = g_0 = G_N(f_0).$$

From Analysis we know that

$$\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty} c_k x^k$$

with

$$c_k = \sum_{i=0}^{k} a_i b_{k-i}.$$

Comment: This is called the Cauchy product of power series.

In the formula connecting the generating functions

Equate coefficients for s^n.

h : $$(n+1) g_{n+1} - a \sum_{k=0}^{n} f_k \cdot (n+1-k) g_{n+1-k}$$

R : $$(a+b) \sum_{k=0}^{n} (k+1) f_{k+1} \cdot g_{n-k}$$

$$= (a+b) \sum_{k=1}^{n+1} k f_k \cdot g_{n-k+1}$$

Rearranging we get

$$(n+1) g_{n+1} - a f_0 (n+1) g_{n+1}$$

$$= a \sum_{k=1}^{n} f_k (n+1-k) g_{n+1-k} +$$

$$(a+b) \sum_{k=0}^{n} (k+1) f_{k+1} g_{n-k}$$

$$= a \sum_{k=1}^{n} f_k (n+1-k) g_{n+1-k}$$

$$+ (a+b) \sum_{k=1}^{n+1} k f_k \cdot g_{n+1-k}$$

$$= a \cdot \sum_{k=1}^{n+1} f_k (n+1-k) g_{n+1-k}$$

$$+ (a+b) \sum_{k=1}^{n+1} k f_k g_{n+1-k}$$

Divide by $(n+1)(1-a f_0)$ to get

$$g_{n+1} = \frac{1}{1-a f_0} \sum_{k=1}^{n+1} \left(a + \frac{bk}{n+1}\right) f_k g_{n+1-k}$$

Theorem 5.8 (Panjer recursion)

We have

$$\boxed{g_{n+1} = \frac{1}{1-a f_0} \sum_{k=1}^{n+1} \left(a + \frac{bk}{n+1}\right) f_k g_{n+1-k}}$$

6. The central limit theorem

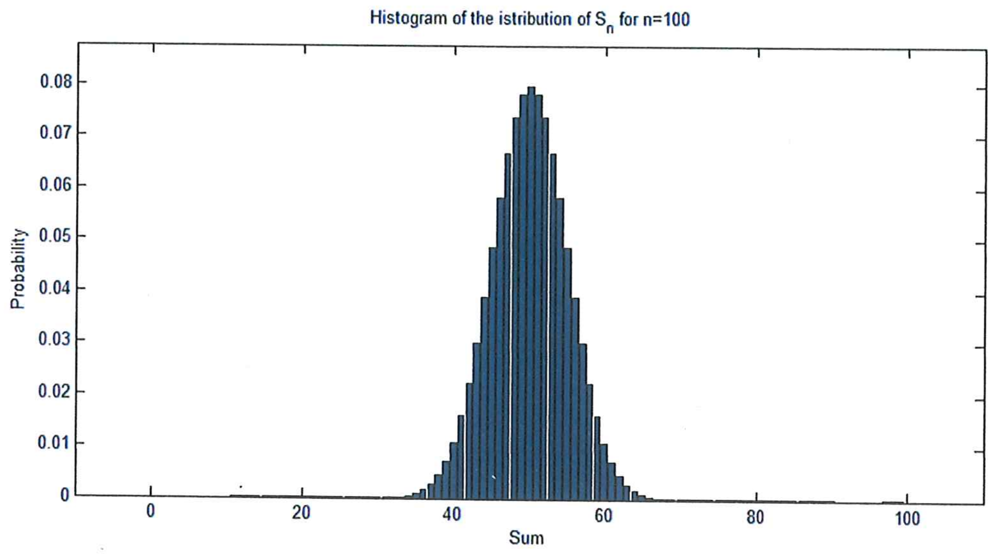
We will be interested in the distribution of sums $S_n = X_1 + X_2 + \cdots + X_n$. For some types of distributions we know the answer but in general this question is difficult to answer. Moreover, in statistics we need to approximate such distributions even if we do not exactly know the distributions of $X_1, X_2, \ldots\ldots$

The setup we will look at will be: $X_1, X_2, \ldots$ are independent, equally distributed random variables. We denote $S_n = X_1 + X_2 + \cdots + X_n$.

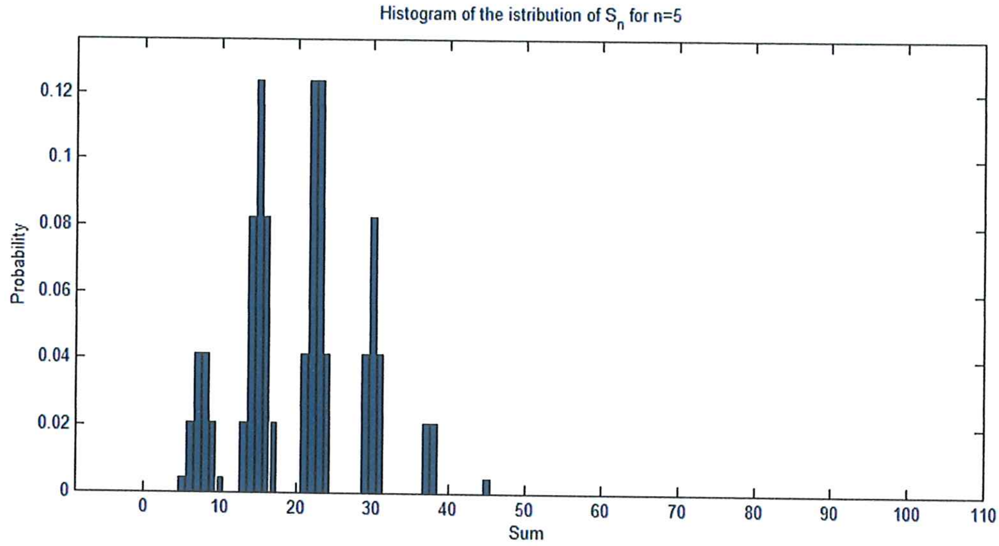
Let us look at examples of distributions of S_n for simple distributions.

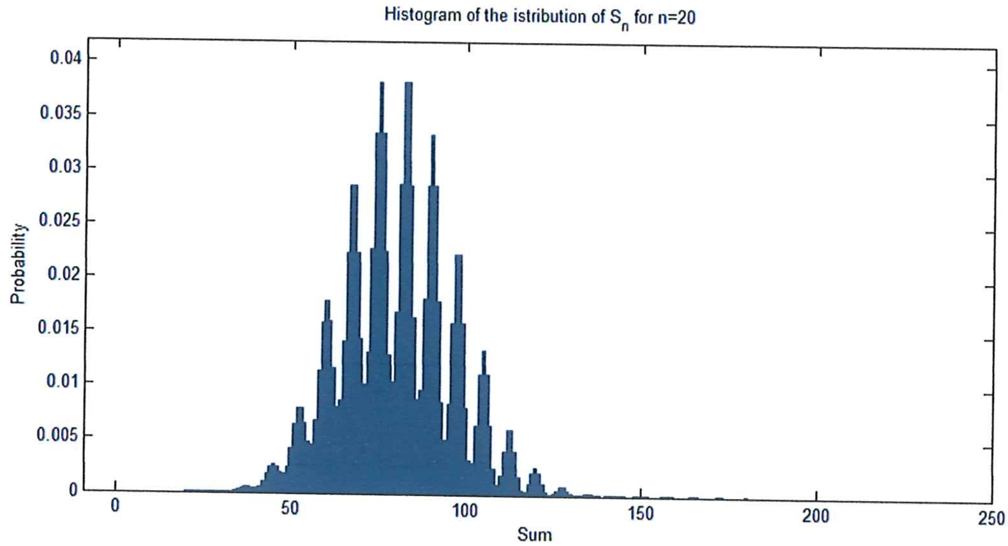
We will look at a few examples of distributions of S_n for different distributions of X_1 and different n .

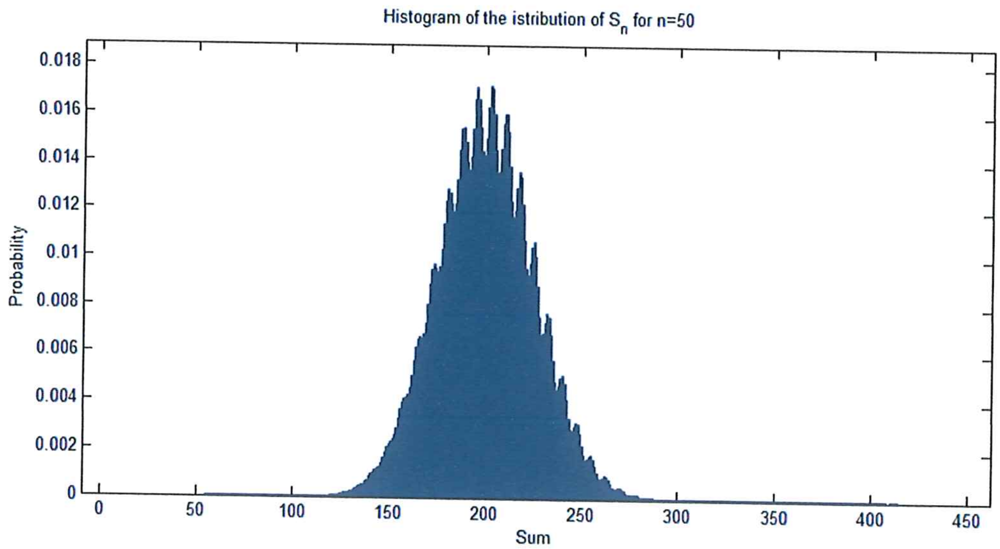
1. Let $P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$. Take $n = 100$. Let $S_n = X_1 + \cdots + X_n$. The histogram of the distribution of S_n is:

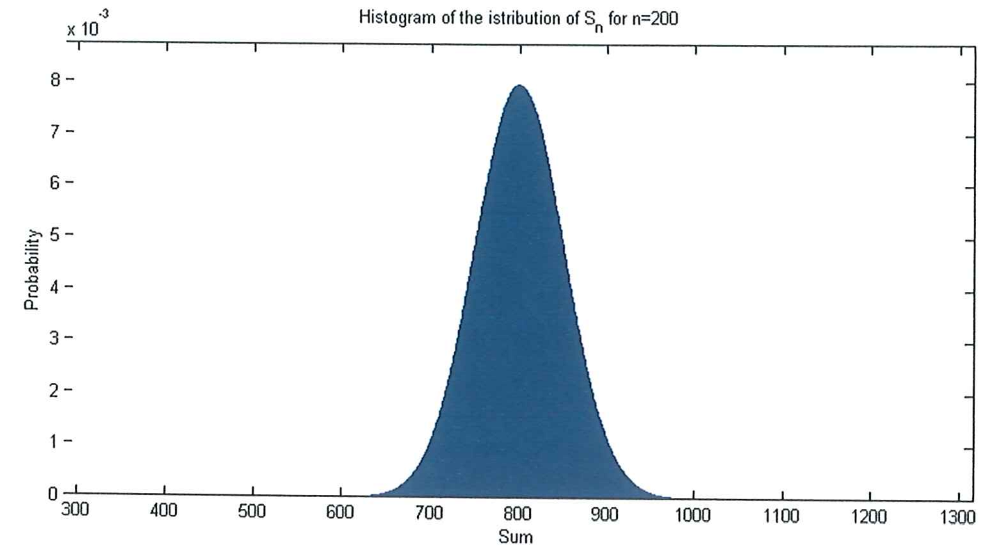


2. Let $P(X_1 = 1) = P(X_1 = 2) = P(X_1 = 9) = 1/3$. Let $n = 5,20,50,200$.

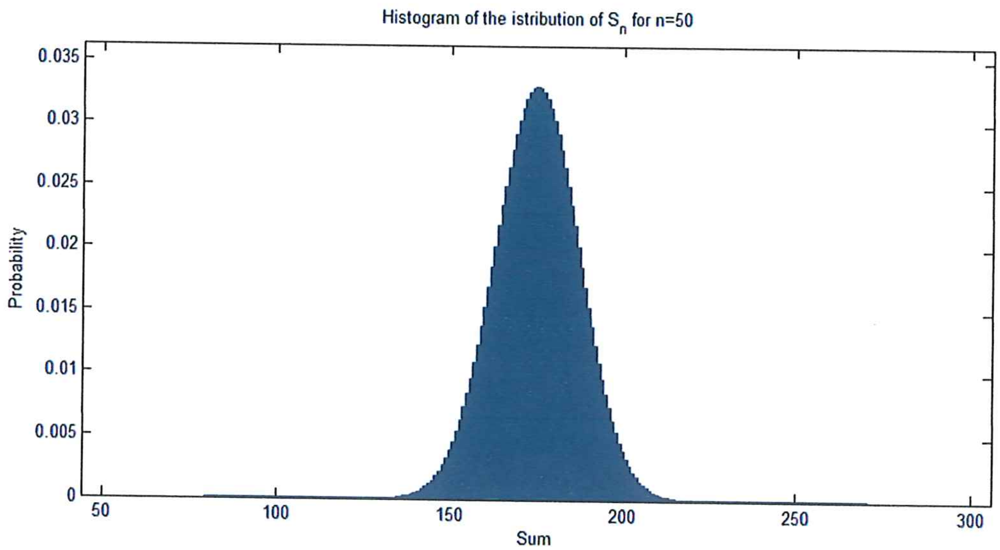

Histogram of the istribution of S_n for n=5
Probability
Sum
0
0.02
0.04
0.06
0.08
0.1
0.12
0
10
20
30
40
50
60
70
80
90
100
110


Histogram of the istribution of S_n for n=20
Probability
Sum
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0
50
100
150
200
250

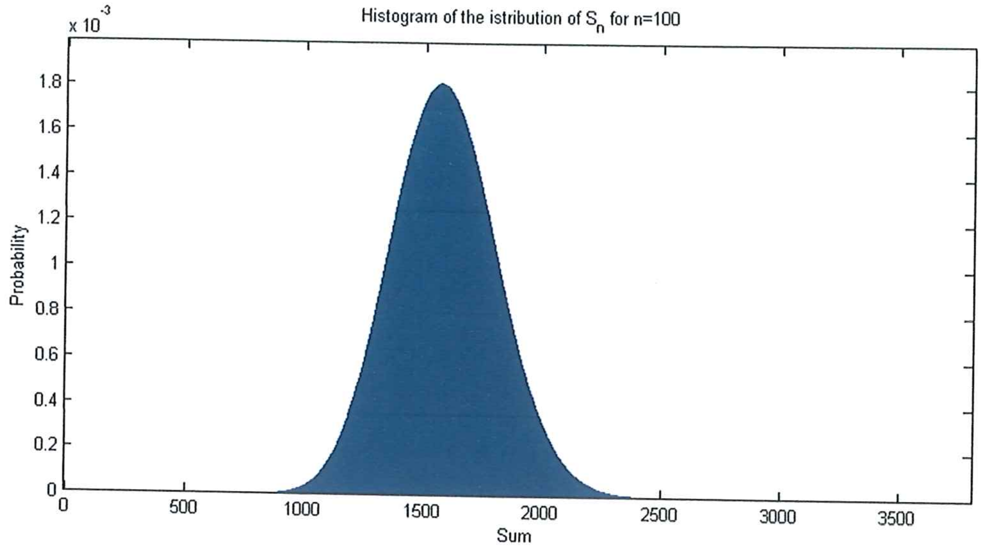

Histogram of the istribution of S_n for n=50
Probability
Sum
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
0
50
100
150
200
250
300
350
400
450



3. Take $P(X_1 = k) = {}^1\!/_6$ for $k = 1,2,\dots,6$. Take $n = 50$.



4. Take $P(X_1 = 2^k) = {}^1\!/_7$ for $k = 0,1,23,4,5,6$. Take $n = 100$.


Histogram of the istribution of S_n for n=100
x 10^{-3}
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Probability
0
500
1000
1500
2000
2500
3000
3500
Sum

From the examples we infer that the distribution of S_n is similar to the normal distribution. This is not clear in itself. But if we accept this observation we need to find the normal distribution that fits the histogram of S_n well.

Idea: We match the top of the two distributions which means that the mean of the normal distribution will be $E(S_n)$. We match the dispersion by choosing the second parameter to be $var(S_n)$.

Denote:

$$E(X_1) = \mu \qquad var(X_1) = \sigma^2$$

$$E(S_n) = n\, E(X_1) = \nu$$

$$var(S_n) = n\, var(X_1) = \tau^2.$$

To approximate the distribution we can say

$$P(a \leq S_n \leq b) \approx \frac{1}{\sqrt{2\pi}\cdot\tau} \int_a^b e^{-\frac{(x-\nu)^2}{2\tau^2}}\,dx .$$

The area of columns in the histogram between a and b ~~are~~ is exactly $P(a \leq S_n \leq b)$. We superimpose a curve closely following the histogram and replace the area of columns with the integral under the curve.

To turn the above into a matematical theorem we will reformulate.

Take $a = \nu + \alpha\cdot\tau$ and $b = \nu + \beta\cdot\tau$.

We compute

$$P(a \leq S_n \leq b)$$

$$= P(\nu + \alpha \cdot \tau \leq S_n \leq \nu + \beta \cdot \tau)$$

$$\approx \frac{1}{\sqrt{2\pi}\,\tau} \int_{\nu+\alpha\cdot\tau}^{\nu+\beta\cdot\tau} e^{-\frac{(x-\nu)^2}{2\tau^2}}\, dx$$

New variable: $\frac{x-\nu}{\tau} = u$

$$= \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-u^2/2}\, du.$$

On the other hand we have

$$P(\nu + \alpha \cdot \tau \leq S_n \leq \nu + \beta \cdot \tau)$$

$$= P\left(\alpha \leq \frac{S_n - \nu}{\tau} \leq \beta\right)$$

$$= P\left(\alpha \leq \frac{S_n - E(S_n)}{\sqrt{var(S_n)}} \leq \beta\right)$$

<u>Definition</u>: The expression

$$\tilde{S}_n = \frac{S_n - E(S_n)}{\sqrt{var(S_n)}}$$

is called the standardized sum.

We noticed that the approximation is „better" if n is „large". We expect the mathematical form to include limits.

<u>Theorem 6.1</u> (central limit theorem)

Let $X_1, X_2, \dots$ be independent equally distributed random variables with $E(X_1) = \mu$ and $var(X_1) = \sigma^2 < \infty$. Let $S_n = X_1 + \dots + X_n$. For any $\alpha < \beta$ we have

$$\lim_{n \to \infty} P\left(\alpha \le \frac{S_n - E(S_n)}{\sqrt{var(S_n)}} \le \beta\right)$$

$$= \Phi(\beta) - \Phi(\alpha)$$

where Φ is the distribution function of the standard normal distribution.

Comments:

(i) we will prove the theorem in several steps. It is true as it is formulated but we will impose the additional assumption $E(|X_1|^3) < \infty$.

(ii) It is enough to prove

$$\lim_{n \to \infty} P\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \beta \right) = \Phi(\beta)$$

for $\beta \in \mathbb{R}$.

(iii) In the limit we get equality. For finite n we use the limit as an approximation.

To prove the central limit theorem we need the following result.

<u>Theorem 6.2</u> (Lindeberg-Bergman)
Let $X_1, X_2, \dots, X_n$ be independent and such that $\mathrm{var}(X_1 + X_2 + \dots + X_n) = 1$ and $E(X_1 + \dots + X_n) = 0$. Assume that $E(|X_k|^3) < \infty$ for all $k = 1, 2, \dots, n$.

Let f be a three times continuously differentiable function such that $|f(x)|, |f'(x)|, |f''(x)|, |f'''(x)| \leq M$ for some $M < \infty$ and all $x \in \mathbb{R}$.

Let $S_n = X_1 + X_2 + \dots + X_n$. Then

$$|E(f(S_n)) - E(f(Z))|$$

$$\leq \frac{1}{6} M \left(1 + \sqrt{\frac{8}{\pi}}\right) E(|X_1|^3 + \dots + |X_n|^3).$$

for $Z \sim N(0,1)$.

Proof: Without loss of generality we can assume $E(X_k) = 0$ for all $k = 1, 2, \dots, n$. Let $Z_1, Z_2, \dots, Z_n$ be independent and independent of $X_1, X_2, \dots, X_n$ and such that

$Z_k \sim N(0, \mathrm{var}(X_k))$, $k = 1, 2, \dots, n$.

Since $\mathrm{var}(X_1) + \dots + \mathrm{var}(X_n) = 1$ by assumption we have that

$$Z = Z_1 + Z_2 + \dots + Z_n \sim N(0,1).$$

Define

$$a_1 = E[f(Z_1 + Z_2 + \dots + Z_n)] - E[f(X_1 + Z_2 + \dots + Z_n)]$$

$$a_2 = E[f(X_1 + Z_2 + \dots + Z_n)] - E[f(X_1 + X_2 + \dots + Z_n)]$$

$$\vdots$$

$$a_n = E[f(X_1 + \dots + X_{n-1} + Z_n)] - E[f(X_1 + X_2 + \dots + X_n)]$$

By triangle inequality we have

$$|E[f(X_1 + \quad + X_n)] - E[f(Z_1 + \quad + Z_n)]| \leq \sum_{k=1}^{n} |a_k|$$

By Taylor we have

$$f(x+h) - f(x) = f'(x)\cdot h + \frac{1}{2} f''(x) h^2 + r$$

where $r = \frac{1}{6} f'''(\xi) h^3$ for some ξ between x and $x+h$. By our assumption $|r| \leq \frac{1}{6} \cdot M \cdot |h|^3$.

Define

$$Y_1 = Z_2 + Z_3 + \dots + Z_n$$

$$Y_2 = X_1 + Z_3 + \dots + Z_n$$

$$Y_3 = X_1 + X_2 + Z_4 + \dots + Z_n$$

$$Y_n = X_1 + X_2 + \quad + X_{n-1}$$

Note that Y_k is independent of (X_k, Z_k) for all $k = 1, 2, \dots, n$.

We use Taylor's expansion around Y_k to get

$$E[f(X_1 + \quad + X_{k-1} + Z_k + \cdots + Z_n)]$$

$$= E[f(Y_k) + f'(Y_k) Z_k + \tfrac{1}{2} f''(Y_k) Z_k^2 + R_k]$$

and

$$E[f(X_1 + \cdots + X_k + Z_{k+1} + \cdots + Z_n)]$$

$$= E[f(Y_k) + f'(Y_k) X_k + \tfrac{1}{2} f''(Y_k) X_k^2 + \tilde{R}_k]$$

Subtracting we get

$$a_k = E[f'(Y_k)(Z_k - X_k) + \tfrac{1}{2} f''(Y_k)(Z_k^2 - X_k^2)$$
$$+ R_k - \tilde{R}_k]$$

$$= E[f'(Y_k)(Z_k - X_k)]$$
$$+ E[\tfrac{1}{2} f''(Y_k)(Z_k^2 - X_k^2)]$$
$$+ E[R_k - \tilde{R}_k]$$

By independence

$$E[f'(Y_k)(Z_k - X_k)]$$

$$= E[f'(Y_k)] \underbrace{E[Z_k - X_k]}_{= 0 \text{ by assumption}}$$

$$= 0$$

and

$$E[f''(Y_k)(Z_k^2 - X_k^2)]$$

$$= E[f''(Y_k)] \underbrace{E[Z_k^2 - X_k^2]}_{= 0 \text{ by assumption}}$$

$$= 0.$$

We are left with

$$a_k = E[R_k - \tilde{R}_k]$$

Since $\mp X \leq |X|$ we have $|E(X)| \leq E(|X|)$

so

$$|a_k| \leq E[|R_k - \tilde{R}_k|]$$

But

$$|R_k| \leq \frac{1}{6} M \cdot |X_k|^3$$

$$|\tilde{R}_k| \leq \frac{1}{6} M |Z_k|^3$$

so

$$|R_k - \tilde{R}_k| \leq \frac{1}{6} M (|X_k|^3 + |Z_k|^3).$$

It follows

$$E[|R_k - \tilde{R}_k|] \leq \frac{1}{6} \cdot M (E(|X_k|^3) + E(|Z_k|^3)).$$

A standard calculation gives that for $Z \sim N(0, \sigma^2)$ we have

$$E(|Z|^3) = \sqrt{\frac{8}{\pi}} \cdot \sigma^3$$

We then have

$$E(|Z_k|^3) = \sqrt{\frac{8}{\pi}} \, \mathrm{var}(X_k)^{3/2}$$

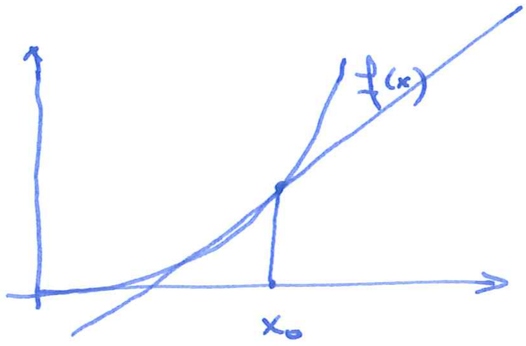
$$= \sqrt{\frac{8}{\pi}} \cdot E(X_k^2)^{3/2}.$$

$$= \sqrt{\frac{8}{\pi}} \, E(|X_k|^2)^{3/2}$$

Since $f(x) = x^{3/2}$ is convex,

the function is above its tangent.

Figure:



We have for $x_0 > 0$

$$f(x) = x^{3/2} \geq \underbrace{f'(x_0)(x - x_0) + f(x_0)}_{\text{tangent}}$$

Take $x_0 = E(|X_k|^2)$. We have

$$(|X_k|^2)^{3/2} \geq f'(x_0)\,(|X_k|^2 - x_0) + x_0^{3/2}$$

Taking expectations we get

$$E(|X_k|^3) \geq E(|X_k|^2)^{3/2}$$

$$= (\mathrm{var}(X_k))^{3/2}$$

Taking all inequalities we get

$$E(|X_k|^3) + E(|Z_k|^3)$$

$$\leq \left(1 + \sqrt{\frac{8}{\pi}}\right) E(|X_k|^3)$$

Finally,

$$\sum_{k=1}^{n} |a_k| \leq \sum_{k=1}^{n} \frac{1}{6} \cdot M \left(1 + \sqrt{\frac{8}{\pi}}\right) E(|X_k|^3) \quad \blacksquare$$

The inequality is valid for arbitrary $X_1, X_2, \ldots, X_n$ provided they are independent. If

$X_1, X_2, \ldots, X_n$ are independent and equally distributed define

$$X_k' = \frac{X_k - E(X_k)}{\sqrt{var(S_n)}}$$

for $S_n = X_1 + X_2 + \cdot + X_n$. Note that

$$X_1' + X_2' + \cdots + X_n' = \frac{S_n - E(S_n)}{\sqrt{var(S_n)}} = \tilde{S}_n$$

We have that $X_1', X_2', \dots, X_n'$ are independent, $E(X_k') = 0$ and $\mathrm{var}(X_1' + \cdot + X_n') = 1$.

Theorem 6.2 implies

$$|E[f(\tilde{S}_n)] - E[f(Z)]|$$

$$\leq \frac{1}{6} M\left(1 + \sqrt{\frac{8}{\pi}}\right) \cdot n \cdot E(|X_1'|^3)$$

But

$$E(|X_1'|^3) = E\left(\left|\frac{X_1 - E(X_1)}{\sqrt{n}\sqrt{\mathrm{var}(X_1)}}\right|^3\right)$$

$$= \frac{1}{n^{3/2}\,\mathrm{var}(X_1)^{3/2}} E(|X_1 - E(X_1)|^3)$$

If $\gamma = E(|X_1 - E(X_1)|^3) < \infty$ we have

$$|E[f(\tilde{S}_n)] - E[f(Z)]|$$

$$\leq \frac{1}{6} \cdot M\left(1 + \sqrt{\frac{8}{\pi}}\right) \cdot \frac{1}{\sqrt{n}} \cdot \frac{\gamma}{\mathrm{var}(X_1)^{3/2}}$$

$$\to 0, \quad \text{so } n \to \infty.$$

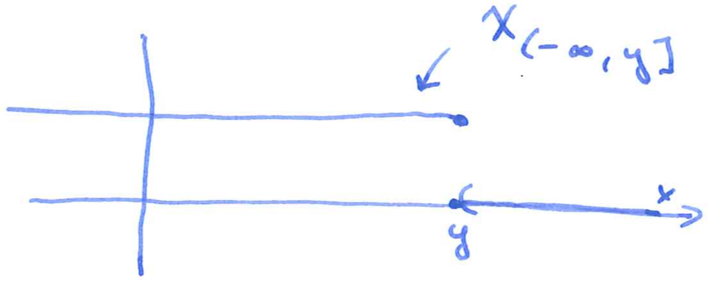
We will prove the central limit theorem under the additional assumption that $\gamma = E(|X_1 - E(X_1)|^3) < \infty$.

Comment : If $E(|X_1|^3) < \infty$ then $\gamma < \infty$.

Proof : Let $\varepsilon > 0$. The distribution function Φ of $Z \sim N(0,1)$ is continuous so for a fixed $\beta \in \mathbb{R}$ there is a $\delta > 0$ such that $|\Phi(x) - \Phi(\beta)| < \varepsilon$ for $|x - \beta| < \delta$.

Denote by $\chi_{(-\infty, y]}$ the indicator function of the interval $(-\infty, y]$.

Figure :



Analysis 1 gives: there are functions $f^{-\varepsilon}$ and f^{ε} with values on $[0,1]$ such that:

(i) $f^{-\varepsilon}, f^{\varepsilon}$ are three times continuously differentiable.

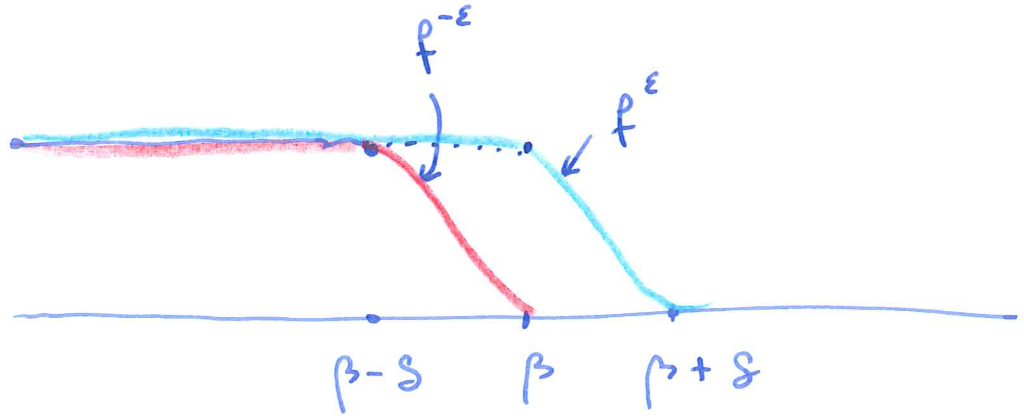
(ii) derivatives up to the third are bounded by $M < \infty$

(iii)

$$\chi_{(-\infty, \beta-\delta]} \leq f^{-\varepsilon} \leq \chi_{(-\infty, \beta]}$$

$$\leq f^{\varepsilon} \leq \chi_{(-\infty, \beta+\delta]}$$

Figure:



We have

$$E(\chi_{(-\infty,y]}(\tilde{S}_n))$$

$$= P(\tilde{S}_n \le y)$$

and similarly

$$E(\chi_{(-\infty,y]}(Z)) = P(Z \le y) = \Phi(y).$$

Inequalities in (iii) imply

$$\Phi(\beta-\delta) \le E(f^{-\varepsilon}(Z))$$

$$\le \Phi(\beta) \le E(f^{\varepsilon}(Z))$$

$$\le \Phi(\beta+\delta)$$

But $E(f^{-\varepsilon}(\tilde{S}_n)) \to E(f^{-\varepsilon}(Z))$

$E(f^{\varepsilon}(\tilde{S}_n)) \to E(f^{\varepsilon}(Z))$

as $n \to \infty$.

For sufficiently large n we will have

$$\Phi(\beta-\delta) - \varepsilon \leq P(\tilde{S}_n \leq \beta) \leq \Phi(\beta+\delta) + \varepsilon.$$

or

$$\Phi(\beta) - 2\varepsilon \leq P(\tilde{S}_n \leq \beta) \leq \Phi(\beta) + 2\varepsilon.$$

This proves that

$$\lim_{n\to\infty} P(\tilde{S}_n \leq \beta) = \Phi(\beta).$$

Examples: Typically we want to estimate probabilities of the form

$$P(a \leq S_n \leq b).$$

we compute

$$P(a \leq S_n \leq b)$$

$$= P\big(a - E(S_n) \leq S_n - E(S_n) \leq b - E(S_n)\big)$$

$$= P\left(\underbrace{\frac{a - E(S_n)}{\sqrt{var(S_n)}}}_{\alpha} \leq \frac{S_n - E(S_n)}{\sqrt{var(S_n)}} \leq \underbrace{\frac{b - E(S_n)}{\sqrt{var(S_n)}}}_{\beta} \right)$$

$$= P\left(\alpha \leq \frac{S_n - E(S_n)}{\sqrt{var(S_n)}} \leq \beta \right)$$

$$\overset{CLT}{\approx} \Phi(\beta) - \Phi(\alpha)$$

(i) Let $X_1, X_2, \ldots$ be independent and $X_k \sim$ Bernoulli(p). We know that $S_n = X_1 + X_2 + \ldots + X_n \sim Bin(n, p)$ so $E(S_n) = n \cdot p$ and $var(S_n) = npq$.

Assume $n = 10{,}000$ and $p = 1/2$ and $a = 4950$ and $b = 5050$. We have

$$P(4950 \leq S_n \leq 5050)$$

$$= P\left(\frac{4950 - 5000}{50} \leq \frac{S_n - E(S_n)}{\sqrt{var(S_n)}} \leq \frac{5050 - 5000}{50} \right)$$

$$\approx P(-1 \leq Z \leq 1)$$

Statistical software gives

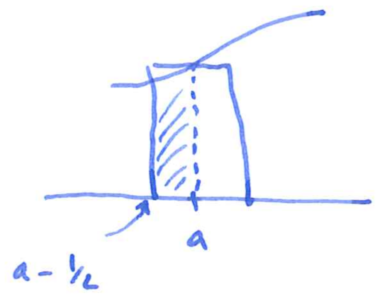
$$P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1)$$

$$\doteq 0.6827$$

The exact probability is 0.6875.

If $X_1, X_2, \ldots$ are integer valued we can improve the approximation by changing a to $a - 1/2$ and b to $b + 1/2$.

Figure:



Changing a to $a - 1/2$ adds the "half" of the column over a. This correction is called correction for continuity.

Using this correction we get

$$P(4950 \le S_n \le 5500)$$

$$\approx \Phi(1.0100) - \Phi(-1,0100)$$

$$\doteq 0.6875 .$$

This is accurate to 4 decimals!

(ii) Let $X_1, X_2, \ldots$ be independent

and $P(X_1 = 1) = P(X_1 = 2) = P(X_1 = 9) = 1/3$

Let $n = 300$. We find

$$E(S_{300}) = 300 \cdot 4 = 1200$$

$$var(S_{300}) = 300 \cdot \frac{38}{3} = 3800$$

We approximate

$$P(1,100 \le S_{300} \le 1,300)$$

$$= P\left(\frac{1100-1200}{\sqrt{3800}} \le \frac{S_{300} - E(S_{300})}{\sqrt{3800}} \le \frac{1300-1200}{\sqrt{3800}}\right)$$

$$\approx \Phi(1.622) - \Phi(-1.622)$$

$$\doteq 0.8952$$

The exact probability using the fast Fourier transform turns out to be 0.8970. If we include the continuity correction we get 0.8970!

Can we say anything about the accuracy approximation? The answer is yes but the proof is demanding.

<u>Theorem 6.3</u> (Berry-Esséen) Let $\gamma = E(|X_1 - E(X_1)|^3)$ and keep all the assumptions of Theorem 6.1. Then

$$\sup_{x \in \mathbb{R}} \left| P\left(\frac{S_n - E(S_n)}{\sqrt{\operatorname{var}(S_n)}} \leq x \right) - \Phi(x) \right|$$

$$\leq \frac{C \cdot \gamma}{\sqrt{n} \operatorname{var}(X_1)^{3/2}}$$

where $C < 0.4748$.

For a proof see

Shevtsova, I., On the accuracy of the normal approximation for sums of independent symmetric random variables, Dokl. Akad. Nauk 443 (2012), no. 6, 671–676.

THE END