Statistics and Probability - Probability Part

Roula Tsonaka, PhD

s.tsonaka@lumc.nl

Department of Biomedical Data Sciences, Medical Statistics Section

Leiden University Medical Center, The Netherlands

Lecture 1: Sample spaces

Outline

1. Introduction

2. Sample spaces

3. Algebra of set theory

Outline

1 Introduction

2 Sample spaces

3 Algebra of set theory

Probability, chance, randomness

- The science of *Statistics* studies **random phenoma**.
- The outcome of random phenoma is not known beforehand.
- We will study the basics of random phenoma: how to describe them in a mathematical manner and understand them.
- Probability theory plays a central role in understanding random phenoma.

Probability, chance, randomness

- ***Genetics:*** model for mutations to explain variability in the population.
- ***Medicine - decision making***: choice of treatment (expensive vs cheap).
- ***Medicine:*** interpretation of test outcome, e.g., Down's syndrome screening test, mammography.
- ***Actuarial Science:*** development of insurance product depends on tools from probability theory.
- ***Common sense:*** e.g., birthday problem.

Outline

1 Introduction

2 Sample spaces

3 Algebra of set theory

Experiments and sample spaces

- **Experiments** are situations for which the outcomes occur randomly.
- **Sample space** (Ω) is the set of all possible outcomes (ω).
- $\boxed{\text{Example 1:}}$
  - Experiment: Driving to work a commuter passes a traffic light.
  - Sample space: commuter stops (s) or continues at traffic light (c), $\Omega = \{c, s\}$.
- $\boxed{\text{Example 2:}}$
  - Experiment: commuter passes three traffic lights.
  - Sample space: $\Omega = \{ccc, scc, csc, ccs, ssc, scs, css, sss\}$
  - Note that order may matter here.

Experiments and sample spaces - Cont'd

- Example 3:
  - Experiment: The number of print jobs of a mainframe computer.
  - Sample space: $\Omega = \{0, 1, 2, 3, \ldots\}$.
- Example 4:
  - Experiment: The length of time between successive earthquakes that are greater in magnitude than a certain threshold.
  - Sample space: $\Omega = \{t \mid t \geq 0\}$.

Events

- **Events** are subsets of the sample space.
- $\boxed{\text{Example:}}$
  - In Example 2: a commuter stops at first traffic light.
  - Sample space: $\Omega = \{ccc, scc, csc, ccs, ssc, scs, css, sss\}$.
  - $A = \{sss, ssc, scc, scs\}$.
- Events or subsets are often denoted by italic uppercase letters.
- Notation: A is subset of Ω: $A \subset \Omega$.

Outline

1 Introduction

2 Sample spaces

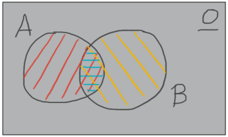
3 Algebra of set theory

Algebra of set theory

- **Union of two events A and B:** $C = A \cup B$
  C is event that either A occurs or B occurs or both occur.
- **Intersection of two events A and B:** $C = A \cap B$
  C is event that both A and B occur.
- **Complement of event A:** $C = A^c$
  C is the event that A does not occur.
- **Empty set:** $\varnothing$, is set without elements.
- **Disjoint events:** A and B are disjoint when $A \cap B = \varnothing$.
- **Partition of Ω:** is a collection of disjoint events $\{A_1, A_2, \ldots, A_n\}$ with $\cup_{i=1}^{n} = \Omega$.

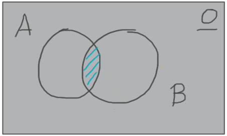
Algebra of set theory - Illustration I

- **Union of two events A and B:**



$A \cup B$

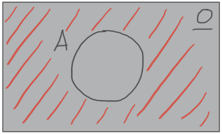
- **Intersection of two events A and B:**



$A \cap B$

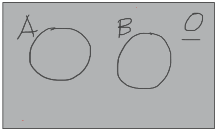
Algebra of set theory - Illustration II

- **Complement of event A:**



A^c

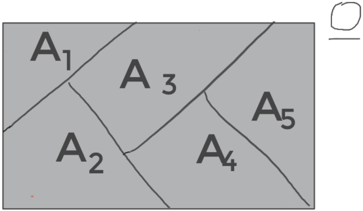
- **Disjoint events:**



$A \cap B = \emptyset$

Algebra of set theory - Illustration II

- **Partition of Ω:**



$$\bigcup_{i=1}^{5} = \Omega$$

Set operations

- **Commutative Laws**: Order of sets does not matter.
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$

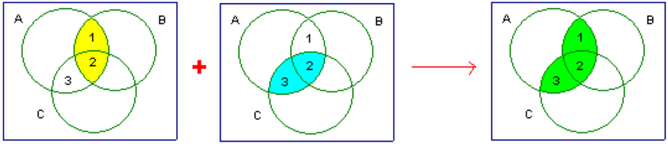
- **Associative Laws**: How to perform multiple unions and intersections.
  - $(A \cup B) \cup C = A \cup (B \cup C)$
  - $(A \cap B) \cap C = A \cap (B \cap C)$

- **Distributive Laws**: How to mix unions and intersections.
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  - $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$

- **De Morgan's Law**: How to compliment unions and intersections.
  - $(A \cap B)^C = A^C \cup B^C$
  - $(A \cup B)^C = A^C \cap B^C$

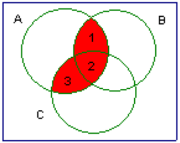
- Venn diagrams illustrate these laws and are very useful when you have to make more complex calculations.

Illustration of Distributive Law



$A \cap B$ + $A \cap C$ → $(A \cap B) \cup (A \cap C)$

Which is the same as



$A \cap (B \cup C)$

The distributive law in sets: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Lecture 2: Probability

Outline

1. Probability measure

2. Computing probabilities: Counting Methods

Outline

1. Probability measure

2. Computing probabilities: Counting Methods

Probability measure

- A **probability measure** on Ω is a function from subsets of Ω to real numbers with the following properties:

  1. $P(\Omega) = 1$.
  2. If $A \subset \Omega$ then $P(A) \geq 0$.
  3. If A and B are disjoint then $P(A \cup B) = P(A) + P(B)$.

     $\Rightarrow$ This can be generalized to many disjoint subsets.

Probability measure

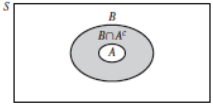
- $\boxed{\text{Example:}}$ A dice is loaded (i.e., not all outcomes are equally likely) such that the probability that the number i shows up is $K \times i$, $i = 1, 2, \ldots, 6$ where K is a constant.
  - What is the value of K?
  - What is the probability that a number greater than 3 shows up?

$\Rightarrow$ $K = 1/21$

$\Rightarrow$ $P(A) = 15/21$

Probability measure - Properties

- $P(A^c) = 1 - P(A)$.
- $P(\varnothing) = 0$.
- If $A \subset B$ then $P(A) \leq P(B)$.



- **Addition Law:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probability measure - Properties

- $\boxed{\text{Example:}}$ A recent survey shows that 40% of the first year students were in the top 10% of their high school class, 65% are female, 25% are both female and belonged to the top 10% of their high school class.

  What is the probability that a student selected randomly from this class either was on the top 10% of his/her high school class or is female or both?

⇒ *Answer:* The probability of this event is 0.80.

Outline

1 Probability measure

2 Computing probabilities: Counting Methods

Computing probabilities

- For finite sample spaces if we know the probabilities of the outcomes of sample space Ω then we can compute the probability of a subset (event) by adding the probabilities of the outcomes of the subset.

- $\boxed{\text{Example:}}$ A woman has given birth to two babies.
  - The sample space for the sex of the babies is: $\Omega = \{bg, gb, gg, bb\}$.
  - Each outcome has probability 0.25.
  - Let A the event of at least one girl: $A = \{bg, gb, gg\} \Rightarrow P(A) = 0.75$.

Computing probabilities- Cont'd

- If all outcomes have equal probabilities:

$$P(A) = \frac{\text{\# ways A occurs}}{\text{total \# of outcomes}}$$

- If we record number of girls delivered by the woman then $\Omega = \{0,1,2\}$ and $P(A) \neq \frac{2}{3}$ because the outcomes in Ω do not have equal probabilities.

Probability measure - Properties

- $\boxed{\text{Example:}}$ Suppose that a fair coin is thrown twice. Let A denote the event of heads on the first toss, and let B denote the event of heads on the second toss. The sample space is: $\Omega = \{hh, ht, th, tt\}$. We assume that each elementary outcome in Ω is equally likely and has probability 1/4.
  What is the probability of the event C that heads comes up on the first toss or on the second toss?

$\Rightarrow$ *Answer:* The probability of this event is 0.75.

Simpson's paradox - Example

- University of California, Berkeley was sued for bias against women who had applied for admission to graduate schools:

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

- But when examining the individual departments:

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Women tended to apply to competitive departments with low rates of admission, whereas men tended to apply to less-competitive departments with high rates of admission.

Computing probabilities

- Listing the elements of sample spaces/events not always possible.
- For large sample spaces, to count number of elements we use alternative methods.
- We distinguish sampling:
  - Without replacement.
  - With replacement.
- Order matters or not:
  - Permutations (ordered samples).
  - Combinations (unordered samples).

Computing probabilities

- Permutations with replacement.
- Permutations without replacement.
- Combinations without replacement.
- Combinations with replacement.

Counting methods - The multiplication principle

- If one experiment has n possible outcomes and another experiment had m possible outcomes, then there are $m \times n$ possible outcomes/configurations for the two experiments.
- $\boxed{\text{Example:}}$ You have 5 T-shirts and 2 pairs of trousers. You can make $5 \times 2 = 10$ outfits.
- This can be generalized to multiple experiments.
- $\boxed{\text{Example:}}$ Tossing 6 coins. There $2^6 = 64$ outcomes.
- $\boxed{\text{Example:}}$ A standard lock has a dial with tick marks for 10 numbers from 0 to 9. A sequence of three numbers must be dialed in the correct order to open the lock. Each of the 10 numbers may appear in each of the three positions regardless of what the other two positions contain. How many possible triplets are there?

Counting methods - Permutations

- Suppose we have a set of 10 numbers $C = \{0, 1, \cdots, 9\}$ and we want to select 3 and out them in an order. How many triplets are there?

- Depends on:
  - **Sampling with replacement:** duplications are allowed. Then the answer is 10^3.
  - **Sampling without replacement:** no duplications are allowed. Then the answer is $10 \times 9 \times 8$.

Counting methods - Permutations

- A **permutation** is an ordered arrangement of outcomes.
- The number of orderings of n different items is $n! = n \times (n-1) \times \ldots \times 1$.
- The number of orderings of k items selected **without replacement** from n different items is $\frac{n!}{(n-k)!} = n \times (n-1) \times \ldots \times (n-k+1)$.
- The number of orderings of k items selected **with replacement** from n different items is n^k.

Note: The symbol $n!$ (is read as “n factorial") is defined as $n! = n \cdot (n-1) \cdot \ldots \cdot (2) \cdot (1)$. Also: $1! = 1$ and $0! = 1$.

Counting methods - Permutations

- $\boxed{\text{Example:}}$ In how many ways can five children be lined up?

- $\boxed{\text{Example:}}$ Suppose that from ten children, five are to be chosen and lined up. How many different lines are possible?

Counting methods - Permutations

- Example: In US, in some states, license plates have six characters: three letters followed by three numbers. How many distinct such plates are possible?

- Example: If all sequences of six characters are equally likely, what is the probability that the license plate for a new car will contain no duplicate letters or numbers?

Counting methods - Combinations

- Instead of counting permutations (all possible orderings of outcomes), we count subsets i.e. **combinations** (unordered).
- **In combinations order is not important**.
- $\boxed{\text{Example:}}$ Consider a set $\{a,b,c,d\}$:
  - The number of subsets of size 2 is: $\{a,b\}$, $\{a,c\}$, $\{a,d\}$, $\{b,c\}$, $\{b,d\}$ and $\{c,d\}$.
  - $\{a,b\}$ and {b, a} in combinations is the same subset but not in permutations.

Counting methods - Combinations

- The number of unordered samples of size k that can be chosen from a set of n objects without replacement is
$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}.$$
This is known as the Binomial coefficient.

- The number of unordered samples of size k that can be chosen from a set of n objects with replacement is $\frac{(n-1+k)!}{(n-1)!k!}$.

Counting methods - Combinations

- $\boxed{\text{Example:}}$ Up until 1991, a player of the California state lottery could win the jackpot prize by correctly choosing the 6 numbers from 1 to 49 that the lottery officials randomly picked. What is the probability of winning?

Counting methods

- Sampling k from n objects:

	ordered sampling	unordered sampling
replacement	n^k	$\frac{(n-1+k)!}{(n-1)!k!}$
no replacement	$\frac{n!}{(n-k)!}$	$\frac{n!}{k!(n-k)!}$

Lecture 3: Conditional Probability

Outline

1. Conditional probabilities

2. Bayes' Theorem

3. Law of total probability

4. Independence

Outline

1. Conditional probabilities

2. Bayes' Theorem

3. Law of total probability

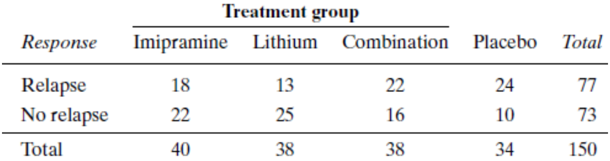
4. Independence

Example I

- A woman wants to know her chance of developing breast cancer.
- Several studies have shown that the chance of getting breast cancer among women with two first degree affected relatives is three times as large compared to women without a positive family history.
- It makes sense to compute conditional probabilities, i.e., the probability of getting breast cancer given your family history.
- Two experiments with two sample spaces:
    - Family history with outcomes yes and no.
    - Breast cancer with outcomes yes and no.
- We can also consider the two experiments together: four outcomes (see Lecture 2).

Example II

- 150 patients with episodes of depression randomized to 4 treatment groups: Placebo and treatments I, L and C.
- Relapse may happen within two to three years.
- For a patient in placebo what is the conditional probability that the patient had a relapse?



	Treatment group				
Response	Imipramine	Lithium	Combination	Placebo	*Total*
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

Figure: Results of the clinical depression study.

Example II - Cont'd

- A the event that the patient had a relapse.
- B the event that the patient received the placebo.
- From the Table $P(B) = 34/150$, $P(A \bigcap B) = 24/150$ and $P(A \mid B) = 24/34 = 0.706$.
- For a patient in lithium $P(C) = 38/150$, $P(A \bigcap C) = 13/150$, and $P(A \mid C) = 13/38 = 0.342$.
- **Knowing which treatment a patient received seems to make a difference in the probability of relapse.**

Definition

- **Conditional Probability:** Let A and B be two events with probability $P(B) \neq 0$. The probability of A given B is $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

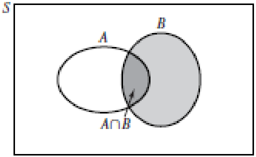
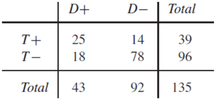


Figure: The outcomes in the event B that also belong to the event A.

Applications

- Digitalis Example:



	D+	*D*−	*Total*
T+	25	14	39
T−	18	78	96
Total	43	92	135

Figure: T^+, T^- test positive or negative and D^+, D^- toxicity present or not.

- $P(T+|D+)$ = sensitivity and $P(T-|D-)$ = specificity.

- **Multiplication Law:** Let A and B be two events with probability $P(B) \neq 0$. Then $P(A \cap B) = P(A|B)P(B)$.

Example

- An urn contains three red balls and one blue ball. Two balls are selected without replacement. What is the probability that they are both red?

Outline

1 Conditional probabilities

2 Bayes' Theorem

3 Law of total probability

4 Independence

Bayes' Theorem

- **The Bayes' theorem gives us the a way to reverse the conditional probability of events** A and B because $\boxed{\text{P(A}|\ B) \neq P(B \mid A)}$.

- For any two events A and B with $P(A) > 0$ and $P(B) > 0$ $\boxed{P(B \mid A) = \frac{P(A|B)P(B)}{P(A)}.}$

Outline

1. Conditional probabilities
2. Bayes’ Theorem
3. Law of total probability
4. Independence

Law of total probability

- This law allows to **compute the probability of an event by decomposing it in smaller events**.

- Let $\{B_1, \ldots, B_n\}$ a partition of Ω, i.e., $B_1, \ldots, B_n$ are disjoint events and $\bigcup_{i=1}^n B_i = \Omega$. Then for any event A,

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_n)P(B_n).$$

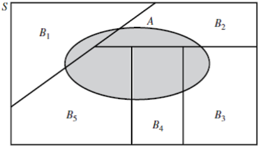


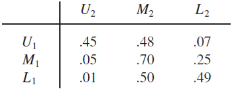
Figure: The intersection of A with events $B_1, B_2, \ldots$.

Law of total probability - Examples

- An urn contains three red balls and one blue ball. Two balls are selected without replacement. What is the probability that a red ball is selected on the second draw?

Law of total probability - Examples

- Suppose that occupations are grouped into upper (U), middle (M), and lower (L) levels. U_1 denotes the event that a father's occupation is upper-level; U_2 denotes the event that a son's occupation is upper-level, etc. Suppose also that of the father's generation, 10% are in U, 40% in M, and 50% in L.



	U_2	M_2	L_2
U_1	.45	.48	.07
M_1	.05	.70	.25
L_1	.01	.50	.49

Figure: Table with conditional probabilities: $P(U_2 \mid U_1), P(M_2 \mid U_1), \ldots$.

What is the probability that a son in the next generation is in U?

Bayes' Theorem Revisited

- Let A and the partition $\{B_1, \ldots, B_n\}$, then using the law of total probability:

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

- The Bayes' rule gives us the a way to reverse the conditional probability of events A and B because $P(A \mid B) \neq P(B \mid A)$.

Bayes' Theorem - Example

- We consider again the occupational mobility example where we now ask a different question: If a son has occupational status U_2, what is the probability that his father had occupational status U_1?

Bayes' Theorem - Terminology

- $P(B_i)$ is typically called **prior probability**.
- Joint probabilities: $P(A \cap B_i) = P(A|B_i)P(B_i)$.
- The conditional probability $P(B_i|A)$ is also called **posterior probability.**

To compute posterior probabilities you may use the table:

Events	B_1	$\cdots$	B_n
Prior probabilities	$P(B_1)$	$\cdots$	$P(B_n)$
Conditional probabilities	$P(A|B_1)$	$\cdots$	$P(A|B_n)$
Joint probabilities	$P(A|B_1)P(B_1)$	$\cdots$	$P(A|B_n)P(B_n)$
Posterior probabilities	$P(B_1|A)$	$\cdots$	$P(B_n|A)$

What is sum of the prior probabilities in line 1?
What is sum of the joint probabilities in line 3?

Outline

1. Conditional probabilities

2. Bayes' Theorem

3. Law of total probability

4. Independence

Independence

- $\boxed{\text{Definition:}}$ A and B are said to be **independent** events if $P(A \cap B) = P(A) \times P(B)$.

- Note that under independence $P(A \mid B) = P(A)$.

  ⇒ the probability of A does not change even after we learn that B has occurred.

Independence - Example

- A card is selected randomly from a deck. Let A denote the event that it is an ace and D the event that it is a diamond. Are the events A and D independent?

Mutually independence

- We consider multiple events.
- Definition: **Pairwise independence** means that any two events are independent.
- Definition: **Mutually independence** means that for any subcollection i the independence rule holds:

$$P(A_{i1} \cap \ldots \cap A_{in}) = P(A_{i1}) \times \ldots \times P(A_{in}).$$

- **Pairwise independence does not guarantee mutual independence.**

Mutually independence - Example

- A fair coin is tossed twice. Let A denote the event of heads on the first toss, B the event of heads on the second toss, and C the event that exactly one head is thrown. A and B are clearly independent.

  Are the events A, B and C mutually independent?

Lecture 4: Discrete random variables

Outline

1. Discrete random variables

2. Bernoulli random variable

3. Binomial Distribution

4. Geometric Distribution

5. Poisson Distribution

Outline

1. Discrete random variables

2. Bernoulli random variable

3. Binomial Distribution

4. Geometric Distribution

5. Poisson Distribution

Random variables - Examples

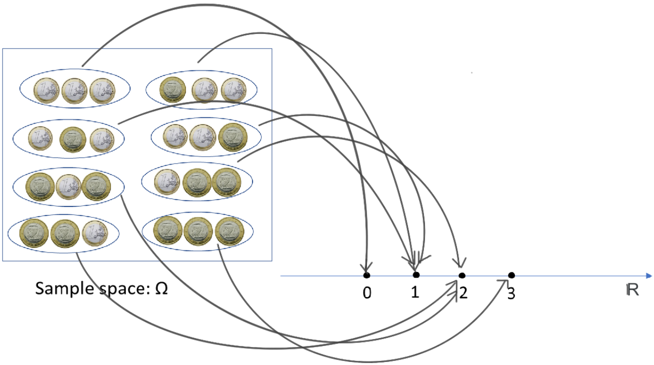
- $\boxed{\text{Example 1:}}$ A coin is thrown two times. Sample space: $\Omega = \{hh, ht, th, tt\}$.
  Examples of random variables defined on Ω:
  - the number of heads X with support $S_X = \{0, 1, 2\}$.
  - the number of tails Y with support $S_Y = \{0, 1, 2\}$.
  - the number of heads minus the number of tails $Z = X - Y$, $S_Z = \{-2, -1, 0, 1, 2\}$.
- $\boxed{\text{Example 2:}}$ The coin is flipped until a Head pops-up. Let W the random variable for the number of Tails before the first Head.
  - Which is the sample space of this experiment?
  - Which is the support of W?

Random variable

- Definition: A **random variable** X is a function which is defined on a sample space Ω and takes values in the real line $\mathbb{R}, X : \Omega \to \mathbb{R}$.
- Random variables are denoted by uppercase letters e.g. X, Y and Z and their observed values by lowercase letters x, y and z.
- The set of values that X takes, forms its support S_X.
- In the following, we will often use the abbreviation rv.
- A random variable X that can only take a finite number k of values $x_1, \ldots, x_k$ or at most a countably infinite number of values $x_1, x_2, \ldots$ is called a **discrete** random variable.

Random variable - Illustration

- A coin is flipped three times: Let the random variable X which counts the number of heads.



- A random variable is a mapping from the outcomes in Ω to the numbers in $\mathbb{R}$, e.g., $X(HHH) = 3$.
- The function X is called random because it takes its values with a certain probability.

Frequency function

- $\boxed{\text{Example:}}$ The probabilities of the number of heads in 3 tosses of a coin (if the coin is fair!) are:
  - $P(X=0)=P(TTT)=\frac{1}{8}$
  - $P(X=1)=P(HTT)+P(THT)+P(TTH)=\frac{3}{8}$
  - $P(X=2)=P(HHT)+P(HTH)+P(THH)=\frac{3}{8}$
  - $P(X=3)=P(HHH)=\frac{1}{8}$

  Note $\sum_{x\in\{0,1,2,3\}} P(X=x)=1$.

- $\boxed{\text{Definition:}}$ Let $x_1, x_2, \ldots$ values of X. The **frequency function** (ff) or **probability mass function** (pmf) of X is a function which specifies the probability of obtaining a number x_i and is denoted by:

$$p(x_i)=p_X(x_i)=P(X=x_i) \text{ with } \sum_i p_X(x_i)=1.$$

Cumulative distribution function

- $\boxed{\text{General definition:}}$ Let $x_1, x_2, \ldots$ values of X. The **cumulative distribution function** (cdf) of a random variable X is defined as follows:

$$F_X(x_k) = P(X \leq x_k) = \sum_{i=1}^{k} p_X(x_i).$$

- Note that the cdf F_X has the following properties:
  - F_X is non-decreasing.
  - The minimum of F_X is 0.
  - The maximum of F_X is 1.

- Capital letters are used for the cdf e.g. $F_X(x)$ or $F(x)$ and lower case for the ff e.g. $p_X(x)$ or $p(x)$.

Frequency and cumulative distribution function

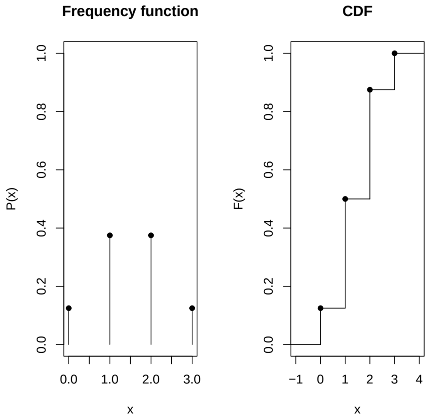


Figure: The probability mass function for coin experiment and its cumulative distribution function.

- The cdf is 0 for any value smaller than the first value that X takes.
- The cdf is 1 for any value larger than the last value that X takes.

Cumulative distribution function

- $\boxed{\text{Example:}}$ The cdf for the 3 coin tosses example

$$F_X(x) = P(X \le x) = \begin{cases} 0, \ \ x < 0 \\ \frac{1}{8}, \ \ x \in [0,1) \\ \frac{1}{8} + \frac{3}{8} = \frac{4}{8}, \ \ x \in [1,2) \\ \frac{4}{8} + \frac{3}{8} = \frac{7}{8}, \ \ x \in [2,3) \\ 1, \ \ x \ge 3. \end{cases}$$

Expected value of a discrete random variable

- General definition: If X is a discrete random variable with frequency function $p_X(x)$, the **expected value** of X denoted by $E(X)$ is

$$E(X) = \sum_i x_i \cdot p_X(x_i).$$

- $E(X)$ is also referred to as the **mean** of X and is often denoted by μ or μ_X.

Expected value of a discrete random variable

- $\boxed{\text{Example:}}$ Let X denote the number of heads in 3 tosses of a fair coin with $p(X)$:
    - $P(X=0)=P(TTT)=\frac{1}{8}$
    - $P(X=1)=P(HTT)+P(THT)+P(TTH)=\frac{3}{8}$
    - $P(X=2)=P(HHT)+P(HTH)+P(THH)=\frac{3}{8}$
    - $P(X=3)=P(HHH)=\frac{1}{8}$

- The mean of X is $\mu=\sum_{x=0}^{3} x\cdot p_X(x)=0\cdot\frac{1}{8}+1\cdot\frac{3}{8}+2\cdot\frac{3}{8}+3\cdot\frac{1}{8}=1.5$

- $\boxed{\text{Interpretation:}}$
  If we repeat the same experiment many times and record number of heads each time, the sample mean will be close to 1.5.

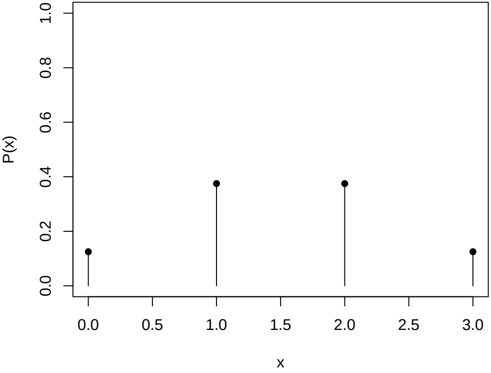
- $E(X)$ is a measure of location.

Expected value of a discrete random variable

- Illustration

```
x <- sample(0:3, size = 100,
            prob = c(1/8, 3/8, 3/8, 1/8), replace = TRUE)
mean(x)
[1] 1.52
```

- Distribution of number of heads in 3 tosses



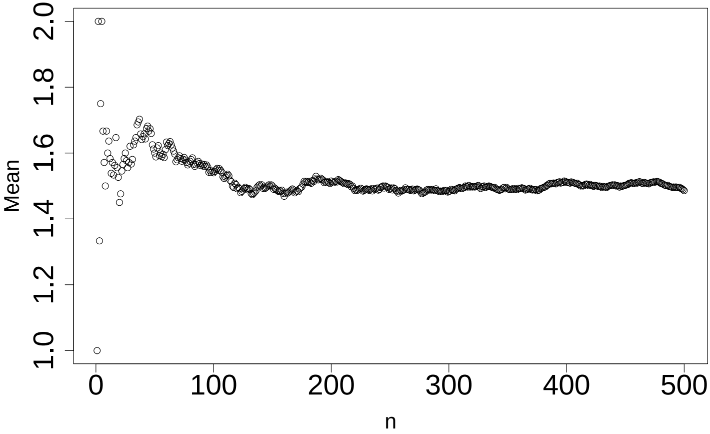
Expected value of a discrete random variable

- Illustration - Sample mean as repetitions of experiment increase

```
s <- 0
out <- numeric()
N <- seq(1, 500, 1)
for (n in N) {
     x <- sample(0:3, size = 1,
            prob = c(1/8, 3/8, 3/8, 1/8), replace = TRUE)
     s <- s + x
     out <- rbind(out, s/n)
     plot(N[1:n], out, xlab = "n", ylab = "Mean")
 }
```

Expected value of a discrete random variable

- Mean number of heads in 3 tosses versus repetitions of experiment



Properties of expectation of a random variable

- Let the random variable X with frequency function $p_X(x)$ and the random variable $Y = g(X)$.
- We know that $E(Y) = \sum_y y \cdot h_Y(y)$, with $h_Y(y)$ the frequency function of Y. This means that we need to derive first $h_Y(y)$. Luckily it holds: $\boxed{E(Y) = \sum_x g(x) \cdot p_X(x).}$
- **No need to derive the form of $h_Y(y)$!**
- Note that $E[g(X)] \neq g[E(X)]$.
- $\boxed{\text{Example:}}$ Let X take on values 1 and 2, each with probability 1/2, then $E(X) = 3/2$. Let $Y = 1/X$ then $E(Y) = 0.75 \neq 1/E(X) = 2/3$.

Properties of expectation of a random variable

- $E(c \cdot X) = c \cdot E(X)$, where c is a constant.

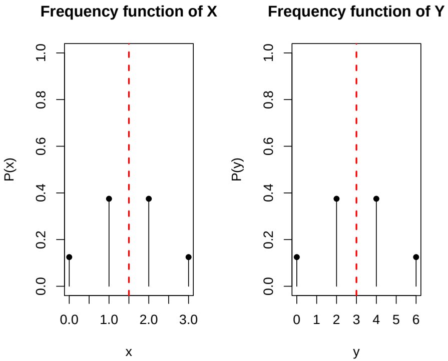


Figure: The probability mass function becomes wider with fixed probabilities, $c = 2$.

Properties of expectation of a random variable

- $E(X+d) = E(X) + d$, where d is a constant.

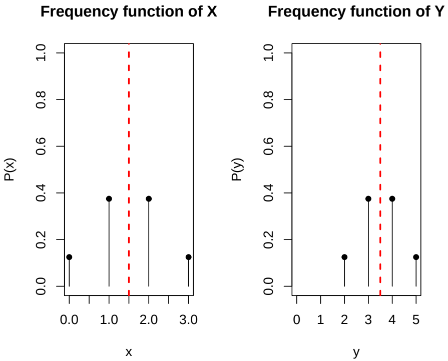


Figure: The probability mass function is shifted with fixed probabilities, $d = 2$.

Variance of a discrete random variable

- $\boxed{\text{Definition:}}$ If X is a discrete random variable with frequency function $p_X(x)$ and expected value $\mu = E(X)$, then the variance of X is

$$Var(X) = E\left[(X-\mu)^2\right] = \sum_i (x_i - \mu)^2 p_X(x_i).$$

- The **standard deviation** is the square root of the variance: $SD(X) = \sqrt{Var(X)}$.

- The variance is denoted by σ^2 and the standard deviation by σ.

- It quantifies the spread of the values of X around its mean.

- Equivalently,

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

Variance of a discrete random variable

- $\boxed{\text{Example:}}$ Let X denote the number of heads in 3 tosses of a fair coin with $p(X)$:
  - $P(X=0)=P(TTT)=\frac{1}{8}$
  - $P(X=1)=P(HTT)+P(THT)+P(TTH)=\frac{3}{8}$
  - $P(X=2)=P(HHT)+P(HTH)+P(THH)=\frac{3}{8}$
  - $P(X=3)=P(HHH)=\frac{1}{8}$
- The variance of X is
  $\sigma^2=\sum_{x=0}^{3}(x-1.5)^2\cdot p(x)=(0-1.5)^2\cdot\frac{1}{8}+(1-1.5)^2\cdot\frac{3}{8}+(2-1.5)^2\cdot\frac{3}{8}+(3-1.5)^2\cdot\frac{1}{8}=0.75$

Properties of the variance of a random variable

- $Var(X) = E(X^2) - [E(X)]^2$
- For any constant c,
$$Var(c \cdot X) = c^2 Var(X).$$
- For any constant d,
$$Var(X + d) = Var(X).$$

Outline

1 Discrete random variables

2 Bernoulli random variable

3 Binomial Distribution

4 Geometric Distribution

5 Poisson Distribution

Bernoulli random variable

- X is a **Bernoulli** random variable if it takes on only two values e.g., 0 and 1, with probabilities $1-p$ and p.
- **Bernoulli frequency function** is:

$$p_X(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x=0 \text{ or } x=1 \\ 0 & \text{otherwise} \end{cases}$$

Bernoulli random variable

- If A is an event, then the **indicator random variable**, I_A, takes on the value 1 if A occurs and the value 0 if A does not occur:

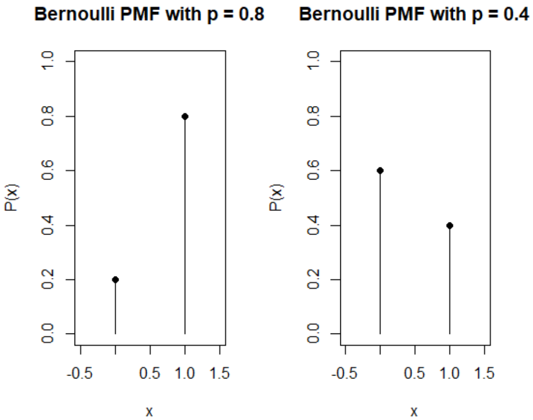
$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

  I_A is a Bernoulli random variable.

- Other examples : Flip a coin, if head then $x = 1$ (“success"), if tail $x = 0$ (“failure"). With probability of “success" $p = 1/2$.

- **Mean:** p, **Variance:** $p(1-p)$.

Frequency function of Bernoulli random variable



Outline

1. Discrete random variables
2. Bernoulli random variable
3. Binomial Distribution
4. Geometric Distribution
5. Poisson Distribution

Binomial Distribution

- Let $X_1 \cdots X_n$ be n independent Bernoulli variables with parameter p, then $Y = X_1 + \ldots + X_n$ is a Binomial random variable with parameters p and n.

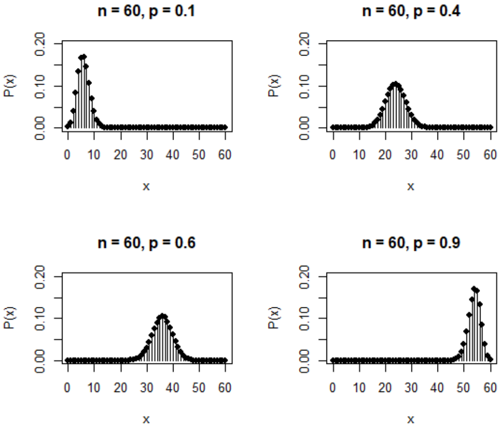
- The **Binomial frequency function** is given by

$$p_Y(k) = P(Y = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

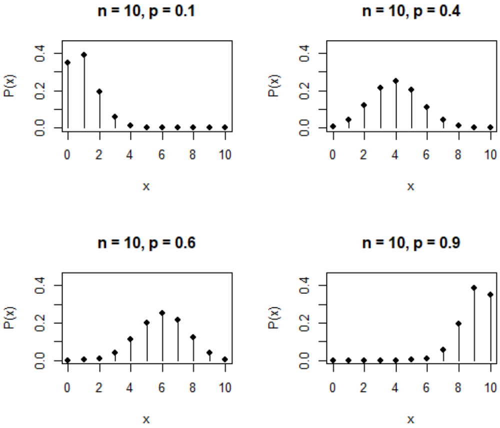
- $\boxed{\text{Example:}}$ Head = “success", Tail = “Failure", prob $p = 1/2$. Let Y be the total number of tails ("successes") in 5 trials. Then Y is a **binomial** random variable with parameters $p = \frac{1}{2}$ and $n = 5$.

- **Mean:** np, **Variance:** $np(1-p)$.

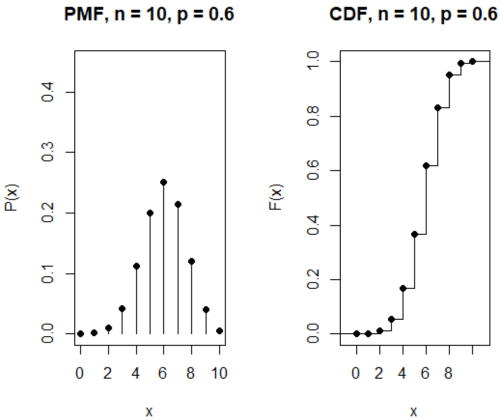
Frequency function of Binomial random variable



Frequency function of Binomial random variable



Frequency function and CDF of Binomial random variable



Binomial Distribution

- Visualize it

```
library(TeachingDemos)
vis.binom()
```

- Example: Tay-Sachs is a rare genetic disease affecting mainly infants and children, of Jewish or eastern European extraction. If a couple are both carriers of the Tay-Sachs disease, a child of theirs has probability 0.25 of being born with the disease. If such a couple has four children, what is the frequency function for the number of children who will have the disease?

R implementation - Distributions

For every distribution there are four functions. The commands for each distribution are prepended with a letter to indicate the functionality:

- “d" returns the height of the probability mass function
- “p" returns the cumulative distribution function
- “q" returns the inverse cumulative distribution function (quantiles)
- “r" returns randomly generated numbers

Check `help(Distributions)`

R implementation - Binomial distribution

- Frequency function, CDF, quantile function, and simulate random variables
  `dbinom(x, size, prob), pbinom, qbinom, and rbinom`

- $\boxed{\text{Example 1:}}$
  Let a family with $n = 5$ children. Each child can be a girl with probability $p = 1/2$.
  - What is the probability that $X = 3$ out of $n = 5$ children are girls?
  - What is the mean number of girls and $Var(X)$?

R implementation - Binomial distribution

- $\boxed{\text{Example 2:}}$
  Roll 12 dices simultaneously, and let X the number of 6's that appear. What is the probability of getting seven, eight, or nine 6's?

⇒ Let Y = {Get a 6 on one roll}, then $P(Y) = 1/6$ and the rolls constitute Bernoulli trials. Then $X \sim Binom(n = 12, p = 1/6)$ and we want
$P(7 \leq X \leq 9) = P(X \leq 9) - P(X < 7) = P(X \leq 9) - P(X \leq 6)$.

Outline

1. Discrete random variables
2. Bernoulli random variable
3. Binomial Distribution
4. Geometric Distribution
5. Poisson Distribution

Geometric distribution

- Flip a coin (Bernoulli trial) many times until you have a first head (“success"). On each trial, “success" occurs with probability p.
- Let X the total number of trials up to and including the first success. X follows a **Geometric** distribution with parameter p and its **frequency function** is given by:

$$P(X=k)=(1-p)^{k-1}p,\ k=1,2,3,\dots$$

- The geometric distribution is constructed by an **infinite** sequence of Bernoulli variables with parameter p the probability of success in each trial.

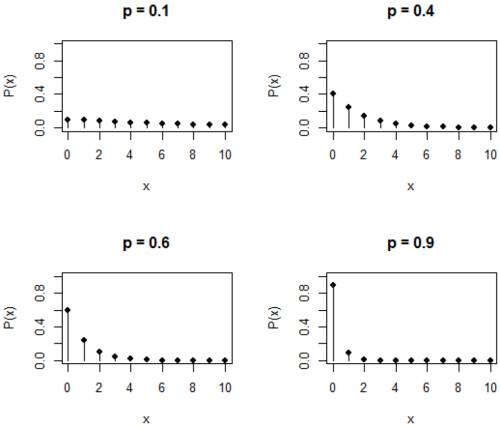
Geometric distribution

- **Mean:** $1/p$, **Variance:** $(1-p)/p^2$
- $\boxed{\text{Example:}}$ X is the number of offspring for parents who wish to have a girl. What is the frequency function of X.
- Alternative definition: Let X be the number of failures before a success (this is used in R)

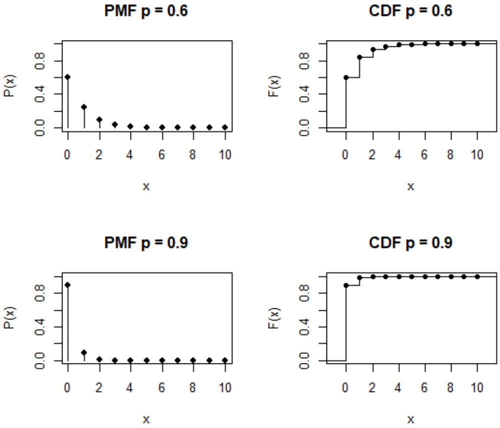
$$P(X=k)=(1-p)^k p,\ k=0,1,2,3,\ldots$$

- In this case, **Mean:** $(1-p)/p$, **Variance:** $(1-p)/p^2$

Frequency function of Geometric random variable



Frequency function and CDF of Geometric random variable



R implementation - Geometric distribution

- Frequency function, CDF, quantile function, and simulate random variables

```
dgeom(x, prob), pgeom, qgeom, and rgeom

sample. <- rgeom(100, 1/2)
summary(sample.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0     0.0     0.9     1.0     5.0
sd(sample.)
[1] 1.184922

hist(sample., breaks=seq(-0.5,6.5, 1),
     col='light grey', border='grey', xlab="")
```

R implementation - Geometric distribution

- Example 1: The Pittsburgh Steelers place kicker, Jeff Reed, made 81.2% of his attempted field goals in his career up to 2006. Assuming that his successive field goal attempts are approximately Bernoulli trials, find the probability that Jeff misses at least 5 field goals before his first successful goal.

$\Rightarrow$ If X = {the number of missed goals until Jeff's first success}, then $X \sim$ Geometric$(p = 0.812)$ and we want $P(X \geq 5) = P(X > 4)$.

```
pgeom(4, prob = 0.812, lower.tail = FALSE)
[1] 0.0002348493
```

Note: pay attention to the `lower.tail = FALSE` argument check `?pgeom`.

- Example 2: What is the probability that a family needs to have $x = 10$ births until the child is a girl? (i.e., 10 births with boys)

```
dgeom(10, 1/2)
[1] 0.0004882812
```

Outline

1. Discrete random variables

2. Bernoulli random variable

3. Binomial Distribution

4. Geometric Distribution

5. Poisson Distribution

Poisson Distribution

- The random variable X describes the number of times an event occurs in a given interval of time.
- Suppose we get on the average $\lambda = 4$ pieces of mail per day.
- Spread: sometimes a little more, sometimes a little less, once in a while nothing at all.
- Given only the average rate λ, for a certain period of observation, the **Poisson** distribution specifies how likely it is that the count will be 3, or 5, or 11, or any other number, during one period of observation.
- That is, it predicts the degree of spread around a known average rate of occurrence.

Poisson distribution

Examples of events that may be modelled as a Poisson distribution include:

- The number of phone calls arriving at a call center per minute.
- The number of goals in sports involving two competing teams.
- The number of mutations in a given stretch of DNA after a certain amount of radiation.

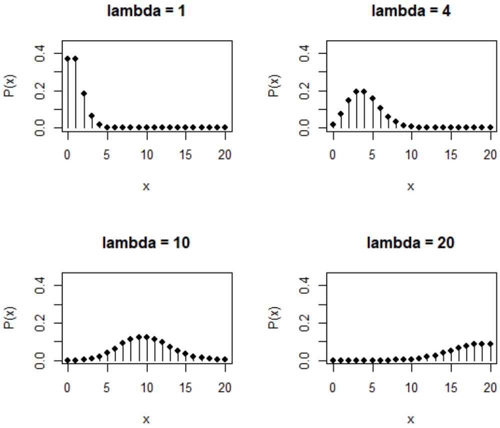
Poisson distribution

- The **Poisson frequency function** with parameter $\lambda (\lambda > 0)$ is given by:

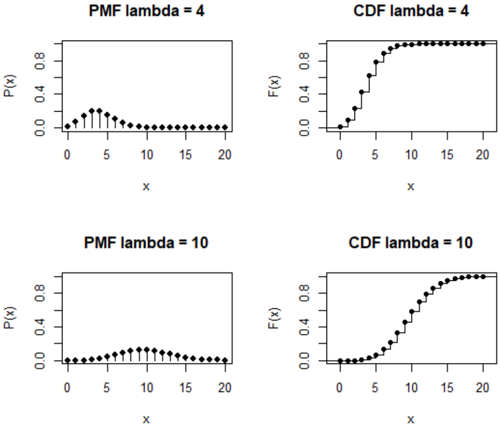
$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \ldots$$

- **Mean:** λ, **Variance:** λ.
- The Poisson distribution is the limit of the Binomial distribution:
  - $n \rightarrow \infty$
  - $p \rightarrow 0$ such that $np = \lambda$
- Note that for the Poisson distribution the probability of zero events is $e^{-\lambda}$ and the probability of at least one event is $1 - e^{-\lambda}$.
- Example: Two dice are rolled 100 times, and the number of double sixes, X, is counted. What is the distribution of X?

Frequency function of Poisson random variable



Frequency function and CDF of Poisson random variable



Poisson distribution

- $\boxed{\text{Example 1:}}$ On average five cars arrive at a car wash per hour. Let X the number of cars that arrive from 10am to 11am. Then $X \sim$ Poisson$(\lambda = 5)$. What is the probability that no car arrives during this period?

$\Rightarrow$ $P(X = 0) = e^{-5} \approx 0.0067$

- $\boxed{\text{Example 2:}}$ Suppose the car wash above is open from 8AM to 6PM, and let Y the number of customers that appear in this period. What is the distribution of Y? What is the probability that there are between 48 and 50 customers?

$\Rightarrow$ $P(48 \leq Y \leq 50) = 0.168$

R implementation - Poisson distribution

- Frequency function, CDF, quantile function, and simulate random variables:
  `dpois(x, lambda), ppois, qpois, and rpois`

Lecture 5: Continuous random variables

Outline

1. Continuous random variables

2. Normal distribution

3. Exponential distribution

4. X^2 Distribution

Outline

1. Continuous random variables

2. Normal distribution

3. Exponential distribution

4. X^2 Distribution

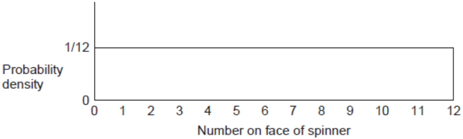
Continuous random variables

- In many applications we are interested in random variables that can take values on a continuum.
- Examples:
  - The glycose levels of diabetes patients.
  - Lifetime of an electronic component.
  - The IQ levels of children with divorced parents.
- For discrete rv's, their distribution is described by the frequency function $P(Y = y), y \in S_Y$.
- **Here instead of frequency function we talk about probability density function $f_X(x)$.**

Continuous random variables - Example

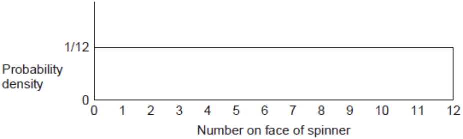
- Assume spinner wheel with 12 numbers ⇒ spinner can stop equally likely at any point on the circle.

- Infinite points on the circle ⇒ probability of a specific point virtually 0.
- The probabilities it stops between 1 and 2, and between 2 and 3 are equal $\Rightarrow \frac{1}{12}$.
- The distribution of probabilities is then:



Continuous random variables - Example - Cont'd

- The distribution of probabilities is then:



- The height of the continuous line (1/12) is called **probability density**.
- The area of the rectangle is 1: length of base $\times$ height = 12 $\times \frac{1}{12}$.
- The proportion of total area between 1 and 2 and between 2 and 3 is 1/12 $\Rightarrow$ probability spinner stops between these 2 numbers.
- This is the uniform distribution.

Uniform Distribution

- The uniform density on the interval $[0,12]$ is defined as follows:

$$f_X(x) = \begin{cases} \frac{1}{12}, & 0 \leq x \leq 12 \\ 0, & x < 0 \text{ or } x > 12 \end{cases}$$

- The uniform density on a general interval $[a,b]$ is:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x < a \text{ or } x > b \end{cases}$$

Probability density function

- $\boxed{\text{Properties}}$
  - The pdf at x should never give a negative value: $f_X(x) \geq 0$.
  - The area under the curve of the function f_X should be 1. Formally, we write it as: $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
- **The probability that X falls in the interval** (a,b) is the area under the curve of the density function f_X between a and b:

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

- Note that $\boxed{P(X = c) = 0}$ and that $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$.
- Note that this is not true for a discrete random variable!!

Cumulative distribution function

- The **cumulative distribution function** of a continuous random variable X is defined in the same way as for a discrete random variable:

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

- It is the area under the curve of f_x up to x.

- $\boxed{\text{Properties:}}$

  - The cdf is non-decreasing.
  - The cdf is 0 for values lower than the lowest value X can take.
  - The cdf is 1 for values higher than the highest value X can take.

- The cdf can be used to calculate the probability that X falls in an interval $\boxed{\text{P(a < X < b) = F}_X(b) - F_X(a)}$ = $\int_a^b f(u)du$.

- F is primitive or anti-derivative of f: Outside of the scope of the course.

Uniform Distribution - Revisited

- The uniform density on the interval $[0,12]$ is defined as:

$$f_X(x) = \begin{cases} \frac{1}{12}, & 0 \leq x \leq 12 \\ 0, & x < 0 \text{ or } x > 12 \end{cases}$$

- The uniform density on a general interval $[a,b]$ is:

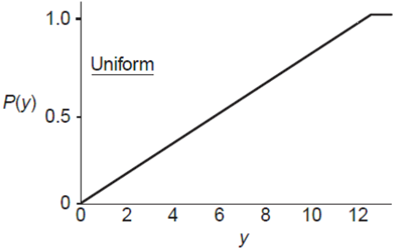
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x < a \text{ or } x > b \end{cases}$$

- The cdf of this density is:

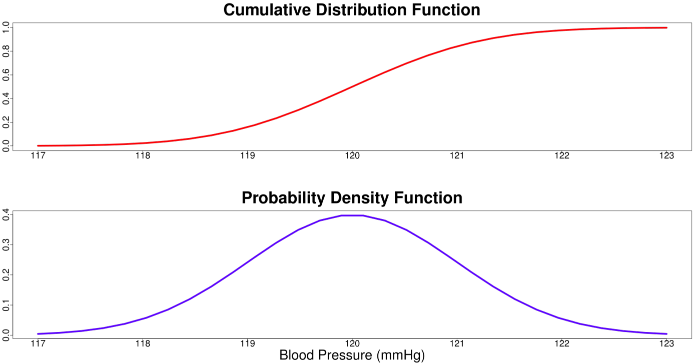
$$F_X(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x \geq b \end{cases}$$

Uniform Distribution - Revisited

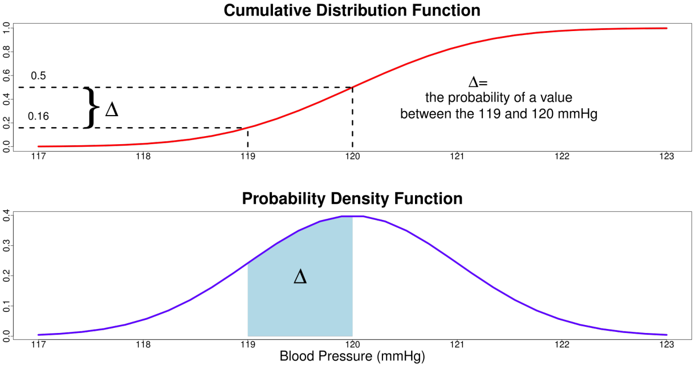
- Uniform CDF:



Cumulative distribution vs pdf



Cumulative distribution vs pdf



Expected value and variance of a continuous random variable

- The mean and the variance of a continuous rv is computed as in the discrete case, but the summation is replaced by the integral.
- If X is a continuous random variable with pdf $f(x)$:
  - The **expected value** of X is $E(X) = \int_{-\infty}^{+\infty} x_i \cdot f(x_i) dx_i$.
  - The **variance** of X is $Var(X) = \int_{-\infty}^{+\infty} (x_i - \mu)^2 \cdot f_X(x_i) dx_i$.
- $E(X)$ is also referred to as the **mean** of X and is often denoted by μ or μ_X. It is a measure of location.
- $Var(X)$ is a measure of spread around the mean.

R implementation - Cumulative distribution vs pdf

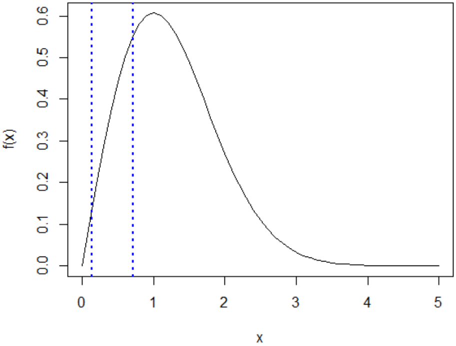
- $\boxed{\text{Example:}}$ Let X have pdf $f_X(x) = x \cdot \exp^{-x^2/2}$, $x > 0$.
  - Find $P(0.14 \leq X \leq 0.71)$.

```
# probability
f <- function(x) x * exp(-x^2/2)
prob. <- integrate(f, lower = 0.14, upper = 0.71)
prob.
x <- seq(from  = 0, to = 5, length.out = 51)
plot(x, f(x), type = "l")
abline(v = 0.14, lty = 3, lwd = 2, col = "blue")
abline(v = 0.71, lty = 3, lwd = 2, col = "blue")
```

- Compute $E(X)$ and $Var(X)$.

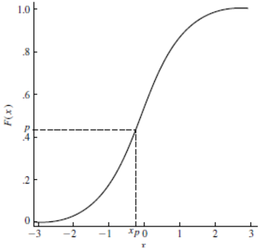
```
# mean
xf <- function(x) x * x * exp(-x^2/2)
mean.x <- integrate(xf, lower = 0, upper = 100)
# 1.253314
# variance
x2f <- function(x) (x - mean.x$ value)^2 * x * exp(-x^2/2)
var.x <- integrate(x2f, lower = 0, upper = 100)
#0.4292037
```

R implementation - Cumulative distribution vs pdf



Quantiles

- Other useful quantities to summarize the distribution of a random variable are the quantiles.
- The pth **quantile** x_p is defined as $F(x_p) = P(X \leq x_p) = p$, i.e., the point x_p of X at which the cdf F_X gets p.
- Popular quantiles are:
  - **median** or 50% quantile: i.e., x_p for which $P(X \leq x_p) = 0.50$.
  - $\mathbf{25}\%$ and $\mathbf{75}\%$ **quantiles:** i.e., x_p for which $P(X \leq x_p) = 0.25$ and $P(X \leq x_p) = 0.25$, respectively.



Quantiles

- $\boxed{\text{Example:}}$ Let X the score on an IQ test, which follows a distribution with cdf $F_X(x)$. What is the lowest possible IQ score x that a person can have and still be in the top 1% of all IQ scores?

⇒ $\boxed{\textit{Answer:}}$ If a person is in top 1%, then this means that 99% of the people have lower IQ scores. So we want a value x such that $F_X(x) = P(X \leq x) = 0.99$, i.e., we want the 99% quantile.

If X is Uniform on $[0,1]$ then we may use the R function `qunif(0.99)`.

R implementation - Uniform distribution

R functions:

- `dunif(x, min=0, max=1)` provides density $f(x)$ of Uniform variable on interval [0,1].
- `runif(N, min=0, max=1)` provides N trials for a Uniform variable.
- `qunif(p, min=0, max=1)` gives the quantile corresponding to the probability p.
- `punif(x, min=0, max=1)` gives the $P(X \leq x)$.

Outline

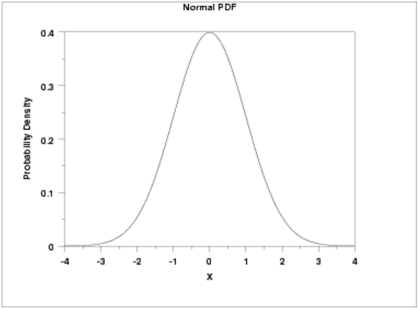
1. Continuous random variables
2. Normal distribution
3. Exponential distribution
4. X^2 Distribution

Normal distribution

The normal distribution or Gaussian distribution plays a central role in probability and statistics. Its form is bell shaped.

Examples are

- the distribution of IQ scores in the population



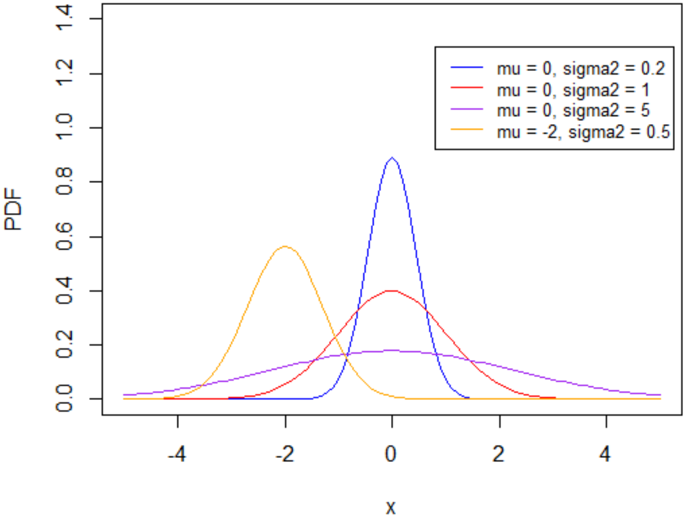
Normal distribution

- $\boxed{\text{Definition:}}$ For $-\infty < x < \infty$, $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$
- The distribution depends on two parameters:
  - $-\infty < \mu < \infty$ the mean which determines its location.
  - $\sigma > 0$ the standard deviation which determines the spread.
- *Notation:* X follows normal distribution with parameters μ and σ:

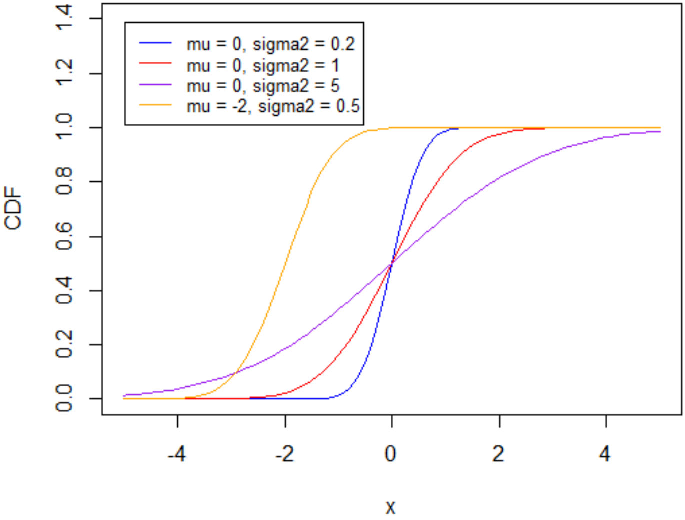
$$X \sim N(\mu, \sigma^2).$$

- f with $\mu = 0$ and $\sigma = 1$ is the **standard normal density**.

Examples of Normal distribution - PDF



Examples of Normal distribution - CDF



R

R functions:

- `dnorm(x, mean = 0, sd = 1)` provides $f(x)$ for standard normally distributed variable.
- `qnorm(p, mean = 0, sd = 1)` provides quantiles for standard normally distributed variable.
- `rnorm(N, mean = 0, sd = 1)` provides N trials for a standard normally distributed variable.
- `pnorm(x, mean = 0, sd = 1)` gives gives the $P(X \leq x)$.

R implementation - Quantiles

- The αth quantile, i.e., $P(Z \leq z_\alpha) = \alpha$, can be calculated as: `qnorm(`α`, lower.tail = TRUE)`.
- Equivalently we may use `qnorm(1 -` α`, lower.tail = FALSE)`.
- $\boxed{\text{Example:}}$ Find the values $z_{0.025}$, $z_{0.01}$, and $z_{0.005}$.

```
qnorm(c(0.025, 0.01, 0.005), lower.tail = TRUE)
[1] -1.959964 -2.326348 -2.575829
qnorm(c(0.975, 0.99, 0.995), lower.tail = FALSE)
```

- Note the `lower.tail` argument!
- Similarly compute quantiles for any Normal rv by setting the arguments `mean` and `sd`.
- For the standard Normal distribution it holds: $P(Z \leq z_\alpha) = 1 - P(Z \leq -z_\alpha)$.

Functions of a random variable

- Let X a continuous random variable with $f_X(x)$ and $F_X(x)$ its pdf and cdf, respectively.
- We often are interested on another rv Y which is function of X, e.g., $Y = a \cdot X + b$.
- Values of the cdf $F_Y(y)$ can be computed using the cdf of X as follows:

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(a \cdot X + b \leq y) \\ &= P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

- This result is useful because we can now compute probabilities for any Normal rv using the cdf of the standard Normal distribution.
- Let X follow the standard Normal distribution, then $P(Y \leq y) = \Phi\left(\frac{y-b}{a}\right)$, where $\Phi(.)$ is the standard Normal cdf.

Functions of a normal random variable

- Using the properties of the expectation and variance:

  If $X \sim N(\mu,\, \sigma^2)$ and $Y = \alpha X + b$, then $Y \sim N(\alpha\mu + b,\, \alpha^2\sigma^2)$.

- If we set $\alpha = \frac{1}{\sigma}$ and $b = \frac{-\mu}{\sigma}$ then $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$.

- Thus we can “standardize" any Normal rv by subtracting the mean and dividing with the sd $\Rightarrow$ Compute probabilities for any Normal rv using the probabilities of the standard normal distribution, which are given in tables.

Functions of a normal random variable

- For any normal rv X with parameters μ and σ we have:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

and

$$P(x_0 < X < x_1) = F_X(x_1) - F_X(x_0) = \Phi\left(\frac{x_1 - \mu}{\sigma}\right) - \Phi\left(\frac{x_0 - \mu}{\sigma}\right).$$

- With $\Phi(.)$ the cdf of the standard Normal distribution.

Functions of a normal random variable

- $\boxed{\text{Example:}}$

  Scores on an IQ test are approximately normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. Here we are referring to the distribution of IQ scores over a very large population. An individual is selected at random.

  What is the probability that his score X satisfies $120 < X < 130$?

- **Solution:** $P(120 < X < 130) = 0.069$.

Outline

1. Continuous random variables
2. Normal distribution
3. Exponential distribution
4. X^2 Distribution

Exponential distribution

- Exponential random variables are used to model the lifetime of electronic devices, waiting times, survival analysis, etc.

- The continuous random variable X has an exponential distribution, with parameter λ, if its **probability density function** is given by:

  $$f_X(x) = \lambda \exp\{-\lambda x\}, \quad x \geq 0, \quad \lambda > 0$$

  λ is known as the rate parameter.

- λ must be positive, since it measures rate of events e.g. if the time X between phone calls is Exponential, λ is the average rate of phone calls per hour.

- **Mean:** $\mu = 1/\lambda$ and **Variance:** $\sigma^2 = 1/\lambda^2$.

- **Cumulative Distribution Function:**

  $$F_X(t) = Pr(X \leq t) = \int_0^t \lambda \exp\{-\lambda x\} dx = 1 - \exp\{-\lambda t\}, \quad t \geq 0.$$

Exponential distribution

- A commonly used alternative parametrization is to define the pdf of an exponential distribution as
$$f_X(x) = \frac{1}{\beta}\exp\{-\frac{x}{\beta}\}, \quad x \geq 0, \quad \beta > 0.$$

- In this case $\beta = \frac{1}{\lambda}$ and is called the scale parameter.

- **Mean:** $\mu = \beta$ and **Variance:** $\sigma^2 = \beta^2$.

- **Cumulative Distribution Function:**
$$F_X(t) = Pr(X \leq t) = \int_0^t \frac{1}{\beta}\exp\{-\frac{1}{\beta}x\}dx = 1 - \exp\{-\frac{1}{\beta}t\}, \quad t \geq 0$$

- To avoid confusion: **It is important to check the parameterization used!** For the rest we will use the first parameterization with λ.

Exponential distribution

- $\boxed{\text{Important property:}}$ Exponential rvs are “**memoryless**" because they “forget" how old they are at every instant:

$$P(X > s+t \mid X > s) = P(X > t), \quad \text{for any } s,\ t > 0$$

- However long you wait, the time until the next occurrence has the same distribution.

- Due to the memoryless property, the Exponential distribution is not a good model for human lifetimes: the probability that a 16-year-old will live at least 10 more years is not the same as the probability that an 80-year-old will live at least 10 more years.

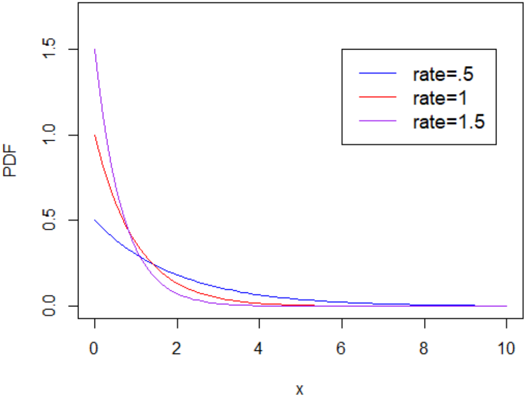
- The Exponential distribution has a constant failure rate $\frac{1}{\lambda}$ which is not realistic for humans.

Exponential distribution

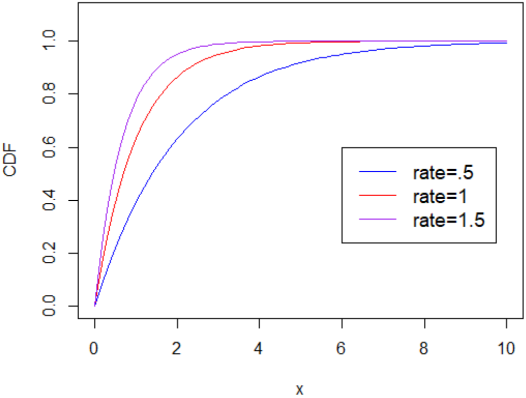
- The exponential distribution is closely related to the Poisson distribution.
- Reminder: A Poisson rv describes the number of occurrences of a random event in a fixed time interval, e.g. the number of customers in a day.
- An Exponential rv measures the time from now until the first occurrence of the event, e.g. time until the next customer enters the shop ⇒ it models times between events of a Poisson process.
- R implementation:
- PDF, CDF, quantile function, and simulate random variables

```
dexp(x, rate = 1), pexp, qexp, and rexp
```

Exponential distribution - Visualization



Exponential distribution - Visualization



Outline

1 Continuous random variables

2 Normal distribution

3 Exponential distribution

4 X^2 Distribution

X^2 Distribution

- The pdf of a chi-square distribution with p degrees of freedom is:

$$f_X(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2}, \; x > 0,$$

where $\Gamma(.)$ is the gamma function.

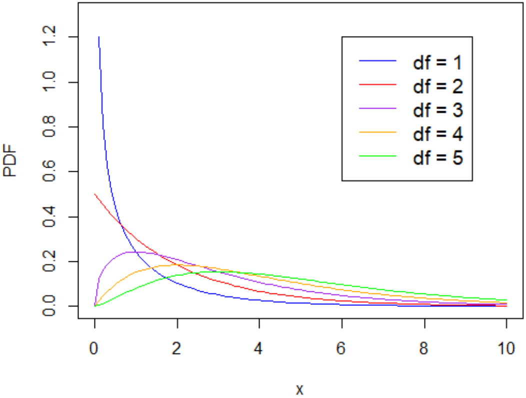
- If $Z \sim N(\mu = 0, \, \sigma^2 = 1)$ then $Z^2 \sim \chi^2_1$.

- R implementation:

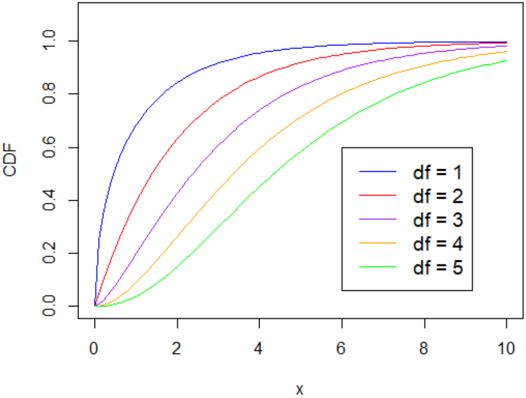
- PDF, CDF, quantile function, and simulate random variables:

```
dchisq(x, df), pchisq, qchisq, and rchisq
```

X^2 distribution - Visualization



X^2 distribution - Visualization



Lecture 6: Joint distributions

Outline

1. Introduction
2. Joint Distributions - Discrete random variables
3. Joint Distributions - Continuous random variables
4. Covariance
5. Correlation
6. Independent random variables

Outline

1. Introduction
2. Joint Distributions - Discrete random variables
3. Joint Distributions - Continuous random variables
4. Covariance
5. Correlation
6. Independent random variables

Examples of joint distributions

- In many applications we are interested in the distribution of two or more random variables:
    - The joint distribution of heights of fathers and sons.
    - The joint distribution of age and cholesterol levels.
    - The joint distribution of a variable X defined as presence $(X = 1)$ or absence $(X = 0)$ of a disease and Y the number of affected relatives of the same person.
    - The joint distribution of height Y and the variable X denoting gender; $X = 1$ for females, $X = 0$ otherwise.

Outline

1. Introduction
2. Joint Distributions - Discrete random variables
3. Joint Distributions - Continuous random variables
4. Covariance
5. Correlation
6. Independent random variables

Joint Distribution - Discrete Random variables

- Let X and Y be two discrete random variables.
- Let also X and Y take values the $x_1, x_2, x_3, \cdots$ and $y_1, y_2, y_3, \cdots$, respectively.
- Their **joint frequency function** is defined as:

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

with $\sum_i \sum_j p(x_i, y_j) = 1$.

Joint frequency function - Example

- Let’s consider the sample space of three times tossing a fair coin: $\{hhh, hht, hth, thh, htt, tht, tth, ttt\}$.
- Let X be 1 if the first toss is a head and let Y be the total number of heads.
- Then the joint frequency function of X and Y can be obtained by counting:

	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

- Note that the sum of all cells sum to one.

Joint frequency function - Example

- We may compute all kind of probabilities from this table, i.e., from joint frequency function.
- For example $P(X<Y)$, $P(X=0)$ or $P(X=1)$.
- We find these probabilities by simply summing the appropriate cells: $P(X=0)=P(X=0,Y=0)+P(X=0,Y=1)+P(X=0,Y=2)+P(X=0,Y=3)$.
- Note $P(X=0)$ and $P(X=1)$ form the **marginal frequency function** of X, $p_X(x)$.
- We derive the marginal ff of X by summing over the values of Y: $\boxed{p_X(x)=\sum_i p(x,y_i).}$
- These results can be generalized to more than two variables.

Outline

1 Introduction

2 Joint Distributions - Discrete random variables

3 Joint Distributions - Continuous random variables

4 Covariance

5 Correlation

6 Independent random variables

Joint Distribution -Continuous random variables

- Let X and Y be continuous random variables.
- Their joint distribution can be described using the **joint probability density function** $f(x,y)$.
- Similar to the univariate case, $f(x,y)$ satisfies the following conditions:
  - $f(x,y) \geq 0$ for $-\infty < x < \infty$ and $-\infty < y < \infty$.
  - $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$.
- Similar to the univariate case, joint probabilities are computed using integration.
- The marginal pdf of X is derived by integrating the joint pdf $f(x,y)$ with respect to $f_Y(y)$.

CDF

- The cdf for two random variables X and Y (either discrete or continuous) is (properties in Lecture 4)
  $\boxed{F(x,y) = P(X \leq x, Y \leq y)}$

- We can use the cdf to compute probabilities, e.g.,:
  $P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1)$

- Similar formula's can be written for cdf's of more than 2 random variables.

CDF - Illustration

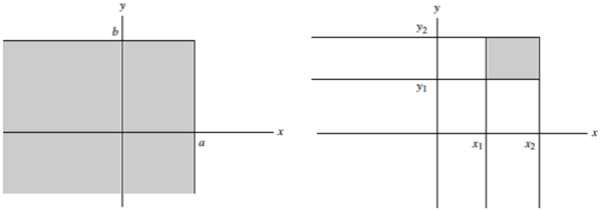


Figure: $F(a,b) = P(X \le a, Y \le b)$ and $P(x_1 < X \le x_2, y_1 < Y \le y_2)$.

Outline

1. Introduction
2. Joint Distributions - Discrete random variables
3. Joint Distributions - Continuous random variables
4. Covariance
5. Correlation
6. Independent random variables

Covariance

- The main motivation to form joint distributions of random variables is to study their dependence.

- **Covariance of two random variables** is a measure of their joint variability or degree of association/dependence.

- Example:
  Let X the number of mutations for sibling 1 and Y be the number of mutations for sibling 2. The covariance of X and Y tells us how they vary together.

Covariance

- $\boxed{\text{Definition:}}$ If X and Y are jointly distributed random variables with expectations μ_X and μ_Y respectively, the covariance of X and Y is
$$Cov(X,Y) \;=\; E[(X-\mu_X)\cdot(Y-\mu_Y)].$$
- It can be shown that: $Cov(X,Y) = E(X\cdot Y) - E(X)\cdot E(Y)$.

R implementation - Covariance

- Let X and Y be the number of mutations carried by sibling 1 and sibling 2, respectively. The joint distribution $P(X,Y)$ of X and Y is

```
PXY
   0    1    2
0 0.3  0.1  0.01
1 0.1  0.3  0.06
2 0.01 0.05 0.07
```

- Let also $E(X) = E(Y) = 0.7$.

- Thus the expectation of $X \cdot Y$ is

```
EXY <- PXY[2, 2] + 2 * PXY[2, 3] + 2 * PXY[3, 2] + 4 * PXY[3, 3]
EXY
[1] 0.8
```

- The covariance of X and Y is

```
COVXY <- EXY - EX * EY
COVXY
[1] 0.27
```

Covariance of linear combinations of random variables

- $Cov(a+X,Y) = Cov(X,Y)$
- $Cov(aX,bY) = abCov(X,Y)$
- $Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z)$
- $Cov(aW+bX,cY+dZ) = acCov(W,Y) + bcCov(X,Y) + adCov(W,Z) + bdCov(X,Z)$

Covariance of linear combinations of random variables

- It holds: $Var(X) = Cov(X,X)$ and thus:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

- $\boxed{Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)}.$

Outline

1 Introduction

2 Joint Distributions - Discrete random variables

3 Joint Distributions - Continuous random variables

4 Covariance

5 Correlation

6 Independent random variables

Correlation

- $\boxed{\text{Definition:}}$ If X and Y are jointly distributed random variables and the variances and covariances are nonzero, then the **correlation coefficient** of X and Y, denoted by ρ is:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

- Note that the correlation coefficient is dimensionless and invariant to linear transformations of X and Y.

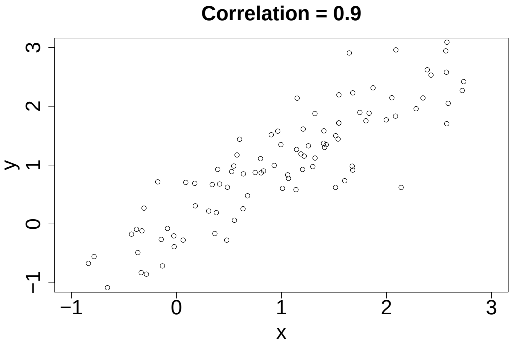
- We often write it as: $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ and use $\sigma_{XY} = \rho \sigma_X \sigma_Y$.

Correlation

- $-1 \leq \rho \leq 1$.
- We say:
  - X and Y are positively correlated if $\rho(X,Y) > 0$.
  - X and Y are negatively correlated if $\rho(X,Y) < 0$.
  - X and Y are uncorrelated if $\rho(X,Y) = 0$.
- $Cov(X,Y)$ and $\rho(X,Y)$ must have the same sign: both are positive, or both are negative, or both are zero.
- The correlation coefficient ρ is a measure of linear association between X and Y.

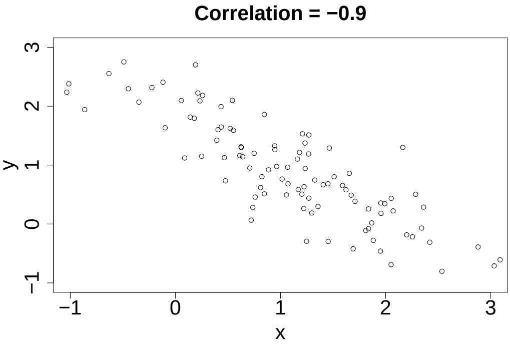
Visualization of correlated normal distributed variables

```
mu1 <- 1
sig1 <- 1
mu2 <- 1
sig2 <- 1
rho <- 0.9
x <- rnorm(100, mu1, sqrt(sig1))
y <- rnorm(100, (mu2 + rho * (sqrt(sig2)/sqrt(sig1)) * (x - mu1)),
           sqrt(sig2 * (1 - rho^2)))
plot(x, y, xlim = c(-1, 3), ylim = c(-1, 3))
```



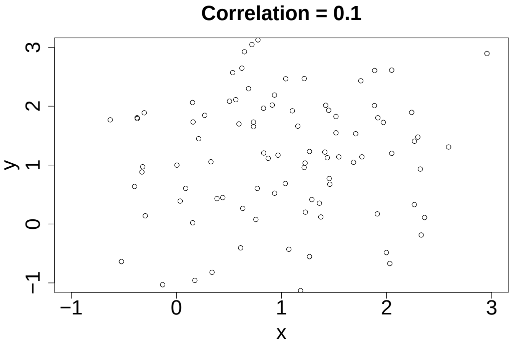
Visualization of correlated normal distributed variables

```
mu1 <- 1
sig1 <- 1
mu2 <- 1
sig2 <- 1
rho <- -0.9
x <- rnorm(100, mu1, sqrt(sig1))
y <- rnorm(100, (mu2 + rho * (sqrt(sig2)/sqrt(sig1)) * (x - mu1)),
           sqrt(sig2 * (1 - rho^2)))
plot(x, y, xlim = c(-1, 3), ylim = c(-1, 3))
```



Visualization of correlated normal distributed variables

```
mu1 <- 1
sig1 <- 1
mu2 <- 1
sig2 <- 1
rho <- 0.1
x <- rnorm(100, mu1, sqrt(sig1))
y <- rnorm(100, (mu2 + rho * (sqrt(sig2)/sqrt(sig1)) * (x - mu1)),
           sqrt(sig2 * (1 - rho^2)))
plot(x, y, xlim = c(-1, 3), ylim = c(-1, 3))
```



Bivariate Normal Distribution

- Let X heights of fathers and Y the heights of sons.
- The joint distribution of X and Y is the bivariate normal distribution with pdf:
  $f(x,y) =$
  $$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_X\sigma_Y}\right\}\right],$$

  for $(x,y) \in \Re^2$
- It depends on 5 parameters:
  - μ_X and σ_X describe location and spread of X.
  - μ_Y and σ_Y describe location and spread of Y.
  - ρ describes how “correlated" X and Y are.

Bivariate Normal Density

- The pdf of the bivariate normal distribution using vectors and matrices:

$$f_{X,Y}(\mathbf{z}) = \frac{1}{2\pi \mid \Sigma \mid^{1/2}} \text{exp} \left\{ -\frac{1}{2} (\mathbf{z} - \mu)^T \Sigma^{-1} (\mathbf{z} - \mu) \right\}$$

- $\mathbf{z} = (x, y)$
- $\mu = (\mu_X, \mu_Y)$
- $\Sigma = \left[\begin{array}{cc} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_Y\sigma_X & \sigma_Y^2 \end{array} \right]$

R implementation - Multivariate normal distribution

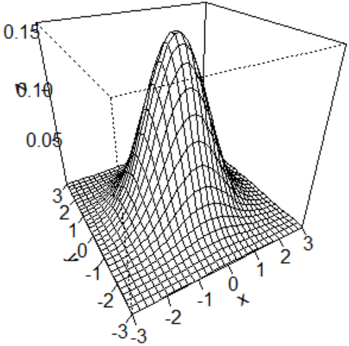
- R functions `dmvnorm(.)` and `rmvnorm(.)` in R package `mvtnorm` for pdf and simulate random vectors.

- 3-D plot of bivariate normal pdf

```
library(mvtnorm)
sigma.x <- 1
sigma.y <- 1
rho <- 0
Sigma <- cbind(c(sigma.x^2, rho * sigma.x * sigma.y),
               c(rho * sigma.x * sigma.y, sigma.y^2))
x <- y <- seq(from = -3, to = 3, length.out = 30)
f <- function(x, y) dmvnorm(cbind(x, y), mean = c(0, 0),
        sigma = Sigma)
z <- outer(x, y, FUN = f)
persp(x, y, z, theta = -30, phi = 30, ticktype = "detailed")
```

R implementation - Multivariate normal distribution

- 3-D plot of bivariate normal pdf



R implementation - Multivariate normal distribution

- 3-D plot of bivariate normal pdf

```
library(rgl)
dnorm2d<-function(x,y,mu1,mu2,sigma1,sigma2,rho){
  xoy = ((x-mu1)^2/sigma1^2 -
           2*rho * (x-mu1)/sigma1 * (y-mu2)/sigma2 +
           (y-mu2)^2/sigma2^2)/(2 * (1 - rho^2))
  density = exp(-xoy)/(2 * pi *sigma1*sigma2*sqrt(1 - rho^2))
  density
}
x<-seq(-5,5,by=0.1)
y<-seq(-5,5,by=0.1)
ff3<-function(x,y){
  dnorm2d(x,y,0,0,1,1,0)}
open3d()
z<-outer(x,y,ff3)
persp3d(x,y,z,col="green",main="ff3")
```

R implementation - Multivariate normal distribution

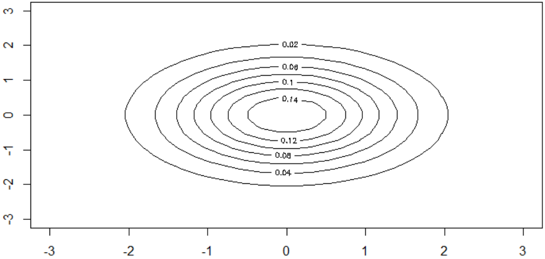
- R functions `dmvnorm(.)` and `rmvnorm(.)` in R package `mvtnorm` for pdf and simulate random vectors.

- Contour plot of bivariate normal pdf

```
library(mvtnorm)
sigma.x <- 1
sigma.y <- 1
rho <- 0
Sigma <- cbind(c(sigma.x^2, rho * sigma.x * sigma.y),
               c(rho * sigma.x * sigma.y, sigma.y^2))
x <- y <- seq(from = -3, to = 3, length.out = 30)
f <- function(x, y) dmvnorm(cbind(x, y), mean = c(0, 0),
        sigma = Sigma)
z <- outer(x, y, FUN = f)
contour(x, y, z)
```

R implementation - Multivariate normal distribution

- R functions `dmvnorm(.)` and `rmvnorm(.)` in R package `mvtnorm` for pdf and simulate random vectors.
- Contour plot of bivariate normal pdf



Bivariate Normal Distribution

- The bivariate normal distribution may be intimidating at first, but it is more tractable compared to other multivariate distributions.
- If $(X,Y) \sim N_2(\mu,\Sigma)$ then $X \sim N(\mu_X,\sigma_X^2)$ and $Y \sim N(\mu_Y,\sigma_Y^2)$.
- If $X \sim N(\mu_X,\sigma_X^2)$ and $Y \sim N(\mu_Y,\sigma_Y^2)$ their joint distribution is not necessarily bivariate normal.

Outline

1. Introduction
2. Joint Distributions - Discrete random variables
3. Joint Distributions - Continuous random variables
4. Covariance
5. Correlation
6. Independent random variables

Independent random variables

- $\boxed{\text{Definition:}}$ The random variables $X_1, X_2, \cdots X_n$ are said to be **independent** if their joint cdf factors into the product of their marginal cdf's:
  $F(x_1, x_2, \cdots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n)$, for all $x_1, x_2, \cdots, x_n$
- For discrete random variables, “cdf factors" is equivalent to state that “their joint frequency function factors into the marginal frequency functions".
- For continuous random variables, “cdf factors" is equivalent to state that “their joint density function factors into the marginal density functions".

Independent random variables

- $\boxed{\text{Example:}}$ Roll a fair dice twice. The joint frequency function is:

$$f_{XY}(x,y) = \frac{1}{36}, x = 1, \ldots, 6, y = 1, \ldots, 6.$$

The marginals are $f_X(x) = 1/6, x = 1, \ldots, 6$ and $f_Y(y) = 1/6, y = 1, \ldots, 6$

X and Y are independent since for every x and y in the joint support the joint frequency function satisfies:

$$f_{X,Y}(x,y) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = f_X(x) f_Y(y).$$

Expectation of functions of random variables

- Remember that if X and Y are independent $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$.
- It also holds that if X and Y are independent $E(X \cdot Y) = E(X) \cdot E(Y)$.
- It can be shown that if X and Y are independent random variables and $g(.)$ and $h(.)$ are fixed functions, then $E[g(X) \cdot h(Y)] = E[g(X)] \cdot E[h(Y)]$.
- If X and Y are independent random variables then $Cov(X,Y) = \rho(X,Y) = 0$.
- The converse is not true as a general rule:

  Two dependent random variables can be uncorrelated. Even though Y is a function of X, it is possible that $\rho(X,Y) = 0$: **They are just not linearly dependent!**

Expectations of Linear Combinations of Random Variables

- If $X_1, \cdots, X_n$ are jointly distributed random variables with expectations $E(X_i)$ and Y is a linear function of the X_i, $Y = a + \sum_{i=1}^{n} b_i X_i$, then

$$E(Y) = a + \sum_{i=1}^{n} b_i E(X_i)$$

- A nice application of this result is the computation of the expected value of a binomial random variable.

Expectations of Linear Combinations of Random Variables

- $\boxed{\text{Example:}}$ Let $Y \sim$ Binomial(n, p) with ff $p_Y(y) = \left(\begin{array}{c} n \\ y \end{array}\right) p^y (1-p)^{n-y}$.

  We can compute the expected value as: $E(Y) = \sum_{y=0}^{n} y \left(\begin{array}{c} n \\ y \end{array}\right) p^y (1-p)^{n-y}$ which is not straightforward.

- Instead notice that $Y = \sum_{i=1}^{n} X_i$ with $X_i \sim$ Bernoulli(p).
- We know that $E(X_i) = 0 \times p + 1 \times (1-p) = p$ and thus we can derive $E(Y)$ using the last property of the expected value: $E(Y) = \sum_{i=1}^{n} E(X_i) = n \times p$.
- Similarly for the variance: $Var(Y) = \sum_{i=1}^{n} Var(X_i) = n \times p \times (1-p)$.

Lecture 7: Conditional distributions

Outline

1. Conditional frequency function

2. Conditional Expectation

Outline

1. Conditional frequency function

2. Conditional Expectation

Examples

- In many applications we are interested in the distribution of X given that Y takes a specific value y:
  - The distribution of the height of sons (X) given that the father's height is $< 160cm$ (Y).
  - The distribution of a person's cholesterol level (X) given his age y.
  - The distribution of a random variable X defined as presence ($X = 1$) or absence ($X = 0$) of a disease given the number of affected relatives (y).
- We use the **conditional pmf or pdf** to update our knowledge for the distribution of X after we have observed that $Y = y$.

Conditional frequency function

- Let X and Y be jointly distributed discrete random variables, the **conditional probability that** $X = x$ **given that** $Y = y$, if $p_Y(y) > 0$ is:

$$\boxed{P(X = x \mid Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{p_{XY}(x,y)}{p_Y(y)}}$$

- $P(X = x \mid Y = y) = 0$, if $p_Y(y) = 0$.

- Note that for $p_Y(y) > 0$, $p_{X|Y}(x \mid y)$:
  - is a function of x only because y is kept fixed.
  - it is a frequency function since it is non-negative and sums to 1.
  - if X and Y are independent it equals the marginal pmf of X.

- For the continuous case we replace the pmf/ff by the pdf.

Joint frequency function - Illustration

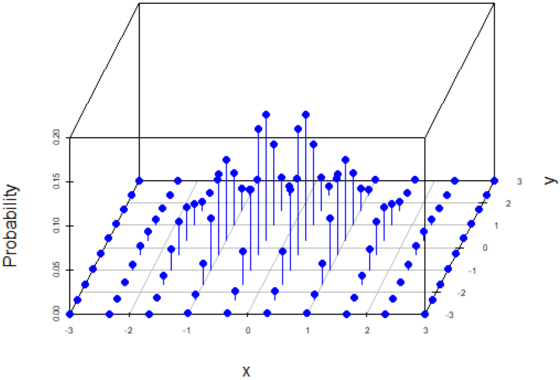


Figure: $p_{X,Y}(x,y) = P(X = x, Y = y)$ for $x \in S_X$ and $y \in S_Y$.

Conditional frequency function - Illustration

- Knowing that $X = 1$ means that we consider only the red joint probabilities and rescale them by dividing with $P(X = 1)$ which is the sum of the red spikes.

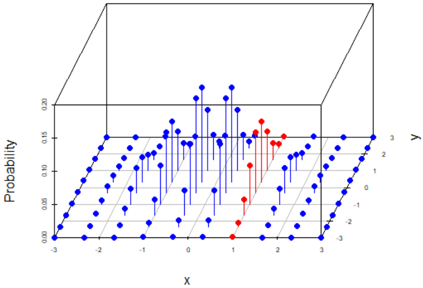


Figure: $p_{Y|X}(y \mid 1) = P(Y = y \mid X = 1)$ for $y \in S_Y$.

Marginal frequency function - Illustration

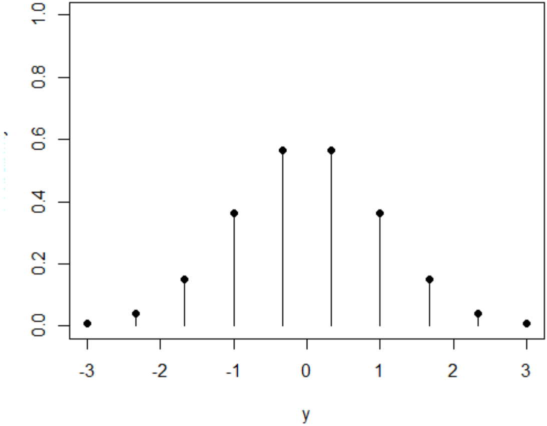


Figure: $p_Y(y) = \sum_x P(X = x, Y = y)$ for $y \in S_Y$.

Marginal frequency function

- From the multiplication law we have:

$$p_{XY}(x,y) = p_{Y|X}(y \mid x)p_X(x)$$

- We can derive the marginal frequency function via the conditional frequency function:

$$\boxed{p_Y(y) = \sum_x p_{Y|X}(y \mid x)p_X(x)}$$

  as the weighted mean of $p_{Y|X}(y \mid x)$ with weights $p_X(x)$.

- For the reverse conditioning: $p_{XY}(x,y) = p_{X|Y}(x \mid y)p_Y(y)$

- The marginal frequency function is derived as the weighted mean of $p_{X|Y}(x \mid y)$ with weights $p_Y(y)$: $\boxed{p_X(x) = \sum_y p_{X|Y}(x \mid y)p_Y(y).}$

Discrete case

- Example: Let's consider the sample space of three times tossing a fair coin: $\{hhh, hht, hth, thh, htt, tht, tth, ttt\}$. Let X be the rv that the first toss is a head and let Y be the total number of heads.

- The joint frequency function of X and Y is:

	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

- Which is the conditional ff of X given $Y = 1$?

$$P(X=0|Y=1)=\frac{2}{3}$$

$$P(X=1|Y=1)=\frac{1}{3}$$

$$E(X|Y=1)=0\times\frac{2}{3}+1\times\frac{1}{3}=\frac{1}{3}$$

$$E(X^2|Y=1)=0^2\times\frac{2}{3}+1^2\times\frac{1}{3}=\frac{1}{3}$$

Outline

1. Conditional frequency function

2. Conditional Expectation

Conditional Expectation

- Let X discrete random variable with frequency function $p_X(x)$. The expectation of X is a weighted average of its possible values with weights $p_X(x)$:

$$E(X) = \sum_x x \cdot p_X(x).$$

- The expectation of X given the event that another variable Y takes the value y is called **conditional expectation of X given $Y = y$**:

$$\underline{E(X \mid Y = y) = \sum_x x \cdot p_{X|Y}(x \mid y)}$$

  and is derived as a weighted average of the possible values of X with updated weights $p_{X|Y}(x \mid y)$.

- For the continuous case, the summation is replaced by integration.

Properties of the Conditional Expectation

- Say we are interested in a function $h(X)$ of the random variable X:

$$\underbrace{E[h(X) \mid Y = y] = \sum_{x} h(x) \cdot p_{X|Y}(x \mid y)}$$

- Similarly for the continuous case.

Conditional Variance

- Let X discrete random variable with frequency function $p_X(x)$. The variance of X is:

$$Var(X) = \sum_x (x - \mu_x)^2 \cdot p_X(x) \quad \text{or} \quad Var(X) = E(X^2) - [E(X)]^2 .$$

- The **conditional variance of X given $Y = y$** is:

$$\underline{Var(X \mid Y = y) = \sum_x (x - \mu_x)^2 \cdot p_{X|Y}(x \mid y)}$$

or

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

Conditional Expectation and Variance

- Both quantities summarize the information given by the conditional ff.
- Let a survey on a number of households where the number of members and the number of cars they own is recorded.
  - Say we sample a household at random and learn the number of its members.
  - What would then be the expected number of cars they own?
  - The conditional expectation allows us to predict the number of cars using the available information.
- Computing $E(X)$ may not always be straightforward and conditional expectations can be used to break a complex problem into easier to compute pieces.

Lecture 8: Limit Theorems

Outline

1. Introduction

2. Chebyshev's Inequality

3. Law of Large Numbers

4. Central Limit Theorem

Outline

1. Introduction
2. Chebyshev’s Inequality
3. Law of Large Numbers
4. Central Limit Theorem

Introduction

- Bridge between Probability part and Statistics part of this course.

- Focus on mean of independent rv's $X_1, X_2, \ldots, X_n$

$$\underline{\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}}$$

$\Rightarrow$ important concept for statistics.

- To make statements about $\bar{X}_n$ e.g. $P(\bar{X}_n < 10)$ we will use the theorems:
    - Law of Large Numbers
    - Central Limit Theorem

Introduction

- Why we need the limiting theorems?
- $\boxed{\text{Example 1:}}$ 100 students in a class flip a fair coin, $P(X_i = H) = 0.5$. What is the probability that the proportion of heads is between 0.4 and 0.6?

  $\boxed{\text{Answer:}}$ Let $Y = \sum_{i=1}^{100} X_i$ the number of heads in 100 indepenent flips of the coin. Then Y follows Binomial with $n = 100$ and $p = 0.5$:

  $$P(0.4 \leq \frac{Y}{100} \leq 0.6) = P(40 \leq Y \leq 60) = \sum_{i=40}^{60} \binom{100}{i} 0.5^i (1-0.5)^{100-i} = 0.9648.$$

- The calculation is simple in this case because we know $Y \sim Binomial(100, 0.5)$.

Introduction

- Why we need the limiting theorems?

- $\boxed{\text{Example 2:}}$ Let us assume that in the Rabobank office in Rotterdam the waiting time in hours U_i for the clients on a Monday is $U_i \sim Uniform(0,1)$.

  What is the probability that the mean waiting time for the 100 clients that visited the bank on that day is more than 10 minutes, $P(\frac{\sum_{i=1}^{100} U_i}{100} \geq 1/6)$?

- The calculation is not that straightforward as in Example 1.

Outline

1 Introduction

2 Chebyshev’s Inequality

3 Law of Large Numbers

4 Central Limit Theorem

Chebyshev's Inequality

切比雪夫不等式

- Let X be a random variable with known $E(X)$ and $Var(X) = \sigma^2$ but **unknown distribution**.

- The probability that a random variable X differs from its mean by at least k standard deviations is less than or equal to $1/k^2$.

$$\boxed{P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}}.$$

Chebyshev's Inequality

- $\boxed{\text{Example:}}$ Let X the rv for the number of words in an article of a statistical journal. We know nothing about the distribution of X, but only that it has an average of 1000 words per article and standard deviation 200.

  What is the probability that the article has between 600 and 1400 words (i.e., within $k = 2$ standard deviations of the mean)?

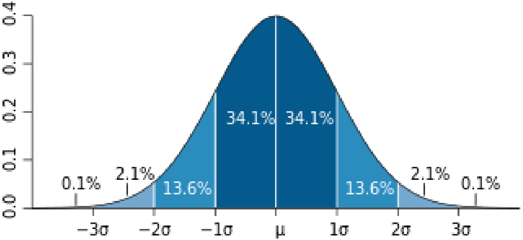
  $$P(600 \leq X \leq 1400) = P\left(-2 \leq \frac{X-1000}{200} \leq 2\right) = 1 - P\left(\left|\frac{X-1000}{200}\right| > 2\right) \geq 1 - \frac{1}{2^2} \geq \frac{3}{4}$$

  $\boxed{\text{Answer:}}$ From Chebyshev's inequality, this probability must more than 75% because there is less than 1/4 chance to be outside that range.

Chebyshev's Inequality

- For any data set, at least
  - 75% of the data lies within two standard deviations from the mean and
  - 88.9% of the data lies within three standard deviations from the mean
- For a bell shaped population distribution, approximately:
  - 68% lie within one standard deviation of the mean.
  - 95% lie within two standard deviations of the mean.
  - 99.7% lie within three standard deviation of the mean.

68 - 95 - 99.7 rule



Outline

1. Introduction

2. Chebyshev's Inequality

3. Law of Large Numbers

4. Central Limit Theorem

Law of Large numbers

- The Chebyshev inequality gives a crude bound on how high the probability is that the sample mean will be close to μ.

- Theorem: Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The sample mean will be close to μ if the sample size is sufficiently large.

$$n \to \infty \Rightarrow \bar{X} \to \mu$$

- **No knowledge of the probability distribution function of $\sum_i X_i$ is needed.**

$$E(\bar{X}) = \frac{\sum_{i=1}^{n} E(X)}{n} = \frac{n\mu}{n} = \mu$$

$$Var(\bar{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Law of Large numbers

- $\boxed{\text{R Example 1:}}$ Sample X_i, $i = 1, \ldots, n$ from a $Gamma(\alpha, \beta)$ with $\alpha = 0.5$, $\beta = 1$. It is known $E(X_i) = \alpha \cdot \beta$ and $Var(X_i) = \alpha \cdot \beta^2$. Compute the mean of X_i when $n = 10, 20, 50, 100, 500, 1000$. What do you observe?

- $\boxed{\text{R Example 2:}}$ Sample X_i, $i = 1, \ldots, n$ from a $N(0, \sigma^2)$ with $\sigma^2 = 3$. Compute the mean of X_i when $n = 10, 20, 50, 100, 500, 1000$. What do you observe?

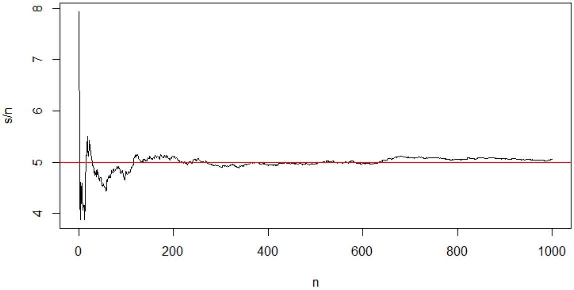
R implementation - Law of Large numbers

- Illustration

```
n <- 1:1000
x <- rchisq(1000, 5)
s <- cumsum(x)
plot(s/n, xlab="n", type="l")
abline(h = 5, col = "red")
```

R implementation - Law of Large numbers

- Illustration



Outline

1 Introduction

2 Chebyshev's Inequality

3 Law of Large Numbers

4 Central Limit Theorem

Central Limit Theorem (CLT)

- Theorem: Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables having a common distribution with mean μ and variance σ^2. We can compute probabilities for the sample mean using the cdf of the Normal distribution:

$$\boxed{\bar{X} \sim N(\mu, \sigma^2/n)}$$

- Note: **shape of the underlying population's distribution is not mentioned**

CLT

- How big does n have to be? It depends on the population distribution
  - if it is normal, then the CLT holds for small samples.
  - if it is not too unusual, then the CLT holds for samples of 30 or more.
  - if it is unusual (e.g., very long tails), then the CLT may require 100 or more observations.

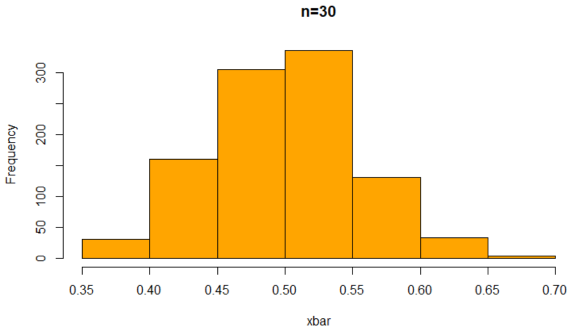
R implementation - CLT

- Use R to simulate 1000 times the sampling distribution of the mean, $\bar{X}$, of 30, and 1000 observations from the uniform distribution. Create a histogram and determine the mean and standard deviation of these simulations.

```
xbar <- numeric(1000)
for (i in 1:1000){x <- runif(30); xbar[i] <- mean(x)}
hist(xbar, col = "orange", main="n=30")
mean(xbar)
sd(xbar)
```

R implementation - CLT

- Illustration



R implementation - CLT

- Illustration

```
library(animation)
ani.options(interval = 0.5)
par(mar = c(3, 3, 1, 0.5), mgp = c(1.5, 0.5, 0), tcl = -0.3)
lambda = 4
f <- function(n) rpois(n, lambda)
clt.ani(FUN = f, mean = lambda, sd = lambda)
```

CLT

- $\boxed{\text{Example:}}$ The number of files stored in the home directory in the MSTAT department has mean μ = 7 and standard deviation σ = 5. If we check 50 employees in the department, what is the probability that $\bar{X}$ will be greater than 8?

$\bar{X} \sim N(7, \frac{5^2}{50})$

$\Rightarrow$ $\bar{X}$ is approximately normal with mean 7 and s.d. $5/\sqrt{(}50) = 0.707$

$\Rightarrow$ By CLT $P(\bar{X} > 8) = P(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > (8-7)/0.707) = P(Z > 1.41) = 0.0793.$

$$P(\bar{X}>8) = P\left(\frac{\bar{X}-7}{\frac{5}{\sqrt{50}}} > \frac{1}{\frac{5}{\sqrt{50}}}\right) = 1-\Phi\left(\frac{\sqrt{50}}{5}\right) = 1-\Phi(\sqrt{2}) = 0.0793$$

CLT

p $p(1-p)$

- A Binomial rv X with parameters n and p is the sum of independent Bernoulli rvs. Thus, we can assume

  → not $\bar{X}$

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

  where $\mu = n \cdot p$ and $\sigma^2 = n \cdot p \cdot (1-p)$. The approximation is best when $n \cdot p > 5$ and $n \cdot (1-p) > 5$.

- Example: Suppose that a coin is tossed 100 times. What is the probability that it lands heads up more than 60 times?

  Let X Binomial rv for the number of heads with $n = 100$ trials and probability of success $p = 0.5$. We know that $E(X) = n \cdot p = 50$ and $Var(X) = n \cdot p \cdot (1-p) = 25$. We can assume that $X \sim N(50, 25)$. Thus,

  $$\begin{aligned} P(X \geq 60) &= P\left(\frac{X-50}{5} \geq \frac{60-50}{5}\right) \\ &\approx 1 - \Phi(2) = 0.0228. \end{aligned}$$