Introduction to intelligent systems

Statistics

Mikkel N. Schmidt

Technical University of Denmark,
DTU Compute, Department of Applied Mathematics and Computer Science.

Overview

1. Descriptive statistics
2. Probability distributions
3. Population and sample
4. Central limit theorem
5. Standard error of the mean
6. Confidence intervals
7. Tasks

Feedback group

- Magnus Alexander Mollatt van Capel
- Rasmus Johansen Rieneck
- Christian Rahbæk Warburg
- Clara Louise Brodt

Learning objectives

I Descriptive and inferential statistics: Population, sample, statistic, parameters, estimator.

II Population mean and standard deviation.

II Sample estimate of mean and standard deviation.

II Confidence interval for mean and proportion.

II Sample size estimate for mean an proportion.

I Understand the concepts and definitions, and know their application. Reason about the concepts in the context of an example. Use correct technical terminology.

II As above plus: Read, manipulate, and work with technical definitions and expressions (mathematical and Python code). Carry out practical computations. Interpret and evaluate results.

Descriptive statistics

Mean and variance of a population

Population mean Average value of the population

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Population variance Average squared distance to the mean

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

Population standard deviation Square root of variance

Exercise: Population mean and variance

Consider a population of $N = 3$ observations.

$$x = \{1, 4, 10\}$$

- What is the population mean μ_x and variance σ_x^2?

Definitions

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)^2$$

Exercise: Population mean and variance

Consider a population of $N = 3$ observations.

$$x = \{1, 4, 10\}$$

- What is the population mean μ_x and variance σ_x^2?

Definitions

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

Solution

$$\mu_x = \frac{1}{3}(1 + 4 + 10) = 5$$

$$\sigma_x^2 = \frac{1}{3}\left((1-5)^2 + (4-5)^2 + (10-5)^2\right) = 14$$

Probability distributions

Probability distributions

Probability distribution

- Mathematical function that gives the probability of each possible outcome in an experiment
- Describes a random phenomenon in terms of probabilities of events
- Can be used as an infinite population

Probability distributions

Probability distribution

- Mathematical function that gives the probability of each possible outcome in an experiment
- Describes a random phenomenon in terms of probabilities of events
- Can be used as an infinite population

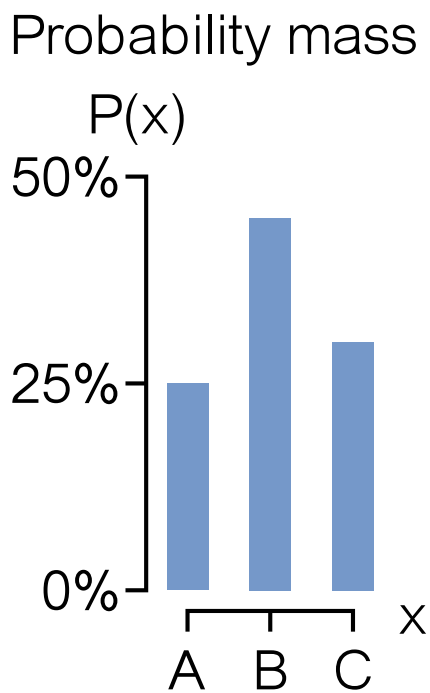
Discrete outcomes

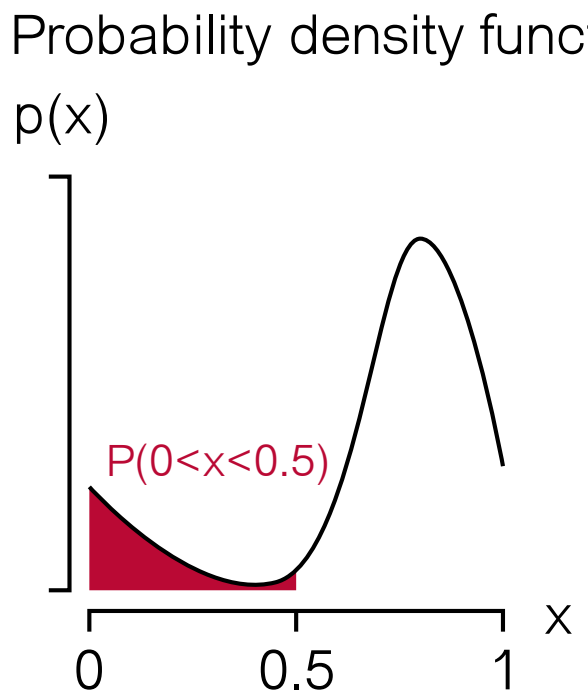
- Probability mass function provides probability of each outcome.

Continuous outcomes

- Probability density function (PDF) provides *relative* probability of each outcome.
- The probability of any particular outcome is zero.
- The probability of an outcome within some range is the area under the PDF curve.

Probability distributions





Mean and variance of a population / distribution

Finite population

- Sum over all values in population
- Weigh each equally by $1/N$

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

Mean and variance of a population / distribution

Finite population

- Sum over all values in population
- Weigh each equally by $1/N$

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

Discrete distribution

- Sum over all K possible outcomes
- Weigh each by their probability

$$\mu_x = \sum_{k=1}^{K} P(x_k) \cdot x_k$$

$$\sigma_x^2 = \sum_{i=1}^{K} P(x_k) \cdot (x_k - \mu_x)^2$$

Mean and variance of a population / distribution

Finite population

- Sum over all values in population
- Weigh each equally by $1/N$

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

Discrete distribution

- Sum over all K possible outcomes
- Weigh each by their probability

$$\mu_x = \sum_{k=1}^{K} P(x_k) \cdot x_k$$

$$\sigma_x^2 = \sum_{i=1}^{K} P(x_k) \cdot (x_k - \mu_x)^2$$

Continuous distribution

- Integral over all possible outcomes
- Weigh each by their probability density

$$\mu_x = \int_{-\infty}^{\infty} p(x) \cdot x \cdot \mathrm{d}x$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} p(x) \cdot (x - \mu_x)^2 \cdot \mathrm{d}x$$

Population and sample

Population and sample

Population Entire set of entities under study

- Can be a hypothetical population
- Typically impossible to survey/measure entire population
- May be infinite

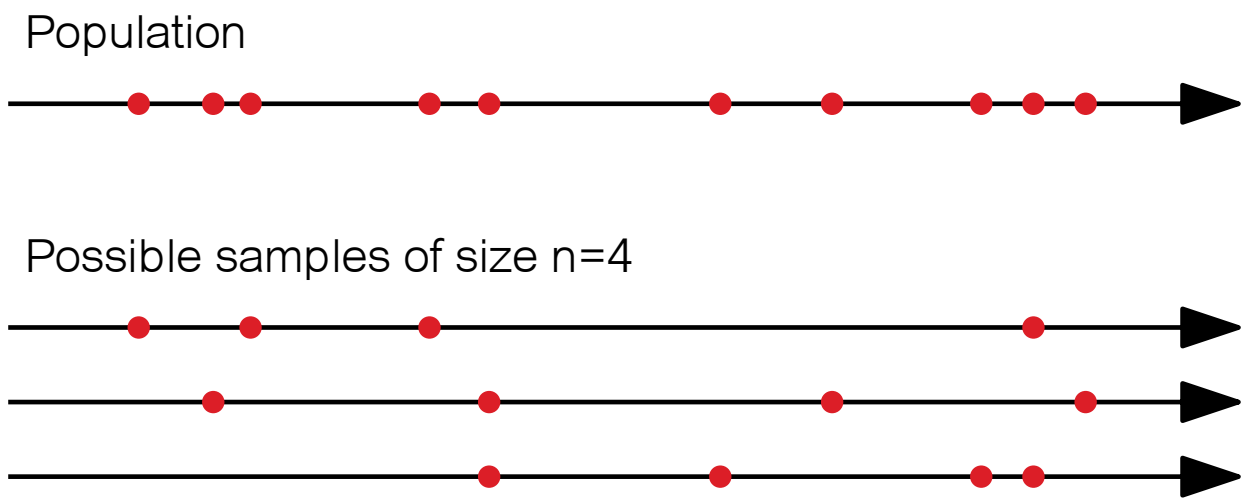
Sample Random subset of entire population

- When studying a population using a sample, we may obtain slightly different answers depending on the sample

Terms used in inferential statistics

Population	The entire set of items of interest.
Sample	A subset of the population.
Statistic	A numerical value computed from the sample.
Parameter	A characteristic of the population under study.
Estimator	A statistic used to estimate a parameter.

Example: Population and sample



Sample estimate of mean and variance

Sample estimate of mean Average value of the sample

$$m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample estimate of variance Estimate of average squared distance to the mean. Notice, we divide by $n-1$, as opposed to n

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2$$

Why divide by $n-1$?

Sample estimate of variance

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2$$

Intuition

- Sample mean is always between the lowest and highest value in the sample—the population mean need not be
- Variance observed in a sample tends to underestimate the population variance
- We would like the *average of all possible sample variances* to equal the population variance

Exercise: Why divide by $n-1$?

Consider a population of $N = 3$ observations

$$x = \{1, 4, 10\}$$

with population mean and variance

$$\mu_x = 5 \qquad \sigma_x^2 = 14$$

- List all possible ordered samples with replacement of size $n = 2$. (Hint: There are 9 such possible samples)

Exercise: Why divide by $n-1$?

Consider a population of $N = 3$ observations

$$x = \{1, 4, 10\}$$

with population mean and variance

$$\mu_x = 5 \qquad \sigma_x^2 = 14$$

- List all possible ordered samples with replacement of size $n = 2$. (Hint: There are 9 such possible samples)

Solution

The 9 possible samples are

$$\{1,1\}, \{1,4\}, \{1,10\}, \{4,1\}, \{4,4\}, \{4,10\}, \{10,1\}, \{10,4\}, \{10,10\}$$

Exercise: Why divide by $n-1$? (II)

Consider a population of $N = 3$ observations

$$x = \{1, 4, 10\}$$

with population mean and variance

$$\mu_x = 5 \qquad \sigma_x^2 = 14$$

Sample estimate

$$m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2$$

- Compute the sample estimate of the mean and variance, m_x and s_{n-1}^2 for each possible sample

$$\{1,1\}, \{1,4\}, \{1,10\}, \{4,1\}, \{4,4\}, \{4,10\}, \{10,1\}, \{10,4\}, \{10,10\}$$

- What is the average sample estimate of the mean and variance (averaged over all possible samples)?

Exercise: Why divide by $n-1$? (II)

Solution

Sample	m_x	s_x^2
$\{1,1\}$		
$\{1,4\}$		
$\{1,10\}$		
$\{4,1\}$		
$\{4,4\}$		
$\{4,10\}$		
$\{10,1\}$		
$\{10,4\}$		
$\{10,10\}$		

Exercise: Why divide by $n-1$? (II)

Solution

Sample	m_x	s_x^2
$\{1,1\}$	$\frac{1+1}{2}=1$	$\frac{(1-1)^2+(1-1)^2}{2-1}=0$
$\{1,4\}$		
$\{1,10\}$		
$\{4,1\}$		
$\{4,4\}$		
$\{4,10\}$		
$\{10,1\}$		
$\{10,4\}$		
$\{10,10\}$		

Exercise: Why divide by $n-1$? (II)

Solution

Sample	m_x	s_x^2
$\{1,1\}$	$\frac{1+1}{2}=1$	$\frac{(1-1)^2+(1-1)^2}{2-1}=0$
$\{1,4\}$	$\frac{1+4}{2}=2.5$	$\frac{(1-2.5)^2+(4-2.5)^2}{2-1}=4.5$
$\{1,10\}$		
$\{4,1\}$		
$\{4,4\}$		
$\{4,10\}$		
$\{10,1\}$		
$\{10,4\}$		
$\{10,10\}$		

Exercise: Why divide by $n-1$? (II)

Solution

Sample	m_x	s_x^2
$\{1,1\}$	$\frac{1+1}{2}=1$	$\frac{(1-1)^2+(1-1)^2}{2-1}=0$
$\{1,4\}$	$\frac{1+4}{2}=2.5$	$\frac{(1-2.5)^2+(4-2.5)^2}{2-1}=4.5$
$\{1,10\}$	$\frac{1+10}{2}=5.5$	$\frac{(1-5.5)^2+(10-5.5)^2}{2-1}=40.5$
$\{4,1\}$		
$\{4,4\}$		
$\{4,10\}$		
$\{10,1\}$		
$\{10,4\}$		
$\{10,10\}$		

Exercise: Why divide by $n-1$? (II)

Solution

Sample	m_x	s_x^2
$\{1,1\}$	$\frac{1+1}{2}=1$	$\frac{(1-1)^2+(1-1)^2}{2-1}=0$
$\{1,4\}$	$\frac{1+4}{2}=2.5$	$\frac{(1-2.5)^2+(4-2.5)^2}{2-1}=4.5$
$\{1,10\}$	$\frac{1+10}{2}=5.5$	$\frac{(1-5.5)^2+(10-5.5)^2}{2-1}=40.5$
$\{4,1\}$	2.5	4.5
$\{4,4\}$	4	0
$\{4,10\}$	7	18
$\{10,1\}$	5.5	40.5
$\{10,4\}$	7	18
$\{10,10\}$	10	0

Exercise: Why divide by $n-1$? (II)

Solution

Sample	m_x	s_x^2
$\{1,1\}$	$\frac{1+1}{2}=1$	$\frac{(1-1)^2+(1-1)^2}{2-1}=0$
$\{1,4\}$	$\frac{1+4}{2}=2.5$	$\frac{(1-2.5)^2+(4-2.5)^2}{2-1}=4.5$
$\{1,10\}$	$\frac{1+10}{2}=5.5$	$\frac{(1-5.5)^2+(10-5.5)^2}{2-1}=40.5$
$\{4,1\}$	2.5	4.5
$\{4,4\}$	4	0
$\{4,10\}$	7	18
$\{10,1\}$	5.5	40.5
$\{10,4\}$	7	18
$\{10,10\}$	10	0

Average s_x^2 over all possible samples

$$\text{avg}(m_x) = \frac{1+2.5+5.5+2.5+4+7+5.5+7+10}{9} = \frac{45}{9} = 5 = \mu_x$$

$$\text{avg}(s_x^2) = \frac{0+4.5+40.5+4.5+0+18+40.5+18+0}{9} = \frac{126}{9} = 14 = \sigma_x^2$$

Sampling distribution

The *sampling distribution* is the distribution of sample means that occurs as we draw samples (of size n) from the population

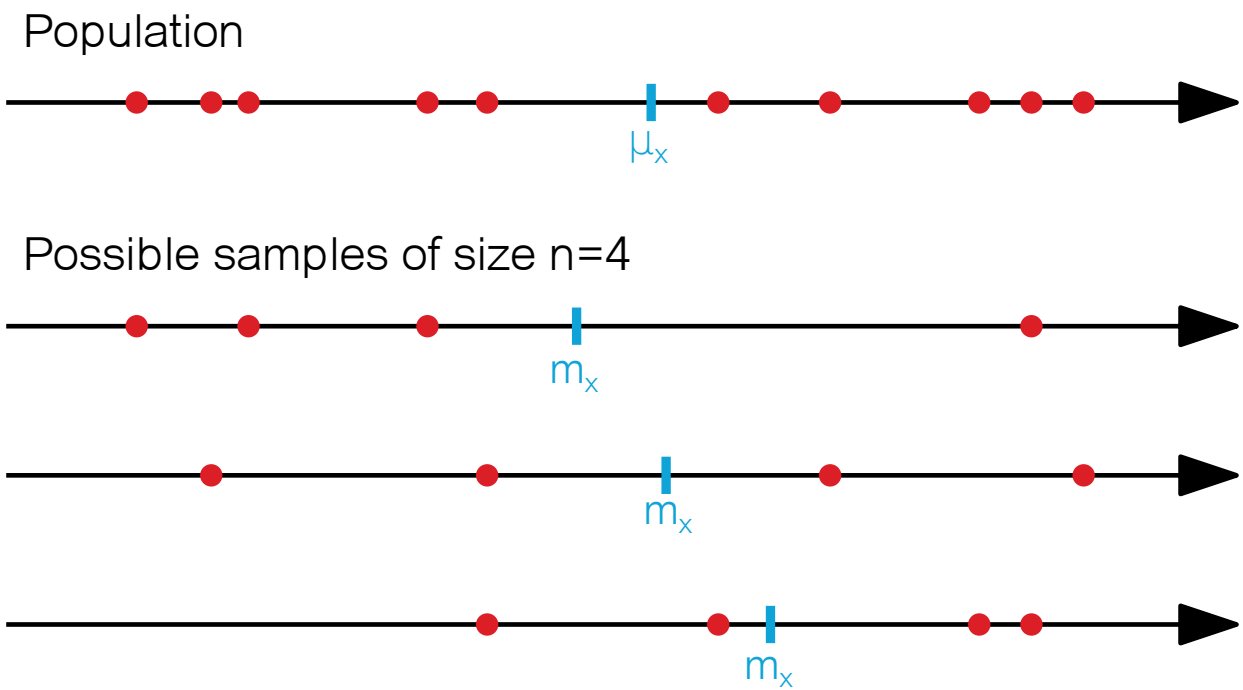
Example We want to study how many kilometers DTU students commute every morning

- We cannot ask all students in the population, so we randomly choose a sample of $n = 10$ persons
- We then compute the mean distance travelled for the 10 persons

Sampling If we then repeat this 5 times, each time asking another sample of 10 persons, we might get the following results

Sample number	Mean number of kilometers
1	5.5
2	7.6
3	2.1
4	6.2
5	1.9

Example: Population and sample means



Central limit theorem

Central limit

What happens to the sampling distribution of the mean, as we increase the sample size n?

- The sample tends to follow a normal distribution
- The sample mean tends to cluster closer around the population mean

Demo: Central limit theorem

Demo: `02-CentralLimitTheorem.ipynb` notebook

Standard error of the mean

Standard error of the mean

We have seen that, as we increase the sample size

- The sample tends to follow a normal distribution
- The sample mean tends to cluster closer around the population mean

If we know the population variance and the sample size n, is there a way to predict the variance of the sampling distribution?

Variance of the sample mean

$$\sigma^2_{m_x} = \frac{\sigma^2}{n}$$

Standard error of the mean

$$\sigma_{m_x} = \frac{\sigma}{\sqrt{n}}$$

Standard error of the mean: Proof

Variance of a sum of independent random variables

$$\mathrm{var}(x_1 + x_2 + \cdots + x_n) = \mathrm{var}(x_1) + \mathrm{var}(x_2) + \cdots + \mathrm{var}(x_n)$$

Variance scales quadratically

$$\mathrm{var}(a \cdot x) = a^2 \cdot \mathrm{var}(x)$$

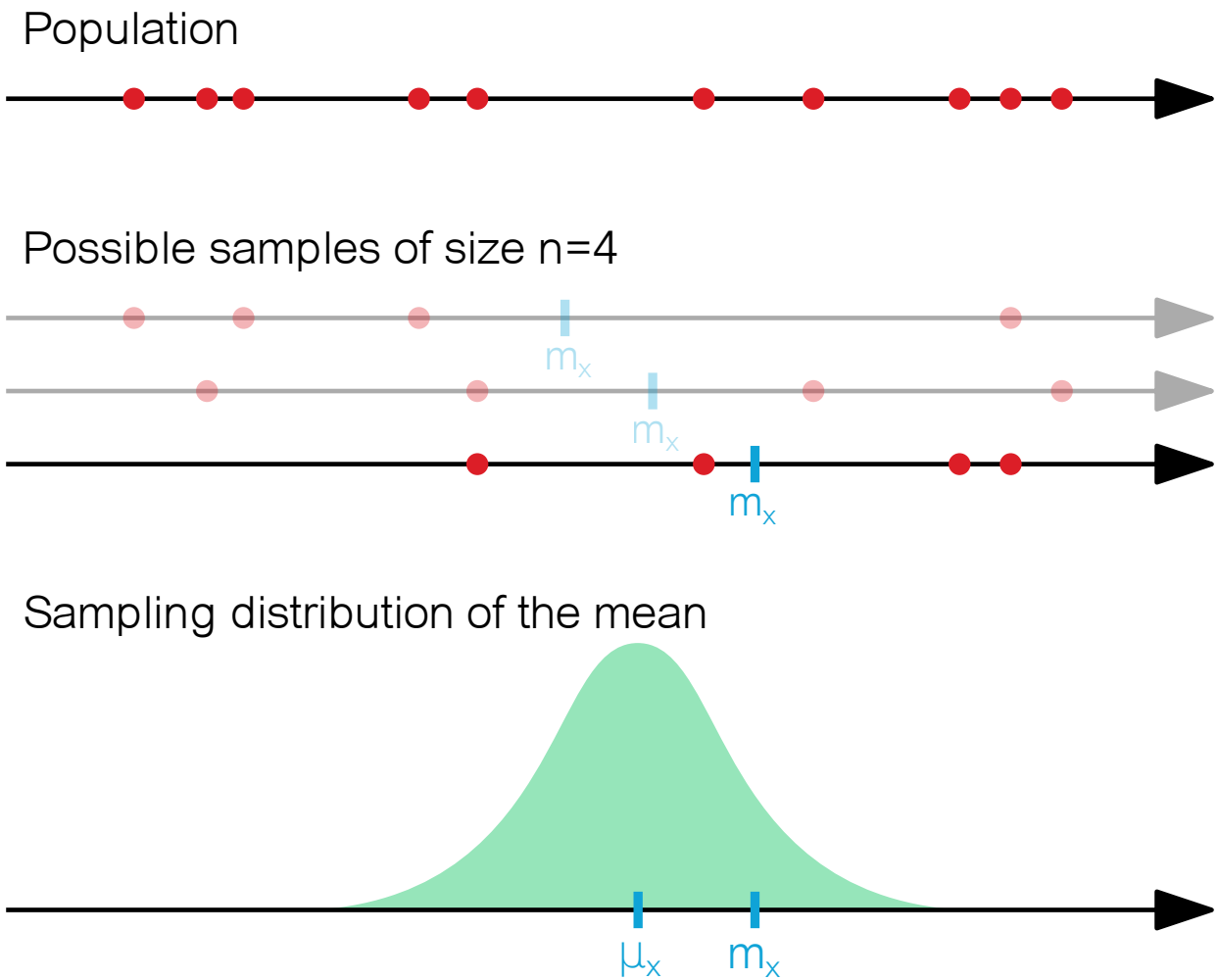
Variance of the sample mean

$$\sigma^2_{m_x} = \mathrm{var}(m_x) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \frac{\sigma^2}{n}$$

Using the central limit theorem

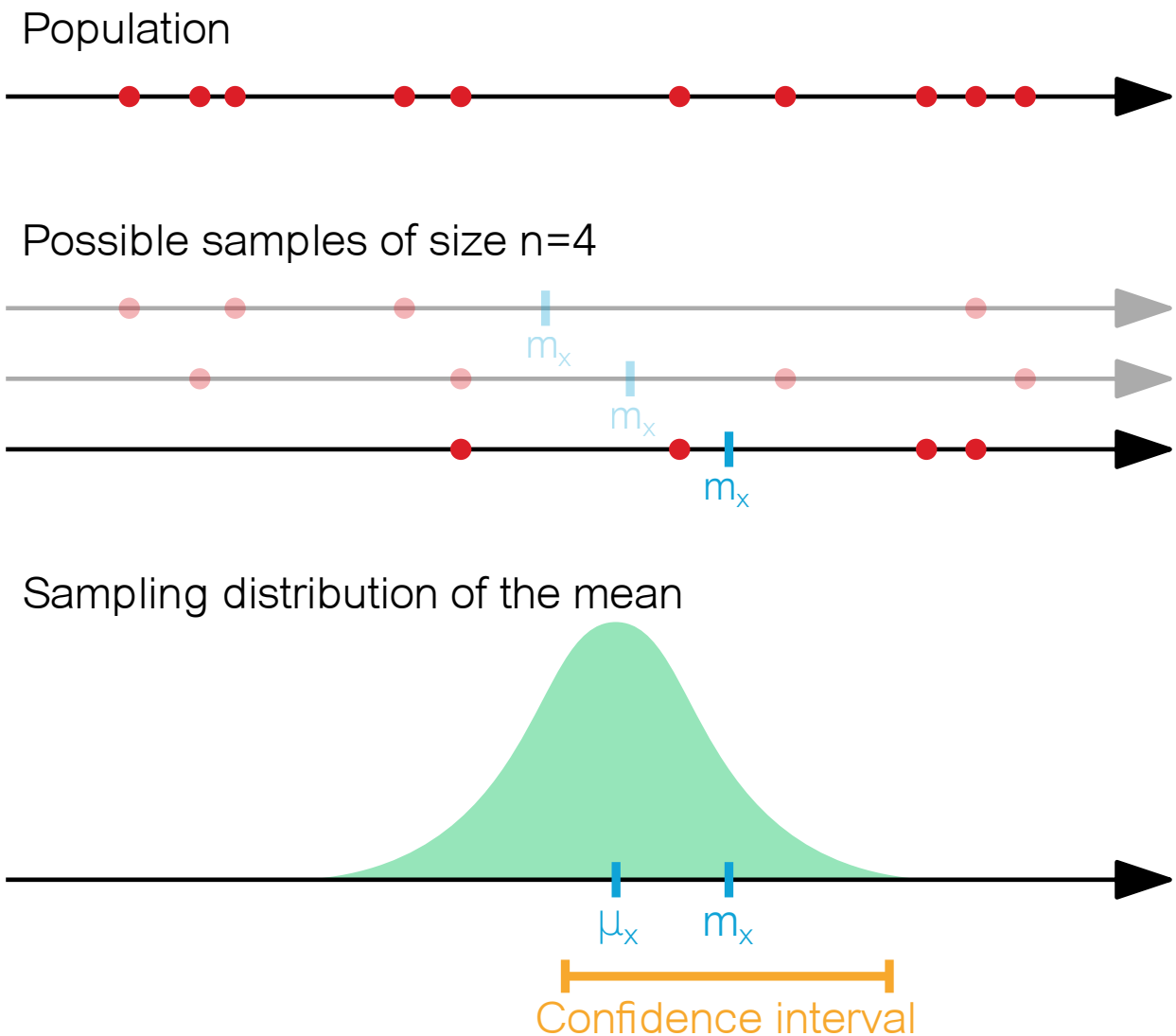
- We know the sample means approximately follow a normal distribution (when the sample size is moderately large)
- We can calculate the mean and variance of that distribution (it is simply $\mu_{m_x} = \mu$ and $\sigma^2_{m_x} = \frac{\sigma^2}{n}$)
- We can use this to predict which values of the sample mean are most likely

Example: Population and sample



Confidence intervals

Example: Population and sample



Confidence interval

Confidence interval

point estimate $\pm$ margin of error

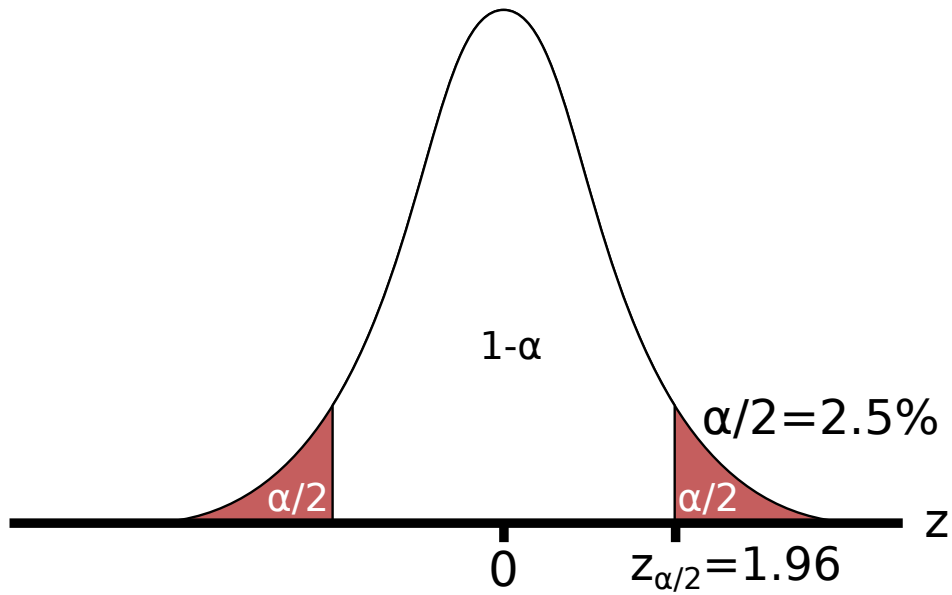
$$m_x \pm \underbrace{z_{\alpha/2}}_{\text{critical value}} \cdot \underbrace{\sqrt{\frac{\sigma^2}{n}}}_{\text{standard error}}$$

Confidence interval for mean We want to estimate the population mean μ

- Sample n observations
- Compute the sample mean $m_x = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Choose confidence level, e.g. $1 - \alpha = 95\%$, and look up critical value
- Compute standard error and multiply by critical value

Critical value

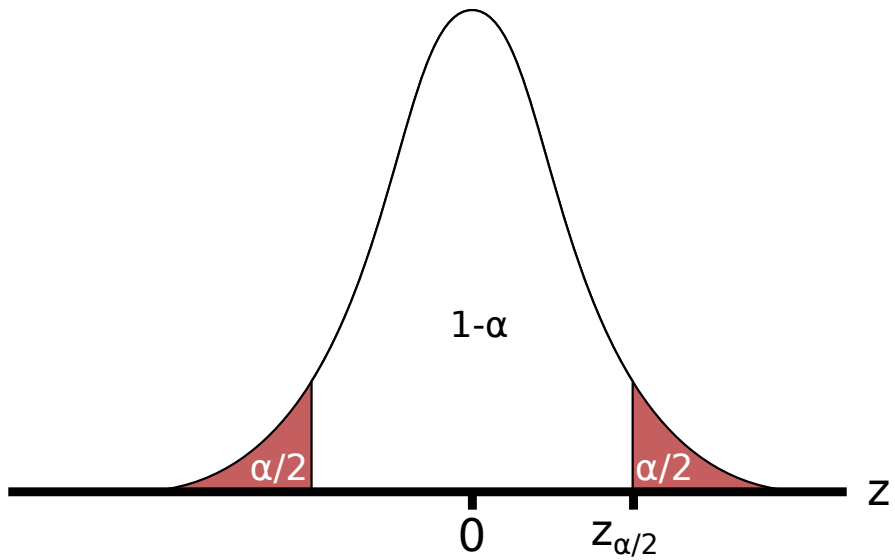
Example: Critical value for $1 - \alpha = 95\%$ interval



Look up in a table or compute in Python using `scipy.stats.norm.ppf`

Table of critical values

Central area $1-\alpha$	Tail area $\alpha/2$	Critical value $z_{\alpha/2}$
50%	0.25	0.67
90%	0.05	1.64
95%	0.025	1.96
99%	0.005	2.58



Interpretation of confidence interval

- Expresses error due to random sampling
- A larger sample size gives a smaller margin of error
- Is defined at a desired confidence level, e.g. $1 - \alpha = 95\%$
- 95% confidence means that out of 100 samples, we expect that 95 times the true population parameter is within the margin of error

Exercise: Mean and variance of a 6-sided dice

Mean and standard deviation of a discrete distribution

- Sum over all possible outcomes
- Weigh each by their probability

$$\mu_x = \sum_{k=1}^{K} P(x_k) \cdot x_k \qquad \sigma_x^2 = \sum_{k=1}^{K} P(x_k) \cdot (x_k - \mu)^2$$

- What is μ_x and σ_x^2 for a normal 6-sided dice?
  $K = 6, \quad x_1 = 1, x_2 = 2, \ldots, x_6 = 6, \quad P(x_1) = P(x_2) = \cdots = P(x_6) = \frac{1}{6}$

Exercise: Mean and variance of a 6-sided dice

Mean and standard deviation of a discrete distribution

- Sum over all possible outcomes
- Weigh each by their probability

$$\mu_x = \sum_{k=1}^{K} P(x_k) \cdot x_k \qquad \sigma_x^2 = \sum_{k=1}^{K} P(x_k) \cdot (x_k - \mu)^2$$

- What is μ_x and σ_x^2 for a normal 6-sided dice?
  $K = 6, \quad x_1 = 1, x_2 = 2, \ldots, x_6 = 6, \quad P(x_1) = P(x_2) = \cdots = P(x_6) = \frac{1}{6}$

Solution

$$\mu_x = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = 3.5$$

$$\sigma_x^2 = \frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\right)$$

$$= \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \approx 2.917$$

Exercise: Confidence interval of 10 dice throws

- Throw a 6-side dice 10 times and record the results
  (e.g. use `www.random.org/dice`)
- Compute the 50% confidence interval for the mean
  Express it as a range [low, high]

Confidence interval

$$m_x \pm \underbrace{z_{\alpha/2}}_{\text{critical value}} \cdot \underbrace{\sqrt{\frac{s_x^2}{n}}}_{\text{standard error}}$$

$$(z_{0.25} = 0.67)$$

Exercise: Confidence interval of 10 dice throws

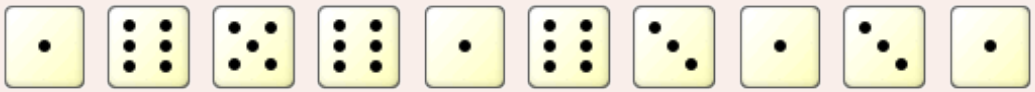
- Throw a 6-side dice 10 times and record the results
  (e.g. use `www.random.org/dice`)
- Compute the 50% confidence interval for the mean
  Express it as a range [low, high]

Confidence interval

$$m_x \pm \underbrace{z_{\alpha/2}}_{\text{critical value}} \cdot \underbrace{\sqrt{\frac{s_x^2}{n}}}_{\text{standard error}}$$

$$(z_{0.25} = 0.67)$$

Solution example



Exercise: Confidence interval of 10 dice throws

- Throw a 6-side dice 10 times and record the results
  (e.g. use `www.random.org/dice`)
- Compute the 50% confidence interval for the mean
  Express it as a range [low, high]

Confidence interval

$$m_x \pm \underbrace{z_{\alpha/2}}_{\text{critical value}} \cdot \underbrace{\sqrt{\frac{s_x^2}{n}}}_{\text{standard error}}$$

$$(z_{0.25} = 0.67)$$

Solution example

$$m_x = \tfrac{1}{10}(1+6+5+6+1+6+3+1+3+1) = \tfrac{33}{10} = 3.3$$

$$s_x^2 = \tfrac{1}{10-1}\left((1-3.3)^2 + (6-3.3)^2(5-3.3)^2 + + \cdots + (1-1)^2\right) \approx 5.12$$

Confidence interval

$$m_x \pm z_{\alpha/2} \cdot \sqrt{\frac{s_x^2}{n}} = 3.3 \pm 0.67 \cdot \sqrt{\frac{5.12}{10}} = 3.3 \pm 0.48$$

$$[2.82, 3.78]$$

The population mean is 3.5 and we expect 50% of the computed confidence intervals to include it

Estimating a proportion

Bernoulli distribution

$$P(k) = \pi^k \cdot (1-\pi)^{1-k} = \begin{cases} (1-\pi), & k = 0 \\ \pi, & k = 1 \end{cases}$$

What is the mean and variance?

Estimating a proportion

Bernoulli distribution

$$P(k) = \pi^k \cdot (1-\pi)^{1-k} = \begin{cases} (1-\pi), & k = 0 \\ \pi, & k = 1 \end{cases}$$

What is the mean and variance?

Mean

$$\mu_k = \sum_{k \in \{0,1\}} P(k) \cdot k$$

$$= \underbrace{(1-\pi) \cdot 0}_{k=0} + \underbrace{\pi \cdot 1}_{k=1} = \pi$$

Estimating a proportion

Bernoulli distribution

$$P(k) = \pi^k \cdot (1-\pi)^{1-k} = \begin{cases} (1-\pi), & k=0 \\ \pi, & k=1 \end{cases}$$

What is the mean and variance?

Mean

$$\begin{aligned} \mu_k &= \sum_{k \in \{0,1\}} P(k) \cdot k \\ &= \underbrace{(1-\pi) \cdot 0}_{k=0} + \underbrace{\pi \cdot 1}_{k=1} = \pi \end{aligned}$$

Variance

$$\begin{aligned} \sigma_k^2 &= \sum_{k \in \{0,1\}} (k-\mu)^2 \cdot P(k) \\ &= \underbrace{(0-\pi)^2 \cdot (1-\pi)}_{k=0} + \underbrace{(1-\pi)^2 \cdot \pi}_{k=1} \\ &= \pi^2(1-\pi) + (1+\pi^2-2\pi)\pi \\ &= \pi^2 - \pi^3 + \pi + \pi^3 - 2\pi^2 = \pi - \pi^2 = \pi(1-\pi) \end{aligned}$$

Confidence interval for proportion

Confidence interval for proportion

$$\text{point estimate} \pm \text{margin of error}$$

$$p \pm \underbrace{z_{\alpha/2}}_{\text{critical value}} \cdot \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{standard error}}$$

Confidence interval for proportion

- Estimate unknown true proportion parameter, π
- Sample n observations
- Compute the sample proportion $p = \frac{\#\text{correct}}{n}$
- Choose confidence level, e.g. $1 - \alpha = 95\%$, and look up critical value.
- Compute standard error and multiply by critical value

Sample size for proportion

- The equation for the confidence interval can be solved for the sample size
- This gives a formula for the required sample size to give a desired margin of error

Sample size for proportion

$$n = z^2_{\alpha/2} \frac{p(1-p)}{e^2}$$

n: Required sample size
p: Expected population proportion
Unknown, but we can substitute best guess or worst case $p = 50\%$
e: Desired margin of error
α: Significance level

Tasks

Tasks

Tasks today

- Start working on lab report 1: Read description on DTU Learn
- Today's feedback group
  - Magnus Alexander Mollatt van Capel
  - Rasmus Johansen Rieneck
  - Christian Rahbæk Warburg
  - Clara Louise Brodt

Lab report hand in

- Lab 1: Image recognition (Deadline: Thursday 14 September 20:00)

Next time

- Algorithms. Preparation: Read the note "Algorithms" + solve problems