Statistical inference

Prof. F. Assis

Statistical inference

- Random Sampling: sampling Distribution of the Mean
- Statistical inference: frequentist inference, Bayesian inference
- Estimation: parameter estimation, random variable estimatiom.
- Point Estimators: bias, the mean squared error (MSE),
- Maximum Likelihood Estimation (MLE): the *log-likelihood function*. Maximum Likelihood Estimation for Gaussian distribution.

Statistical inference

- The Maximum Likelihood Estimation for Gaussian distribution (for independente random variables).
- Discriminant functions: quadratic discriminant, linear discriminant.
- Confidence interval (frequentist inference): interval estimation, confidence level, confidence limits.
- Confidence Intervals for the sample mean.
- Choice of Sample Size for the sample mean (frequentist inference).

Statistical inference

- Statistical Hypothesis: null hypothesis, alternative hypothesis.
- Statistical Hypothesis: Hypothesis-testing methodology.
- Statistical Hypothesis: Type I error, Type II error.

A statistic definition

- A statistic is any quantity computed from values in a sample obtained from a continuous or discrete process which is considered for a statistical purpose.

- Statistical purposes include estimating a population parameter (mean, variance, any order moment etc.), describing a sample (process modeling), or evaluating a hypothesis.

Random Sampling

Random Sampling

- The random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, are a **random sample** of size n if (a) the X_i's are independent random variables and (b) every X_i has *the same probability distribution* (???).

  - Each numerical value in the data is the observed value of a random variable.
  - Furthermore, the random variables are usually assumed to be **independent and identically distributed** (**i.i.d.**).

- These random variables are known as a *random sample*.

Random Sampling

- The observed data are also referred to as a *random sample*, but the use of the same phrase should not cause any confusion.
- The assumption of a random sample is extremely important. If the sample is not random and is based on judgment or is flawed in some other way, statistical methods will not work properly and will lead to incorrect decisions.
- The primary purpose in taking a random sample is to obtain information about the unknown population parameters.

Random Sampling

- A **statistic** is any function of the observations in a random sample.
  - For example, if $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, is a random sample of size n, the **sample mean** X, the **sample variance** S^2 (a variance estimation), and the **sample standard deviation** S (a standard deviation estimation) are statistics.
  - Because a statistic is a random variable, it has a probability distribution.
- The probability distribution of a statistic is called a **sampling distribution**.

Random Sample

The collection of random variables $X_1, X_2, X_3, \ldots, X_n$ is said to be a **random sample** of size n if they are independent and identically distributed (i.i.d.), i.e.,

1. $X_1, X_2, X_3, \ldots, X_n$ are independent random variables, and
2. they have the same distribution, i.e,

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x), \qquad \text{for all } x \in \mathbb{R}.$$

Random Sample

- Assume that $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_n$ are a random sample.

1. the X_i's are independent;
2. $F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F_X(x)$;
3. $EX_i = EX = \mu < \infty$;
4. $0 < \mathrm{Var}(X_i) = \mathrm{Var}(X) = \sigma^2 < \infty$.

The sample mean: Properties

The sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

1. $E\overline{X} = \mu$.
2. $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.
3. Weak Law of Large Numbers (WLLN):

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \epsilon) = 0.$$

4. Central Limit Theorem: The random variable

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \qquad \text{for all } x \in \mathbb{R}$$

where $\Phi(x)$ is the standard normal CDF.

Example: Sampling Distribution of the Mean

- The probability distribution of $\bar{X}$ is called the **sampling distribution of the mean**.
- The sampling distribution of a statistic depends on:
  - the distribution of the population,
  - the size of the sample, and
  - the method of sample selection.

Example: Sampling Distribution of the Mean

- Consider determining the sampling distribution which we will call it as the sample mean $\bar{X}$.
- Suppose that a random sample of size n is taken from a normal population with mean μ and variance σ^2.

Example: Sampling Distribution of the Mean

- Then because linear functions of independent, normally distributed random variables are also normally distributed, we conclude that the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

(observe that the random variables are independent and identically distributed (i.i.d.))

- has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

- and variance

$$\sigma^2_{\bar{X}} = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Example: Sampling Distribution of the Mean

- If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately Gaussian distribution with mean μ and variance σ^2/n if the sample size n is large.
- This is one of the most useful theorems in statistics, called the **Central Limit Theorem** (as we saw in the module 5).

We repeat: The Central Limit Theorem (CLT)

If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2 and if $\bar{X}$ is **the sample mean**, the limiting form of the distribution of

$$Z_n = \frac{\bar{X}_n - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution with the zero-mean, unit variance.

Statistical inference

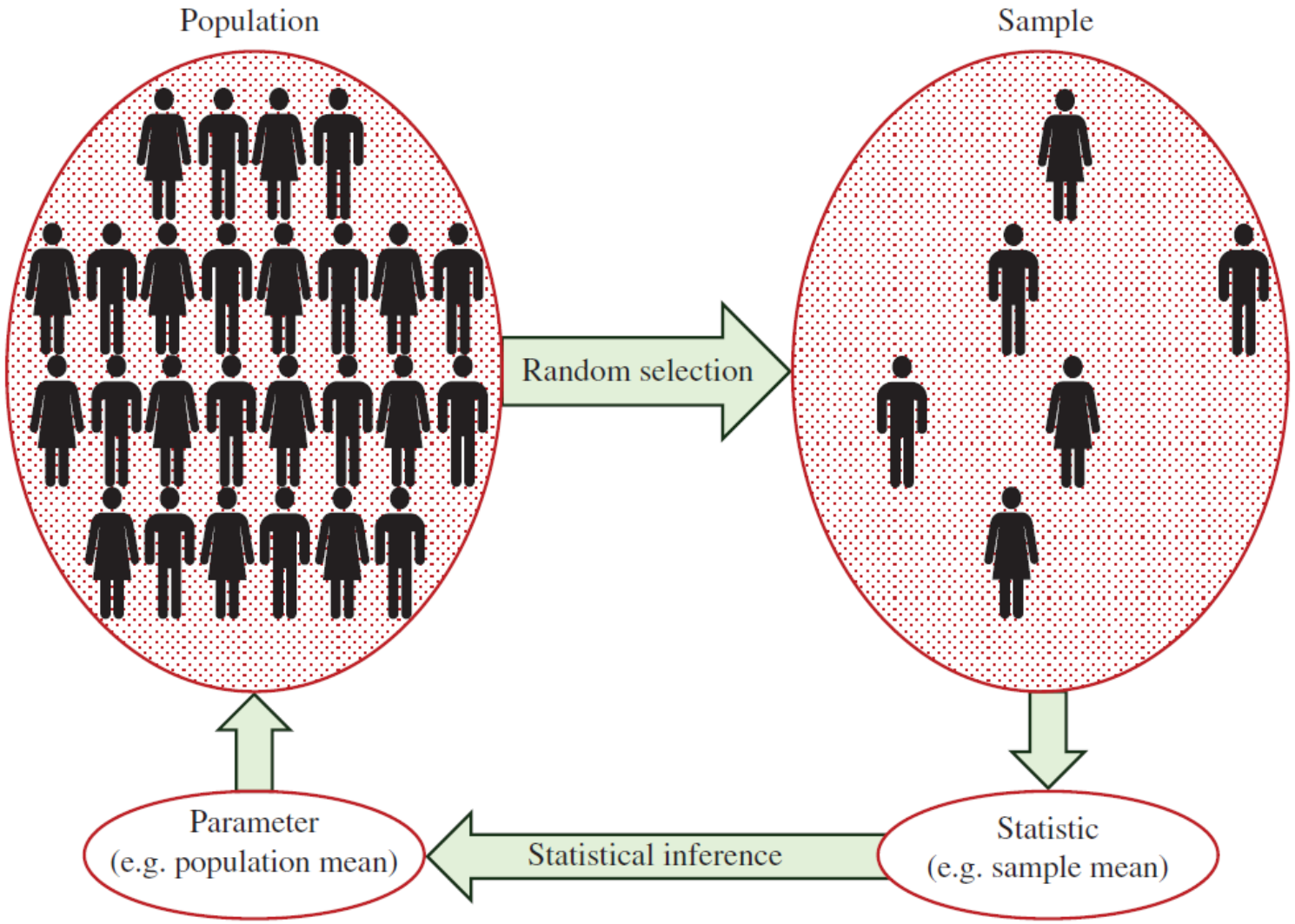
GPDS
Grupo de Processamento Digital de Sinais

Prof. F. Assis

- **Statistical inference** is a collection of methods that deal with drawing conclusions from data that are prone to random variation.

Random Sampling and Statistical inference

- **Statistical inference** is concerned with making **decisions** about a population based on the information contained in **a random sample** from that population.



Statistical inference

- Frequentist Inference
- Bayesian Inference

Frequentist Inference

- In this approach, the unknown quantity θ is assumed to be a fixed quantity.
- That is, θ is a **deterministic** (non-random) quantity that is to be estimated by the observed data.
- For example, in the polling problem stated above we might consider θ as the percentage of people who will vote for a certain candidate, call him/her Candidate A.

Frequentist Inference

- After asking n randomly chosen voters, we might estimate θ by

$$\widehat{\Theta} = \frac{\boldsymbol{A}}{n} \; (relative\ frequency)$$

- Consider $\boldsymbol{A}$ as the number of people (among the randomly chosen voters) who say they will vote for Candidate A.

- Although θ is assumed to be a non-random quantity, our estimate of θ, which we show by $\widehat{\Theta}$ is a random variable, because it depends on our random sample.

Bayesian Inference

- In the Bayesian approach the unknown quantity Θ is assumed to be a random variable, and we assume that we have some initial guess about the distribution of Θ.
- After observing the data, we update the distribution of Θ using **Bayes' Rule**.

Bayesian Inference

the unknown quantity

initial guess

prior

likelihood

posterior

$$P(C|x) = \frac{P(C)P(\boldsymbol{x}|C)}{P(x)}$$

evidence

Estimation

Prof. F. Assis

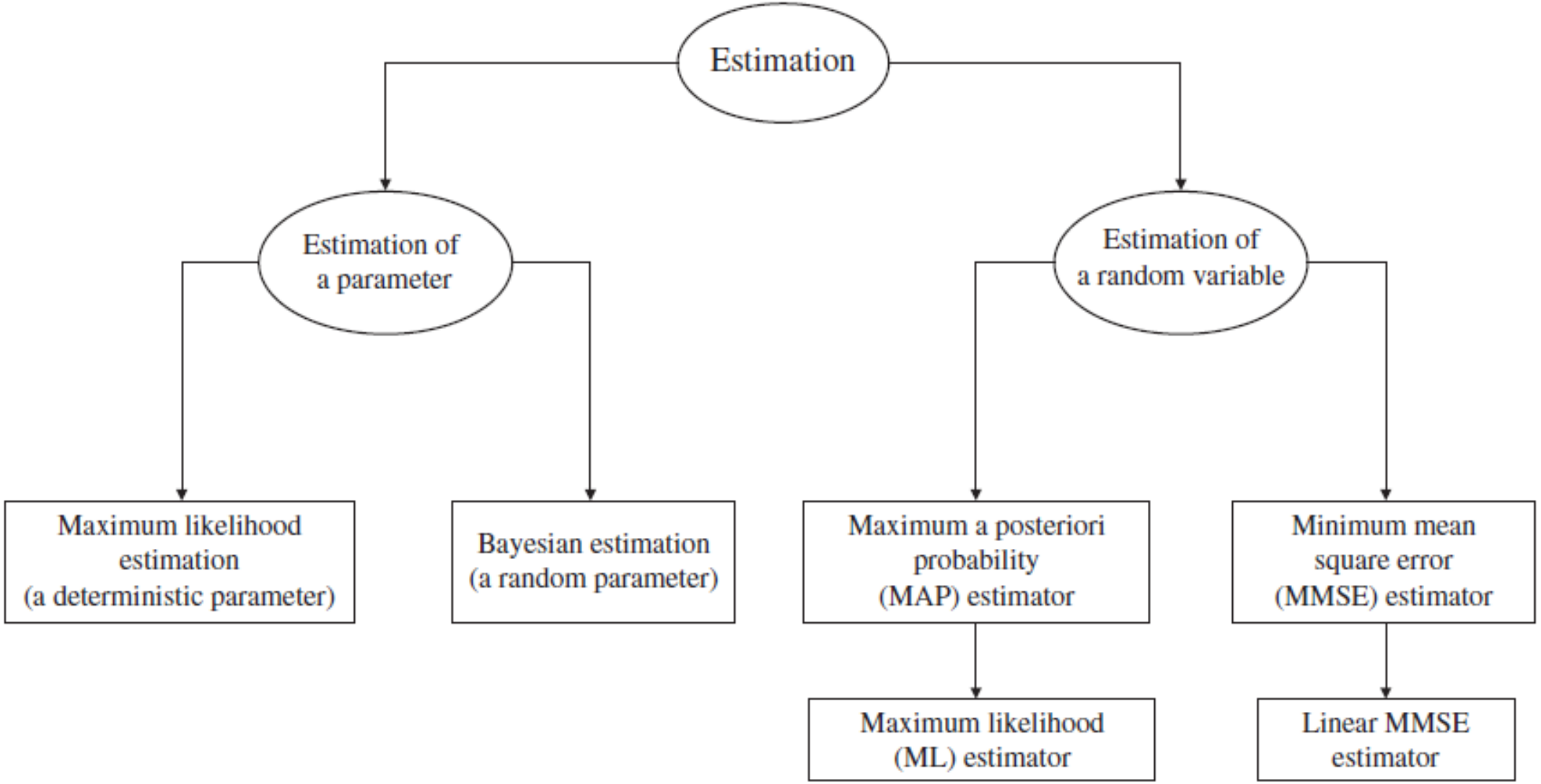
Estimation

- Parameter estimation is concerned with the **estimation of a parameter of a random variable** (example: mean and variance).

Estimation

- Estimation of a random variable deals with the value of an **inaccessible random variable** in terms of the **observation of an accessible random variable**.

Estimation problem



Parameter Estimation

- An estimate that specifies a single value is called a ***point estimate*.**

- An estimate that specifies a range of values is called an ***interval estimate*.**

Parameter Estimation

- A statistic that is a function of the random vector

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$$

and it is used to estimate an unknown parameter θ is called an ***estimator*** $\widehat{\Theta}$.

Parameter Estimation

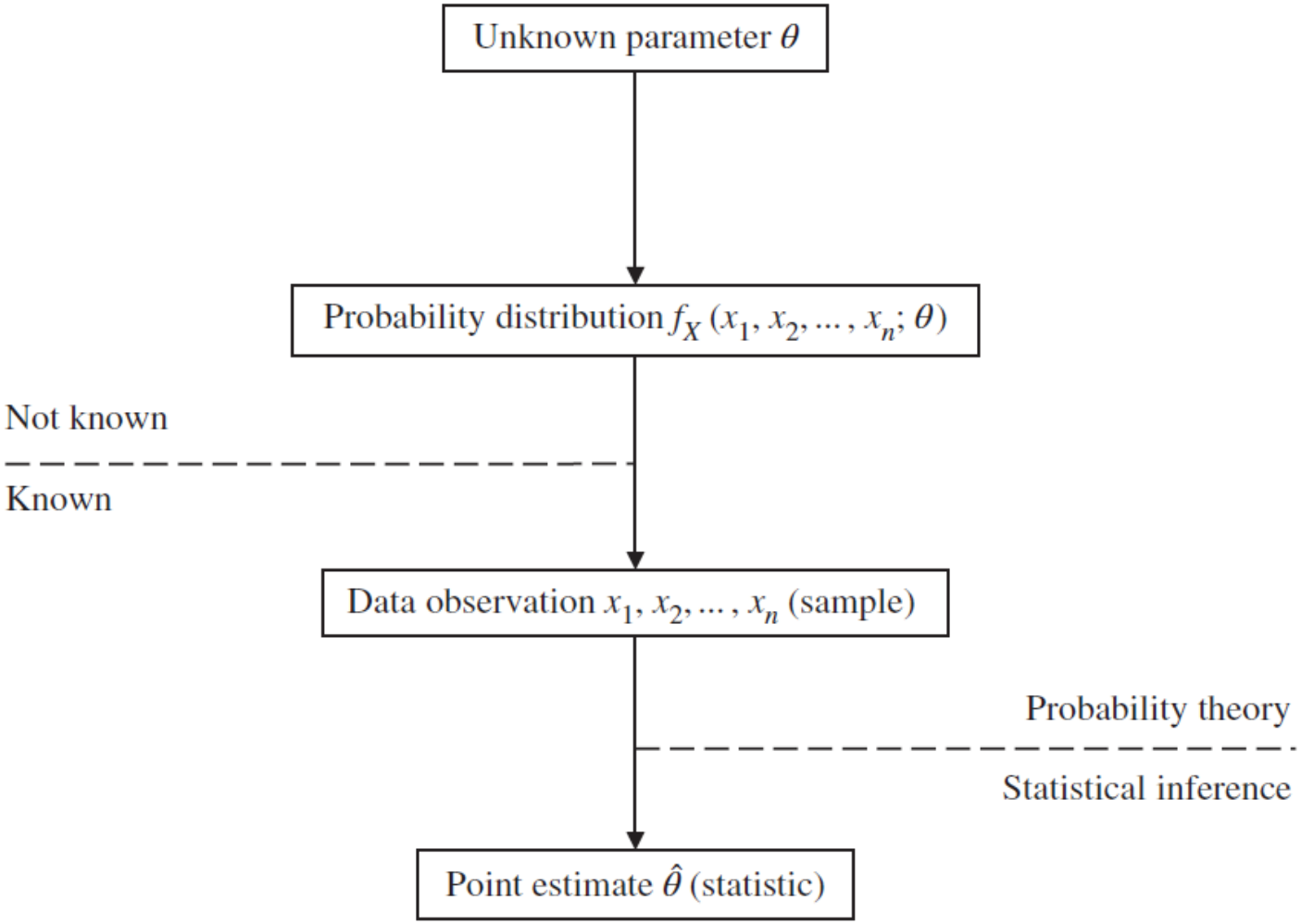
- The estimator $\widehat{\Theta}$ itself is a random variable because it depends on the random sample.

- After **making *n* observations**, the set $x_1, x_2, \ldots, x_n$ represents the values of **the observed data**. How can this tell the truth about the entire process?

Parameter Estimation

- The realized value of an estimator for a particular sample of data, $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, is called an ***estimate*** of θ, denoted by $\hat{\theta}$.

Relationship between a point estimate $\hat{\theta}$ and an unknown parameter θ



Parameter Estimation

- An estimator is a measured value based on a random sample to estimate a parameter, as its **value highly depends** on which particular sample is obtained.

Point Estimators

- A point estimate of an unknown parameter θ is a statistic $\hat{\theta}$ that represents the **best guess** at the value of θ.
- It is done according to some approach **measurement metrics**.

Point Estimators

- An estimator for a population parameter is **unbiased** (impartial) if the difference between the expected value of the estimator and the true value of the parameter being estimated is zero.
- Otherwise, the estimator is said to be **biased**.

Point Estimators

- The estimator $\widehat{\Theta}$ is an ***unbiased estimator*** of the parameter θ, if we have

$$E\left[\widehat{\Theta}\right] = \theta$$

The **bias** of point estimator

- Consider $\widehat{\Theta} = g(X_1, X_2, \cdots, X_n)$ be a point estimator for θ. The **bias** of point estimator $\widehat{\Theta}$ is defined by

$$\mathrm{B}\left[\widehat{\Theta}\right] = E\left[\widehat{\Theta} - \theta\right] = E\left[\widehat{\Theta}\right] - \theta$$

- We say that $\widehat{\Theta}$ is an **unbiased** of estimator of θ if

$$\mathrm{B}\left[\widehat{\Theta}\right] = 0$$

for all possible values of θ.

Exercise 6.1

- Suppose that $\boldsymbol{X}$ is a random variable with mean μ and variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a **random sample** of size n from the population represented by $\boldsymbol{X}$.
- Show that the sample mean $\bar{X}$ and sample variance S^2, computed using the equation below, are unbiased estimators of μ and σ^2, respectively.

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Exercise 6.1: solution

- First the **sample mean** and we already have showed that $E(\bar{X}) = \mu$.
- Therefore, the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ.

$$
\begin{aligned}
E[\overline{X}] &= \frac{EX_1 + EX_2 + \ldots + EX_n}{n} \\
&= \frac{nEX}{n} \text{ (since } EX_i = EX) \\
&= EX.
\end{aligned}
$$

Exercise 6.1: solution

- Now consider the sample variance computed in the way shown below. We have

$$E\,[S^2] = \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E\left[{X_i}^2 - 2\bar{X}X_i + \bar{X}^2\right] = \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[\,{X_i}^2] - nE[\bar{X}^2]\right\}$$

Exercise 6.1: solution

$$E\,[S^2] = \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[\,X_i{}^2] - nE[\bar{X}^2]\right\}$$

- The last equality follows the equation for the mean of a linear function (module 5).
- However, because $E[X_i{}^2] = \mu^2 + \sigma^2$and, $\mathrm{E}[\bar{X}^2] = \mu^2 + \frac{\sigma^2}{n}$, we have

$$E\,[S^2] = \frac{1}{n-1}\left\{\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right\}$$

$$= \frac{1}{n-1}(\mu^2 + \sigma^2 - n\mu^2 - \sigma^2) = \sigma^2$$

→ Conclusion: S^2 computed in that way is an **unbiased** estimator.

The **mean squared error** (MSE)

- The **mean squared error** (MSE) of a point estimator $\widehat{\Theta}$, shown by $MSE(\widehat{\Theta})$, is defined as

$$\mathrm{MSE}[\widehat{\Theta}] = E\left[\left(\widehat{\Theta} - \theta\right)^2\right]$$

The mean squared error (MSE)

- If $\widehat{\Theta}$ is a point estimator for θ

$$\mathrm{MSE}[\widehat{\Theta}] = E\left[\left(\widehat{\Theta} - \theta\right)^2\right] = Var\left(\widehat{\Theta} - \theta\right) + \left(E\left[\widehat{\Theta} - \theta\right]\right)^2 = Var\left(\widehat{\Theta}\right) + B\left(\widehat{\Theta}\right)^2$$

- where $\mathrm{B}\left(\widehat{\Theta}\right) = E\left[\widehat{\Theta} - \theta\right] = E\left[\widehat{\Theta}\right] - \theta$ is the bias of $\widehat{\Theta}$.

- Just remembering: $Var(X + a) = Var(X)$

$$Var(X) = E[(X - \mu_{\boldsymbol{X}})^2] = E[\boldsymbol{X}^2] - (E[\boldsymbol{X}])^2$$

The **mean squared error** (MSE)

- Let $\widehat{\Theta}_1, \widehat{\Theta}_1, \ldots, \widehat{\Theta}_n, \ldots$ be a sequence of point estimators of θ.

- We say that $\widehat{\Theta}_n$ is a **consistent** estimator of θ, if

$$\lim_{n\to\infty} P\left(\left|\widehat{\Theta}_n - \theta\right| \geq \varepsilon\right) = 0, \qquad \text{for all } \varepsilon > 0$$

Point Estimator for Mean

- Based on the Weak Law of Large Numbers (WLLN), the average of $\boldsymbol{X}$ is often a **reasonable point estimator** for the mean.

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Exercise 6.2

- Let $X_1, X_2, \ldots, X_n$ be a **random sample** of size n (***i.i.d.***) from the population represented by $\boldsymbol{X}$, with mean $E(X_i) = \mu = \theta$, and variance $\mathrm{Var}(X_i) = \sigma^2$.

- Show that $\widehat{\Theta}_n = \bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$ (arithmetic average) is a consistent estimator of θ.

Exercise 6.2: solution

- We need to show that

$$\lim_{n\to\infty} P\left(\left|\widehat{\Theta}_n - \theta\right| \geq \varepsilon\right) = 0, \qquad for\ all\ \varepsilon > 0$$

- Using the **Weak Law of Large Numbers**. In particular, we can use Chebyshev's inequality to rewritten below

$$P\left(|X - E[X]| \geq b\right) \leq \frac{Var(X)}{b^2}$$

Exercise 6.2: solution

- We can use Chebyshev's inequality to write

$$P(|\overline{X} - \theta| \geq \epsilon) \leq \frac{\mathrm{Var}(\overline{X})}{\epsilon^2}$$
$$= \frac{\sigma^2}{n\epsilon^2},$$

- which goes to 0 as $n \to \infty$. Therefore, $\widehat{\Theta}_n = \bar{X}$ is a consistent estimator of θ.

Exercise 6.3

- An article in the *Journal of Heat Transfer* (Trans. ASME, Sec. C, 96, 1974, p. 59) described a new method of measuring the thermal conductivity of Armco iron (a company). Using a temperature of 100°F and a power input of 550 watts, the following 10 measurements of **thermal conductivity** (in Btu/hr-ft-°F) were obtained:

$$41.60, 41.48, 42.34, 41.95, 41.86,$$
$$42.18, 41.72, 42.26, 41.81, 42.04$$

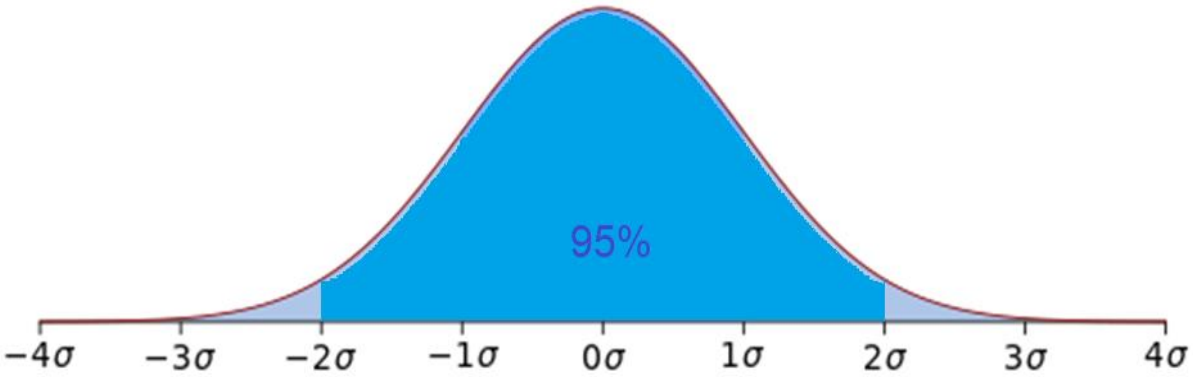
- A point estimate of the mean thermal conductivity at 100°F and 550 watts is the sample mean or $\bar{x} = 41.924 \text{ Btu / hr-ft-°F}$
- We want to estimate the value of $\sigma_{\bar{X}}$.

Exercise 6.3: solution

- The standard error of the sample mean is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, and because σ is unknown, we may replace it by the sample standard deviation estimation S = 0.284 (computed from the random sample) to obtain the estimated standard error of $\bar{X}$ as

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898 \text{ Btu/hr-ft-°F}$$

Exercise 6.3: solution



- Practical Interpretation: Notice that the standard deviation error is about 0.2 percent of the sample mean, implying that we have obtained a relatively precise point estimate of thermal conductivity (frequentist inference interpretation).

- If we can assume that thermal conductivity is normally distributed, 2 times the standard deviation error is $2\sigma_{\overline{X}} = 2(0.0898) = 0.1796$, and we are highly confident (95%) that the true mean thermal conductivity is within the interval [41.924, 42.104] Btu/hr-ft-°F.

Point Estimator for Variance

- Since $E[(X-\mu_{\boldsymbol{X}})^2]$, it is suggested (in this case) the following estimator for the variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2$$

Investigating the **bias** of point estimator in a random sample

- Consider $X_1, X_2, X_3, \ldots, X_n$ be a **random sample (iid)** with mean $E[X_i] = \mu$, and variance $\mathrm{Var}[X_i] = \sigma^2$.

- Suppose that we use the estimator $\bar{S}^2$ for variance

$$\bar{S}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k^2 - 2X_k\bar{X} + \bar{X}^2\right)$$

Investigating the **bias** of point estimator in a random sample

$$\bar{S}^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - 2\bar{X}\frac{1}{n}\sum_{k=1}^{n} X_k + \bar{X}^2\frac{1}{n}\sum_{k=1}^{n} 1$$

$$= \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}^2$$

$$= \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2\right)$$

Investigating the **bias** of point estimator in a random sample

- where the mean estimator is: $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

- To find the bias of this estimator

$$B(\bar{S}^2) = E(\bar{S}^2 - \sigma^2) = E(\bar{S}^2) - \sigma^2$$

- Observe that $Var(\boldsymbol{X}) = E[\boldsymbol{X}^2] - (E[\boldsymbol{X}])^2$

- Then $E[\boldsymbol{X}^2] = (E[\boldsymbol{X}])^2 + Var(\mathbf{X}) = \mu^2 + \sigma^2$

Investigating the **bias** of point estimator in a random sample: Observe that is a frequentist inference

- Since it is a **random sample**: $E[\boldsymbol{X}_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ (**i.i.d.**).

$$E[\bar{\boldsymbol{X}}^2] = (E[\bar{\boldsymbol{X}}])^2 + Var(\bar{\boldsymbol{X}}) = \mu^2 + \frac{1}{n}\sigma^2$$

- Replacing the last result in the expression $E(\bar{S}^2)$

$$E(\bar{S}^2) = \frac{1}{n}E\left[\left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2\right)\right] = \frac{1}{n}\left(\sum_{k=1}^{n} E[X_k^2] - nE[\bar{X}^2]\right)$$

$$= \frac{1}{n}\left\{n(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{1}{n}\sigma^2\right)\right\} = \frac{n-1}{n}\sigma^2$$

Investigating the **bias** of point estimator in a random sample

- We conclude that is $\bar{S}^2$ a **biased estimator of the variance** in a random sample, and $B(\bar{S}^2)$ can be expressed as

$$B(\bar{S}^2) = E(\bar{S}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

Investigating the **bias** of point estimator in a random sample

- Since $E(\bar{S}^2) = \frac{n-1}{n}\sigma^2$, it is possible to obtain an unbiased estimator S^2 by mutipling the estimator of σ^2 by $\frac{n}{n-1}$

$$S^2 = \frac{n}{n-1}\bar{S}^2 = \frac{1}{n-1}\left(\sum_{k=1}^{n}(X_k - \bar{X})^2\right)$$

The Maximum Likelihood Estimator (MLE)

The Maximum Likelihood Estimator (MLE)

- Maximum likelihood estimation selects as its estimate the unknown, but deterministic, value of the parameter that maximizes the probability of the observed data.

The Maximum Likelihood Estimator (MLE)

- Maximum likelihood estimators **are not optimal in general**.
- It can be implemented for complex estimation problems, and it can also yield **optimal performance** when the **data set is large enough**, i.e. it is asymptotically efficient.

The Maximum Likelihood Estimator (MLE)

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution with a parameter θ. Given that we have observed $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, a maximum likelihood estimate of θ, shown by $\hat{\theta}_{ML}$ is a value of θ that maximizes the likelihood function

$$L(x_1, x_2, \cdots, x_n; \theta).$$

A maximum likelihood estimator (MLE) of the parameter θ, shown by $\hat{\Theta}_{ML}$ is a random variable $\hat{\Theta}_{ML} = \hat{\Theta}_{ML}(X_1, X_2, \cdots, X_n)$ whose value when $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ is given by $\hat{\theta}_{ML}$.

Maximum Likelihood Estimators for iid random variables

- Suppose we have a sequence of ***n*** **iid** observed values of a particular **random sample**, $x_1, x_2, \ldots, x_n$, for the random variable X, and assume θ is the parameter of interest to be estimated.
- We now define the ***likelihood function*** $L(\theta)$ that is a function of both the n observed sample values and the unknown, but deterministic (frequentist inference), parameter θ, as follows (???):

$$L(\theta)=\begin{cases} f_X(x_1,x_2,\ldots,x_n;\theta)=f_X(x_1;\theta)f_X(x_2;\theta)\cdots f_X(x_n;\theta)=\prod_{i=1}^{n} f_X(x_i;\theta) \\ \text{where } X \text{ is a continuous random variable} \\ p_X(x_1,x_2,\ldots,x_n;\theta)=p_X(x_1;\theta)p_X(x_2;\theta)\cdots p_X(x_n;\theta)=\prod_{i=1}^{n} p_X(x_i;\theta) \\ \text{where } X \text{ is a discrete random variable} \end{cases}$$

Log-Likelihood function

- The ***log-likelihood function*** is defined as

$$\ln L(\theta) = \begin{cases} \sum_{i=1}^{n} \ln f_X(x_i;\theta) \text{ where } X \text{ is a continuous random variable} \\ \sum_{i=1}^{n} \ln p_X(x_i;\theta) \text{ where } X \text{ is a discrete random variable} \end{cases}$$

Maximum Likelihood Estimation

- Likelihood of θ given the **i.i.d.** sample $\mathcal{X}$ drawn from $\mathrm{P}(x|\theta)$

$$L(\mathcal{X}|\theta) \equiv P(\mathcal{X}|\theta) = \prod_k p(x_k|\theta)$$

- Log likelihood

$$\mathcal{L}(\mathcal{X}|\theta) \equiv \ln\{L(\mathcal{X}|\theta)\} = \sum_k \ln\{P(x_k|\theta)\}$$

- Maximum likelihood estimator (MLE)

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\mathcal{X}|\Theta)$$

Maximum Likelihood Estimators

- The maximum likelihood method selects the estimator value $\widehat{\Theta}_{ML} = \hat{\theta}$, where θ is the **parameter** value that maximizes the value of $L(\theta)$ or $\mathcal{L}(\theta)$,whichever is easier to maximize.
- The ***maximum likelihood estimation*** helps select the parameter value $\hat{\theta}$ that maximizes the **probability** of the particular observation sequence $x_1, x_2, \ldots, x_n$.

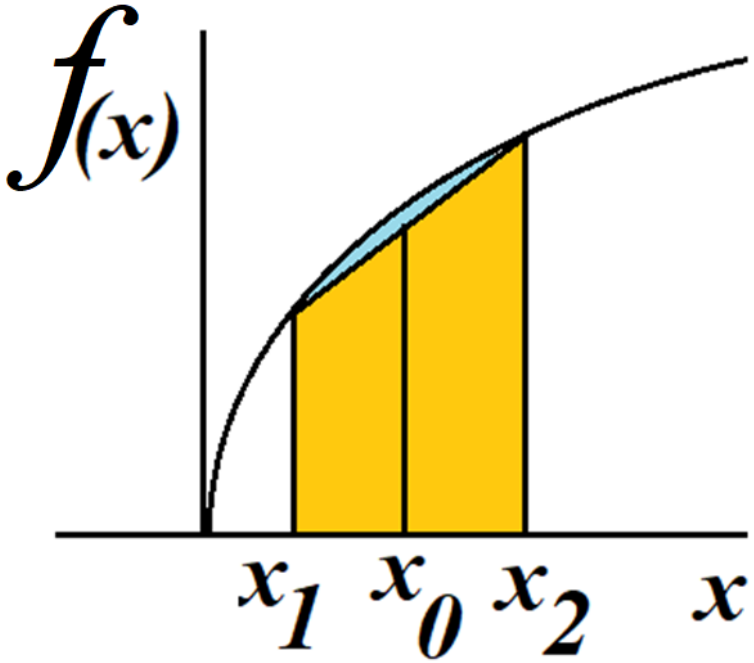
Maximum Likelihood Estimation for **Gaussian distribution**

- Let us say we have an independent and identically distributed (iid) sample of a random variable with $\boldsymbol{N}$ instances $X = \{x_k\}_{k=1}^{N}$.
- Consider the instances are drawn from some known probability density family, $P(x|\theta)$, where $\boldsymbol{X} \sim \mathcal{N}(x|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$

$$f_X(x) \approx P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

Observe: from module 3

- Comment: The approximation in calculating probabilities for continuous variables can be computed considering the trapezium area below



(1) Regular Polygon: The infinitesimal area segments can be approximated by rectangles.

$$P(x < X \leq x + \delta) \approx f_X(x)\delta$$

Or (2) Irregular polygon : $P(x_1 \leq \mathrm{x} \leq x_2) =$

$\int_{x_1}^{x_2} f_X(x)dx \cong \frac{\{f(x_2)+f(x_1)\}(x_2-x_1)}{2} = f(x_0)(x_2 - x_1)$ -- The trapezium Euler integration method.

Maximum Likelihood Estimation for Gaussian distribution (for independent random variables)

- We want to find θ that makes sampling $X = \{x_k\}_{k=1}^{N}$ from $P(x|\theta)$ as likely as likelihood possible.
- Because $X = \{x_k\}_{k=1}^{N}$ are independent, the *likelihood* of parameter θ given sample $\boldsymbol{X}$ is the product of the likelihoods of the individual points:

$$L(\mathcal{X}|\theta) \equiv P(\mathcal{X}|\theta) = \prod_{k=1}^{N} P(x_k|\theta)$$

Maximum Likelihood Estimation for Gaussian distribution

- In *maximum likelihood estimation*, we are interested in finding θ that estimation makes X the most likely to be drawn.

- We can maximize the **log** (*ln*) of the likelihood without changing the value where it takes its maximum.

Maximum Likelihood Estimation for Gaussian distribution

- $ln(\cdot)$ converts the product into a sum and leads to further computational simplification when certain densities are assumed, for example, containing exponents.

Maximum Likelihood Estimation for Gaussian distribution (for independente random variables)

$$\mathcal{L}(\mathcal{X}|\theta) \equiv \ln\{L(\mathcal{X}|\theta)\} = ln\left\{\prod_{k=1}^{N} P(x_k|\theta)\right\} = ln\left\{\prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right\}$$

$$= \sum_{k=1}^{N} ln\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right\}$$

$$= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2$$

Maximum Likelihood Estimation for Gaussian distribution (for independente random variables)

- The Maximum log-likelihood estimator (MLE) is

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\mathcal{X}|\theta)$$

- To obtain the Maximum log-likelihood estimator (MLE) for expectation $\hat{\mu}$ and variance $\hat{\sigma}^2$

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu} = 0 \qquad \frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma^2} = 0$$

Maximum Likelihood Estimation for Gaussian distribution

- The **Maximum log-likelihood estimator** (MLE) for **mean** is

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \mu} = \frac{\partial}{\partial \mu}\left\{\frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2\right\} = \frac{1}{\sigma^2}\sum_{k=1}^{N}(x_k-\mu) = 0$$

$$= \sum_{k=1}^{N} x_k - \sum_{k=1}^{N}\mu = 0 \quad \Rightarrow \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

Maximum Likelihood Estimation for Gaussian distribution

- The **Maximum log-likelihood estimator** (MLE) for **variance** is

$$\frac{\partial \mathcal{L}(\mathcal{X}|\theta)}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left\{-\frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2\right\}$$

$$= -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{k=1}^{N}(x_k-\mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k-\mu)^2$$

Bayesian Inference: it makes sense

the unknown quantity

initial guess

prior

likelihood

posterior

$$P(C|x) = \frac{P(C)P(\boldsymbol{x}|C)}{P(x)}$$

evidence

Asymptotic Properties of Maximum Likelihood Estimator (MLE)

Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from a distribution with a parameter θ. Let $\hat{\Theta}_{ML}$ denote the maximum likelihood estimator (MLE) of θ. Then, under some mild regularity conditions,

1. $\hat{\Theta}_{ML}$ is asymptotically consistent, i.e.,

$$\lim_{n\to\infty} P(|\hat{\Theta}_{ML} - \theta| > \epsilon) = 0.$$

   \item $\hat{\Theta}_{ML}$ is asymptotically unbiased, i.e.,

$$\lim_{n\to\infty} E[\hat{\Theta}_{ML}] = \theta.$$

2. As n becomes large, $\hat{\Theta}_{ML}$ is approximately a normal random variable. More precisely, the random variable

$$\frac{\hat{\Theta}_{ML} - \theta}{\sqrt{\mathrm{Var}(\hat{\Theta}_{ML})}}$$

   converges in distribution to $N(0,1)$.

Discriminant Functions

GPDS
Grupo de Processamento Digital de Sinais
Prof. F. Assis

Obtaining discriminant functions

- Consider the random variables modeled as

$$X \sim \mathcal{N}(x|\mu_X, \Sigma_X) \Rightarrow f_X(X) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}\{(x-\mu_X)^T \Sigma_X^{-1} (x-\mu_X)\}}$$

- The vector that characterizes the expectancy is given by $\mu_X = \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_d] \end{bmatrix}$

- and the covariance matrix can be written as

$$\Sigma_X = \mathrm{E}[(X - E[X])^T (X - E[X])] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1d} & \sigma_{2d} & \cdots & \sigma_d^2 \end{bmatrix}$$

Obtaining discriminant functions

- where

$$\sigma_{ji} = \sigma_{ij} = Cov\left(X_i, X_j\right) = (X_i - E[X_i])\left(X_j - E\left[X_j\right]\right)$$

$$\sigma_{\mathrm{i}}^2 = E\left[(X_i - E[X_i])^2\right]$$

Obtaining discriminant functions

- Taking the naïve-Bayes parametric classifier

the unknown quantity

initial guess

prior

likelihood

posterior

$$P(C|x) = \frac{P(C)P(\boldsymbol{x}|C)}{P(x)} = \frac{P(C_k)P(\mathbf{x}|C_k)}{\sum_{i=1}^{d} P(C_i)P(\mathbf{x}|C_i)}$$

evidence

Obtaining discriminant functions

- (From Exercise 3.3) We want to maximize the probability of correctly estimating the class (diagnosis), that is:

$$\text{Diagnosis} = \max_k \{P(C_k|x_i)\}$$

- where

$$P(C_0|x_i) = \frac{P(x_i|C_0)P(C_0)}{P(x_i|C_0)P(C_0) + P(x_i|C_1)P(C_2) + P(x_i|C_2)P(C_2)}$$

$$P(C_1|x_i) = \frac{P(x_i|C_1)P(C_1)}{P(x_i|C_0)P(C_0) + P(x_i|C_1)P(C_2) + P(x_i|C_2)P(C_2)}$$

$$P(C_2|x_i) = \frac{P(x_i|C_2)P(C_2)}{P(x_i|C_0)P(C_0) + P(x_i|C_1)P(C_2) + P(x_i|C_2)P(C_2)}$$

Obtaining discriminant functions

- In the classification process, the denominator is a common element for the estimation of $P(C_i|\mathbf{x})\ \forall\ i$.
- So, it can be discarded in the classification process.

- We can define a discriminant function as $g_k(x) = P(x_i|C_k)P(C_0)$

- Taking the logarithm on both sides of the last equation

$$G_k\ (x_i) = ln\{P(C_k)P(x_i|C_{\boldsymbol{k}})\} = \ ln\{P(C_k)\} + ln\{P(x_i|C_{\boldsymbol{k}})\}$$

$$\text{Diagnosis} = \max_k\{G_k(x_i)\}$$

Obtaining discriminant functions

- Assuming a Gaussian Model for input data ($\boldsymbol{x} \equiv x_i$): $X \sim \mathcal{N}(\boldsymbol{x}|\mu_X, \Sigma_X)$

$$G_k(x) = ln\left\{\frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}}\, e^{-\frac{1}{2}\{(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\mu_k)\}}\right\} + \; ln\{P(C_k)\}$$

$$G_k(x) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}ln|\Sigma_k| \; - \; \frac{1}{2}\{(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\mu_k)\} + \; ln\{P(C_k)\}$$

Obtaining discriminant functions

- We can also discard the constant term common to all discriminative functions $\{-\frac{d}{2}\ln(2\pi)\}$, resulting in

$$G_k(\boldsymbol{x}) = -\frac{1}{2}ln|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}-\mu_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\mu_k) + ln\{P(C_k)\}$$

- Expanding the array product

$$G_k(\boldsymbol{x}) = -\frac{1}{2}ln|\Sigma_i| - \frac{1}{2}\left(\boldsymbol{x}^T\Sigma_k^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T\Sigma_k^{-1}\mu_k + \mu_k^T\Sigma_k^{-1}\mu_k\right) + ln\{P(C_k)\}$$

Obtaining discriminant functions: **quadratic discriminant**

- We can define the quadratic discriminant from the last equation as

$$G_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W}_k \boldsymbol{x} + \boldsymbol{w}_k{}^T \boldsymbol{x} + w_{\mathrm{k0}}$$

- where

$$\boldsymbol{W_k} = -\frac{1}{2}\Sigma_i^{-1}\,\{a\ matrix\ (d\ \mathrm{x}\ d)\}$$

$$\boldsymbol{w_k} = \Sigma_i^{-1}\mu_k\,\{a\ column\ vector\ (d\ \mathrm{x}\ 1)\}$$

$$w_{k0} = -\frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}ln|\Sigma_k| + \ ln\{P(C_k)\}$$

- Observe that

$$\boldsymbol{x}^T\Sigma_k^{-1}\mu_k = \left(\Sigma_k^{-1}\mu_k\right)^T\boldsymbol{x}$$

Obtaining the **linear discriminant**

- Observe the **quadratic discriminant**

$$G_k(x) = -\frac{1}{2} ln|\Sigma_i| - \frac{1}{2}(x-\mu_k)^T \Sigma_i^{-1}(x-\mu_k) + ln\{P(C_i)\}$$

- If it is assumed that all classes have the same **covariance matrix**, $(-\frac{1}{2} ln|\Sigma_i| = constant)$ the **quadratic discriminant** reduces to

$$G_k(x) = -\frac{1}{2}(\boldsymbol{x}-\mu_k)^T \Sigma_k^{-1}(\boldsymbol{x}-\mu_k) + ln\{P(C_k)\}$$

$$G_k(x) = -\frac{1}{2}\left(\boldsymbol{x}^T \Sigma_k^{-1}\boldsymbol{x} - 2\boldsymbol{x}^T \Sigma_k^{-1}\mu_k + \mu_k^T \Sigma_k^{-1}\mu_k\right) + ln\{P(C_i)\}$$

Obtaining the **linear discriminant**

- Note that, in this case, after this simplification the quadratic term $x^T \Sigma_k^{-1} x$ also cancels out since it is common to all G_i discriminant functions.
- Consequently, for this case the quadratic discriminant can be rewritten as a first-order function.
- This representation is called a **linear discriminant** function.

$$\hat{G}_k(\boldsymbol{x}) = \boldsymbol{x}^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \ ln\{P(C_k)\}$$

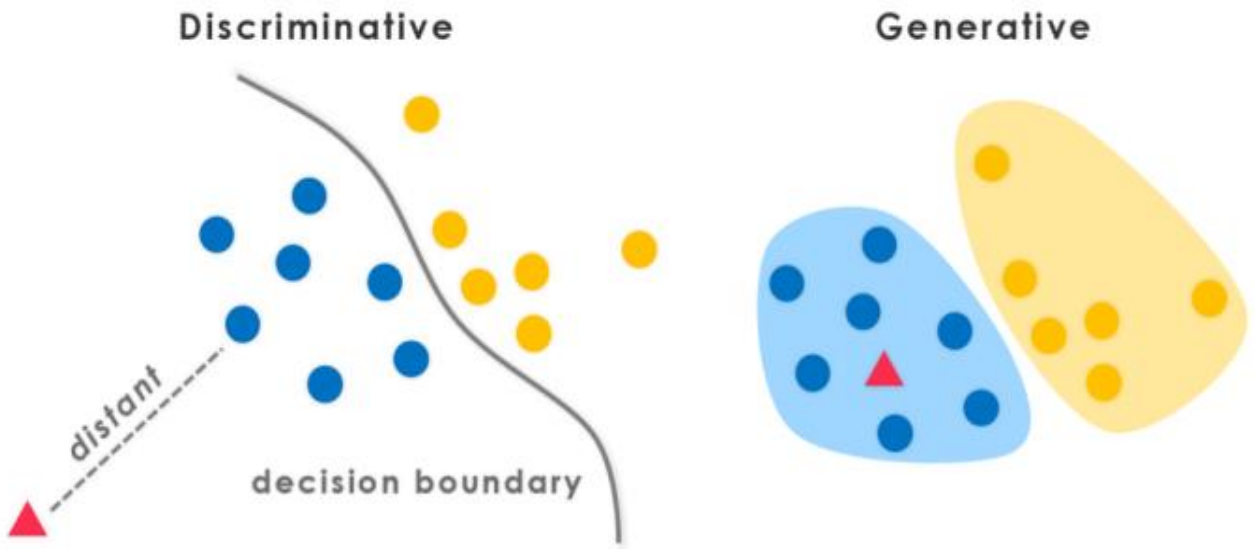
- In compact form the linear discriminant function can be expressed as

$$\hat{G}_k(\boldsymbol{x}) = \boldsymbol{w}_k{}^T \boldsymbol{x} + w_{\mathrm{k0}}$$

$$\text{and } \boldsymbol{w_k} = \Sigma_k^{-1} \mu_k; w_{\mathrm{k0}} = -\frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \ ln\{P(k)\}$$

Discriminative and Generative algorithms

- Discriminative Learning (linear and quadratic discriminants): A discriminative classifier tries to model by just depending on the observed data.



- Generative Learning (Naïve Bayes): Generative approaches try to build a model of the positives and a model of the negatives. A decision boundary is formed where one model becomes more likely. As these create models of each class they can be used for generation. To create these models, a generative learning algorithm learns the joint probability distribution $f(x_1, x_2, \ldots, x_n)$

Confidence interval
Confidence level
Sample Size

Confidence interval (frequentist inference)

- The point estimate $\hat{\theta}$ alone does not give much information about θ.
- Without additional information, we do not know how close $\hat{\theta}$ is to the real θ.

Confidence interval

- There is a range (interval) of $\hat{\theta}$.
- The concept of **interval estimation** leads us to a precision measure of our estimator.
- In this interval for θ, there is a: *high estimate* $\hat{\theta}_h$ and a low estimate $\hat{\theta}_l$.

Confidence interval

- An estimate that specifies a range of values is called an ***interval estimate***.
- Assuming with a **high probability of $(1-\alpha)$**, the true value lies in the estimated range of the possible values for the parameter.
- **$(1-\alpha) \times 100\%$** is then called the ***confidence level*** or ***confidence coefficient***.
- The end-points of a confidence interval are called the ***confidence limits***.

Confidence interval

- The **length** of the reported interval (***confidence limits***), $(\hat{\theta}_h - \hat{\theta}_l)$, shows the precision with which we can estimate θ.

- The **confidence level** shows how confident we are about the measurement interval.

Confidence interval

- Consider $X_1, X_2, X_3, \ldots, X_n$ **a random sample** from a distribution with a parameter θ that is to be estimated.

- Suppose that we have observed

$$X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n.$$

Confidence interval

- An **interval estimator** with **confidence level** $(1-\alpha)$ consists of two estimators

$$\widehat{\Theta}_l = \widehat{\Theta}_l(x_1, x_2, \dots x_n) \quad \widehat{\Theta}_h = \widehat{\Theta}_h(x_1, x_2, \dots x_n)$$

- Such that $P\left(\widehat{\Theta}_l \leq \theta \leq \widehat{\Theta}_h\right) \geq 1-\alpha$

- for every possible value of the estimated parameter θ.

Confidence interval

- It can also be equivalently written as

$$P\left(\theta \in [\widehat{\Theta}_l, \widehat{\Theta}_h]\right) \geq 1 - \alpha$$

- We say that $[\widehat{\Theta}_l, \widehat{\Theta}_h]$ is a **$(1-\alpha) \times 100\%$ confidence interval** for the estimated parameter θ.

Confidence interval

- The randomness in these approaching is due to $\widehat{\Theta}_l$ and $\widehat{\Theta}_h$, not to the parameter θ.

- The parameter θ is the unknown quantity which is assumed to be non-random (frequentist inference).

Confidence interval

- The estimator Θ may have been designed by deterministic rules.

- On the other hand, $\widehat{\Theta}_l$ and $\widehat{\Theta}_h$ are random variables because they are functions of the observed random variables

$$X_1, X_2, X_3, \ldots, X_n.$$

Confidence interval

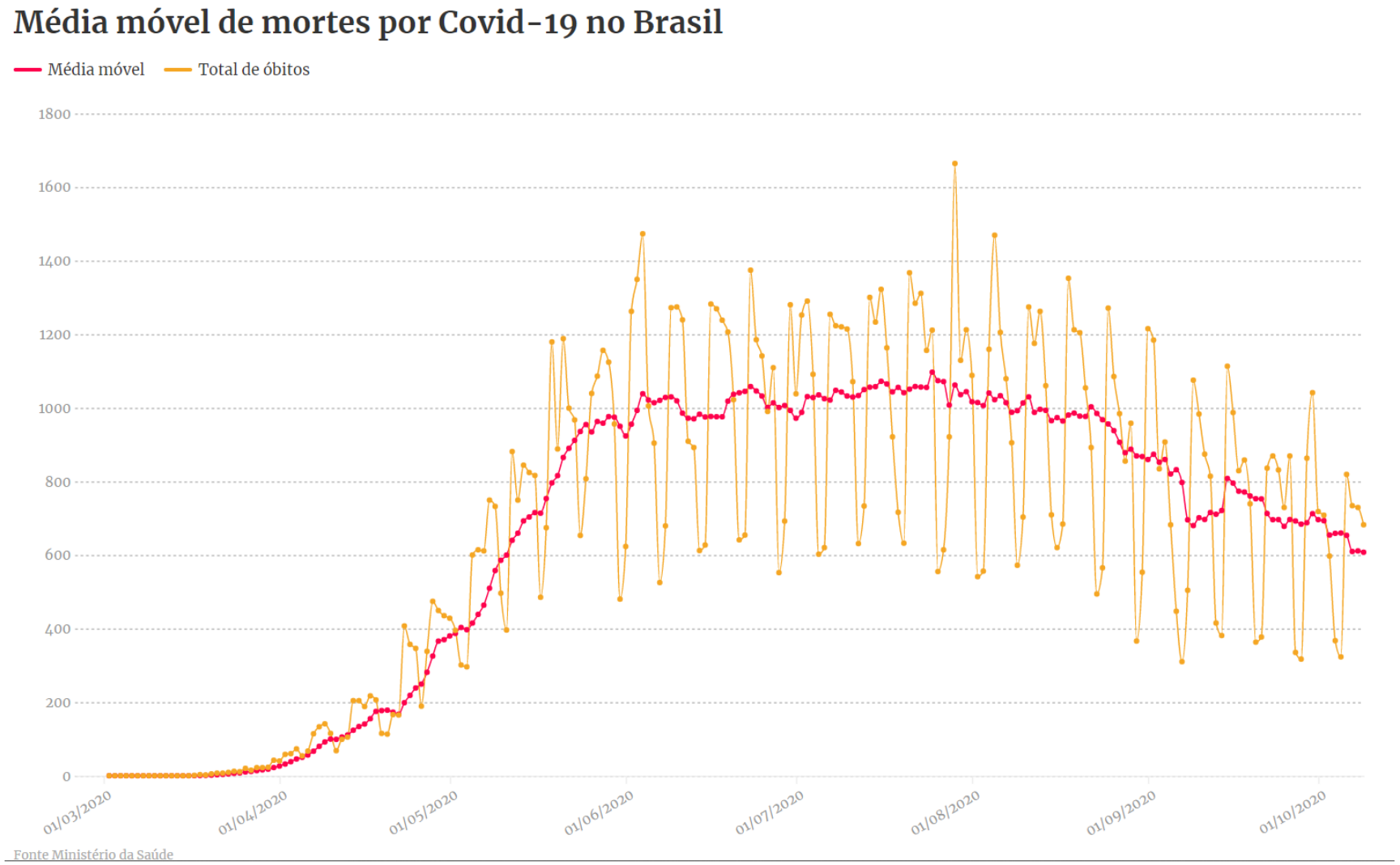
- Example - The moving average estimator:

$$\bar{X}_N = y[n]$$
$$= \frac{1}{N}\sum_{k=0}^{N-1} x[n-k] = \frac{1}{N}(x[n] + x[n-1] + \cdots + x[n-N+1])$$

- The integer variable **k** specifies the sample delay value.
- The integer variable **n** specifies the time index of the output value $y[n]$ of the moving average estimator.

Confidence interval

- Example



Confidence interval

- The confidence intervals and confidence levels are numerical measures of the accuracy.

- A narrower confidence interval reflects a more accurate estimate.

- A higher confidence level indicates a more reliable estimate.

Finding Interval Estimators

- Let $\boldsymbol{X}$ be a continuous Gaussian random variable with CDF $F_X(x) = \mathbf{P}(X \leq x)$.

- Suppose that we are interested in finding two values x_h and x_l such that

$$\mathrm{P}(x_l \leq X \leq x_h) \geq 1 - \alpha.$$

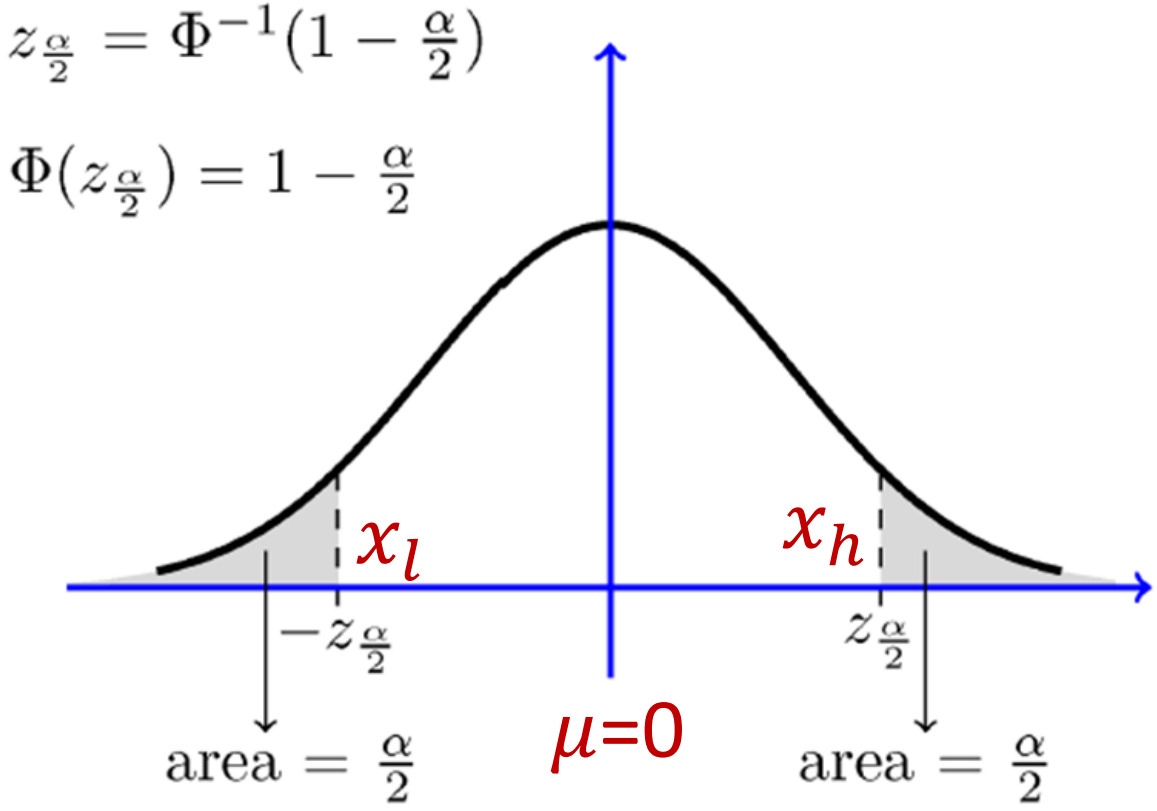
Finding Interval Estimators

- One way to do this, is to chose x_l and x_h such that

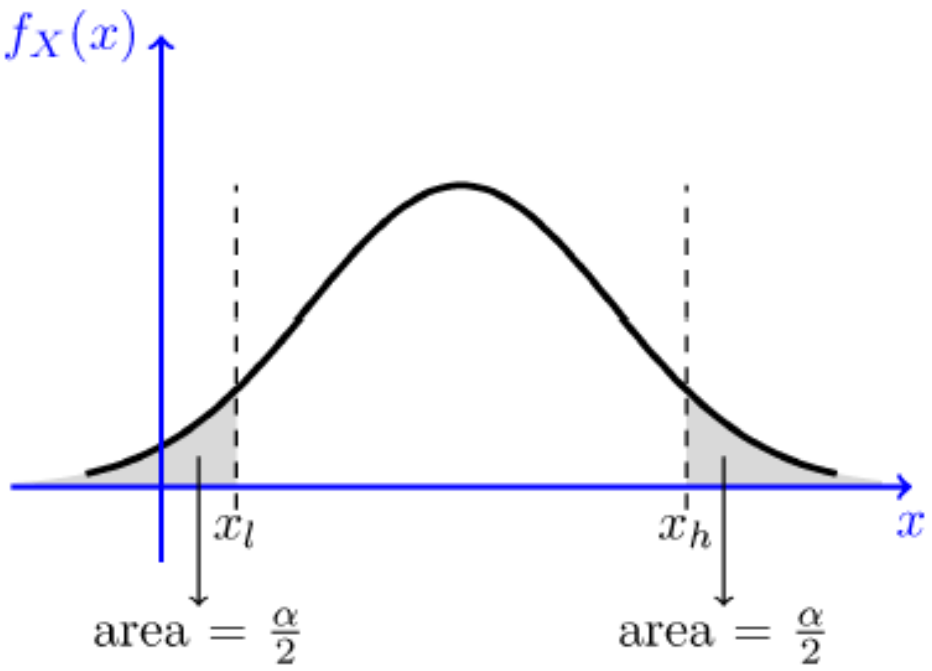
$$P(X \leq x_l) = \frac{\alpha}{2} \; and \; P(X \geq x_h) = \frac{\alpha}{2}$$

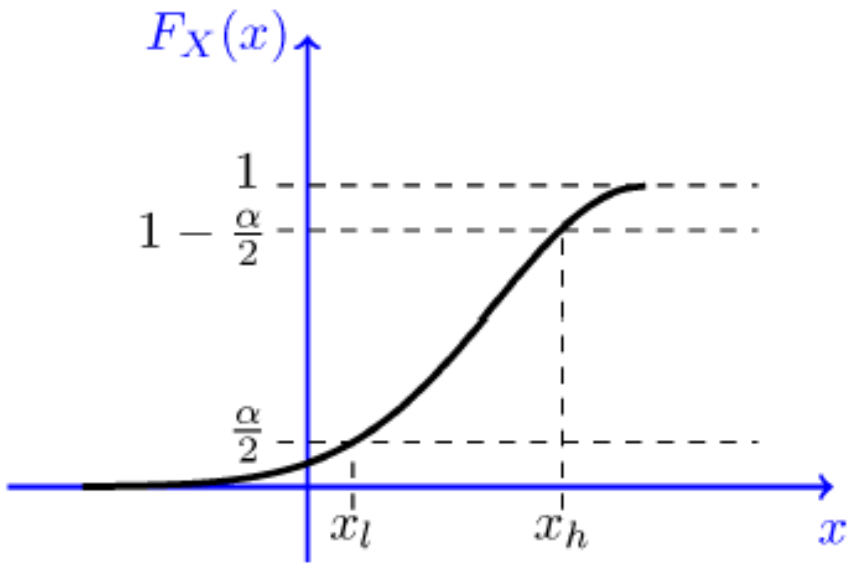
- We can rewrite these equations by using the inverse function F_X^{-1} as

$$x_l = F_X^{-1}\left(\frac{\alpha}{2}\right), \text{and } x_h = F_X^{-1}\left(1 - \frac{\alpha}{2}\right)$$



Finding Interval Estimators





$[x_l, x_h]$ is a $(1-\alpha)$ interval for X, that is,

$$P\Big(x_l \leq X \leq x_h\Big) = 1 - \alpha.$$

Example: Confidence Intervals for the sample mean $\bar{X}_n$ (frequentist inference)

- We assume that the sampling is either from an infinite population or with replacement from a finite population so as to make the arithmetic simple.

- We also assume either the observed samples $X_1, X_2, \ldots, X_n$ are **iid** Gaussian random variables or ***n*** is large enough to allow us to approximate the sample mean as a Gaussian random variable using **the central limit theorem**.

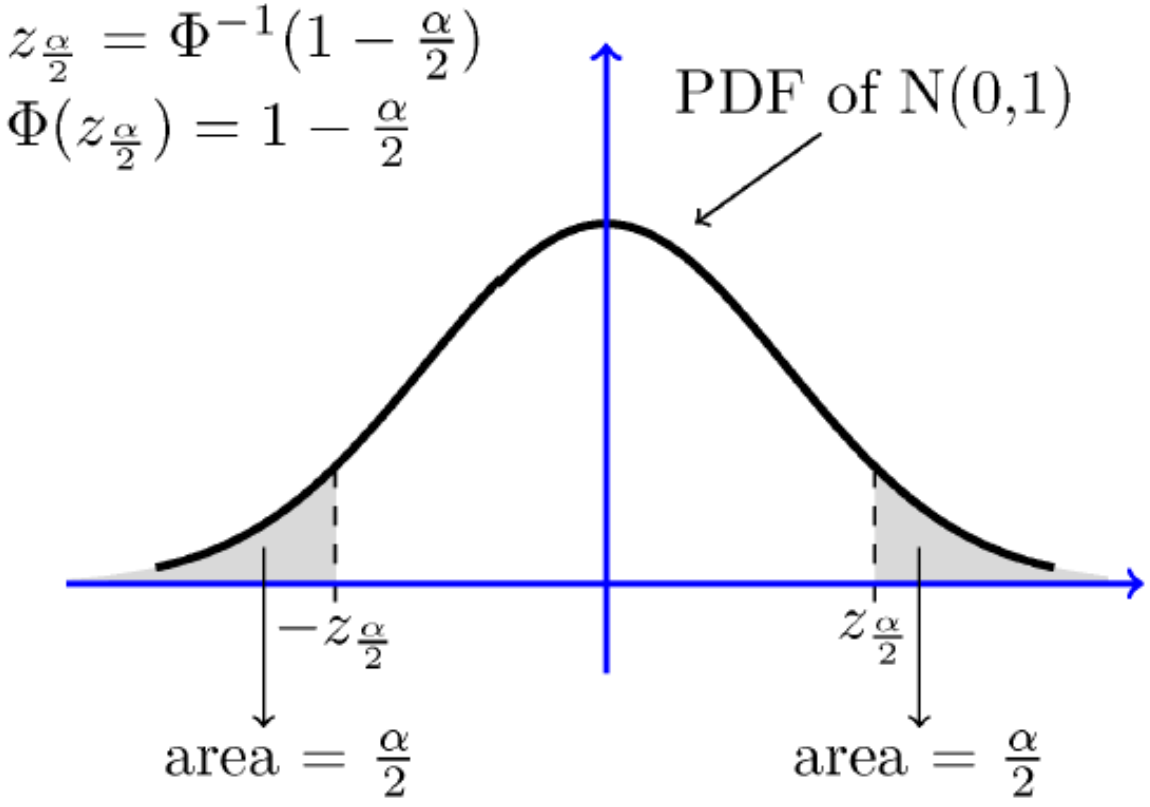
Example: Confidence Intervals for the sample mean $\bar{X}_n$

- Observe that the **sample mean** $\bar{X}_n$ is a Gaussian random variable with unknown mean μ_X and known variance $\frac{\sigma_X^2}{n}$.
- We then have a randomic variable $Z_n = \frac{\bar{X}_n - \mu_X}{\sigma_X \sqrt{n}}$ (from Central Limit Theorem) as the standard (zero-mean, unit-variance) Gaussian random variable:

$$Z_n \sim \mathcal{N}(z|\mu, \sigma^2) = \mathcal{N}(z|0,1).$$

Example: Confidence Intervals for the sample mean

- Based on the Central Limit Theorem, we need to recover x_l (from $-Z_{\alpha/2}$) and x_h (from $Z_{\alpha/2}$ in such a way that we have the following:



$$P(-Z_{\alpha/2} \leq Z_n \leq Z_{\alpha/2}) = 1 - \alpha$$

Example: Confidence Intervals for the sample mean

- Our problem can also be postulated in the following ways:

$$P\left(Z_n \in [-Z_{\alpha/2}, Z_{\alpha/2}]\right) \geq 1-\alpha$$

$$P\left(-Z_{\frac{\alpha}{2}} \leq Z_n = \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \leq Z_{\alpha/2}\right) \geq 1-\alpha$$

Example: Confidence Intervals for the sample mean

- Now manipulate the quantities inside the brackets by (1) multiplying through by $\frac{\sigma_X}{\sqrt{n}}$,

$$P\left(-\frac{\sigma_X}{\sqrt{n}}Z_{\frac{\alpha}{2}} \leq \bar{X}_n - \mu_X \leq \frac{\sigma_X}{\sqrt{n}}Z_{\alpha/2}\right) \geq 1-\alpha$$

Example: Confidence Intervals for the sample mean

- (2) subtracting $\bar{X}_n$ from each term

$$P\left(-\bar{X}_n - \frac{\sigma_X}{\sqrt{n}} Z_{\frac{\alpha}{2}} \leq -\mu_X \leq -\bar{X}_n + \frac{\sigma_X}{\sqrt{n}} Z_{\frac{\alpha}{2}}\right) \geq 1 - \alpha$$

- and (3) multiplying through by -1.

$$P\left(\bar{X}_n - \frac{\sigma_X}{\sqrt{n}} Z_{\frac{\alpha}{2}} \leq \mu_X \leq \bar{X}_n + \frac{\sigma_X}{\sqrt{n}} Z_{\frac{\alpha}{2}}\right) \geq 1 - \alpha$$

Example: Confidence Intervals for the sample mean

- Assumptions: A random sample $X_1, X_2, X_3, \ldots, X_n$ is given from a distribution with known variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$ for a large n.
- The $(1-\alpha) \times 100\%$ confidence level for the confidence interval obtained for the mean $\mu_X = E[X_i]$ is thus given by

$$\left[\bar{X}_n - Z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}, \bar{X}_n + Z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}\right]$$

Exercise 6.4

- ASTM (**American Society for Testing and Materials**) Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (Joule - J) on specimens of A238 steel cut at 60ºC are as follows:

  64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3.

- Assume that impact energy is normally distributed with $\sigma = 1$J. We want to find a 95% confidence interval for μ, the mean impact energy. The required quantities are $z_{\alpha/2} = z_{0.025} = 1.96$ (for $\mathcal{N}(z|0,1)$), $n = 10$, $\sigma = 1$, and $\bar{x} = 64.46$.

- ----------

- (In Portuguese): O padrão E23 define métodos de teste padrão para teste de impacto de barra entalhada de materiais metálicos. A técnica "Charpy V-notch (CVN)" mede a energia de impacto e, é frequentemente usada para determinar se um material experimenta ou não uma transição dúctil-para-frágil com a diminuição da temperatura. Dez medições de energia de impacto (J) em amostras de aço A238 cortadas a 60ºC são as seguintes:

  64,1, 64,7, 64,5, 64,6, 64,5, 64,3, 64,6, 64,8, 64.2, e 64,3.

- Suponha que a energia do impacto seja normalmente distribuída com $\sigma = 1$J. Queremos encontrar um intervalo de confiança de 95% para μ, a energia média do impacto. As quantidades necessárias são $z_{\alpha/2} = z_{0,025} = 1,96$ (para $\mathcal{N}(z|0,1)$), $n = 10$, $\sigma = 1$ e $\bar{x} = 64.46$.

Exercise 6.4: solution

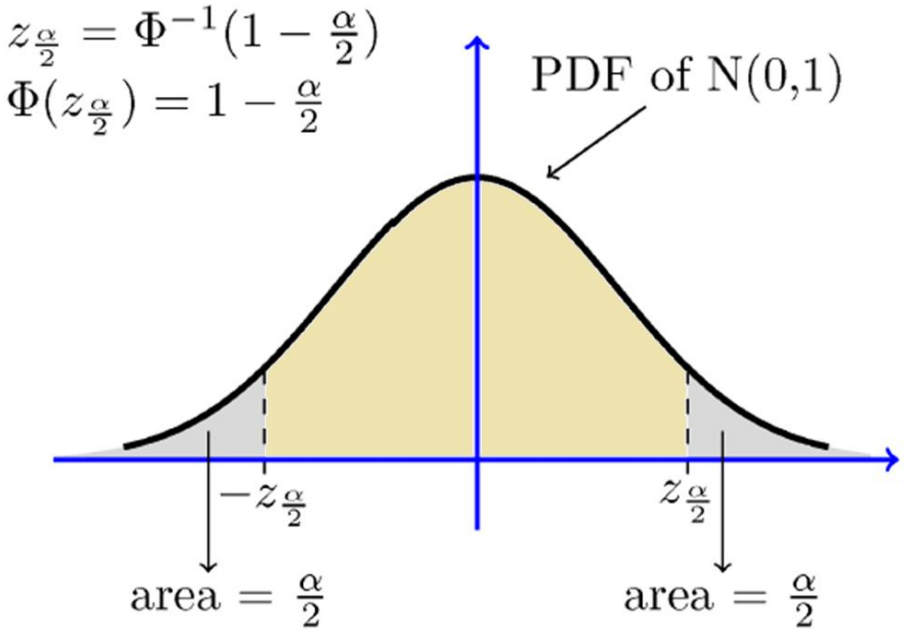
- The resulting 95% confidence interval is found from equation below as follows (remember, it is the sample mean):

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$64.46 - 1.96 \frac{1}{\sqrt{10}} \leq \mu \leq 64.46 + 1.96 \frac{1}{\sqrt{10}}$$

$$63.84 \leq \mu \leq 65.08$$



Exercise 6.4: solution

- Practical Interpretation: Based on the sample data, a range of highly plausible values for mean impact energy for A238 steel at 60°C is $63.84\ J \leq \mu \leq 65.08 J$.

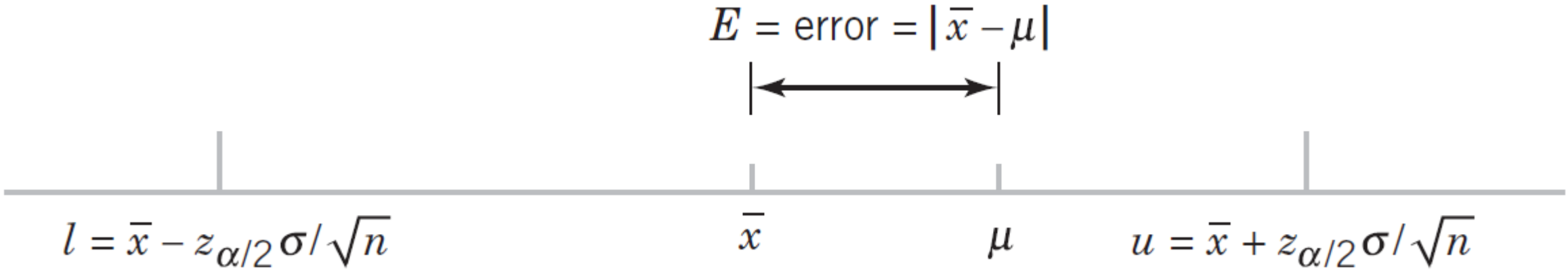
"We are 95% confident that the mean impact energy of A238 steel cut at 60ºC belongs to the interval: $63.84\ J \leq \mu \leq 65.08 J$."

Choice of Sample Size

- The range (precision) of the confidence interval in equation

$$\left[\bar{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma_X}{\sqrt{n}}, \bar{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma_X}{\sqrt{n}}\right] \text{ is } 2z_{\frac{\alpha}{2}}\frac{\sigma_X}{\sqrt{n}}.$$

- This means that in using $\bar{X}$ to estimate μ, the error $E = |\bar{X} - \mu|$ is less than or equal to $z_{\frac{\alpha}{2}}\frac{\sigma_X}{\sqrt{n}}$ n with confidence $100(1-\alpha)\%$. This is shown graphically in figure below.



Choice of Sample Size

- In situations whose sample size can be controlled, we can choose n so that we are 100 x (1 − α)% confident that the error in estimating μ is less than a specified bound on the error E.
- The appropriate sample size is found by choosing **n** such that $z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}} = E$. Solving this equation gives the following formula for n.
- If x is used as an estimate of μ, we can be 100 x (1 − α)% confident that the error $E = |\bar{X} - \mu|$ will not exceed a specified amount E when the sample size is

$$n = \left(z_{\frac{\alpha}{2}} \frac{\sigma_X}{E} \right)^2$$

Exercise 6.5

- To illustrate the use of this procedure, consider the CVN test described in Exercise 6.4 and suppose that we want to determine how many specimens must be tested to ensure that the 95% confidence interval (CI) on μ for A238 steel cut at 60°C has a length of at most 1.0J.

- Because the bound-on error in estimation E is one-half of the length of the confidence interval, to determine **n**, we choose $E = 0.5$, $\sigma = 1$, and $z_{\alpha/2} = 1.96$.

Exercise 6.5

- The required sample size is,

$$n = \left(z_{\frac{\alpha}{2}} \frac{\sigma_X}{E}\right)^2 = \left(1.96 \frac{1}{0.5}\right)^2 = 15{,}37$$

- and because **n** must be an integer, the required sample size is **n** = 16.

Confidence Level - Confidence Interval - Sample Size (for the sample mean)

Notice the general relationship between **sample size**, desired length of the **confidence interval** 2E, **confidence level** $100 \times (1 - \alpha)\%$, and standard deviation σ:

- As the desired length of the **confidence interval** 2E decreases, the required sample size **n** increases for a fixed value of σ and specified confidence level.

Confidence Level - Confidence Interval - Sample Size (for the sample mean)

- As σ increases, the required sample size **n** increases for a fixed desired length 2E of **confidence interval** and specified **confidence level**.

- As the **level of confidence** increases, the required **sample size n** increases for fixed desired length 2E of **confidence interval** and standard deviation σ.

Exercise 6.6

- A random experiment using the Naive Bayes algorithm for diagnosing breast cancer was run 50 times.
- The original dataset features file has: 699 instances.
- Percentage of Data Bank used for training: 90.0% - 629 instances (70 instances used for evaluate the algorithm).
- Number of classes in the dadaset is 2 (benign, malignant), and the input data space dimension is 9.
- Assume that the accuracy has a Gaussian distribution.

Exercise 6.6: solution

- The average accuracy: 96.17% ($\bar{X} = 96.17$). The standard deviation of measured average accuracy was 1.72%. ($S = \sigma_{\bar{X}} = 1.72$).
- a) What is the **confidence interval** for to have a **confidence level** equal to 95%?
- b) What is the **confidence interval** for to have a **confidence level** equal to 99.7%?

- It are also measured when run 50 times

  Maximum Accuracy: 100.00%,

  Minimum Accuracy: 91.43%

Exercise 6.6: solution (frequentist inference)

- For a Gaussian random variable, based on the **Central Limit Theorem** (CLT)

$$Z_n = \frac{\bar{X}_n - \mu_X}{\sigma_{\bar{X}}}$$

- Using **the sample mean** $E[X] = \mu_X = \mu_{\bar{X}}, \quad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$

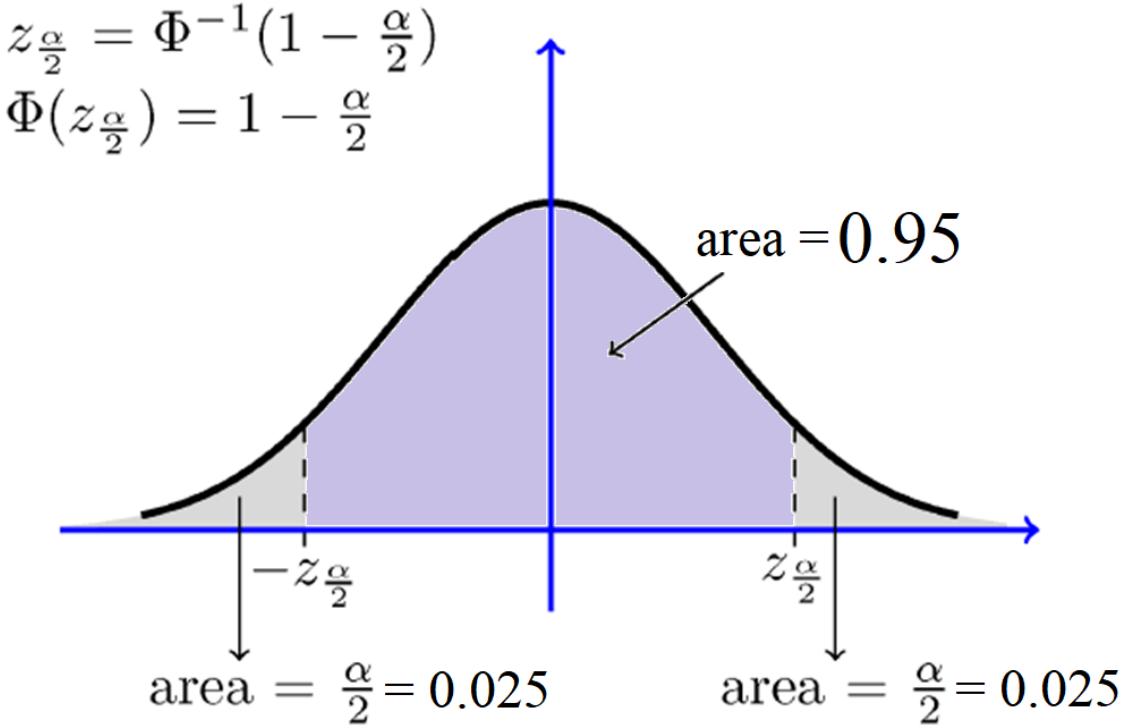
$$Z_n = \frac{\bar{X}_n - \mu_X}{n\sigma_X/\sqrt{n}} = \frac{X_1 - \mu_{X_1} + X_2 - \mu_{X_2} + \cdots X_n - \mu_{X_n}}{\sqrt{n}\sigma_X}$$

$$= \frac{X_1 + X_2 + \cdots X_n - n\mu_X}{\sqrt{n}\sigma_X}$$

Exercise 6.6: solution

- We need to find the values of accuracy $[x_l, x_h]$ that is a 95% **confidence interval** for the estimated parameter θ.
- Observe that 95% = **(1−α) × 100%.** Thus, the **confidence level** is **(1−α) = 0.95** and $\alpha = 0.05$
- It means

$$P(x_l \leq \mu_X \leq x_h) \geq 1 - \alpha = 0.95$$



Exercise 6.6: solution

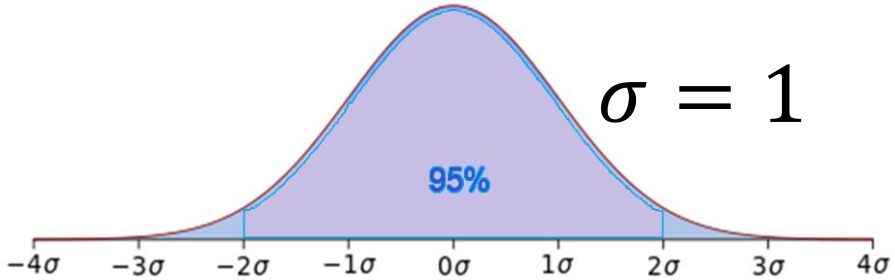
- This is a known result:

$$P\left(\bar{X}_n - SZ_{\frac{\alpha}{2}} \leq \mu_X \leq \bar{X}_n + SZ_{\frac{\alpha}{2}}\right) \geq 1 - \alpha = 0.95$$

- where $Z_{\frac{\alpha}{2}}$ is the upper $Z_{\frac{\alpha}{2}}$ percentage point of the standard normal distribution.

$$Z_n \sim \mathcal{N}(z|\mu, \sigma^2) = \mathcal{N}(z|0,1).$$

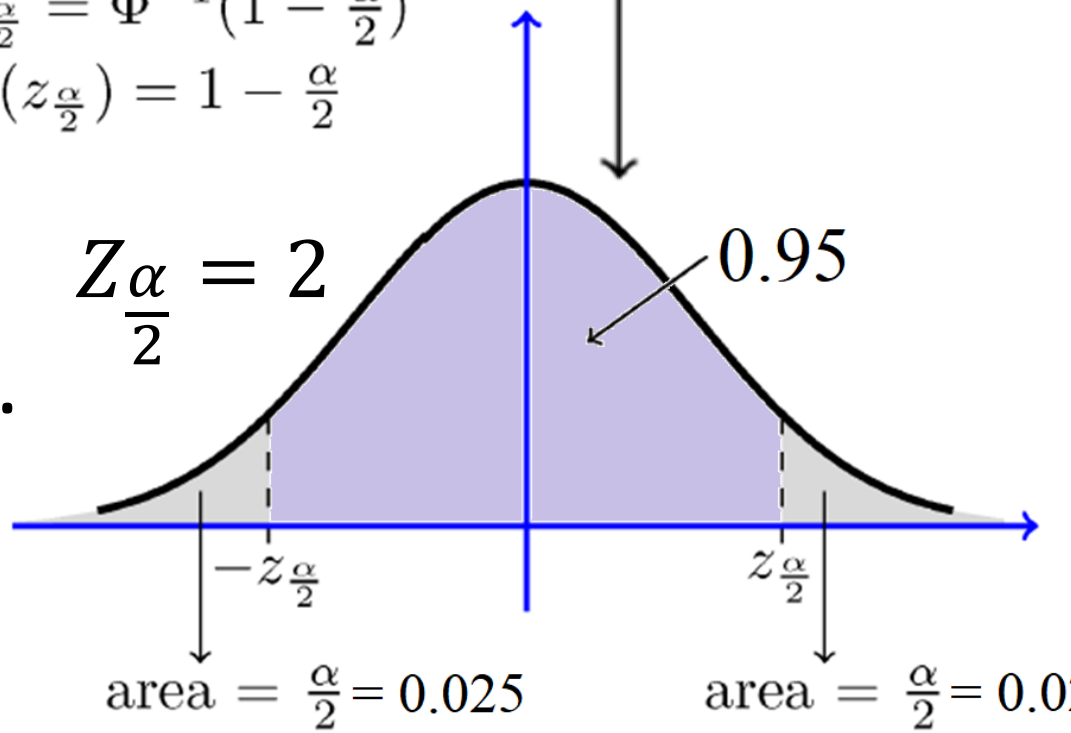
- $\sigma_{\bar{X}} = \bar{S} = 1.72$ and $\bar{X}_n = 96.17$ are already computed



$\sigma = 1$

$$z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \tfrac{\alpha}{2})$$
$$\Phi(z_{\frac{\alpha}{2}}) = 1 - \tfrac{\alpha}{2}$$

$Z_{\frac{\alpha}{2}} = 2$



Exercise 6.6: solution

- Subtituting the recpective values

$$P(\bar{X}_n - 2\sigma_{\bar{X}} \leq \mu_X \leq \bar{X}_n + 2\sigma_{\bar{X}}) \geq 1 - \alpha = 0.95$$

$$P(96.17 - 3.44 \leq \mu_X \leq 96.17 + 3.44) \geq 0.95$$

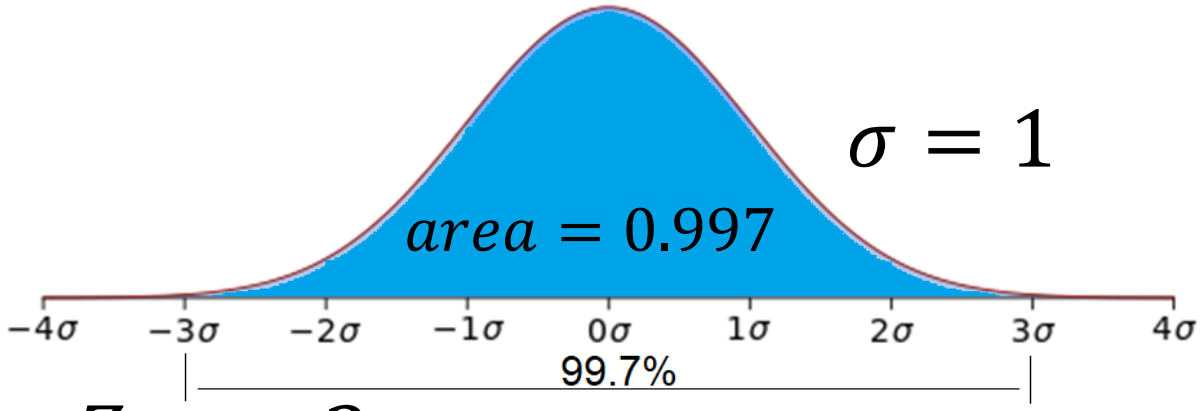
$$P(92.73 \leq \mu_X \leq 99.61) \geq 0.95$$

- The **confidence interval** [92.73%,99.61%] has a **confidence level** greater or equal 0.95 (it means 95% of confidence).
- Observe the computed maximum accuracy: 100.00% and minimum accuracy: 91.43%. The range taken the 50 simulations is [91.43, 100.0]. What happened?

$$Z_n \sim \mathcal{N}(z|\mu,\sigma^2) = \mathcal{N}(z|0,1).$$

Exercise 6.6: solution

- b) What is the **confidence interval** for to have a **confidence level** greater or equal to 99.7%?



$-Z_{\frac{\alpha}{2}} = -3$ $\qquad$ $Z_{\frac{\alpha}{2}} = 3$

$$P(\bar{X}_n - 3\sigma_{\bar{X}} \leq \mu_X \leq \bar{X}_n + 3\sigma_{\bar{X}}) \geq 1 - \alpha = 997$$

$$P(96.17 - 5.16 \leq \mu_X \leq 96.17 + 5.16) \geq 0.997$$

$$P(91.01 \leq \mu_X \leq 101.33) \geq 0.997$$

- The **confidence interval** [91.01%, 101.33%] has a **confidence level** greater or equal to 99.7%.

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

- Suppose that the population of interest has a normal distribution with unknown mean μ and unknown variance σ^2.
- Assume that a random sample of size n, say, $X_1, X_2, \ldots, X_n$, is available, and let $\bar{X}$ and S^2 be the sample mean and variance, respectively.
- We wish to construct a two-sided confidence interval on μ:

  If the variance σ^2 is known, we know that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution.

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

- When σ^2 is unknown, a logical procedure is to replace σ with the sample standard deviation estimations S (frequentist inference).
- The random variable Z now becomes $T = \frac{\bar{x}-\mu}{S/\sqrt{n}}$.
- A logical question is what effect replacing σ with S has on the distribution of the random variable T:
  - If n is large, the answer to this question is "very little," and we can proceed to use the confidence interval based on the normal distribution.
  - However, n is usually small in most engineering problems, and in this situation, a different distribution must be employed to construct the Confidence Interval .

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a Gaussian distribution

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

- has a *Student's t-distribution (or simply the t-distribution)* with $(n - 1)$ degrees of freedom, then $\bar{X}$ and S are estimated by

$$\bar{X} = \frac{1}{n}\sum_{m=1}^{n} x_m \qquad S^2 = \frac{1}{n-1}\sum_{m=1}^{n}(x_m - \bar{X})^2$$

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

- The *Student's t-distribution* is

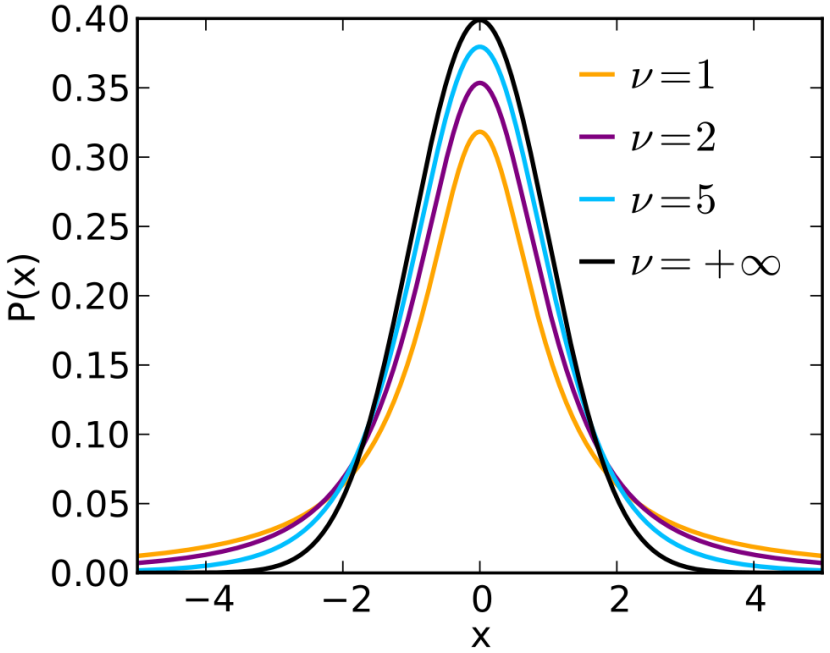
$$f(x) = \frac{\Gamma\left[(k+1)/2\right]}{\sqrt{\pi k}\,\Gamma(k/2)} \cdot \frac{1}{\left[\left(x^2/k\right)+1\right]^{(k+1)/2}} \qquad -\infty < x < \infty$$

- where k is the number of degrees of freedom ($k = n - 1$), and n the number of prototypes.

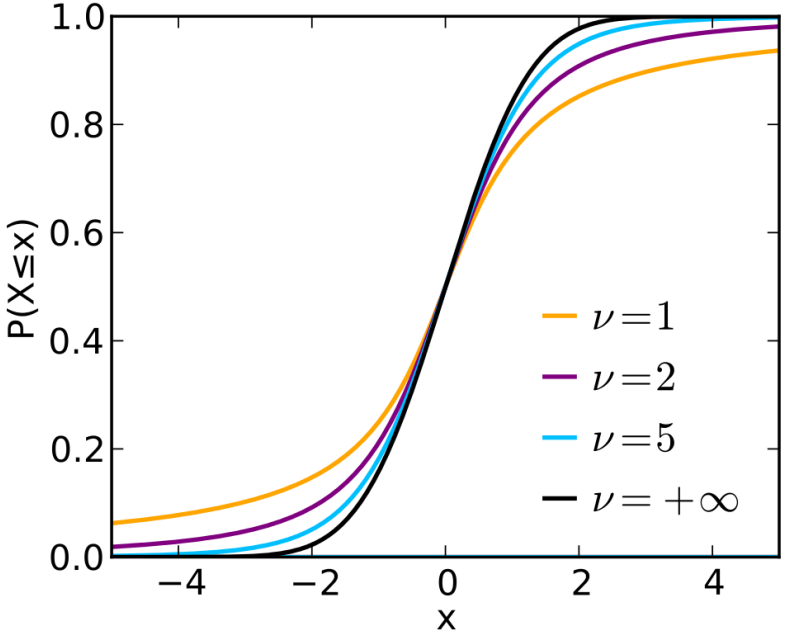
Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

- If we take a sample of n observations from a normal distribution, then the t-distribution with $\upsilon = n - 1$ degrees of freedom can be defined as the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation, after multiplying by the standardizing term $\sqrt{n}$. In this way, the t-distribution can be used to construct a confidence interval for the true mean.

Probability density function



Cumulative distribution function



Statistician William Sealy Gosset, known as "Student"

Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

- where $\upsilon = n - 1$ is the number of degrees of freedom and Γ represent the gamma function

For $\nu > 1$ even,

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots 5\cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4\cdot 2}.$$

For $\nu > 1$ odd,

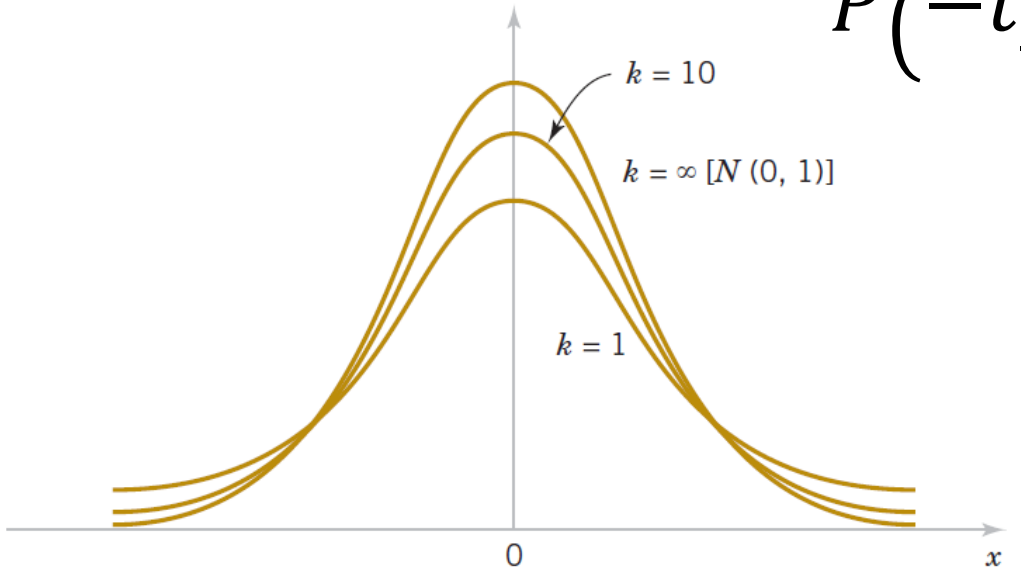
$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots 4\cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5\cdot 3}.$$

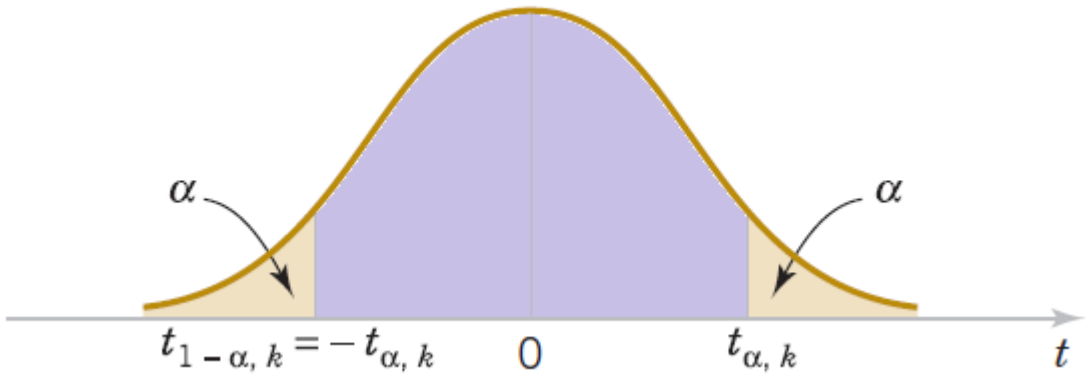
1. $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx;$
2. $\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx = \frac{\Gamma(\alpha)}{\lambda^\alpha}, \qquad \text{for } \lambda > 0;$
3. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha);$
4. $\Gamma(n) = (n-1)!, \text{ for } n = 1, 2, 3, \cdots;$
5. $\Gamma(\frac{1}{2}) = \sqrt{\pi}.$

Student's t-distribution Confidence Interval on μ

- It is easy to find a 100 x (1 – α)% confidence interval on the mean of a normal distribution with unknown variance
- Letting $t_{\frac{\alpha}{2},n-1}$ be the upper $100\frac{\alpha}{2}$ percentage point of the *t*-distribution with $n - 1$ degrees of freedom, we may write

$$P\left(-t_{\frac{\alpha}{2},n-1} \leq T \leq t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha$$





Student's t-distribution Confidence Interval on μ

- The last equation can be rewritten for mean sample as

$$P\left(-t_{\frac{\alpha}{2},n-1} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

$$P\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2},n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

- This leads to the following definition of the 100 x (1 – α)% two-sided confidence interval on μ.

Student's t-distribution Confidence Interval on μ

- If $\bar{X}$ and S are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a **100(1 – α)% confidence interval on μ** is given by

$$\bar{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2},n-1} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2},n-1}$$

$$\bar{X} = \frac{1}{n}\sum_{m=1}^{n} x_m \qquad S^2 = \frac{1}{n-1}\sum_{m=1}^{n}(x_m - \bar{X})^2$$

Exercise 6.7

- A diagnosis of breast cancer is implemented using the **Decision Tree** algorithm .

- The original dataset features file has: 699 instances.
- Percentage of Data Bank used for training: 75.0% - 524 instances (175 instances used for validation).
- Number of classes in the dataset is and the input data space dimension is 9.

Exercise 6.7

- Three random experiments are performed to cross-validate the classifier performance with 10, 20 and 100 simulations.
- For mean sample, the Means and variance in the three experiments are estimated.
- Obtain the confidence interval for the 90%, 95%, 98% and 99.7% confidence levels
- For mean sample of accuracy use Gaussian and T-Studenty probability distribution.

Exercise 6.7: Solution

- Let's consider the following scenario:

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

$$\bar{X} = \frac{1}{n}\sum_{m=1}^{n} x_m \qquad S^2 = \frac{1}{n-1}\sum_{m=1}^{n}(x_m - \bar{X})^2$$

- **100(1 – α)% confidence interval on μ** is: $[\bar{X} - \lambda\bar{S}, \bar{X} + \lambda\bar{S}]$.

Decision Tree Classification Algorithm: 10 simulations

- **Average: 93.77, Standard Deviation: 1.97, Minimum Accuracy : 90.86, Maximum Accuracy: 96.57**
- Confidence Interval for mean sample of accuracy using **Gaussian** probability distribution:
- Confidence level: 90.0%, Confidence Interval: [90.53, 97.01]
- Confidence level: 95.0%, Confidence Interval: [89.91, 97.63]
- Confidence level: 98.0%, Confidence Interval: [89.19, 98.35]
- Confidence level: 99.7%, Confidence Interval: [87.93, 99.62]

- Confidence Interval for mean sample of accuracy using **T-Student** probability distribution:
- Confidence level: 90.0%, Confidence Interval: [90.20, 97.34]
- Confidence level: 95.0%, Confidence Interval: [89.38, 98.16]
- Confidence level: 98.0%, Confidence Interval: [88.33, 99.21]
- Confidence level: 99.7%, Confidence Interval: [86.11, 100.00]

Decision Tree Classification Algorithm: 20 simulations

- **Average: 93.86, Standard Deviation: 1.56, Minimum Accuracy : 90.29, Maximum Accuracy: 96.00**
- Confidence Interval for mean sample of accuracy using **Gaussian** probability distribution:
- Confidence level: 90.0%, Confidence Interval: [90.99, 96.87]
- Confidence level: 95.0%, Confidence Interval: [90.43, 97.44]
- Confidence level: 98.0%, Confidence Interval: [89.77, 98.09]
- Confidence level: 99.7%, Confidence Interval: [88.62, 99.24]

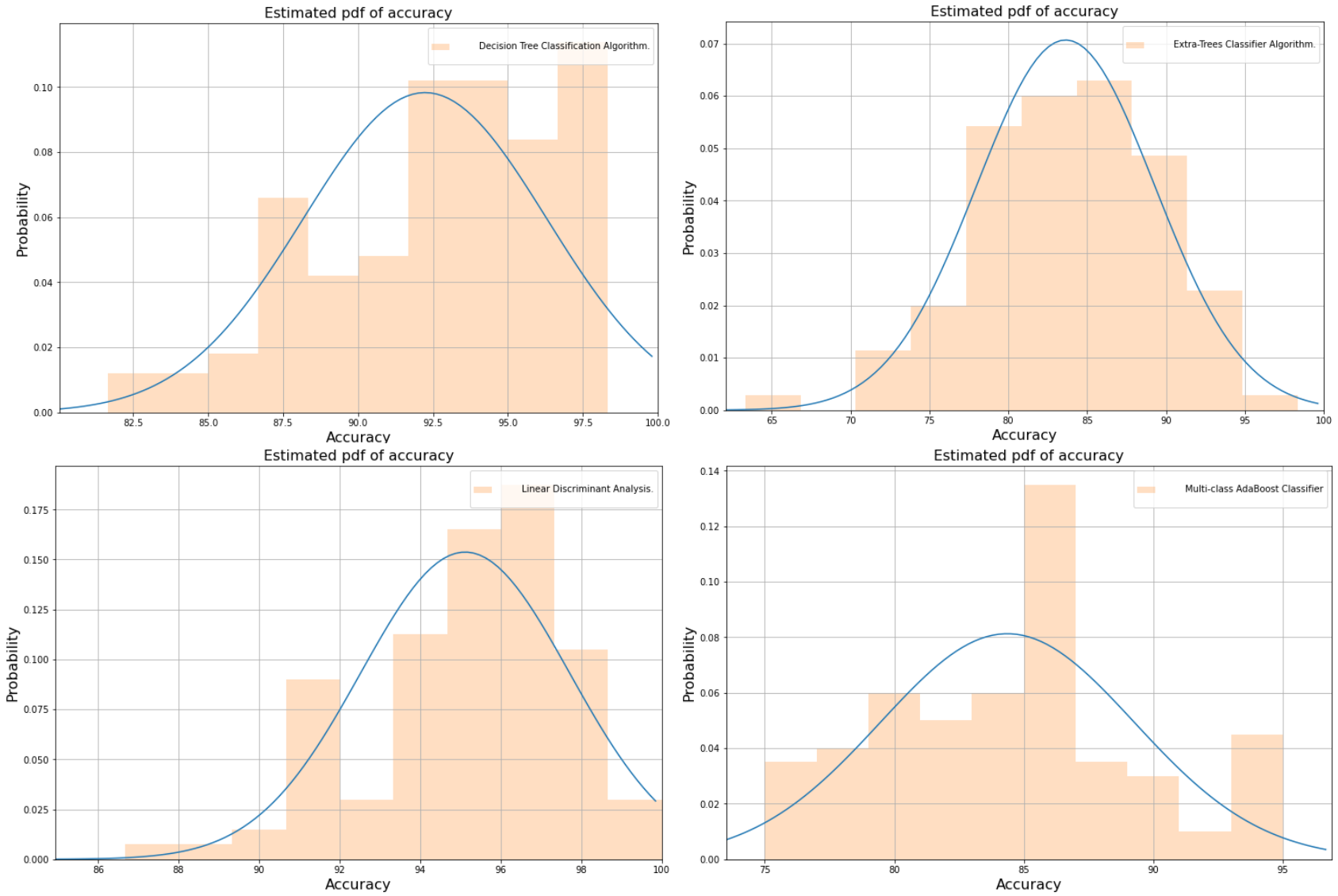
- Confidence Interval for mean sample of accuracy using **T-Student** probability distribution:
- Confidence level: 90.0%, Confidence Interval: [90.96, 96.90]
- Confidence level: 95.0%, Confidence Interval: [90.38, 97.48]
- Confidence level: 98.0%, Confidence Interval: [90.70, 98.16]
- Confidence level: 99.7%, Confidence Interval: [88.49, 99.37]

Decision Tree Classification Algorithm: 100 simulations

- **Average: 94.13, Standard Deviation: 1.64, Minimum Accuracy : 90.29, Maximum Accuracy: 97.71**
- Confidence Interval for mean sample of accuracy using **Gaussian** probability distribution:
- Confidence level: 90.0%, Confidence Interval: [91.43, 96.82]
- Confidence level: 95.0%, Confidence Interval: [90.91, 97.34]
- Confidence level: 98.0%, Confidence Interval: [90.31, 97.94]
- Confidence level: 99.7%, Confidence Interval: [89.26, 98.99]

- Confidence Interval for mean sample of accuracy using **T-Student** probability distribution:
- Confidence level: 90.0%, Confidence Interval: [91.41, 96.85]
- Confidence level: 95.0%, Confidence Interval: [90.87, 97.38]
- Confidence level: 98.0%, Confidence Interval: [90.25, 98.00]
- Confidence level: 99.7%, Confidence Interval: [89.14, 99.11]

Gaussian pdf approaching



Statistical Hypothesis

GPDS
Grupo de Processamento Digital de Sinais

Prof. F. Assis

Evaluating a hypothesis

- We will classify hypotheses into two types: statistical and non-statistical.

A hypothesis

Statistical hypothesis

Non-statistical hypothesis

Evaluating a hypothesis

- A statistical hypothesis has something to say about the behavior of one or more random variables.
- If the hypothesis is well-defined but no random phenomenon is involved, then we call such hypotheses as non-statistical hypotheses.
- Many of the physical laws or mathematical conjectures are non-statistical hypotheses.

Evaluating a hypothesis

- If a hypothesis is about the parameter(s) of well-defined distributions, then we call them parametric hypotheses.
- All other statistical hypotheses will be called non-parametric hypotheses.

A statistical hypothesis → Parametric hypothesis

A statistical hypothesis → Non-parametric hypothesis

Evaluating a hypothesis

- We are interested in testing an assertion about a population based on a random sample $\boldsymbol{X_n}$ (frequentist inference).

- This assertion is stated in the form of a hypothesis about the underlying distribution of $\boldsymbol{X}$, and the objective of the test is to accept or reject the hypothesis based on the observed data $\boldsymbol{X_n}$.

Evaluating a hypothesis: What do we want? Something as:

- Is the dice biased?
- Does the product build by an industry meets its technical specifications?
- Is the probabilistic model for describing a noise that is present in a stochastic process correct?
- Does a new manufacturing process produce a “new and improved” batteries that last longer?

Statistical Hypothesis

In a statistical hypothesis, a methodology to assess a statistical model based a random experiment sample.

If possible, hypotheses are formulated to be tested as a single value for a parameter.

Statistical Hypothesis

- Many problems in engineering require that we decide which of two competing claims or statements about some parameter is true.

- The statements are **called hypotheses**, and the decision-making procedure is called **hypothesis testing**.

- This is one of the most useful aspects of **statistical inference**, because many types of decision-making problems, tests, or experiments in the engineering world can be formulated as hypothesis-testing problems.

Statistical Hypothesis

- A **statistical hypothesis** is a statement about the parameters of one or more populations.

- Because we use **probability distributions** to represent **populations**, a **statistical hypothesis** may also be thought of as a statement about the probability distribution of a random variable.

Statistical Hypothesis

Significance testing is a process through which a formally stated hypothesis about the population parameter can be accepted or rejected.

Through significance testing, the credibility (plausibility) of a specific statement can thus be assessed.

Statistical Hypothesis

Unfortunately, in general, it **is not possible to make correct decisions at all times**, as **the available data may be limited** and/or **distorted**, i.e. the observed data may not be **sufficient and/or accurate**.

"The data sample is not telling the true story about the process under study"

In a **statistical hypothesis**, the set of **possible outcomes** is divided into a set of values for which we **accept the hypothesis** and another set of values for which we **reject it**.

Statistical Hypothesis

There are two hypotheses required for **hypothesis testing**: $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$.

The ***null hypothesis***, denoted by $\boldsymbol{H_0}$, is a statement that generally specifies a particular value for the parameter being studied.

The ***alternative hypothesis***, denoted by $\boldsymbol{H_1}$, which is a statement that specifies either an alternative value or a range of alternative values.

Statistical Hypothesis

Sample observations are collected to determine if the data support $\boldsymbol{H_0}$ or $\boldsymbol{H_1}$, not to determine whether $\boldsymbol{H_0}$ or $\boldsymbol{H_1}$ is true.

The decision between the two hypotheses is made on the basis of the value of a statistic called the ***test statistic***, which is a function of the observations in a **random sample**.

Statistical Hypothesis

- If the results observed in a random sample are considerably different from the results expected under the hypothesis, then the observed differences are **significant**.

The following sequence of steps in applying hypothesis-testing methodology is recommended.

- **1. Parameter of interest:** From the problem context, identify the parameter of interest.
- **2. Null hypothesis, H_0:** State the null hypothesis, H_0.
- **3. Alternative hypothesis, H_1:** Specify an appropriate alternative hypothesis, H_1.
- **4. Test statistic:** Determine an appropriate test statistic.
- **5. Reject H_0 if:** State the rejection criteria for the null hypothesis.
- **6. Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
- **7. Draw conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context.

Statistical Hypothesis: Example

- For example, consider the air crew escape system. Suppose that we are interested in the **burning rate** of **the solid propellant**.
- **Burning rate** is **a random variable** that can be described by **a probability distribution**.
- Suppose that our interest focuses on **the mean burning rate** (a parameter of this distribution).
- Specifically, we are interested in deciding whether or not the mean burning rate is 50 centimeters per second. We may express this formally as

$$H_0: \mu = 50 \text{ centimeters per second} \quad H_1: \mu \neq 50 \text{ centimeter per seconds}$$

Statistical Hypothesis: Example

- **The statement H0**: $\mu = 50$ centimeters per second is called the **null hypothesis**. This is a claim that is initially assumed to be true.

- **The statement H1**: $\mu \neq 50$ centimeters per second is called **the alternative hypothesis** and it is a statement that **condradicts the null hypothesis**.

Statistical Hypothesis: Example

- Because the alternative hypothesis specifies values of μ that could be either greater or less than 50 centimeters per second, it is called a two-sided alternative hypothesis.
- In some situations, we may wish to formulate a one-sided alternative hypothesis, as in

$$H_0: \mu = 50 \text{ centimeters per second} \quad H_0: \mu = 50 \text{ centimeters per second}$$

or

$$H_1: \mu < 50 \text{ centimeter per seconds} \quad H_1: \mu > 50 \text{ centimeters per second}$$

Statistical Hypothesis: Example

- We will always state **the null hypothesis** as an **equality claim**.
- However, when **the alternative hypothesis** is stated with the $<$,
  - the implicit claim in **the null hypothesis** can be taken as $\geq$ and
  - when **the alternative hyphothesis** is stated with the $>$, the implicit claim in **the null hypothesis** can be taken as $\leq$.
- It is important to remember that hypotheses are always statements about **the population or distribution under study**, not statements about **the sample**.

Statistical Hypothesis: Example

- A procedure leading to a decision about the **null hypothesis** is called a **test of a hypothesis**.
- **Hypothesis-testing procedures** rely on using the information in a **random sample** from **the population of interest**.
- **Testing the hypothesis** involves taking **a random sample**, computing a **test statistic** from **the sample data**, and then using the **test statistic** to make a decision about the **null hypothesis**.
- If this information **is consistent** with the **null hypothesis**, we will **not reject** it.
- However, if this information **is inconsistent** with the **null hypothesis**, we will conclude that **the null hypothesis is false** and reject it in favor of **the alternative hypothesis**.

Statistical Hypothesis

In order to specify the decision rule, the observation space, which is an ***n*-dimensional Euclidean space**, is partitioned into two disjoint regions ***A*** *(Acceptance Region)* and ***R*** *(Rejection Region)*.

If we have $\boldsymbol{x} \in \boldsymbol{A}$, we will then decide on $\boldsymbol{H_0}$.The region ***A*** is thus known as the ***acceptance region*** or ***region of nonsignificance***.

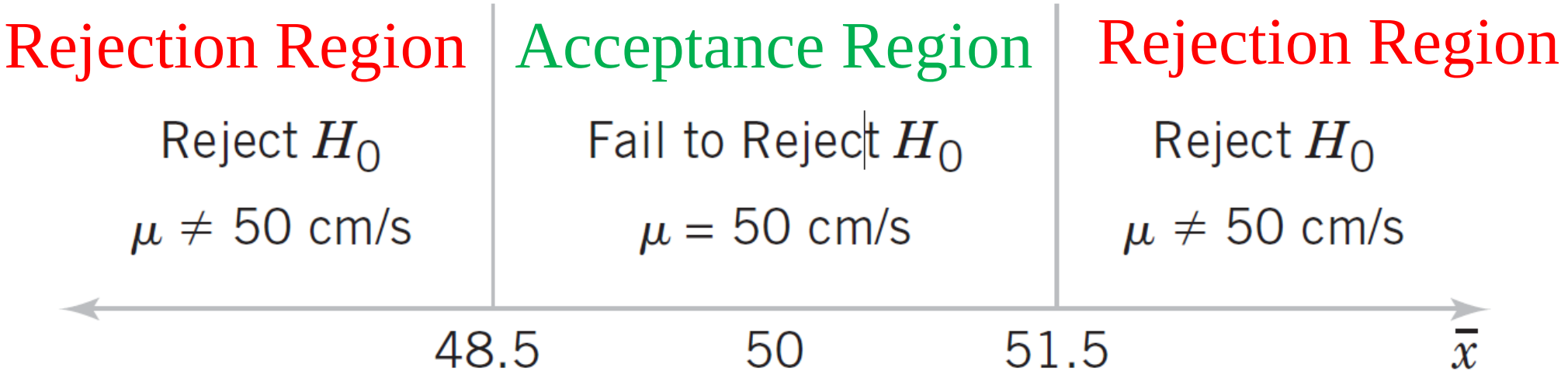
If we have $\boldsymbol{x} \in \boldsymbol{R}$, we will then decide on $\boldsymbol{H_1}$.The region ***R*** is thus known as the ***rejection region*** or ***critical region*** or ***region of significance***.

Statistical Hypothesis: Example

- Suppose that a sample of n = 10 specimens is tested and that the sample mean burning rate x is observed.
- The sample mean $\bar{x}$ is an estimate of the true population mean μ.
- A value of the sample mean $\bar{x}$ that falls close to the hypothesized value of μ = 50 centimeters per second does not conflict with the null hypothesis that the true mean μ is really 50 centimeters per second.
- On the other hand, a sample mean that is considerably different from 50 centimeters per second is evidence in support of the alternative hypothesis H_1. Thus, the sample mean is the test statistic in this case.

Statistical Hypothesis: Example

- The **sample mean** (n = 10) can take on many different values. Suppose that if $48.5 \leq \bar{x} \leq 51.5$, we will **not reject the null hypothesis** H_0: $\mu = 50$, and if either $\bar{x} < 48.5$ or $\bar{x} > 51.5$, we will **reject the null hypothesis** in favor of the alternative hypothesis H_1: $\mu \neq 50$.



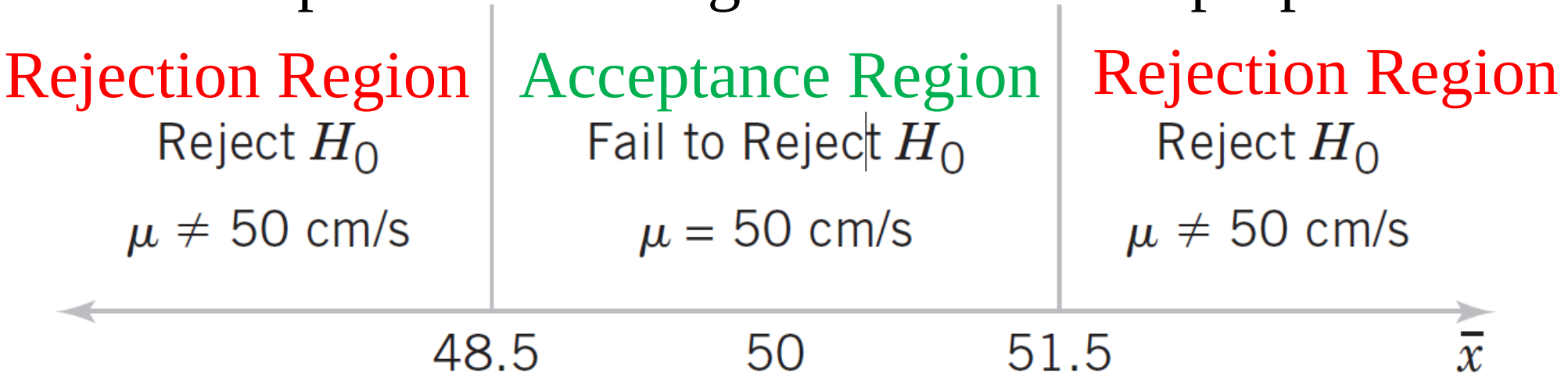
Statistical Hypothesis

There are two types of errors, known as ***Type I error*** and ***Type II error*** that can occur when the decision rule is followed:

Type I error: Reject H_0 when H_0 is true (false alarm or **false positive**)

Type II error: Accept H_0 when H_0 is false (miss or **false negative**)

Example: The burning rate of the solid propellant



the critical region boundaries: 48.8 and 51.5

Statistical Hypothesis

- Sometimes **the type I error probability** is called the **significance level**, the **α-error**, or **the size of the test**.

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

(false positive)

- The **probability of a type II error**, which we will denote by β.

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

(false negative)

Statistical Hypothesis: Example

- In the propellant burning rate example, **a type I error** will occur when either $\bar{x} > 51.5$ or $\bar{x} < 48.5$ when the true mean burning rate really is $\mu = 50$ centimeters per second.
- Suppose that the standard deviation of burning rate is $\sigma_X = 2.5$ centimeters per second and that the burning rate has a distribution for which the conditions of **the central limit theorem** apply, so the distribution of the sample mean is approximately normal with mean $\mu = 50$ and standard deviation $\frac{\sigma_X}{\sqrt{n}} = \frac{2.5}{\sqrt{10}} = 0.79$ (frequentist inference).

By the central limit theorem: $Z_n = \dfrac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}}$

Statistical Hypothesis: Example

- The probability of making a **type I error** (or the significance level of our test) is equal to the sum of the areas that have been shaded **in red color** (in the tails of the normal distribution). We may find this probability as

$$\alpha = P(\bar{X} < 48.5 \text{ when } \mu = 50) + P(\bar{X} > 51.5 \text{ when } \mu = 50)$$

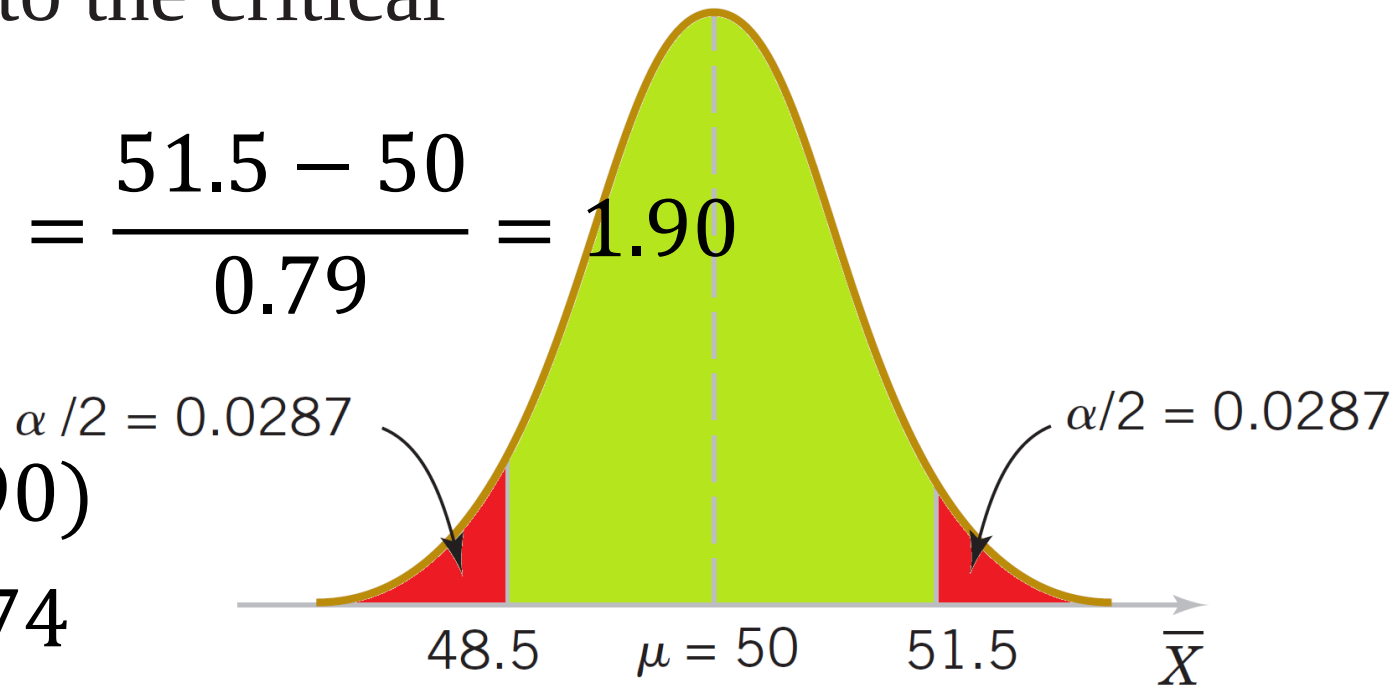
- The z-values that correspond to the critical values 48.5 and 51.5 are

$$z_1 = \frac{48.5 - 50}{0.79} = -1.9 \qquad z_2 = \frac{51.5 - 50}{0.79} = 1.90$$

- Therefore,

$$\alpha = P(z < -1.9) + P(z > 1.90)$$

$$= 0.0287 + 0.0287 = 0.0574$$



Statistical Hypothesis: Example

- This implies that 5.74% of all random samples would lead to rejection of the hypothesis H_0: $\mu = 50$ centimeters per second when the true mean burning rate is really 50 centimeters per second.
- We could also reduce **α** by increasing the sample size. If $n = 16$,

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{2.5}{\sqrt{16}} = 0.625$$

$$\alpha = P\left(z < \frac{48.5 - 50}{0.625}\right) + P(z > \frac{51.5 - 50}{0.625})$$

$$= P(z < -2.40) + P(z > 2.40) = 0.0164$$

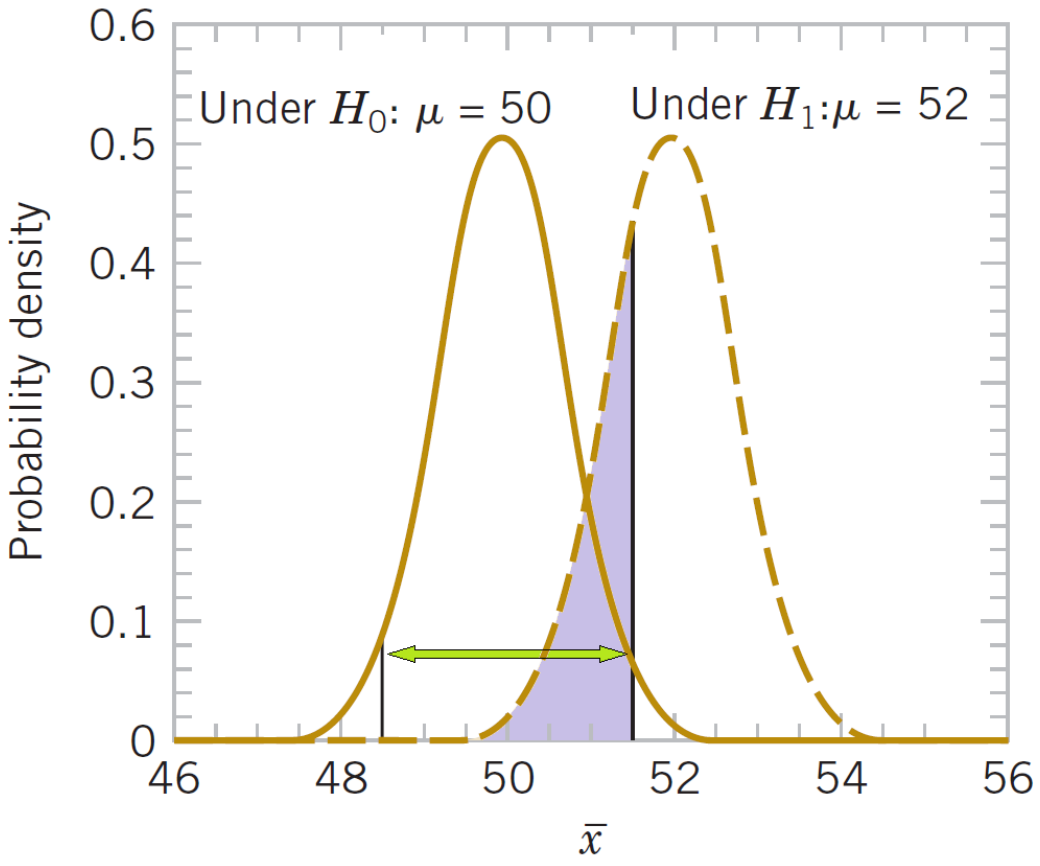
The probability of a **type I error (false positive)** is 1.64%

Statistical Hypothesis: Example

- To calculate β (sometimes called the β-error), we must have a specific **alternative hypothesis**; that is, we must have a particular value of μ.
- For example, suppose that it is important to reject the **null hypothesis** H_0: $\mu = 50$ whenever the mean burning rate μ is greater than 52 centimeters per second or less than 48 centimeters per second.
- We could calculate **the probability of a type II error β** (**false negative**) for the values $\mu = 52$ and $\mu = 48$ and use this result to tell us something about how the test procedure would perform.
- Specifically, how will the test procedure work if we wish to detect, that is, reject H_0, for a mean value of $\mu = 52$ or $\mu = 48$?
- Because of **symmetry**, it is necessary to evaluate only one of the two cases—say, find **the probability of accepting the null hypothesis** H_0: $\mu = 50$ centimeters per second when the **true mean** is $\mu = 52$ centimeters per second.

Statistical Hypothesis: Example

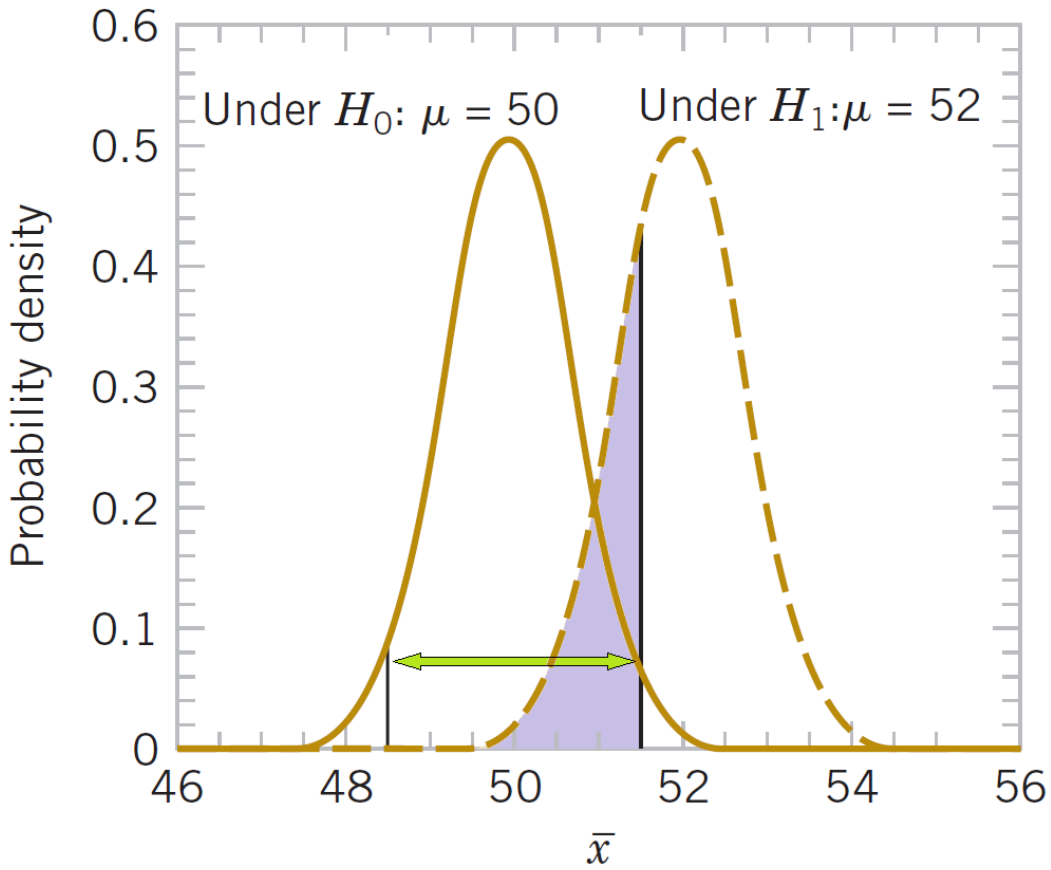
- The figure will help us calculate the probability of **type II error β (false negative)**.
- The normal distribution on the figure on the left is the distribution of the test statistic $\bar{X}$ when the null hypothesis H_0: μ = 50 is true (this is what is meant by the expression "under H_0: μ = 50"), and the normal distribution on the right is the distribution of $\bar{X}$ when the alternative hypothesis is true, and the value of the mean is 52 (or "under H_1: μ = 52").



Statistical Hypothesis: Example

- A **type II error** (fail to reject H_0 when H_0 is false - **false negative**) will be committed if the sample mean $\bar{X}$ falls between 48.5 and 51.5 (the critical region boundaries) when $\mu = 52$.

- As seen in the figure on the right, this is just the probability that $48.5 \leq \bar{X} \leq 51.5$ when the true mean is $\mu = 52$, or the shadec area under the normal distribution centered at $\mu = 52$.



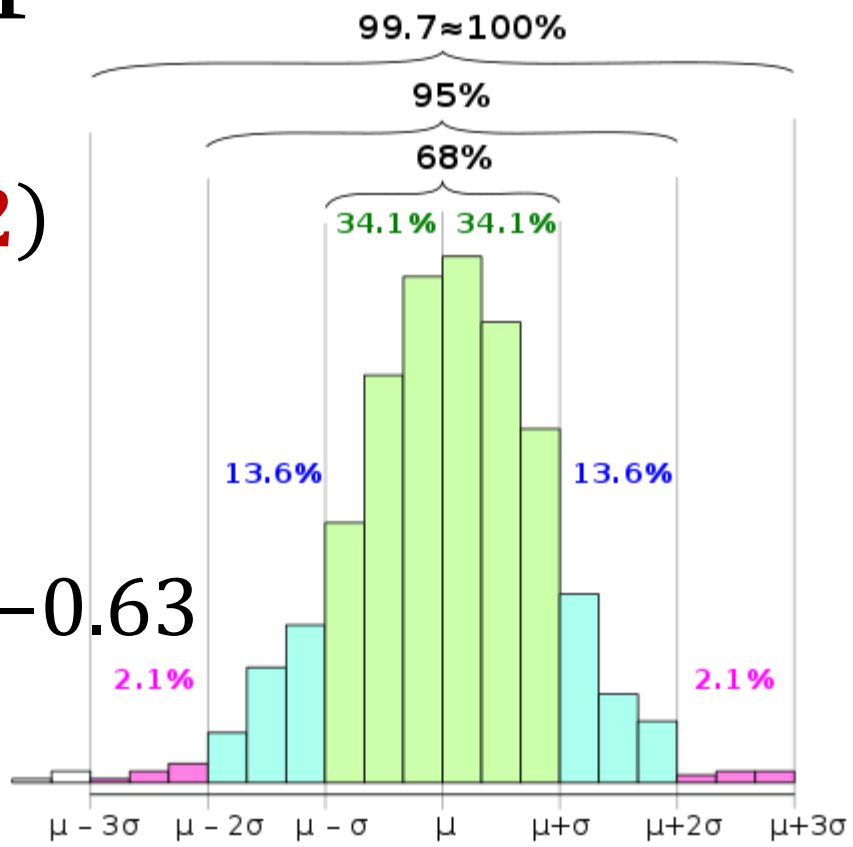
Statistical Hypothesis: Example

- β can be computed by

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ when } \boldsymbol{\mu = 52})$$

- The z-values corresponding to 48.5 and 51.5 when the true μ = 52 are

$$z_1 = \frac{48.5 - 52}{0.79} = -4.43 \; z_2 = \frac{51.5 - 52}{0.79} = -0.63$$



- Terefore,

$$\beta = P(-4.43 \leq Z \leq -0.63) = \mathrm{P}(Z \leq -0.63) - \mathrm{P}(Z \leq -4.43)$$

$$= 0.2643 - 0.000 = 0.2643$$

the probability of a type II error β (**false negative**) = 26.43%

Statistical Hypothesis: Example

- Thus, if we are testing H_0: $\mu = 50$ against H1: $\mu \neq 50$ with $n = 10$ and the true value of the mean is $\mu = 52$, the probability that we will **fail to reject** the false null hypothesis is 0.2643.
- By symmetry, if the true value of the mean is $\mu = 48$, the value of β will also be 0.2643.
- **The probability of making a type II error β (false negative) increases rapidly as the true value of μ approaches the hypothesized value**. For example, if the true value of the mean is $\mu = 50.5$ and the hypothesized value is H_0: $\mu = 50$. The true value of μ is very close to 50, and the value for β is

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ when } \boldsymbol{\mu = 50.5})$$

$$z_1 = \frac{48.5 - 50.5}{0.79} = -2.53 \qquad z_2 = \frac{51.5 - 50.5}{0.79} = -1.27$$

Statistical Hypothesis: Example

- Therefore,

$$\beta = P(-2.53 \leq Z \leq Z1.27) = \mathrm{P}(Z \leq 1.27) - \mathrm{P}(Z \leq -2.53)$$

$$= 0.8980 - 0.0057 = 0.8923$$

- Thus, the **type II error probability** (**false negative**) is much higher for the case in which the true mean is 50.5 centimeters per second than for the case in which the mean is 52 centimeters per second.

Statistical Hypothesis

- Suppose we have a random sample

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$$

of a random variable X whose pdf $f_{\boldsymbol{X}}(\boldsymbol{x}, \theta)$ depends on a population parameter θ, which is unknown but not random, and the observed data represented by the vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

- We want to test a hypothesis about a parameter θ of the random variable X.
- The objective of a significance test is to accept or reject the null hypothesis H_0.

Exercício 6.8

- You have a coin, and you would like to check whether it is fair or not.
- In the random experiment the coin is tossed 100 times.
- More specifically, let θ be **the probability of heads**, $\theta = P(H)$. You have two hypotheses:
    - H_0 (the null hypothesis): The coin is fair, i.e. $\theta = \theta_0 = ½$.
    - H_1 (the alternative hypothesis): The coin is not fair, i.e., $\theta \neq ½$.

Exercício 6.8: solution

- We need to design a test to either accept H_0 or H_1.
- To check whether the coin is fair or not, we perform the following experiment.
- In the random experiment, we toss the coin 100 times and record the number of heads.
- Let $\boldsymbol{X}$ be the number of heads that we observe, so

$$X \sim Binomial(100, \theta).$$

Exercício 6.8: solution

- The problem Contextualization - Unordered Sampling (Combinations) without Replacement **(**ordering does not matter and repetition is not allowed) choose ***k*** elements from a set with a total of ***n*** elements.
- ***n* independent Bernoulli** trials: A random variable corresponding to a binomial distribution

$$P(\boldsymbol{k}\ times \text{ is } true\ and\ (\boldsymbol{n}-\boldsymbol{k})\ times \text{ is } false)$$

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \text{ for } 0 \leq k \leq n.$$

Exercício 6.8: solution

- Now, if H_0 is true, then $\theta = \theta_0 = ½$, so **we expect the number of heads to be close to 50**.
- Thus, intuitively we can say that if we observe close to 50 heads we should accept H_0, otherwise we should reject it.

Exercício 6.8: solution

- We suggest the following criteria: If $|X - 50|$ is less than or equal to some threshold, we accept H_0.
- On the other hand, if $|X - 50|$ is larger than the threshold we reject H_0 and accept H_1.

Exercício 6.8: solution

- Let's call that threshold γ.

$$\text{If } |X - 50| \leq \gamma, \text{ accept } H_0.$$

$$\text{If } |X - 50| > \gamma, \text{ accept } H_1.$$

- **To choose γ properly it is necessary to state some requirements for our test**, an important factor here is probability of error.

Exercício 6.8: solution

- One way to make an error is when we reject H_0 while in fact it is true.
- We call this *type I error*. More specifically, this is the event that $|X - 50| > \gamma$ when H_0 is true.

$$P(\text{type I error}) = P(|X - 50| > \gamma \mid H_0).$$

- We read this as "*the probability that* $|X - 50| > \gamma$ *when* H_0 *is true*".

Exercício 6.8: solution

- To be able to decide what γ needs to be, we can choose a desired value for P(type I error). For example, **we might want to have a test for which**

$$P(\text{type I error}) \leq \alpha = 0.05$$

- α is called *the level of significance*. To satisfy the desired level of significance we can choose

$$P(|X - 50| > \gamma \mid H_0) = \alpha = 0.05$$

Exercício 6.8: solution

- Since we know the distribution of X under H_0, i.e., $X|H_0 \sim Binomial\ (100, \theta = ½)$, we should be able to choose γ such that

$$P(|X - 50| > \gamma \mid H_0) = \alpha = 0.05$$

- Using the central limit theorem (CLT), we can create a new random variable **Y** by combining the linear variable X.
- For large values of n, we can approximate a *Binomial* (n, θ) distribution by a normal distribution, $\mathcal{N}(0, 1)$.

Exercício 6.8: solution

- We can say that for large values of n, if $\boldsymbol{X} \sim Binomial(n, \theta_0 = ½)$, then (using **CLT**)

$$Y = \frac{X - \mu_X}{\sigma_X/\sqrt{n}} = \frac{\bar{X} - n\mu_X}{\sqrt{n\sigma_X^2}} = \frac{\bar{X} - n\theta_0}{\sqrt{n\,\theta_0(1-\theta_0)}}$$

- For $n = 100$ and $\theta_0 = ½$, we have

$$\mathrm{Y} = \frac{X - 100\,\frac{1}{2}}{\sqrt{100\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{X - 50}{5}$$

- $\boldsymbol{Y}$ is a Gaussian variable $\mathcal{N}(0,1)$.

Exercício 6.8: solution

- Thus, to be able to use the CLT, instead of looking at X directly, we can look at Y. Note that

$$P(type\ I\ error) = P(|X - 50| > \gamma | H_0)$$
$$= P\left(\left|\frac{X-50}{5}\right| > \frac{\gamma}{5}\middle| H_0\right) = P\left(|Y| > \frac{\gamma}{5}\middle| H_0\right)$$

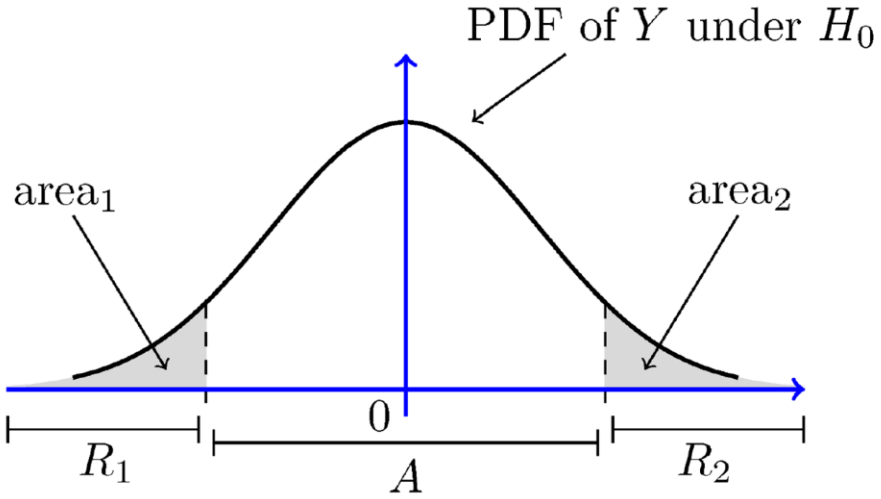
- If $|Y| \leq \frac{\gamma}{5}$, accept H_0.
- If $|Y| > \frac{\gamma}{5}$, accept H_1.

Exercício 6.8: solution

- We have the following condition

$$\alpha = P\left(|Y| > \frac{\gamma}{5}\right) = 1 - P\left(-\frac{\gamma}{5} \leq Y \leq \frac{\gamma}{5}\right)$$
$$= 1 - \left\{\Phi\left(\frac{\gamma}{5}\right) - \Phi\left(-\frac{\gamma}{5}\right)\right\} = 1 - \left\{\Phi\left(\frac{\gamma}{5}\right) - \left[1 - \Phi\left(\frac{\gamma}{5}\right)\right]\right\}$$
$$= 2 - 2\Phi\left(\frac{\gamma}{5}\right)$$

- Using $\Phi\left(\frac{\gamma}{5}\right) = 1 - \Phi\left(-\frac{\gamma}{5}\right)$



Exercício 6.8: solution

- We obtain

$$\alpha = 2 - 2\Phi\left(\frac{\gamma}{5}\right) = 0.05$$

$$\Phi\left(\frac{\gamma}{5}\right) = \frac{1.95}{2} \Rightarrow \frac{\gamma}{5} = \Phi^{-1}(0.975) = 1.96$$

Exercício 6.8: solution

- Thus, we conclude

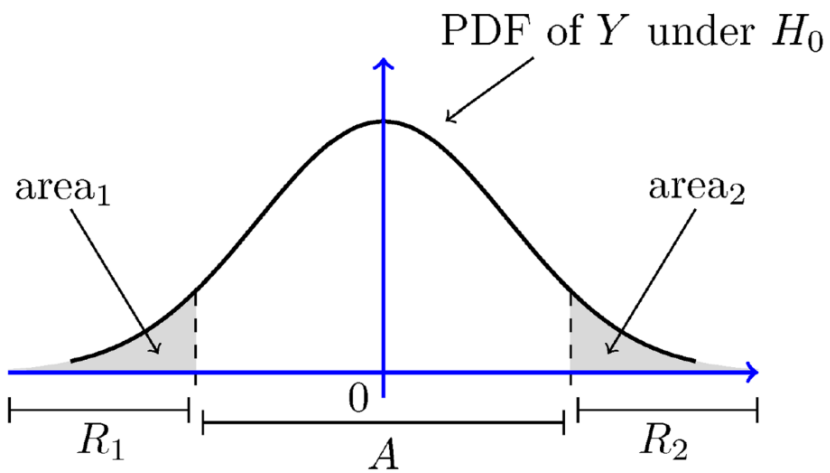
$$\text{If } |Y| \leq \frac{\gamma}{5} = 1.96, \text{ accept } H_0.$$

$$\text{If } |Y| > \frac{\gamma}{5} = 1.96, \text{ accept } H_1.$$

$$\gamma > 9.8 \approx 10 \;\Rightarrow |X - 50| \;\geq\; 10$$

- *Accept*: $\theta = P(H) = \frac{1}{2}$ if the number of Heads (H) obtained in the experiment is between: $40 \leq X \leq 60$

Exercício 6.8: solution



$$R = (-\infty, -1.96) \cup (1.96, \infty)$$

$$P(type\ I\ error)$$
$$area_1 + area_2$$

The boundaries that mark the beginning of the critical region are called ***critical values***.

Accept H_0 if $x \in A$ ***(Acceptance Region)***

Reject H_0 if $x \in R = R_1 U\ R_2$ ***(Rejection Region)***

- Note that not rejecting the hypothesis is not the same as accepting the hypothesis.
- When we do not reject the hypothesis, we are saying the hypothesis is plausible, but it is essential to note that there can be other equally plausible hypotheses.

Statistical Hypothesis

The probabilities of Type I and Type II errors can also be written as

$$P(\boldsymbol{x} \in R \mid H_0) = \alpha$$

$$P(\boldsymbol{x} \in A \mid H_1) = \beta$$

where α is called the ***level of significance*** or ***size of the rejection region***, and $(1-\beta)$ is known as the ***power of the test***.

Statistical Hypothesis

		Decision based on the data	
		Reject H_0	**Do not reject H_0**
Truth (unknown)	H_0 is true	Type I error (false rejection) Probability $= \alpha$	Correct decision Probability $= 1 - \alpha$
	H_0 is false	Correct decision Probability $= 1 - \beta$	Type II error (false acceptance) Probability $= \beta$

Statistical Hypothesis - **Normality Tests**

- Tests whether a data sample has a Gaussian distribution.
- **Hypothesis**
  - H0: the sample has a Gaussian distribution.
  - H1: the sample does not have a Gaussian distribution.
- **Assumptions**: Observations in each sample are independent and identically distributed (iid).
- **Tests: Shapiro-Wilk Test**, **D'Agostino's K^2 Test**, **Anderson-Darling Test.**

Statistical Hypothesis - **Correlation Tests**

- Tests whether two samples have a linear relationship (collinearity).
- **Tests: Pearson's Correlation Coefficien**t - Tests whether two samples have a linear relationship.
- **Hypothesis**
  - H0: the two samples are independent.
  - H1: there is a dependency between the samples.
- **Assumptions**:
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample are normally distributed.
  - Observations in each sample have the same variance.

Statistical Hypothesis - **Correlation Tests**

- Tests whether two samples have a linear relationship (collinearity).
- **Spearman's Rank Correlation, Kendall's Rank Correlation -** Tests whether two samples have a monotonic relationship;
- **Hypothesis**
  - H0: the two samples are independent.
  - H1: there is a dependency between the samples.
- **Assumptions**:
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample can be ranked.

Statistical Hypothesis - **Correlation Tests**

- Tests whether two samples have a linear relationship (collinearity).
- **Chi-Squared Test -** Tests whether two categorical variables are related or independent.
- **Hypothesis**
  - H0: the two samples are independent.
  - H1: there is a dependency between the samples.
- **Assumptions**:
  - Observations used in the calculation of the contingency table are independent.
  - 25 or more examples in each cell of the contingency table.

Statistical Hypothesis - **Stationary Tests**

- Check if a time series (a signal) is stationary or not.
- **Tests**: **Augmented Dickey-Fuller** - Tests whether a time series has a unit root, e.g. has a trend or more generally is autoregressive;
- **Hypothesis**
  - H0: a unit root is present (series is non-stationary).
  - H1: a unit root is not present (series is stationary).
- **Assumptions**: Observations in are temporally ordered.

Statistical Hypothesis - **Stationary Tests**

- Check if a time series (a signal) is stationary or not.
- **Tests**: **Kwiatkowski-Phillips-Schmidt-Shin -** Tests whether a time series is trend stationary or not.
- **Hypothesis**
  - H0: the time series is trend-stationary.
  - H1: the time series is not trend-stationary.
- **Assumptions**: Observations in are temporally ordered.

Statistical Hypothesis - **Parametric Statistical Hypothesis Tests**

- Statistical tests that can be used to compare data samples. Tests whether the means of two independent samples are significantly different.

- **Student's t-test** -Tests whether the means of two independent samples are significantly different; **Analysis of Variance Test (ANOVA)** - Tests whether the means of two or more independent samples are significantly different.

- **Interpretation**
  - H0: the means of the samples are equal.
  - H1: the means of the samples are unequal.
- **Assumptions**:
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample are normally distributed.
  - Observations in each sample have the same variance.

Statistical Hypothesis - **Parametric Statistical Hypothesis Tests**

- Statistical tests that can be used to compare data samples. Tests whether the means of two independent samples are significantly different.

- **Paired Student's t-test** - Tests whether the means of two paired samples are significantly different; **Repeated Measures ANOVA Test** - Tests whether the means of two or more paired samples are significantly different.

- **Interpretation**
  - H0: the means of the samples are equal.
  - H1: the means of the samples are unequal.
- **Assumptions**:
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample are normally distributed.
  - Observations in each sample have the same variance.
  - Observations across each sample are paired.

Statistical Hypothesis - **Nonparametric Statistical Hypothesis Tests**

- **Mann-Whitney U Test, Kruskal-Wallis H Test**: Tests whether the distributions of two independent samples are equal or not.
- **Assumptions**
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample can be ranked.
- **Interpretation**
  - H0: the distributions of both samples are equal.
  - H1: the distributions of both samples are not equal.

Statistical Hypothesis - **Nonparametric Statistical Hypothesis Tests**

- **Wilcoxon Signed-Rank Test, Friedman Test**: Tests whether the distributions of two paired samples are equal or not.
- **Assumptions**
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample can be ranked.
  - Observations across each sample are paired.
- **Interpretation**
  - H0: the distributions of both samples are equal.
  - H1: the distributions of both samples are not equal.

Statistical Hypothesis - **Nonparametric Statistical Hypothesis Tests**

- **Kruskal-Wallis H Test:** Tests whether the distributions of two or more independent samples are equal or not.
- **Assumptions**
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample can be ranked.
- **Interpretation**
  - H0: the distributions of both samples are equal.
  - H1: the distributions of both samples are not equal.

Statistical Hypothesis - **Nonparametric Statistical Hypothesis Tests**

- **Friedman Test:** Tests whether the distributions of two or more paired samples are equal or not.

- **Assumptions**
  - Observations in each sample are independent and identically distributed (iid).
  - Observations in each sample can be ranked.
  - Observations across each sample are paired.

- **Interpretation**
  - H0: the distributions of all samples are equal.
  - H1: the distributions of one or more samples are not equal.

Statistical Hypothesis

- **Normality Tests**: Shapiro-Wilk Test, D'Agostino's K^2 Test, Anderson-Darling Test.
- **Correlation Tests**: Pearson's Correlation Coefficient, Spearman's Rank Correlation, Kendall's Rank Correlation, Chi-Squared Test.
- **Stationary Tests**: Augmented Dickey-Fuller, Kwiatkowski-Phillips-Schmidt-Shin,
- **Parametric Statistical Hypothesis Tests**: Student's t-test, Paired Student's t-test, Analysis of Variance Test (ANOVA), Repeated Measures ANOVA Test.
- **Nonparametric Statistical Hypothesis Tests**: Mann-Whitney U Test, Wilcoxon Signed-Rank Test, Kruskal-Wallis H Test, Friedman Test.

The End of Module 6