CENG 222 Probability and Statistics

Statistical Inference

Nesli Erdoğmuş & Burak Galip Aslan

Contents

- **9.1 Introduction**
    - **9.1.1 Method of moments**
    - **9.1.2 Method of maximum likelihood**
    - **9.1.3 Estimation of standard errors**
- 9.2 Confidence intervals
    - 9.2.1 Construction of confidence intervals: a general method
    - 9.2.2 Confidence interval for the population mean
    - 9.2.3 Confidence interval for the difference between two means
- 9.4 Hypothesis testing
    - 9.4.1 Hypothesis and alternative
    - 9.4.2 Type I and Type II errors: level of significance
    - 9.4.3 Level *α* tests: general approach

Statistical inference

In this lecture, we learn how

1. *to estimate parameters* of the distribution.

   Methods we have seen before mostly concern measure of location (mean, median, quantiles) and variability (variance, standard deviation, interquartile range).

   As we know, this does not cover all possible parameters, and thus, we still lack a **general methodology** of estimation.

Statistical inference

In this lecture, we learn how

2. *to construct confidence intervals*.

Any estimator, computed from a collected random sample instead of the whole population, is understood as only an **approximation** of the corresponding parameter.

Instead of one estimator that is subject to a sampling error, it is often more reasonable to produce an **interval** that will contain the true population parameter with a certain known high probability.

Statistical inference

In this chapter, we learn how

3. *to test hypotheses*.

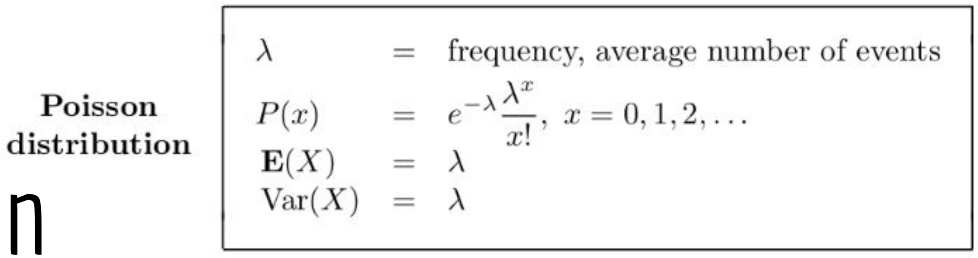
That is, we shall use the collected sample to **verify** statements and claims about the population. As a result of each test, a statement is either rejected on basis of the observed data or accepted (not rejected).

Sampling error in this analysis results in a possibility of wrongfully accepting or rejecting the hypothesis; however, we can **design tests** to control the probability of such errors.

9.1 Parameter estimation

By now, we have learned a few elementary ways to determine the family of distributions. We take into account the nature of our data, basic description, and range; propose a suitable family of distributions; and support our conjecture by looking at a histogram.

In this section, we learn how to **estimate parameters of distributions**. As a result, a large family will be reduced to just one distribution that we can use for performance evaluation, forecasting, etc.



Poisson distribution

λ	$=$	frequency, average number of events
$P(x)$	$=$	$e^{-\lambda}\frac{\lambda^x}{x!}, \ x = 0, 1, 2, \ldots$
$\mathbf{E}(X)$	$=$	λ
$\mathrm{Var}(X)$	$=$	λ

9.1 Parameter estimation

Example 9.1 (Poisson). For example, consider a sample of computer chips with a certain type of rare defects. The number of defects on each chip is recorded. This is the number of rare events, and thus, it should follow a Poisson distribution with some parameter λ.

We know that λ = $\mathbf{E}(X)$ is the expectation of a Poisson variable (Section 3.4.5). Then, should we estimate it with a sample mean $\overline{X}$? Or, should we use a sample variance s^2 because λ also equals $\mathrm{Var}(X)$?

9.1 Parameter estimation

Questions raised in these examples **do not have unique answers**. Statisticians developed a number of estimation techniques, each having certain optimal properties.

Two rather popular methods are discussed in this section:

- method of moments, and
- method of maximum likelihood.

9.1.1 Method of moments

DEFINITION 9.1

The **k-th population moment** is defined as

$$\mu_k = \mathbf{E}(\boldsymbol{X}^k).$$

The **k-th sample moment**

$$m_k = 1/n \sum_{i=1:n} \mathbf{X}_i^k$$

estimates μ_k from a sample $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$.

The first sample moment is the sample mean $\bar{\mathbf{X}}$.

9.1.1 Method of moments

Moments describe how the probability mass of a random variable is distributed.

- The zeroth moment, total mass, quantifies the fact that all distribution’s have a total mass of one.
- The first moment, the mean, specifies the distribution’s location, shifting the center of mass left or right.
- The second moment, variance, specifies the scale or spread; loosely speaking, flatter or more spread out distributions are “more random”.

https://gregorygundersen.com/blog/2020/04/11/moments/

9.1.1 Method of moments

Moments describe how the probability mass of a random variable is distributed.

- The third moment, skewness (*çarpıklık*), quantifies the relative size of the two tails of a distribution; the sign indicates which tail is bigger and the magnitude indicates by how much.
- The fourth moment, kurtosis (*basıklık*), captures the absolute size of the two tails.

https://gregorygundersen.com/blog/2020/04/11/moments/

9.1.1 Method of moments

Higher standardized moments simply recapitulate (*özetlemek*) the information in skewness and kurtosis; by convention, we ignore these in favor of the third and fourth standardized moments.

Moments are important theoretically because they provide an alternative way to fully and uniquely specify a probability distribution, a fact that is intuitive if you understand how moment’s quantify a distribution’s location, spread, and shape.

https://gregorygundersen.com/blog/2020/04/11/moments/

9.1.1 Method of moments

Central moments are computed similarly, after centralizing the data, that is, subtracting the mean.

DEFINITION 9.2

For $k \geq 2$, the **k-th population central moment** is defined as

$$\mu'_k = \mathbf{E}(X - \mu_1)^k \ .$$

The **k-th sample central moment**

$$m'_k = 1/n \sum_{i=1:n}(\mathbf{X}_i - \bar{\mathbf{X}})^k$$

estimates μ'_k from a sample $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$.

9.1.1 Method of moments

Remark: The second population central moment is variance Var($\mathcal{X}$). The second sample central moment is sample variance, although (n−1) in its denominator is now replaced by n. We mentioned that estimation methods are not unique. For unbiased estimation of σ^2 = Var($\mathcal{X}$), we use

$$s^2 = 1/(n-1) \sum_{i=1:n}(\mathbf{X}_i - \bar{\mathbf{X}})^2$$

However, method of moments and method of maximum likelihood produce a different version,

$$S^2 = m'_2 = 1/n \sum_{i=1:n}(\mathbf{X}_i - \bar{\mathbf{X}})^2$$

We’ll see other estimates of σ^2 as well.

9.1.1 Method of moments

Estimation

Method of moments is based on a simple idea. Since our sample comes from a family of distributions $\{F(\theta)\}$, we choose such a member of this family whose properties are close to properties of our data. Namely, we shall match the moments. To estimate k parameters, equate the first k population and sample moments,

$$\begin{cases} \mu_1 = m_1 \\ \quad\vdots \\ \mu_k = m_k \end{cases}$$

9.1.1 Method of moments

Estimation

$$\begin{cases} \mu_1 = m_1 \\ \vdots \\ \mu_{k\,=} = m_k \end{cases}$$

The left-hand sides of these equations depend on the distribution parameters. The right-hand sides can be computed from data. The **method of moments estimator** is the solution of this system of equations.

9.1.1 Method of moments

Example 9.3 (Poisson). To estimate parameter λ of Poisson(λ) distribution, we recall that

$$\mu_1 = \mathbf{E}(\boldsymbol{X}) = \lambda.$$

There is only one unknown parameter, hence we write one equation,

$$\text{(population) } \mu_1 = \lambda = m_1 = \overline{X}. \text{ (sample)}$$

“Solving” it for λ, we obtain

$$\hat{\lambda} = \overline{X},$$

the method of moments estimator of λ.

9.1.1 Method of moments

Simplicity is the main attractive feature of the method of moments. If it is easier, one may opt to equate central moments.

9.1.1 Method of moments

Example 9.5 (Pareto). A two-parameter Pareto distribution has a cdf

$$F(\boldsymbol{x}) = 1 - (\boldsymbol{x}/\sigma)^{-\theta} \text{ for } \boldsymbol{x} > \sigma.$$

How should we compute method of moments estimators of σ and θ?

We have not seen Pareto distribution so far, so we'll have to compute its first two moments.

9.1.1 Method of moments

Example 9.5 (Pareto). $F(\boldsymbol{x}) = 1 - (\boldsymbol{x}/\sigma)^{-\theta}$ for $x > \sigma$

We start with the density

$$f(\boldsymbol{x}) = F'(\boldsymbol{x}) = \theta/\sigma\ (\boldsymbol{x}/\sigma)^{-\theta-1} = \theta\sigma^{\theta}\boldsymbol{x}^{-\theta-1}$$

and use it to find the expectation

$$\mu_1 \quad = \mathbf{E}(\mathcal{X}) = \int_{\sigma}\boldsymbol{x}f(\boldsymbol{x})d\boldsymbol{x} = \theta/\sigma^{\theta}\int_{\sigma}\boldsymbol{x}^{-\theta}d\boldsymbol{x}$$

$$= \theta/\sigma^{\theta}(\boldsymbol{x}^{-\theta+1})/\ (-\theta+1)\ |_{\boldsymbol{x}=\sigma} = (\theta\sigma)\ /\ (\theta-1), \text{ for } \theta>1.$$

9.1.1 Method of moments

Example 9.5 (Pareto). $F(\boldsymbol{x}) = 1 - (\boldsymbol{x}/\sigma)^{-\theta}$ for $x > \sigma$

$\mu_1 = \theta\sigma / (\theta-1)$, for $\theta>1$.

and the second moment

$\mu_2 = \mathbf{E}(\boldsymbol{\mathcal{X}}^2) = \int_\sigma \boldsymbol{x}^2 f(\boldsymbol{x}) d\boldsymbol{x} = \theta/\sigma^\theta \int_\sigma \boldsymbol{x}^{-\theta+1} d\boldsymbol{x} = (\theta\sigma^2) / (\theta-2)$, for $\theta>2$.

For $\theta\leq1$, a Pareto variable has an infinite expectation, and for $\theta\leq2$, it has an infinite second moment.

9.1.1 Method of moments

Example 9.5 (Pareto). $F(\boldsymbol{x}) = 1 - (\boldsymbol{x}/\sigma)^{-\theta}$ for $x > \sigma$

$\mu_1 = (\theta\sigma) / (\theta-1)$, for $\theta>1$.

$\mu_2 = (\theta\sigma^2) / (\theta-2)$, for $\theta>2$.

Then we solve the method of moments equations

$$\mu_1 = (\theta\sigma) / (\theta-1) = m_1$$

$$\mu_2 = (\theta\sigma^2) / (\theta-2) = m_2$$

and find that

$$\hat{\theta} = \sqrt{(m_2/(m_2-m_1^2))}+1 \text{ and } \hat{\sigma} = m_1(\hat{\theta}-1)/\hat{\theta}$$

9.1.1 Method of moments

On rare occasions, when k equations are not enough to estimate k parameters, we'll consider higher moments.

Example 9.6 (Normal). Suppose we already know the mean μ of a Normal distribution and would like to estimate the variance σ^2. Only **one** parameter σ^2 is unknown; however, the first method of moments equation

$$\mu_1 = m_1$$

does not contain σ^2 and therefore does not produce its estimate. We then consider the second equation, say,

$$\mu'_2 = \sigma^2 = m'_2 = S^2 \text{ which gives us } \hat{\sigma}^2 = S^2.$$

9.1.2 Method of maximum likelihood

Another interesting idea is behind the **method of maximum likelihood estimation**.

Since the sample $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ has already been observed, we find such parameters that maximize the probability (likelihood) for this to happen.

In other words, we make the event that has already happened to be as likely as possible. This is yet another way to make the chosen distribution consistent with the observed data.

9.1.2 Method of maximum likelihood

DEFINITION 9.3 **Maximum likelihood estimator** is the parameter value that **maximizes the likelihood** of the observed sample. For a discrete distribution, we maximize the joint pmf of data P($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$). For a continuous distribution, we maximize the joint density f($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$).

9.1.2 Method of maximum likelihood

Discrete case

For a discrete distribution, the probability of a given sample is the joint pmf of data,

$$P\{\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)\} = P(\mathbf{X}) = P(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \prod_{i=1:n} P(\mathbf{X}_i)$$

because in a simple random sample, all observed $\mathbf{X}_i$ are independent.

To maximize this likelihood, we consider the critical points by taking derivatives with respect to all unknown parameters and equating them to 0.

9.1.2 Method of maximum likelihood

Discrete case

The maximum can only be attained at such parameter values θ where the derivative $\partial/\partial\theta\ P(\mathbf{X})$ equals 0, or, where it does not exist, at the boundary of the set of possible values of θ.

A nice computational shortcut is to take logarithms first. Differentiating the sum

$$\ln \prod_{i=1:n} P(\mathbf{X}_i) = \sum_{i=1:n} \ln P(\mathbf{X}_i)$$

is easier than differentiating the product. Besides, log is an increasing function, so the likelihood $P(\mathbf{X})$ and the log-likelihood $\ln P(\mathbf{X})$ are maximized by exactly the same parameters.

$$\ln \prod_{i=1}^{n} P(X_i) = \sum_{i=1}^{n} \ln P(X_i)$$

9.1.2 Method of maximum likelihood

Example 9.7 (Poisson). The pmf of Poisson distribution is

$$P(x) = e^{-\lambda} \frac{\lambda^x}{x!},$$

and its logarithm is $\ln P(\boldsymbol{x}) = -\lambda + \boldsymbol{x}\ln\lambda - \ln(\boldsymbol{x}!)$.

Thus, we need to maximize

$$\ln P(\mathbf{X}) = \sum_{i=1:n} -(\lambda + \mathbf{X}_i \ln\lambda) + C = -n\lambda + \ln\lambda \sum_{i=1:n} \mathbf{X}_i + C$$

where $C = -\sum \ln(\boldsymbol{x}!)$ is a constant that does not contain the unknown parameter λ.

9.1.2 Method of maximum likelihood

Example 9.7 (Poisson).

Find the critical point(s) of this log-likelihood. Differentiating it and equating its derivative to 0, we get

$$\partial/\partial\lambda \ln P(\mathbf{X}) = -n + 1/\lambda \sum_{i=1:n} \mathbf{X}_i = 0.$$

This equation has only one solution

$$\hat{\lambda} = 1/n \sum_{i=1:n} \mathbf{X}_i = \bar{\mathbf{X}}$$

Since this is the only critical point, and the likelihood converges to 0 as $\lambda \downarrow 0$ or $\lambda \uparrow \infty$, we conclude that $\hat{\lambda}$ is the maximizer. Therefore, it is the maximum likelihood estimator of λ.

9.1.2 Method of maximum likelihood

Continuous case

In the continuous case, the probability to observe exactly the given number X = x is 0, as we know from Chapter 4. Instead, the method of maximum likelihood will maximize the probability of observing "almost" the same number.

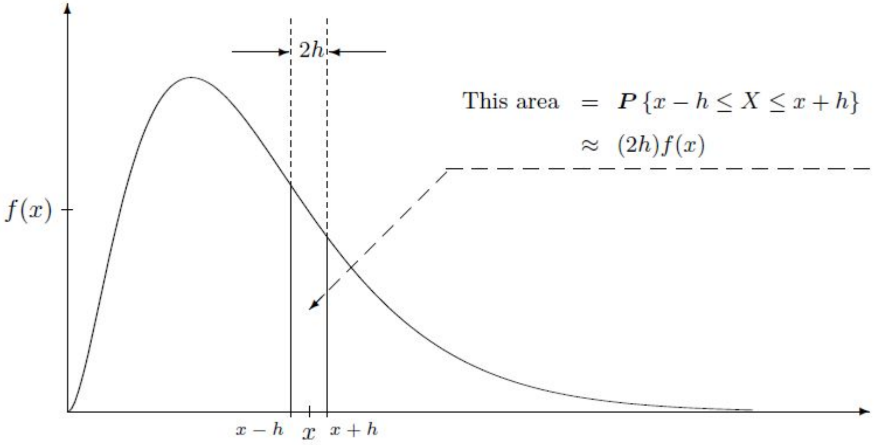


FIGURE 9.1: *Probability of observing "almost" $X = x$.*

9.1.2 Method of maximum likelihood

Continuous case

For a very small h, $P\{x - h < X < x + h\} = \int_{x-h}^{x+h} f(y)dy \approx (2h)f(x)$

The probability of observing a value close to x is proportional to the density $f(x)$. Then, for a sample $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, the maximum likelihood method will maximize the joint density $f(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$.

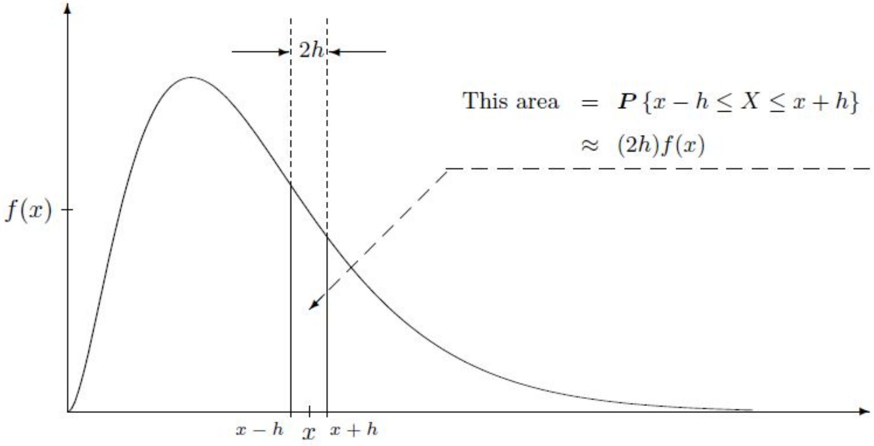


FIGURE 9.1: *Probability of observing "almost" $X = x$.*

$$\ln \prod_{i=1}^{n} P(X_i) = \sum_{i=1}^{n} \ln P(X_i)$$

9.1.2 Method of maximum likelihood

Example 9.8 (Exponential). The Exponential density is $f(\boldsymbol{x})=\lambda e^{-\lambda \boldsymbol{x}}$, so the log-likelihood of a sample can be written as

$$\ln f(\mathbf{X}) = \sum_{i=1:n}\ln(\lambda e^{-\lambda \mathbf{X}i}) = \sum_{i=1:n}(\ln\lambda - \lambda \mathbf{X}_i) = n\ln\lambda - \lambda\sum_{i=1:n}\mathbf{X}_i$$

Taking its derivative with respect to the unknown parameter λ, equating it to 0, and solving for λ, we get

$$\partial/\partial\lambda \ln f(\mathbf{X}) = n/\lambda - \sum_{i=1:n}\mathbf{X}_i = 0,$$

resulting in

$$\hat{\lambda} = n / \sum_{i=1:n}\mathbf{X}_i = 1 / \bar{\mathbf{X}}.$$

9.1.2 Method of maximum likelihood

Sometimes the likelihood has no critical points inside its domain, then it is maximized at the boundary.

Example 9.9 (Uniform). Based on a sample from Uniform(0, *b*) distribution, how can we estimate the parameter *b*?

The Uniform(0, b) density is f($\boldsymbol{x}$)=1/b for 0 ≤ $\boldsymbol{x}$ ≤ b.

It is decreasing in b, and therefore, it is maximized at the the smallest possible value of b, which is $\boldsymbol{x}$.

For a sample ($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$), the joint density f($\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$)=1/b^n for 0 ≤ $\mathbf{X}$ ≤ b also attains its maximum at the smallest possible value of b which is now the largest observation.

9.1.2 Method of maximum likelihood

Sometimes the likelihood has no critical points inside its domain, then it is maximized at the boundary.

Example 9.9 (Uniform). Based on a sample from Uniform(0, *b*) distribution, how can we estimate the parameter *b*?

Indeed, b≥$\mathbf{X}_i$ for all i only if b≥max($\mathbf{X}_i$). If b<max($\mathbf{X}_i$), then f($\mathbf{X}$)=0, and this cannot be the maximum value. Therefore, the maximum likelihood estimator is $\hat{b}$ = max($\mathbf{X}_i$).

9.1.2 Method of maximum likelihood

Example 9.10 (Pareto). For the Pareto distribution in Example 9.5, the log-likelihood is

$$\ln f(\mathbf{X}) = \sum_{i=1:n}\ln(\theta\sigma^{\theta}\mathbf{X}_i^{-\theta-1}) = n\ln\theta + n\theta\ln\sigma - (\theta+1)\sum_{i=1:n}\ln\mathbf{X}_i$$

for $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \geq \sigma$. Maximizing this function over both σ and θ, we notice that it always increases in σ. Thus, we estimate σ by its largest possible value, which is the smallest observation,

$$\hat{\sigma} = \min(\mathbf{X}_i).$$

We can substitute this value of σ into the log-likelihood and maximize with respect to θ,

9.1.2 Method of maximum likelihood

Example 9.10 (Pareto).

$$\ln f(\mathbf{X}) = \sum_{i=1:n} \ln(\theta\sigma^{\theta}\mathbf{X}_i^{-\theta-1}) = n\ln\theta + n\theta\ln\sigma - (\theta+1)\sum_{i=1:n}\ln\mathbf{X}_i$$

we estimate σ by its largest possible value, which is the smallest observation,

$$\hat{\sigma} = \min(\mathbf{X}_i).$$

We can substitute this value of σ into the log-likelihood and maximize with respect to θ,

$$\partial/\partial\lambda\theta \ln f(\mathbf{X}) = n/\theta + n \ln\hat{\sigma} - \sum_{i=1:n}\ln\mathbf{X}_i = 0$$

$$\hat{\theta} = n / (\sum_{i=1:n}\ln\mathbf{X}_i - n \ln \hat{\sigma}) = n / (\sum_{i=1:n}\ln(\mathbf{X}_i/\hat{\sigma}))$$

9.1.2 Method of maximum likelihood

Maximum likelihood estimators are rather popular because of their nice properties.

Under mild conditions, these estimators are consistent, and for large samples, they have an approximately Normal distribution.

Often in complicated problems, finding a good estimation scheme may be challenging whereas the maximum likelihood method always gives reasonable solution.

9.1.3 Estimation of standard errors

How good are the estimators that we learned today?

Standard errors can serve as measures of their accuracy. To estimate them, we derive an expression for the standard error and estimate all the unknown parameters in it.

9.1.3 Estimation of standard errors

Example 9.11 In Examples 9.3 and 9.7, we found the method of moments and maximum likelihood estimators of the Poisson parameter λ. Both estimators appear to be equal the sample mean $\hat{\lambda} = \bar{\mathbf{X}}$. Let us now estimate the standard error of $\hat{\lambda}$.

Solution. There are at least two ways to do it.

On one hand, $\sigma=\sqrt{\lambda}$ for the Poisson(λ) distribution, so

$$\sigma(\hat{\lambda}) = \sigma(\bar{X}) = \sigma/\sqrt{n} = \sqrt{(\lambda/n)}.$$

Estimating λ by $\bar{X}$, we obtain $s_1(\hat{\lambda}) = \sqrt{(\bar{\mathbf{X}}/n)} = \sqrt{(\Sigma \mathbf{X}_i)}/n$

9.1.3 Estimation of standard errors

Example 9.11 In Examples 9.3 and 9.7, we found the method of moments and maximum likelihood estimators of the Poisson parameter λ. Both estimators appear to be equal the sample mean $\hat{\lambda} = \bar{\mathbf{X}}$. Let us now estimate the standard error of $\hat{\lambda}$.

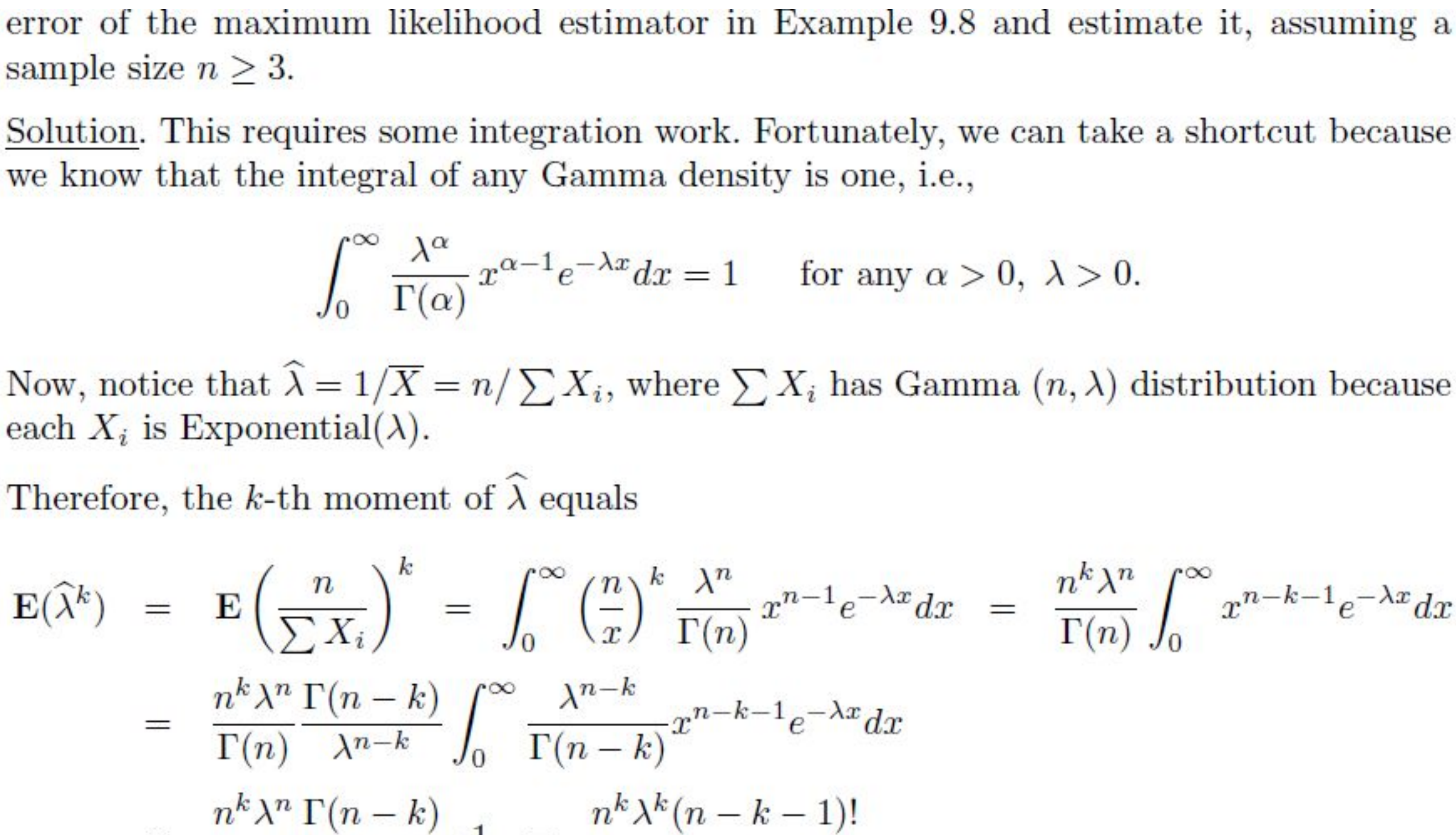
Solution. There are at least two ways to do it.

On the other hand, we can use the sample standard deviation and estimate the standard error of the sample mean as in Example 8.17

$$s_2(\hat{\lambda}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n(n-1)}}$$

9.1.3 Estimation of standard errors

Example 9.12 (Estimation of the Exponential parameter). Derive the standard



error of the maximum likelihood estimator in Example 9.8 and estimate it, assuming a sample size $n \geq 3$.

Solution. This requires some integration work. Fortunately, we can take a shortcut because we know that the integral of any Gamma density is one, i.e.,

$$\int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx = 1 \qquad \text{for any } \alpha > 0,\ \lambda > 0.$$

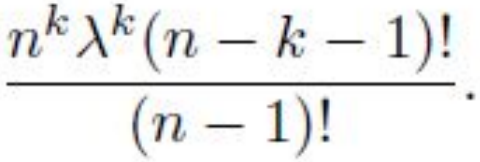
Now, notice that $\widehat{\lambda} = 1/\overline{X} = n/\sum X_i$, where $\sum X_i$ has Gamma (n, λ) distribution because each X_i is Exponential(λ).

Therefore, the k-th moment of $\widehat{\lambda}$ equals

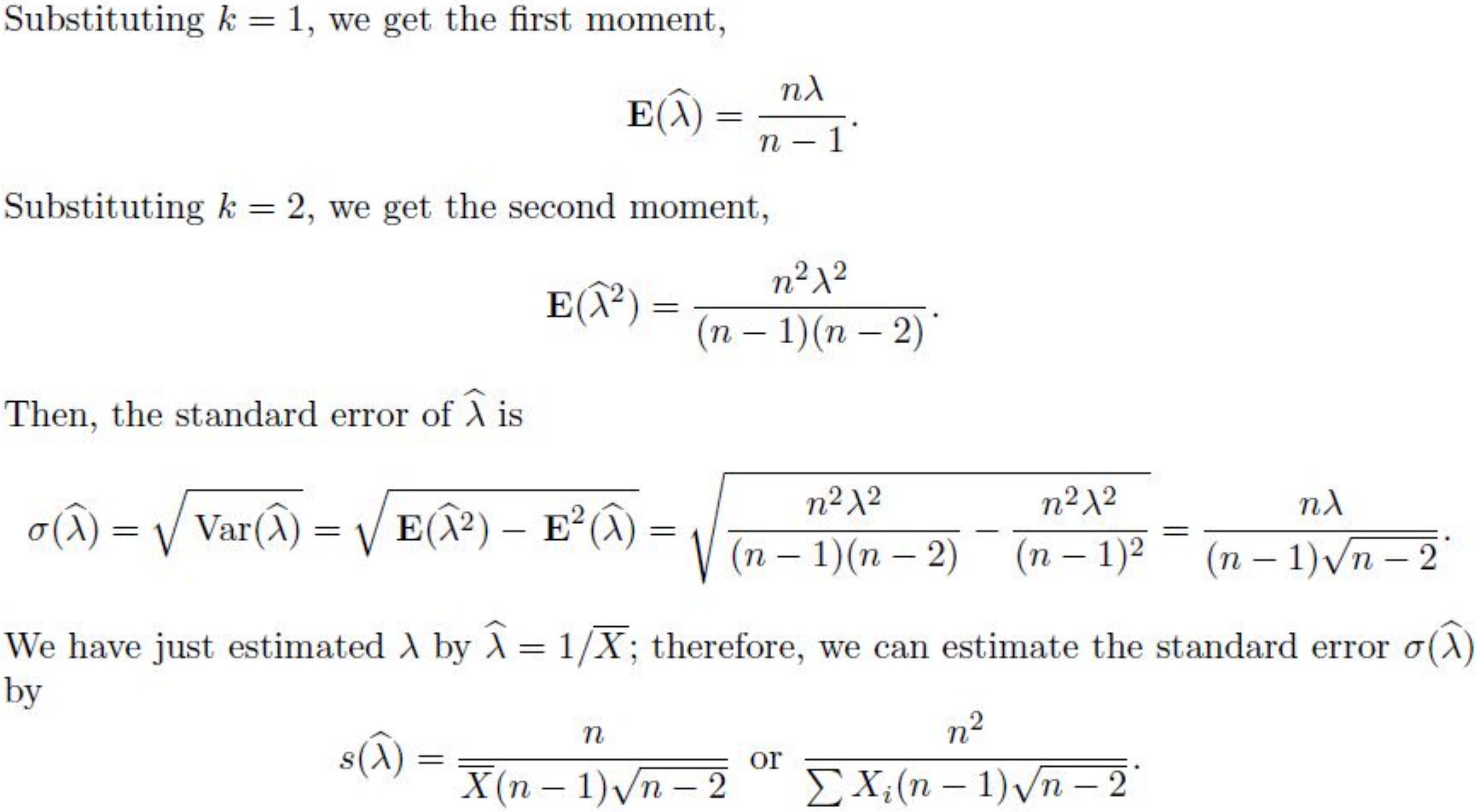
$$\begin{aligned} \mathbf{E}(\widehat{\lambda}^k) &= \mathbf{E}\left(\frac{n}{\sum X_i}\right)^k = \int_0^\infty \left(\frac{n}{x}\right)^k \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x} dx = \frac{n^k \lambda^n}{\Gamma(n)} \int_0^\infty x^{n-k-1} e^{-\lambda x} dx \\ &= \frac{n^k \lambda^n}{\Gamma(n)} \frac{\Gamma(n-k)}{\lambda^{n-k}} \int_0^\infty \frac{\lambda^{n-k}}{\Gamma(n-k)} x^{n-k-1} e^{-\lambda x} dx \end{aligned}$$

$$n^k \lambda^n\, \Gamma(n-k) \qquad n^k \lambda^k (n-k-1)!$$

9.1.3 Estimation of standard errors



$$\frac{n^k \lambda^k (n-k-1)!}{(n-1)!}.$$



Substituting $k = 1$, we get the first moment,

$$\mathbf{E}(\widehat{\lambda}) = \frac{n\lambda}{n-1}.$$

Substituting $k = 2$, we get the second moment,

$$\mathbf{E}(\widehat{\lambda}^2) = \frac{n^2\lambda^2}{(n-1)(n-2)}.$$

Then, the standard error of $\widehat{\lambda}$ is

$$\sigma(\widehat{\lambda}) = \sqrt{\operatorname{Var}(\widehat{\lambda})} = \sqrt{\mathbf{E}(\widehat{\lambda}^2) - \mathbf{E}^2(\widehat{\lambda})} = \sqrt{\frac{n^2\lambda^2}{(n-1)(n-2)} - \frac{n^2\lambda^2}{(n-1)^2}} = \frac{n\lambda}{(n-1)\sqrt{n-2}}.$$

We have just estimated λ by $\widehat{\lambda} = 1/\overline{X}$; therefore, we can estimate the standard error $\sigma(\widehat{\lambda})$ by

$$s(\widehat{\lambda}) = \frac{n}{\overline{X}(n-1)\sqrt{n-2}} \text{ or } \frac{n^2}{\sum X_i(n-1)\sqrt{n-2}}.$$

Contents

- 9.1 Introduction
    - 9.1.1 Method of moments
    - 9.1.2 Method of maximum likelihood
    - 9.1.3 Estimation of standard errors
- **9.2 Confidence intervals**
    - **9.2.1 Construction of confidence intervals: a general method**
    - **9.2.2 Confidence interval for the population mean**
    - **9.2.3 Confidence interval for the difference between two means**
- 9.4 Hypothesis testing
    - 9.4.1 Hypothesis and alternative
    - 9.4.2 Type I and Type II errors: level of significance
    - 9.4.3 Level *α* tests: general approach

9.2 Confidence intervals

When we report an estimator $\hat{\theta}$ of a population parameter θ, we know that most likely $\hat{\theta} \neq \theta$ due to a sampling error. We realize that we have estimated θ *up to some error*.

- Nobody understands the internet connection of 11 megabytes per second as exactly 11 megabytes going through the network every second.
- Nobody takes a meteorological forecast as the promise of exactly the predicted temperature.

9.2 Confidence intervals

- Then how much can we trust the reported estimator?
- How far can it be from the actual parameter of interest?
- What is the probability that it will be reasonably close?
- And if we observed an estimator $\hat{\theta}$, then what can the actual parameter θ be?

To answer these, statisticians use **confidence intervals**, which contain parameter values that deserve some confidence, given the observed data.

9.2 Confidence intervals

DEFINITION 9.4 An interval [a, b] is a (1− α)100% **confidence interval** for the parameter θ if it contains the parameter with probability (1 − α),

$$P\{a \leq \theta \leq b\} = 1 - \alpha.$$

The **coverage probability** (1−α) is also called a **confidence level**.

- Let us take a moment to think about this definition.
- The probability of a random event {a ≤ θ ≤ b} has to be (1 − α). What **randomness** is involved in this event?

9.2 Confidence intervals

DEFINITION 9.4 An interval [a, b] is a (1− α)100% **confidence interval** for the parameter θ if it contains the parameter with probability (1 − α),

$$P\{a \le \theta \le b\} = 1- \alpha.$$

The **coverage probability** (1−α) is also called a **confidence level**.

- The population parameter θ is **not** random.
- But, the **interval** is computed from random data, and therefore, it is random. The coverage probability refers to the chance that our interval covers a constant parameter θ.

9.2 Confidence intervals

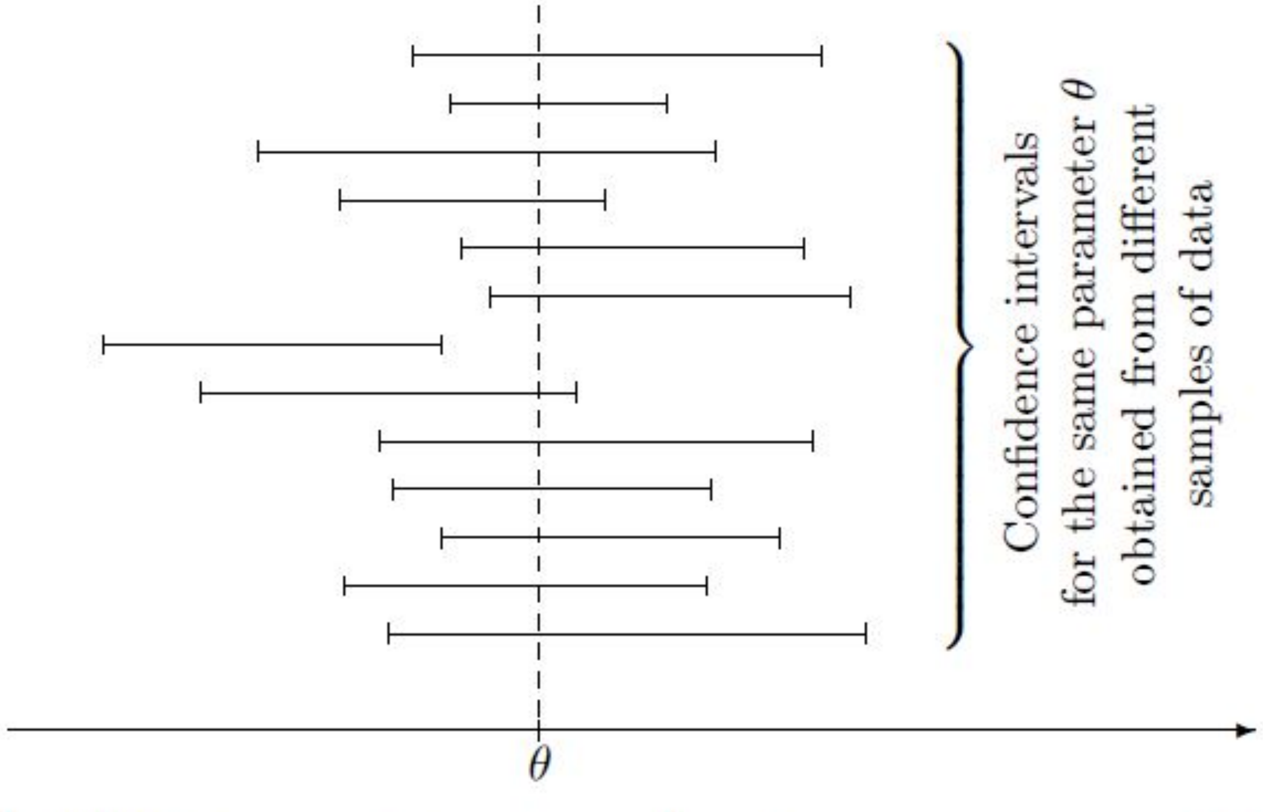


FIGURE 9.2: *Confidence intervals and coverage of parameter* θ.

9.2 Confidence intervals

Suppose that we collect many random samples and produce a confidence interval from each of them. If these are $(1-\alpha)100\%$ confidence intervals, then we expect $(1-\alpha)100\%$ of them to cover θ and $100\alpha\%$ of them to miss it.

In Figure 9.2, we see one interval that does not cover θ. No mistake was made in data collection and construction of this interval. It missed the parameter only **due to a sampling error**.

9.2 Confidence intervals

It is therefore wrong to say, “*I computed a 90% confidence interval, it is* [3, 6]. *Parameter belongs to this interval with probability* 90%.” The parameter is constant; it either belongs to the interval [3, 6] (with probability 1) or does not. In this case, 90% refers to the **proportion of confidence intervals** that contain the unknown parameter in a long run.

9.2.1 Construction of confidence intervals

Given a sample of data and a desired confidence level ($1 - \alpha$), how can we construct a confidence interval $[a, b]$ that will satisfy the coverage condition ($P\{a \le \theta \le b\} = 1 - \alpha$) ?

9.2.1 Construction of confidence intervals

We start by estimating parameter θ. Assume there is an **unbiased** estimator $\hat{\theta}$ that has a **Normal** distribution. When we standardize it, we get a Standard Normal variable

$$Z = \frac{\widehat{\theta} - \mathbf{E}(\widehat{\theta})}{\sigma(\widehat{\theta})} = \frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})},$$

where E($\hat{\theta}$) = θ because $\hat{\theta}$ is unbiased, and σ($\hat{\theta}$) is its standard error. This variable falls between the Standard Normal quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ with probability **1-α**. The quantiles are denoted as:

$$-\boldsymbol{z}_{\alpha/2} = q_{\alpha/2}$$
$$\boldsymbol{z}_{\alpha/2} = q_{1-\alpha/2}$$

9.2.1 Construction of confidence intervals

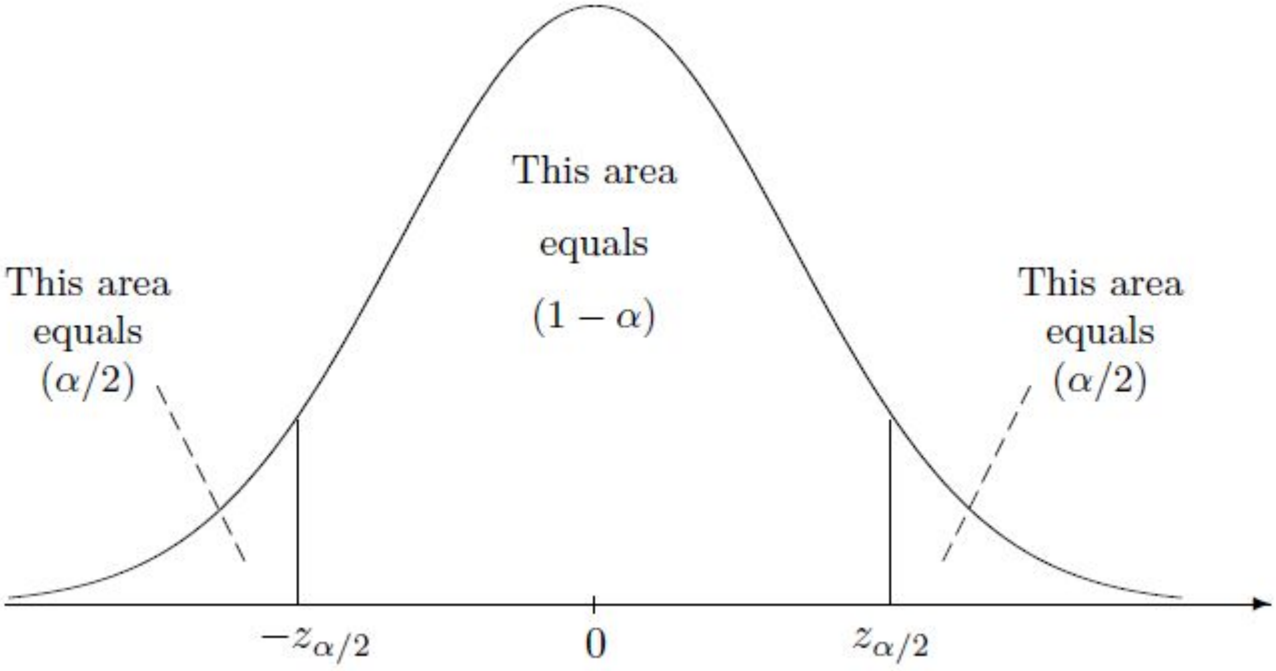

This area equals $(1-\alpha)$
This area equals $(\alpha/2)$
This area equals $(\alpha/2)$
$-z_{\alpha/2}$
0
$z_{\alpha/2}$

FIGURE 9.3: *Standard Normal quantiles* $\pm z_{\alpha/2}$ *and partition of the area under the density curve.*

9.2.1 Construction of confidence intervals

$$P\{-\boldsymbol{z}_{\alpha/2} \leq \mathcal{Z} \leq \boldsymbol{z}_{\alpha/2}\} = 1-\alpha$$

$$P\{-\boldsymbol{z}_{\alpha/2} \leq (\hat{\theta} - \theta) / \sigma(\hat{\theta}) \leq \boldsymbol{z}_{\alpha/2}\} = 1-\alpha$$

$$P\{\hat{\theta} - \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta})\} = 1-\alpha$$

Problem is solved!

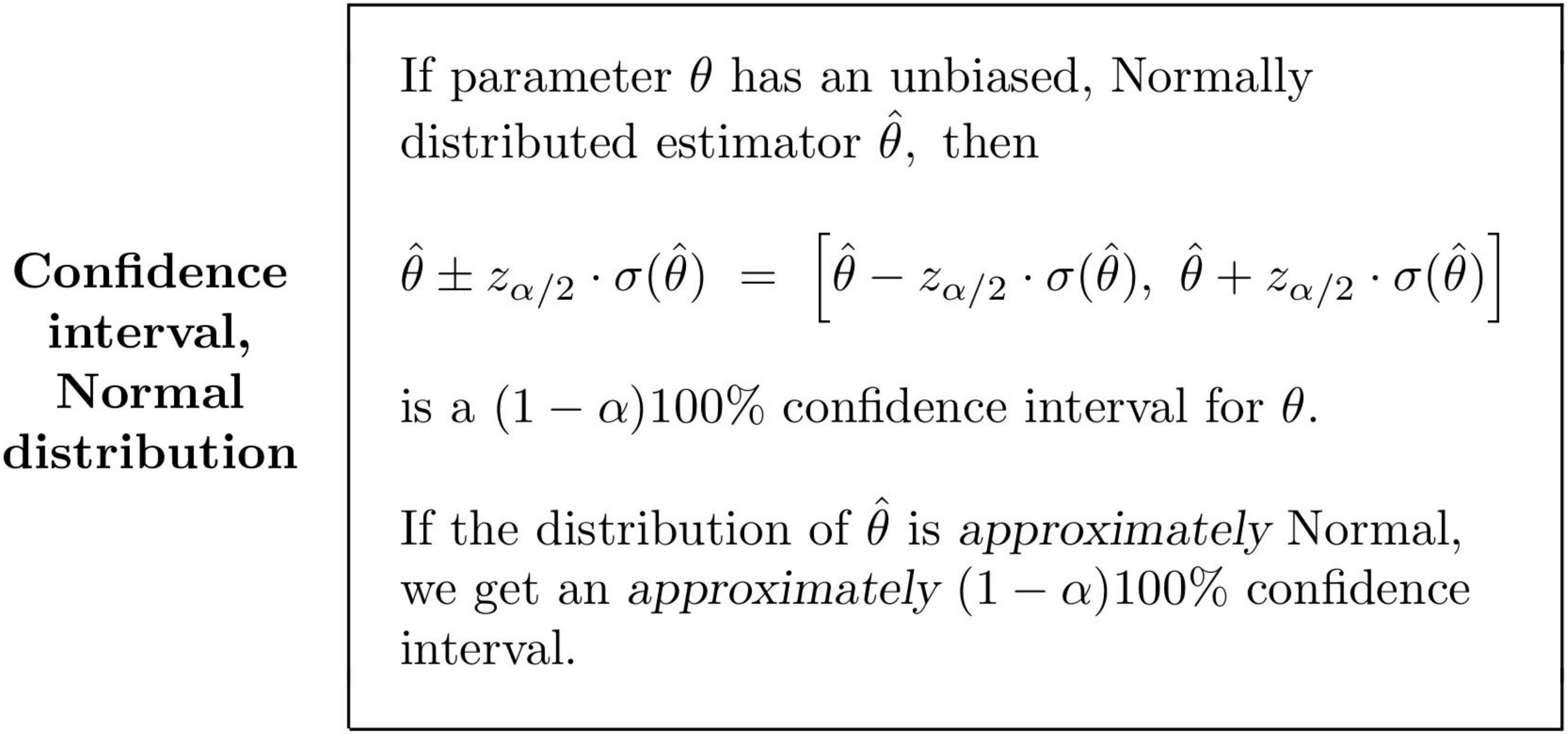
$$a = \hat{\theta} - \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta})$$

$$b = \hat{\theta} + \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta})$$

such that

$$P\{a \leq \theta \leq b\} = 1-\alpha$$

9.2.1 Construction of confidence intervals



Confidence interval, Normal distribution

If parameter θ has an unbiased, Normally distributed estimator $\hat{\theta}$, then

$$\hat{\theta} \pm z_{\alpha/2} \cdot \sigma(\hat{\theta}) \ = \ \left[\hat{\theta} - z_{\alpha/2} \cdot \sigma(\hat{\theta}),\ \hat{\theta} + z_{\alpha/2} \cdot \sigma(\hat{\theta})\right]$$

is a $(1-\alpha)100\%$ confidence interval for θ.

If the distribution of $\hat{\theta}$ is *approximately* Normal, we get an *approximately* $(1-\alpha)100\%$ confidence interval.

9.2.1 Construction of confidence intervals

$$P\{\hat{\theta} - \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta}) \leq \theta \leq \hat{\theta} + \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta})\} = 1-\alpha$$

- In this formula, $\hat{\theta}$ is the **center of the interval**, and $z_{\alpha/2} \cdot \sigma(\hat{\theta})$ is **the margin**.
- The margin of error is often reported along with poll and survey results.
- In newspapers and press releases, it is usually computed for a 95% confidence interval.

9.2.1 Construction of confidence intervals

- We have seen quantiles ± $\boldsymbol{z}_{\alpha/2}$ in inverse problems (Example 4.12 on p. 91).
- Now, in **confidence estimation**, and also, in the next section on **hypothesis testing**, they will play a crucial role as we'll need to attain the desired confidence level α. The most commonly used values are
  - $\boldsymbol{z}_{0.10} = 1.282$
  - $\boldsymbol{z}_{0.05} = 1.645$
  - $\boldsymbol{z}_{0.025} = 1.960$
  - $\boldsymbol{z}_{0.01} = 2.326$
  - $\boldsymbol{z}_{0.005} = 2.576$

$\boldsymbol{z}_{\alpha} = q_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the value of a Standard Normal variable $\mathcal{Z}$ that is exceeded with probability α.

9.2.1 Construction of confidence intervals

Several important applications of this general method are discussed below. In each problem, we

a. find an unbiased estimator of θ,

b. check if it has a Normal distribution,

c. find its standard error $\sigma(\hat{\theta}) = \mathrm{Std}(\hat{\theta})$,

d. obtain quantiles $\pm \boldsymbol{z}_{\alpha/2}$ from the table of Normal distribution (Table A4 in the Appendix), and finally,

e. apply the rule: $[\hat{\theta} - \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta}),\ \hat{\theta} + \boldsymbol{z}_{\alpha/2} \cdot \sigma(\hat{\theta})]$ is a $(1-\alpha)100\%$ confidence interval for θ.

9.2.2 Confidence interval for the population mean

Let us construct a confidence interval for the population mean

$$\theta = \mu = \mathbf{E}(X).$$

Start with an estimator: $\hat{\theta} = \bar{X} = 1/n \sum_{i=1:n} X_i$

1. If a sample X comes from Normal distribution, then $\bar{X}$ is also Normal, and the rule can be applied.
2. If a sample comes from any distribution, but **the sample size n is large**, then $\bar{X}$ has an approximately Normal distribution according to the Central Limit Theorem. Then the rule gives an **approximately** $(1-\alpha)100\%$ confidence interval.

9.2.2 Confidence interval for the population mean

We have previously derived

$$E(\bar{X}) = \mu$$

$$\sigma(\bar{X}) = \sigma/\sqrt{n}.$$

Then, the rule reduces to the following (1−α)100% confidence interval for μ:

$$\bar{X} \pm \boldsymbol{z}_{\alpha/2} \cdot (\sigma/\sqrt{n})$$

9.2.2 Confidence interval for the population mean

Example 9.13. Construct a 95% confidence interval for the population mean based on a sample of measurements

2.5, 7.4, 8.0, 4.5, 7.4, 9.2

if measurement errors have Normal distribution, and the measurement device guarantees a standard deviation of σ=2.2.

Solution. This sample has size n=6 and sample mean **$\bar{X}$ = 6.50**. To attain a confidence level of 1−α=0.95, we need α=0.05 and α/2=0.025. Considering $\boldsymbol{z}_{0.025}$=1.960:

$\bar{X} \pm \boldsymbol{z}_{\alpha/2} \cdot (\sigma/\sqrt{n}) = 6.50 \pm (1.960) \cdot (2.2/\sqrt{6}) =$ **6.50 ± 1.76** = **[4.74, 8.26]**

9.2.3 Confidence interval for the difference between two means

Under the same conditions as in the previous section,

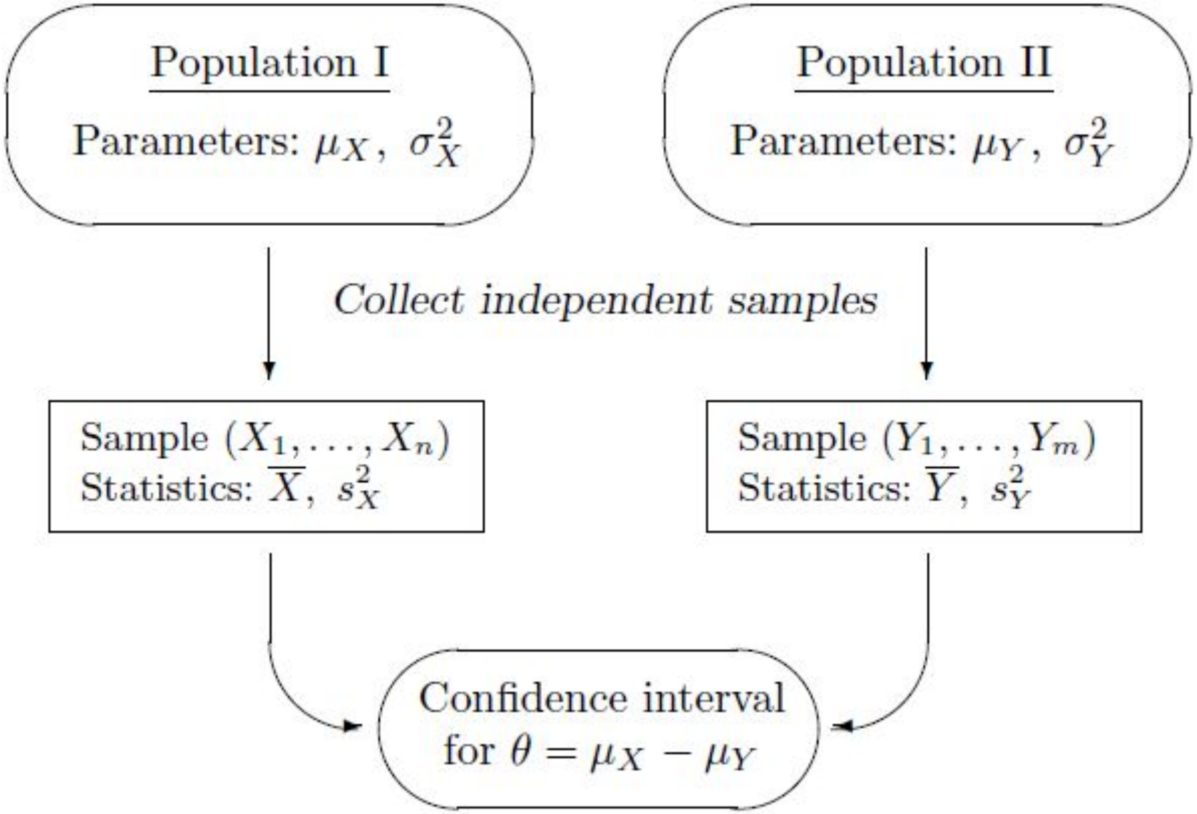
- Normal distribution of data or
- sufficiently large sample size,

we can construct a confidence interval for the difference between two means.

This problem arises when we compare two populations.

It may be a comparison of two materials, two suppliers, two service providers, two communication channels, two labs, etc.

9.2.3 Confidence interval for the difference between two means


Population I
Parameters: μ_X, σ_X^2
Population II
Parameters: μ_Y, σ_Y^2
Collect independent samples
Sample $(X_1, \ldots, X_n)$
Statistics: $\overline{X}$, s_X^2
Sample $(Y_1, \ldots, Y_m)$
Statistics: $\overline{Y}$, s_Y^2
Confidence interval
for $\theta = \mu_X - \mu_Y$

9.2.3 Confidence interval for the difference between two means

- $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ from one population,
- $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$ from the other population.

Suppose that the two samples are collected independently of each other.

To construct a confidence interval for the difference between population means

$$\theta = \mu_{\mathcal{X}} - \mu_{\mathcal{Y}},$$

we complete the steps (a)–(e).

9.2.3 Confidence interval for the difference between two means

a. Propose an estimator of θ: $\hat{\theta} = \bar{X} - \bar{Y}$.

- It is natural to come up with this estimator because $\bar{X}$ estimates $\mu_{\mathcal{X}}$ and $\bar{Y}$ estimates $\mu_{\mathcal{Y}}$.

b. Check that $\hat{\theta}$ is unbiased.

$$\mathbf{E}(\hat{\theta}) = \mathbf{E}(\bar{X} - \bar{Y}) = \mathbf{E}(\bar{X}) - \mathbf{E}(\bar{Y}) = \mu_{\mathcal{X}} - \mu_{\mathcal{Y}} = \theta$$

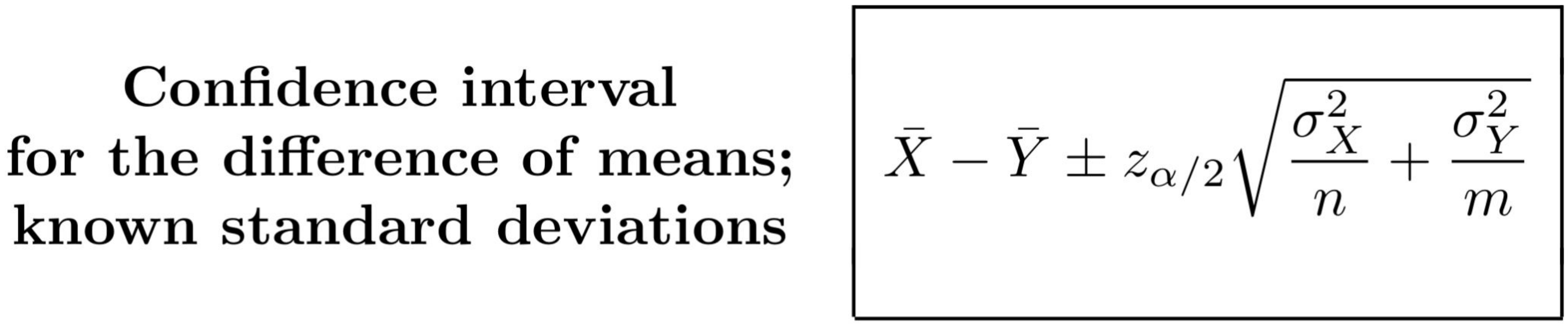
c. Check that $\hat{\theta}$ has a Normal or approximately Normal distribution. This is true if the observations are Normal or both sample sizes m and n are large.

9.2.3 Confidence interval for the difference between two means

d. Find the standard error of θ (using independence of X and Y)

$$\sigma(\hat{\theta}) = \sqrt{\mathbf{Var}(\bar{X} - \bar{Y})} = \sqrt{\mathbf{Var}(\bar{X}) + \mathbf{Var}(\bar{Y})} = \sqrt{\sigma_{x}^{2}/n + \sigma_{y}^{2}/m}$$

e. Find quantiles $\pm \boldsymbol{z}_{\alpha/2}$ and compute the confidence interval. This results in the following formula.



Confidence interval for the difference of means; known standard deviations

$$\boxed{\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

9.2.3 Confidence interval for the difference between two means

Example 9.14 (Effect of an upgrade). A manager evaluates effectiveness of a major hardware upgrade by running a certain process 50 times before the upgrade and 50 times after it. Based on these data, the average running time is 8.5 minutes before the upgrade, 7.2 minutes after it. Historically, the standard deviation has been 1.8 minutes, and presumably it has not changed. Construct a **90% confidence interval** showing how much the mean running time reduced due to the hardware upgrade.

9.2.3 Confidence interval for the difference between two means

Solution. We have n=m=50, $\sigma_{\chi}=\sigma_{y}=1.8$, $\bar{X} = 8.5$, and $\bar{Y} = 7.2$. Also, the confidence level (1−α) equals 0.9, hence α/2=0.05, and $\boldsymbol{z}_{\alpha/2} = \boldsymbol{z}_{0.05} = 1.645$.

The distribution of times may not be Normal; however, due to large sample sizes, the estimator $\hat{\theta} = \bar{X} - \bar{Y}$ is approximately Normal by the Central Limit Theorem. Thus, the derived formula is applicable, and a 90% confidence interval for the difference of means $(\mu_{\chi} - \mu_{y})$ is

$8.5 - 7.2 \pm (1.645)\sqrt{1.8^2/50 + 1.8^2/50)} = \mathbf{1.3 \pm 0.6} = \mathbf{[0.7, 1.9]}$

Contents

- 9.1 Introduction
    - 9.1.1 Method of moments
    - 9.1.2 Method of maximum likelihood
    - 9.1.3 Estimation of standard errors
- 9.2 Confidence intervals
    - 9.2.1 Construction of confidence intervals: a general method
    - 9.2.2 Confidence interval for the population mean
    - 9.2.3 Confidence interval for the difference between two means
- **9.4 Hypothesis testing**
    - **9.4.1 Hypothesis and alternative**
    - **9.4.2 Type I and Type II errors: level of significance**
    - **9.4.3 Level α tests: general approach**

9.4 Hypothesis testing

A vital role of Statistics is in verifying statements, claims, conjectures, and in general - **testing hypotheses**. Based on a random sample, we can use Statistics to verify whether

- a system has not been infected,
- a hardware upgrade was efficient,
- the average number of concurrent users increased by 2000 this year,
- the average connection speed is 54 Mbps, as claimed by the internet service provider,
- the number of errors in software is independent of the manager's experience, etc.

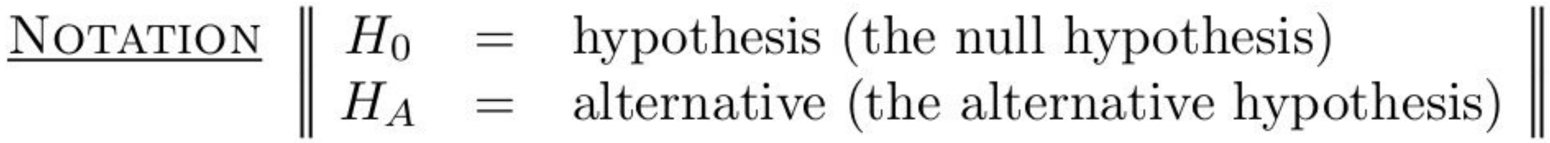
9.4 Hypothesis testing

Testing statistical hypotheses has wide applications far beyond Computer Science. These methods are used

- to prove efficiency of a new medical treatment,
- safety of a new automobile brand,
- innocence of a defendant,
- authorship of a document,
- to establish cause-and-effect relationships,
- to identify factors that can significantly improve the response,
- to detect information leaks,
- and so forth.

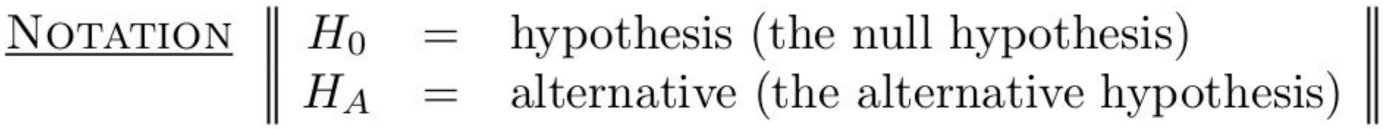
9.4.1 Hypothesis and alternative

To begin, we need to state exactly what we are testing. These are **hypothesis** and **alternative**.



NOTATION ‖ H_0 = hypothesis (the null hypothesis)
H_A = alternative (the alternative hypothesis) ‖

- *H_0* and *H_A* are simply two mutually exclusive statements. Each test results either in acceptance of *H_0* or its rejection in favor of *H_A*.



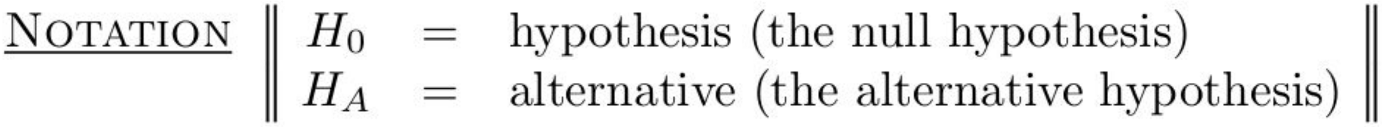
NOTATION

H_0 = hypothesis (the null hypothesis)

H_A = alternative (the alternative hypothesis)

9.4.1 Hypothesis and alternative

- A **null hypothesis** is always an equality, absence of an effect or relation, some “normal,” usual statement that people have believed in for years.
- In order to overturn the common belief and to reject the hypothesis, we need **significant evidence**. Such evidence can only be provided by data.
- Only when such evidence is found, and when it strongly supports the alternative H_A, can the hypothesis H_0 be rejected in favor of H_A.



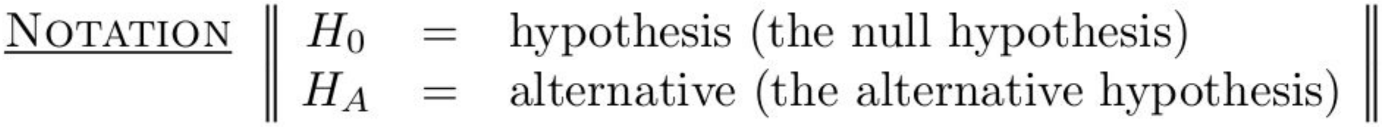
NOTATION

H_0 = hypothesis (the null hypothesis)

H_A = alternative (the alternative hypothesis)

9.4.1 Hypothesis and alternative

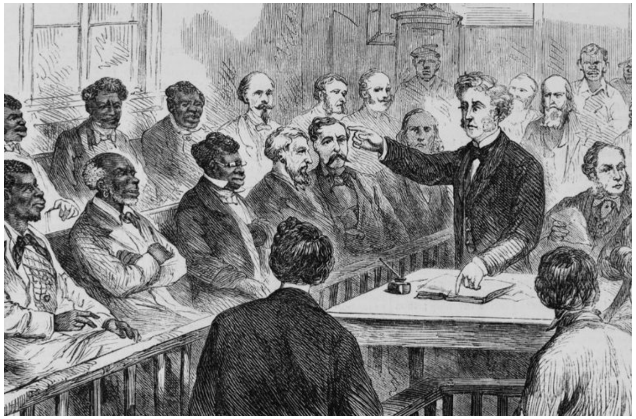
- Based on a random sample, a statistician cannot tell whether the hypothesis is true or the alternative. We need to see the entire population to tell that.
- The purpose of each test is to determine whether the data provides sufficient evidence **against** H_0 **in favor** of H_A.



NOTATION ‖ H_0 = hypothesis (the null hypothesis) ‖
H_A = alternative (the alternative hypothesis) ‖

9.4.1 Hypothesis and alternative

This is similar to a criminal trial. Typically, the jury cannot tell whether the defendant committed a crime or not. It is not their task. They are only required to determine if the presented evidence against the defendant is sufficient and convincing. By default, called *(varsayım)* **presumption of innocence**, insufficient evidence leads to acquittal *(beraat)*.



9.4.1 Hypothesis and alternative

Example 9.22. To verify that the the average connection speed is 54 Mbps, we test the hypothesis $H_0 : \mu = 54$ against the two-sided alternative $H_A : \mu \neq 54$, where μ is the average speed of all connections.

However, if we worry about a low connection speed only, we can conduct a one-sided test of

$$H_0 : \mu = 54 \quad \text{vs} \quad H_A : \mu < 54.$$

In this case, we only measure the amount of evidence supporting the one-sided alternative $H_A : \mu < 54$. In the absence of such evidence, we gladly accept the null hypothesis.

9.4.1 Hypothesis and alternative

DEFINITION 9.6

Alternative of the type H_A : $\mu \neq \mu_0$ covering regions on both sides of the hypothesis ($H_0 : \mu = \mu_0$) is a **two-sided alternative**.

Alternative H_A : $\mu < \mu_0$ covering the region to the left of H_0 is **one-sided, left-tail**.

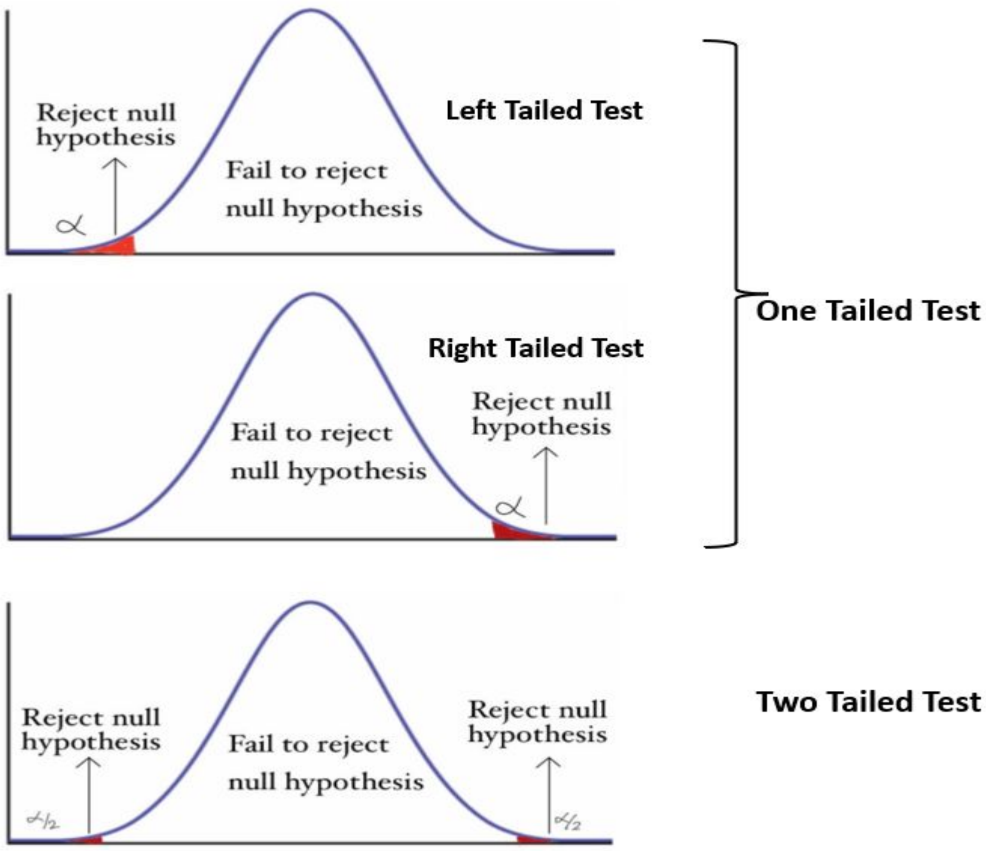
Alternative H_A : $\mu > \mu_0$covering the region to the right of H_0 is **one-sided, right-tail**.

9.4.1 Hypothesis and alternative

Null Hypothesis(H0) :$\mu \geq value$
Alternative Hypothesis(H_A) : $\mu < value$

Null Hypothesis(H0) :$\mu \leq value$
Alternative Hypothesis(H_A) : $\mu > value$

Null Hypothesis(H0) :$\mu = value$
Alternative Hypothesis(H_A) : $\mu \neq value$



Hypothesis Testing: An Intuitive Explanation | by Renu Khandelwal | Medium

9.4.1 Hypothesis and alternative

Example 9.23. To verify whether the average number of concurrent users increased by 2000, we test

$$H_0 : \mu_2 - \mu_1 = 2000 \quad \text{vs} \quad H_A : \mu_2 - \mu_1 \neq 2000,$$

where μ_1 is the average number of concurrent users last year, and μ_2 is the average number of concurrent users this year.

Depending on the situation, we may replace the two-sided alternative $H_A : \mu_2 - \mu_1 \neq 2000$ with a one-sided alternative $H_A^{(1)} : \mu_2 - \mu_1 < 2000$ or $H_A^{(2)}$ A : $\mu_2 - \mu_1 > 2000$.

9.4.1 Hypothesis and alternative

Example 9.23. To verify whether the average number of concurrent users increased by 2000, we test

$$H_0 : \mu_2 - \mu_1 = 2000 \quad \text{vs} \quad H_A : \mu_2 - \mu_1 \neq 2000,$$

where μ_1 is the average number of concurrent users last year, and μ_2 is the average number of concurrent users this year.

The test of H_0 against $H_A^{(1)}$ evaluates the amount of evidence that the mean number of concurrent users changed by fewer than 2000. Testing against $H_A^{(2)}$, we see if there is sufficient evidence to claim that this number increased by more than 2000.

9.4.1 Hypothesis and alternative

Example 9.24. To verify if the proportion of defective products is at most 3%, we test

$$H_0 : p = 0.03 \quad \text{vs} \quad H_A : p > 0.03,$$

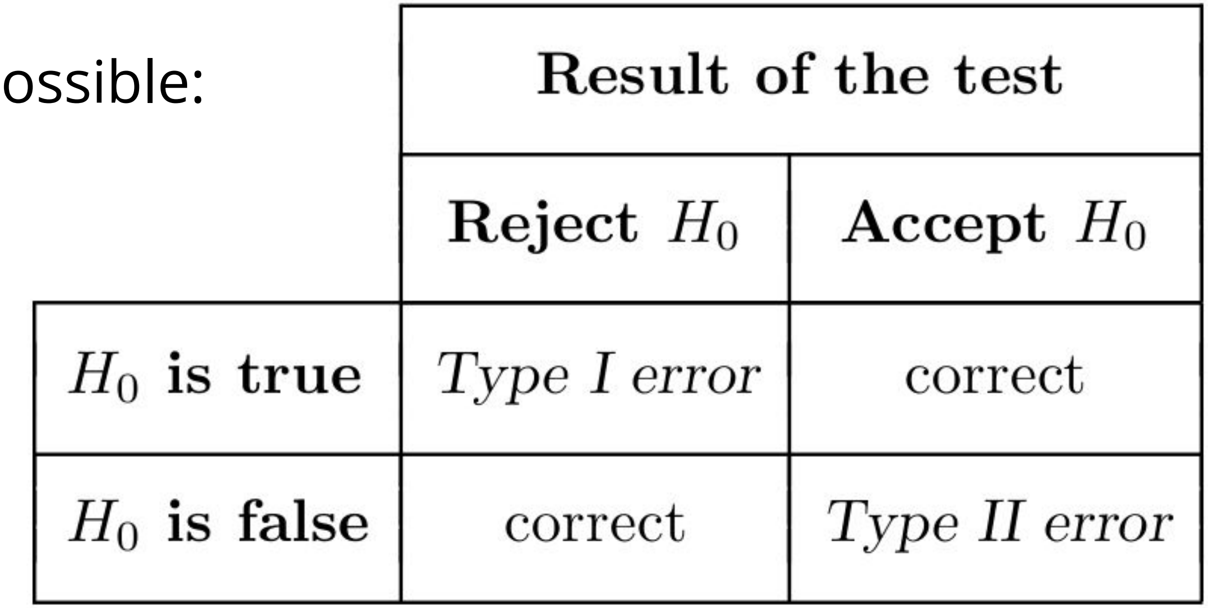
where p is the proportion of defects in the whole shipment.

Why do we choose the right-tail alternative $H_A : p > 0.03$? That is because we reject the shipment only if significant evidence supporting this alternative is collected. If the data suggest that $p < 0.03$, the shipment will still be accepted.

9.4.2 Type I and Type II errors: level of significance

When testing hypotheses, we realize that all we see is a random sample. Therefore, with all the best statistics skills, our decision to accept or to reject H_0 may still be wrong. That would be a **sampling error**.

Four situations are possible:



	Result of the test	
	Reject H_0	**Accept** H_0
H_0 **is true**	*Type I error*	correct
H_0 **is false**	correct	*Type II error*

9.4.2 Type I and Type II errors: level of significance

In two of the four cases, the test results in a correct decision. Either we accepted a true hypothesis, or we rejected a false hypothesis. The other two situations are sampling errors.

DEFINITION 9.7

A **type I error** occurs when we reject the true null hypothesis.

A **type II error** occurs when we accept the false null hypothesis.

Each error occurs with a certain probability that we hope to keep small. A good test results in an erroneous decision only if the observed data are somewhat extreme.

9.4.2 Type I and Type II errors: level of significance

A type I error is often considered more dangerous and undesired than a type II error. Making a type I error can be compared with convicting an innocent defendant or sending a patient to a surgery when (s)he does not need one.

For this reason, we shall design tests that bound the probability of type I error by a preassigned small number α. Under this condition, we may want to minimize the probability of type II error.

9.4.2 Type I and Type II errors: level of significance

DEFINITION 9.8

Probability of a type I error is the **significance level** of a test,

$$\alpha = \mathrm{P}\,\{\text{reject } H_0 \mid H_0 \text{ is true}\}\,.$$

Probability of rejecting a false hypothesis is the **power** of the test,

$$p(\theta) = \mathrm{P}\,\{\text{reject } H_0 \mid \theta;\, H_A \text{ is true}\}\,.$$

It is usually a function of the parameter θ because the alternative hypothesis includes a set of parameter values. Also, the power is the probability to avoid a Type II error.

9.4.2 Type I and Type II errors: level of significance

- Typically, hypotheses are tested at significance levels as small as 0.01, 0.05, or 0.10, although there are exceptions.
- Testing at a low level of significance means that only a large amount of evidence can force rejection of H_0.
- Rejecting a hypothesis at a very low level of significance is done with a lot of confidence that this decision is right.

9.4.3 Level α tests: general approach

A standard algorithm for a level α test of a hypothesis H_0 against an alternative H_A consists of 3 steps.

Step 1. Test statistic

Testing hypothesis is based on a **test statistic** T, a quantity computed from the data that has some known, tabulated distribution F_0 if the hypothesis H_0 is true.

Test statistics are used to discriminate between the hypothesis and the alternative. When we verify a hypothesis about some parameter θ, the test statistic is usually obtained by a suitable transformation of its estimator $\hat{\theta}$.

9.4.3 Level α tests: general approach

Step 2. Acceptance region and rejection region

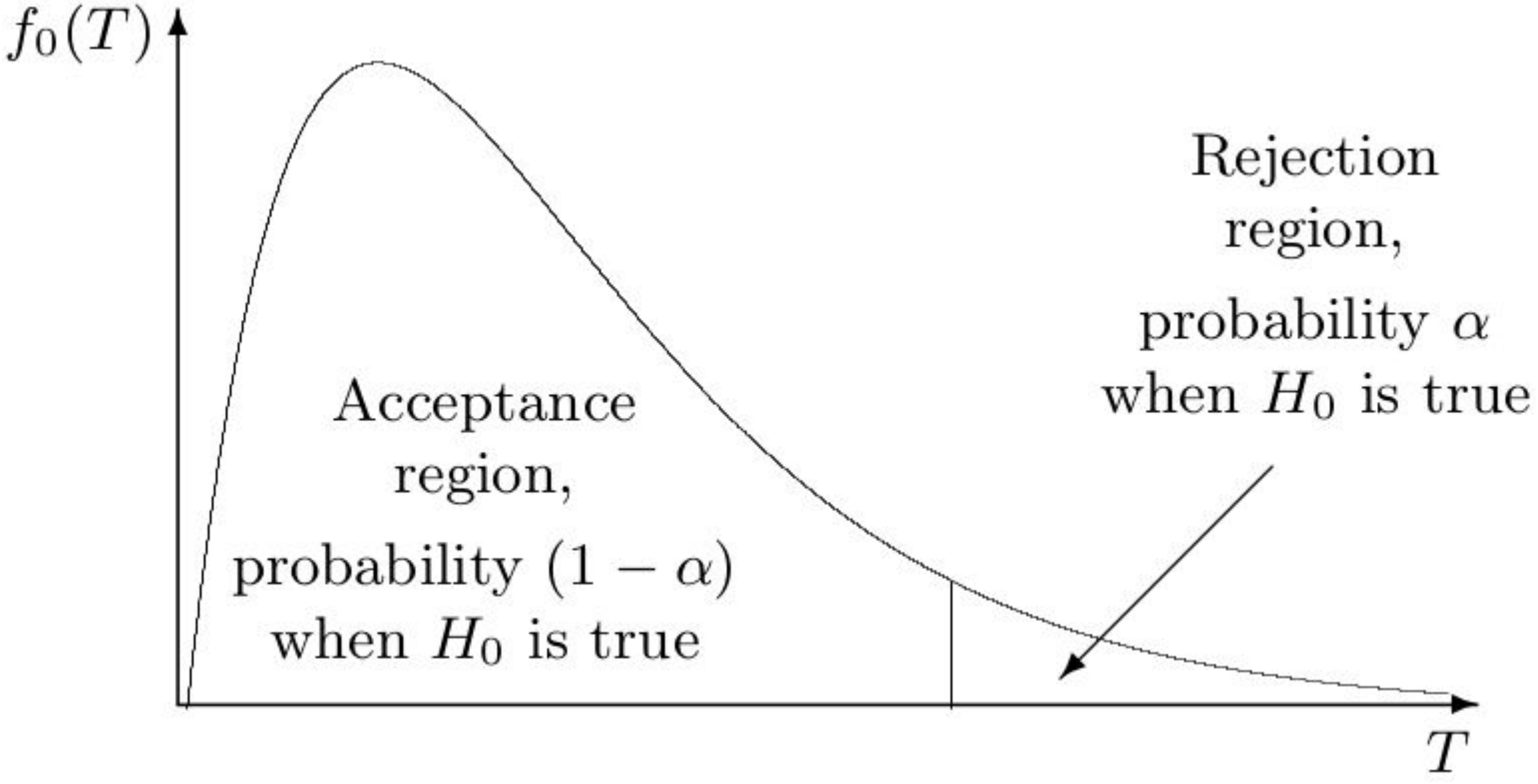
Next, we consider the **null distribution** F_0. This is the distribution of test statistic T when the hypothesis H_0 is true.

If it has a density f_0, then the whole area under the density curve is 1, and we can always find a portion of it whose area is α, as shown in the next figure. It is called **rejection region** ($\mathcal{R}$).

The remaining part, the complement of the rejection region, is called **acceptance region** ($\mathcal{A}=\mathcal{R}$). By the complement rule, its area is $(1 - \alpha)$.

9.4.3 Level α tests: general approach

Step 2. Acceptance region and rejection region


$f_0(T)$
Rejection region, probability α when H_0 is true
Acceptance region, probability $(1-\alpha)$ when H_0 is true
T

9.4.3 Level α tests: general approach

Step 2. Acceptance region and rejection region

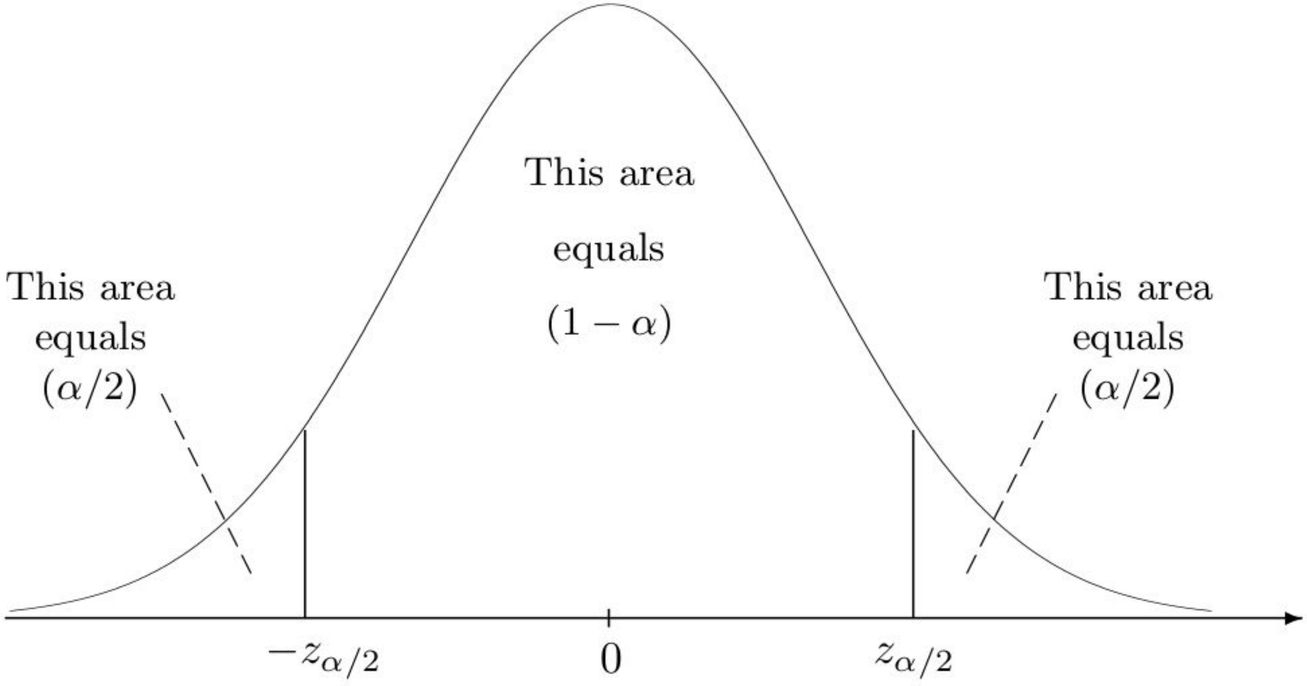
These regions are selected in such a way that the values of test statistic T in the rejection region provide a stronger support of H_A than the values $T \in \mathcal{A}$.

For example, suppose that T is expected to be large if H_A is true. Then the rejection region corresponds to the right tail of the null distribution F_0 shown in the Figure.

9.4.3 Level α tests: general approach

Step 2. Acceptance region and rejection region

As another example, look at Figure 9.3 on p. 249:



9.4.3 Level α tests: general approach

Step 2. Acceptance region and rejection region

If the null distribution of T is Standard Normal, then the area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ equals exactly $(1 - \alpha)$. The interval

$$\mathcal{A} = (-z_{\alpha/2}, z_{\alpha/2})$$

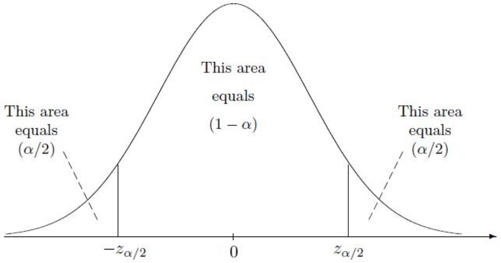
can serve as a level α acceptance region for a two-sided test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$.

The remaining part consists of two symmetric tails,

$$\mathcal{R} = \bar{\mathcal{A}} = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, +\infty);$$

this is the rejection region.

9.4.3 Level α tests: general approach



Step 2. Acceptance region and rejection region

Areas under the density curve are probabilities, and we conclude that

$$P\{T \in \text{acceptance region} \mid H_0\} = 1 - \alpha$$

and

$$P\{T \in \text{rejection region} \mid H_0\} = \alpha.$$

9.4.3 Level α tests: general approach

Step 3: Result and its interpretation

- Accept the hypothesis H_0 if the test statistic T belongs to the acceptance region.
- Reject H_0 in favor of the alternative H_A if T belongs to the rejection region.

Our acceptance and rejection regions guarantee that the significance level of our test is

$$\begin{aligned} \text{Significance level} &= \boldsymbol{P}\{\text{ Type I error }\} &&= \boldsymbol{P}\{\text{ Reject } \mid H_0\} \\ &= \boldsymbol{P}\{T \in \mathcal{R} \mid H_0\} &&= \alpha. \end{aligned}$$

Therefore, indeed, we have a level α test!

9.4.3 Level α tests: general approach

Step 3: Result and its interpretation

The interesting part is to interpret our result correctly.

- Notice that conclusions like “My level *α* test accepted the hypothesis. Therefore, the hypothesis is true with probability (1−*α*)” are wrong!
- Statements H_0 and H_A are about a non-random population, and thus, the hypothesis can either be true with probability 1 or false with probability 1.

If the test rejects the hypothesis, all we can state is that the data provides sufficient evidence against H_0 and in favor of H_A. It may either happen because H_0 is not true, or because our sample is too extreme. The latter, however, can only happen with

9.4.3 Level α tests: general approach

Step 3: Result and its interpretation

The interesting part is to interpret our result correctly.

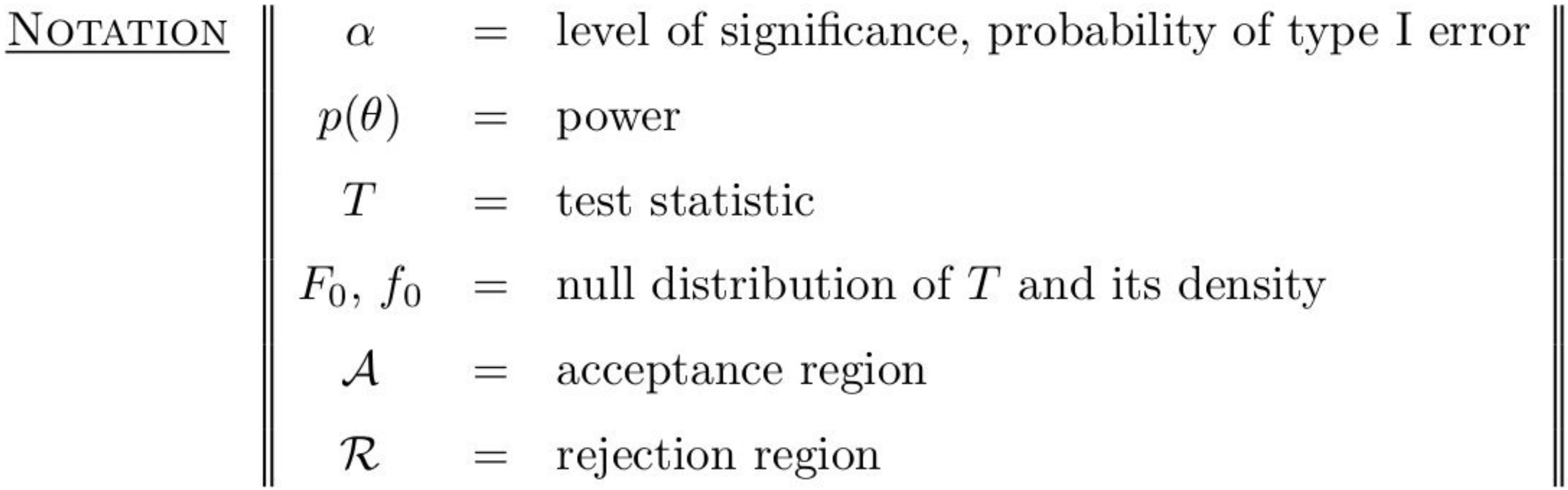
- If the test rejects the hypothesis, all we can state is that the data provides sufficient evidence against H_0 and in favor of H_A.
- It may either happen because H_0 is not true, or because our sample is too extreme. The latter, however, can only happen with probability α.

9.4.3 Level α tests: general approach

Step 3: Result and its interpretation

- If the test accepts the hypothesis, it only means that the evidence obtained from the data is not sufficient to reject it.
- In the absence of sufficient evidence, by default, we accept the null hypothesis.

9.4.3 Level α tests: general approach



NOTATION

α	=	level of significance, probability of type I error
$p(\theta)$	=	power
T	=	test statistic
$F_0,\ f_0$	=	null distribution of T and its density
$\mathcal{A}$	=	acceptance region
$\mathcal{R}$	=	rejection region

9.4.4 Rejection regions and power

Our construction of the rejection region guaranteed the desired significance level α, as we have previously proved.

However, one can choose many regions that will also have probability α. Among them, **which one is the best choice**?

To avoid type II errors, we choose such a rejection region that will likely cover the test statistic T in case if the alternative H_A is true. This maximizes the power of our test because we'll rarely accept H_0 in this case.

9.4.4 Rejection regions and power

Then, we look at our test statistic T under the alternative. Often

a. a right-tail alternative forces T to be large,
b. a left-tail alternative forces T to be small,
c. a two-sided alternative forces T to be either large or small

(although it certainly depends on how we choose T).

9.4.4 Rejection regions and power

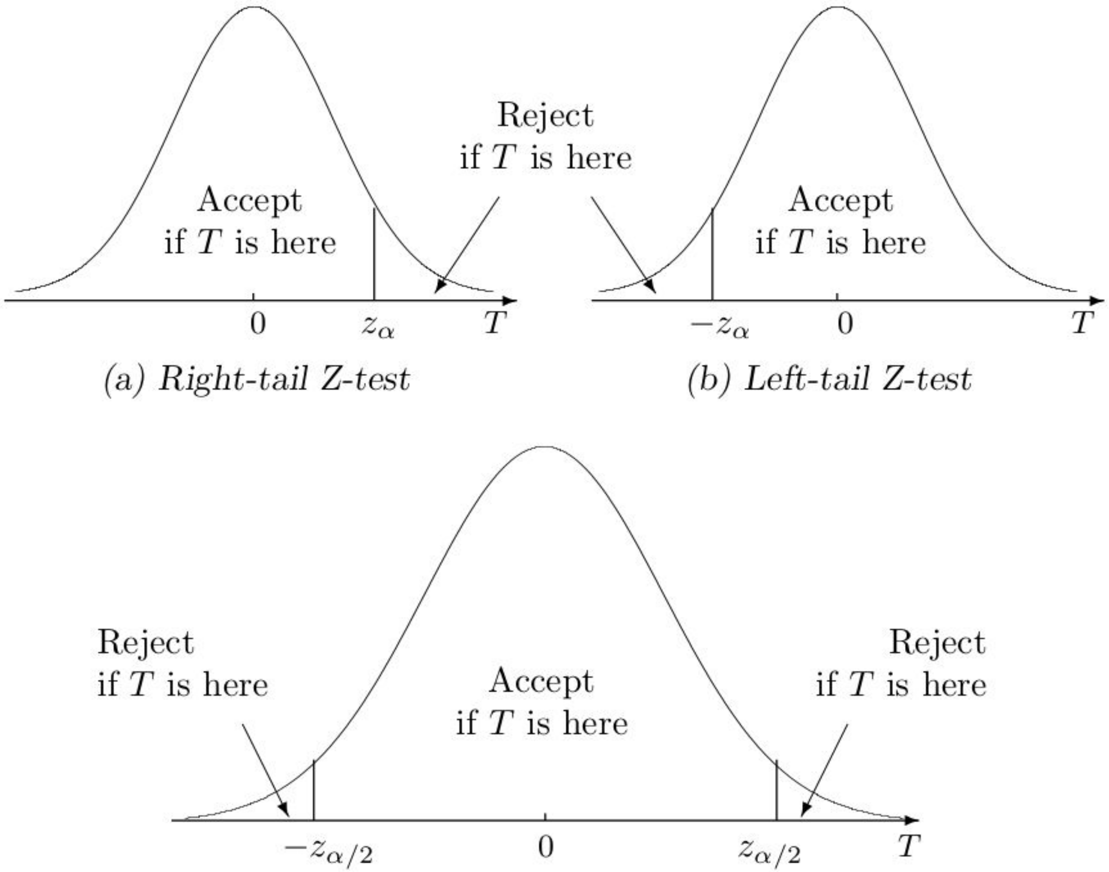
If this is the case, it tells us exactly when we should reject the null hypothesis:

a. For a right-tail alternative, the rejection region $\mathcal{R}$ should consist of large values of T . Choose R on the right, $\mathcal{A}$ on the left.

b. For a left-tail alternative, the rejection region R should consist of small values of T . Choose $\mathcal{R}$ on the left, $\mathcal{A}$ on the right.

c. For a two-sided alternative, the rejection region $\mathcal{R}$ should consist of very small and very large values of T. Let $\mathcal{R}$ consist of two extreme regions, while $\mathcal{A}$ covers the middle.

9.4.4 Rejection regions and power



(a) *Right-tail Z-test*

(b) *Left-tail Z-test*

(c) *Two-sided Z-test*

9.4.5 Standard Normal null distribution (Z-test)

An important case, in terms of a large number of applications, is when the null distribution of the test statistic is **Standard Normal**.

The test in this case is called a **Z-test**, and the test statistic is usually denoted by Z.

9.4.5 Standard Normal null distribution (Z-test)

A level α test with a right-tail alternative should

$$\begin{cases} \text{reject } H_0 \text{ if } \mathcal{Z} \geq \boldsymbol{z}_\alpha \\ \text{accept } H_0 \text{ if } \mathcal{Z} < \boldsymbol{z}_\alpha \end{cases}$$

The rejection region in this case consists of large values of $\mathcal{Z}$ only, $\mathcal{R} = [\boldsymbol{z}_\alpha , +\infty)$, $\mathcal{A} = (-\infty, \boldsymbol{z}_\alpha)$.

Under the null hypothesis, Z belongs to $\mathcal{A}$ and we reject the null hypothesis with probability

$$P \{ T \geq \boldsymbol{z}_\alpha \mid H_0 \} = 1 - \Phi(\boldsymbol{z}_\alpha) = \alpha,$$

making the probability of false rejection (type I error) equal α.

9.4.5 Standard Normal null distribution (Z-test)

A level α test with a left-tail alternative should

$$\begin{cases} \text{reject } H_0 \text{ if } \mathcal{Z} \leq \boldsymbol{z}_\alpha \\ \text{accept } H_0 \text{ if } \mathcal{Z} > \boldsymbol{z}_\alpha \end{cases}$$

The rejection region in this case consists of small values of $\mathcal{Z}$ only, $\mathcal{R} = (-\infty, -\boldsymbol{z}_\alpha)$, $\mathcal{A} = [-\boldsymbol{z}_\alpha, +\infty)$.

Under the null hypothesis, Z belongs to $\mathcal{A}$ and we reject the null hypothesis with probability

$$P\{ T \leq -\boldsymbol{z}_\alpha \mid H_0 \} = \Phi(-\boldsymbol{z}_\alpha) = \alpha,$$

making the probability of false rejection (type I error) equal α.

9.4.5 Standard Normal null distribution (Z-test)

A level α test with a two-sided alternative should

$$\begin{cases} \text{reject } H_0 \text{ if } |\mathcal{Z}| \leq \boldsymbol{z}_{\alpha/2} \\ \text{accept } H_0 \text{ if } |\mathcal{Z}| > \boldsymbol{z}_{\alpha/2} \end{cases}$$

The rejection region in this case consists of very small and very large values of $\mathcal{Z}$, $\mathcal{R} = (-\infty, -\boldsymbol{z}_{\alpha/2}] \cup [\boldsymbol{z}_{\alpha/2}, +\infty)$, $\mathcal{A} = (-\boldsymbol{z}_{\alpha/2}\ , \boldsymbol{z}_{\alpha/2})$.

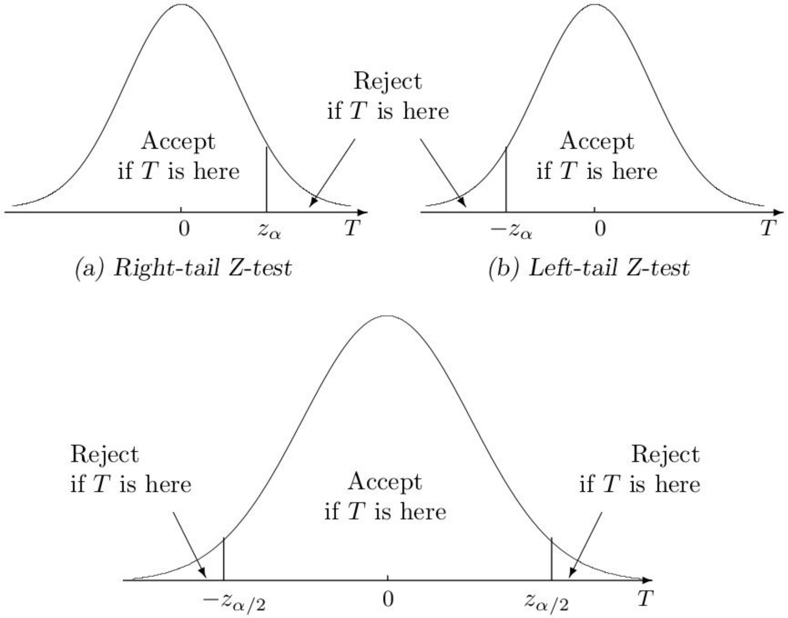
Under the null hypothesis, Z belongs to $\mathcal{A}$ and we reject the null hypothesis with probability

$$P\ \{\ T \leq -\boldsymbol{z}_{\alpha/2} \cup T \geq \boldsymbol{z}_{\alpha/2} \mid H_0\ \} = \Phi(-\boldsymbol{z}_{\alpha/2}) + 1 - \Phi(\boldsymbol{z}_{\alpha/2}) = \alpha,$$

making the probability of false rejection (type I error) equal α.

9.4.5 Standard Normal null distribution (Z-test)

- For a two-sided test, divide α by two and use $\boldsymbol{z}_{\alpha/2}$;
- for a one-sided test, use $\boldsymbol{z}_{\alpha}$ keeping in mind that the rejection region consists of just one piece.



(a) *Right-tail Z-test*

(b) *Left-tail Z-test*

(c) *Two-sided Z-test*

9.4.5 Standard Normal null distribution (Z-test)

Now consider testing a hypothesis about a population parameter θ. Suppose that its estimator $\hat{\theta}$ has Normal distribution, at least approximately, and we know $E(\hat{\theta})$ and $Var(\hat{\theta})$ if the hypothesis is true.

Then the test statistic

$$\mathcal{Z} = \hat{\theta} - \mathbf{E}(\hat{\theta}) / \mathbf{Std}(\hat{\theta})$$

has Standard Normal distribution, and we can use what we have learned so far to construct acceptance and rejection regions for a level α test. We call $\mathcal{Z}$ a **Z-statistic**.

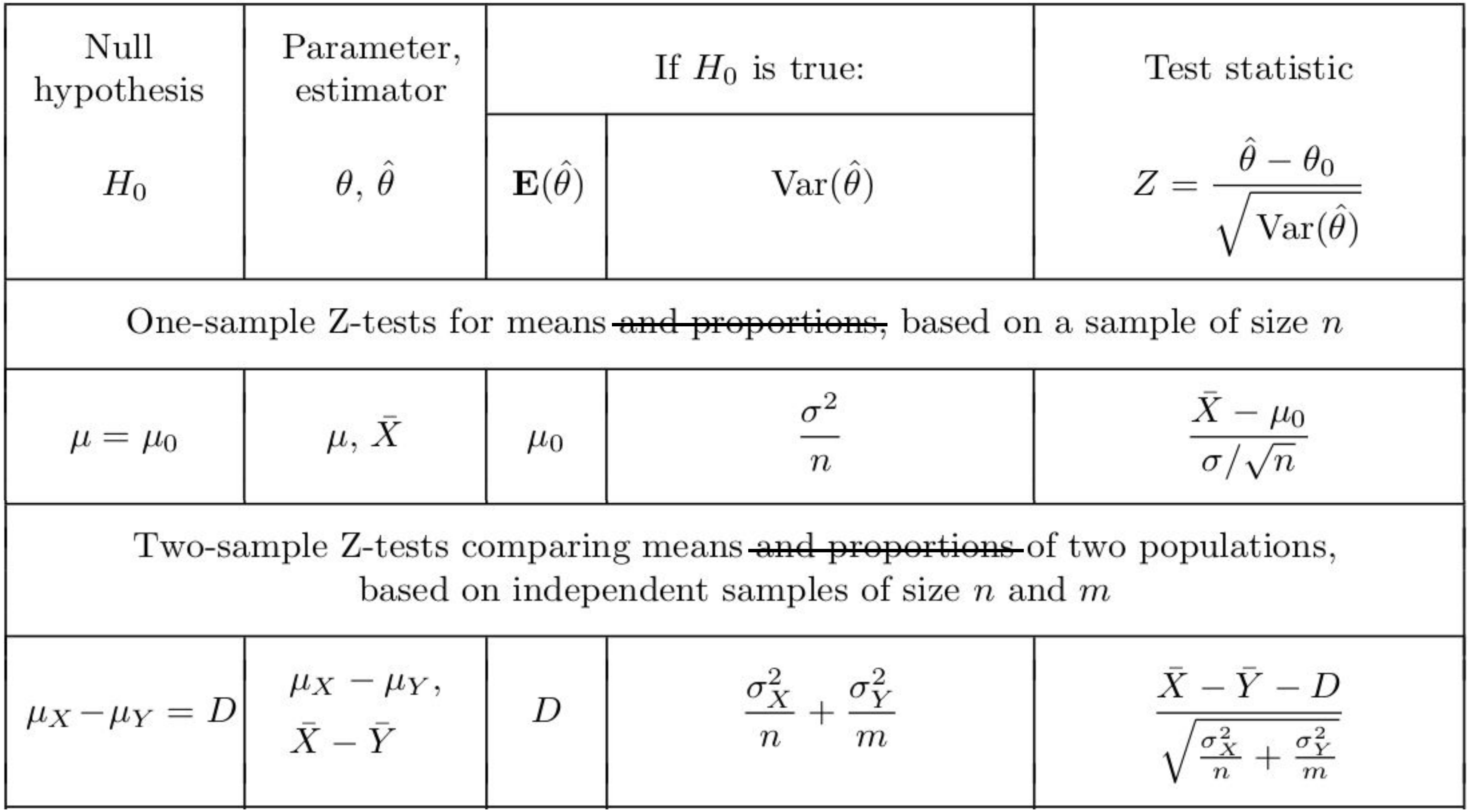
9.4.6 Z-tests for means ~~(and proportions)~~

As we already know,

- sample means have Normal distribution when the distribution of data is Normal;
- sample means have approximately Normal distribution when they are computed from large samples (the distribution of data can be arbitrary).

For these cases, we can use a Z-statistic and rejection regions to design powerful level α tests.

9.4.6 Z-tests for means ~~(and proportions)~~



Null hypothesis	Parameter, estimator	If H_0 is true:		Test statistic
H_0	θ, $\hat{\theta}$	$\mathbf{E}(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$Z = \dfrac{\hat{\theta} - \theta_0}{\sqrt{\mathrm{Var}(\hat{\theta})}}$
One-sample Z-tests for means ~~and proportions,~~ based on a sample of size n				
$\mu = \mu_0$	μ, $\bar{X}$	μ_0	$\dfrac{\sigma^2}{n}$	$\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$
Two-sample Z-tests comparing means ~~and proportions~~ of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y = D$	$\mu_X - \mu_Y$, $\bar{X} - \bar{Y}$	D	$\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$	$\dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$

9.4.6 Z-tests for means ~~(and proportions)~~

Example 9.25 (Z-test about a population mean). The number of concurrent users for some internet service provider has always averaged 5000 with a standard deviation of 800. After an equipment upgrade, the average number of users at 100 randomly selected moments of time is 5200. Does it indicate, at a 5% level of significance, that the mean number of concurrent users has increased? Assume that the standard deviation of the number of concurrent users has not changed.

9.4.6 Z-tests for means ~~(and proportions)~~

Example 9.25 (Z-test about a population mean).

$\mu_i = 5000$ $\sigma_i = \sigma_f = 800$ $\bar{X} = 5200$ $n = 100$

$H_0 : \mu_f = 5000$ $H_A : \mu_f > 5000$ $\alpha = 0.05$

Test statistic:

$Z = (\bar{X} - \mu_i) / (\sigma_f/\sqrt{n}) = (5200\text{-}5000) / (800/\sqrt{100}) = 2.5$

$\boldsymbol{z}_\alpha = \boldsymbol{z}_{0.05} = 1.645$

reject H_0 if $\mathcal{Z} \geq 1.645$
accept H_0 if $\mathcal{Z} < 1.645$

9.4.6 Z-tests for means ~~(and proportions)~~

Example 9.25 (Z-test about a population mean).

$\mu_i = 5000$ $\sigma_i = \sigma_f = 800$ $\bar{X} = 5200$ $n = 100$

$H_0 : \mu_f = 5000$ $H_A : \mu_f > 5000$ $\alpha = 0.05$

Test statistic:

$Z = (\bar{X} - \mu_i) / (\sigma_f/\sqrt{n}) = (5200\text{-}5000) / (800/\sqrt{100}) = 2.5$

$\boldsymbol{z}_\alpha = \boldsymbol{z}_{0.05} = 1.645$

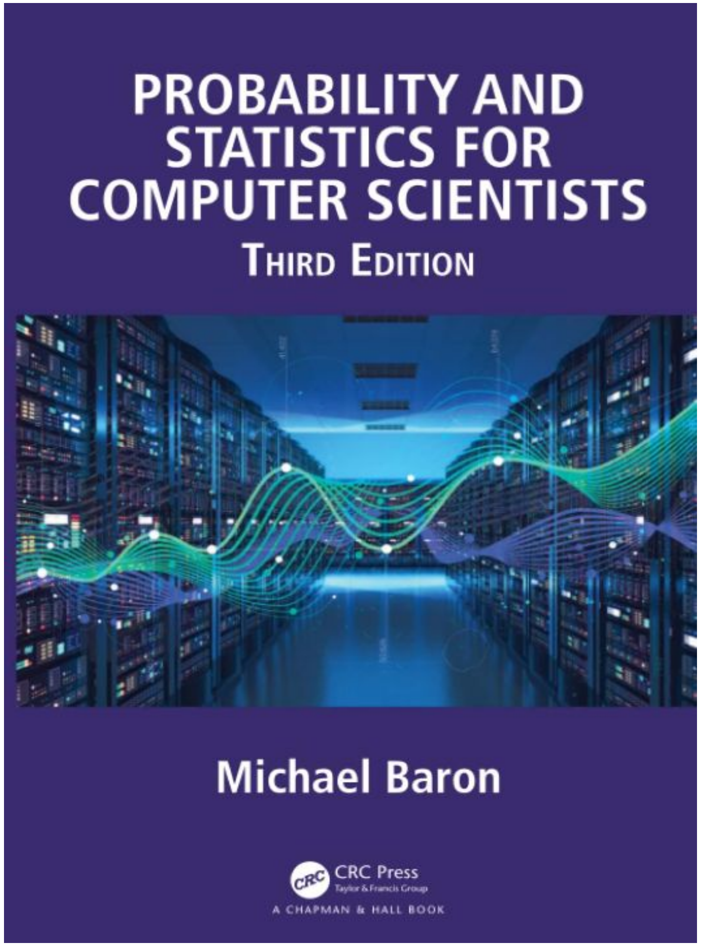
$$\begin{cases} \text{reject } H_0 \text{ if } \mathcal{Z} \geq 1.645 \\ \text{accept } H_0 \text{ if } \mathcal{Z} < 1.645 \end{cases}$$

The data provided sufficient evidence in favor of the alternative hypothesis that the mean number of users has increased.

References

WIKIPEDIA

Google images



PROBABILITY AND STATISTICS FOR COMPUTER SCIENTISTS

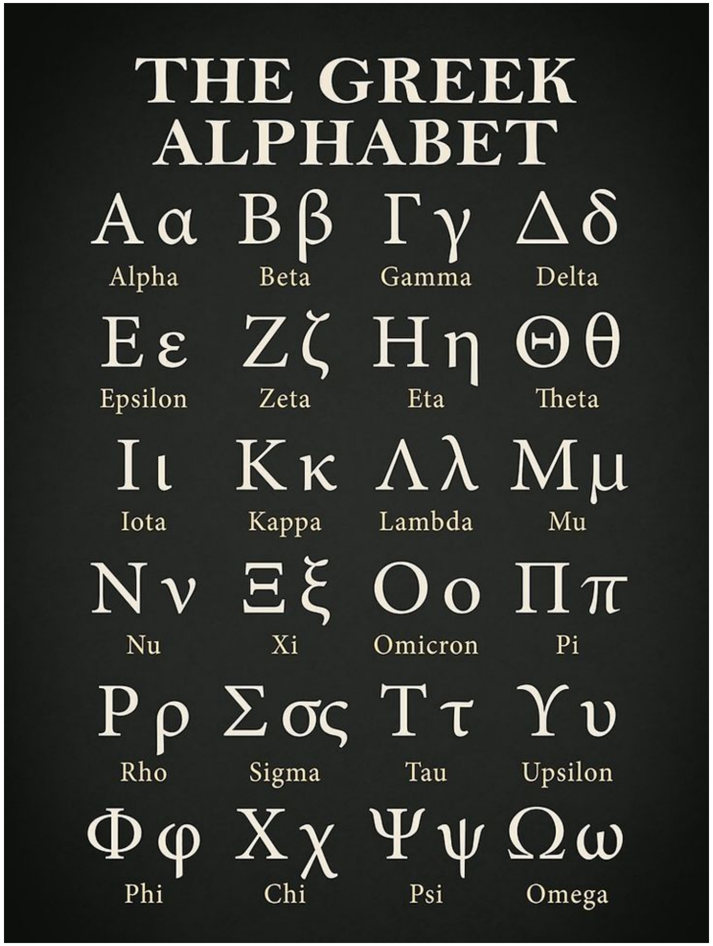
THIRD EDITION

Michael Baron

CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Appendix



THE GREEK ALPHABET

Α α	Β β	Γ γ	Δ δ
Alpha	Beta	Gamma	Delta
Ε ε	Ζ ζ	Η η	Θ θ
Epsilon	Zeta	Eta	Theta
Ι ι	Κ κ	Λ λ	Μ μ
Iota	Kappa	Lambda	Mu
Ν ν	Ξ ξ	Ο ο	Π π
Nu	Xi	Omicron	Pi
Ρ ρ	Σ σ ς	Τ τ	Υ υ
Rho	Sigma	Tau	Upsilon
Φ φ	Χ χ	Ψ ψ	Ω ω
Phi	Chi	Psi	Omega

Normal distribution

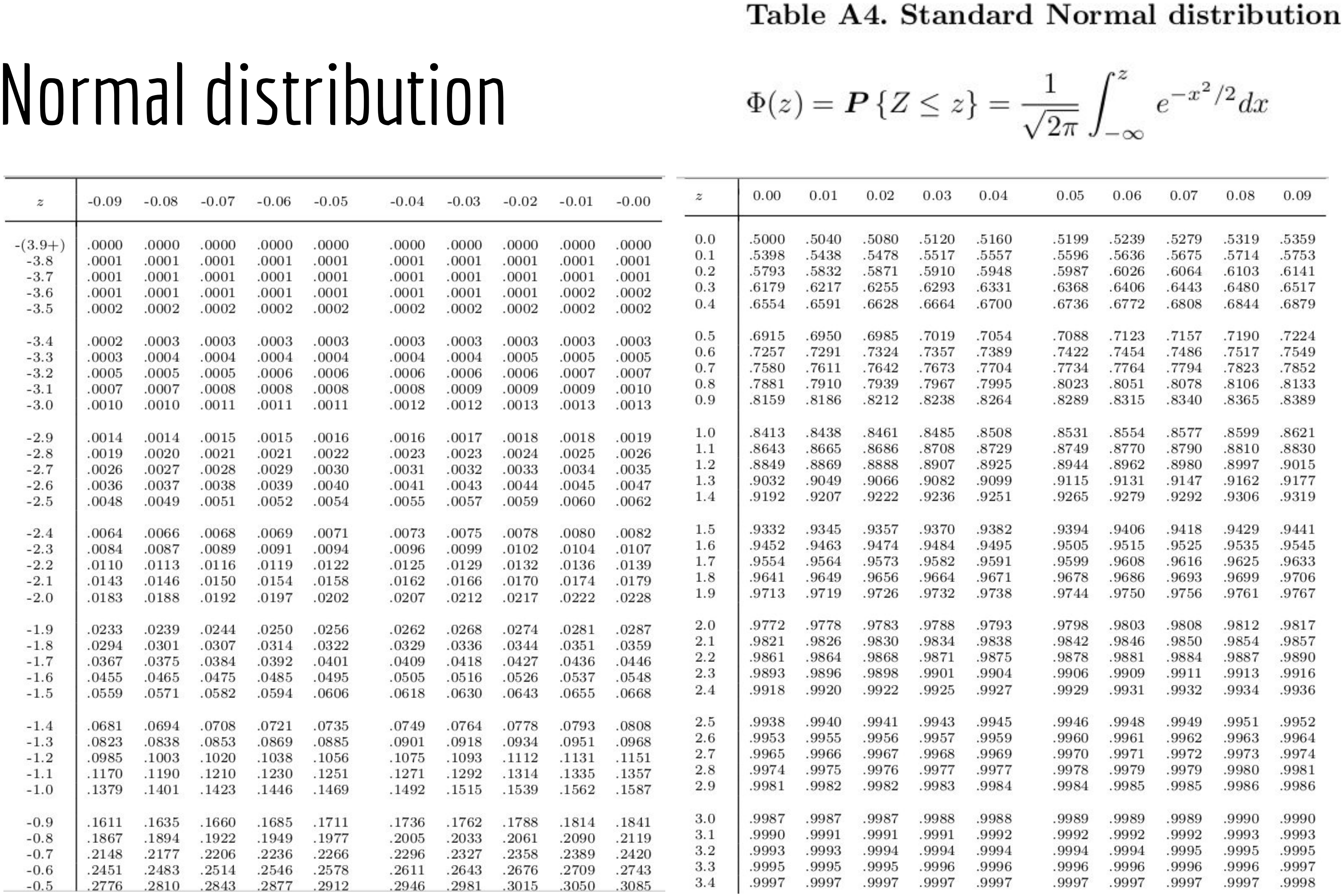


Table A4. Standard Normal distribution

$$\Phi(z) = \boldsymbol{P}\{Z \le z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00
-(3.9+)	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.6	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
-3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007
-3.1	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
-3.0	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
-2.9	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019
-2.8	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026
-2.7	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035
-2.6	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047
-2.5	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062
-2.4	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082
-2.3	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107
-2.2	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139
-2.1	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179
-2.0	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228
-1.9	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287
-1.8	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359
-1.7	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446
-1.6	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548
-1.5	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668
-1.4	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808
-1.3	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968
-1.2	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151
-1.1	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357
-1.0	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587
-0.9	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841
-0.8	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119
-0.7	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420
-0.6	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743
-0.5	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998