Y-DATA
SCHOOL OF DATA SCIENCE

Probability and Statistics for Data Science

Lecture 4 – Parameter estimation

From probability to statistics

- In the first three lectures, we acquired important tools from probability
- Today, we will use these tools to develop statistical methods
- Specifically, we will learn how to estimate a parameter and to point at some useful properties we would like for our estimators

Today

- Sample
- (Parametric) statistical models
- Likelihood function
- Maximum likelihood estimation
- Method of moments
- Goodness of estimation and MSE
- Interval estimation

Sample and statistical model

- A *sample* is a vector of random variables $X = (X_1, \dots, X_n)$
- We normally believe that the observations are independent and identically distributed (*i.i.d.*)
- These assumptions are not always correct, but this framework allows us to develop statistical tools for various situations.
- A *statistical model* is a family of distribution $\mathbb{F}$ from which we believe that the data are generated.

Parametric vs. non-parametric models

- We will assume that $\mathbb{F}$ is a parametric family.
- That is, we assume that we know the type of distribution up to its unknown parameter(s) $\theta \in \Theta$, where Θ is a *parameter space*.
- For example, we can assume that $X_i \sim Pois(\theta), \theta > 0$.
- When the number of unknown parameters is infinite, we say that the model is non-parametric. This will NOT be the case in our course.

To sum up:

We will deal with observations $X_1, \dots, X_n$ i.i.d. random variables with joint distribution $F_\theta(x)$ that belongs to a parametric family of distributions $\mathbb{F}_\theta, \theta \in \Theta$.

What can we do with our model?

In statistical inference, many problems can be identified as being one of three types:

- Point estimation

Providing a single “best guess” of some quantity. Normally θ or some function of θ.

- Confidence intervals

Providing an interval that traps θ with some desired probability.

- Hypothesis testing

Providing evidence to reject some default theory.

Today, we will focus on the first two types of statistical inference.

Point Estimation

Likelihood function

- Assume that we have a sample $X_1, \dots, X_n$ from some discrete distribution with one unknown parameter θ. Define the likelihood function to be
$$L(\theta; \boldsymbol{x}) = P_\theta(X_1 = x_1, \dots, X_n = x_n)$$
- Interpretation: The probability to observe the given data $\boldsymbol{x}$, for any possible value $\theta \in \Theta$**.**
- For example, if for a given x, $L(\theta_1; x) > L(\theta_2; x)$ we can say that the value θ_1 for θ is more suited to the data than θ_2.
- According to this logic, it makes sense to estimate θ by maximizing the likelihood function with respect to θ.

Maximum likelihood estimator

- **Def:** The maximum likelihood estimator (MLE) $\hat{\theta}$ of θ is
$$\hat{\theta} = argmax_{\theta \in \Theta} L(\theta; x)$$
- In the case of an i.i.d. sample $X_1, \dots, X_n$, the likelihood function is
$$L(\theta; x) = \prod_{i=1}^{n} P_\theta(X_i = x_i)\text{, if } X_i's \text{ are discrete}$$
$$L(\theta; x) = \prod_{i=1}^{n} f_\theta(x_i)\text{, if } X_i's \text{ are continuous}$$
- In practice, it is usually easier to find the maximum of the log-likelihood function (discrete case is similar)
$$l(\theta; x) = \ln L(\theta; x) = \sum_{i=1}^{n} \ln f_\theta(x_i)$$

MLE example – Bernoulli distribution

Consider a sample of i.i.d. random variables $X_1, \ldots, X_n$ from the distribution $Ber(\theta)$.

The likelihood function is

$$L(\theta; x) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \; = \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)}$$

$$= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$$

The log-likelihood is then

$$l(\theta; x) = \left(\sum_i x_i\right) \ln \theta + \left(n - \sum_i x_i\right) \ln(1-\theta)$$

MLE example – Bernoulli distribution

$$l(\theta; x) = \left(\sum_i x_i\right) \ln \theta + \left(n - \sum_i x_i\right) \ln(1 - \theta)$$

Differentiating $l(\theta; x)$ w.r.t. θ yields

$$l'(\theta; x) = \frac{\sum_i x_i}{\theta} + \frac{n - \sum_i x_i}{1 - \theta}$$

Equating the derivative to 0, we get

$$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$$

(it is left to check that the second derivative is negative to verify that $\hat{\theta}$ is indeed a maximum.

Estimation – method of moments

- We express the population moments of the distribution of the data in terms of its unknown parameter and equate them to the corresponding sample moments. The parameters are estimated by the solutions of the resulting equations.
- More precisely, consider the first p moments as functions of

$\theta = (\theta_1, \dots, \theta_p)$,

$$\mu_j(\theta) = E\left(X_1^j\right), j = 1, \dots, p$$

The corresponding sample moment is $m_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$.

Now we can obtain a system of p equations $\mu_j(\theta) = m_j, j = 1, \dots, p$. Solving this system implies the method of moments estimator $\hat{\theta} = \left(\hat{\theta}_1, \dots, \hat{\theta}_p\right)$.

MOM – mean and variance example

Assume that we have some i.i.d. sample $X_1, \dots, X_n$ from a distribution with first two finite moments and we want to estimate $\theta_1 = E(X_1)$ and $\theta_2 = Var(X_1)$.

Applying the method of moments, yield the following system of equations (remember that we need $\mu_j(\theta) = E\left(X_1^j\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^j = m_j$)

$$\mu_1(\theta) = E(X_1) = \theta_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu_2(\theta) = E(X_1^2) = Var(X_1) + E^2(X_1) = \theta_2 + \theta_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

This implies,

$$\hat{\theta}_1 = \bar{X}_n, \hat{\theta}_2 = \overline{X^2}_n - (\bar{X}_n)^2$$

MOM – uniform example

Consider an i.i.d. sample $X_1, \dots, X_n \sim U(\{0, \dots, \theta\})$. The MLE in this case is $\hat{\theta}_{MLE} = X_{max}$. Let's estimate θ using the method of moments. In this case,

$$\mu_1(\theta) = E(X_1) = \frac{\theta}{2} = m_1 = \bar{X}_n$$

Therefore, $\hat{\theta}_{MOM} = 2\bar{X}_n$.

This estimator is problematic. If we have observations $0.2, 0.6, 2, 0.4$, then $\hat{\theta}_{MOM} = 1.6$. This doen't make sense since one of the observations is bigger that $\hat{\theta}_{MOM}$.

Goodness of estimation

- We have considered two estimation methods. Of course, there are more.
- How can we compare between estimators? Which performs better? Does the best estimator exists?
- First, we need to define what “better” means. That is, define a measure of goodness of estimation.
- Today we will talk about the MSE as the measure of choice.
- We will also get to know additional proporties that we would like our estimator to have.

MSE

- **Def:** The *mean squared error* (MSE) of the estimator $\hat{\theta}$ is
$$MSE\left(\hat{\theta}, \theta\right) = E\left(\hat{\theta} - \theta\right)^2$$
- A good estimator is one with low values of MSE.
- Warning: the MSE depends on the value of the unknown parameter θ! This means that when comparing the MSE of two estimators, we may get that one of them is better for some values of θ, and vice versa for another region of θ.
- Yet, if we have two estimators such that $MSE\left(\hat{\theta}_1, \theta\right) < MSE\left(\hat{\theta}_2, \theta\right)$ for all $\theta \in \Theta$, we would prefer $\hat{\theta}_1$ over $\hat{\theta}_2$.

MSE decomposition

- **Theorem:** The MSE can be written as
$$MSE\left(\hat{\theta},\theta\right) = Var_{\theta}\left(\hat{\theta}\right) + \left(E\left(\hat{\theta}\right) - \theta\right)^2 = Var_{\theta}\left(\hat{\theta}\right) + b\left(\hat{\theta}\right)^2$$
- **Def:** The bias of the estimator $\hat{\theta}$ is $b\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$. We say that an estimator is unbiased if $b\left(\hat{\theta}\right) = 0$ for any value of θ.

Example: For an i.i.d. sample $X_1, \dots, X_n \sim N(\mu, 1\ )$, compute the bias of the estimator $\hat{\mu} = \bar{X}_n$ for μ.

Sol: It holds that $E_{\mu}(\hat{\mu}) = \mu$. Therefore, $\hat{\mu}$ is unbiased for μ.

The variance of $\hat{\mu}$ is: $Var_{\mu}(\hat{\mu}) = \frac{1}{n}$.

To sum up, $MSE(\hat{\mu}, \mu) = \frac{1}{n}$.

MSE example

For a sequence of i.i.d. Bernoulli trials $X_1, \dots, X_n \sim Ber(p)$, consider the following estimators for p:

$$\hat{p}_1 = \bar{X}_n, \qquad \hat{p}_2 = \frac{\sum_i X_i + 1}{n+2}, \qquad \hat{p}_3 = X_1$$

- $\hat{p}_1$ and $\hat{p}_3$ are both unbiased.
- $\hat{p}_2$ is biased but has smaller variance.

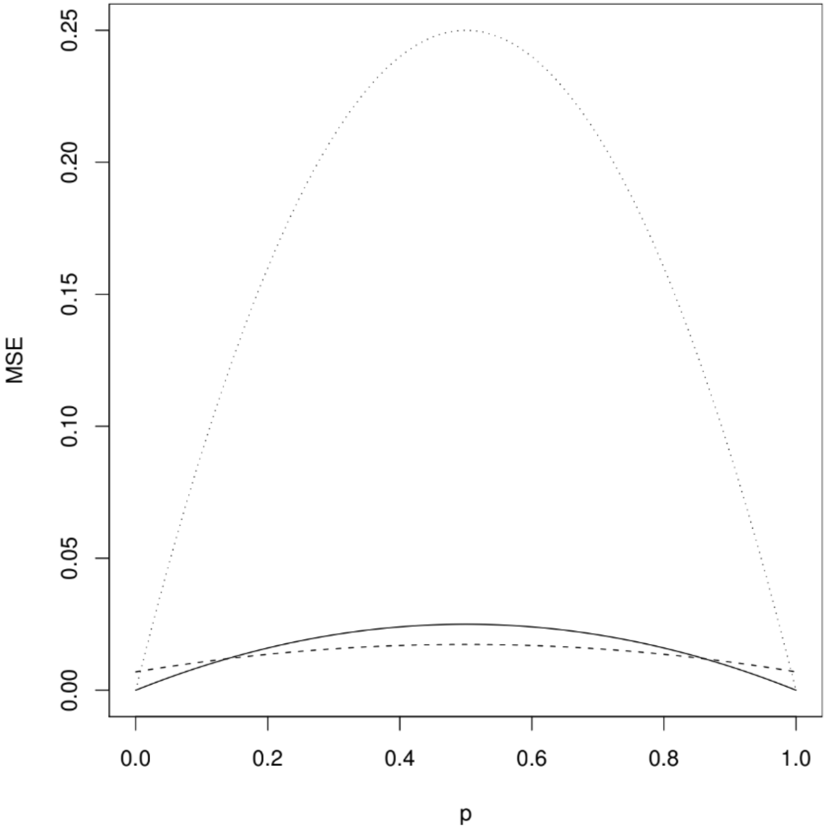


Figure 2.3: MSE for three estimators of p in $\mathscr{B}(10,p)$: $MSE(\hat{p}_1,p)$ (solid line), $MSE(\hat{p}_2,p)$ (dashed line), and $MSE(\hat{p}_3,p)$ (dotted line).

Unbiasedness is nice but is not everything!

Desirable properties of estimators

- Low MSE $MSE\left(\hat{\theta},\theta\right) = E\left(\hat{\theta}-\theta\right)^2$
- Unbiasedness $\mathrm{E}_{\theta}\hat{\theta} = \theta$ (but may increase variance!)
- Consistency: $\hat{\theta}_n \to \theta$ in probability, as $n \to \infty$ (as sample size grows)

So, why MLE?

- Under some regularity conditions, the MLE is consistent.
- Functional invariance: If $\hat{\theta}$ is the MLE of θ, then $g\left(\hat{\theta}\right)$ is the MLE of $g(\theta)$ for a known function g
- Efficiency: Under appropriate conditions, the MLE is asymptotically normal.

Confidence Intervals

Confidence intervals

- So far, we tried to estimate the exact value of the unknown parameter θ
- In most cases, even if our estimate is very good it will probably not be equal to θ. There will always be some error.
- That’s why we are going to build an interval that covers the real parameter with high probability
- Such interval is called a *confidence interval* and the probability that it contains the parameter is called the *confidence level*
- We would like our interval to be as short as possible
- It turns out that there’s a tradeoff between the confidence level and the lengh of the interval

Confidence interval

- **Def:** A confidence interval (CI) for the unknown parameter θ with confidence level $1-\alpha$ based on the random sample $X_1, \dots, X_n$ is a random interval $[C_1(X_1, \dots, X_n), C_2(X_1, \dots, X_n)]$ such that for all $\theta \in \Theta$

$$P\big(C_1(X_1, \dots, X_n) \le \theta \le C_2(X_1, \dots, X_n)\big) = 1 - \alpha$$

Remarks:

1. α is a probability – usually a small probability such as 0.05 or 0.01.

α is set before computing the CI.

2. Remember! C_1, C_2 are functions of the sample and therefore random.

3. Sometimes we will require for simplicity that

$$P(C_1 \le \theta \le C_2) \ge 1 - \alpha$$

Last remark about CI’s

It is important to remember that NOT for every sample the event

$$\{C_1 \leq \theta \leq C_2\}$$

will hold.

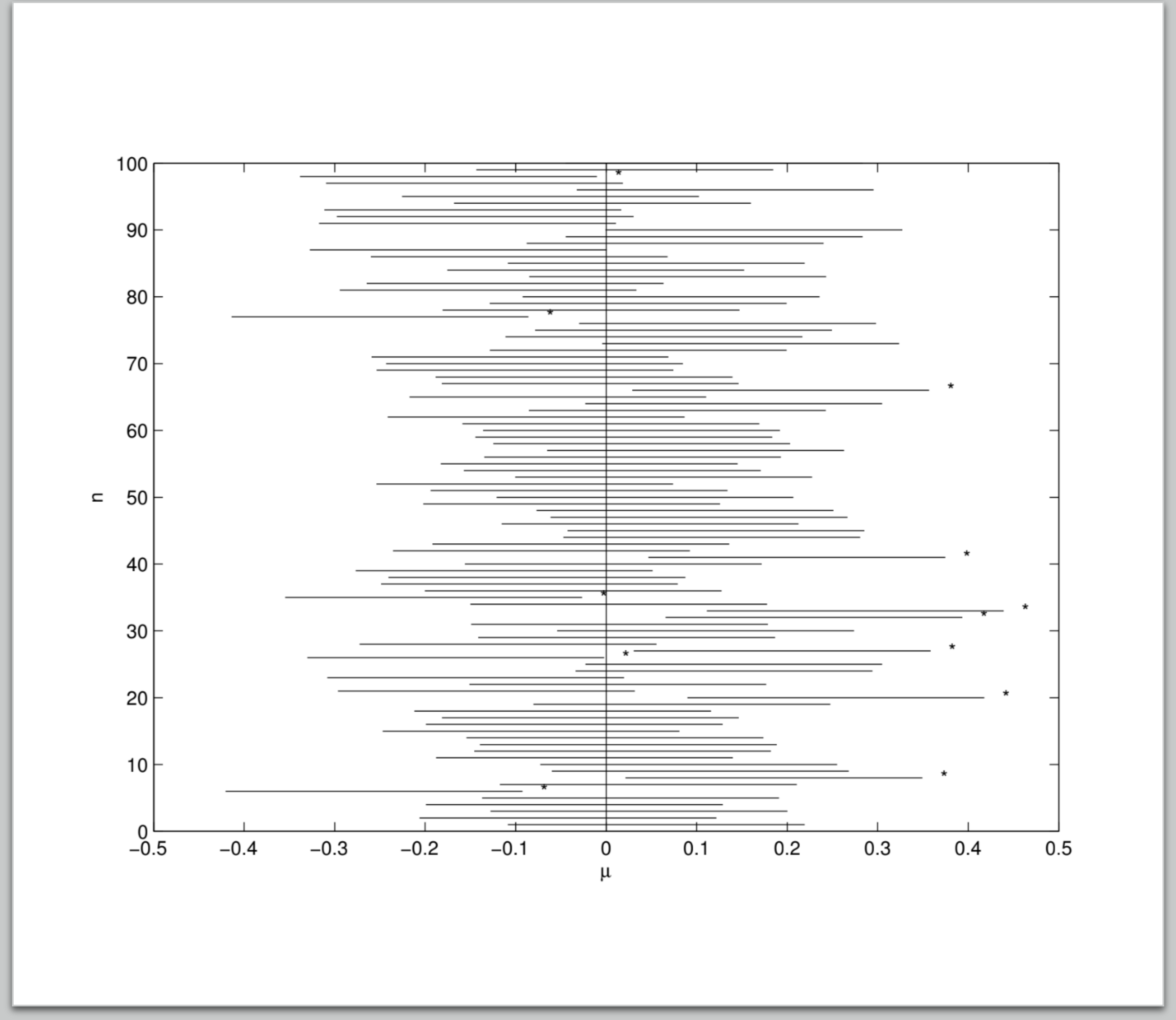
This will happen on $(1-\alpha)\%$ of the samples.

Moreover, we can’t know for our specific sample whether the true parameter is covered or not.

Visualization of CI

Here we have 90% CI's for the unknown parameter μ from 100 independent sample $N(\mu, 1)$ of size 100.

The true value of the parameter was $\mu = 0$. Approx. 10% do not include the true value.



Example: Normal sample with known variance

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and assume that σ^2 is known. We want to construct a CI for μ with confidence level $1 - \alpha$.

We will try to find some symmetric interval around the point estimator for μ – which is the mean. That is, we will find c such that

$$P(\bar{X}_n - c \leq \mu \leq \bar{X}_n + c) = 1 - \alpha$$

To this end,

$$1 - \alpha = P(\bar{X}_n - c \leq \mu \leq \bar{X}_n + c) = P(-c \leq \bar{X}_n - \mu \leq c)$$

$$= P\left(-\frac{c}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{c}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right)$$

Example: Normal sample with known variance

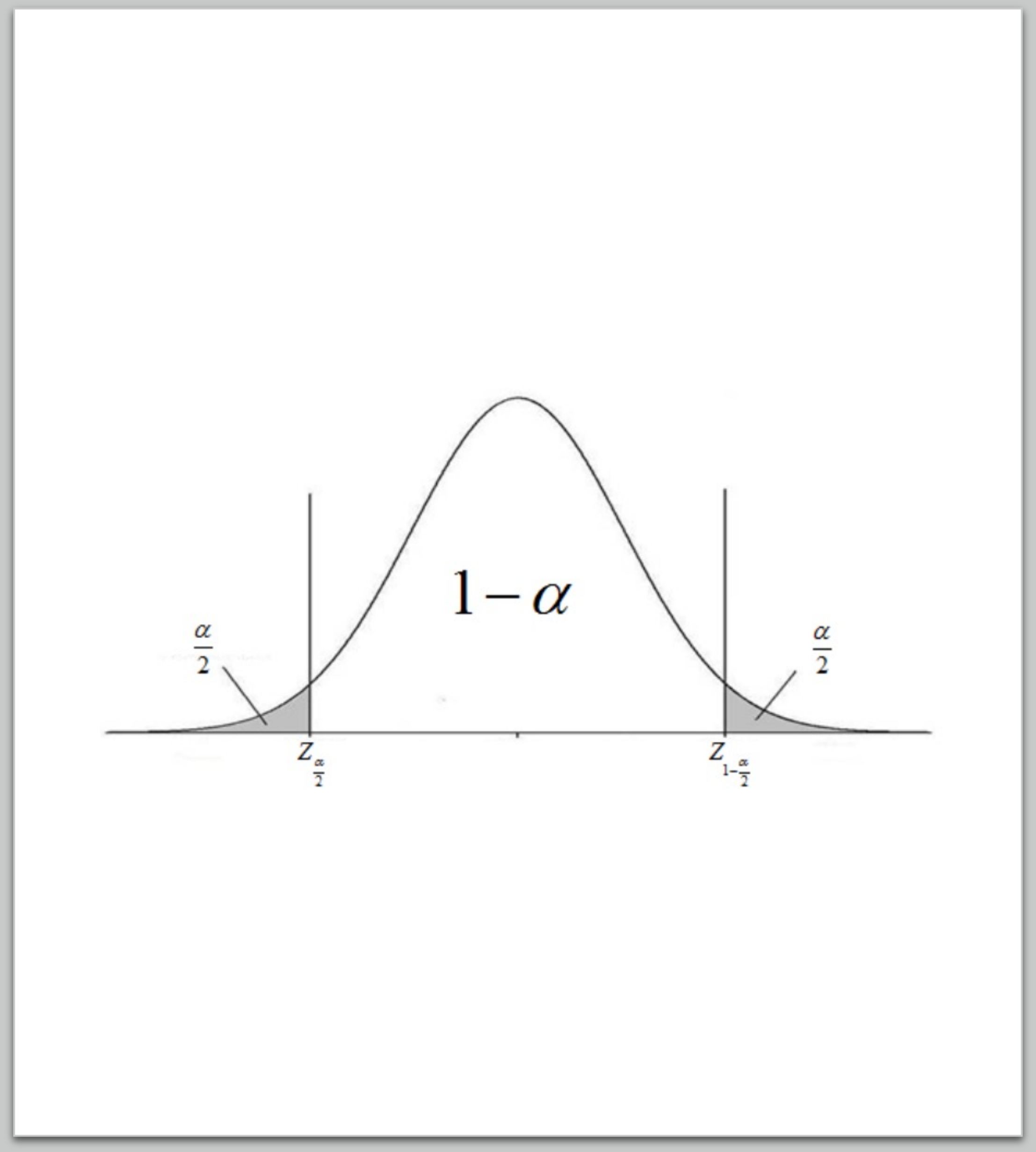
- We get that for this case,

$$c = z_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$$

And the CI is $CI = \left[\bar{X}_n - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right] = \left[\bar{X}_n \pm \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$

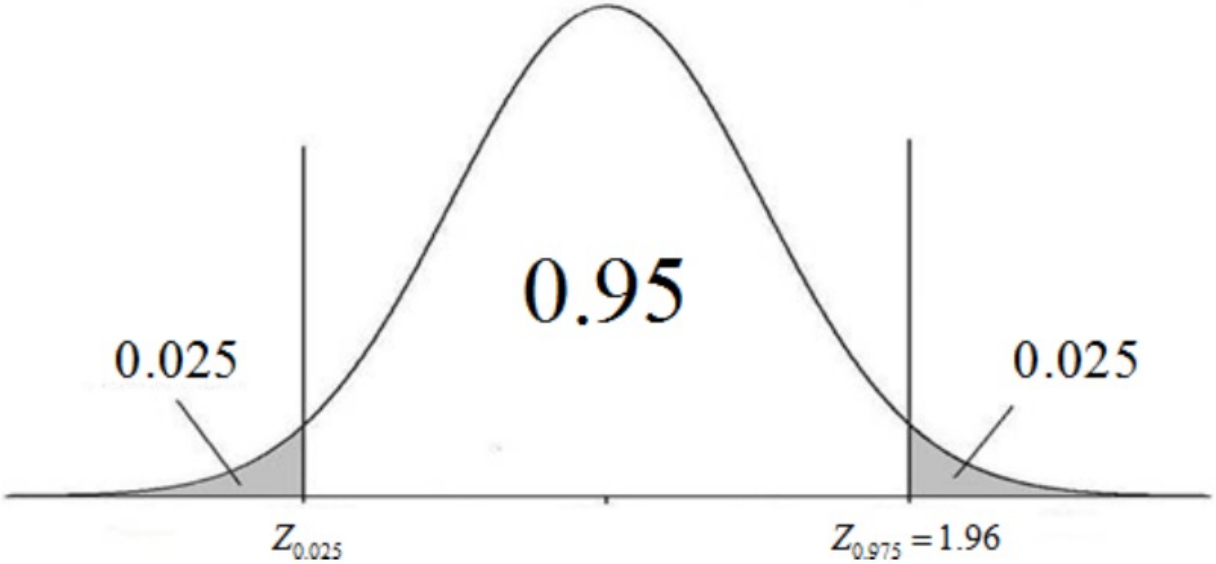
The length of the CI is $L = C_2 - C_1 = 2 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$.

Note, as n increases, the length is smaller. As $1 - \frac{\alpha}{2}$ increases, the length increases.



Example for the example

- Suppose that we're looking for a 95% CI for the normal mean μ. That is, $\alpha = 0.05$. So, the quantile $z_{1-\frac{\alpha}{2}} = 1.96$.
- The CI is given by $\bar{X}_n \pm 1.96\sigma/\sqrt{n}$



Example: Normal sample with unknown mean and variance

- **Def:** If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ are independent, and denote the unbiased estimator for the variance by $S_n^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2$. Then the distribution of the random variable

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

Is called t distribution with $n - 1$ degrees of freedom.

- When the variance is also unknown in a normal sample, the $1 - \alpha$ CI for μ is given by,

$$\bar{X}_n \pm \frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}$$

Note, the length in this case is also a RV: $L = 2\frac{S_n}{\sqrt{n}} t_{n-1,1-\alpha/2}$

Asymptotic CI

- What if our sample is not normally distributed? In many cases, we can use the central limit theorem to construct a CI.
- Let $X_1, \dots, X_n$ be an i.i.d. sample with mean $E(X)$ and variance $Var(X)$. Then, a $1 - \alpha$ asymptotic CI for $E(X)$ is given by

$$\left[\bar{X}_n \pm \frac{\sqrt{\hat{V}ar(X)}}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$$

- That's true because

$$P\left(\frac{\bar{X}_n - E(X)}{\frac{\sqrt{\hat{V}ar(X)}}{\sqrt{n}}} \leq x\right) \approx \Phi(x)$$

Example: CI for the probability in Ber(p) dist.

Let our sample be $X_1, \dots, X_n \sim Ber(p)$.

In this case, $E(X) = p, Var(X) = p(1-p)$. Thus,

$$\hat{p} = \bar{X}_n, \hat{V}ar(X) = \hat{p}(1-\hat{p})$$

The asymptotic confidence interval for p is then given by

$$\left[\hat{p} \pm \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$$

Sample size planning

- What is the number of observations needed to construct a $1-\alpha$ CI with length of at most L_0?
- For the case of normal sample with known variance, we know that

$$L = 2\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

- Thus, the answer is $n = 4z_{1-\alpha/2}^2 \frac{\sigma^2}{L_0^2}$
- When the variance is unknown, L is random and we can’t compute n. A way to approach it is assuming that the variance is known. Another way is to upper bound the variance and use the bound as the estimator of the variance.

CI for a function of the parameter

- Let $[C_1, C_2]$ be the $1 - \alpha$ CI for θ, and suppose we want to construct a $1 - \alpha$ CI for $g(\theta)$ (the function g is known).
- If g is monotone increasing, then $[g(C_1), g(C_2)]$ is a $1 - \alpha$ CI for $g(\theta)$.
  $P\big(g(C_1) \leq g(\theta) \leq g(C_2)\big)$
  $= P\left(g^{-1}(g(C_1)) \leq g^{-1}(g(\theta)) \leq g^{-1}\big(g(C_2)\big)\right) = P(C_1 \leq \theta \leq C_2)$
  $= 1 - \alpha$
- If g is monotone decreasing, then $[g(C_2), g(C_1)]$ is a $1 - \alpha$ CI for $g(\theta)$.

CI’s everywhere!

We covered some types of confidence interval.

The goal was to give you the tools to understand the rational behind confidence intervals, so that when you meet a different CI, you know what does it mean.

i.e.

- CI for the variance
- CI for mean differences
- Simultaneous CI’s

A glimpse into the Bayesian approach

- Under the Bayesian approach, the parameter is a random variable as well.
- This allows us to incorporate priop belief regarding the parameter.
- The sample at hand is then $X|\boldsymbol{\theta} = \theta$, and we need to consider that in our calculations.
- The inference on θ is based on the posterior distribution $\theta|X$.
- In some way, we have some belief regarding θ, and we “update” this belief using the data.

References

- Abramovich, Felix, and Ya'acov Ritov. “Statistical theory: a concise introduction. CRC Press, 2013.
- Bertsekas, Dimitri P. and Tsitsiklis, John N.. "Introduction to Probability." 2008 .
- Wasserman, Larry. *All of statistics : a concise course in statistical inference*. New York: Springer, 2010.
- Haviv, Moshe. Introduction to Descriptive Statistics and Probability, 2021.
- Micha Mandel’s lecture notes in course 52221 at the Hebrew University.