Review of probability and statistics

ECON306 – Slides 1
Stock and Watson Ch. 2–3

Bruno Salcedo

The Pennsylvania State University

Summer 2013

[0]

Phenomena involving chance

- Rolling a dice to see which number comes out on top
- Tomorrow's weather
- The winner of the Superbowl
- The price of a ton of wheat on a given location at a given time
- The value of $\sin(\pi^2)$ (uncertainty)
- The exact temperature of a CPU (random number generators)

Randomness (frequentist)

- An experiment is an activity that:
  - Is performed with the intention of measuring the value of a variable
  - The environment can be replicated, so that the experiment can be repeated
- An experiment is said to be random when the outcome of each realization cannot be predicted
- The probability of an outcome is defined as the proportion of times that the outcome would result if the experiment were repeated infinitely many times

Uncertainty (Bayesian)

- An event is uncertain if its occurrence is unknown
  - Many conceivable worlds are consistent with our observations
  - We are often ignorant of many characteristics of the actual world
- People constantly make choices under uncertainty
  - A is more likely than B for you, if you would prefer vetting on A than betting on B
  - If enough comparisons are made, we can recover quantitative probabilities from subjective beliefs (Savage, 1954)
- The frequentist approach is objective
- The Bayesian approach is subjective

Probability (mathematical)

Definition

A probability space consists of states, events and probabilities:

- *A set of possible states of the world (or outcomes)*

$$\Omega = \{\omega_1, \omega_2, \dots \omega_m\}$$

- *An event is a set of states* $E \subseteq \Omega$
- *A probability function or measure* $\Pr$ *is a function that assigns a number* $\Pr(E)$ *to each event* E
- *Probability functions must satisfy:*
  - $0 \leq \Pr(E) \leq 1$
  - $\Pr(\Omega) = 1$ *and* $\Pr(\emptyset) = 0$
  - *If* $E \cap F = 0$ *then* $\Pr(E \cup F) = \Pr(E) + \Pr(F)$

Probability of an event

- Let $E = \{\omega_1, \omega_2, \ldots, \omega_k\} \subseteq \Omega$ be an event
- Notice that we can write E as the union of singleton events:

$$E = \{\omega_1\} \cup \{\omega_2\} \cup \ldots \cup \{\omega_k\}$$

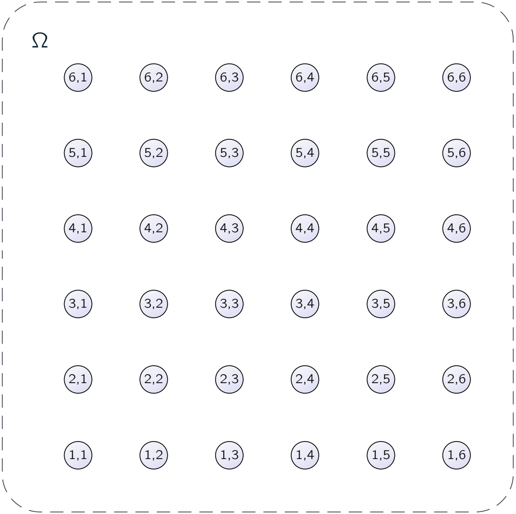
- Also notice that if $\omega_i \neq \omega_j$, then $\{\omega_i\} \cap \{\omega_j\} = \emptyset$
- The probability of E thus equals the sum of the probability of its elements

$$\begin{aligned} \Pr(E) &= \sum_{\omega \in E} \Pr(\{\omega\}) \\ &= \Pr(\{\omega_1\}) + \Pr(\{\omega_2\}) + \ldots + \Pr(\{\omega_k\}) \end{aligned}$$

- Implication: to specify a probability function Pr, it is sufficient to specify the probability of singleton events $\{\omega\}$

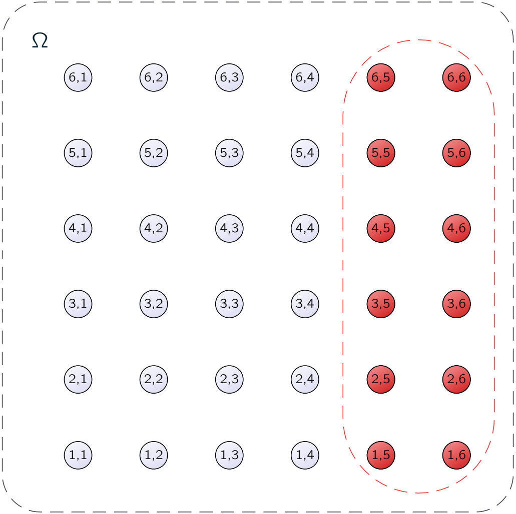
Example: Rolling two dice

State space Ω



Example: Rolling two dice

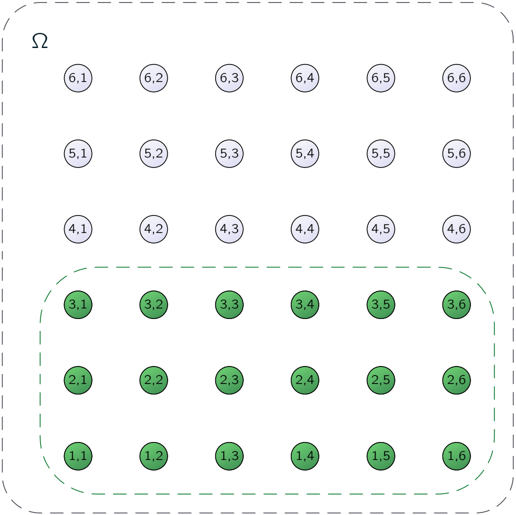
Events



$E_1 = \{\ \omega \mid \text{second dice is greater than 4}\ \}$

Example: Rolling two dice

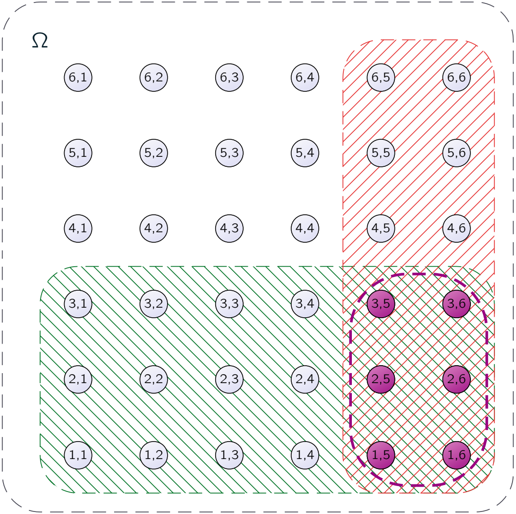
Events



$$E_2 = \{\, \omega \mid \text{first dice is less than } 4 \,\}$$

Example: Rolling two dice

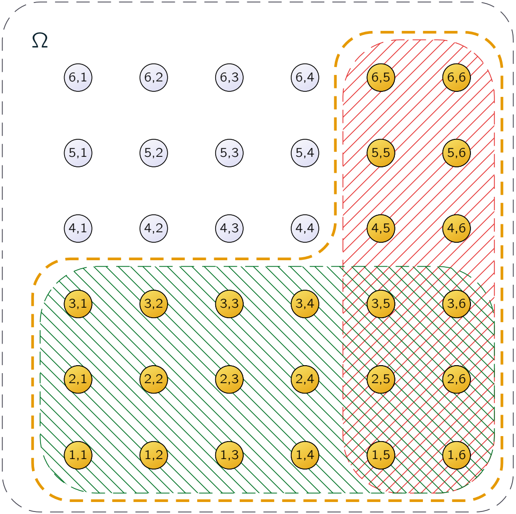
Events



$E_3 = E_1 \cap E_3 = \{\ \omega \mid$ first dice is less than 4 AND second dice is greater than 4 $\}$

Example: Rolling two dice

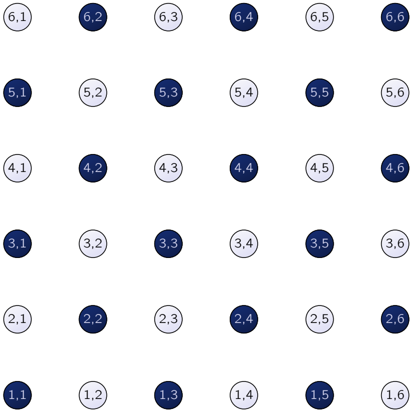
Events



$E_3 = E_1 \cup E_3 = \{\, \omega \mid \text{first dice is less than 4 OR second dice is greater than 4} \,\}$

Example: Rolling two dice

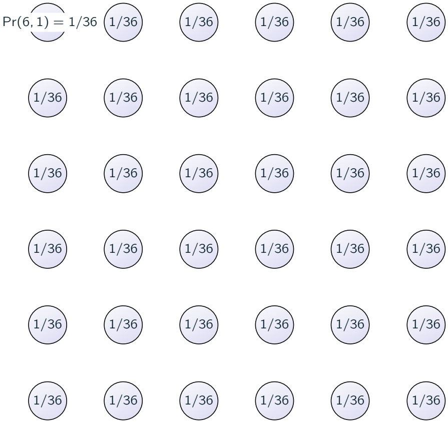
Events



$$E = \{\, \omega \mid \text{sum of the dice is odd} \,\}$$

Example: Rolling two dice

Uniform probability function $\Pr(\omega) = 1/36$



Example: Rolling two dice

Uniform probability function $\Pr(\omega) = 1/36$

$$\Pr(E_1) = \sum_{\omega \in E_1} \Pr(\omega) = \frac{1}{3}$$

Example: Rolling two dice

A different probability function Pr

4/81	4/81	4/81	4/81	4/81	16/81
1/81	1/81	1/81	1/81	1/81	4/81
1/81	1/81	1/81	1/81	1/81	4/81
1/81	1/81	1/81	1/81	1/81	4/81
1/81	1/81	1/81	1/81	1/81	4/81
1/81	1/81	1/81	1/81	1/81	4/81

Example: Rolling two dice

A different probability function Pr

$$\Pr(E_1) = \sum_{\omega \in E_1} \Pr(\omega) = \frac{45}{81} = \frac{5}{9} \approx 0.56$$

Random variables

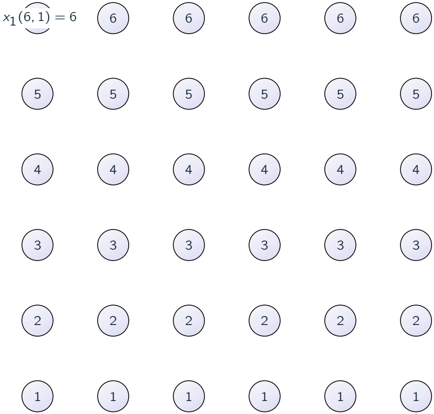
- The states of the world are abstract objects without many structure (e.g. a state of the world could be “Green Bay wins the Superbowl”)
- We can add structure by mapping the states of the world into mathematical objects, such as real numbers

Definition

Given a probability space, a random variable is a function $x : \Omega \to \mathbb{R}$ *that assigns a real number to each possible state of the world*

Example: Rolling two dice

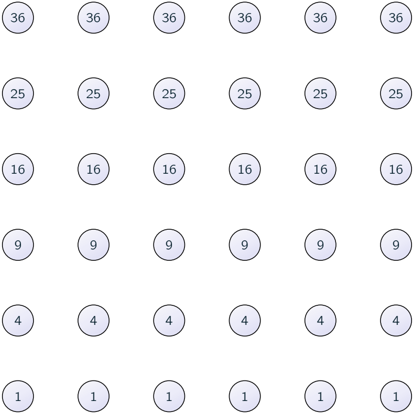
Random variable



$x_1 =$ number from the first dice

Example: Rolling two dice

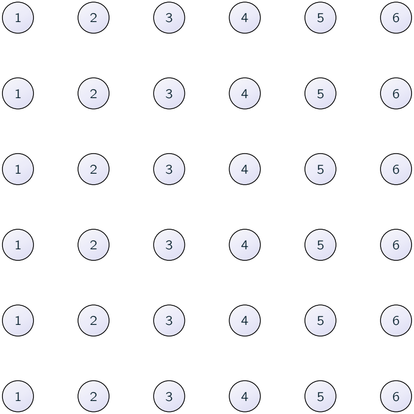
Random variable



x_1^2 = number from the first dice squared

Example: Rolling two dice

Random variable



x_2 = number from the second dice

Example: Rolling two dice

Random variable

7	8	9	10	11	12
6	7	8	9	10	11
5	6	7	8	9	10
4	5	6	7	8	9
3	4	5	6	7	8
2	3	4	5	6	7

$x_3 = \text{sum of the two dice} = x_1 + x_2$

Example: Rolling two dice

Random variable

7	3.5	1	0	3.3	12
6	7	e	9	10	11
3.6	6	13	2	11	10
1	π	42	$\sqrt{5}$	4	2.2
4	3	0	0	0	7.8
7.5	11	4	1	1	0

$x_4 =$ arbitrarily assigned

Random variables vs. states of the world

- If we where only interested in one random variable, we could identify outcomes with the corresponding values

- Random variables are useful because we can define different random variables in the same probability state
- This enables to ask questions about the relationship between different random variables

- For instance, it is clear that if we learn something about x_1, then we automatically learn something about x_1^2 and about $x_3 = x_1 + x_2$
- Furthermore, if we can somehow directly influence x_1, then we can indirectly influence x_1^2 and x_3

Support

- Informally, the support of a random variable is the set of possible values it can take
- For example
  - The support of x_1 and x_2 is $\{1, 2, 3, 4, 5, 6\}$
  - The support of x_1^2 is $\{1, 4, 9, 16, 25, 36\}$
  - The support of x_3 is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
- A random variable is discrete if its support is finite (or countable)
- A random variable is continuous otherwise

Distribution for discrete random variables

- The probability distribution of a random variable specifies the probability of each value in its support:

$$\Pr(x = \xi) = \Pr\left(\left\{\, \omega \;\middle|\; x(\omega) = \xi \,\right\}\right)$$

- The cumulative probability distribution of ξ is the probability that the value of x is less or equal than ξ:

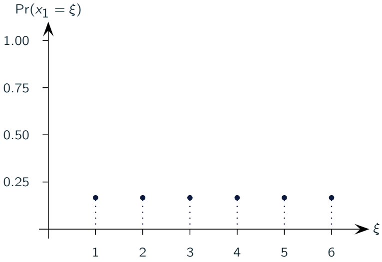
$$\begin{aligned} F(\xi) = \Pr(x \le \xi) &= \Pr\left(\left\{\, \omega \;\middle|\; x(\omega) \le \xi \,\right\}\right) \\ &= \sum_{\substack{\zeta \in X \\ \zeta \le \xi}} \Pr(x = \zeta) \end{aligned}$$

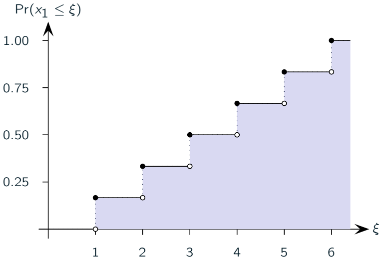
where X denotes the support of x

Example: Rolling two dice

Uniform probability distributions

ξ	1	2	3	4	5	6
$\Pr(x_1 = \xi)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_1 \leq \xi)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

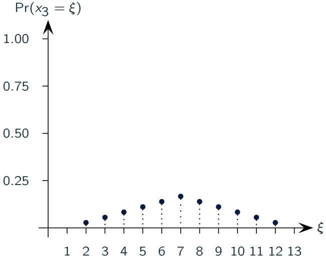


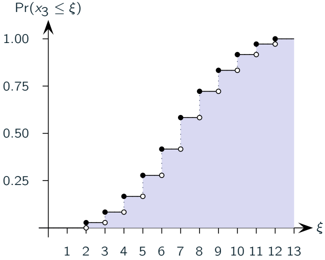


Example: Rolling two dice

Non-uniform probability distributions

ξ	2	3	5	5	6	7	8	9	10	11	12
$\Pr(x_3 = \xi)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$\Pr(x_3 \leq \xi)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1





Example: Bernouli distribution

- Suppose you flip a coin and x is the random variable which assigns 1 to heads and 0 to tail
- The probability function is described by a single paramenter p: the probability of a head ($p = 1/2$ for fair coins)
- The probability distribution of x is:

$$\Pr(x = 1) = p \qquad \text{and} \qquad \Pr(x = 0) = 1 - p$$

- The cumulative probability distribution is:

$$F(\xi) = \begin{cases} 0 & \text{if} \quad \xi < 0 \\ 1 - p & \text{if} \quad 0 \leq \xi < 1 \\ 1 & \text{if} \quad \xi \geq 1 \end{cases}$$

Probability for continuous random variables

- The cumulative distribution is defined as before

$$F(\xi) = \Pr(x \leq \xi)$$

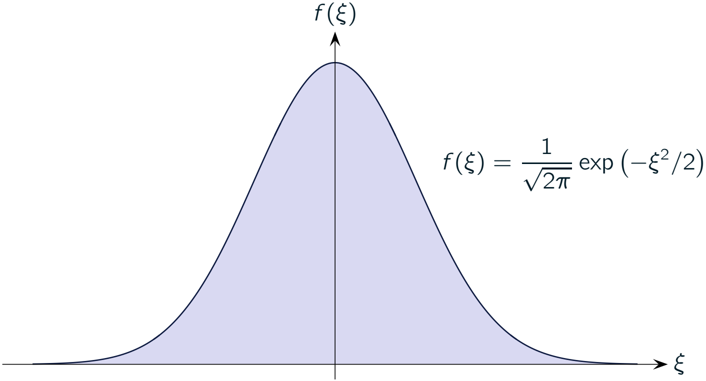
- The probability density $f(x)$ describes the rate at which probability is accumulated
- The Probability that x is between ξ_1 and ξ_2 corresponds to the area below the graph of f between ξ_1 and ξ_2

- Using calculus $f = dF/d\xi$ and:

$$\Pr(\xi_1 \leq x \leq \xi_2) = F(\xi_2) - F(\xi_1) = \int_{\xi_1}^{\xi_2} f(\zeta)\, d\zeta$$

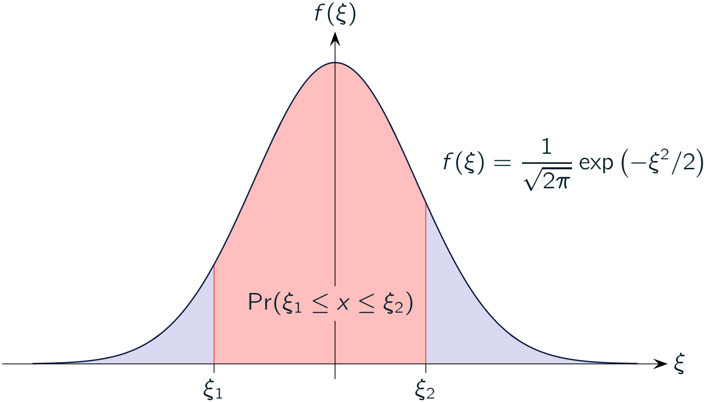
Example: Normal standard distribution

Density function



Example: Normal standard distribution

Area under the curve



Expected value

- The expected value, mean or expectation is a measure of centrality
- For discrete random variable it is given by:

$$\mu_x = \mathbb{E}[\, x \,] = \sum_{\xi \in X} \Pr(x = \xi) \cdot \xi$$

- The expected value of a function of a discrete random variable is:

$$\mathbb{E}[\, f(x) \,] = \sum_{\xi \in X} \Pr(x = \xi) \cdot f(\xi)$$

- The expected value is a linear operator meaning that:

$$\mathbb{E}[\, ax + by \,] = a\mathbb{E}[\, x \,] + b\mathbb{E}[\, y \,]$$

Variance

- The variance of a random variable is a measure of dispersion

- It is defined in terms of expectations:

$$\sigma_x^2 = \mathbb{V}[\,x\,] = \mathbb{E}\left[\,(x - \mu_x)^2\,\right]$$

- A useful way to compute variance is using the formula:

$$\sigma_x^2 = \mathbb{E}\left[\,x^2\,\right] - \mu_x^2$$

- The variance satisfies:

$$\mathbb{V}[\,x + y\,] = \mathbb{V}[\,x\,] + \mathbb{V}[\,y\,] + 2\mathbb{C}[\,x\,,\ y\,]$$
$$\mathbb{V}[\,ax\,] = a^2\mathbb{V}[\,x\,]$$

Example: Rolling two dice

Expectation and variance of x_1

ξ	1	2	3	4	5	6
$\Pr(x_1 = \xi)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_1 \leq \xi)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

$$\begin{aligned}\mathbb{E}[\,x_1\,] &= \Pr(x=1)\cdot 1 + \Pr(x=2)\cdot 2 + \ldots + \Pr(x=6)\cdot 6\\ &= \frac{1}{6}+\frac{2}{6}+\frac{3}{6}+\frac{4}{6}+\frac{5}{6}+\frac{6}{6} = \frac{21}{6} = 3.5\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\,x_1^2\,\right] &= \Pr(x=1)\cdot 1^2 + \Pr(x=2)\cdot 2^2 + \ldots + \Pr(x=6)\cdot 6^2\\ &= \frac{1}{6}+\frac{4}{6}+\frac{9}{6}+\frac{16}{6}+\frac{30}{6}+\frac{36}{6} = \frac{21}{6} = 16\end{aligned}$$

$$\mathbb{V}[\,x_1\,] = \mathbb{E}\left[\,x_1^2\,\right] - \left(\mathbb{E}[\,x_1\,]\right)^2 = 16 - 12.25 = 3.75$$

Example: Rolling two dice

Expectation and variance of x_3

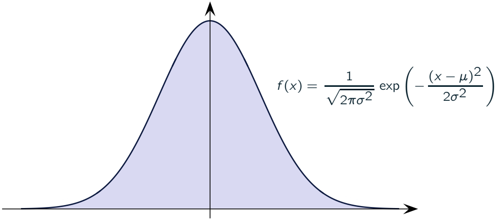
ξ	2	3	4	5	6	7	8	9	10	11	12
$\Pr(x_3 = \xi)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$\Pr(x_3 \leq \xi)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

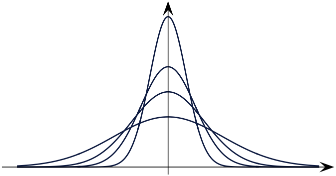
$$\mathbb{E}[\,x_3\,] = \frac{2}{36} + \frac{6}{36} + \frac{12}{36} + \frac{20}{36} + \ldots + \frac{12}{36} = \frac{252}{36} = 7$$

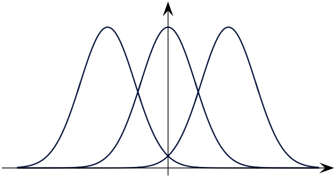
$$\mathbb{E}\left[\,x_3^2\,\right] = \frac{1}{36}\cdot 2^2 + \frac{2}{36}\cdot 3^2 + \frac{3}{36}\cdot 4^2 + \frac{4}{36}\cdot 5^2 + \ldots + \frac{1}{36}\cdot 12^2 = \frac{1974}{36} = 58.8\bar{3}$$

$$\mathbb{V}[\,x_3\,] = \mathbb{E}\left[\,x_3^2\,\right] - \big(\mathbb{E}[\,x_3\,]\big)^2 = 58.8\bar{3} - 49 = 9.8\bar{3}$$

Example: Normal distribution $N(\mu, \sigma^2)$




Changing σ^2


Changing μ

Conditional probability

- Knowing information about one event (or random variable) may convey information about other events (or random variables)
- e.g. if we know that $x_1 \geq 3$ then we know that $x_1^1 = 9$ and that $x_3 = x_1 + x_2 \geq 6$

Baye's rule

The conditional probability of E given F is:

$$\Pr(E|F) = \frac{\Pr(E \text{ and } F)}{\Pr(F)}$$

Independence

- Two events are independent if the occurrence of one of them does not affect the probability of the other

- In terms of conditional probability this means that:

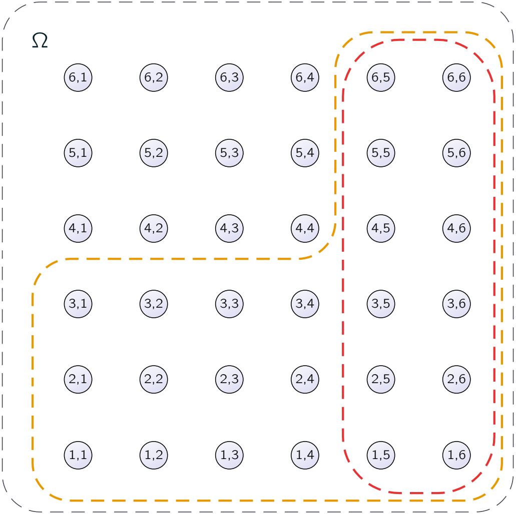
$$\Pr(E|F) = \Pr(E)$$

- Using Baye's rule, E and F are independent if and only if:

$$\Pr(E \text{ and } F) = \Pr(E) \cdot \Pr(F)$$

Example: Rolling two dice

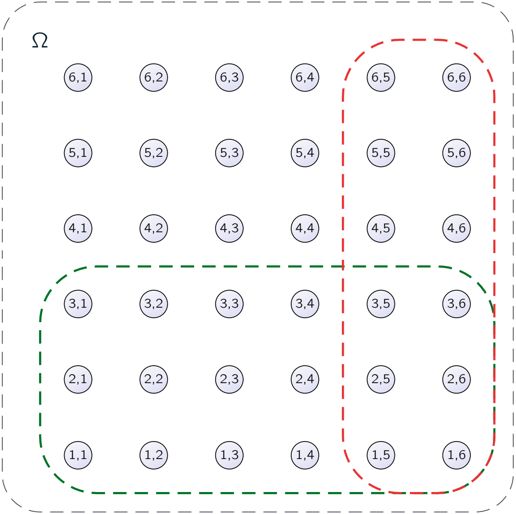
Conditional probabilities



$$\Pr(E_1|E_4) = \frac{\Pr(E_1 \cap E_4)}{\Pr(E_4)} = \frac{1/3}{2/3} = \frac{1}{2}$$

Example: Rolling two dice

Independent events



$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)} = \frac{1/6}{1/3} = \frac{1}{3} = \Pr(E_1)$$

Joint and marginal distributions

- Let x and y be random variables on the same probability space
- The joint distribution of x and y specifies for each pair of numbers ξ and ψ the probability:

$$\Pr(x = \xi \text{ and } y = \psi) = \Pr\left(\{\omega \mid x(\omega) = \xi \text{ and } y(\omega) = \psi\}\right)$$

- The marginal distributions $\Pr(x = \xi)$ and $\Pr(y = \psi)$ can be obtained from the joint:

$$\Pr(y = \psi) = \sum_{\xi \in X} \Pr(x = \xi \text{ and } y = \psi)$$

- The converse is false: the joint distribution cannot be obtained from the marginals

Example: Rolling two dice

Joint and marginal distributions

	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$	$x_1 = 5$	$x_1 = 6$	
$x_3 = 2$	1/36	0	0	0	0	0	1/36
$x_3 = 3$	1/36	1/36	0	0	0	0	2/36
$x_3 = 4$	1/36	1/36	1/36	0	0	0	3/36
$x_3 = 5$	1/36	1/36	1/36	1/36	0	0	4/36
$x_3 = 6$	1/36	1/36	1/36	1/36	1/36	0	5/36
$x_3 = 7$	1/36	1/36	1/36	1/36	1/36	1/36	6/36
$x_3 = 8$	0	1/36	1/36	1/36	1/36	1/36	5/36
$x_3 = 9$	0	0	1/36	1/36	1/36	1/36	4/36
$x_3 = 10$	0	0	0	1/36	1/36	1/36	3/36
$x_3 = 11$	0	0	0	0	1/36	1/36	2/36
$x_3 = 12$	0	0	0	0	0	1/36	1/36
	1/6	1/6	1/6	1/6	1/6	1/6	

Example: Rolling two dice

Joint and marginal distributions

	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$x_2 = 5$	$x_2 = 6$	
$x_1 = 1$	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$x_1 = 2$	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$x_1 = 3$	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$x_1 = 4$	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$x_1 = 5$	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$x_1 = 6$	1/36	1/36	1/36	1/36	1/36	1/36	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

Example: Rolling two dice

Same marginals different joint

	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$	$x_1 = 5$	$x_1 = 6$	
$x_1 = 1$	1/6	0	0	0	0	0	1/6
$x_1 = 2$	0	1/6	0	0	0	0	1/6
$x_1 = 3$	0	0	1/6	0	0	0	1/6
$x_1 = 4$	0	0	0	1/6	0	0	1/6
$x_1 = 5$	0	0	0	0	1/6	0	1/6
$x_1 = 6$	0	0	0	0	0	1/6	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

Example

Same marginals different joint

	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	
$x_1 = 1$	1/12	0	0	0	0	1/12	1/6
$x_1 = 2$	1/12	1/12	0	0	0	0	1/6
$x_1 = 3$	0	1/12	1/12	0	0	0	1/6
$x_1 = 4$	0	0	1/12	1/12	0	0	1/6
$x_1 = 5$	0	0	0	1/12	1/12	0	1/6
$x_1 = 6$	0	0	0	0	1/12	1/12	1/6
	1/6	1/6	1/6	1/6	1/6	1/6	

Conditional distributions and independence

- Just like conditional probabilities, we can define conditional distributions as:

$$\Pr(x = \xi | y = \psi) = \frac{\Pr(x = \xi \text{ and } y = \psi)}{\Pr(y = \psi)}$$

- And say that x and y are independent if:

$$\Pr(x = \xi | y = \psi) = \Pr(x = \xi)$$

- Independence is equivalent to requiring that the joint distribution equals the product of the marginals:

$$\Pr(x = \xi \text{ and } y = \psi) = \Pr(x = \xi) \cdot \Pr(y = \psi)$$

Example: Rolling two dice

Conditional distributions

ξ	2	3	5	5	6	7	8	9	10	11	12
$\Pr(x_3 = \xi)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$\Pr(x_3 = \xi \vert x_1 = 1)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0	0
$\Pr(x_3 = \xi \vert x_1 = 2)$	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0
$\Pr(x_3 = \xi \vert x_1 = 3)$	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0
$\Pr(x_3 = \xi \vert x_1 = 4)$	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0
$\Pr(x_3 = \xi \vert x_1 = 5)$	0	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0
$\Pr(x_3 = \xi \vert x_1 = 6)$	0	0	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

x_1 and x_3 are NOT independent

Example: Rolling two dice

Conditional distributions

ξ	1	4	9	16	25	36
$\Pr(x_1^2 = \xi)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_1^2 = \xi \mid x_1 = 1)$	1	0	0	0	0	0
$\Pr(x_1^2 = \xi \mid x_1 = 2)$	0	1	0	0	0	0
$\Pr(x_1^2 = \xi \mid x_1 = 3)$	0	0	1	0	0	0
$\Pr(x_1^2 = \xi \mid x_1 = 4)$	0	0	0	1	0	0
$\Pr(x_1^2 = \xi \mid x_1 = 5)$	0	0	0	0	1	0
$\Pr(x_1^2 = \xi \mid x_1 = 6)$	0	0	0	0	0	1

x_1 and x_1^2 are NOT independent

Example: Rolling two dice

Conditional distributions

ξ	1	2	3	4	5	6
$\Pr(x_2 = \xi)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_2 = \xi \mid x_1 = 1)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_2 = \xi \mid x_1 = 2)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_2 = \xi \mid x_1 = 3)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_2 = \xi \mid x_1 = 4)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_2 = \xi \mid x_1 = 5)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
$\Pr(x_2 = \xi \mid x_1 = 6)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

x_1 and x_2 are independent

Covariance

- The covariance of a random variable is a measure of the linear association between them
- It is defined in terms of expectations:

$$\sigma_{xy} = \mathbb{C}[\,x\,,\,y\,] = \mathbb{E}[\,(x-\mu_x)(y-\mu_y)\,]$$

- A useful way to compute covariance is using the formula:

$$\sigma_{xy} = \mathbb{E}[\,xy\,] - \mu_x\mu_y$$

- For every random variable $\mathbb{C}[\,x\,,\,x\,] = \sigma_x^2$
- For every two random variables $\mathbb{C}[\,x\,,\,y\,] = \mathbb{C}[\,y\,,\,x\,]$

Correlation

- The correlation is a normalization of the covariance:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- The correlation always is between -1 and 1
- x and y are uncorrelated if $\rho_{xy} = 0$

- Independence implies uncorrelation
- Uncorrelation does NOT imply independence
- Correlation only measures linear association
- Correlation means linear association, NOT slope
- Correlation does NOT imply a causal relation

Example: Rolling two dice

Variance-covariance matrix

	x_1 [3.00]	x_2 [3.00]	x_1^2 [153.40]	x_3 [6.00]	x_4 [53.93]
x_1	0.333	0.000	0.046	0.167	0.008
x_2	0.000	0.333	0.000	0.167	0.001
x_1^2	0.046	0.000	0.007	0.023	0.001
x_3	0.167	0.167	0.023	0.167	0.005
x_4	0.008	0.001	0.001	0.005	0.019

Example: Rolling two dice

Correlation matrix

	x_1 [3.00]	x_2 [3.00]	x_1^2 [153.40]	x_3 [6.00]	x_4 [53.93]
x_1	1.000	0.000	0.979	0.707	0.107
x_2	0.000	1.000	0.000	0.707	0.009
x_1^2	0.979	0.000	1.000	0.692	0.062
x_3	0.707	0.707	0.692	1.000	0.082
x_4	0.107	0.009	0.062	0.082	1.000

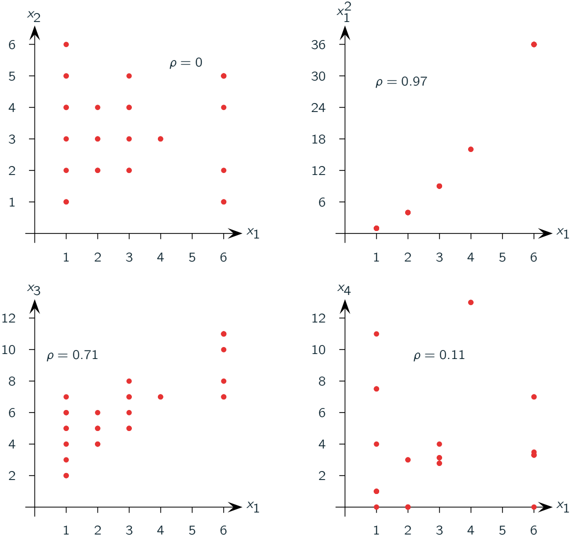
Example: Rolling two dice

Random sample

	x_1	x_2	x_1^2	x_3	x_4
1	4	3	16	7	13
2	1	2	1	3	11
3	4	3	16	7	13
4	2	2	4	4	3
5	1	5	1	6	1
6	6	5	36	11	3.3
7	6	1	36	7	7
8	2	4	4	6	0
9	1	4	1	5	1
10	6	5	36	11	3.3
11	3	5	9	8	4
12	1	1	1	2	7.5
13	1	4	1	5	1
14	3	4	9	7	e
15	6	5	36	11	3.3
16	6	2	36	8	3.5
17	2	2	4	4	3
18	6	1	36	7	7
19	1	5	1	6	1
20	3	3	9	6	42
21	3	4	9	7	e
22	2	3	4	5	0
23	3	2	9	5	π
24	1	3	1	4	4
25	6	4	36	10	0
26	3	2	9	5	π
27	6	5	36	11	3.3
28	1	6	1	7	0
29	2	3	4	5	0
30	1	1	1	2	7.5

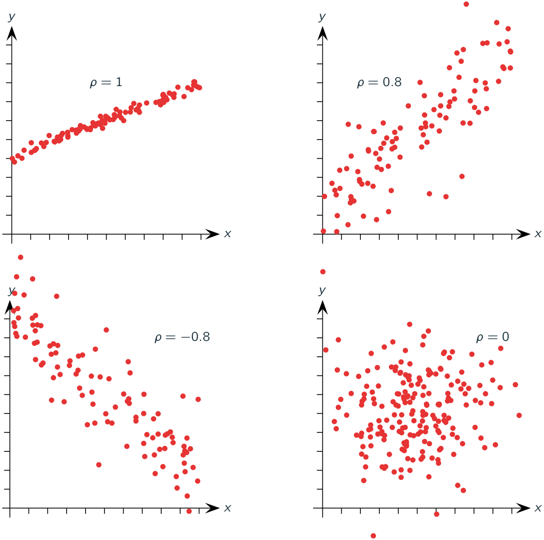
Example: Rolling two dice

Scatterplots



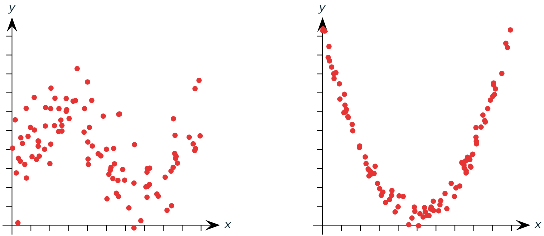
Example: Scatterplots

Correlation



Example: Scatterplots

Correlation and independence



[0]

Statistical science

- Statistics is the science of using data to learn about the world around us
- The main three objectives of statistics are:
  - **Estimation** – quantifying relations between different variables
  - **Inference** – testing whether theoretical relations hold in real life
  - **Forecasting** – predicting the future realizations of variables
- For data to be useful we need to make assumptions on the data-generating process

Random samples

Definition

A *random sample* *is a sequence* $\{x_1, x_2, \ldots, x_n\}$ *of mutually independent and identically distributed (i.i.d.) random variables*

- Mutual independence is more than pairwise independence
  - It is not sufficient that x_i and x_j are independent for all i and j
  - The entire joint distribution should equal the product of the marginal distributions
- Some examples of random samples
  - The sequence of outcomes from repeating a random experiment
  - The characteristics of different objects of a population selected randomly
- Most of the course we will assume that datasets are realizations of random samples

Statistics

Definition

A statistic is a function that maps each possible outcome of a random sample to a real number

- Statistics are random variables (being functions of random variables)
- The probability distribution of a statistic is called the sampling distribution
- Most of the theory of Statistics is based on understanding sampling distributions

Some commonly used statistics

- The ample mean $\bar{x}$ is the average value in the sample:

$$\bar{x} = \mathbb{E}_n\left[\, x_i \,\right] = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- If the random variables x_i are normally distributed, then $\bar{x}$ has a Student t-distribution

- The sample variance s_x^2 is the average deviation from the sample mean

$$s_x^2 = \mathbb{V}_n\left[\, x_i \,\right] = \mathbb{E}_n\left[\, (x_i - \bar{x})^2 \,\right] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\, \bar{x}\,(1 - \bar{x})$$

- The jth order statistics $x_{(j)}$ is the jth highest value in the sample, e.g.:

$$x_{(1)} = \max\{x_1, x_2, \ldots, x_n\}$$

Example: Rolling two dice

Statistics

	x_1	x_2	x_1^2	x_3	x_4
1	4	3	16	7	13
2	1	2	1	3	11
3	4	3	16	7	13
4	2	2	4	4	3
5	1	5	1	6	1
6	6	5	36	11	3.3
7	6	1	36	7	7
8	2	4	4	6	0
9	1	4	1	5	1
10	6	5	36	11	3.3
11	3	5	9	8	4
12	1	1	1	2	7.5
13	1	4	1	5	1
14	3	4	9	7	e
15	6	5	36	11	3.3
16	6	2	36	8	3.5
17	2	2	4	4	3
18	6	1	36	7	7
19	1	5	1	6	1
20	3	3	9	6	42
21	3	4	9	7	e
22	2	3	4	5	0
23	3	2	9	5	π
24	1	3	1	4	4
25	6	4	36	10	0
Sample mean	3.20	3.28	14.08	6.48	5.62
Sample variance	4.00	1.88	212.66	5.84	71.28
Maximum	6	5	36	11	42

Example: Rolling two dice

Statistics (different sample)

	x_1	x_2	x_1^2	x_3	x_4
1	4	2	16	6	6
2	1	6	1	7	0
3	4	3	16	7	13
4	3	3	9	6	42
5	1	1	1	2	7.5
6	5	6	25	11	11
7	2	6	4	8	7.8
8	1	3	1	4	4
9	3	3	9	6	42
10	5	3	25	8	e
11	3	5	9	8	4
12	4	2	16	6	6
13	6	6	36	12	12
14	6	2	36	8	3.5
15	1	3	1	4	4
16	4	1	16	5	3.6
17	5	3	25	8	e
18	6	2	36	8	3.5
19	4	1	16	5	3.6
20	3	3	9	6	42
21	5	4	25	9	9
22	5	4	25	9	9
23	4	6	16	10	10
24	3	5	9	8	4
25	1	6	1	7	0
Sample mean	3.56	3.56	15.32	7.12	10.12
Sample variance	2.76	3.01	129.48	5.03	156.06
Maximum	6	6	36	12	42

Empirical distribution

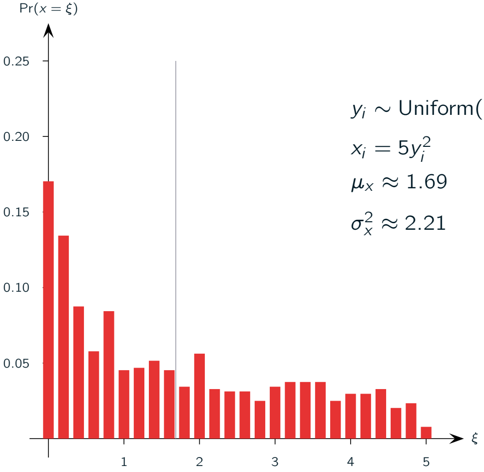
- A useful statistic is the sample relative frequency of each value in the support:

$$f_n(\xi) = \frac{\#\{x_i \mid x_i = \xi\}}{n}$$

- The collection of such statistics $\{f_n(\xi)\}$ constitutes a probability distribution, its called the empirical distribution
- Empirical distributions are usually represented using histograms

Example: A strange random variable

Empirical distribution



$y_i \sim \text{Uniform}(0, 1)$

$x_i = 5y_i^2$

$\mu_x \approx 1.69$

$\sigma_x^2 \approx 2.21$

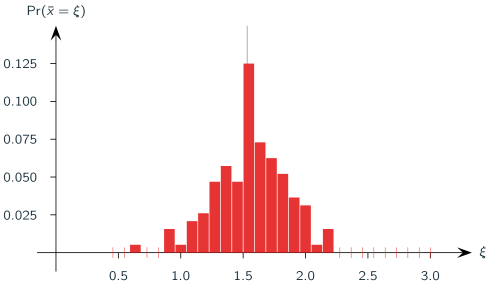
Approximating the sample distribution

- Except for rare cases (e.g. normal distributions), sample distributions are difficult to obtain
- Sometimes they can be approximated using simulation methods (bootstrap)
- Another approach for "large samples" is to approximate using asymptotic distributions
- It is much easier to determine what happens to the sampling distribution when n becomes large

Example: A strange random variable

Bootstrap

- The distribution of x_i is hard to obtain, the distribution of $\bar{x}$ is harder
- Simulate realizations of a random sample $\{x_1, \ldots, x_n\}$ with $n = 25$, and compute $\bar{x}$
- Repeat this process 120 times to generate an empirical distribution for $\bar{x}$



Estimators

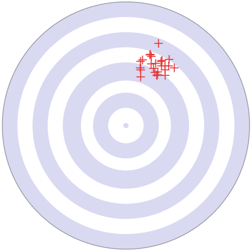
- We are often interested in the quantitative values of unknown parameters
- e.g. the mean number of defective products, the price elasticity of demand, the gravitational acceleration
- We want generate "good" estimates from the data

- An estimator estimator for a parameter is a statistic that is used as a proxy for it's true value
- The realized value of an estimator is called an estimate

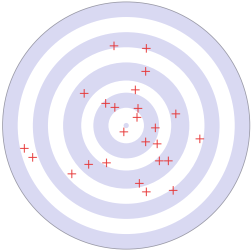
- Typically we use hats or Latin letters to denote estimators, e.g. $\hat{\varepsilon}$ and e for estimators of ε

Desirable properties

We want our estimators to be both as accurate and as precise as possible



Precise but inaccurate



Accurate but imprecise

Desirable properties

- **Accuracy**
  - Let θ be an unknown parameter and $\hat{\theta}$ an estimator
  - The bias of $\hat{\theta}$ is defined as $\mathbb{E}\left[\hat{\theta}\right] - \theta$
  - An estimator is unbiased if it has no bias, i.e. if $\mathbb{E}\left[\hat{\theta}\right] = \theta$
- **Precision**
  - The variance of an estimator can be used as a measure of precision
  - Given two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$, we say that $\hat{\theta}$ is more efficient than $\tilde{\theta}$ if $\mathbb{V}\left[\hat{\theta}\right] \leq \mathbb{V}\left[\tilde{\theta}\right]$
  - An unbiased estimator is efficient if is more efficient than any other unbiased estimator

The sample mean

- The sample mean is an unbiased estimator of the mean:

$$\mathbb{E}[\,\bar{x}_n\,] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\,x_i\,] = \frac{1}{n}\sum_{i=1}^{n}\mu_x = \mu_x$$

- Its variance is given by:

$$\mathbb{V}[\,\bar{x}_n\,] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[\,x_i\,] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_x^2 = \frac{1}{n}\sigma_x^2$$

- The sample mean is the best linear unbiased estimator (BLUE) of the mean

Consistency

- Instead of focusing finite sample properties we can ask for asymptotic results
- Can we guarantee that our estimator will be both precise and accurate if the sample is large enough?

- An estimator is consistent if it converges in probability to the true value, we denote this as:

$$\hat{\theta} \xrightarrow[p]{} \theta$$

- It means that the probability that $\hat{\theta}$ is far away from θ becomes arbitrarily small as the size of the sample increases

- Asymptotic efficiency is also important, but people focus on rate of convergence instead

Law of large numbers

- The law of large number essentially states that the sample mean is a consistent estimator of the mean

Law of large numbers

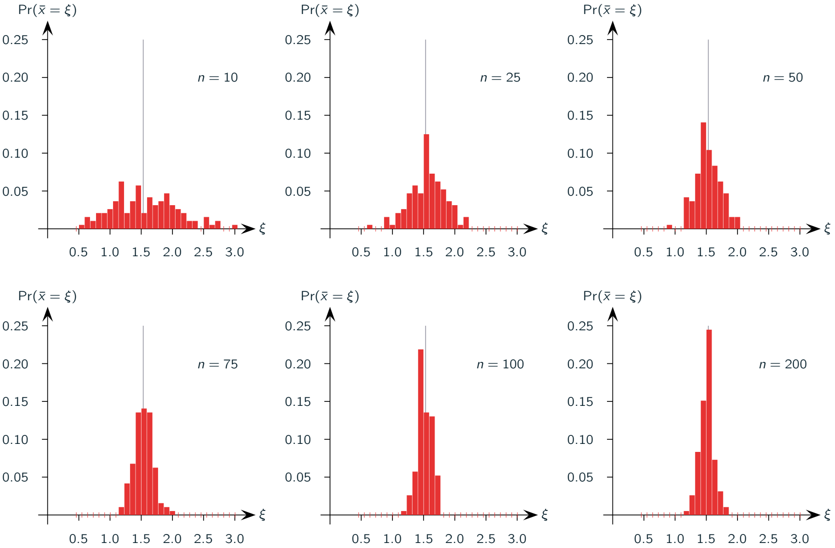
Given a random sample and a function h finite moments:

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \underset{p}{\longrightarrow} \mathbb{E}[\,h(x)\,]$$

- The intuition here is that we can always write $x_i = \mu_x + \varepsilon$
- Every realization contains μ_x, and different errors cancel out
- For frequentist statistics, the LLN can be understood as the definition of randomness itself

Example: A strange random variable

Law of large numbers



Hypotheses

- The second application of statistics is to test whether hypothetical assertions hold in real life

- In order to test an hypothesis we need to specify a counterfactual alternative
- The conjectured hypothesis is called the null hypothesis $\mathscr{H}_0$
- The counterfactual is called the alternative hypothesis $\mathscr{H}_1$

- We will only consider null hypothesis of the form $\theta = \theta_0$ or $\theta \geq \theta_0$ for some unknown parameter θ and some number θ_0
- Similar methodologies can be used to test much more complicated hypothesis

Testing

- A test is a rule to decide whether to reject or not reject an hypothesis based on the realized data
- A test can be thought of as a statistic that takes the values 1 (for not reject) and 0 (for reject)
- Not rejecting does NOT mean accepting
  - It means that there is not sufficient evidence to disprove the hypothesis
  - It does not mean that there is enough evidence to prove it
- Most tests use a test statistic t, and an acceptance region C
- The rule is to accept if and only if the realized value of t lies in C

Methodology

- Suppose that we want to test the hypothesis

$$\mathscr{H}_0: \ \theta = \theta_0 \quad \text{vs.} \quad \mathscr{H}_1: \ \theta \neq \theta_0$$

  and we have a consistent estimator $\hat{\theta}$

- The logic is to ask, if it where true that $\theta = \theta_0$, then what would be the probability of observing the actual/realized sample?

- We know that $\hat{\theta}$ will be close to θ
- Under the null hypothesis $\hat{\theta}$ should be close to θ_0
- Hence we can reject the null hypothesis if the distance $|\hat{\theta} - \theta_0|$ is large enough

p-vaule

- What does ‘large enough’ mean?

- The p-value is the probability, under $\mathscr{H}_0$, of drawing a test statistic at least as adverse to $\mathscr{H}_0$ as the realized one
- In our example, it is the probability that

$$\left|\hat{\theta} - \theta_0\right| \geq \left|\hat{\theta}^{\text{ac}} - \theta_0\right|$$

  where $\hat{\theta}^{\text{ac}}$ is the actual realized value of $\hat{\theta}$

- The p-value is a statistic measuring how likely is the realized sample under $\mathscr{H}_0$
- It quantifies ‘large enough’ in terms of probabilities
- We still have to choose a threshold probability to reject $\mathscr{H}_0$

Choosing a significance level

- There are two things that can go wrong in testing an hypothesis:
  - **Type I Error** – Rejecting a true hypothesis
  - **Type II Error** – Not rejecting a false hypothesis
- There is a trade-off between type I and type II errors
  - **Significance** – Probability of type I error under $\mathscr{H}_0$
  - **Power** – Probability of not committing type II error under $\mathscr{H}_1$
- Usually: choose significance and then maximize power
- This approach requires knowing the sampling distribution of the test statistic

Central limit theorem

- Under very mild conditions, the asymptotic distribution of $\bar{x}$ is normal independently of the distribution of x

Central limit theorem

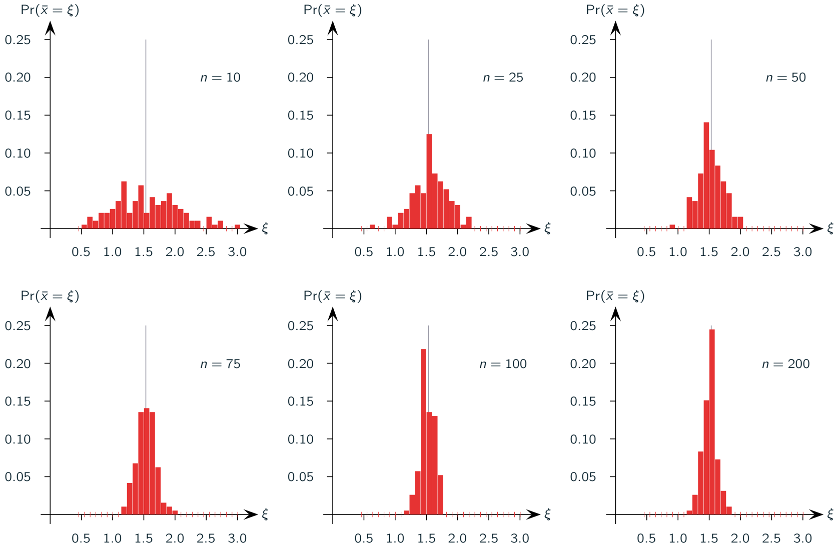
For any random sample with finite moments, the distribution of $\sqrt{n}\,\bar{x}$ approaches $N(\mu_x, \sigma_x^2)$ when n becomes large

- If we normalize $z = (x - \mu_x)/\sigma_x$ then the distribution of $\sqrt{n}\,\bar{z}$ approaches $N(0, 1)$

- This is an amazing result that is not intuitive at all
- Why should $\exp^{-x/2}/\sqrt{2\pi}$ be the function that describes any data generating process?
- It is extremely powerful because it means that we do not need to make parametric assumptions when we have large samples!

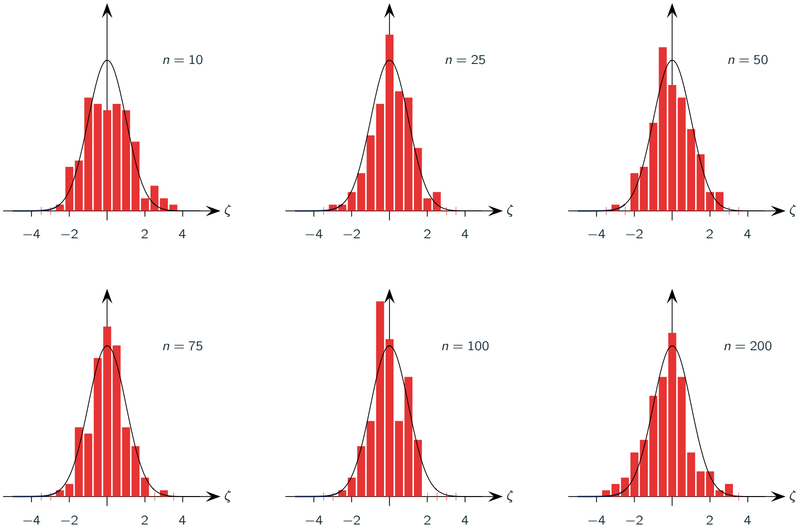
Example: A strange random variable

Law of large numbers



Example: A strange random variable

Central limit theorem



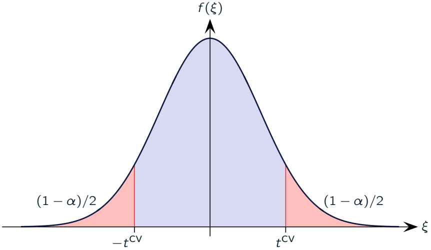
Hypothesis about the mean

$$\mathscr{H}_0: \ \mu_x = m \quad \text{vs.} \quad \mathscr{H}_1: \ \mu_x \neq m$$

$$t = \sqrt{n}\left(\frac{\bar{x}^{\text{ac}} - m}{\sigma_x}\right) \qquad \text{or} \qquad t = \sqrt{n}\left(\frac{\bar{x}^{\text{ac}} - m}{s_x}\right)$$

- Under $\mathscr{H}_0$ the asymptotic distribution of t is $N(0,1)$
- A test of significance α is to reject $\mathscr{H}_0$ if:

$$|t| > t^{\text{cv}} = \Phi^{-1}\big((1-\alpha)/2\big)$$



One sided hypothesis

$$\mathscr{H}_0:\ \mu_x \leq m \quad \text{vs.} \quad \mathscr{H}_1:\ \mu_x > m$$

$$t = \sqrt{n}\left(\frac{\bar{x}^{\text{ac}} - m}{\sigma_x}\right) \qquad \text{or} \qquad t = \sqrt{n}\left(\frac{\bar{x}^{\text{ac}} - m}{s_x}\right)$$

- Under $\mathscr{H}_0$ the asymptotic distribution of t is $N(0,1)$
- A test of significance α is to reject $\mathscr{H}_0$ if:

$$t > t^{\text{cv}} = \Phi^{-1}(\alpha)$$