Begüm Demirdöven - 20210608021

# MATH 236 - İSTATİSTİK

→ **Population**: The collection of all items or things under consideration.

↳ Türkiye'de yapılan secimdeki tüm secmenler ($\mu, \sigma^2, p$)

→ **Sample**: A portion of the population selected for analysis.

↳ Türkiye'de yapılan secimdeki öğrenci olan tüm secmenler ($\hat{\mu}, \hat{\sigma}^2, \hat{p}$)

→ **Parameter**: A summary measure that describes a characteristic of the population. → $\mu, \sigma$ ... → bunlar parametredirler. (Populasyon özellikleri)

→ **Statistic**: A summary measure computed from a sample to describe a characteristic of the population.

↳ $\hat{\mu}, \hat{\sigma}$ ... → bunlar istatistiklerdir. (sample özellikleri)

** Amaç sample'ların istatistikleri kullanarak populasyonların parametrelerine ulaşmaktır.

## Descriptive Statistics (Tanımlayıcı İstatistik)

→ Collecting data (survey)

→ Summarizing data (sample mean)

→ Presenting data (tables, graphs)

## Inferential Statistics (Çıkarımsal İstatistik)

→ Drawing conclusions and/or making decisions concerning a population based only on sample data.

1) Estimation

2) Hypothesis Testing.

## 1) ESTIMATION (Tahmin):

→ Estimate the population mean weight using the sample mean weight. → Örneğin, sample ortalama ağırlığı kullanarak popülasyon ortalama ağırlığını tahmin etmek.

## 2) HYPOTHESIS TESTING:

→ Test the claim that the population mean weight is 140 pounds → Örneğin, nüfusun ortalama ağırlığının 140 pound olduğu iddiasını test etmek.

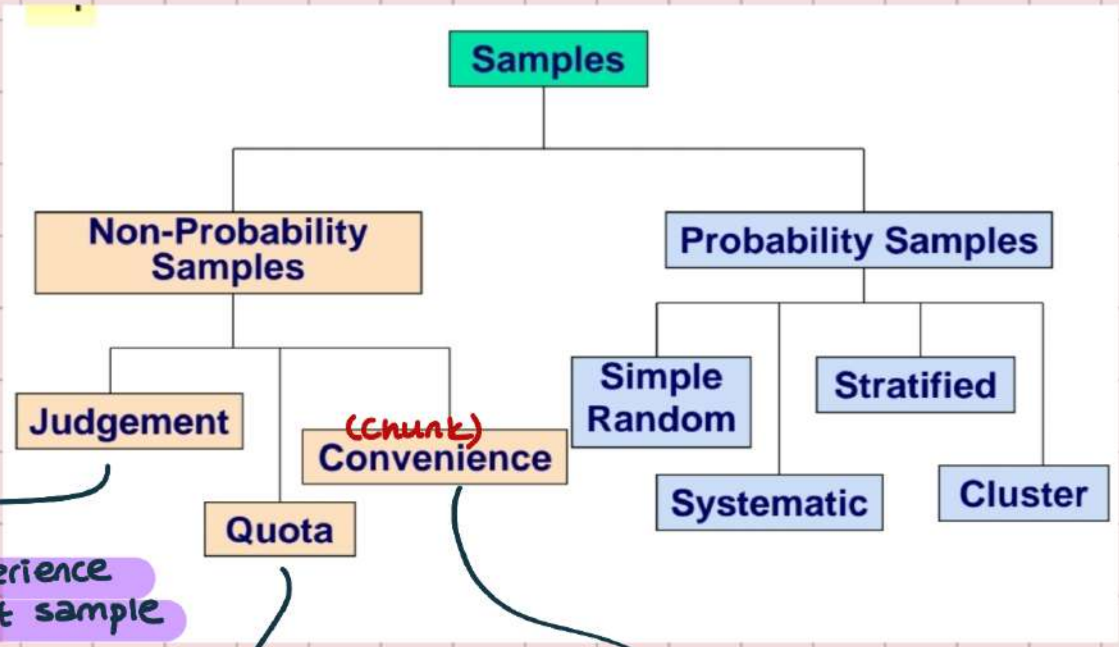
## * TYPES OF SAMPLES USED

1) **Nonprobability Sample:** Items included are chosen without regard to their probability of occurrence.

↳ Dahil edilen item`lar, olma olasılıkları dikkate alınmadan secilir.

2) **Probability Sample:** Items in the sample are chosen on the basis of known probabilities.

↳ Numunedeki item`lar, olma olasılıkları dikkate alınarak secilir.


Samples

Non-Probability Samples

Probability Samples

Judgement

Quota

(chunk) Convenience

Simple Random

Systematic

Stratified

Cluster

Judgement → use experience to select sample

Quota → Similar to stratified sampling except NO Random sampling.

Convenience → use elements most available (hangisi kolaysa o sample`ı secmek)

## * PROBABILITY SAMPLES:

1) **Simple Random Samples:** Every individual or item from the frame has an equal chance of being selected. (Her bir öğeninin bir topluluktan secilebilme şansı eşittir.)

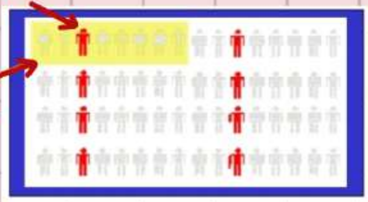
2) **Systematic Samples:** Sample size`ı belirle (n). Bu popülasyonda (N) adet birey bulunsun. Bu topluluğu n ve N değerlerine göre (k) adet gruba böl.

→ $k = N/n$

↳ Bir bireyi random olarak 1. gruptan sec.

↳ k kadar gruptan bu bireyleri secmeye devam et.

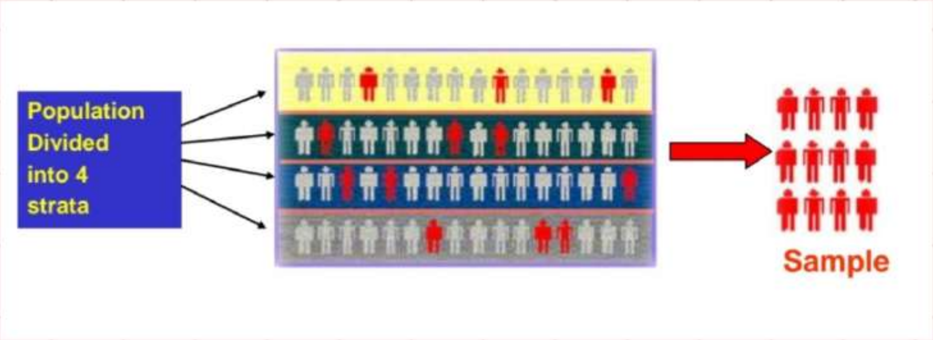
N=64
k=8
n=8



3) Stratified Samples: Divide population into two or more subgroups (called strata) according to some commen characteristics.

↳ A simple random sample is selected from each subgroup, with sample sizes proportional to strata sizes.

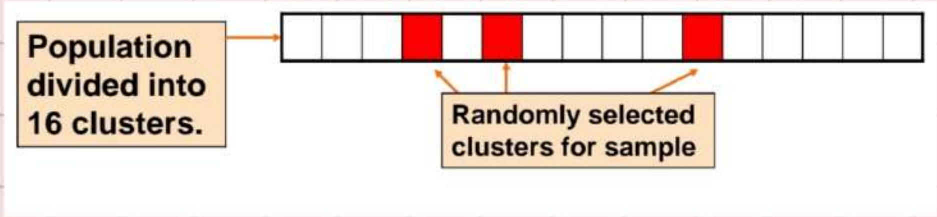
↳ Samples from subgroups are combined into one.



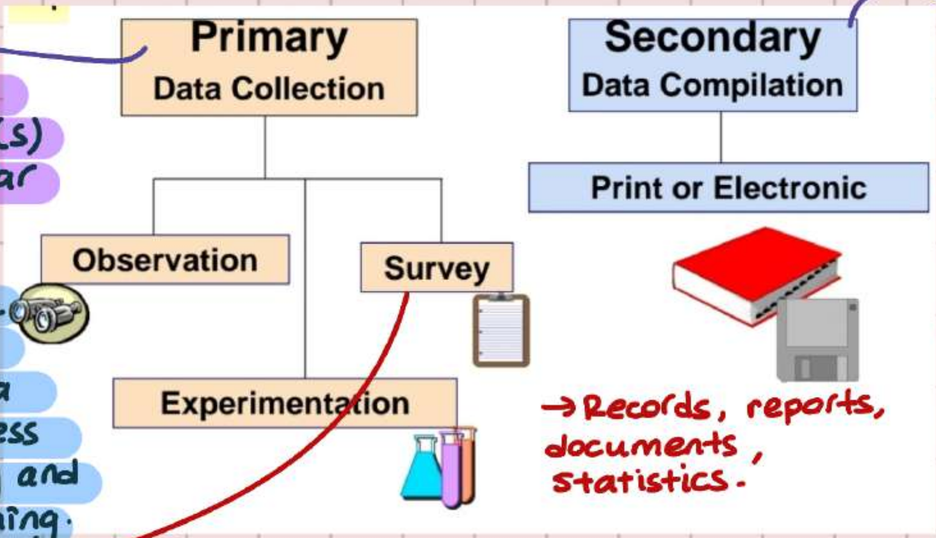
(küme)

4) Cluster Samples: Population is divided into several "clusters", each representative of the population.

↳ A simple random sample of clusters is selected.


Population divided into 16 clusters.
Randomly selected clusters for sample

→ DATA SOURCES:



Data obtained from research(s) on a particular subject.

→ Although the error rate is low, the data collection process can be costly and time consuming.

Data compiled by individuals or institutions for various purposes in advance.

→ Although the data collection process is less costly and requires a short time, its reliability is low.

→ Records, reports, documents, statistics.

• TYPES OF SURVEY ERRORS:

→ Coverage Error or Selection Bias:

↳ var olan bazı gruplar topluluktan çıkartılmışsa ve bir daha hiç seçilme hakları yoksa

→ Non Response Error or Bias:

↳ Cevap vermeyenler, cevap verenlerden farklı olabilir.

→ Sampling Error:

↳ Sample`dan sample`a çeşitlilik her zaman olacaktır.

→ Measurement Error:

↳ Sorunun sorulma şeklinin yanlışlığı, cevaptaki yanlışlıklar, survey yapan kişinin cevaba olan etkisi.

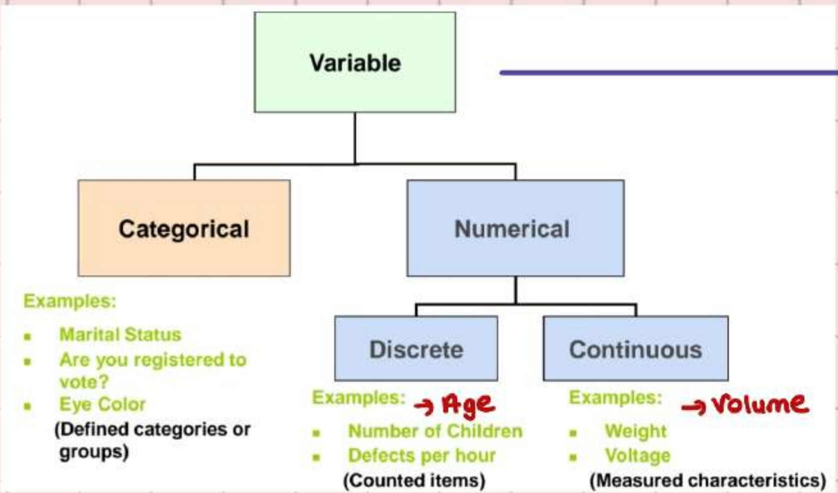
→ SURVEY STEPS:

1) Define Purpose → 2) Design Questionnaire → 3) Select Sample (Design sample Type, Size)

4) Collect data → 5) Prepare data (edit code) → 6) Analyze Data

7) Interpret Findings → 8) Report Results

→ VARIABLE: A variable is a specific characteristic (such as age or weight) of an individual or object. (Income, Voting Rate, GPA, Gender, occupation)



→ Values of the variables are called data.

→ Data is classified into two groups:

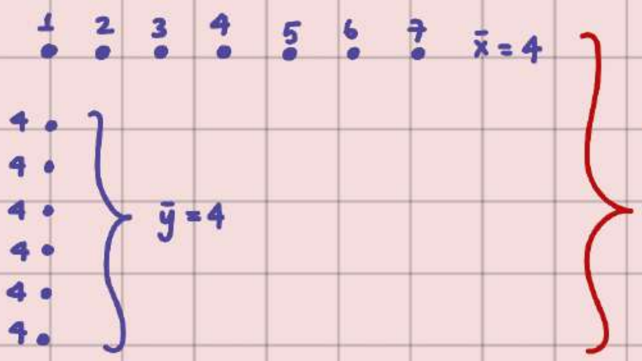
1) Qualitative Data: Cannot be expressed by numerical values. Categories are used to describe.

2) Quantitative Data: Can be obtained by counting or measuring and expressed by numerical values.

* NUMERICAL MEASURES: Quantitative variable`ları tanımlamanın iki farklı yolu vardır:

1) Measures of Location: average, mod, medyan → the purpose of a measure of location is to pinpoint the center of a distribution of data.

2) Measures of Dispersion: Standart devision → Dataların nasıl yayıldığını gösterir.



* İkisinin average'ı aynı ancak dispersion'ları farklı. Hangisinin daha iyi olduğuna daha fazla data ile ulaşırız.

## → MEASURES of LOCATION : The Sample mean and Median

### → ARITHMETIC MEAN :

→ If the data set is the entire population of data, then the population mean, $\mu$, is a parameter given by :

average of population ←

$$\boxed{\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}}$$

→ population values

→ population size

→ If the data set is from a sample, the sample mean $\bar{x}$ (x bar), is a statistic given by :

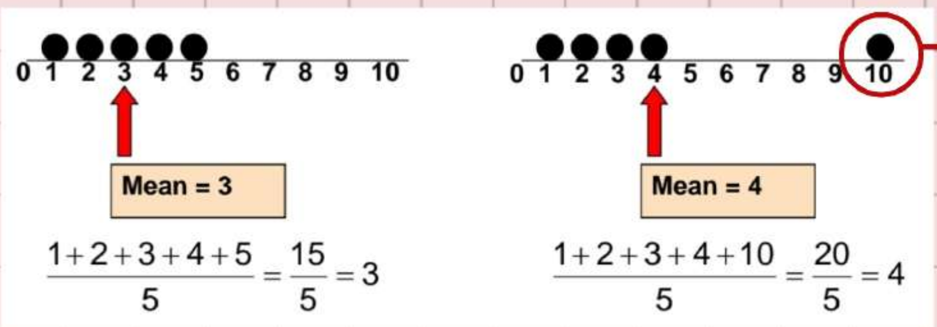
average of sample ←

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}}$$

→ Observed values

→ sample size

* Arithmetic Mean, extreme value'lardan ve outlier'lardan etkilenir.



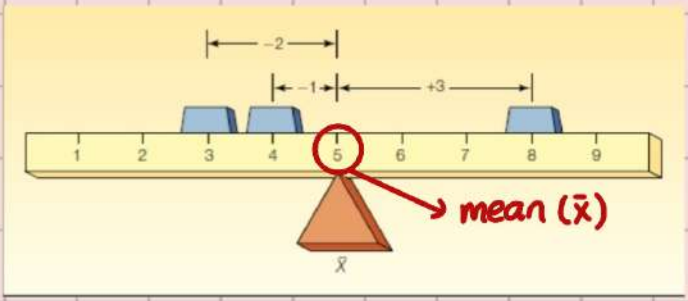
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

→ outlier (sapan değer)

* The sum of the deviations of each value from the mean is zero. Expressed by:

$$\boxed{\sum (x - \bar{x}) = 0}$$

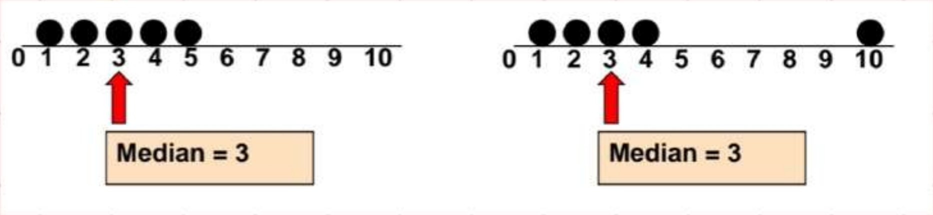


$$\Rightarrow \sum (x-\bar{x}) = (3-5)+(8-5)+(4-5)=0$$

→ MEDIAN:

→ The median is the middle observation of a set of observations that are arranged in increasing (or decreasing) order.

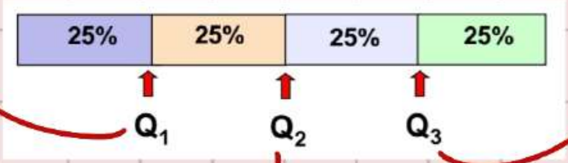
→ Not affected by extreme values.



$$\boxed{\bar{x} = \begin{cases} x_{(n+1)/2} & \text{, if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{, if } n \text{ is even} \end{cases}}$$

→ QUARTILES:

→ Quartiles split the ranked data into 4 segments with an equal number of values per segment.



The first quartile, $Q_1$, is the value for which 25% of the observations are smaller and 75% are larger.

$$\boxed{Q_1 = 0.25(n+1)}$$

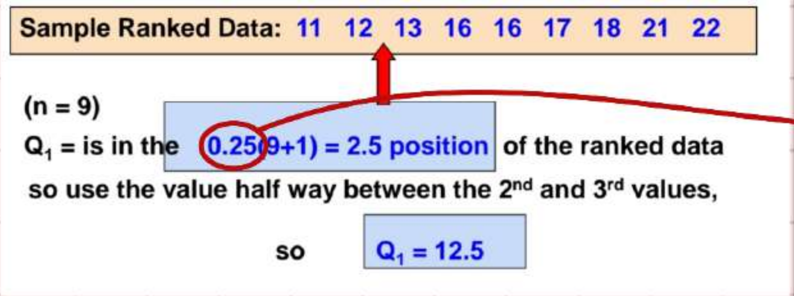
$Q_2$ is the same as the median (50% are smaller, 50% are larger)

$$\boxed{Q_2 = 0.50(n+1)}$$

Only 25% of the observations are greater than the third quartile.

$$\boxed{Q_3 = 0.75(\underbrace{n}_{\text{number of observation}}+1)}$$

Example:



Sample Ranked Data: 11 12 13 16 16 17 18 21 22

(n = 9)

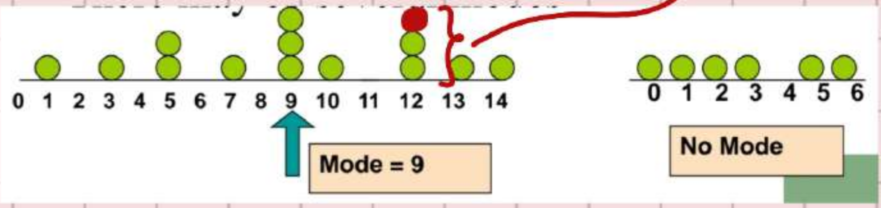
$Q_1$ = is in the 0.25(9+1) = 2.5 position of the ranked data so use the value half way between the 2nd and 3rd values,

so $Q_1 = 12.5$

verilerin %25'i bu değerin altında %75'i bu değerin üstünde kaldı.

→ **MODE:** (En çok tekrarlanan değer)

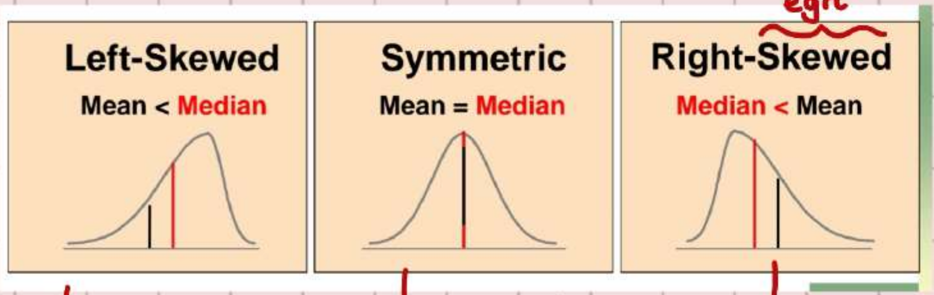
- Value that occurs most often.
- Not affected by extreme values
- There may be no mode.
- There may be several modes.



bu durumda 2 tane modumuz olur.

- Describes how data are distributed.

eğri



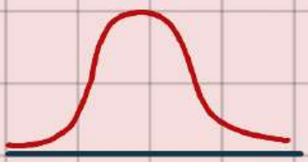
ünide genelin yüksek az kişinin düşük aldığı durum.

Normal Distribution (çan eğrisi)

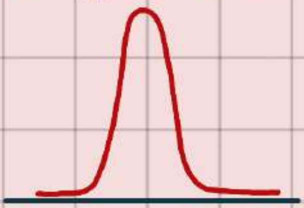
ünide genelin düşük az kişinin yüksek aldığı durum.

→ **VARIABILITY:** (Çeşitlilik)

- Variability indicates how spread out the scores are.
  - When there are large differences among scores, the data are said to contain a lot of variability.
- Consistency is the opposite of the variability.
  (tutarlılık)



High variability
Low predictability



Low variability
High predictability

→ Distributions with the same mean, may have different variability:

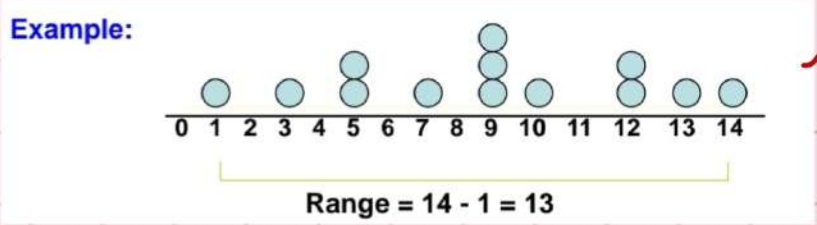
- 0, 2, 6, 10, 12
- 8, 7, 6, 5, 4
- 6, 6, 6, 6, 6

## → RANGE

- The greater spread of the data from the center of the distribution, the larger range will be.
- Difference between the largest and the smallest observations:

$$\boxed{Range = X_{largest} - X_{smallest}}$$

outlier olduğu zaman data pointlerden birisi etkilenir.


Example:

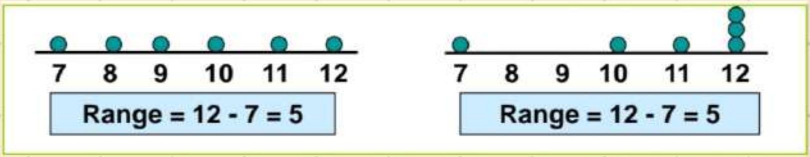
Range = 14 - 1 = 13

### Disadvantages:

- Range data`nın nasıl yayıldığını gösterir. Outlier olduğu durumda range güvenilmez olabilir.

Bu durumdan kacmak icin data artan veya azalan şeklinde sıralanır ve outlier olan değerlerden birkacı cıkartılır.

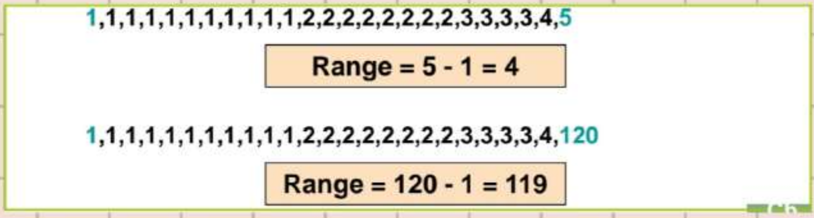
- Ignores the way in which data are distributed:



Range = 12 - 7 = 5

Range = 12 - 7 = 5

- Sensitive to outliers:


1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

## → INTERQUARTILE RANGE:

- Can eliminate some outlier problems by using the interquartile range.
- The interquartile range (IQR) measures the spread in the middle 50% of the data.
- Eliminate high- and low valued observations and calculate the range of the middle 50% of the data.

Interquartile Range = 3rd quartile − 1st quartile

IQR = Q − Q

$$\boxed{IQR = Q_3 - Q_1}$$

→ VARIANCE

→ The average of sum of squared terms is called the variance.

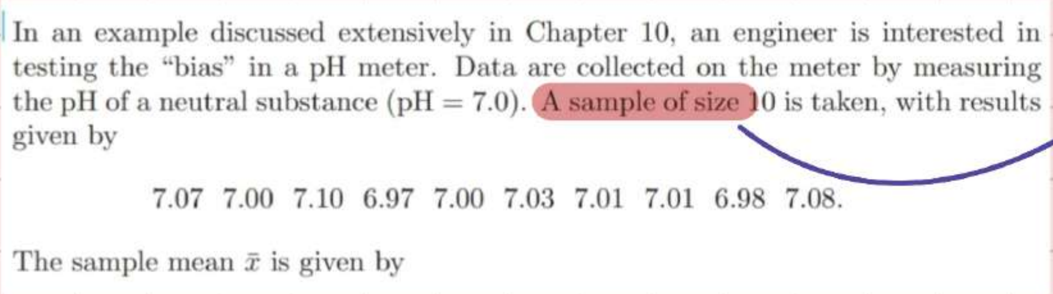
→ The sample variance, denoted by $s^2$, is given by

→ Each value in the data set is used in the calculation.

$$\boxed{s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \longrightarrow \boxed{s = \sqrt{s^2}}$$

↳ standart devision

## Example:



In an example discussed extensively in Chapter 10, an engineer is interested in testing the "bias" in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH = 7.0). A sample of size 10 is taken, with results given by

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08.

The sample mean $\bar{x}$ is given by

Eğer population deseydi $\sigma^2$ formülünü kullanırdık. Bunu kullanmak için de $\mu$'yi ilk olarak buluruz.

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{70.25}{10} = 7{,}025$$

$$s^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{10-1} = \frac{(7.07 - 7.025)^2 + (7 - 7.025)^2 + \ldots + (7.08 - 7.025)^2}{9}$$

→ GRAPHICAL SUMMARY

1) Scatterplot

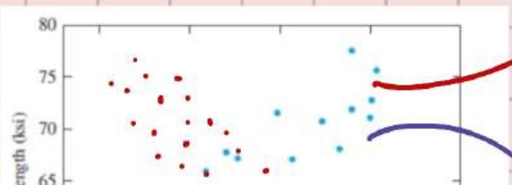
2) Steam and Leaf Display

3) Histogram

4) Box Plot

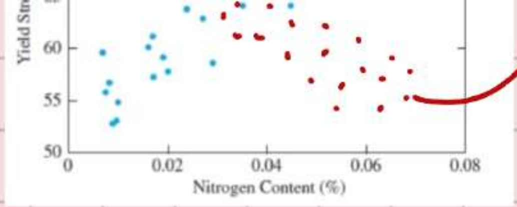
## 1) SCATTERPLOT:

- Data for which items consist of a pair of values is called bivariate. (iki değişkenli)
- The graphical summary for bivariate data is a scatterplot.

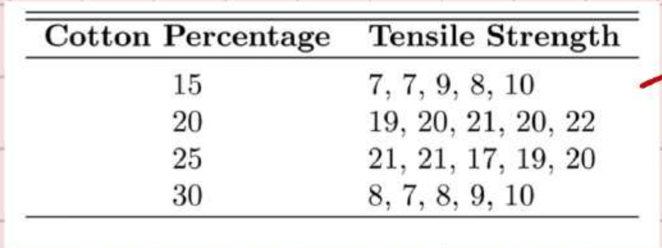


x ve y gibi 2 değişkenimiz varsa ve bunlar arasındaki ilişkiyi ölçmek için kullanılır.

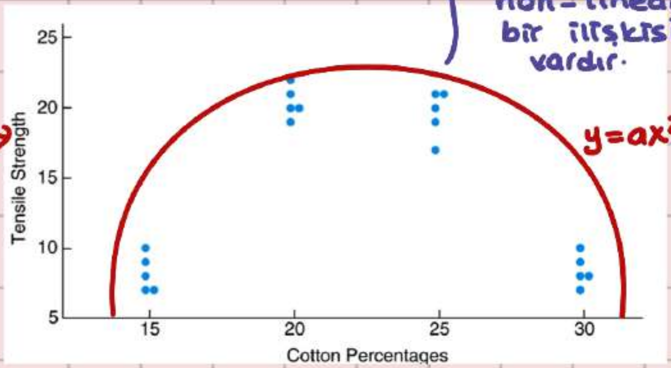
linear : $y=a+bx$



→ negatif plot



| Cotton Percentage | Tensile Strength |
|---|---|
| 15 | 7, 7, 9, 8, 10 |
| 20 | 19, 20, 21, 20, 22 |
| 25 | 21, 21, 17, 19, 20 |
| 30 | 8, 7, 8, 9, 10 |



→ x ve y'lerin non-linear bir ilişkisi vardır.

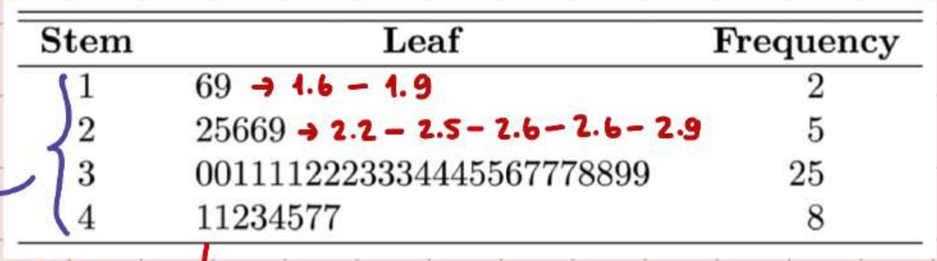
→ en fazla ilk iki basamak alınır.

## 2) STEM and LEAF PLOT

→ Each item in the sample is divided into two parts : a stem, consisting of the leftmost one or two digits ; and the leaf, which consist of the next digit.

100, 101, 105, 109, 118, 113

| STEM | LEAF |
|---|---|
| 10 | 0,1,5,9 |
| 11 | 3,8 |

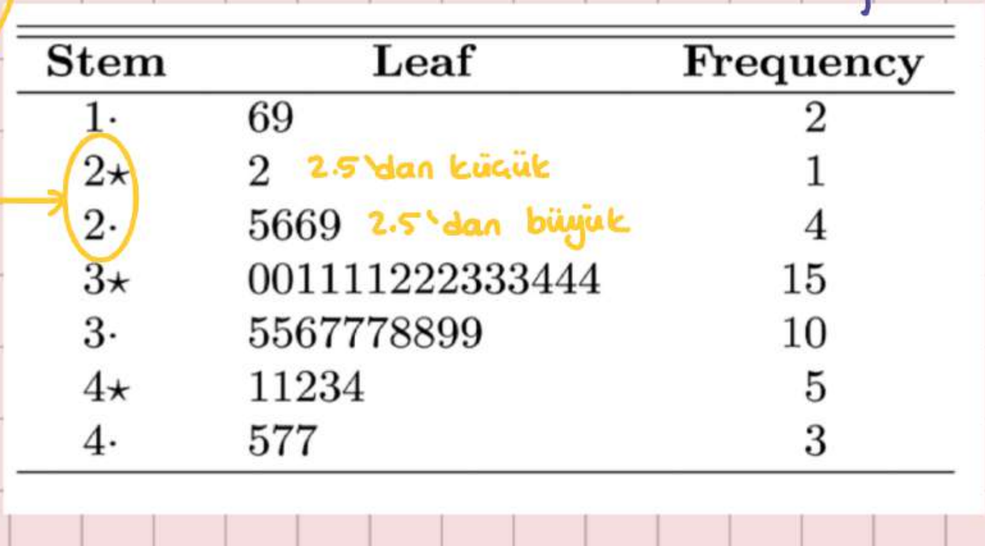


| Stem | Leaf | Frequency |
|---|---|---|
| 1 | 69 → 1.6 – 1.9 | 2 |
| 2 | 25669 → 2.2 – 2.5 – 2.6 – 2.6 – 2.9 | 5 |
| 3 | 0011112223334445567778899 | 25 |
| 4 | 11234577 | 8 |

→ 4.1 – 4.1 – 4.2 – 4.3 – 4.4 – 4.5 – 4.7 – 4.7

[1,2)
[2,3) → [2,2.5) , [2.5,3)
[3,4)
[4,∞)

→ data'yı detaylandırdık.



| Stem | Leaf | Frequency |
|---|---|---|
| 1· | 69 | 2 |
| 2⋆ | 2 2.5'dan küçük | 1 |
| 2· | 5669 2.5'dan büyük | 4 |
| 3⋆ | 001111222333444 | 15 |
| 3· | 5567778899 | 10 |
| 4⋆ | 11234 | 5 |
| 4· | 577 | 3 |

### → FREQUENCY DISTRIBUTION

* A frequency distribution is a list or a table...

*! Use at least 5 but no more than 15-20 intervals (RULE 1)

*! Intervals never overlap.

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

data:

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: 58 - 12 = 46
- Select number of classes: 5 (usually between 5 and 15) → Rule 1'i kullanarak kendimiz seçtik.
- Compute interval width: 10 (46/5 then round up) → length of each interval
- Determine interval boundaries: 10 but less than 20, 20 but less than 30, . . . , 60 but less than 70

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

| Interval | Frequency | Relative Frequency | Percentage |
|---|---|---|---|
| 10 but less than 20 | 3/20 → | .15 | 15 |
| 20 but less than 30 | 6/20 → | .30 | 30 |
| 30 but less than 40 | 5/20 → | .25 | 25 |
| 40 but less than 50 | 4/20 → | .20 | 20 |
| 50 but less than 60 | 2/20 → | .10 | 10 |
| Total | 20 | 1.00 | 100 |

→ total sample

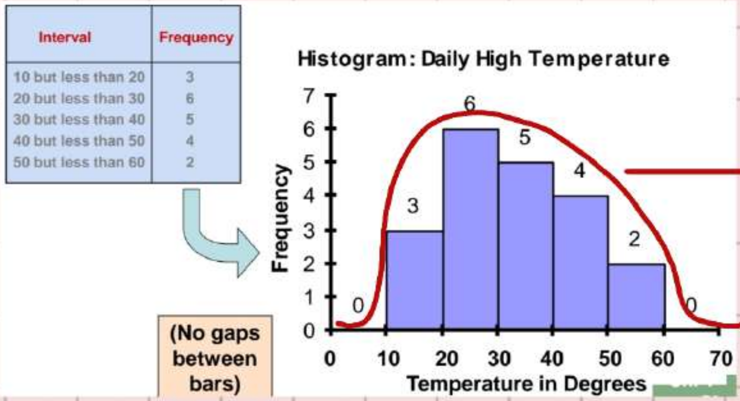
Ch. 1-

→ HISTOGRAM:

→ A graph of data in frequency distribution is called a histogram.

→ The interval endpoints are shown on the horizontal axis.

→ The vertical axis is either frequency, relative frequency, or percentage.

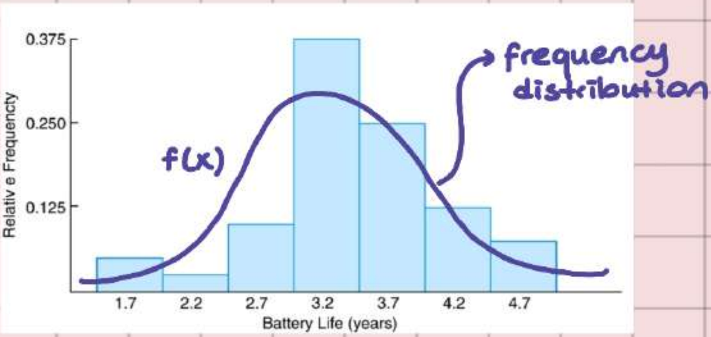


| Interval | Frequency |
|---|---|
| 10 but less than 20 | 3 |
| 20 but less than 30 | 6 |
| 30 but less than 40 | 5 |
| 40 but less than 50 | 4 |
| 50 but less than 60 | 2 |

(No gaps between bars)

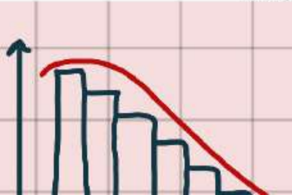
→ shape of distribution'a bu şekilde ulaşırız.

| Class Interval | Class Midpoint | Frequency, $f$ | Relative Frequency |
|---|---|---|---|
| 1.5–1.9 | 1.7 | 2 | 0.050 |
| 2.0–2.4 | 2.2 | 1 | 0.025 |
| 2.5–2.9 | 2.7 | 4 | 0.100 |
| 3.0–3.4 | 3.2 | 15 | 0.375 |
| 3.5–3.9 | 3.7 | 10 | 0.250 |
| 4.0–4.4 | 4.2 | 5 | 0.125 |
| 4.5–4.9 | 4.7 | 3 | 0.075 |



*

*

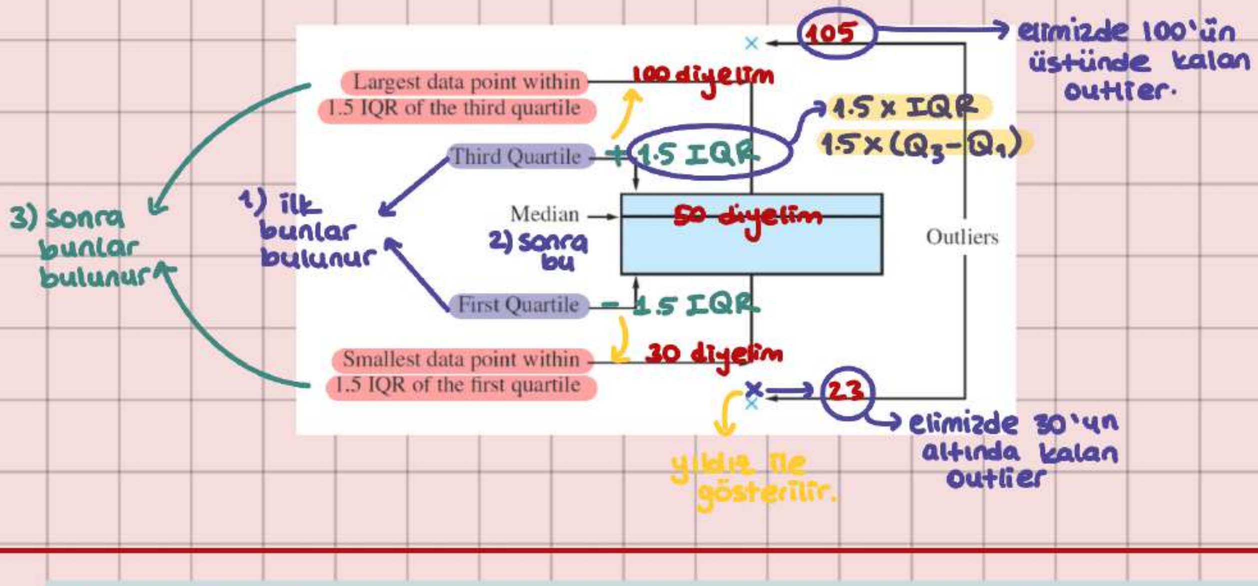


Normal Distribution

Exponential Distribution

## → Box and WHISKER PLOT or Box PLOT

→ A boxplot is a graphic that presents the median, the first and third quartiles, and any outliers present in the sample.



## → DEFINITION OF A RANDOM VARIABLE

→ A random variable assigns a numerical value to each outcome in a sample space

(Rastgele süreclerin sonuclarının gercek sayılarla eşleştirilmesi)

→ Bir olasılık değeri değildir. X, Y, Z gibi büyük harfler ile tanımlanır.

→ Ayrık olasılık Dağılımı (Discreate Prob.)

→ Sürekli olasılık Dağılımı (Continous Prob.)

**Aynık Rastgele Değişken:** Bir şeylerin sayısıdır. (The number of....) → varsa discreate

(Discreate Prob. Distribution)

* Bir para 3 kez atılıyor. → X: Tura gelmesinin sayısı → $X = \{0, 1, 2, 3\}$

Olasılık Deneyi

0 → hiç tura gelmez; 1 → 1 kere gelebilir; 2 → 2 kere gelebilir; 3 → 3 defa gelebilir.

* 4 white 6 Black balls in a box. Two balls are selected → X: number of black balls.

Olasılık deneyi

$P(WW) = P(X=0)$
$P(BW \text{ or } WB) = P(X=1)$
$P(BB) = P(X=2)$

$\Big\}\; X = \{0, 1, 2\}$

0 → hiç black gelmez; 1 → 1 tane gelir; 2 → 2 tane gelir.

* Bir zar 2 kez atılıyor → y: 5 gelmelerin sayısı → $y = \{0, 1, 2\}$

Olasılık Deneyi

**Sürekli Rastgele Değişken:** Sayılamayan durumlara ait bir olaydır.
(Continous Prob. Distributions)

- bir şeyin süresi } the length of...
- bir şeyin miktarı } the amount of... } varsa continous!

- $1 < x < 3$
- $5 \leq y \leq 10$

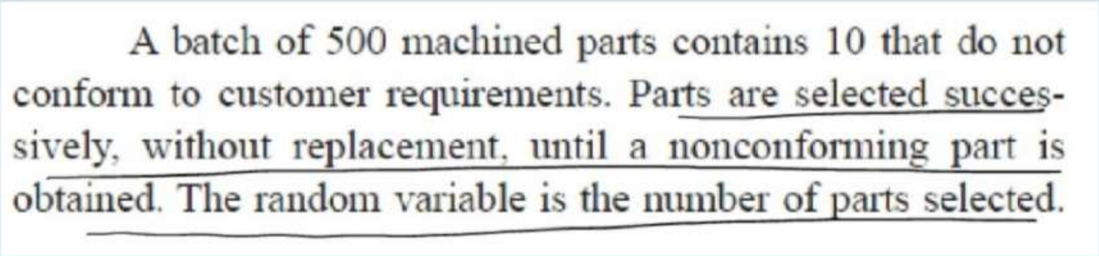
→ x: the number of defective products
→ x: the number of COVID-19 cases in Turkey
→ x: the number of deaths in Turkey
} Ayrık (Discreate Probability Distributions)

→ y: the weight of newborn baby
→ y: proportion of the defective products in a production line
→ y: duration of quarantine because of the pandemic
} Sürekli (Continous Probability Distributions)

* Random variable bir fonk. olduğu için x'in değer aralığından (range) bahsedebiliriz.

- x'in değer kümesi ⇒ Range(x) or $R_x$

Example:



A batch of 500 machined parts contains 10 that do not conform to customer requirements. Parts are selected successively, without replacement, until a nonconforming part is obtained. The random variable is the number of parts selected.

x: # number of parts selected

- 490 conform
- 10 not conform

$R_x = \{1, 2, 3 \ldots\ldots, 490, 491\}$ (Discreate Random Sample)

hep conform seçecek o yüzden 1 ile başladı

490'a kadar sorunsuz seçebilir. 491. kesin non-conform olmak zorunda kalır ve o zaman durur.

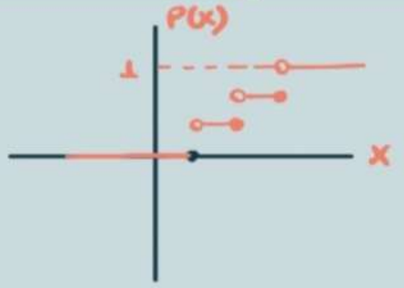
→ CUMULATIVE DISTRIBUTION FUNCTION (cdf) f(x) → (Discreate) kesikli olasılık dağılımı (pmf) (olasılık kütle fonk.)

f(x) → süretli olasılık dağılımı (Continous) (pdf) (olasılık yoğunluk fonk.)

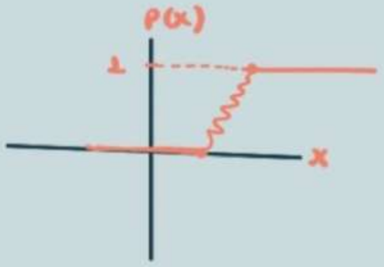
⇒ f(x) 'den ↓ elde edilen F(x) ile yazılan bir fonksiyondur.

**F(x) = birikimli dağılım fonk.**

* Olasılıklar toplanarak 1'e ulaşır ve F(x) oluşur.

* 

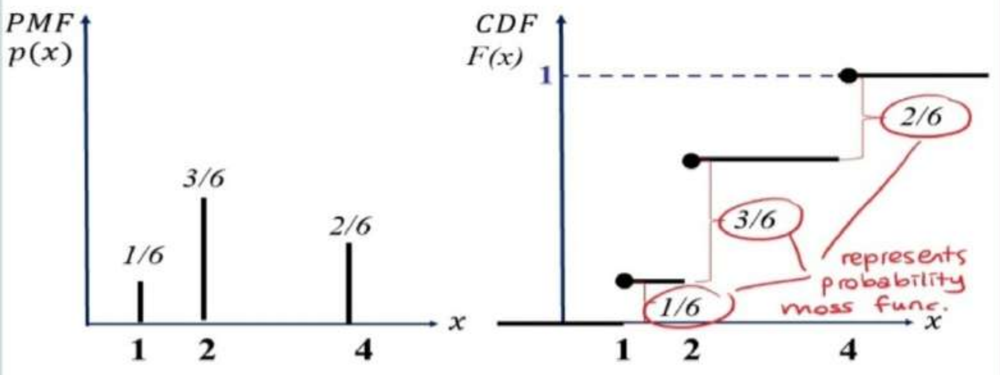
→ Discreate CDF F(x)

* 

→ Continous CDF F(x)

* F(x) daima sürekli fonkdur.

PMF and CDF of a Discrete Random Variable



P(x=1)=1/6
P(x=2)=3/6
P(x=4)=2/6

P(x<1)=0
P(x≤1)=1/6

P(x≤2)=P(x=1)+P(x=2)=4/6
P(4)=2/6

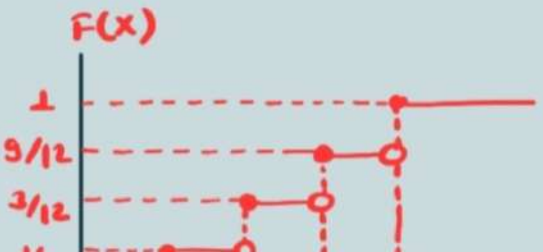
**Example:**

| X | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| f(x) | $\frac{1}{12}$ | $\frac{1}{6}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

3/12 9/12 1

biçiminde olasılık kütle fonksiyonu veriliyor. X rastgele değişkeninin birikimli dağılım fonksiyonunu bulunuz.

$$F(x) = \begin{cases} 0, & x<1 \\ 1/12, & 1 \le x < 2 \\ 3/12, & 2 \le x < 3 \\ 9/12, & 3 \le x < 4 \\ 1, & 4 \le x \end{cases}$$



**Example:**

$$f(x) = \begin{cases} kx, & x = 1,2,4 \\ 0, & \text{diğer durum} \end{cases}$$

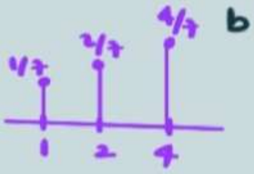
olasılık kütle fonksiyonu veriliyor.

a) k kaçtır?

b) F(x) fonk ve grafiği ?

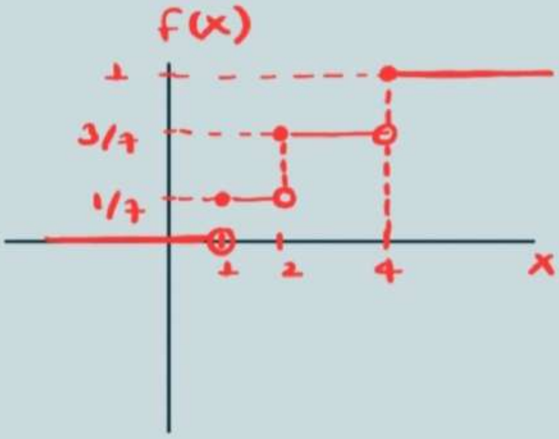
a) $k + 2k + 4k = 1$

$k = 1/7$

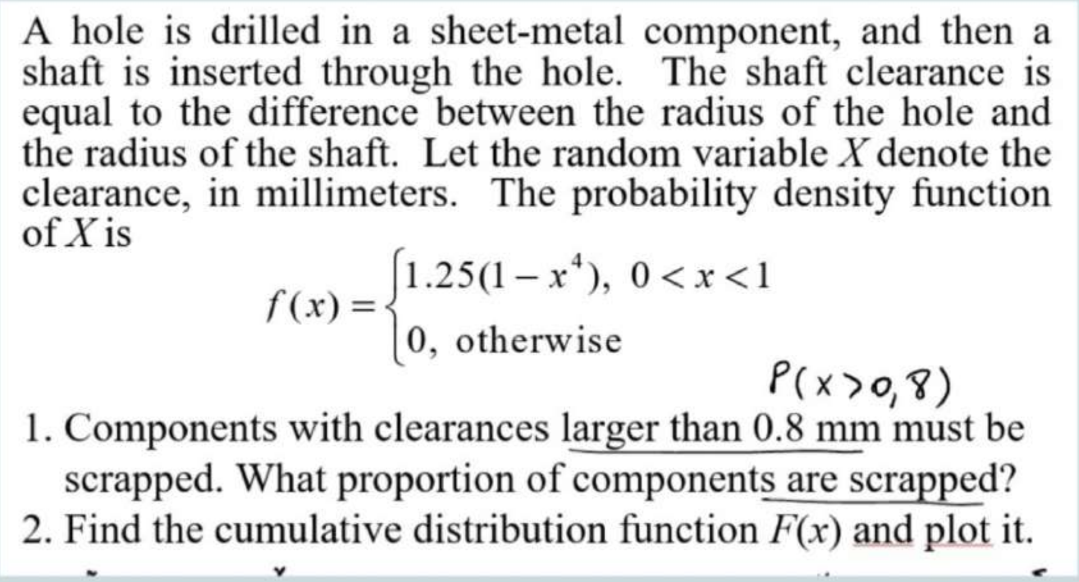


b)

$$F(x) = \begin{cases} 0, & x < 1 \\ 1/7, & 1 \leq x < 2 \\ 3/7, & 2 \leq x < 3 \\ 1, & 4 \leq x \end{cases}$$



**Example:**



A hole is drilled in a sheet-metal component, and then a shaft is inserted through the hole. The shaft clearance is equal to the difference between the radius of the hole and the radius of the shaft. Let the random variable $X$ denote the clearance, in millimeters. The probability density function of $X$ is

$$f(x) = \begin{cases} 1.25(1-x^4), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

$P(x > 0,8)$

1. Components with clearances larger than 0.8 mm must be scrapped. What proportion of components are scrapped?
2. Find the cumulative distribution function $F(x)$ and plot it.

①

$$P(x>0,8) = \int_{0,8}^{1} (1,25)(1-x^4)\,dx = (1,25)\left(x - \frac{x^5}{5} \Big|_{0,8}^{1}\right)$$

②

$$F(x) = \int_{-\infty}^{x} f(x)\,dx \rightarrow F(x) = \int_{0}^{x} (1,25)(1-x^4)\,dx = (1,25)\left(x - \frac{x^5}{5}\right)\Big|_{0}^{x}$$

$$= (1,25).\left(x - \frac{x^5}{5}\right) //$$

$$F(x) = \begin{cases} 0, & x \leq 0 \\ (1,25)\left(x - \frac{x^5}{5}\right), & 0 < x < 1 \end{cases}$$

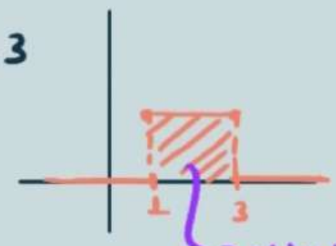
$1 \quad , \; x \geq 1$

→ **CONTINOUS PROBABILITY DISTRIBUTIONS**

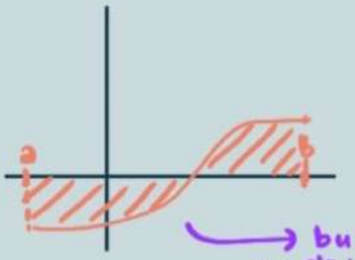
* $X \rightarrow 1 < x < 5$ formunda olması lazım.

→ Sürekli olasılık dağılımları bir parçalı fonk. ile gösterilir.

$$f(x) = \begin{cases} x+1 \; , & 1<x<3 \\ 0 \; , & \text{diğer} \end{cases}$$



altında kalan alan 1 olmalı



→ bu cont prob dis. grafiği olamaz. çünkü; $f(x) \geq 0$ olmalı

**Sürekli Olasılık Dağılımı Olma Şartları:**

① $\int_{-\infty}^{\infty} f(x)\,dx = 1$ olmalı

② $f(x) \geq 0$ olmalı

**Example:**

$$f(x) = \begin{cases} x+1 \; , & 1<x<3 \\ 0 \; , & \text{diğer} \end{cases}$$

fonksiyonu sürekli olasılık dağılımı mıdır?

1. şart

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{1}^{3} (x+1)\,dx = \frac{x^2}{2} + x \Big|_{1}^{3} = \frac{15}{2} - \frac{3}{2} = 6 \neq 1$$

→ Sürekli olasılık dağılımı değildir.

* Sürekli Olasılık Dağılımı = Olasılık Yoğunluk Fonksiyonu (Probability Density Function (pdf))

* $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = \int_{a}^{b} f(x)\,dx$

≤ ve < işareti veya ≥ ve > işareti arasında fark yoktur.

* $P(x \leq a) = P(x < a) = \int_{-\infty} f(x)dx$

* $P(x \geqslant a) = P(x > a) = \int_{a}^{\infty} f(x)dx$

Example:

$$f(x) = \begin{cases} 2/15\, x \ , & 1 < x < 4 \\ 0 \ , & \text{diğer} \end{cases} \quad \text{pdf.}$$

a) $P(x=3) = ?$ 0 (continousda eşitlik olmaz)

b) $P(2 < x < 3) = ?$ $\int_{2}^{3} \frac{2}{15} x\, dx$

c) $P(2 \leq x) = ?$ $\int_{2}^{\infty} f(x)dx = \int_{2}^{4} \frac{2}{15} x\, dx$

Example:

$$f(x) = \begin{cases} cx \ , & 1 < x < 5 \\ 0 \ , & \text{diğer} \end{cases}$$

şeklinde olasılık yoğunluk fonk. verilmiştir.

a) c kaçtır? 1/12

b) $P(X=3)$ kaçtır? 0

c) $P(x < 2) = ?$

d) $P(2 < x \leq 4) = ?$

a) $\int_{-\infty}^{\infty} f(x)dx = 1$ olmalı

$$\int_{1}^{5} cx\, dx = \frac{cx^2}{2} \Big|_{1}^{5} = \frac{25c}{2} - \frac{c}{2} = 1 \rightarrow c = 1/12$$

c) $$P(x < 2) = \int_{1}^{2} \frac{x}{12} dx = \frac{x^2}{24} \Big|_{1}^{2} = \frac{4}{24} - \frac{1}{24} = \frac{1}{8}$$

d) $$P(2 \leq x < 4) = \int_{2}^{4} x/12\, dx = \frac{x^2}{24} \Big|_{2}^{4} = \frac{16}{24} - \frac{4}{24} = \frac{1}{2}$$

$\parallel$

$P(2 < x < 4)$

## → CUMULATIV DISTRIBUTION FUNCTION

* f(x) sürekli olasılık dağılımı ⟶ F(x) birikimli dağılım fonksiyonu

$$\boxed{F(x) = \int_{-\infty}^{x} f(x)\,dx}$$

**Example:**

$$f(x) = \begin{cases} kx, & 1 < x < 3 \\ 0, & \text{diğer} \end{cases}$$

olasılık yoğunluk fonksiyonu veriliyor.

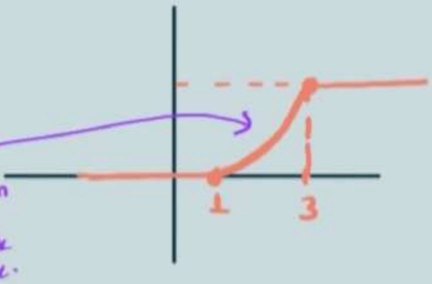
a) k nedir?

b) birikimli dağılım fonk olan F(x)'i bulunuz ve grafiğini çiziniz.

a) $\int_{-\infty}^{\infty} f(x)dx = 1 \rightarrow \int_{1}^{3} kx\,dx = \frac{kx^2}{2}\Big|_{1}^{3} = \frac{9k}{2} - \frac{k}{2} = \frac{8k}{2} = 1 \rightarrow k = 1/4$

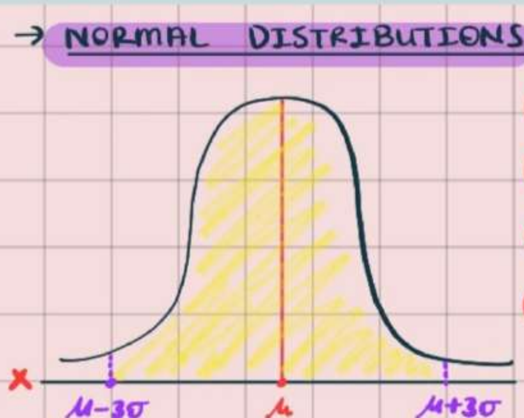
b) $F(x) = \int_{-\infty}^{x} f(x)dx \rightarrow F(x) = \int_{1}^{x} \frac{x}{4}dx = \frac{x^2}{8}\Big|_{1}^{x} = \frac{x^2}{8} - \frac{1}{8}$

$$F(x) = \begin{cases} 0, & x \leq 1 \\ \frac{x^2}{8} - \frac{1}{8}, & 1 < x < 3 \rightarrow \\ 1, & x \geq 3 \end{cases}$$



x²'li olduğu için arasını parabolik yaptık.

## → NORMAL DISTRIBUTIONS



① Çan eğrisinin altında kalan alan 1'dir.

② Maksimum nokta μ ile eşleşir.

③ Simetriktir.

④ μ+3σ ve μ-3σ sağa ve sola gidildiğinde tüm alanın %99,9.... civarlarına ulaşılır.

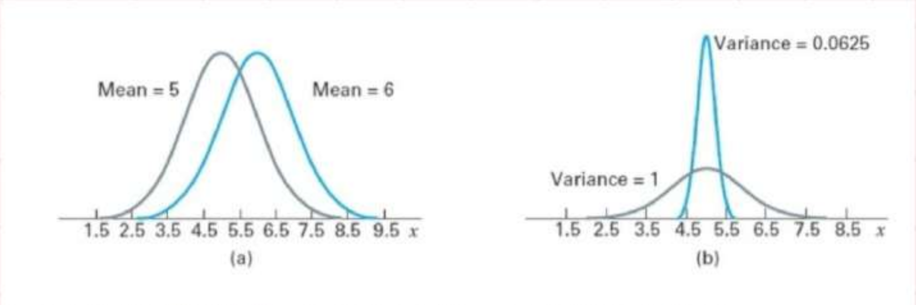
$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \boxed{-\infty < x < \infty}}$$

$\boxed{\mu_x = \mu}$

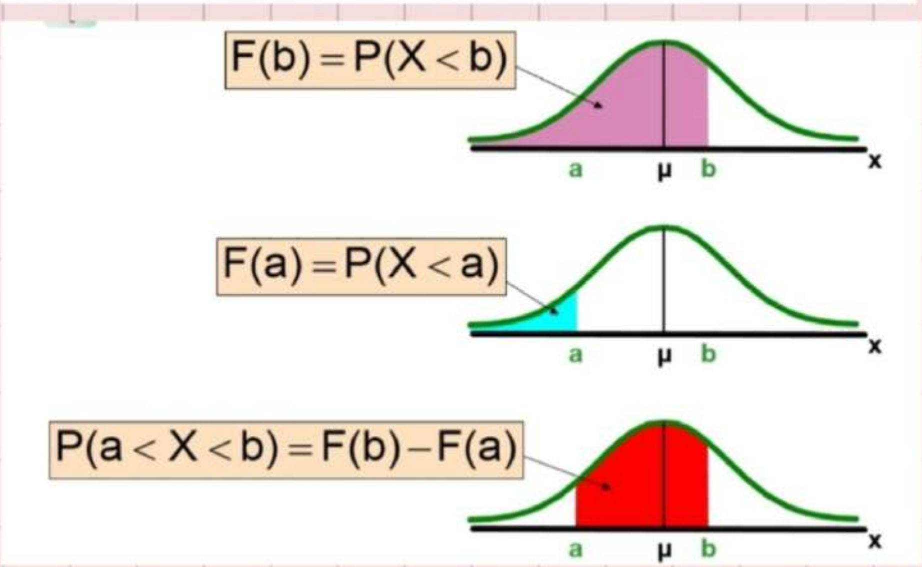
$\boxed{\sigma_x^2 = \sigma^2}$

$\boxed{X \sim N(\mu, \sigma^2)}$

$\int_{-\infty}^{\infty} f(x)\,dx = 1$   $f(x) \geq 0$



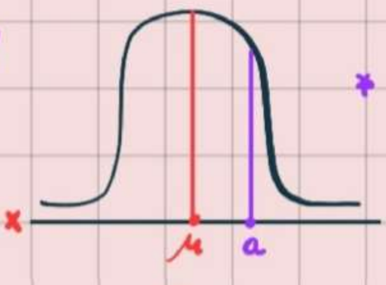
a. Two Normal Distributions with Same Variance but Different Means
b. Two Normal Distributions with Different Variances and Mean = 5



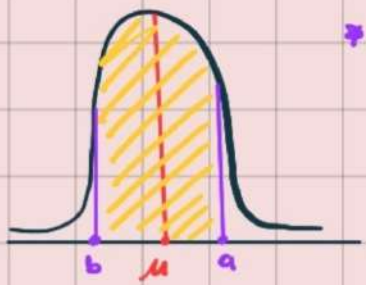
$F(b) = P(X < b)$

$F(a) = P(X < a)$

$P(a < X < b) = F(b) - F(a)$



$P(x=a) = ?$ 0 → (Normal distribution süreklidir. Aralık bildiren olasılıkların değerleri vardır. Eşitlik bildirenler 0'a eşittir.)



$P(a < x \leq b) = ?$

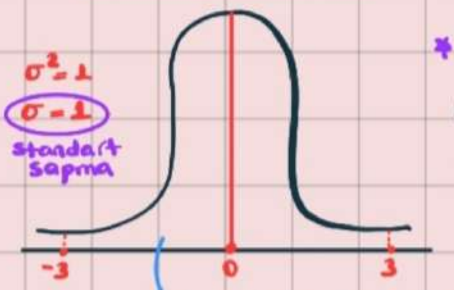
$$\underbrace{\int_a^b \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx}$$

Bu integralin hesaplanması zor olduğu için, bizim yerimize bu integraller hesaplanıp bir z tablosu oluşturulmuştur.

* Normal Dağılım (x'e bağlıdır) → (Normal Dağılım, Standart ND'ye çevrilir.) * Standart Normal Dağılım (z'ye bağlıdır) → z tablosu kullanılarak olasılık hesaplanır.

## → STANDART NORMAL DISTRIBUTION

$$\boxed{z = \frac{x-\mu}{\sigma}}$$ → Standartlaştırma işlemi



* Ortalaması 0 ve varyansı 1 haline getirilen normal dağılımlara standart normal dağılım denir. (Standart Normal Population)

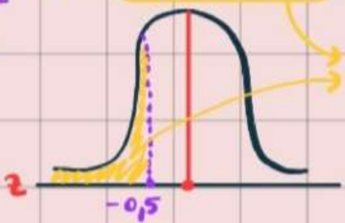
Bunun altında kalan alanları z tablosu ile hesaplarız.

**Ex**

$\overbrace{\text{Ortalaması}}^{\mu}$ 6 ve $\overbrace{\text{standart sapması}}^{\sigma}$ 2 olan bir normal dağılımda $P(x<5) = ?$

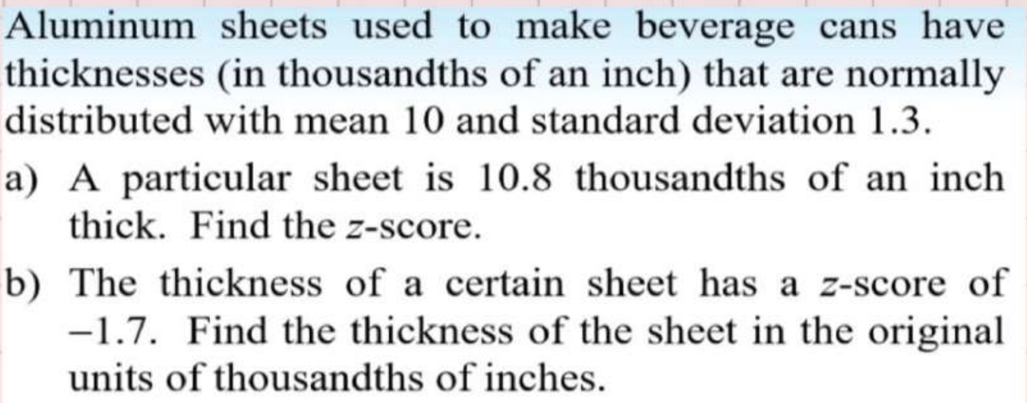
$$P(x<5) \Rightarrow P\left(z < \frac{5-6}{2}\right) = \boxed{P(z<-0{,}5)}$$

$z = \frac{x-\mu}{\sigma}$



Taralı alan bu olasılığı verecek Hesaplanması için de z tablosu kullanılır.

**Ex**



Aluminum sheets used to make beverage cans have thicknesses (in thousandths of an inch) that are normally distributed with mean 10 and standard deviation 1.3.

a) A particular sheet is 10.8 thousandths of an inch thick. Find the *z*-score.

b) The thickness of a certain sheet has a *z*-score of −1.7. Find the thickness of the sheet in the original units of thousandths of inches.

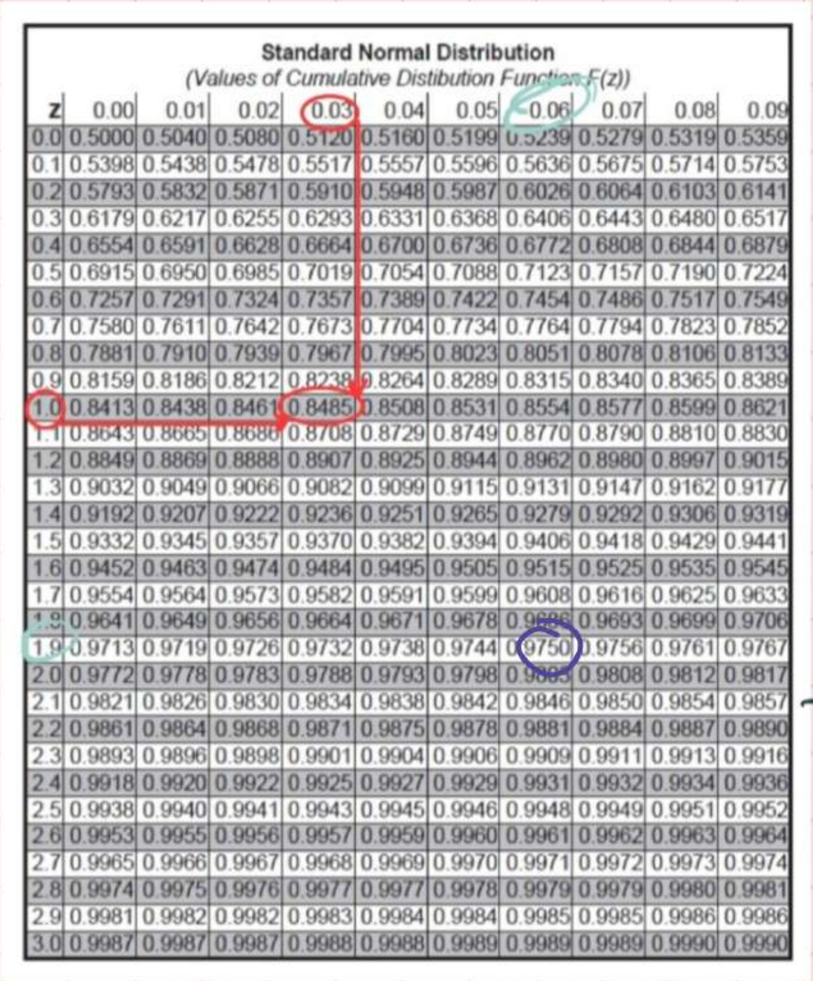
$X \sim N(10, (1.3)^2)$

a) $z = \frac{x-\mu}{\sigma}$ $\quad z = \frac{10.8-10}{1.3} = \frac{0.8}{1.3} = 0.615$

b) $z = \frac{x-10}{1.3} = -1.7$ $\quad x = 1.3 \times (-1.7) + 10$

$x = 12.21$

## → AREAS UNDER THE NORMAL CURVE

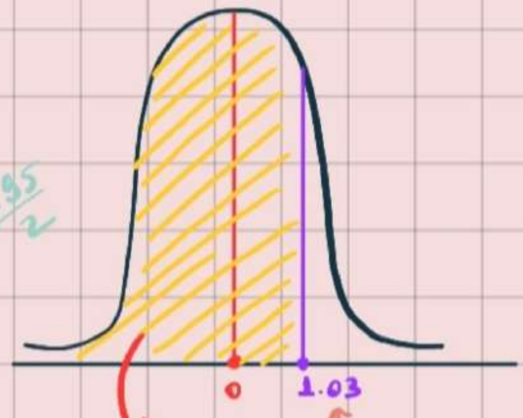


**Standard Normal Distribution**

*(Values of Cumulative Distibution Function F(z))*

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |

$\frac{1.95}{2}$



$1.0 + 0.03 = 1.03$

Bir tablo çizildiğinde ve bir nokta seçildiğinde o nokta tabloyu ikiye böler. Böldüğü kısımlardan büyük olanın alanını bu z tablosu verir.

## Ex

a) Find the area under normal curve to the left of $z = 0.47$.

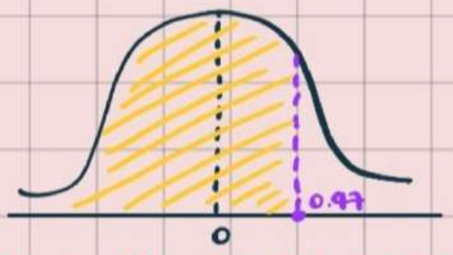
b) Find the area under the curve to the right of $z = 1.38$.

c) Find the area under the curve to the left of $z = -1.55$.

d) Find the area under the normal curve between $z = 0.71$ and $z = 1.28$.

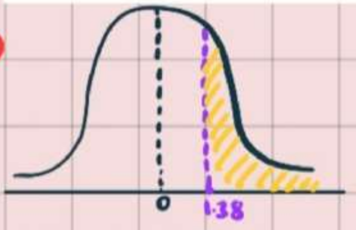
e) What $z$-score corresponds to the 75th percentile of a normal curve?

a)



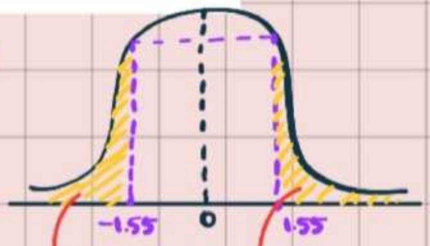
$P(z < 0.47) = F(0.47) = 0.6808$

b)



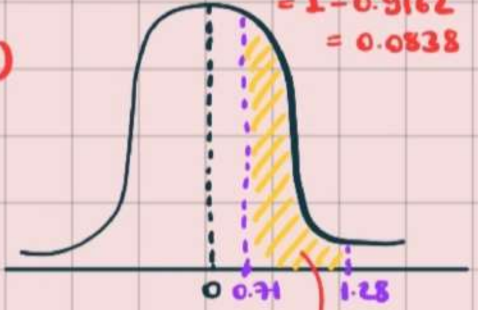
$P(z > 1.38) = 1 - P(z < 1.38)$
$= 1 - F(1.38)$
$= 1 - 0.9162$
$= 0.0838$

c)



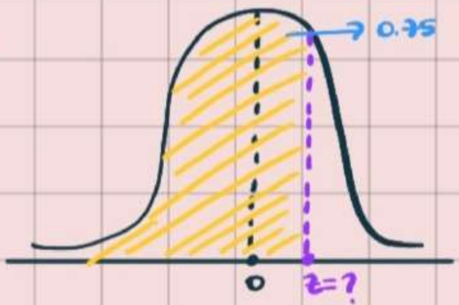
$P(z < -1.55) = P(z > 1.55) = 1 - F(1.55)$

d)



$= P(0.71 < z < 1.28)$

e)



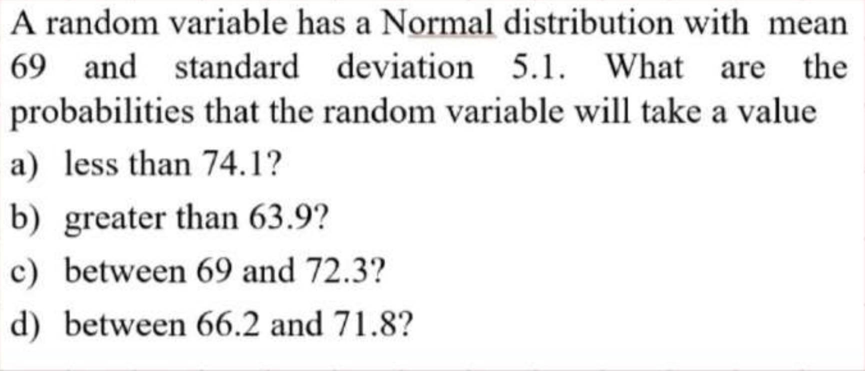
$P(z < ?) = 0.75 \Rightarrow F(?) = 0.75$

$= P(0.71 < z < 1.28)$ $\qquad$ $P(z < z) = 0.75 \Rightarrow F(z) = 0.75$

$= F(1.28) - F(0.71)$ $\qquad$ $z = 0.6745$

$= 0.8997 - 0.7611$

**Ex**



A random variable has a Normal distribution with mean 69 and standard deviation 5.1. What are the probabilities that the random variable will take a value

a) less than 74.1?

b) greater than 63.9?

c) between 69 and 72.3?

d) between 66.2 and 71.8?

a) $P(x < 74.1) = P\left(\frac{x-\mu}{\sigma} < \frac{74.1-69}{5.1}\right)$

$= P(z < 1) = F(1) = 0.8413$

b) $P(x > 63.9) = P\left(\frac{x-\mu}{\sigma} > \frac{63.9-69}{5.1}\right)$

$= P(z > -1) = P(z < 1) = F(1) = 0.8413$

c) $P(69 < x < 72.3) = P\left(\frac{69-69}{5.1} < z < \frac{72.3-69}{5.1}\right)$

$= P(0 < z < \underbrace{0.6471}_{0.65}) = F(0.65) - F(0)$

**Ex**

Bir sınıftaki öğrencilerin boylarının uzunluğu normal dağılmaktadır. Bu sınıftaki öğrencilerin boylarının uzunluğunun ortalaması 160 cm ve standart sapması 5 cm dir.

Standart normal dağılıma çevrilmeli

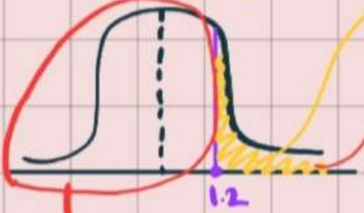
Bu sınıftan seçilen bir öğrencinin boyunun 166 cm'den uzun olma olasılığı kaçtır?

$\mu = 160$ cm

$\sigma = 5$ cm

$P(x > 166) = ?$

$P\left(z > \frac{166-160}{5}\right) = P(z > 1.2)$



$1 - 0.8849 = \underline{0.1151}$

0.8849
(z tablosu burayı

vectr.)

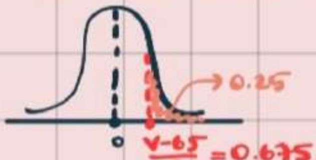
**Ex**

Bir yoldan geçen araçların hızları normal dağılmaktadır. Bu yoldan geçen araçların ortalama hızı 65 km/sa ve standart sapması 2km/sa'tir. Bu yolda belli bir hızın aşılması durumunda araçlara ceza kesildiğine göre, bu yolda izin verilen en yüksek hız kaç km/sa'tir?

$\mu = 65$ $\qquad V=$ ceza kesilen hız?

$\sigma = 2$ $\qquad P(X > V) = 0.25$

$z = \frac{x-\mu}{\sigma}$ $\qquad P\left(z > \frac{V-65}{2}\right) = 0.25$



$\frac{V-65}{2} = 0.675 = \frac{0.67+0.68}{2}$

(Tablodan yaklaşık olan iki değerin ort.'unu aldık)

$V-65 = 1.350$

$\boxed{V = 66.35 \text{km/sa}}$

**Ex**

Lifetimes of batteries in a certain application are Normally distributed with mean 50 hours and standard deviation 5 hours.

Find the probability that a randomly chosen battery lasts between 42 and 52 hours.

$\mu = 50$ $\qquad P(42 < x < 52) = ?$

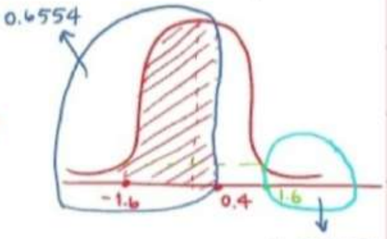
$\sigma = 5$ $\qquad P\left(\frac{42-50}{5} < z < \frac{52-50}{5}\right)$

$P(-1.6 < z < 0.4)$

$P(z < 0.4) - P(z > 1.6)$

$= 0.6554 - 0.0548$

$= 0.6006$



$1 - 0.9452$
$= 0.0548$

**Ex**

A random variable has a Normal distribution with variance 100. Find its mean if the probability that it will take on a value less than 77.5 is 0.8264.

$\sigma^2 = 100$ $\qquad P(x < 77.5)$

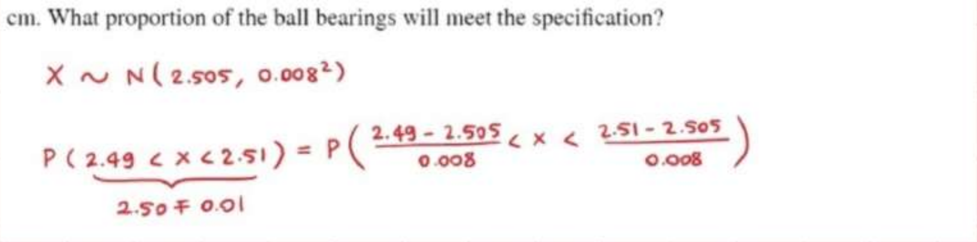
$\sigma = 10$ $\qquad P\left(z < \underbrace{\frac{77.5-\mu}{10}}_{0.94}\right) = 0.8264$

$\frac{77.5-\mu}{10} = 0.94$

$\boxed{\mu = 68.1}$

**Ex**

A process manufactures ball bearings whose diameters are normally distributed with mean 2.505 cm and standard deviation 0.008 cm. Specifications call for the diameter to be in the interval 2.5 ± 0.01

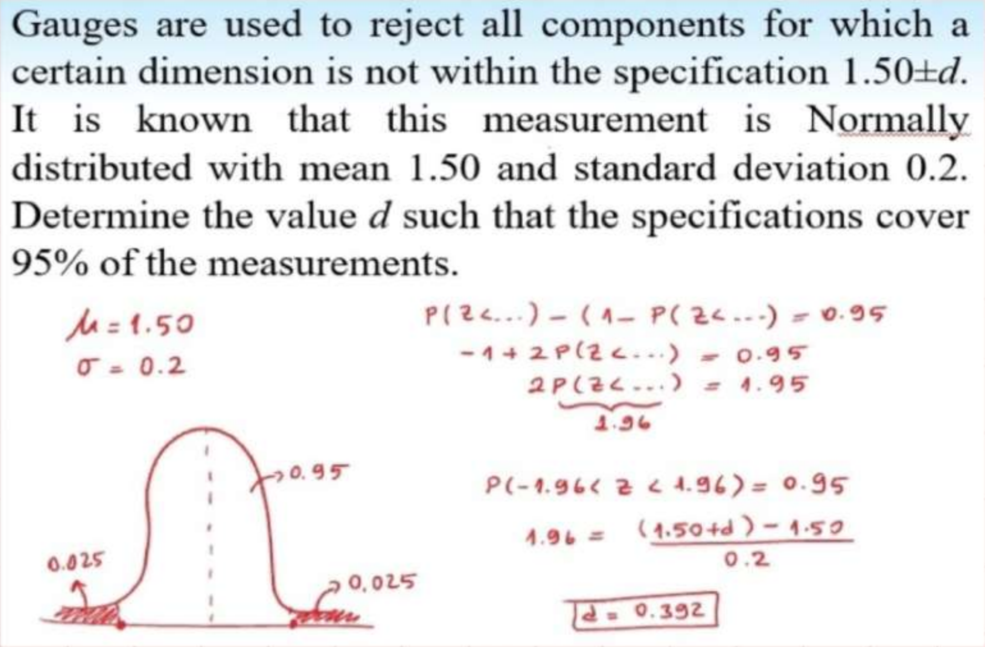


cm. What proportion of the ball bearings will meet the specification?

$$X \sim N(2.505,\ 0.008^2)$$

$$P(\underbrace{2.49 < X < 2.51}_{2.50 \mp 0.01}) = P\left(\frac{2.49-2.505}{0.008} < x < \frac{2.51-2.505}{0.008}\right)$$

## Ex



Gauges are used to reject all components for which a certain dimension is not within the specification 1.50±*d*. It is known that this measurement is Normally distributed with mean 1.50 and standard deviation 0.2. Determine the value *d* such that the specifications cover 95% of the measurements.

$\mu = 1.50$

$\sigma = 0.2$

$$P(Z<\dots) - (1 - P(Z<\dots)) = 0.95$$

$$-1 + 2P(Z<\dots) = 0.95$$

$$\underbrace{2P(Z<\dots)}_{1.96} = 1.95$$

$$P(-1.96 < Z < 1.96) = 0.95$$

$$1.96 = \frac{(1.50+d) - 1.50}{0.2}$$

$$\boxed{d = 0.392}$$

## → RANDOM SAMPLING

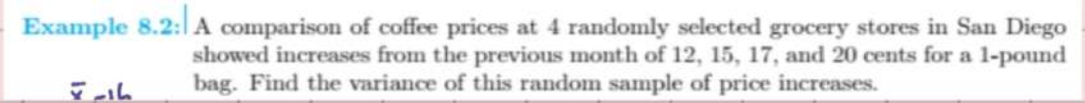
→ Let $X_1, X_2, \dots X_n$ be $n$ independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \dots, X_n$ to be a random sample of size $n$ from the population $f(x)$ and write its joint probability distribution as;

$$\boxed{f(x_1, x_2, \dots, x_n) = f(x_1) . f(x_2) \dots f(x_n)}$$

→ If $S^2$ is the variance of a random sample of size $n$, we may write

$$\boxed{S^2 = \frac{1}{n.(n-1)}\left[n . \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]}$$

Example :



**Example 8.2:** A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag. Find the variance of this random sample of price increases.

$\bar{x}=16$

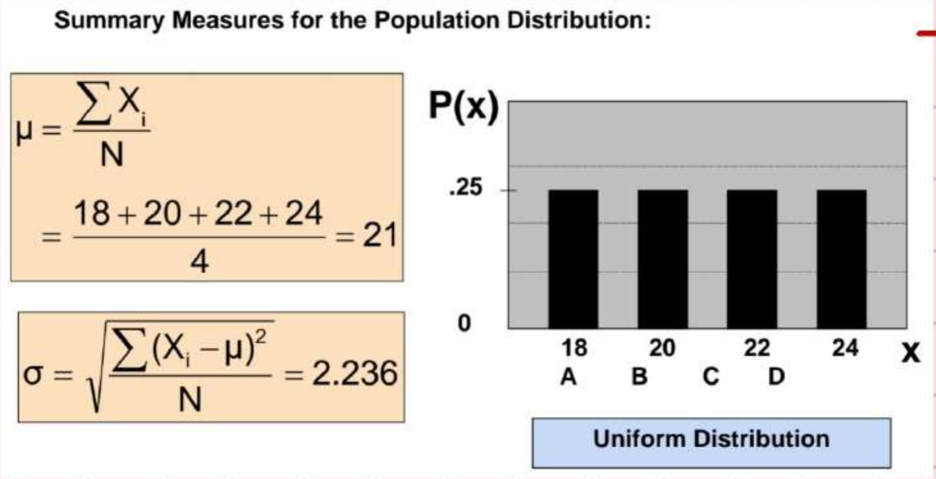
$$\frac{(12-16)^2+(15-16)^2+(17-16)^2+(20-16)^2}{(4-1)} = \hat{\sigma}^2 = s^2 = \frac{\sum(x_i-\bar{x})^2}{(n-1)}$$

→ SAMPLING DISTRIBUTIONS:

→ A sampling distribution is a probability distribution of all of the possible values of a statistic for a given size sample selected from a population

✱ Assume there is a population :

- Population size = 4
- Random variable, X, is age of individuals (Value of X: 18, 20, 22, 24 (years))



Summary Measures for the Population Distribution:

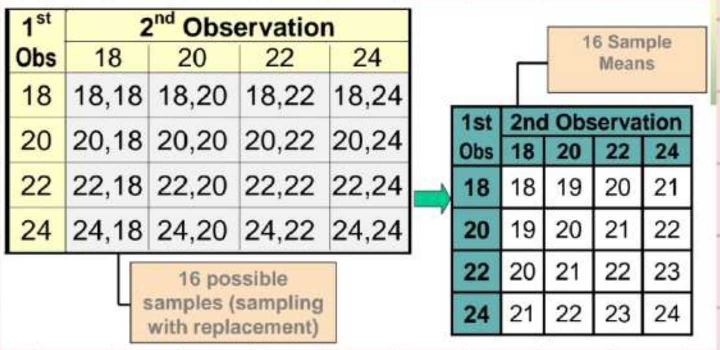
$$\mu = \frac{\sum X_i}{N} = \frac{18+20+22+24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum(X_i-\mu)^2}{N}} = 2.236$$

Uniform Distribution

→ Tek bir seçim yapıldığı durum.

✱ Now consider all possible samples of size n=2 (4 kişi içinden 2 kişi seçiyoruz)

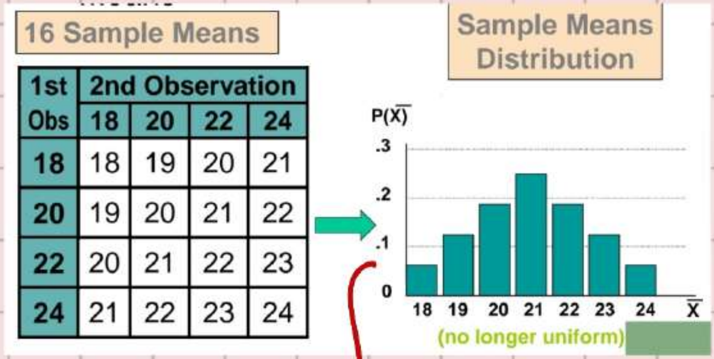


| 1st Obs | 2nd Observation 18 | 20 | 22 | 24 |
|---|---|---|---|---|
| 18 | 18,18 | 18,20 | 18,22 | 18,24 |
| 20 | 20,18 | 20,20 | 20,22 | 20,24 |
| 22 | 22,18 | 22,20 | 22,22 | 22,24 |
| 24 | 24,18 | 24,20 | 24,22 | 24,24 |

16 possible samples (sampling with replacement)

16 Sample Means

| 1st Obs | 2nd Observation 18 | 20 | 22 | 24 |
|---|---|---|---|---|
| 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 20 | 21 | 22 |
| 22 | 20 | 21 | 22 | 23 |
| 24 | 21 | 22 | 23 | 24 |



16 Sample Means

| 1st Obs | 2nd Observation 18 | 20 | 22 | 24 |
|---|---|---|---|---|
| 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 20 | 21 | 22 |
| 22 | 20 | 21 | 22 | 23 |
| 24 | 21 | 22 | 23 | 24 |

Sample Means Distribution

(no longer uniform)

2'li kombinasyonların ortalamasının dağılımı ($\bar{x}$)

→ SAMPLE MEAN:

✱ Let $X_1, X_2, \ldots, X_n$ represent a random sample from a population :

$$\boxed{\bar{X} = \frac{1}{n}\sum_{n=1}^{n} x_i}$$

## → STANDARD ERROR OF THE MEAN:

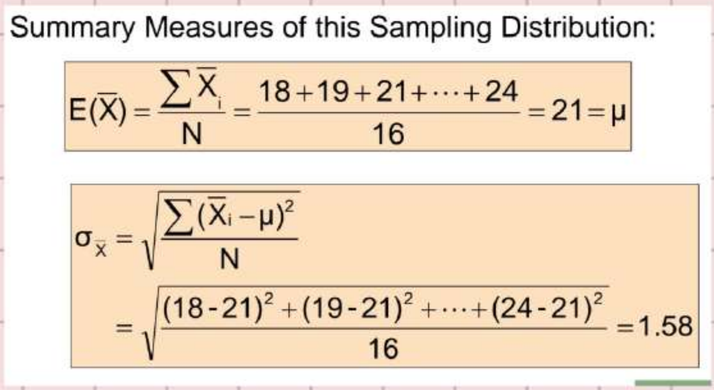
→ Different samples of the same size from the same population will yield different sample means

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

→ sample'ın mean'inin standart devision'u

→ Standart error of the mean

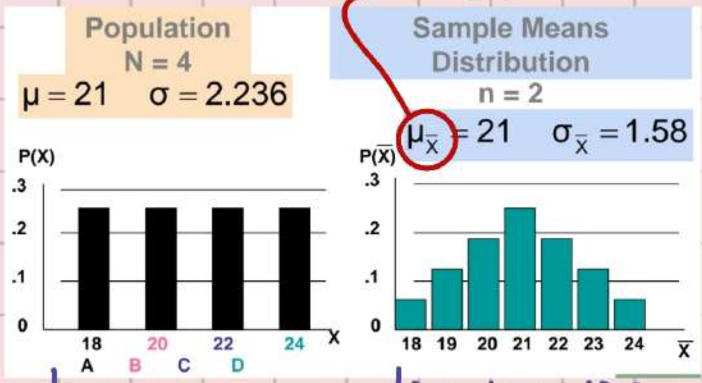
* Note that the standard error of the mean decreases as the sample size increases.



Summary Measures of this Sampling Distribution:

$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18+19+21+\cdots+24}{16} = 21 = \mu$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum(\bar{X}_i-\mu)^2}{N}} = \sqrt{\frac{(18-21)^2+(19-21)^2+\cdots+(24-21)^2}{16}} = 1.58$$



→ $E(\bar{x})$

↳ 1 observation

↳ 2 observation

## If the Population is Normal:

* If a population is normal with mean $\mu$ and standart deviation $\sigma$, the sampling distribution of $\bar{x}$ is also normally distributed with:

$$\boxed{\mu_{\bar{x}} = \mu} \qquad \boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

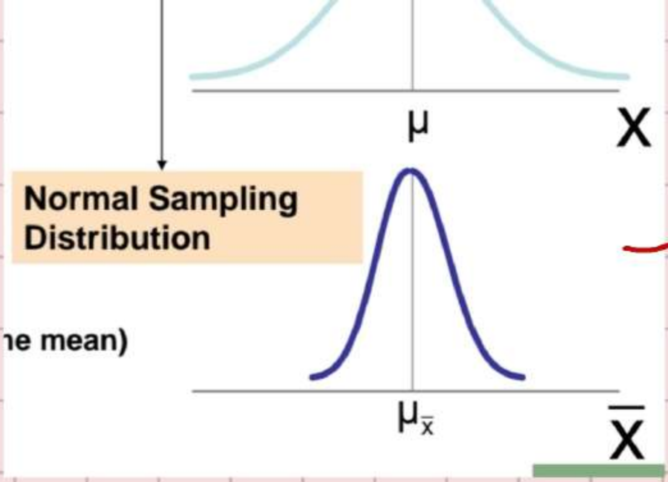
↳ ortalamanın ortalaması

* Z-value for the sampling distribution of $\bar{x}$ :

$$\boxed{z = \frac{(\bar{x}-\mu)}{\sigma_{\bar{x}}} = \frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}}} \qquad \boxed{\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$$

→ $\boxed{E[\bar{x}] = \mu}$

→ both distributions have the same mean ($\bar{x}$ is unbiased)

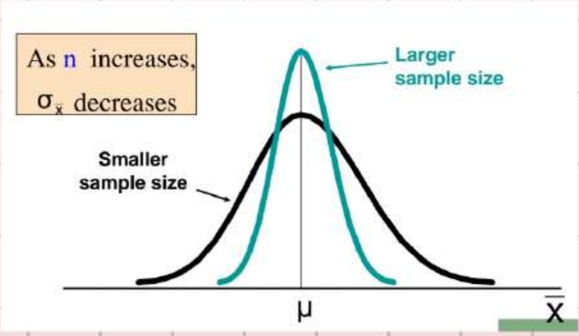


standart error of the mean

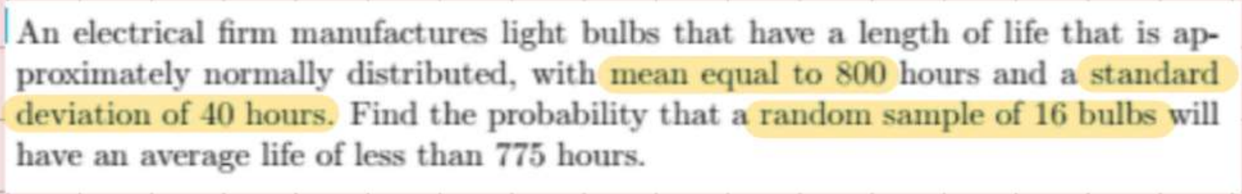
$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

→ the distribution of $\bar{x}$ has a reduced standart deviation

( $\bar{x}$ is unbiased)



Example:



An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

$X \sim N(\mu = 800,\ \sigma = 40)$

↳ lifetime

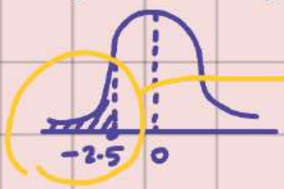
$n = 16$

$P(\bar{x} \leq 775) = ?$

$\bar{X} \sim N\left(\mu = 800,\ \frac{\sigma}{\sqrt{n}} = \frac{40}{4}\right)$

$$P\left\{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{775 - 800}{10}\right\}$$

$P\{z \leq -2.5\} = 1 - P\{z - 2.5\} = 0.0062$



If $X_1, X_2, \dots, X_n$ are independent random variables having normal distributions with means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively, then the random variable

$$\boxed{Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n}$$

has a normal distribution with mean :

$$\boxed{\mu_Y = a_1 \mu_1 + a_2 \mu_2 + \dots + a_n \mu_n}$$

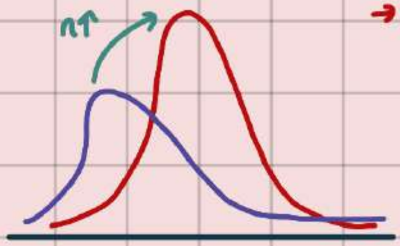
and variance :

$$\boxed{\sigma_Y^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2}$$

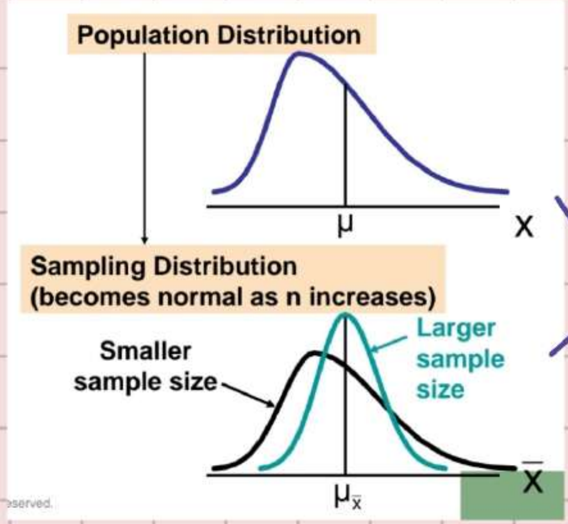
→ CENTRAL LIMIT THEOREM

## → CENTRAL LIMIT THEOREM

- If $\bar{x}$ is the mean of a random sample of size n taken from a population with mean $\mu$ and finite variance $\sigma^2$, then the limiting from of the distribution of $\boxed{z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}$ as $n \to \infty$, is the standard normal distribution $n(z;0.1)$.



- → As the sample size gets large enough the sampling distribution becomes almost normal regardless of shape of population. ( n(sample size) arttıkça distribution normal`e yaklaşır)



→ Sampling distribution properties :

$$\left.\begin{array}{l} \boxed{\mu_{\bar{x}} = \mu} \\ \boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}} \end{array}\right\}$$ sadece sample size değişiyor.

- For most distributions, n>30 will give a sampling distribution that is nearly normal.
- For normal population distributions, the sampling distribution of the mean is always normally distributed.

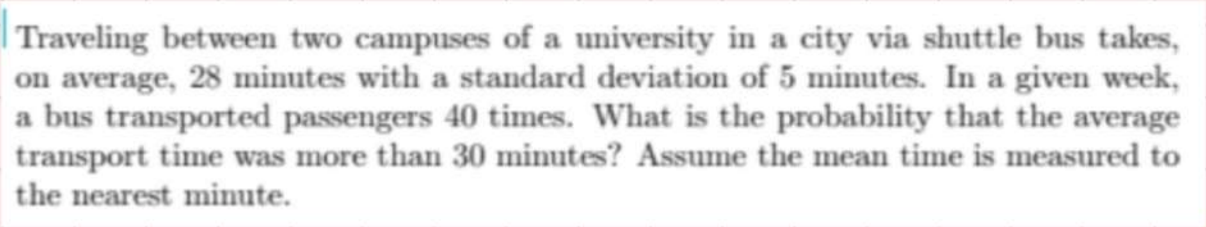
If the Population is NOT Normal :

- We can apply the Central Limit Theorem :
  - Even if the population is not normal, sample means from the population will be approximately normal as long as the sample size is large enough.
    ↳ n>30
    ↳ sample size >30 ise neredeyse normaldir ve normal distribution olarak kabul edilir.

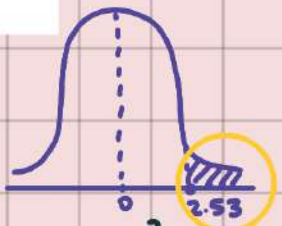
- Properties of the sampling distribution :

$\boxed{\mu_{\bar{x}} = \mu}$ and $\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$

## Example:



Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.



travelling time

$X \to \mu_x = 28$ min

$\sigma_x = 5$ min

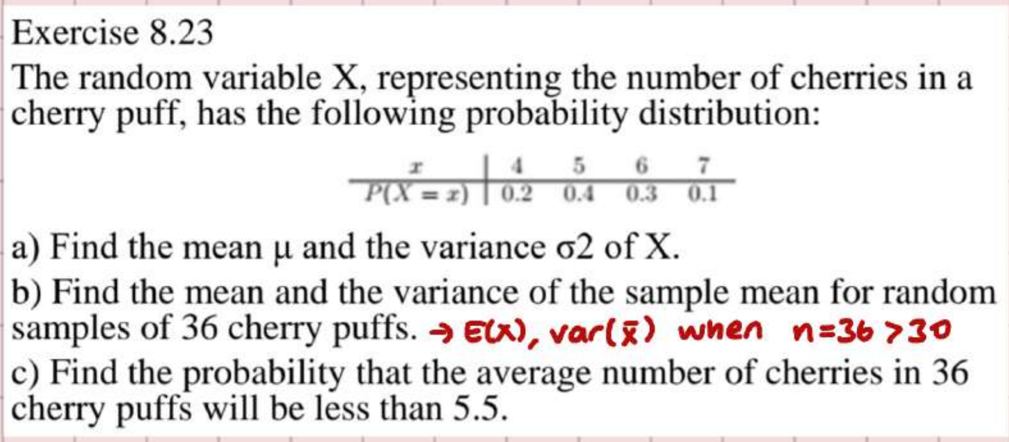
$n = 40 > 30$

$P(\bar{x} > 30) = ?$

$\bar{X} \sim N\left(\mu_x = 28, \frac{\sigma_x}{\sqrt{n}} = \frac{5}{\sqrt{40}}\right)$

$P\left\{ \frac{\bar{x}-28}{5/\sqrt{40}} > \frac{30-28}{5/\sqrt{40}} \right\} = P\{z > 2.53\} =$

## Example:



Exercise 8.23

The random variable X, representing the number of cherries in a cherry puff, has the following probability distribution:

| $x$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| $P(X = x)$ | 0.2 | 0.4 | 0.3 | 0.1 |

a) Find the mean μ and the variance σ2 of X.

b) Find the mean and the variance of the sample mean for random samples of 36 cherry puffs. → $E(\bar{x})$, $var(\bar{x})$ when $n = 36 > 30$

c) Find the probability that the average number of cherries in 36 cherry puffs will be less than 5.5.

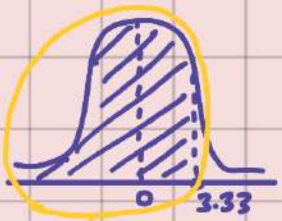
$E(x) = \sum_x x \cdot p(x)$

$Var(x) = \sum_x (x - E(x))^2 \cdot p(x)$

a) $E(x) = (0.2 \times 4) + (0.4 \times 5) + (0.3 \times 6) + (0.1 \times 7) = 5$

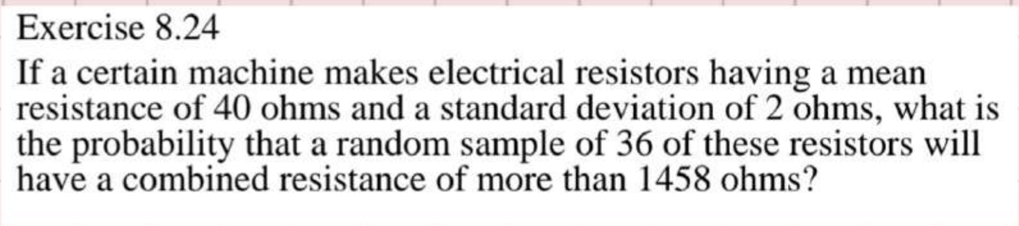
$Var(x) = (4-5)^2 \cdot (0.2) + (5-5)^2 \cdot (0.4) + (6-5)^2 \cdot (0.3) + (7-5)^2 \cdot (0.1) = 0.9$

b) $E(x) = \mu_{\bar{x}} = \mu = 5$ $\quad Var(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.9}{\sqrt{36}} = 0.15$



c) $P(\bar{x} < 5.5) = ? \to P\left\{ \frac{\bar{x}-5}{0.9/\sqrt{36}} < \frac{5.5-5}{0.15} \right\} = P\{z < 3.33\} =$

## Example:



Exercise 8.24

If a certain machine makes electrical resistors having a mean resistance of 40 ohms and a standard deviation of 2 ohms, what is the probability that a random sample of 36 of these resistors will have a combined resistance of more than 1458 ohms?

$\mu_x = 40$ $$P\left\{ \frac{\bar{x}-40}{2/\sqrt{36}} > \frac{40.5-40}{0.33} \right\}$$

$\sigma_x = 2$

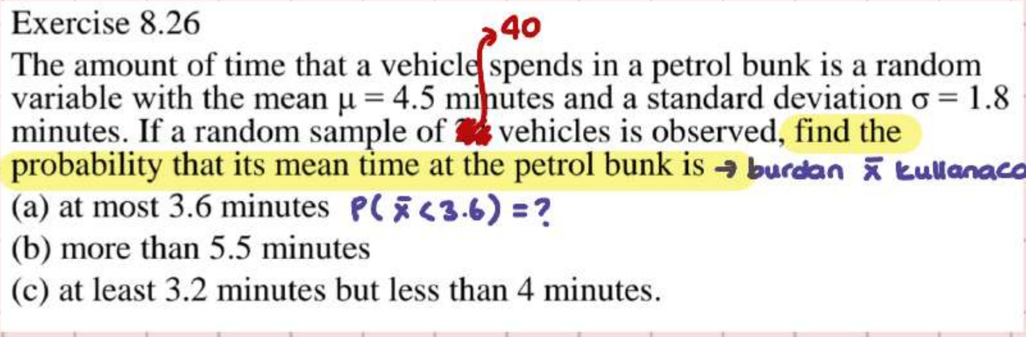
$n = 36$

$P(\bar{x} > \boxed{\frac{1458}{36}})$

↳ ohm başına

$$P\{ z > 1.5 \} = 0.066$$

**Example:**



Exercise 8.26

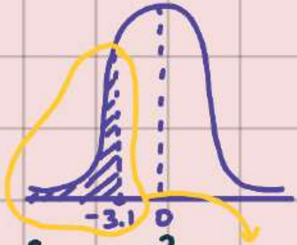
The amount of time that a vehicle spends in a petrol bunk is a random variable with the mean μ = 4.5 minutes and a standard deviation σ = 1.8 minutes. If a random sample of 40 vehicles is observed, find the probability that its mean time at the petrol bunk is → burdan $\bar{x}$ kullanacağımızı anladık

(a) at most 3.6 minutes $P(\bar{x} < 3.6) = ?$

(b) more than 5.5 minutes

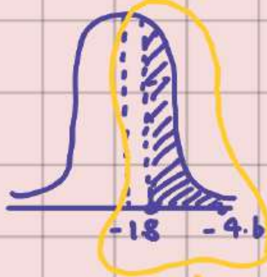
(c) at least 3.2 minutes but less than 4 minutes.

a) $\bar{x} \sim \text{Normal}\left( \mu_{\bar{x}} = 4.5 ,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.8}{\sqrt{40}} \right)$

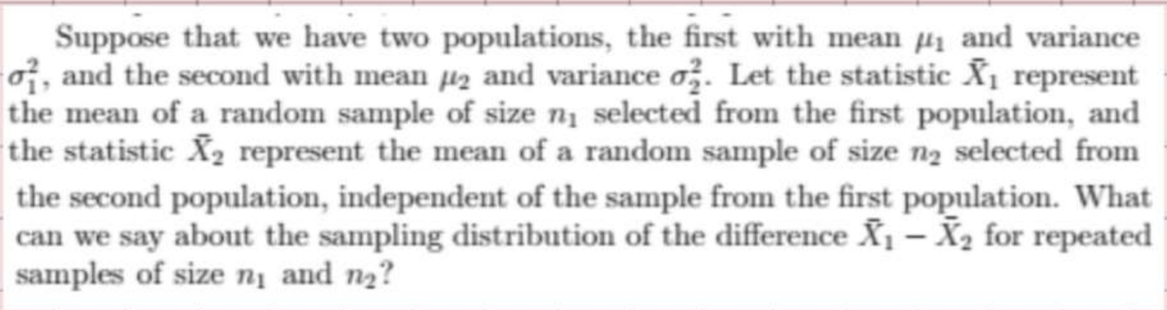


$$P\{ \bar{x} < 3.6 \} = P\left\{ \overbrace{\frac{\bar{x}-4.5}{1.8/\sqrt{40}}}^{z} < \frac{3.6-4.5}{1.8/\sqrt{40}} \right\} = P\{ z < -3.1 \} = 1 - P\{ z < 3.1 \} \approx 0$$

b) $$P(\bar{x} > 5.5) = P\left( \frac{5.5-4.5}{1.8/\sqrt{40}} > \frac{1}{0.28} \right) = P(z > 3.51) =$$



c) $$P(3.2 < \bar{x} < 4) = P\left( \frac{3.2-4.5}{1.8/\sqrt{40}} < z < \frac{4-4.5}{1.8/\sqrt{40}} \right) = P(-4.6 < z < -1.8) =$$



Suppose that we have two populations, the first with mean $\mu_1$ and variance $\sigma_1^2$, and the second with mean $\mu_2$ and variance $\sigma_2^2$. Let the statistic $\bar{X}_1$ represent the mean of a random sample of size $n_1$ selected from the first population, and the statistic $\bar{X}_2$ represent the mean of a random sample of size $n_2$ selected from the second population, independent of the sample from the first population. What can we say about the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ for repeated samples of size $n_1$ and $n_2$?

$\bar{x}_1 \to N\left(\mu, \frac{\sigma_1}{}\right.$ $\Big\}$ $\bar{x}_1 - \bar{x}_2$ : yine Normal Distribution olur.

$$\left( \quad \sqrt{n_1} \right) \rightarrow n_1 > 30$$

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

$$\underbrace{Var(ax+b)}\ = a^2 . Var(x) \rightarrow var(b) = constant = 0$$

$$\bar{X}_2 \rightarrow N\left(\mu, \frac{\sigma_2}{\sqrt{n_2}}\right) \rightarrow n_2 > 30$$

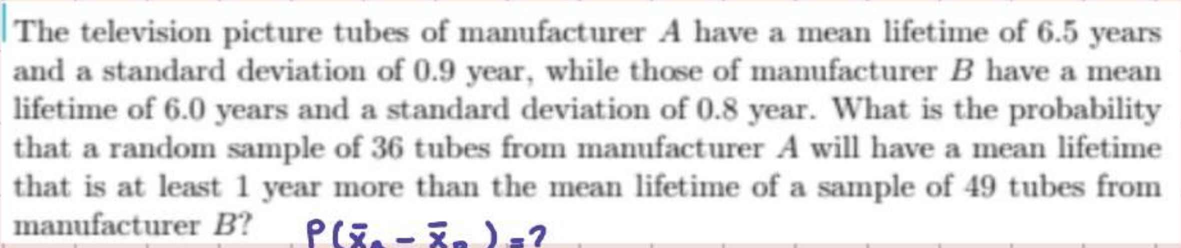
$$Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

! $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$$\boxed{z = \frac{\overset{\text{new } \bar{x}}{(\bar{x}_1 - \bar{x}_2)} - \overset{\text{new average}}{(\mu_1 - \mu_2)}}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}}$$

is approximately a standard normal variable.

**Example:**



The television picture tubes of manufacturer *A* have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer *B* have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer *A* will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer *B*?

$P(\bar{x}_A - \bar{x}_B) = ?$

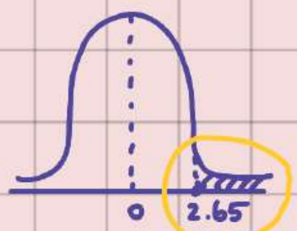
| pop 1 | pop 2 |
|---|---|
| $\mu_A = 6.5$ | $\mu_B = 6.0$ |
| $\sigma_A = 0.9$ | $\sigma_B = 0.8$ |
| $n_A = 36$ | $n_B = 49$ |

$$\bar{X}_A - \bar{X}_B \sim Normal\left( \mu_{A-B} = 0.5 \;,\; \sigma^2_{A-B} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} \right)$$

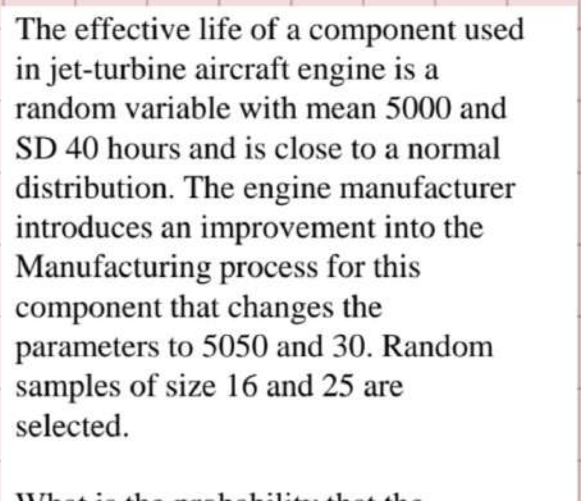
$\mu_{A-B} = 0.5 \rightarrow \mu_A - \mu_B = mean$

$\rightarrow 0.189$ (variance of $\bar{x}_A - \bar{x}_B$)

$$P\left\{ \frac{(\bar{x}_A - \bar{x}_B) - 0.5}{0.189} \geq \frac{1 - 0.5}{0.189} \right\} = P\{z \geq 2.65\} = 1 - 0.9960 = \underline{\underline{0.004}}$$



**Example:**



The effective life of a component used in jet-turbine aircraft engine is a random variable with mean 5000 and SD 40 hours and is close to a normal distribution. The engine manufacturer introduces an improvement into the Manufacturing process for this component that changes the parameters to 5050 and 30. Random samples of size 16 and 25 are selected.

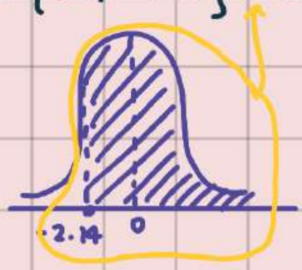
What is the probability that the

| | |
|---|---|
| $\mu_A = 5000$ | $\mu_B = 5050$ |
| $\sigma_A = 40$ | $\sigma_B = 30$ |
| $n_A = 16$ | $n_B = 25$ |

$$\bar{X} \sim Normal\left( \mu_{B-A} = 50 \;,\; \sigma^2_{B-A} = \sqrt{\frac{30^2}{25} + \frac{40^2}{16}} \right) \rightarrow 11.66$$

What is the probability that the difference in the two sample means is at least 25 hours? $P(\bar{x} \geqslant 25) = ?$

$$P\left\{ \frac{\bar{x}-50}{11.66} \geqslant \frac{25-50}{11.66} \right\} = P\{z \geqslant -2.14\} = 0.9840$$



→ **Sampling Distributions of Sample Proportions:**

- $\boxed{P = \text{the proportion of the population having some characteristic}}$

- Sample proportion ($\hat{p}$): provides an estimate of P;

$$\boxed{\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{Sample size}}}$$

- $0 \leq \hat{p} \leq 1$
- $\hat{p}$ has a binomial distribution, but can be approximated by a normal distribution when $\boxed{n.P.(1-P) > 5}$ → böyleyse normale yaklaşır

$\bar{X} \sim \text{Normal}(\mu, \sigma)$

$\mu$ → np  $\sigma$ → $\sigma^2 = np(1-p)$

$\bar{X} \sim \text{Binomial}(n, p)$ → Bunları yerlerine yazınca binomial olan dağılım normale yaklaşır.

n → # of trials

p → success probability

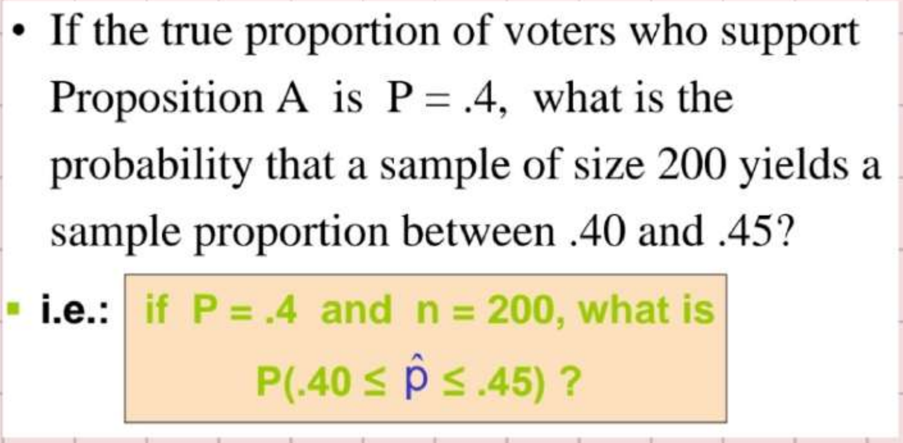
$$\boxed{E(\hat{p}) = P} \qquad \boxed{\sigma^2_{\hat{p}} = Var\left(\frac{X}{n}\right) = \frac{P.(1-P)}{n}}$$ where P = population proportion

- Standardize $\hat{p}$ to a z value with the formula:

$$\boxed{z = \frac{\hat{p}-P}{\sigma_{\hat{p}}} = \frac{\hat{p}-P}{\sqrt{\frac{P(1-P)}{n}}}}$$ P → expected proportion

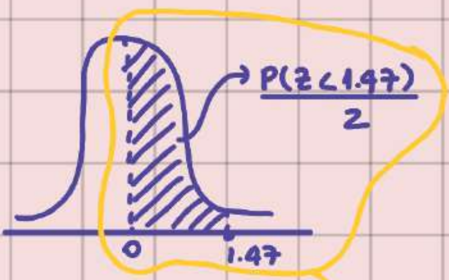
Example:

Example:



- If the true proportion of voters who support Proposition A is P = .4, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?

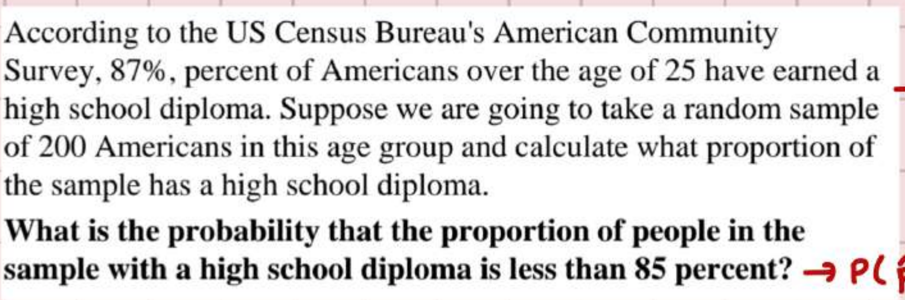
- i.e.: if P = .4 and n = 200, what is $P(.40 \le \hat{p} \le .45)$ ?



$\to \frac{P(z<1.47)}{2}$

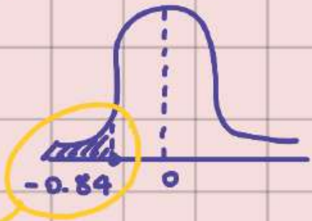
$$P\left\{\frac{0.4-0.4}{\sqrt{\frac{0.24}{200}}} \le \frac{\hat{p}-P}{\sqrt{\frac{P\times(1-P)}{n}}} \le \frac{0.45-0.4}{\sqrt{\frac{0.24}{200}}}\right\} = P\{0 \le z \le 1.47\} =$$

## Example:



According to the US Census Bureau's American Community Survey, 87%, percent of Americans over the age of 25 have earned a high school diploma. Suppose we are going to take a random sample of 200 Americans in this age group and calculate what proportion of the sample has a high school diploma. → $P = 0.87$ , $n=200$

**What is the probability that the proportion of people in the sample with a high school diploma is less than 85 percent?** → $P(\hat{p} < 0.85) = ?$



$$P\left\{\frac{\hat{p}-P}{\sqrt{\frac{P(1-P)}{n}}} < \frac{0.85-0.87}{\sqrt{\frac{0.87\times 0.13}{200}}}\right\} = P(\hat{p} < -0.84) =$$

## * Two Population Proportions:

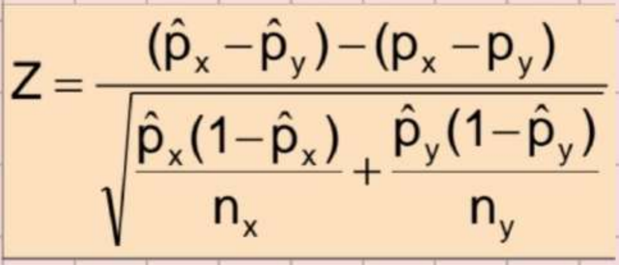
→ **Goal:** For the difference between two population properties;

$$\boxed{P_x - P_y}$$

→ **Assumptions:** Both sample sizes are large;

$$\boxed{n.P.(1-P) > 5}$$

* The random variable:



$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}$$

is approximately normally distributed.

$$z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}$$

Where ;

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

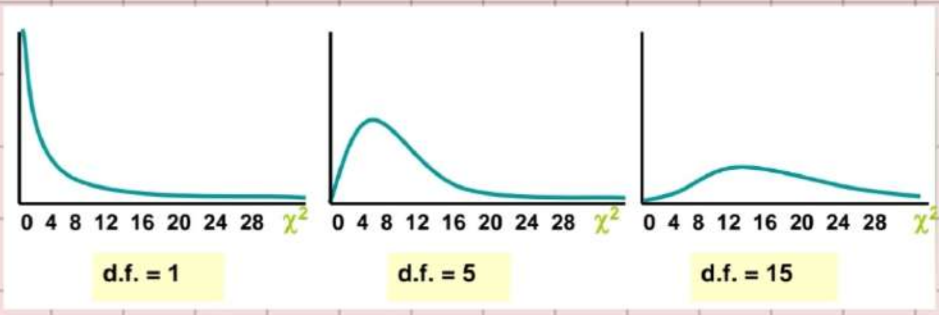
→ CHI-SQUARED DISTRIBUTION:

→ If $S^2$ is the variance of a random sample of size n taken from a normal population having the variance $\sigma^2$, then the statistic;

$$\boxed{\underset{chi}{\chi}^2 = \frac{(n-1).S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\sigma^2}}$$

has a chi-squared distribution with $\vartheta = n-1$ degrees of freedom.

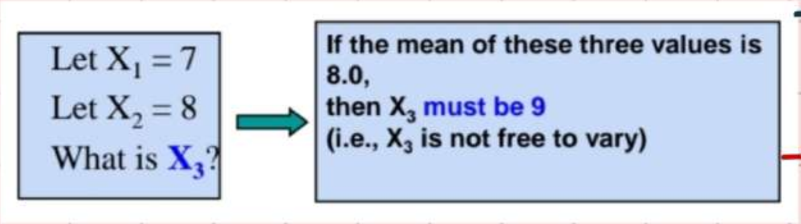
* The chi-square distribution is a family of distributions, depending on degrees of freedom: $\boxed{d.f. = \underset{\hookrightarrow \text{sample size}}{n-1}}$



→ Degrees of Freedom (df): (Serbestlik Derecesi)

* Number of observations that are free to vary after sample mean has been calculated.

Example: Suppose the mean of 3 numbers is 8.0



Let $X_1 = 7$
Let $X_2 = 8$
What is $X_3$?

If the mean of these three values is 8.0,
then $X_3$ must be 9
(i.e., $X_3$ is not free to vary)

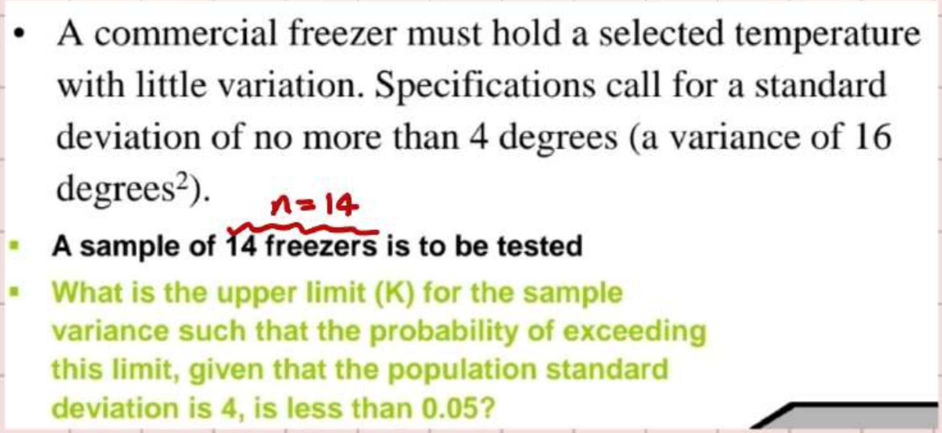
→ Here, n=3, so degrees of freedom = $= n-1 = 3-1 = 2$

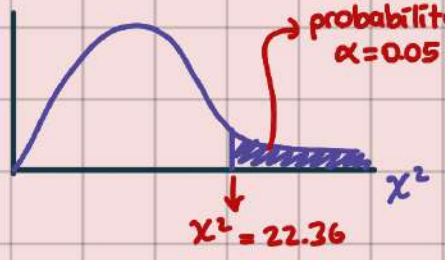
→ (2 values can be any numbers, but the third is not free to vary for a given mean)

* $X_1$ ve $X_2$ herhangi bir sayı oldu ama $X_3$'ü ortalamaya göre bulduk. $X_3$ rastgele değişken olmamış oldu. (Her bir estimation'da 1 df kaybediyoruz)

değişken bulunmuş olur. (Her bir estimation'da 1 df kaybediyoruz)
↳ o yüzden n-1

## Example


- A commercial freezer must hold a selected temperature with little variation. Specifications call for a standard deviation of no more than 4 degrees (a variance of 16 degrees$^2$).
- A sample of ~~14~~ freezers is to be tested n=14
- What is the upper limit (K) for the sample variance such that the probability of exceeding this limit, given that the population standard deviation is 4, is less than 0.05?

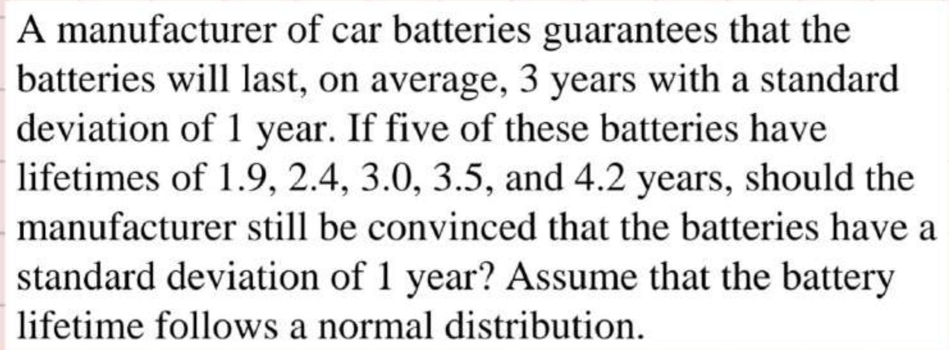


$$P(s^2 > K) = P\left(\frac{(n-1)s^2}{16} > \chi^2_{13}\right) = 0.05$$

$$\frac{(n-1)K}{16} = 22.36 \longrightarrow K = \frac{(22.36).(16)}{(14-1)} = 27.52$$

If $s^2$ from the sample of size n=14 is greater than 27.52, there is strong evidence to suggest the population variance exceeds 16.

## Example


A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

↗ bu değer soruda verilir.
confidence level = α = 0,05
(bu soru için standard devision'un 1 olması isteniyor. α değeri bu isteğin sağlanmama olasılığıdır.) (yani standard devision'un 1'den büyük ve küçük olduğu durum)

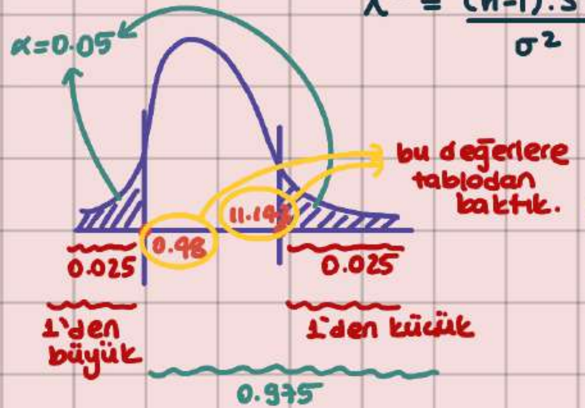
n=5

μ=3 year

σ = 1 year

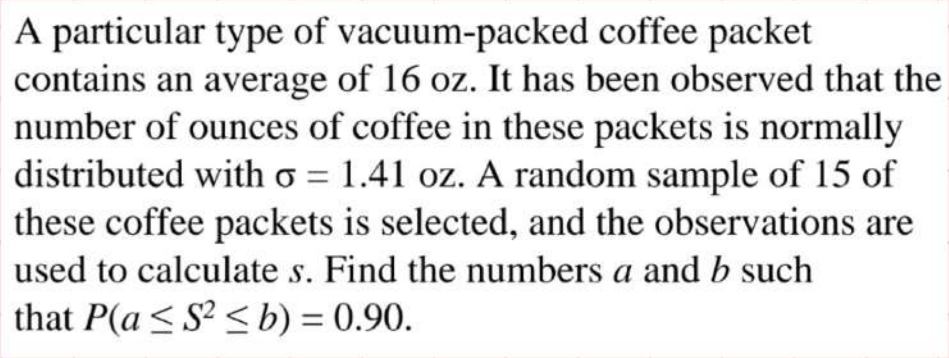
$$s^2 = \frac{\sum_{n=1}^{5}(x_i - \bar{x})^2}{n-1} = 0.851$$

$$\chi^2 = \frac{(n-1).s^2}{\sigma^2} = \frac{(5-1).(0.851)}{1^2} = \boxed{3.26}$$



Bulduğumuz $\chi^2$ değeri grafikte istenen aralıktaysa, standard devision'u 1 olur. yani bu değerin standard devision'u 95% olasılıkla 1'dir.

## Example:

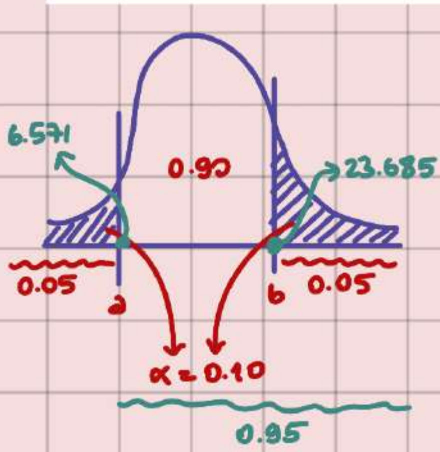


A particular type of vacuum-packed coffee packet contains an average of 16 oz. It has been observed that the number of ounces of coffee in these packets is normally distributed with σ = 1.41 oz. A random sample of 15 of these coffee packets is selected, and the observations are used to calculate *s*. Find the numbers *a* and *b* such that $P(a \le S^2 \le b) = 0.90$.

$\mu = 16$

$\sigma = 1.41$

$n = 15$



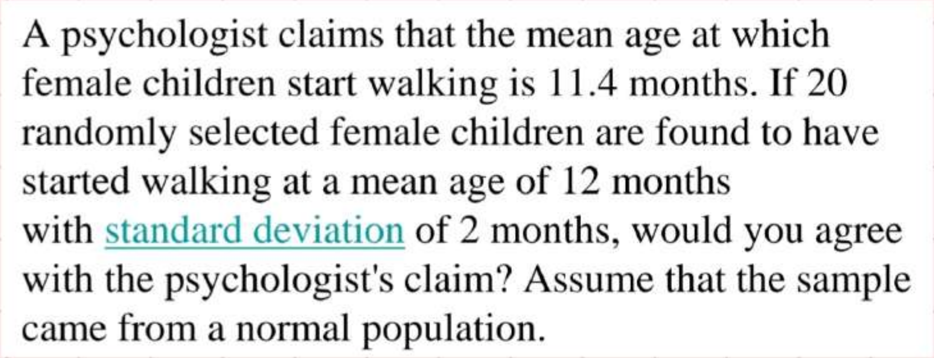
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$s_a^2 = \frac{\overset{6.571}{\chi_a^2} \cdot \sigma^2}{n-1} = 0.933$$

$$s_b^2 = \frac{\overset{23.685}{\chi_b^2} \cdot \sigma^2}{n-1} = 3.363$$

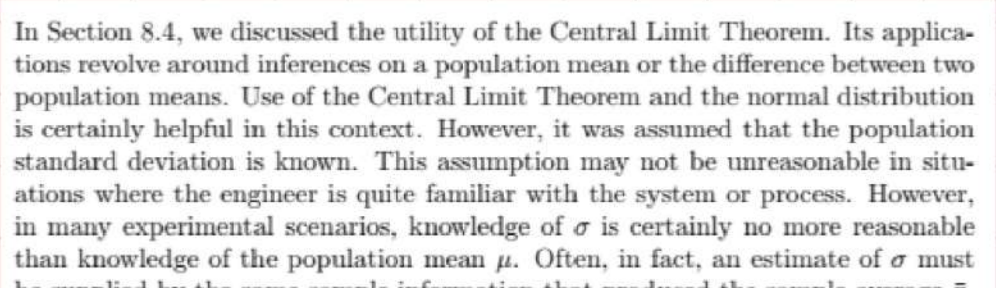
$$P(0.93 \le s^2 \le 3.36) = 0.90$$

Example:

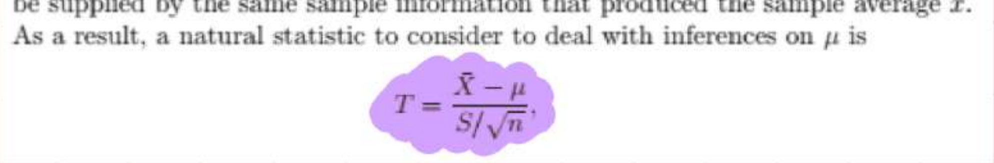


A psychologist claims that the mean age at which female children start walking is 11.4 months. If 20 randomly selected female children are found to have started walking at a mean age of 12 months with standard deviation of 2 months, would you agree with the psychologist's claim? Assume that the sample came from a normal population.

→ t-Distribution:

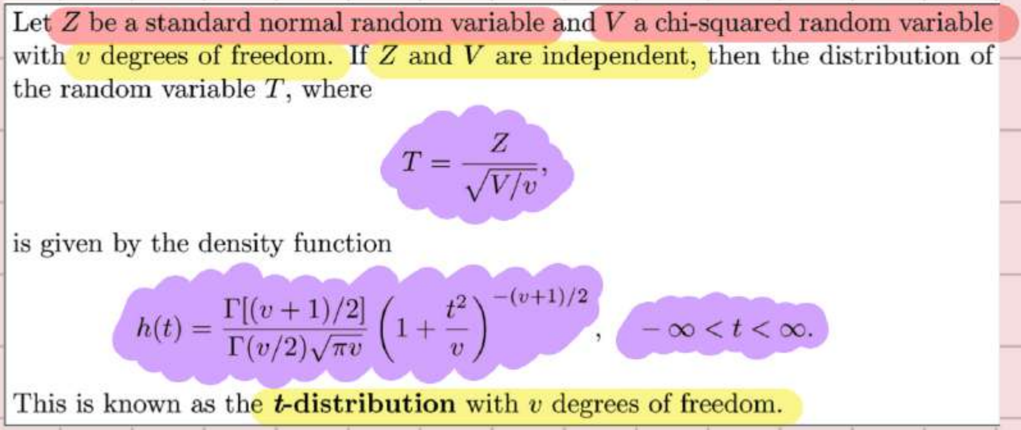


In Section 8.4, we discussed the utility of the Central Limit Theorem. Its applications revolve around inferences on a population mean or the difference between two population means. Use of the Central Limit Theorem and the normal distribution is certainly helpful in this context. However, it was assumed that the population standard deviation is known. This assumption may not be unreasonable in situations where the engineer is quite familiar with the system or process. However, in many experimental scenarios, knowledge of $\sigma$ is certainly no more reasonable than knowledge of the population mean $\mu$. Often, in fact, an estimate of $\sigma$ must be supplied by the same sample information that produced the sample average $\bar{x}$.



be supplied by the same sample information that produced the sample average $\bar{x}$. As a result, a natural statistic to consider to deal with inferences on $\mu$ is

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



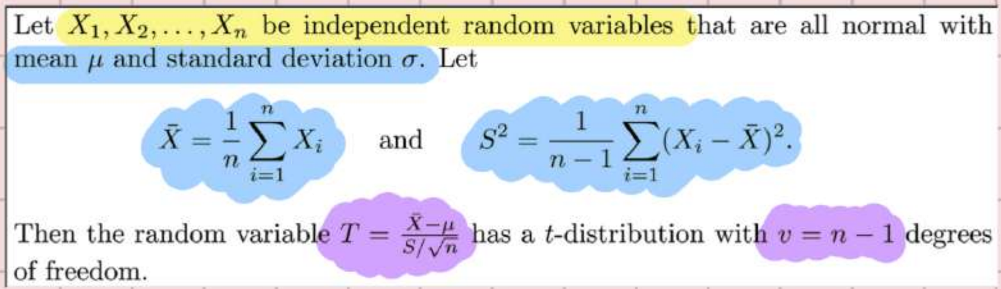
Let $Z$ be a standard normal random variable and $V$ a chi-squared random variable with $v$ degrees of freedom. If $Z$ and $V$ are independent, then the distribution of the random variable $T$, where

$$T = \frac{Z}{\sqrt{V/v}},$$

is given by the density function

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

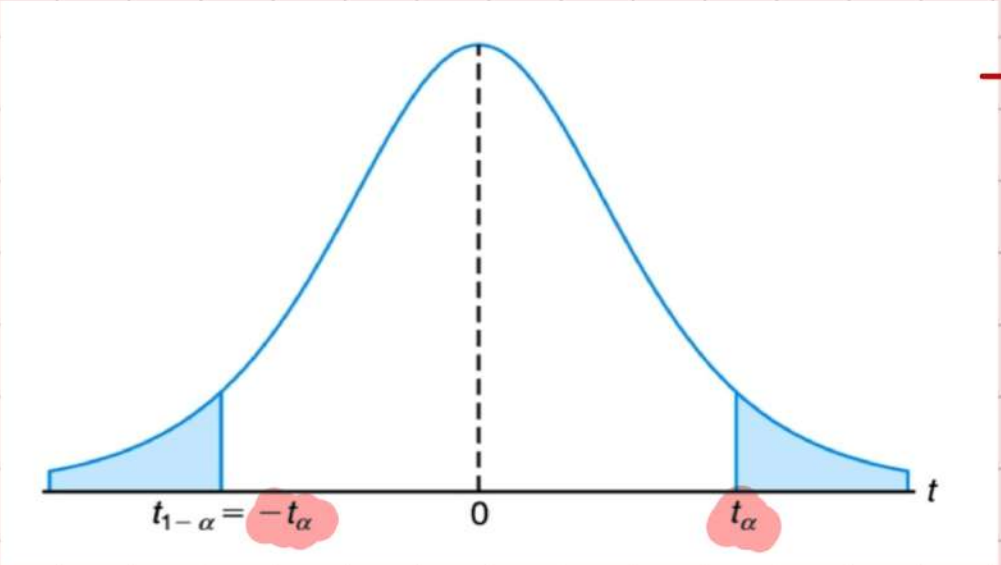
This is known as the **$t$-distribution** with $v$ degrees of freedom.



Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean $\mu$ and standard deviation $\sigma$. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

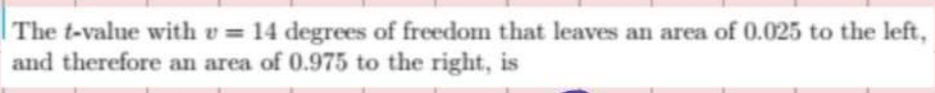
Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a $t$-distribution with $v = n - 1$ degrees of freedom.



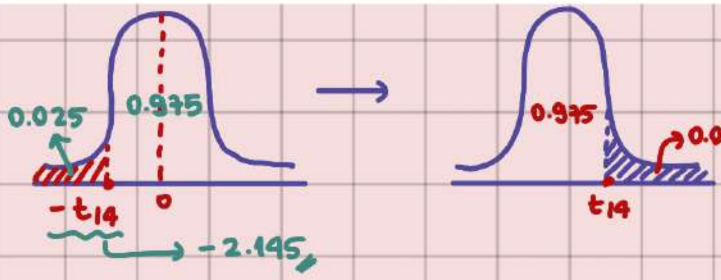
→ t - Distribution simetriktir.

**Example :**

a)



The $t$-value with $v = 14$ degrees of freedom that leaves an area of 0.025 to the left, and therefore an area of 0.975 to the right, is



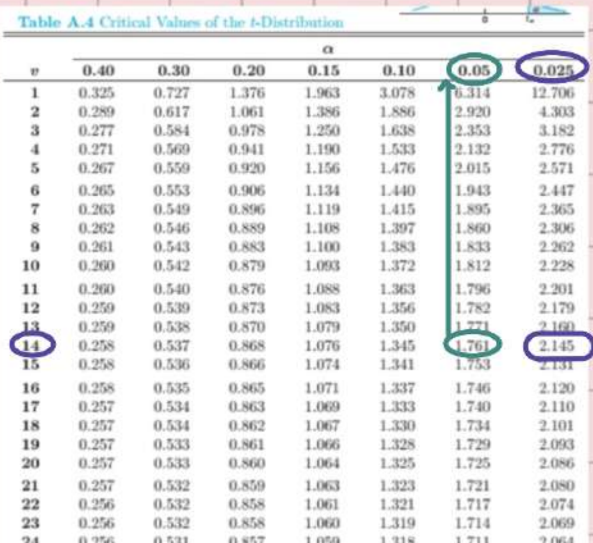
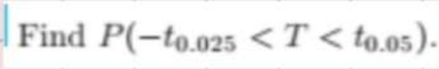


Table A.4 Critical Values of the $t$-Distribution

| | $\alpha$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $v$ | 0.40 | 0.30 | 0.20 | 0.15 | 0.10 | 0.05 | 0.025 |
| 1 | 0.325 | 0.727 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 |
| 2 | 0.289 | 0.617 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 |
| 3 | 0.277 | 0.584 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 |
| 4 | 0.271 | 0.569 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 |
| 5 | 0.267 | 0.559 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 |
| 6 | 0.265 | 0.553 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 |
| 7 | 0.263 | 0.549 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 |
| 8 | 0.262 | 0.546 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 |
| 9 | 0.261 | 0.543 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 |
| 10 | 0.260 | 0.542 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 |
| 11 | 0.260 | 0.540 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 |
| 12 | 0.259 | 0.539 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 |
| 13 | 0.259 | 0.538 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 |
| 14 | 0.258 | 0.537 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 |
| 15 | 0.258 | 0.536 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 |
| 16 | 0.258 | 0.535 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 |
| 17 | 0.257 | 0.534 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 |
| 18 | 0.257 | 0.534 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 |
| 19 | 0.257 | 0.533 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 |
| 20 | 0.257 | 0.533 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 |
| 21 | 0.257 | 0.532 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 |
| 22 | 0.256 | 0.532 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 |
| 23 | 0.256 | 0.532 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 |
| 24 | 0.256 | 0.531 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 |

b)



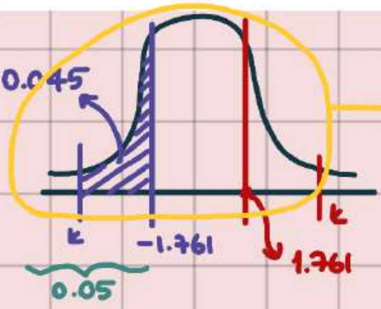
Find $P(-t_{0.025} < T < t_{0.05})$.

→ 1 - 0.05 - 0.025 =
= 0.925

$= 0.325$ //

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 25 | 0.256 | 0.531 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 |
| 26 | 0.256 | 0.531 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 |
| 27 | 0.256 | 0.531 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 |
| 28 | 0.256 | 0.530 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 |
| 29 | 0.256 | 0.530 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 |
| 30 | 0.256 | 0.530 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 |
| 40 | 0.255 | 0.529 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 |
| 60 | 0.254 | 0.527 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 |
| 120 | 0.254 | 0.526 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 |
| ∞ | 0.253 | 0.524 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 |

c) Find $k$ such that $P(k < T < -1.761) = 0.045$ for a random sample of size 15 selected from a normal distribution and $\frac{\overline{X}-\mu}{s/\sqrt{n}}$.



$1 - 0.005 = 0.995$

$\boxed{k = 2.977}$

↳ $0.05 - 0.095 = 0.005$

## Example:

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed $t$-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

iddia ettiğimiz mean

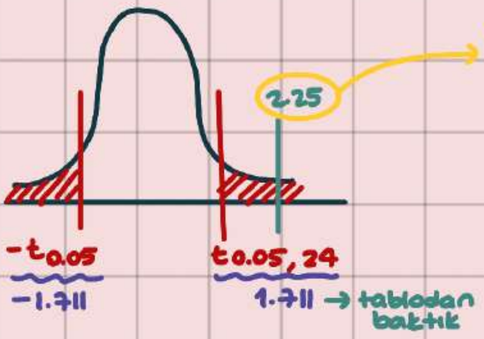
$\mu = 500$ gram

$n = 25$

$\bar{x} = 518$ gram

$S = 40$ , $\nu = 25 - 1 = 24$ degrees of freedom

$$\text{tscore} = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{518-500}{40/\sqrt{25}} = 2.25$$



taralı alan içerisinde, bu yüzden average'ı 500'den farklı.
(%10 olasılıkla 500'den fazla)
$0.05 + 0.05$

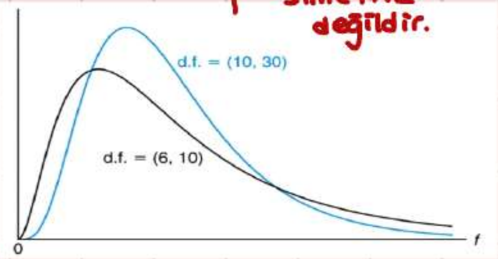
→ **F-Distribution:** two variance olunca kullanılır $(\hat{\sigma}_1^2 / \hat{\sigma}_2^2)$

Let $U$ and $V$ be two independent random variables having chi-squared distributions with $v_1$ and $v_2$ degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density function

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \le 0. \end{cases}$$

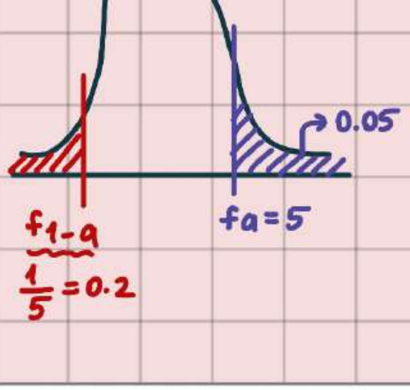
This is known as the **F-distribution** with $v_1$ and $v_2$ degrees of freedom (d.f.).

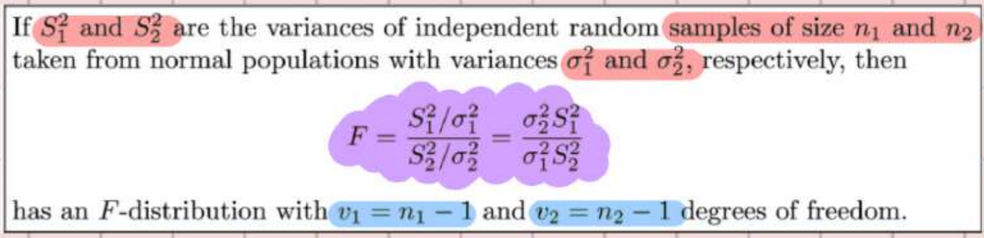
F-Distribution simetrik değildir.



Writing $f_\alpha(v_1, v_2)$ for $f_\alpha$ with $v_1$ and $v_2$ degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$



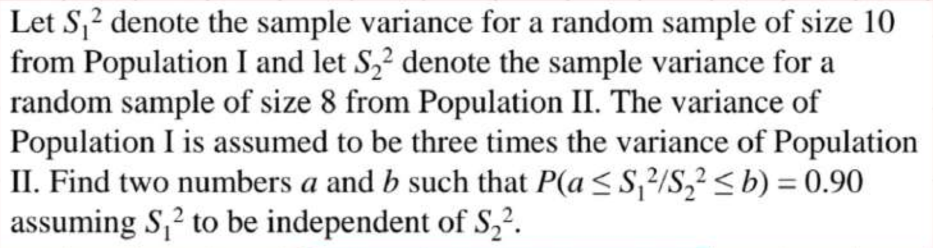


If $S_1^2$ and $S_2^2$ are the variances of independent random samples of size $n_1$ and $n_2$ taken from normal populations with variances $\sigma_1^2$ and $\sigma_2^2$, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an $F$-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

## Example:

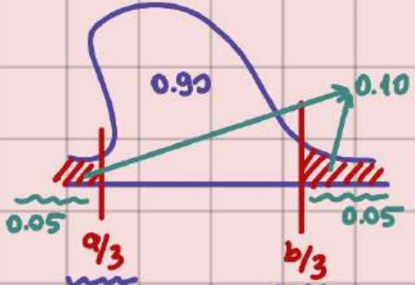


Let $S_1^2$ denote the sample variance for a random sample of size 10 from Population I and let $S_2^2$ denote the sample variance for a random sample of size 8 from Population II. The variance of Population I is assumed to be three times the variance of Population II. Find two numbers $a$ and $b$ such that $P(a \le S_1^2/S_2^2 \le b) = 0.90$ assuming $S_1^2$ to be independent of $S_2^2$.

| pop 1 | pop 2 |
|---|---|
| $S_1^2$ | $S_2^2$ |
| $n_1=10$ | $n_2=8$ |

$\sigma_1^2 = 3\sigma_2^2 \rightarrow \boxed{\frac{\sigma_1^2}{\sigma_2^2} = 3}$

$$F_{dis} \sim \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \rightarrow P\left\{\frac{a}{3} \le \frac{1}{3}\frac{S_1^2}{S_2^2} \le \frac{b}{3}\right\}$$



$$P\left\{\frac{a}{3} \le F \le \frac{b}{3}\right\}$$

$F_{\upsilon_2,\upsilon_1,1-\alpha}$

$\frac{1}{F_{\upsilon_2,\upsilon_1,\alpha}}$

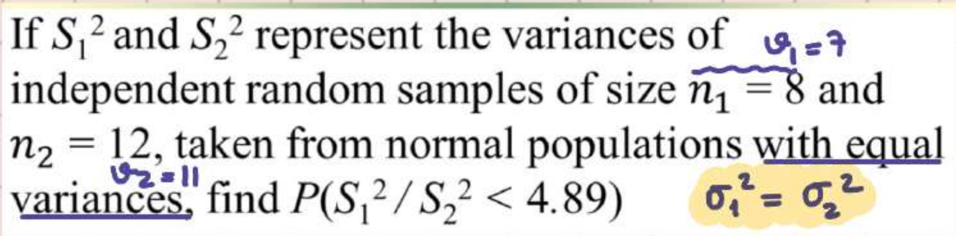
$F_{\underbrace{\upsilon_1}_{9},\underbrace{\upsilon_2}_{7},\alpha} = 3.68$

$b/3 = 3.68$

$\boxed{b = 11.04}$

$$\hookrightarrow \frac{1}{3.29} = F_{\upsilon_2,\upsilon_1,1-\alpha} = \frac{a}{3} \rightarrow \boxed{a = \frac{3}{3.29} = 0.91 /\!/}$$

## Example:



If $S_1^2$ and $S_2^2$ represent the variances of independent random samples of size $\underbrace{n_1 = 8}_{}$ and $n_2 = 12$, taken from normal populations with equal variances, find $P(S_1^2/S_2^2 < 4.89)$

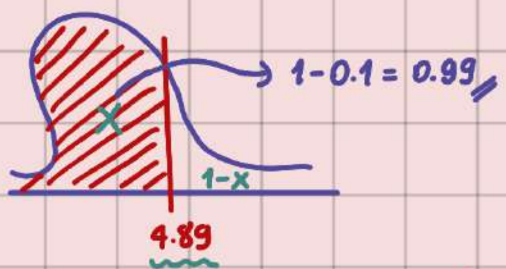
$\upsilon_1 = 7$

$\upsilon_2 = 11$

$\sigma_1^2 = \sigma_2^2$

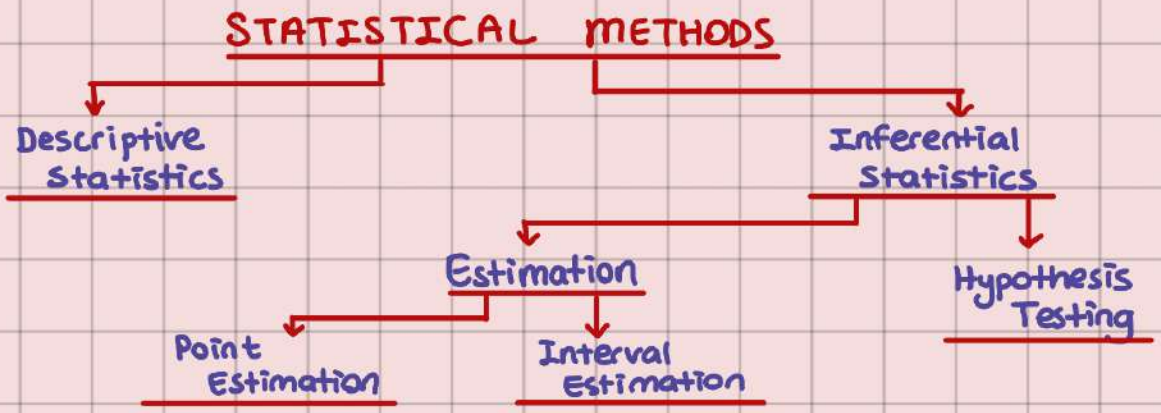
$$F_{dis} \sim \underbrace{\frac{\sigma_2^2}{\sigma_1^2}}_{1}\frac{S_1^2}{S_2^2}$$

$$P\left(1\cdot\frac{S_1^2}{S_2^2} < 4.89\cdot 1\right)$$



$$\boxed{F_{\underbrace{\vartheta_1}_{7},\underbrace{\vartheta_2}_{11},\underbrace{\alpha}_{0.01}} = 4.89}$$

# LECTURE 5



## ① POINT ESTIMATION

→ Provides single value

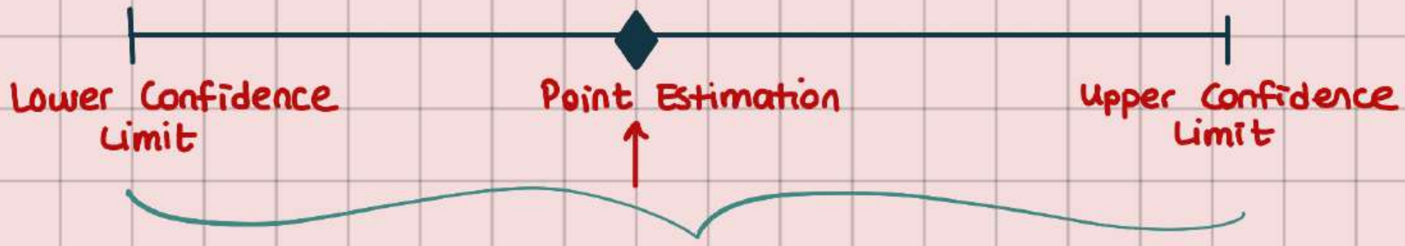
→ Dezavantajı, gerçek değere ne kadar yakın veya uzak olduğunu söyleyememiz.

→ Diyelim ki bir populasyondaki öğrencilerin GPA'lerinin average'larını öğrenmek istiyoruz. Populasyondaki herkese sorayamayacağımız için bir sample alıp onun GPA average'ını bulduk diyelim. Ve tek bir değer söyledik. Yani sample average 3 çıktı, demekki populasyon da 3 dedik. Bu durumda point estimation yapmış olduk. Yani tek bir değer verdik.

## ② INTERVAL ESTIMATION

→ Gerçek değerlere yakın değerleri söyleyebiliriz. (Avantaj)

→ Aynı örneği düşünelim. Sample average'ı 2 ve 5 arasında bulduk diyelim. Bu durumda populasyon average'ı da 2 ve 5 arasında deriz. Yani interval estimation yapmış olduk. (a-b arasında)



⊛ Interval Estimation yapabilmek için ilk olarak point estimation yaparız. Ordan bulduğumuz değere göre bir lower ve upper limit belirleriz. Ardından interval estimation yaparız.

! $\bar{x}$ (sample mean)
$\hat{p}$ (sample proportion) } Point estimation yapabilmek için $\bar{x}$ ve $\hat{p}$ değerlerini kullanırız.

→ UNBIASEDNESS :

! A statistic $\hat{\theta}$ is said to be a unbiased estimator of the parameter $\theta$ if $\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta$.

α $E(\bar{x}) = \mu$
α $E(s^2) = \sigma^2$
α $E(\hat{p}) = p$
} sample'ın sahip olduğu bir değer (mean, average...)'ın expected value'su (average)'ı populasyonun sahip olduğu o değere (mean, variance...) eşitse, orda unbiasness var demektir.

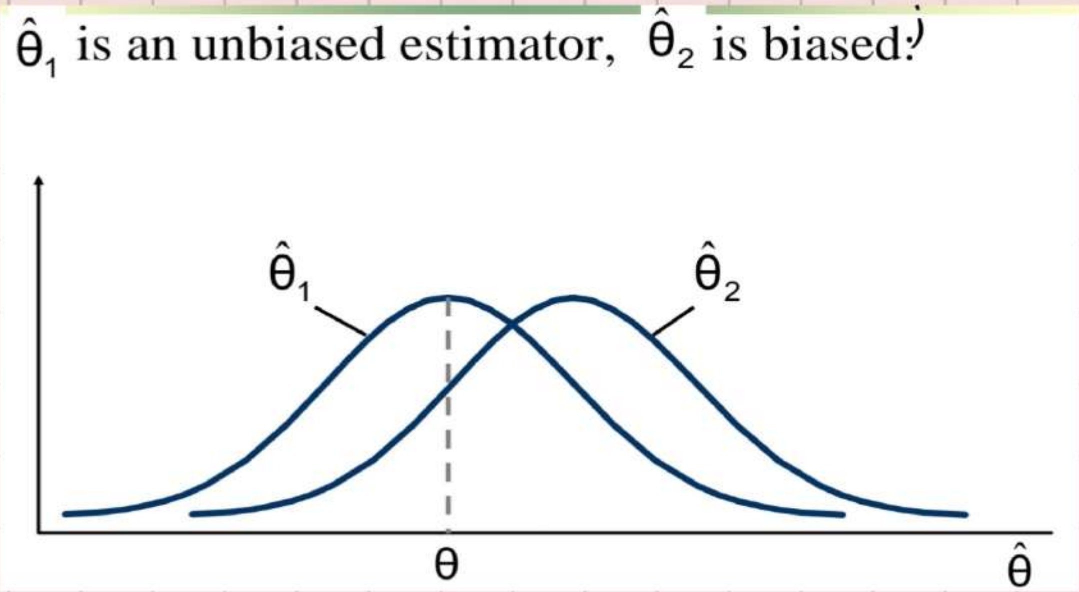
→ BIAS: (yanlılık)

$$\boxed{Bias(\hat{\theta}) = E(\hat{\theta}) - \theta}$$ → yanlılık (bias) derecesi hesaplama

⊛ Eğer estimation unbiased ise bias değeri sıfır olmuş olur. Çünkü zaten $E(\hat{\theta})$ ve $\theta$ eşit olucak. Farkları 0 olur.

- $Bias(\bar{x}) = E(\bar{x}) - \mu$

- $Bias(\bar{x}) = E(\bar{x}) - \mu$
- $Bias(s^2) = E(s^2) - \sigma^2$
- $Bias(\hat{p}) = E(\hat{p}) - p$



$\hat{\theta}_1$ is an unbiased estimator, $\hat{\theta}_2$ is biased.

→ CONSISTENCY

→ Unbiased olmadığı , bias olduğu zaman consistency olur.

→ sample size ↑, $(E(\hat{\theta}) - \theta)$ ↓ → yani sample size (n) arttıkça bias olma özelliği azalır.

→ Bu durumda $\hat{\theta}$, $\theta$'nun consistent estimator'u olur.

→ $\bar{x}$, $\mu$'nün
$s^2$, $\sigma$'nın
$\hat{p}$, $p$'nin
} consist estimator'ları olur yani.

İki estimation'umuz var diyelim $\hat{\theta}_1$ ve $\hat{\theta}_2$. Bunların mean'i $\bar{x}_1$ ve $\bar{x}_2$ diyelim ve population mean'i de $\mu$. İkisi de unbiased. Bu durumda hangisinin daha iyi olduğuna nasıl karar veririz ?

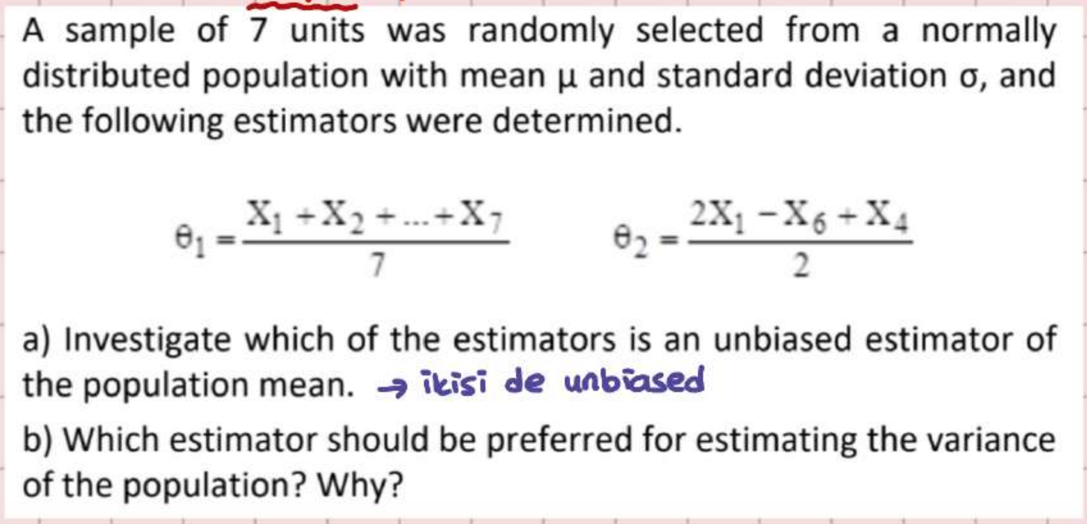
$E(\bar{x}_1) = \mu$
$E(\bar{x}_2) = \mu$

$\hat{\theta}_1$ ve $\hat{\theta}_2$'nin varyanslarına bakarız. Hangisi daha küçükse o daha efficienttir.

! $\boxed{Var(\hat{\theta}_1) < Var(\hat{\theta}_2)}$ → $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$.

! $$\boxed{\text{Relative Efficiency} = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}}$$

## Example



A sample of $\underset{\text{sample size}}{7}$ units was randomly selected from a normally distributed population with mean μ and standard deviation σ, and the following estimators were determined.

$$\theta_1 = \frac{X_1 + X_2 + \ldots + X_7}{7} \qquad \theta_2 = \frac{2X_1 - X_6 + X_4}{2}$$

a) Investigate which of the estimators is an unbiased estimator of the population mean. → ikisi de unbiased

b) Which estimator should be preferred for estimating the variance of the population? Why?

a) $E(\theta_1) \overset{?}{=} \mu$

$E(\theta_2) \overset{?}{=} \mu$

① $$E\left(\frac{x_1+x_2+\ldots+x_7}{7}\right) = \frac{1}{7}\, E(x_1+x_2+\ldots+x_7) = \mu$$

$$= \frac{1}{7}\left(\underbrace{E(x_1)}_{\mu} + \underbrace{E(x_2)}_{\mu} + \ldots + \underbrace{E(x_7)}_{\mu}\right) = \frac{7\mu}{\mu}$$

② $$E\left(\frac{2x_1 - x_6 + x_4}{2}\right) = \frac{1}{2}\, E(2x_1 - x_6 + x_4) = \mu$$

$$= \frac{1}{2}\left(2\underbrace{E(x_1)}_{\mu} - \underbrace{E(x_6)}_{\mu} + \underbrace{E(x_4)}_{\mu}\right) = \frac{2\mu}{2\mu}$$

b) $Var(ax+b) = a^2 \cdot Var(x)$

$$Var(\theta_1) = Var\left(\frac{x_1+x_2+\ldots+x_7}{7}\right) = \frac{1}{49}\cdot Var(\underbrace{x_1}_{\sigma^2}+x_2+\ldots+x_7)$$

$$Var(\theta_1) = \frac{1}{49}\cdot 7\sigma^2 = \frac{\sigma^2}{7}$$

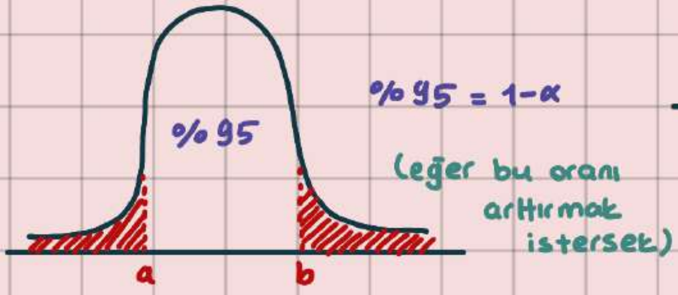
$$Var(\theta_2) = Var\left(\frac{2x_1 - x_6 + x_4}{2}\right) = \frac{1}{4}\cdot Var(2x_1 - x_6 + x_4)$$

$$= \frac{1}{4}\left(4\underbrace{Var(X_1)}_{\sigma^2} + \underbrace{Var(-X_6)}_{\sigma^2} + \underbrace{Var(X_4)}_{\sigma^2}\right) = \frac{4\sigma^2 + \sigma^2 + \sigma^2}{4} = \frac{3\sigma^2}{2}$$

$$\left.\begin{aligned} var(\theta_1) &= \sigma^2/7 \\ var(\theta_2) &= 3\sigma^2/2 \end{aligned}\right\}$$ $\theta_1$ daha kücük

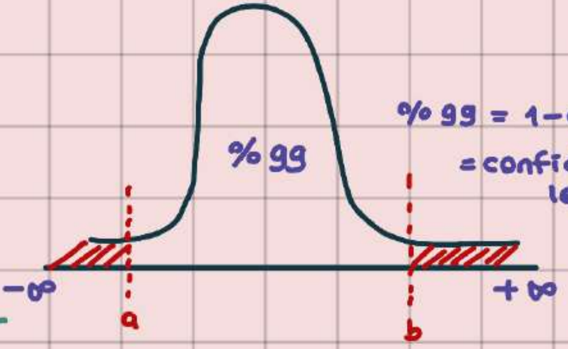
θ daha iyi

## → CONFIDENCE INTERVALS



%95 = 1−α

(eğer bu oranı arttırmak istersek)



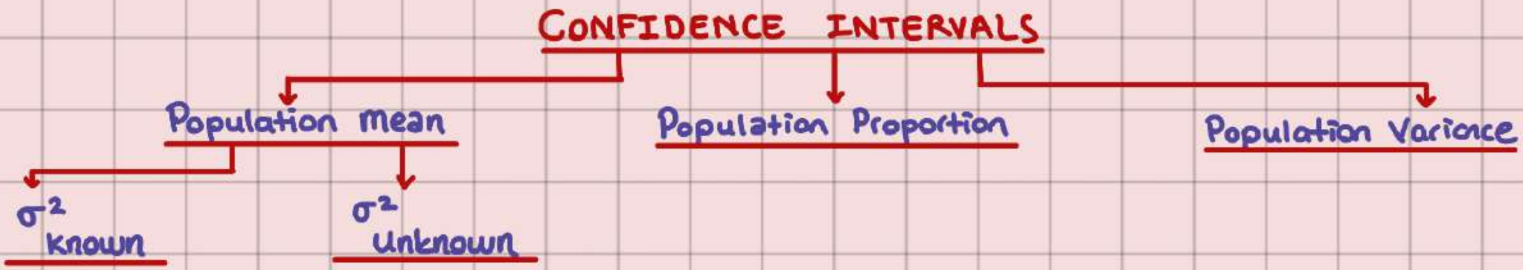
%99 = 1−α
= confidence level

- * Suppose confidence level = 95%
- * Also written (1−α) = 0.95

↳ 100% olması icin −∞'dan +∞'a gitmesi gerekir. Ama bu durum mantıklı ve işlevsel olmaz. (α = 0 iken)

(a−b) = confidence level

↳ confidence level can never be 100% confident.



! The general formula for all confidence intervals is :

$$\boxed{\underbrace{\text{Point Estimate}}_{\text{point estimation for } \mu = \bar{x}} \mp (\text{Reliability Factor})\underbrace{(\text{Standard Error})}_{\text{standart error of the } \mu \to \sigma/\sqrt{n}}}$$

## ① POPULATION MEAN ( $\sigma^2$ KNOWN)

- Population variance $\sigma^2$ is known.
- Population is normally distributed.
- If population is not normal, use large sample.

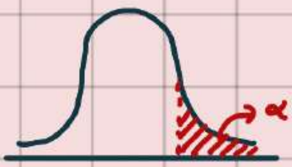
→ **Confidence Interval Estimate**

$$\boxed{\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

Relaibility Factor

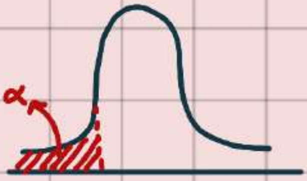
soruda bir distribution verilecek. Bu dağılımın $\bar{x}$`ını bulmak icin sınırlar bulmaya calısıcaz.

* $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ : Upper Limit



Right-sided confidence interval

* $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ : Lower Limit



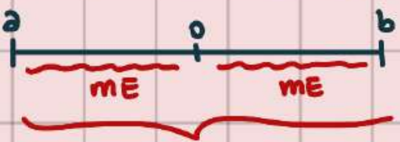
Left-sided confidence interval

! $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ → margin of error

$$\boxed{\bar{x} \mp ME}$$

$$\boxed{ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

⇒ Bu intervaller arasındaki uzaklık :



uzaklık 2ME kadar olur.

→ **The margin of error can be reduced if.**

→ the population standard deviation can be reduced ($\sigma \downarrow$)

→ The sample size is increased ($n\uparrow$)

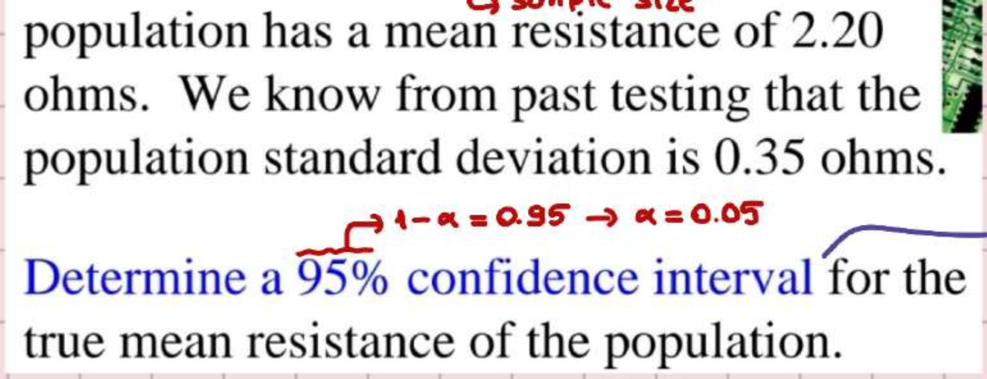
→ The confidence level is decreased $(1-\alpha)\downarrow$ $(\alpha)\uparrow$

**Example:**

→ populasyon normal dağılıyor.

A sample of 11 circuits from a large normal

→ sample size



sample size

population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.

$1-\alpha=0.95 \rightarrow \alpha=0.05$

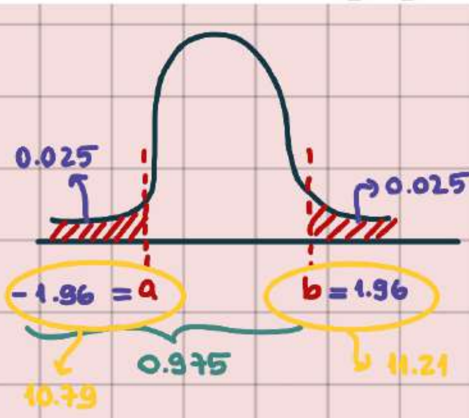
Determine a 95% confidence interval for the true mean resistance of the population.

→ two-sided confidence interval demeliydi.

$n=11$

$\bar{x}=2.20$

$\sigma=0.35$



$a=10.79$ (gerçek değer)
$a=-1.96$ (standart değer)

$b=11.21$ (gerçek değer)
$b=1.96$ (standart değer)

tablodan aldık

$$\bar{x} \mp \underbrace{z_{0.025}}\frac{\sigma}{\sqrt{n}}$$

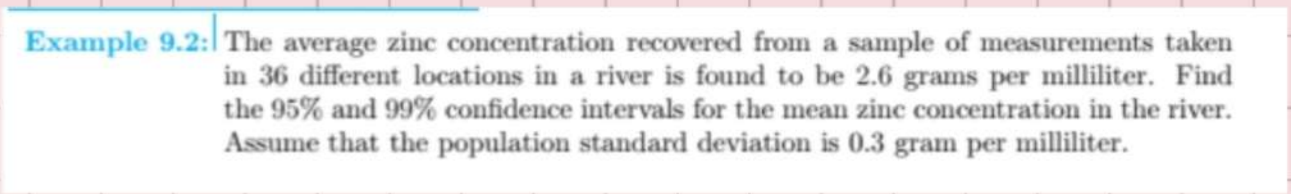
→ one-sided olsaydı $z_{0.05}$ alırdık

$$= 2.20 \mp (1.96)\frac{(0.35)}{\sqrt{11}}$$

$$= [10.79, 11.21]$$

↳ Bu aralıkta 95% olasılıkla true mean değerine sahipiz.

## Example



**Example 9.2:** The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

$\bar{x}=2.6$

$\sigma=0.3$

$n=36$

### 95 %

$\alpha=0.05$

$$\bar{x} \mp z_{0.025}\frac{\sigma}{\sqrt{n}} \rightarrow 2.6 \mp \overset{1.96}{z_{0.025}}\frac{(0.3)}{\sqrt{36}}$$

$$\underbrace{2.6-(1.96)\frac{(0.3)}{\sqrt{36}}}_{2.50} < \mu < \underbrace{2.6+(1.96)\frac{(0.3)}{\sqrt{36}}}_{2.70} \rightarrow [2.50, 2.70]$$

true population mean should be between these two numbers

### 99 %

$\alpha=0.01$

$$\bar{x} \mp z_{0.005}\frac{\sigma}{\quad} \rightarrow 2.6 \mp (2.575)(0.3)$$

%99 ihtimalle population mean'i bu aralıkta olacak.

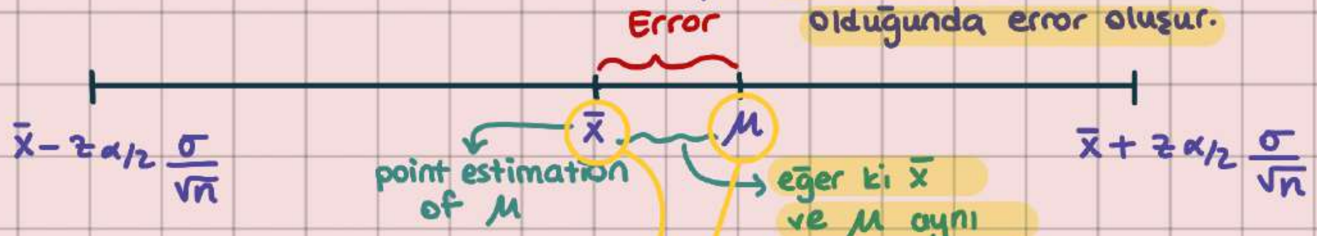
%1 ihtimalle de

$\sqrt{n}$ , $\sqrt{36}$

olmayacak. Bu duruma error denir.

$$2.6-\underbrace{(2.575)\frac{(0.3)}{\sqrt{36}}}_{2.47} < \mu < \underbrace{2.6+(2.575)\frac{(0.3)}{\sqrt{36}}}_{2.73} \rightarrow [2.47, 2.73]$$

## * Error in Estimating $\mu$ by $\bar{x}$

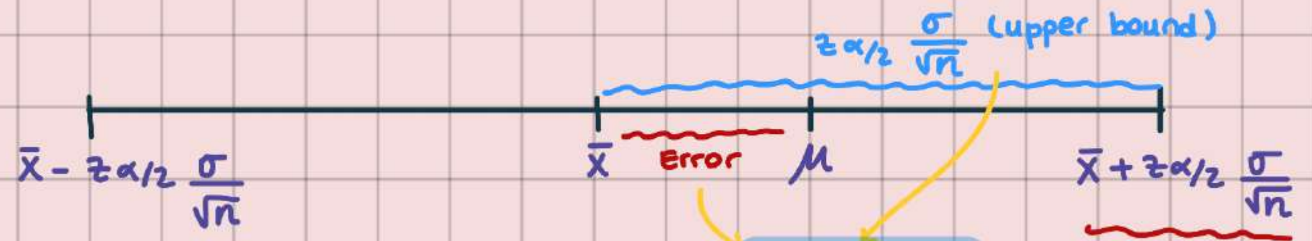


Error
$\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
$\mu$
$\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
point estimation of $\mu$

$\bar{x}$ ve $\mu$ farklı noktalarda olduğunda error oluşur.

eğer ki $\bar{x}$ ve $\mu$ aynı pointte olursa error sıfır olur.

Bu iki değeri matchlemeye çalışırız. ve bunu yaptıkça da sample size'a ulaşırız. (error'ü limitlemeye çalışırız)



$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (upper bound)
$\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
Error
$\mu$
$\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Bunu error ile match etmeye çalışırız.

The error should not exceed these two point. So;

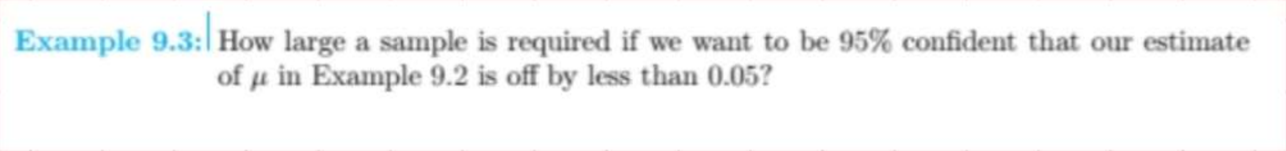
$$\left(\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)-\bar{x}$$

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

$$\boxed{\begin{array}{l} \text{Error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \sqrt{n} = z_{\alpha/2}\frac{\sigma}{E} \rightarrow n = \left(z_{\alpha/2}\frac{\sigma}{e}\right)^2 \end{array}}$$

error'ü upper bound'ın altında tutarak limitlemeye çalışırız. Ve bunu yaparken de sample size ile oynarız. Ve gün sonunda sample size'a ulaşırız.

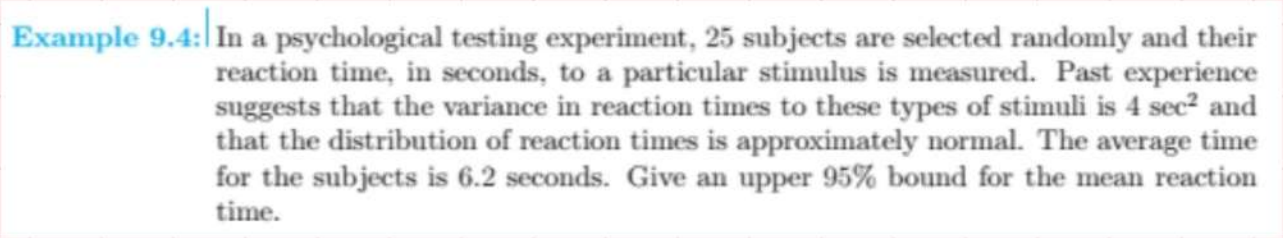
Example



Example 9.3: How large a sample is required if we want to be 95% confident that our estimate of $\mu$ in Example 9.2 is off by less than 0.05?

$$\left(\sqrt{n} = z_{\alpha/2}\,\frac{\sigma}{e}\right)^2$$

## Example



Example 9.4: In a psychological testing experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular stimulus is measured. Past experience suggests that the variance in reaction times to these types of stimuli is 4 sec$^2$ and that the distribution of reaction times is approximately normal. The average time for the subjects is 6.2 seconds. Give an upper 95% bound for the mean reaction time.

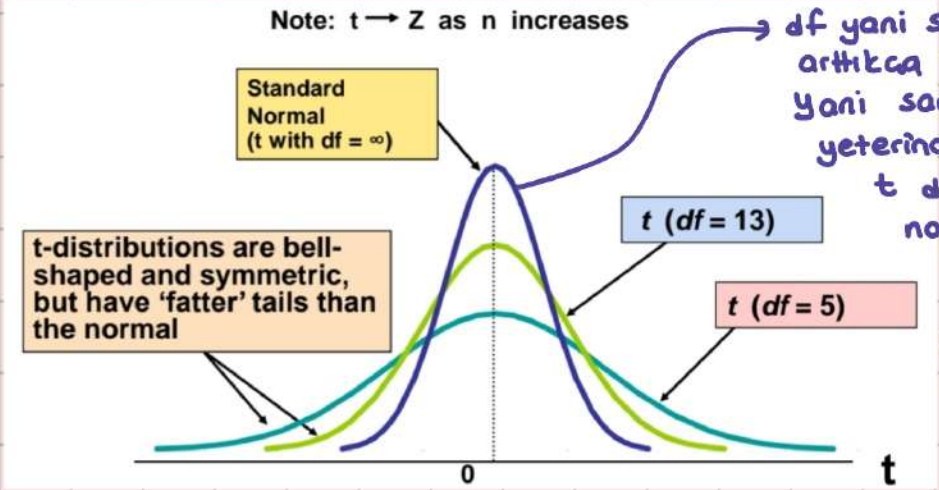
② POPULATION MEAN ( $\sigma^2$ UNKNOWN)

## ② POPULATION MEAN (σ UNKNOWN)

→ If the population standard devision σ is unknown, we can substitute the sample standard devision, s → we can use s because s unbias point estimator of σ

→ So we use the t distribution instead of the normal distribution.

↳ degrees of freedom → $\boxed{df = n-1}$



→ df yani sample size arttıkça bu şekil olur. Yani sample size'ın yeterince büyük olması t distribution'u normal distribution'a çevirir. Bu durumda hangi distribution'u kullandığımızın önemi olmaz.

α Population standard deviation is unknown.

α Population is normally distributed.

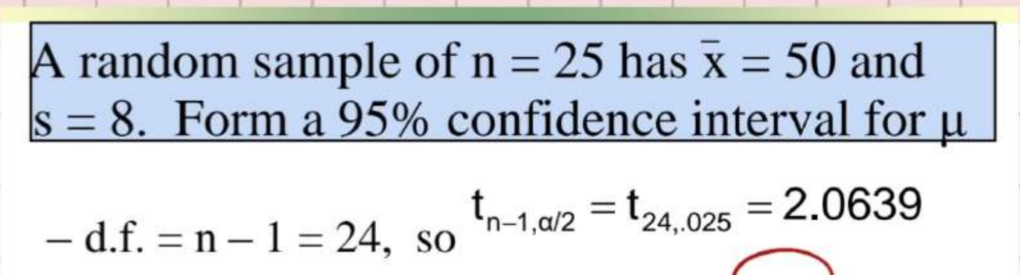
α If population is not normal, use large sample. → sample size 30'dan küçükse t kullanırız, 30'dan büyükse z veya t fark etmez.

! $\boxed{\bar{x} \mp t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}}$ → σ'yı bilmediğimiz için bunu kullandık

$$\boxed{\bar{x} \mp ME}$$

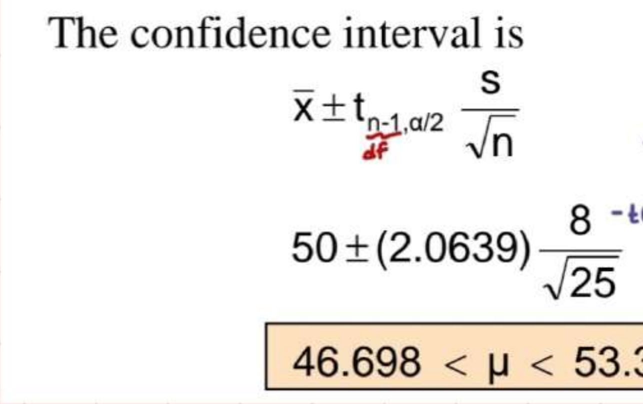
$$\boxed{ME = t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}}$$

Example



A random sample of n = 25 has $\bar{x}$ = 50 and s = 8. Form a 95% confidence interval for μ

– d.f. = n – 1 = 24, so $t_{n-1,\alpha/2} = t_{24,.025} = 2.0639$

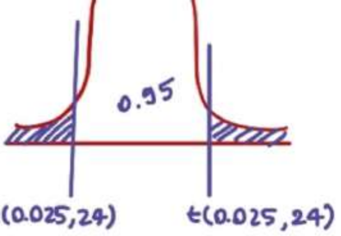


The confidence interval is

$$\bar{x} \pm t_{\underbrace{n-1}_{df},\alpha/2} \frac{s}{\sqrt{n}}$$

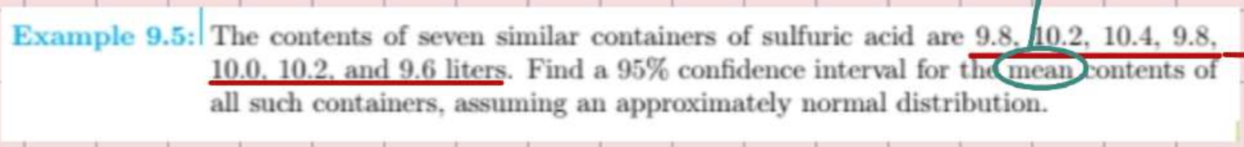
$$50 \pm (2.0639)\frac{8}{\sqrt{25}}$$

$$\boxed{46.698 < \mu < 53.302}$$



confidence interval of mean soruyor. t distribution kullanırız. Cünkü population variance (σ) bilmiyoruz. z dis. kullanamayız. Ayrıca sample size da 30'dan kücük yani kesin t kullanıcaz.

## Example



**Example 9.5:** The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

→ 7 sample

$n=7$

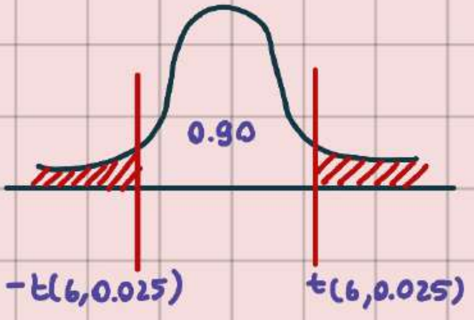
$$\bar{x} = \frac{9.8+10.2+10.4+9.8+10+10.2+9.6}{7} = 10$$

$$s = \frac{\sum_{1}^{7}(x-\bar{x})^2}{n-1} = 0.28$$

$$\bar{x} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = 10 \mp \overbrace{\underbrace{t_{n-1,\alpha/2}}_{2.447}}^{t_{6,0.025}} \frac{(0.28)}{\sqrt{7}}$$

$$= 10 \mp (2.447)\frac{(0.28)}{\sqrt{7}}$$

$$= [9.74, 10.26]$$

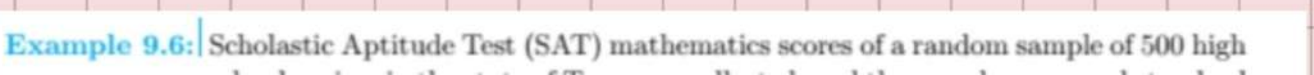


## ⊛ Concept of Large-Sample Confidence Interval

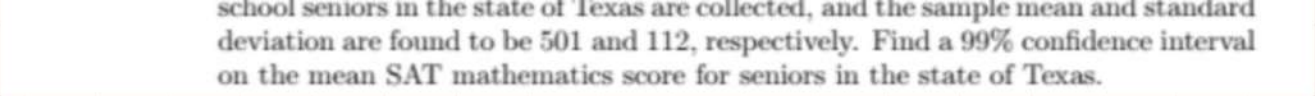
$$\boxed{\bar{x} \mp z_{\alpha/2}\frac{s}{\sqrt{n}}} \Rightarrow$$

Eğer ki σ (population variance) bilinmiyorsa ama n ≥ 30 ise t distribution'u, z'ye convert ederek kullanırız. (Bu durumda t'de kullanabiliriz ama t tablosunda büyük değerler olmadığı icin approctimately olarak z'ye cevirip kullanırız.

## Example



**Example 9.6:** Scholastic Aptitude Test (SAT) mathematics scores of a random sample of 500 high



school seniors in the state of Texas are collected, and the sample mean and standard deviation are found to be 501 and 112, respectively. Find a 99% confidence interval on the mean SAT mathematics score for seniors in the state of Texas.

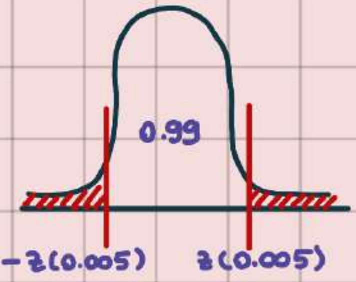
$\bar{x} = 501$

$s = 112$

$n = 500$

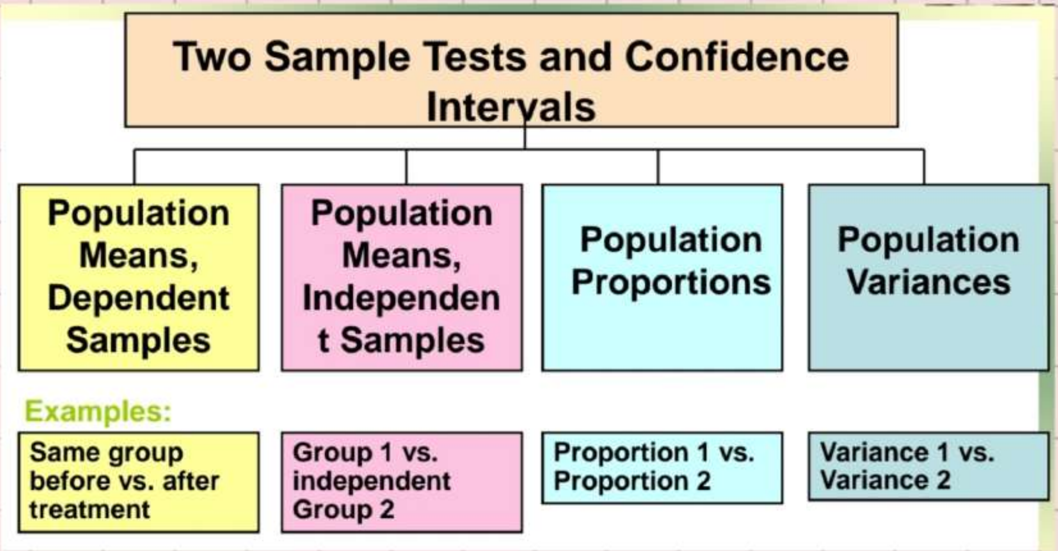
$$\bar{x} \mp z_{\alpha/2} \frac{s}{\sqrt{n}} = 501 \mp \underbrace{z_{0.005}}_{2.575} \frac{112}{\sqrt{500}}$$

$\rightarrow [488.1, 513.9]$



$\rightarrow$ $n \geq 30$ ama population variance bilinmiyor. t'yi z ile convert edip kullanırız. (t dis yine bulabiliriz ama bu kadar büyük bir değeri tabloda bulmak zor olur)

Standart Error of Point Estimation $\left\{ \frac{\sigma}{\sqrt{n}}, \frac{s}{\sqrt{n}} \right\}$ Soruda $\sigma$ ve s 'den hangisi verilirse onu kullanırız.



Two Sample Tests and Confidence Intervals

Population Means, Dependent Samples

Population Means, Independent Samples

Population Proportions

Population Variances

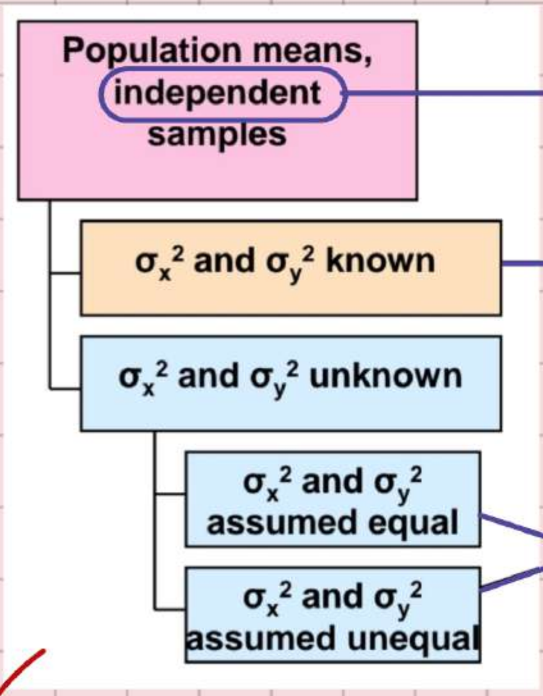
Examples:

Same group before vs. after treatment

Group 1 vs. independent Group 2

Proportion 1 vs. Proportion 2

Variance 1 vs. Variance 2



Population means, independent samples

$\sigma_x^2$ and $\sigma_y^2$ known

$\sigma_x^2$ and $\sigma_y^2$ unknown

$\sigma_x^2$ and $\sigma_y^2$ assumed equal

$\sigma_x^2$ and $\sigma_y^2$ assumed unequal

$\rightarrow$ we are using different samples/ different populations.

$\rightarrow$ z-value kullanılır.

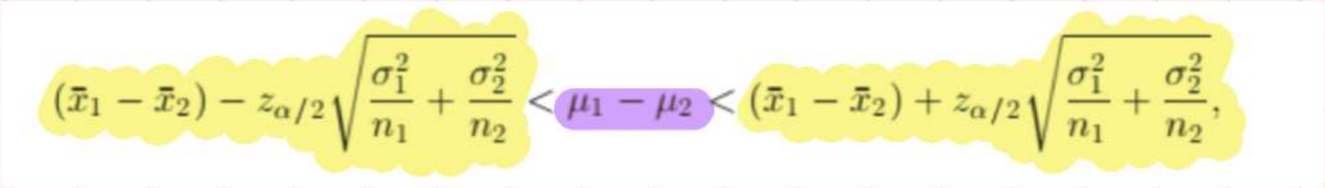
$\rightarrow$ t-distribution kullanılır.

(1) $\sigma_x^2$ and $\sigma_y^2$ known :

→ Samples are randomly and independently drawn

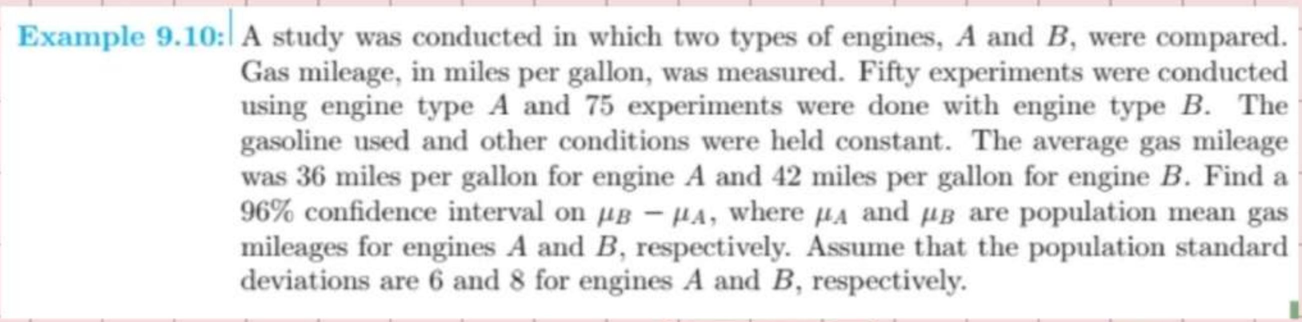
→ Both population distributions are normal

→ Population variances are known



$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

## Example



**Example 9.10:** A study was conducted in which two types of engines, $A$ and $B$, were compared. Gas mileage, in miles per gallon, was measured. Fifty experiments were conducted using engine type $A$ and 75 experiments were done with engine type $B$. The gasoline used and other conditions were held constant. The average gas mileage was 36 miles per gallon for engine $A$ and 42 miles per gallon for engine $B$. Find a 96% confidence interval on $\mu_B - \mu_A$, where $\mu_A$ and $\mu_B$ are population mean gas mileages for engines $A$ and $B$, respectively. Assume that the population standard deviations are 6 and 8 for engines $A$ and $B$, respectively.

$n_A = 50$

$n_B = 75$

$\bar{x}_A = 36$

$\bar{x}_B = 42$

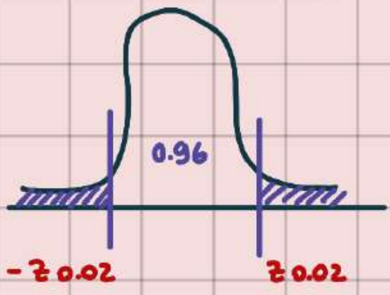
$\sigma_A = 6$

$\sigma_B = 8$

$$= (\bar{x}_B - \bar{x}_A) \mp z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

$$= (42-36) \mp \underset{2.05}{\underline{z_{0.02}}}\sqrt{\frac{36}{50} + \frac{64}{75}}$$

$$= [3.93, 8.57]$$

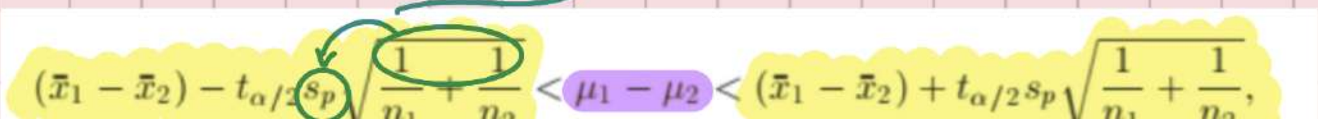


## ② $\sigma_x^2$ and $\sigma_y^2$ unknown and Equal

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown but assumed equal

|  | df | degrees of freedom |
|---|---|---|
| $n_x \longrightarrow$ | $n_x - 1$ | $n_x + n_y - 2$ |
| $n_y \longrightarrow$ | $n_y - 1$ |  |

variance'ı dışarı çıkardık.



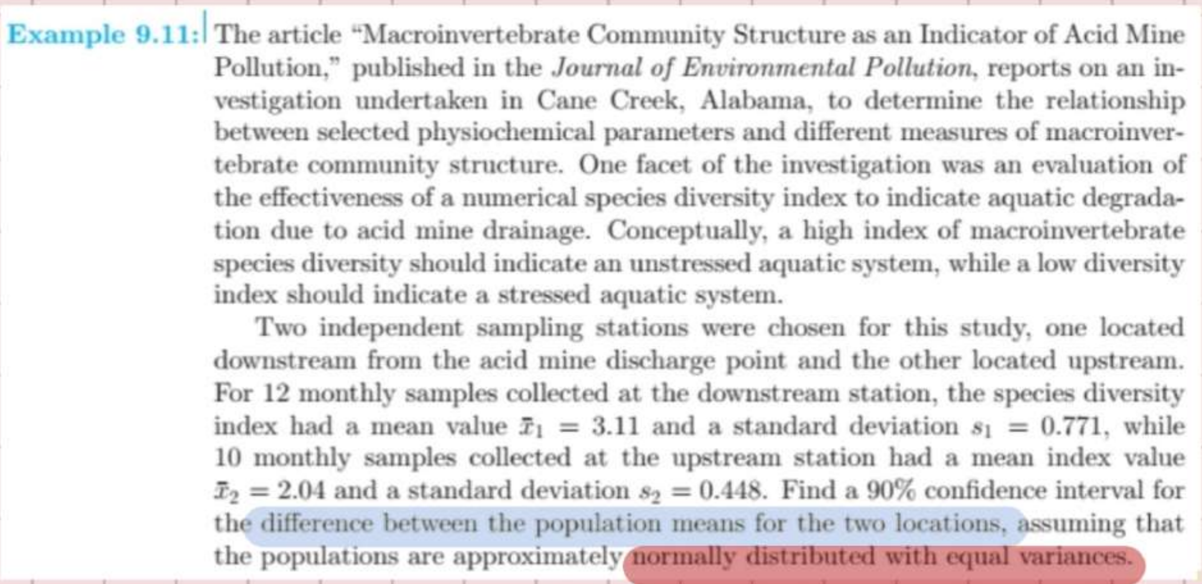
$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

$$S_p^2 = \underbrace{\frac{(n_1-1)}{n_1+n_2-2}}_{\text{weight of 1}} S_1^2 + \underbrace{\frac{(n_2-1)}{n_1+n_2-2}}_{\text{weight of 2}} S_2^2$$

→ variance`lar aynı ama weight`lerine göre hangi sample`dan ne kadar aldığımıza karar veririz.

## Example



**Example 9.11:** The article "Macroinvertebrate Community Structure as an Indicator of Acid Mine Pollution," published in the *Journal of Environmental Pollution*, reports on an investigation undertaken in Cane Creek, Alabama, to determine the relationship between selected physiochemical parameters and different measures of macroinvertebrate community structure. One facet of the investigation was an evaluation of the effectiveness of a numerical species diversity index to indicate aquatic degradation due to acid mine drainage. Conceptually, a high index of macroinvertebrate species diversity should indicate an unstressed aquatic system, while a low diversity index should indicate a stressed aquatic system.

Two independent sampling stations were chosen for this study, one located downstream from the acid mine discharge point and the other located upstream. For 12 monthly samples collected at the downstream station, the species diversity index had a mean value $\bar{x}_1 = 3.11$ and a standard deviation $s_1 = 0.771$, while 10 monthly samples collected at the upstream station had a mean index value $\bar{x}_2 = 2.04$ and a standard deviation $s_2 = 0.448$. Find a 90% confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with equal variances.

$n_1 = 12$

$n_2 = 10$

$\bar{x}_1 = 3.11$

$\bar{x}_2 = 2.04$

$s_1 = 0.771$

$s_2 = 0.448$

$\boxed{\sigma_1^2 = \sigma_2^2}$

$$1.07 \mp \underbrace{(1.725)}_{t_{0.05,20}} \underbrace{(0.646)}_{s_p} \sqrt{\frac{1}{12} + \frac{1}{10}}$$

$n_1+n_2-2$

$$0.593 \leq \mu_1 - \mu_2 \leq 1.547$$

## ③ $\sigma_x^2$ and $\sigma_y^2$ Unknown and Unequal

- Samples are randomly and independently drawn.
- Populations are normally distributed.
- Population variances are unknown and assumed unequal

! Use a t value with $\nu$ degrees of freedom, where

$$\left[\left(\frac{s_x^2}{n}\right) + \left(\frac{s_y^2}{n}\right)\right]^2$$

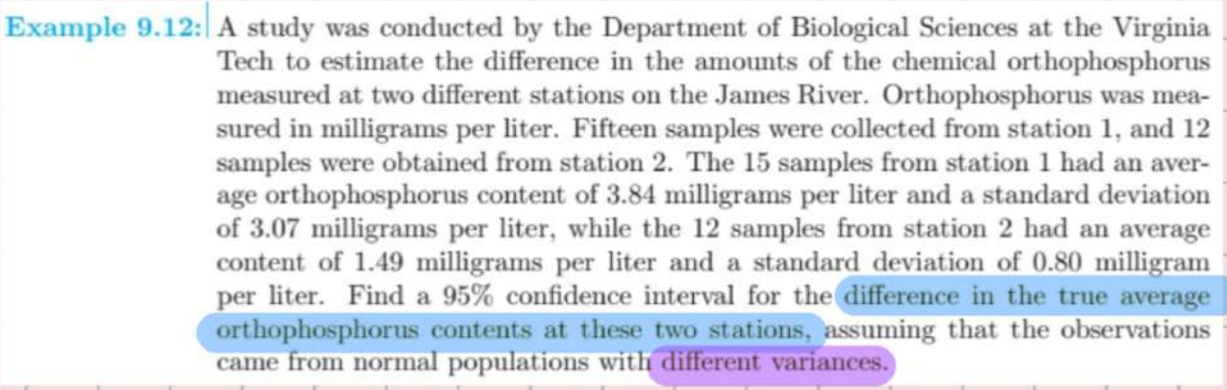
$$t = \frac{(\bar{x} - \bar{y}) - D_0}{}$$

$$v = \frac{\left[\quad n_x \qquad n_y \quad\right]}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$$

$$t = \frac{}{\sqrt{\frac{s_x^2}{n_X}+\frac{s_y^2}{n_Y}}}$$

$$(\bar{x}_1-\bar{x}_2)-t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < \mu_1-\mu_2 < (\bar{x}_1-\bar{x}_2)+t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

## Example



**Example 9.12:** A study was conducted by the Department of Biological Sciences at the Virginia Tech to estimate the difference in the amounts of the chemical orthophosphorus measured at two different stations on the James River. Orthophosphorus was measured in milligrams per liter. Fifteen samples were collected from station 1, and 12 samples were obtained from station 2. The 15 samples from station 1 had an average orthophosphorus content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average content of 1.49 milligrams per liter and a standard deviation of 0.80 milligram per liter. Find a 95% confidence interval for the difference in the true average orthophosphorus contents at these two stations, assuming that the observations came from normal populations with different variances.

$n_1 = 15$ , $n_2 = 12$

$\bar{x}_1 = 3.84$

$\bar{x}_2 = 1.49$

$s_1 = 3.07$

$s_2 = 0.80$

$\boxed{\sigma_1^2 \neq \sigma_2^2}$

$v = 16.3 \approx 16 \qquad = (\bar{x}_1-\bar{x}_2) \mp \underbrace{t_{\alpha/2, v}}_{t_{0.025,\,16}} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

$= 2.35 \mp 2.120 \sqrt{\frac{(3.07)^2}{15}+\frac{(0.80)^2}{12}} \longrightarrow \boxed{0.60 \leq \mu_1-\mu_2 \leq 4.10}$

## → PAIRED OBSERVATIONS (Dependent Samples)

$\boxed{di = x_i - y_i}$ → after observation

↓ before observation

dependent oldukları için ortak ilerleriz. di buluruz ve her işlemi di üzerinden yaparız.

$$d_i = x_i - y_i$$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

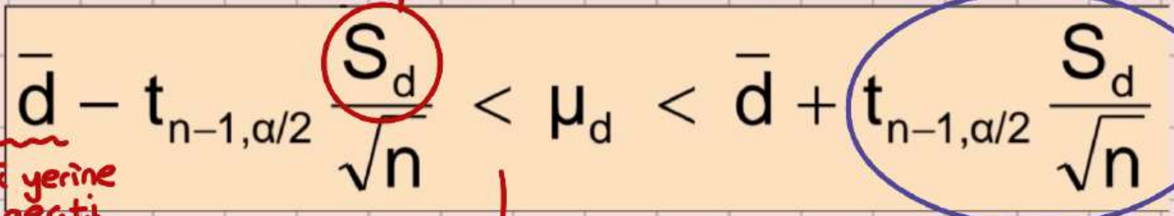
↳ average of di's

$$S_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}}$$

↳ sample standart devision of di's.

σ veya s yerine geçti

ME

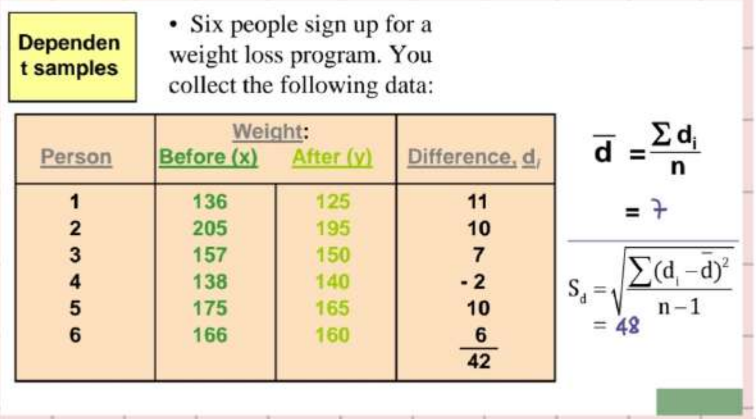


$$\overset{\bar{x}\text{ yerine geçti}}{\bar{d}} - t_{n-1,\alpha/2}\frac{\boxed{S_d}}{\sqrt{n}} < \mu_d < \bar{d} + \boxed{t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}}}$$

x̄ yerine geçti

if σ is unknown and n ≤ 30 ise böyle confidence interval buluruz.

## Example



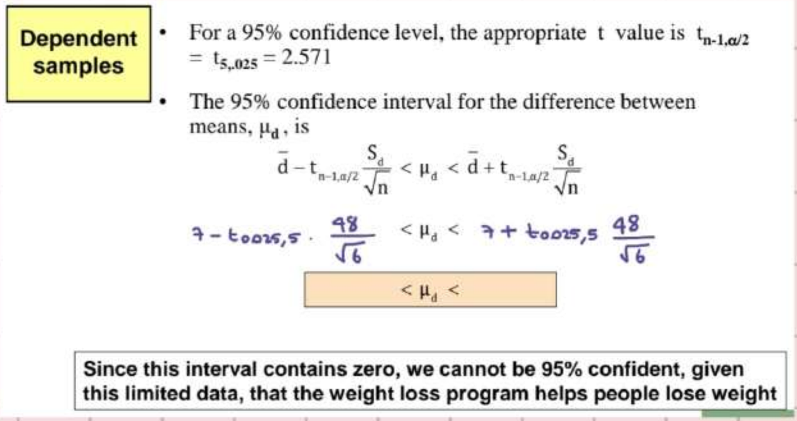
**Dependent samples**

- Six people sign up for a weight loss program. You collect the following data:

| Person | Weight: Before (x) | Weight: After (y) | Difference, $d_i$ |
|---|---|---|---|
| 1 | 136 | 125 | 11 |
| 2 | 205 | 195 | 10 |
| 3 | 157 | 150 | 7 |
| 4 | 138 | 140 | - 2 |
| 5 | 175 | 165 | 10 |
| 6 | 166 | 160 | 6 |
| | | | 42 |

$$\bar{d} = \frac{\sum d_i}{n} = 7$$

$$S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = 48$$



**Dependent samples**

- For a 95% confidence level, the appropriate t value is $t_{n-1,\alpha/2} = t_{5,.025} = 2.571$
- The 95% confidence interval for the difference between means, $\mu_d$, is

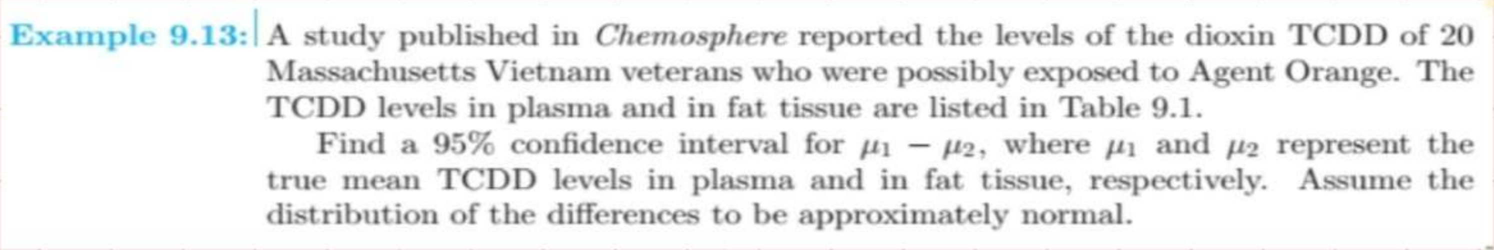
$$\bar{d} - t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}}$$

$$7 - t_{0025,5} \cdot \frac{48}{\sqrt{6}} < \mu_d < 7 + t_{0025,5}\frac{48}{\sqrt{6}}$$

$$< \mu_d <$$

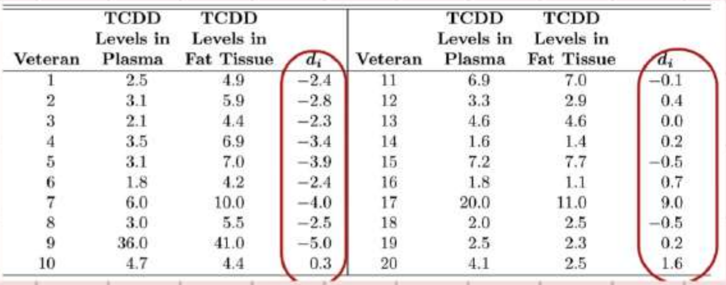
**Since this interval contains zero, we cannot be 95% confident, given this limited data, that the weight loss program helps people lose weight**

## Example



**Example 9.13:** A study published in *Chemosphere* reported the levels of the dioxin TCDD of 20 Massachusetts Vietnam veterans who were possibly exposed to Agent Orange. The TCDD levels in plasma and in fat tissue are listed in Table 9.1.

Find a 95% confidence interval for $\mu_1 - \mu_2$, where $\mu_1$ and $\mu_2$ represent the true mean TCDD levels in plasma and in fat tissue, respectively. Assume the distribution of the differences to be approximately normal.



| Veteran | TCDD Levels in Plasma | TCDD Levels in Fat Tissue | $d_i$ | Veteran | TCDD Levels in Plasma | TCDD Levels in Fat Tissue | $d_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 2.5 | 4.9 | −2.4 | 11 | 6.9 | 7.0 | −0.1 |
| 2 | 3.1 | 5.9 | −2.8 | 12 | 3.3 | 2.9 | 0.4 |
| 3 | 2.1 | 4.4 | −2.3 | 13 | 4.6 | 4.6 | 0.0 |
| 4 | 3.5 | 6.9 | −3.4 | 14 | 1.6 | 1.4 | 0.2 |
| 5 | 3.1 | 7.0 | −3.9 | 15 | 7.2 | 7.7 | −0.5 |
| 6 | 1.8 | 4.2 | −2.4 | 16 | 1.8 | 1.1 | 0.7 |
| 7 | 6.0 | 10.0 | −4.0 | 17 | 20.0 | 11.0 | 9.0 |
| 8 | 3.0 | 5.5 | −2.5 | 18 | 2.0 | 2.5 | −0.5 |
| 9 | 36.0 | 41.0 | −5.0 | 19 | 2.5 | 2.3 | 0.2 |
| 10 | 4.7 | 4.4 | 0.3 | 20 | 4.1 | 2.5 | 1.6 |

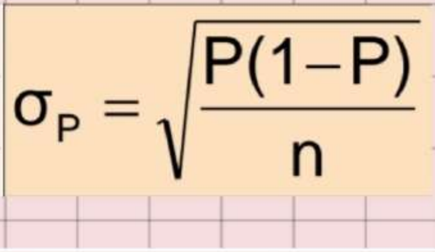
$\bar{d} = -0.87$ $\quad \bar{d} + (t_{0025,19})\frac{S_d}{\sqrt{n}}$

$S_d = 2.9773$

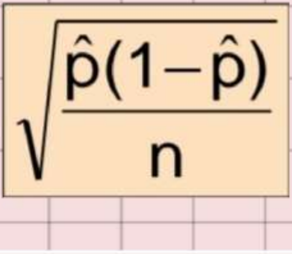
$$-0.87 \mp \underbrace{(t_{0.025,19})}_{2.09}\frac{2.9773}{\sqrt{20}}$$

$[-2.26 \leq \mu_1 - \mu_2 \leq 0.52]$

→ CONFIDENCE INTERVAL for the POPULATION PROPORTION



$$\sigma_P = \sqrt{\frac{P(1-P)}{n}}$$



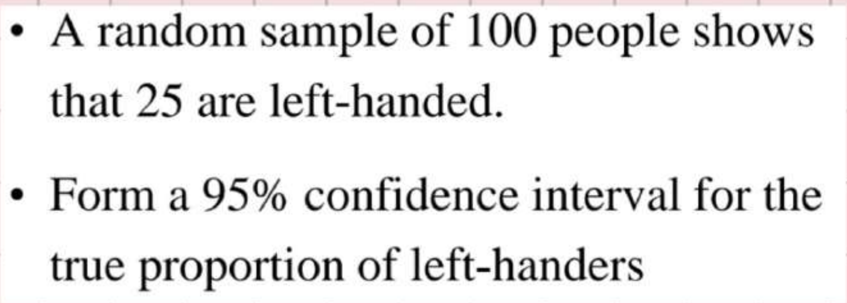
$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

→ where ;

- $z_{\alpha/2}$ is the standard normal value for the level of confidence desired.
- $\hat{p}$ is the sample proportion
- n is the sample size
- $n.P.(1-P) > 5$

## Example



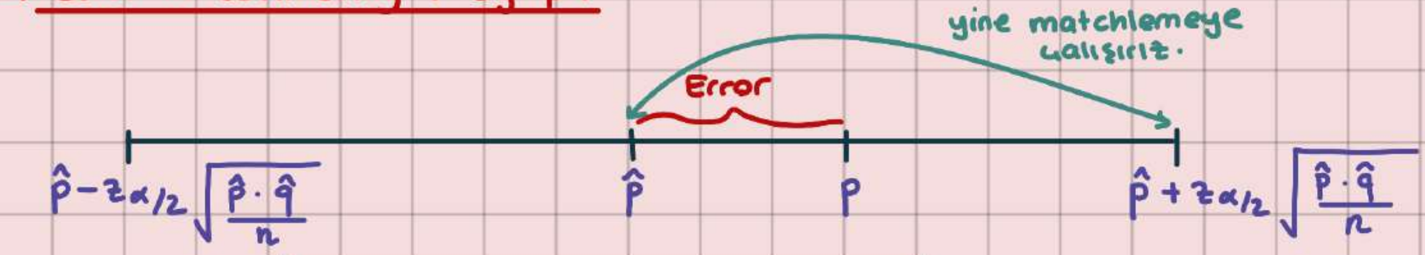
- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers

$$\left.\begin{array}{l} n=100 \\ x=25 \end{array}\right\} \hat{p} = 25/100 = 0.25$$

$$\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}} = 0.25 \mp \underset{1.96}{\underline{z_{0.025}}}\sqrt{\frac{(0.25)(0.75)}{100}}$$

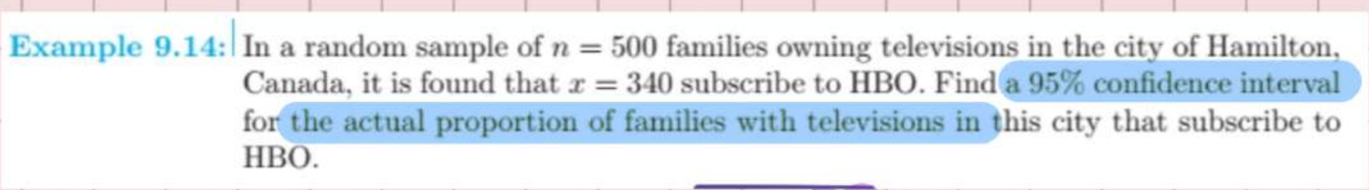
$$0.165 < P < 0.334$$

→ Error in Estimating P by $\hat{p}$ :


yine matchlemeye çalışırız.
Error
$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$
$\hat{p}$
P
$\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$

$$\boxed{\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} = \text{Error}}$$

## Example



**Example 9.14:** In a random sample of $n = 500$ families owning televisions in the city of Hamilton, Canada, it is found that $x = 340$ subscribe to HBO. Find a 95% confidence interval for the actual proportion of families with televisions in this city that subscribe to HBO.

$n = 500$

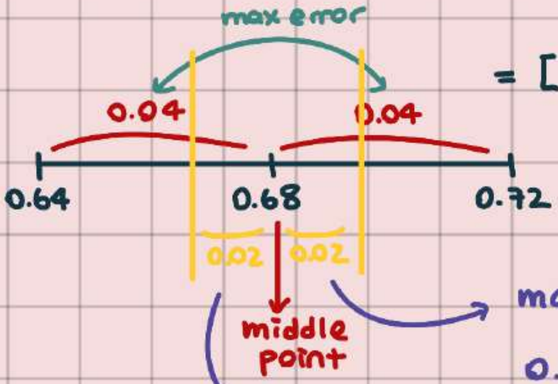
$$\hat{p} \mp z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}$$

$\sqrt{\frac{}{n}}$

$x = 340$

$\hat{p} = 340/500 = 0.68$ $\quad = 0.68 \mp \underset{1.96}{\underline{z_{0.025}}} \sqrt{\frac{(0.68)(0.32)}{500}}$

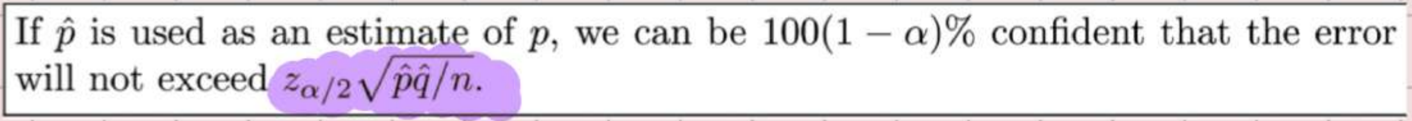
$= [0.639, 0.720]$



max errorü 0.04'den

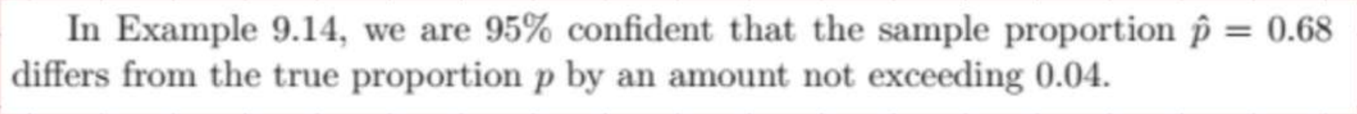
0.02'ye indirmek istiyorum.

O zaman n ile oynamam lazım.



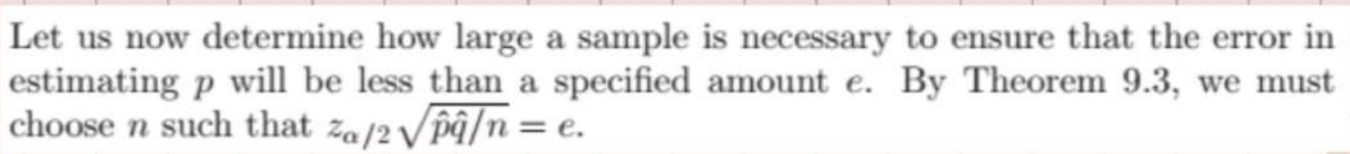
If $\hat{p}$ is used as an estimate of $p$, we can be $100(1-\alpha)\%$ confident that the error will not exceed $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

Example

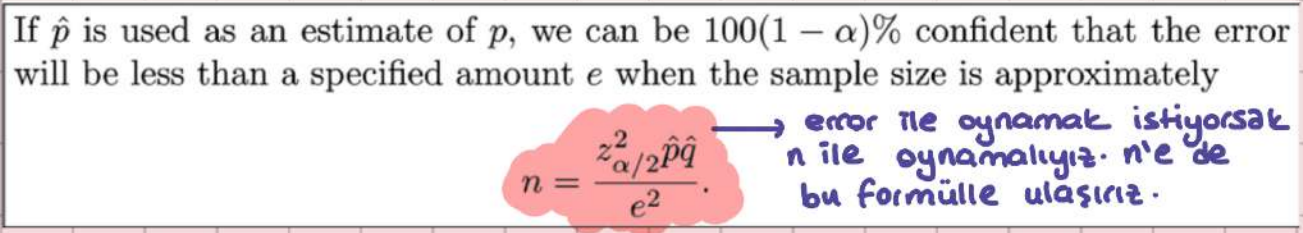


In Example 9.14, we are 95% confident that the sample proportion $\hat{p} = 0.68$ differs from the true proportion $p$ by an amount not exceeding 0.04.

Example



Let us now determine how large a sample is necessary to ensure that the error in estimating $p$ will be less than a specified amount $e$. By Theorem 9.3, we must choose $n$ such that $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} = e$.



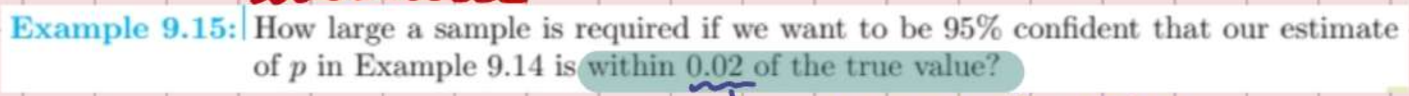
If $\hat{p}$ is used as an estimate of $p$, we can be $100(1-\alpha)\%$ confident that the error will be less than a specified amount $e$ when the sample size is approximately

$$n = \frac{z^2_{\alpha/2}\hat{p}\hat{q}}{e^2}. \longrightarrow$$

error ile oynamak istiyorsak n ile oynamalıyız. n'e de bu formülle ulaşırız.

## Example

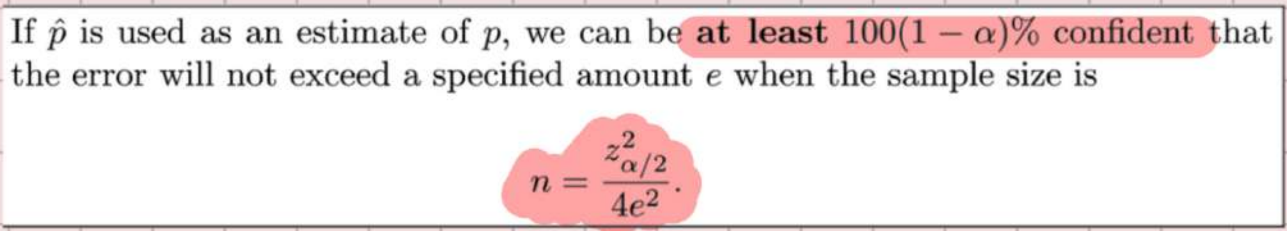
n?



**Example 9.15:** How large a sample is required if we want to be 95% confident that our estimate of $p$ in Example 9.14 is within 0.02 of the true value?

→ max error'ü 0.02'ye limitlemek istiyoruz.

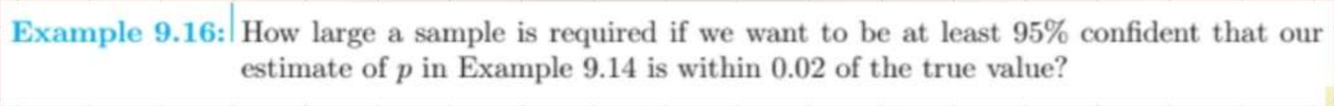
$$\frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2089.9 \approx 2090$$



If $\hat{p}$ is used as an estimate of $p$, we can be **at least** $100(1-\alpha)\%$ confident that the error will not exceed a specified amount $e$ when the sample size is

$$n = \frac{z^2_{\alpha/2}}{4e^2}.$$

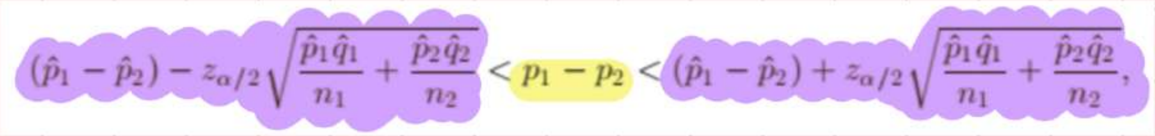
## Example



**Example 9.16:** How large a sample is required if we want to be at least 95% confident that our estimate of $p$ in Example 9.14 is within 0.02 of the true value?

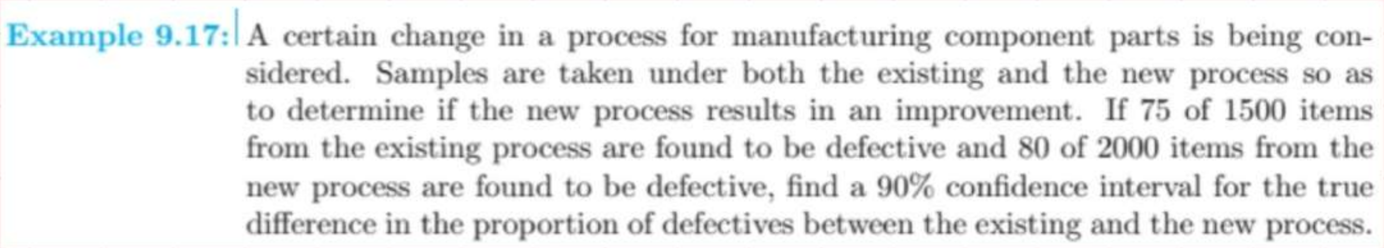
$$n = \frac{(z_{\alpha/2})^2}{4e^2} = \frac{(z_{0.025})^2}{4e^2} = \frac{(1.96)^2}{4.(0.02)^2} = 2.401$$

## Two Samples: Estimating the Difference between Two Proportions:



$$(\hat{p}_1-\hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1-\hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}},$$

## Example



**Example 9.17:** A certain change in a process for manufacturing component parts is being considered. Samples are taken under both the existing and the new process so as to determine if the new process results in an improvement. If 75 of 1500 items from the existing process are found to be defective and 80 of 2000 items from the new process are found to be defective, find a 90% confidence interval for the true difference in the proportion of defectives between the existing and the new process.

$n_1 = 1500$ $\qquad$ $n_2 = 2000$

$x_1 = 75$ $\qquad$ $x_2 = 80$

$\hat{p}_1 = 75/1500 = 0.05$ $\qquad$ $\hat{p}_2 = 80/2000 = 0.04$

$$(0.05 - 0.04) \mp \underset{1.645}{z_{0.05}} \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}}$$

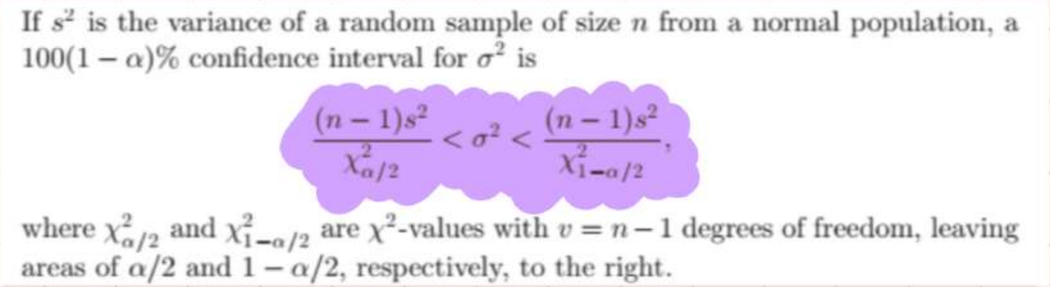
$$\boxed{-0.007 \leq \hat{p}_1 - \hat{p}_2 \leq 0.0217}$$

→ Bu interval 0'ı da içeriyor. Demekki $\hat{p}_1$ ve $\hat{p}_2$ arasında çok bir fark yok.

## ⊛ Single Sample : Estimating the Variance

→ tek bir variance ile ilgileniyorsak onun dağılımı chi-square.

→ Eğer iki variance ile ilgileniyorsak onun dağılımı da f.

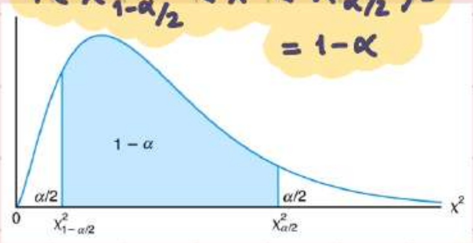


If $s^2$ is the variance of a random sample of size $n$ from a normal population, a $100(1-\alpha)\%$ confidence interval for $\sigma^2$ is

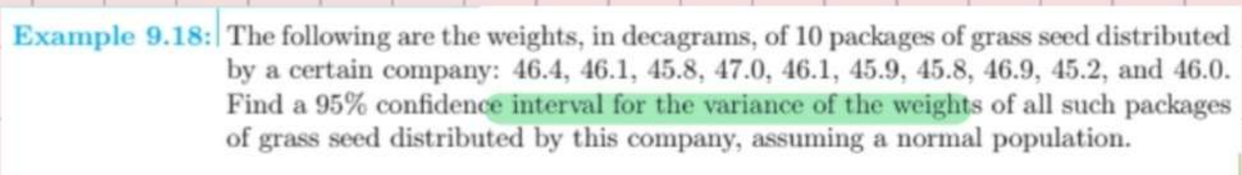
$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are $\chi^2$-values with $v = n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$, respectively, to the right.

$$P(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}) = \\ = 1-\alpha$$



## Example



**Example 9.18:** The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, and 46.0. Find a 95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company, assuming a normal population.

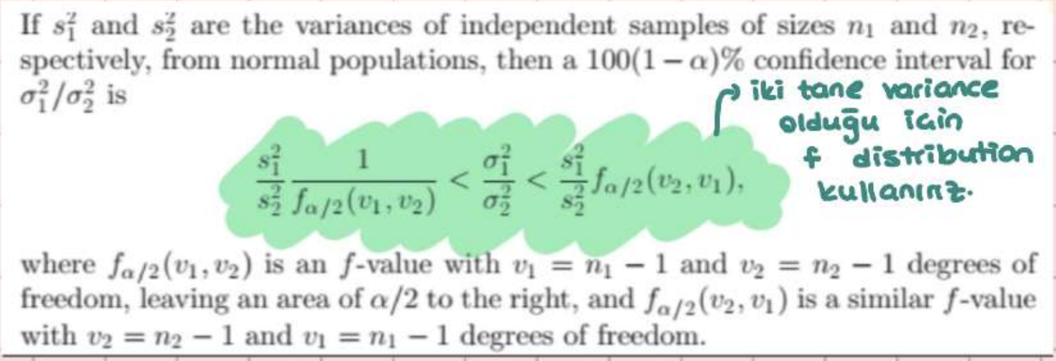
$n = 10$

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = 0.286 \qquad \frac{(n-1)s^2}{\underset{19.02}{\chi^2_{\alpha/2}}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\underset{2.70}{\chi^2_{1-\alpha/2}}}$$

$n-1$

$$\boxed{0.135 \le \sigma^2 \le 0.953}$$

## ⊛ Two Samples : Estimating The Ratio of Two Variances



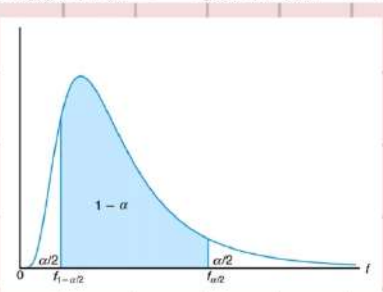
If $s_1^2$ and $s_2^2$ are the variances of independent samples of sizes $n_1$ and $n_2$, respectively, from normal populations, then a $100(1-\alpha)\%$ confidence interval for $\sigma_1^2/\sigma_2^2$ is

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}f_{\alpha/2}(v_2,v_1),$$

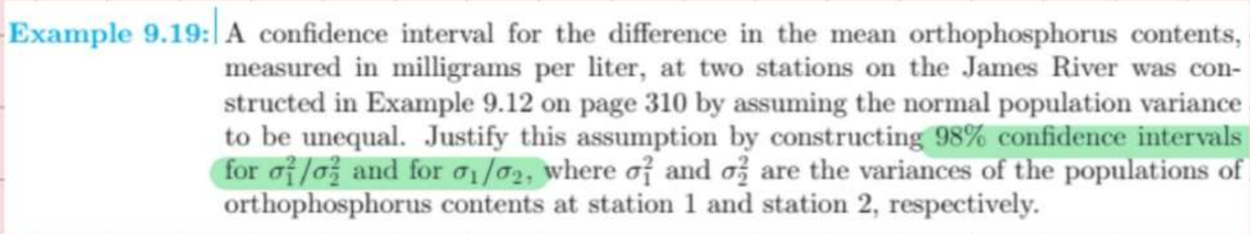
where $f_{\alpha/2}(v_1,v_2)$ is an $f$-value with $v_1 = n_1-1$ and $v_2 = n_2-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right, and $f_{\alpha/2}(v_2,v_1)$ is a similar $f$-value with $v_2 = n_2-1$ and $v_1 = n_1-1$ degrees of freedom.

→ iki tane variance olduğu için f distribution kullanınız.

$P[f_{1-\alpha/2}(v_1,v_2) < F < f_{\alpha/2}(v_1,v_2)] = 1-\alpha$



### Example



**Example 9.19:** A confidence interval for the difference in the mean orthophosphorus contents, measured in milligrams per liter, at two stations on the James River was constructed in Example 9.12 on page 310 by assuming the normal population variance to be unequal. Justify this assumption by constructing 98% confidence intervals for $\sigma_1^2/\sigma_2^2$ and for $\sigma_1/\sigma_2$, where $\sigma_1^2$ and $\sigma_2^2$ are the variances of the populations of orthophosphorus contents at station 1 and station 2, respectively.

$n_1 = 15$

$n_2 = 20$

$s_1 = 3.07$

$s_2 = 0.80$

$$\frac{s_1^2}{s_2^2}\frac{1}{f(\alpha/2)(\vartheta_1,\vartheta_2)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f(\alpha/2)(\vartheta_2,\vartheta_1)$$

$$\boxed{3.425 \le \sigma_1^2/\sigma_2^2 \le 56.991} \rightarrow$$

Bu aralık 1 değerini içerdiği zaman $\sigma_1^2$ ve $\sigma_2^2$ arasında çok bir fark yok demektir. Bu interval için de $\sigma_1^2$ ve $\sigma_2^2$ arasında çok fark yok deriz

## LECTURE 6

### → STATISTICAL HYPOTHESES :

→ Hypothesis yaparken sample ile ilgili bir parameter kullanamayız. Her zaman population ile ilgili kullanmalıyız. (population mean, population proportion)

① Null Hypothesis
② Alternative Hypothesis

} hipotez çeşitleri

## ① NULL HYPOTHESIS ($H_0$)

→ States the assumption (numerical) to be tested.

→ Always contains "=", "≤" or "≥" sign.

→ May or may not be rejected.

## ② ALTERNATIVE HYPOTHESIS ($H_1$)

→ Null Hypothesis'in tersidir.

→ Never contains the "=", "≤" or "≥" sign.

→ May or may not be supported.

→ Is generally the hypothesis that the researcher is trying the support.

## ④ Hypothesis Testing Process:

* Claim: the population mean age is 50.

(Null Hypothesis: $H_0: \mu = 50$)

( Alternative Hypothesis: $H_1: \mu \neq 50$)

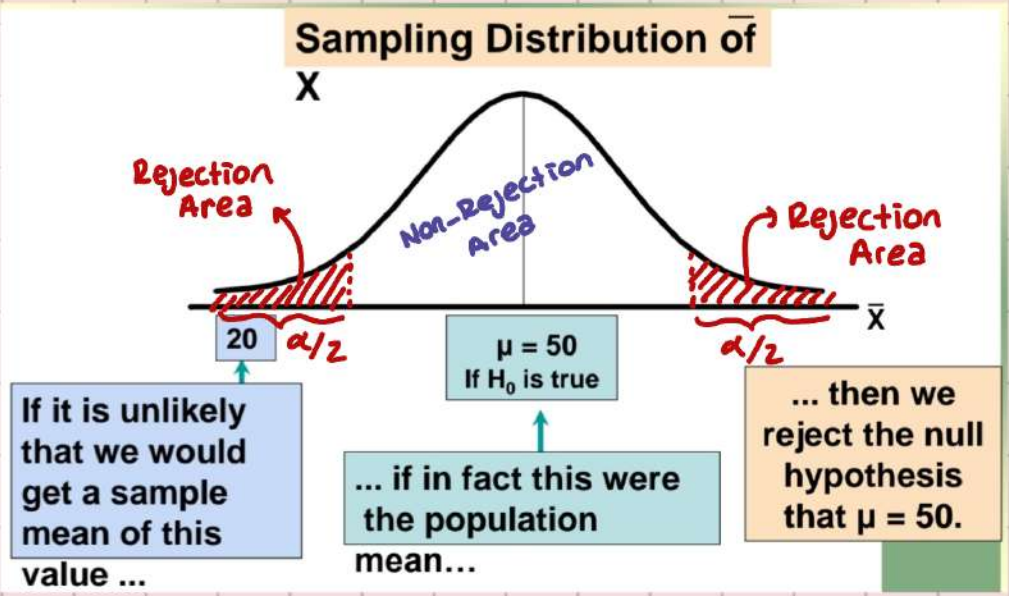
→ Suppose the sample mean age is 20: $\bar{x} = 20$ — aradaki fark çok büyük olduğu için reject.

(Bu durumda reject ederiz - REJECT NULL HYPOTHESIS)

→ $\bar{x} = 48$ gibi bir değer olsaydı direkt reject edemezdik.

Fail to reject edebilirdik. Ama hiçbir zaman accept kullanamayız.

(Reject olup olmayacağını belirlemek için de bir interval belirleriz)



Sampling Distribution of $\bar{X}$

20

μ = 50
If $H_0$ is true

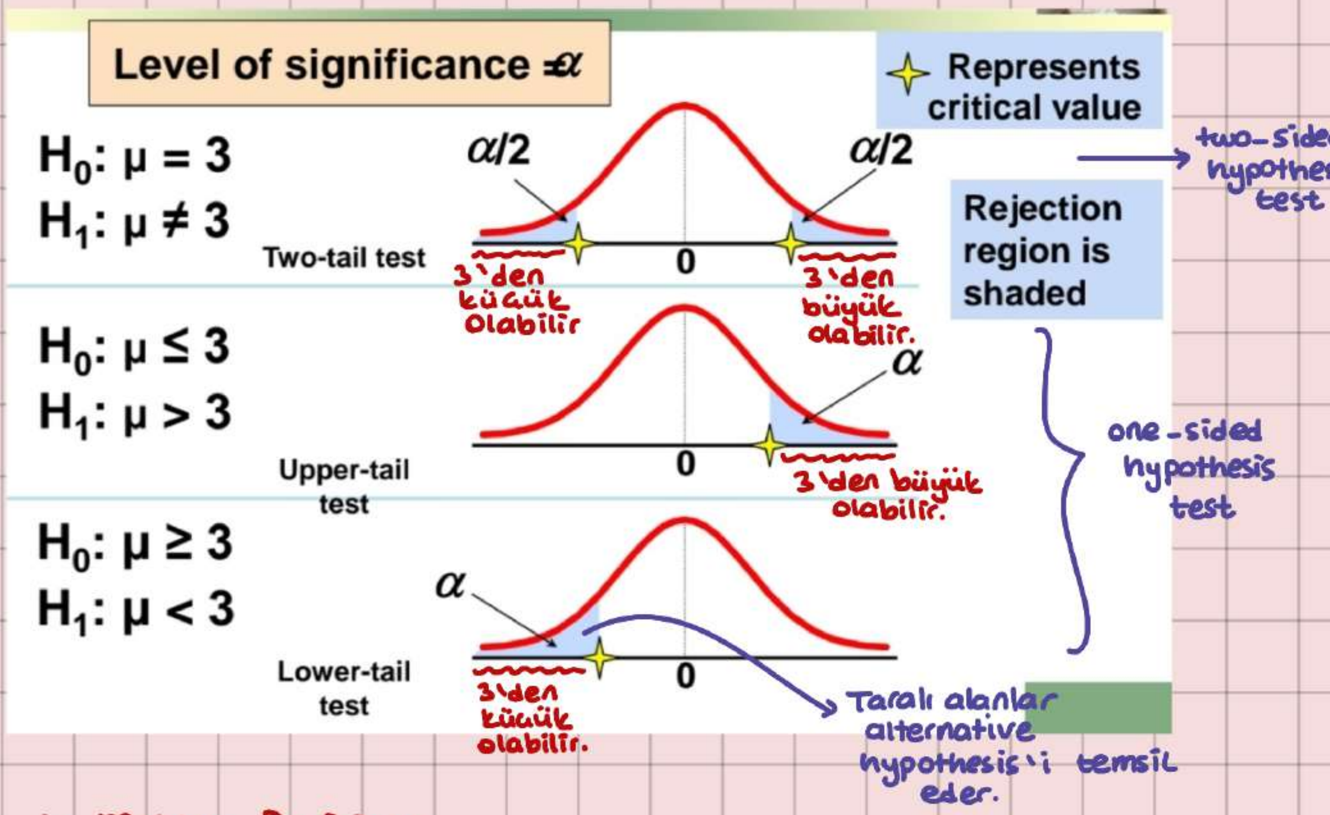
If it is unlikely that we would get a sample mean of this value ...

... if in fact this were the population mean…

... then we reject the null hypothesis that μ = 50.

→ Level of Significance (α):

- α Defines the unlikely values of the sample statistic if the null hypothesis true.
- → Defines rejection region of the sampling distribution.
- → Typical values are 0.01, 0.05 or, 0.10 → en iyi sonucu genelde 0.05 verir.
- → Provides the critical value(s) of the test.



Level of significance = α

Represents critical value

Rejection region is shaded

$H_0: \mu = 3$
$H_1: \mu \neq 3$
Two-tail test

$H_0: \mu \leq 3$
$H_1: \mu > 3$
Upper-tail test

$H_0: \mu \geq 3$
$H_1: \mu < 3$
Lower-tail test

α/2
α/2
0
3'den küçük olabilir
3'den büyük olabilir.
two-sided hypothesis test
α
0
3'den büyük olabilir.
one-sided hypothesis test
α
0
3'den küçük olabilir.
Taralı alanlar alternative hypothesis'i temsil eder.

→ Errors in Making Decisions:

① TYPE I ERROR

- → Bir şey doğruyken ona yanlış demek. (Reject a true null hypothesis)
- → The probability of Type I Error is α
- → Called level of signifiance of the test

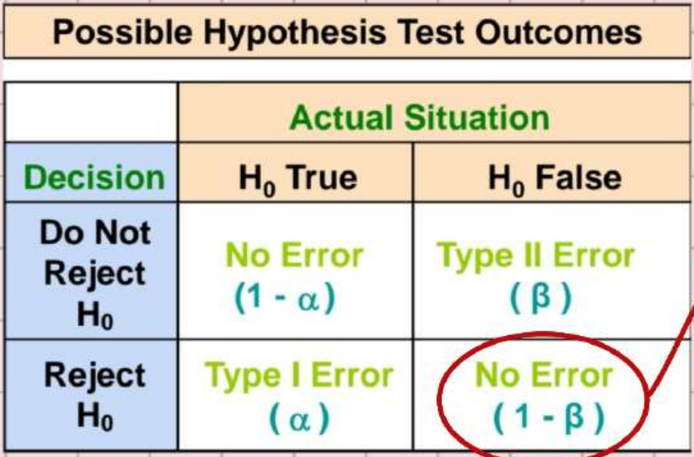
④ Diyelim ki sınıfın not ortalaması 3 dedik. Ve gerçekten 3. Bunu test etmek için bir sample aldık. Ama çok kötü bir almışız ve sample ortalaması 2 geldi. Bu durumda $\mu = 3$'ü reject ederiz. Ama aslında doğruydu ve reject etmemeliydik. Bu type I error olur.

② TYPE II ERROR

- → Bir şey yanlışken ona doğru demek. (Fail to reject a false null hypothesis)
- → The probability of Type II Error is β
- → Sınıf not ortalaması 3 dedik ve test etmek için sample aldık, ama sample için çok iyi öğrencileri seçtik diyelim. Ortalama 3 geldi diyelim.

ve reject etmedik. Ama aslında ortalama 1 ve reject etmeliydik.

Bu durumda type II error yapmış oluruz.



**Possible Hypothesis Test Outcomes**

| | Actual Situation | |
|---|---|---|
| Decision | $H_0$ True | $H_0$ False |
| Do Not Reject $H_0$ | No Error<br>$(1-\alpha)$ | Type II Error<br>$(\beta)$ |
| Reject $H_0$ | Type I Error<br>$(\alpha)$ | No Error<br>$(1-\beta)$ |

Bir şey yanlışken ona yanlış demek $(1-\beta)$ Literature'de → The power of the test olarak geçer.

Type I and Type II errors can not happen at the same time.

- Type I error can only occur if $H_0$ is true.
- Type II error can only occur if $H_0$ is false.

} type I ve type II error mutually exclusive error'dür. (Yani biri gözleniyorsa diğeri gözlenemez)

If Type I error probability $(\alpha)\uparrow$, then Type II error probability $(\beta)\downarrow$ → mutually exclusive events

- $\beta\uparrow$ when $\alpha\downarrow$
- $\beta\uparrow$ when $\sigma\uparrow$ → $\sigma$ arttıkça estimation poorlaşır ve error, yani $\beta$, artar.
- $\beta\uparrow$ when $n\downarrow$ → sample size ne kadar büyük olursa error, yani $\beta$, o kadar küçülür.

→ THE POWER of the TEST

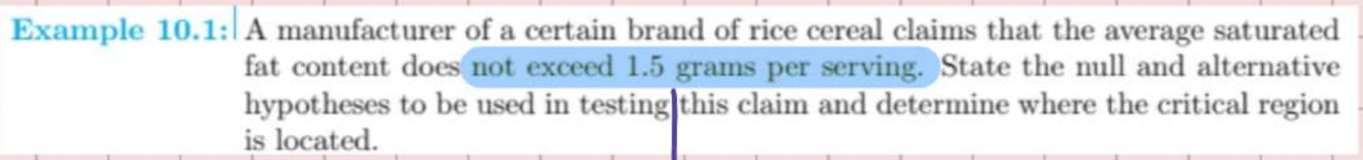
- The power of a test is the probability of rejecting a null hypothesis that is false.

$$\boxed{\text{Power} = P(\text{Reject } H_0 \mid H_1 \text{ is true})}$$

↳ Null Hypothesis'i reject etmenin olasılığı

Power of the test increases as the sample size increases.

Example

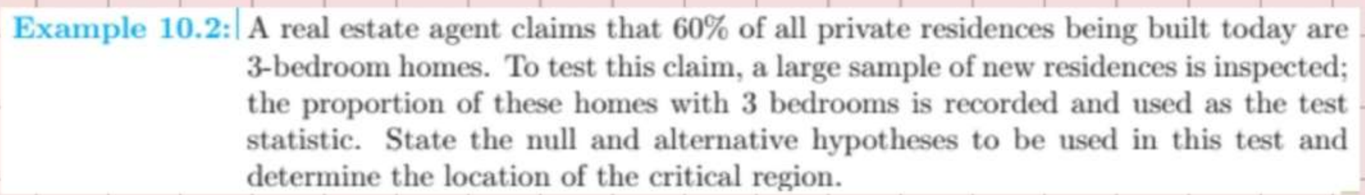


**Example 10.1:** A manufacturer of a certain brand of rice cereal claims that the average saturated fat content does not exceed 1.5 grams per serving. State the null and alternative hypotheses to be used in testing this claim and determine where the critical region is located.

$H_0: \mu \le 1.5$

$H_1: \mu > 1.5$

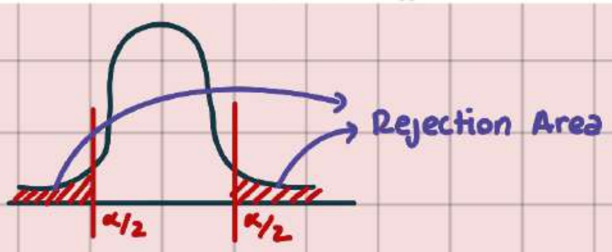
## Example



**Example 10.2:** A real estate agent claims that 60% of all private residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test and determine the location of the critical region.

$H_0 : p = 0.6$

$H_1 : p \neq 0.6$



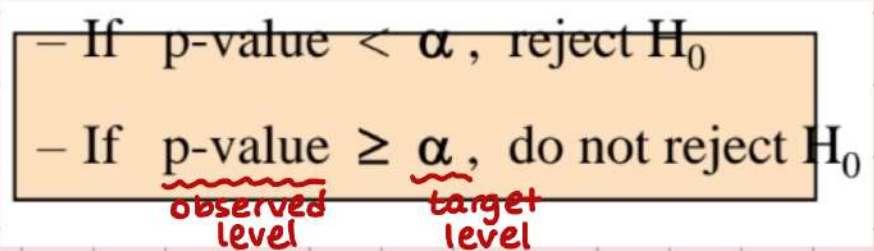
### → Hipotez Testi için 3 method var:

1) Using p-value approach.
2) Standart method.
3) Confidence interval approach.

### ① P-VALUE APPROACH TO TESTING:

→ p-value'yu (probability)'yi hesaplarız. Eğer α'nın altındaysa null hypothesis reject edilir. Eğer üstündeyse reject to fail olur.

→ α = Target Level (Başta hedef olarak belirlenir.)



– If p-value < α , reject $H_0$

– If $\underbrace{\text{p-value}}_{\text{observed level}} \geq \underbrace{\alpha}_{\text{target level}}$, do not reject $H_0$

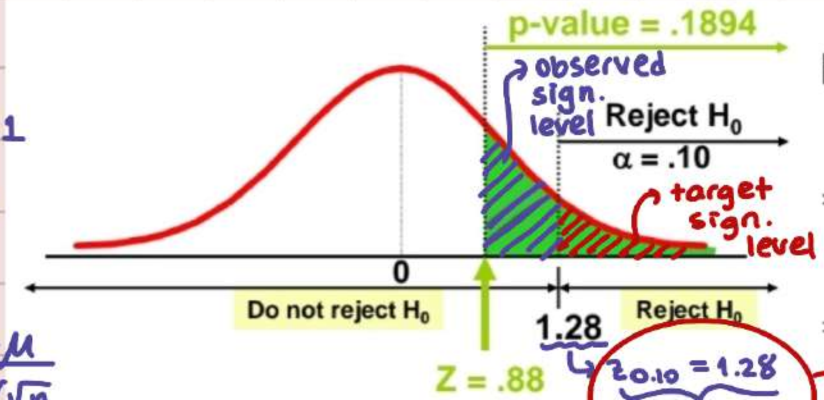
$\mu = 52$

$\bar{x} = 53.1$

$\sigma = 10$

$n = 64$

$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$



$P(\bar{x} \geq 53.1 \mid \mu = 52.0)$

$= P\left(z \geq \dfrac{53.1 - 52.0}{10/\sqrt{64}}\right)$

$= P(z \geq 0.88) = 1 - .8106$

Bundan büyük

$/\sqrt{n}$ critical value $= .1894$

bir değer gelseydi reddederdik.

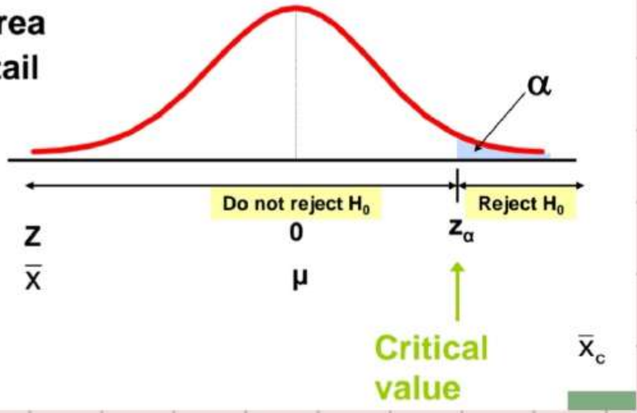
Do not reject $H_0$ since p-value = .1894 > $\alpha$ = .10

↳ fail to reject

## UPPER-TAIL TESTS

- There is only one critical value, since the rejection area is in only one tail

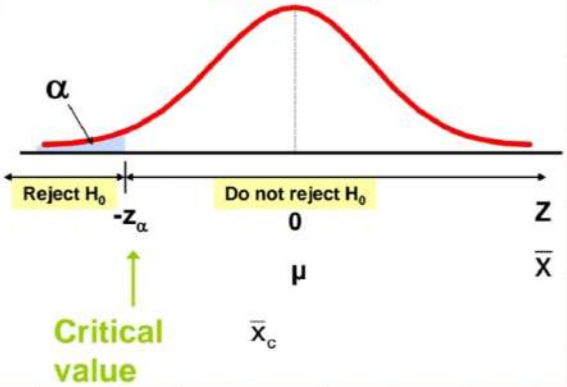
$H_0$: $\mu \le 3$
$H_1$: $\mu > 3$



## LOWER-TAIL TESTS

- There is only one critical value, since the rejection area is in only one tail

$H_0$: $\mu \ge 3$
$H_1$: $\mu < 3$

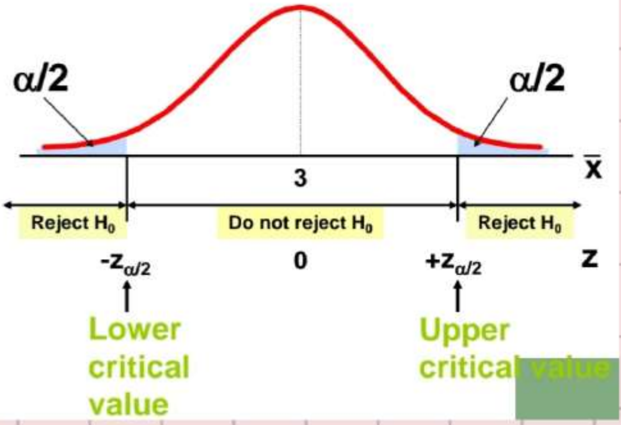


## TWO-TAIL TESTS

- In some settings, the alternative hypothesis does not specify a unique direction

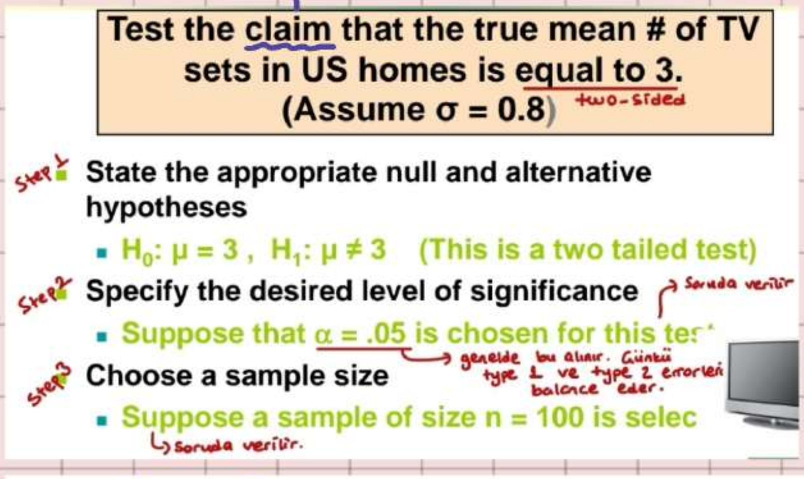
$H_0$: $\mu = 3$
$H_1$: $\mu \neq 3$

- There are two critical values, defining the two regions of rejection



## Example

claim dediği için Ho'a yazdık.



Test the claim that the true mean # of TV sets in US homes is equal to 3. (Assume σ = 0.8) two-sided

step 1 State the appropriate null and alternative hypotheses

- $H_0$: μ = 3 , $H_1$: μ ≠ 3 (This is a two tailed test)

step 2 Specify the desired level of significance → Soruda verilir

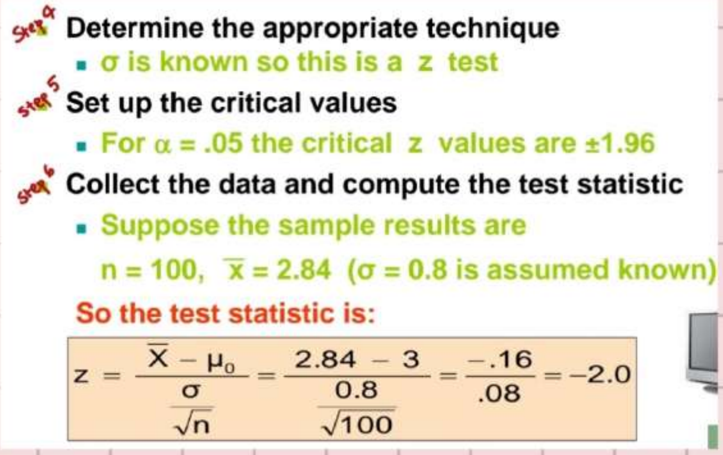
- Suppose that α = .05 is chosen for this tes

→ genelde bu alınır. Çünkü type 1 ve type 2 errorleri balance eder.

step 3 Choose a sample size

- Suppose a sample of size n = 100 is selec

↳ Soruda verilir.



step 4 Determine the appropriate technique

- σ is known so this is a z test

step 5 Set up the critical values

- For α = .05 the critical z values are ±1.96

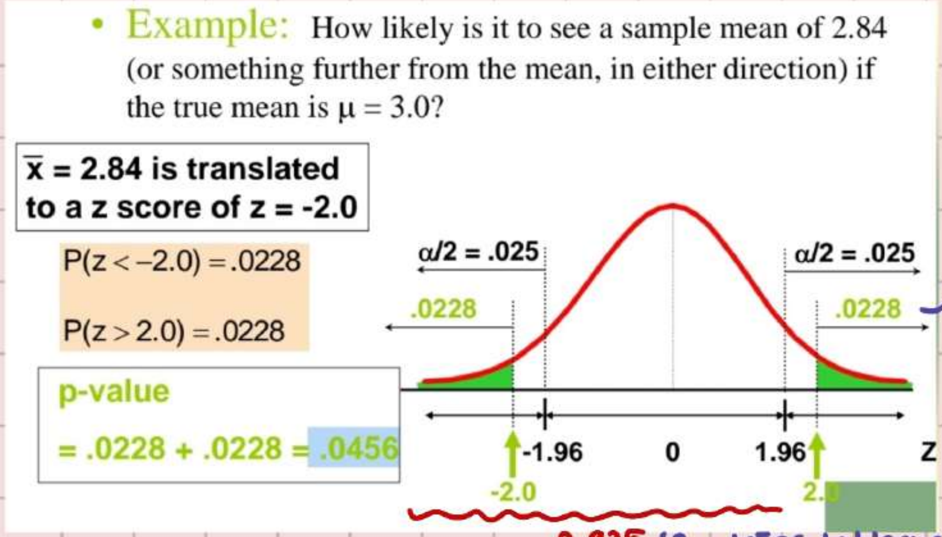
step 6 Collect the data and compute the test statistic

- Suppose the sample results are

n = 100, $\bar{X}$ = 2.84 (σ = 0.8 is assumed known)

So the test statistic is:

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$



- Example: How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is μ = 3.0?

$\bar{x}$ = 2.84 is translated to a z score of z = -2.0

$P(z < -2.0) = .0228$

$P(z > 2.0) = .0228$

p-value = .0228 + .0228 = .0456

0.875 (Bu değer tablonun içindedir ve z valusu 1.96'dır.)

p-value = 0.0228 + 0.0228

p-value = 0.0456 , α = 0.05

→ if p-value < α , reject Ho

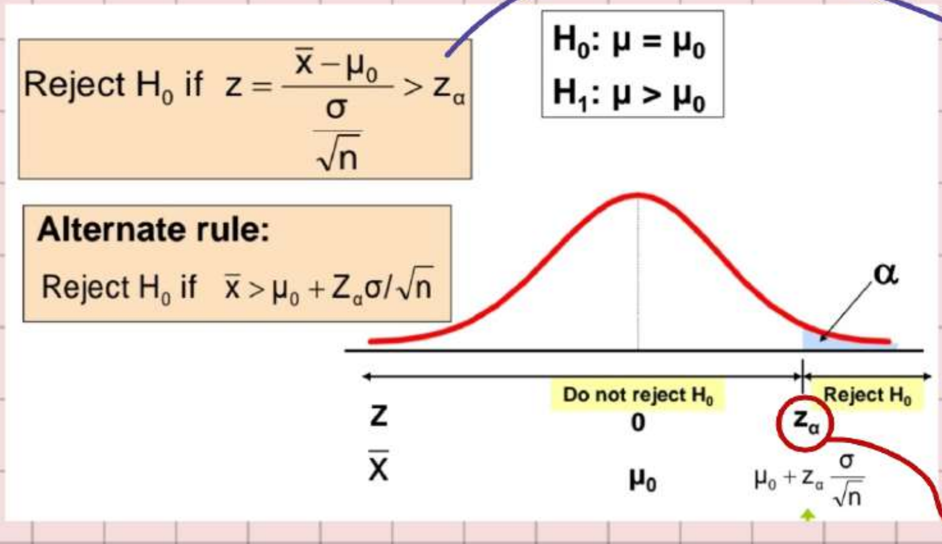
→ if p-value > α, do not reject Ho

0.0456 < 0.05, so we reject the null hypothesis Ho

## → HYPOTHESIS TESTS for the MEAN

- σ Known
- σ Unknown

## ① σ KNOWN ;



Reject $H_0$ if $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha$

$H_0$: μ = $\mu_0$

$H_1$: μ > $\mu_0$

**Alternate rule:**

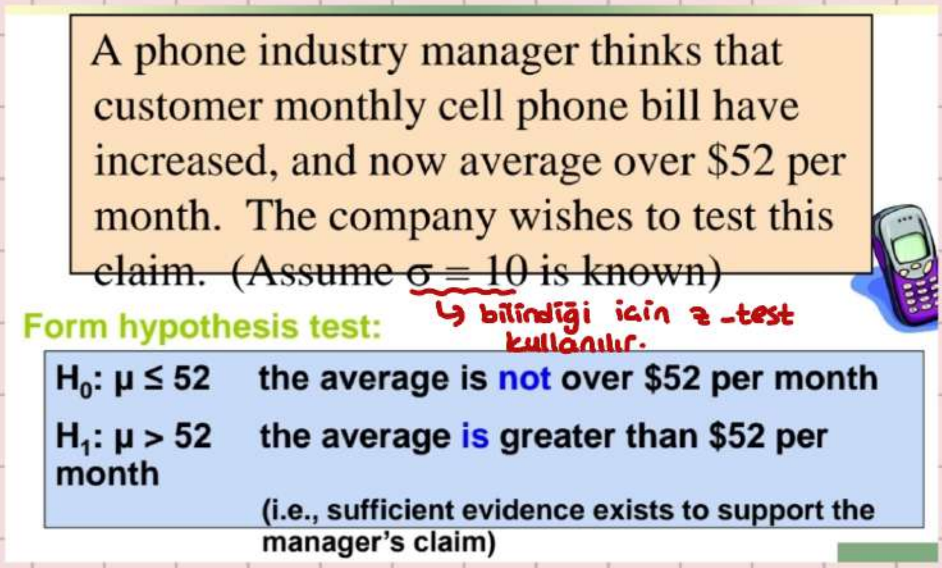
Reject $H_0$ if $\bar{x} > \mu_0 + Z_\alpha \sigma/\sqrt{n}$

Bu standar methoddur.

Eğer ki bu z value'nun olasılığına tablodan bakarsak; bu p-value olur. Eğer p-value < α ise Ho reject edilir.

critical value

## Example



A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over \$52 per month. The company wishes to test this claim. (Assume $\sigma = 10$ is known)

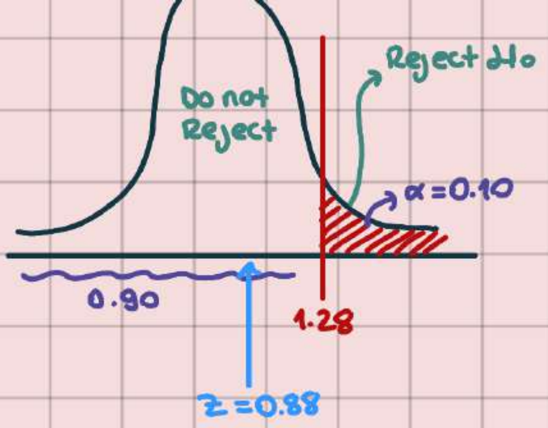
↳ bilindiği için z-test kullanılır.

Form hypothesis test:

$H_0$: $\mu \le 52$  the average is not over \$52 per month

$H_1$: $\mu > 52$  the average is greater than \$52 per month

(i.e., sufficient evidence exists to support the manager's claim)

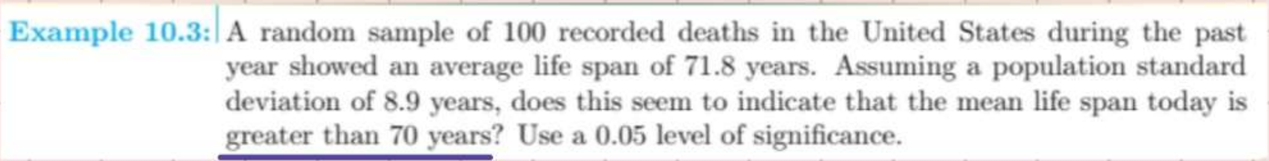


$z=0.88 < 1.28$ (Do not reject $H_0$)

↳ Bu standart yoldur. Bunun z-tablodaki olasılık değerine bakarsak p-value olur ve p-value>α olur. Ve do not reject $H_0$ deriz.

Suppose $\left.\begin{array}{l} n=64 \\ \bar{x}=53.1 \end{array}\right\}$ $z = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \dfrac{53.1-52}{10/\sqrt{64}} = 0.88$

## Example



**Example 10.3:** A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

$n=100$

$\bar{x}=71.8$

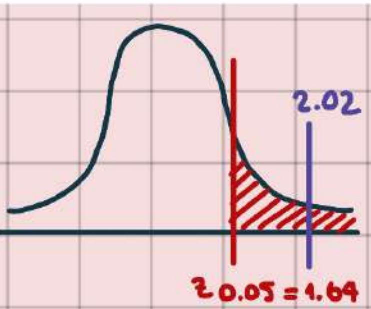
$\sigma=8.9$

$\alpha=0.05$

$H_0 : \mu \le 70$

$H_1 : \mu > 70$

Reject $H_0$

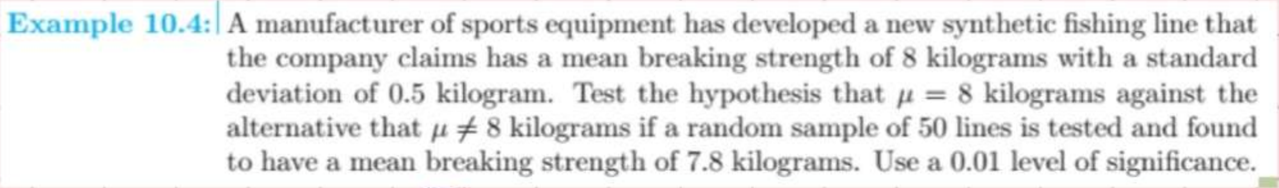


$$z\text{score} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

$$z = \frac{71.8-70}{8.9/\sqrt{100}} = 2.02$$

$z_{0.02}$ = p-value = 0.0217

## Example



**Example 10.4:** A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that $\mu = 8$ kilograms against the alternative that $\mu \neq 8$ kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

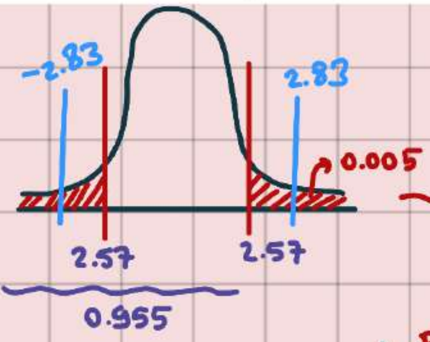
$H_0 : \mu = 8$

$H_1 : \mu \neq 8$

$n=50$

$\sigma=0.5$

$\bar{x}=7.8$

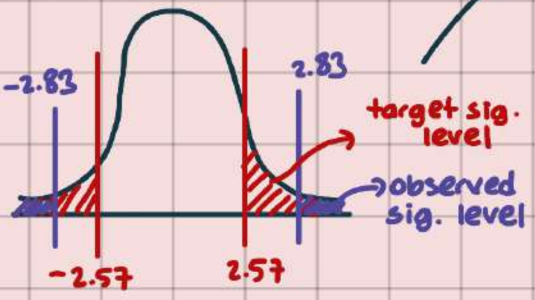


$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{7.8-8}{0.5/\sqrt{50}} = -2.83$$

→ Reject Null Hypothesis ($H_0$) (Standart Approach)

→ P-VALUE APPROACH

z_test

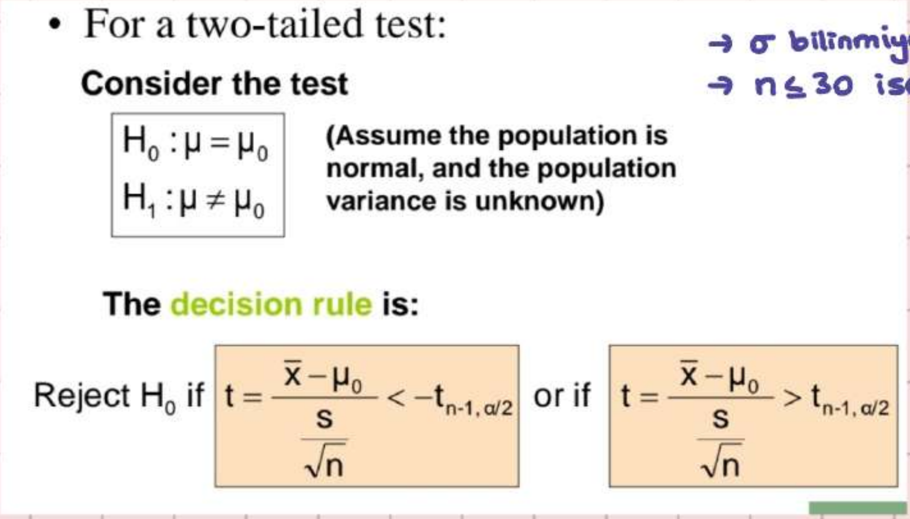


2.83'nin verdiği alanı tablodan buluruz.

p-value = 0.0046 → p-value < α → Reject Ho

(Grafikten de observed sig. level'ın target sig. level'den küçük olduğunu görüp p-value < α diyebiliriz)

## ② σ UNKNOWN:



- For a two-tailed test:

**Consider the test**

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}}$$

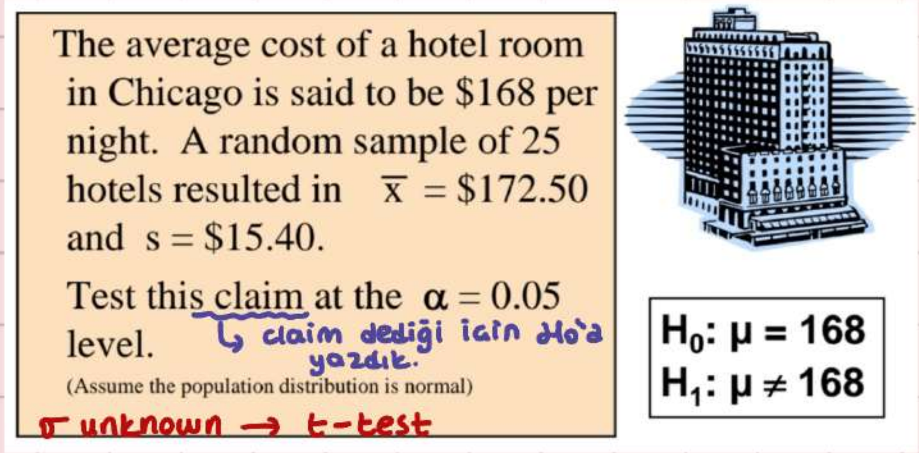
**(Assume the population is normal, and the population variance is unknown)**

**The decision rule is:**

Reject $H_0$ if $t = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < -t_{n-1,\alpha/2}$ or if $t = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} > t_{n-1,\alpha/2}$

→ σ bilinmiyorsa
→ n ≤ 30 ise
} t-test kullanılır.

### Example



The average cost of a hotel room in Chicago is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{x}$ = \$172.50 and s = \$15.40.

Test this claim at the α = 0.05 level.

↳ claim dediği için Ho'a yazdık.

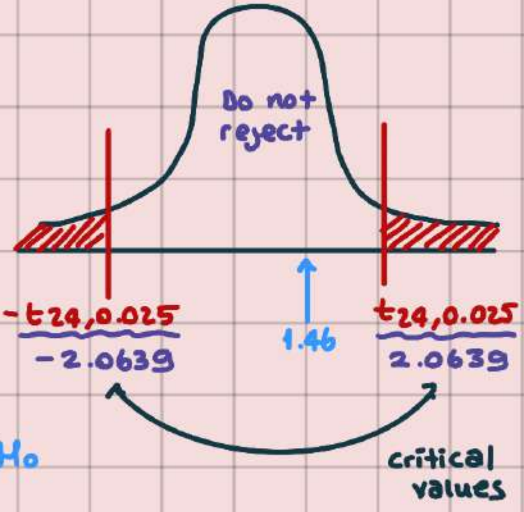
(Assume the population distribution is normal)

σ unknown → t-test

$H_0$: μ = 168
$H_1$: μ ≠ 168

Ho : μ = 168    n = 25

H₁ : μ ≠ 168    α = 0.05



$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{172.5 - 168}{15.40/\sqrt{25}} = 1.46$$

↳ Do not reject Ho

## → HYPOTHESIS TEST for POPULATION PROPORTION

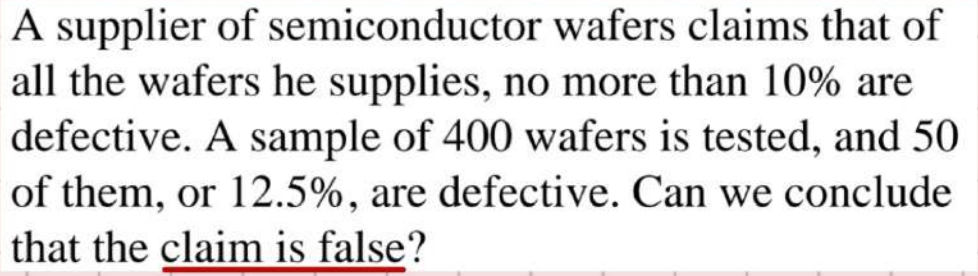
$$\boxed{z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}}$$

α If the alternative hypothesis is : $H_1 : \hat{p} > p_0$, the P-value is the area to the right of z.

α If the alternative hypothesis is : $H_1 : \hat{p} < p_0$, the P-value is the area to the left of z.

α If the alternative hypothesis is : $H_1 : \hat{p} \neq p_0$, the P-value is the sum of the areas in the tails cut off by z and -z.

## Example



A supplier of semiconductor wafers claims that of all the wafers he supplies, no more than 10% are defective. A sample of 400 wafers is tested, and 50 of them, or 12.5%, are defective. Can we conclude that the claim is false?

$H_0 : p \leq 10\,\%$

$H_1 : p > 10\,\%$

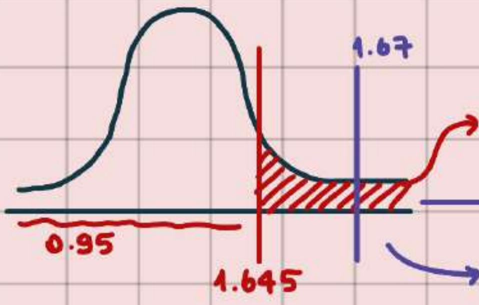
$n = 400$

$\bar{x} = 50$ defective

$\hat{p} = 50/400 = 12.5\,\%$

$n \geq 30 \rightarrow$ z-test

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.125 - 0.10}{\sqrt{\frac{(0.10)(0.90)}{400}}} = 1.67$$

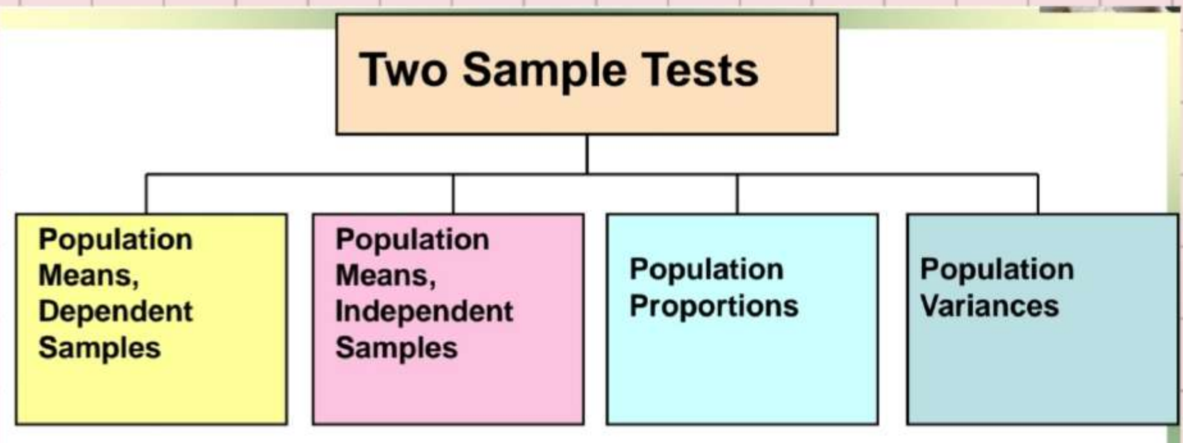


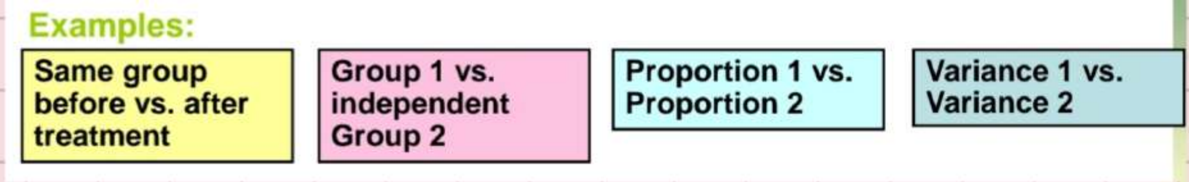
→ 0.05 = α (assume ettik)

→ Reject $H_0$ $p \leq 10\%$

→ p-value approach'da 1.67'nin tablodan alan değerine bakarız. α'dan küçükse reject ederiz.

## ⊛ TWO SAMPLE TESTS


Two Sample Tests
Population Means, Dependent Samples
Population Means, Independent Samples
Population Proportions
Population Variances



**Examples:**

| Same group before vs. after treatment | Group 1 vs. independent Group 2 | Proportion 1 vs. Proportion 2 | Variance 1 vs. Variance 2 |
|---|---|---|---|

→ DEPENDENT SAMPLES

* Both populations are normally distributed.

$$\boxed{d_i = x_i - y_i}$$

$y_i$ → after observation

$x_i$ → before observation

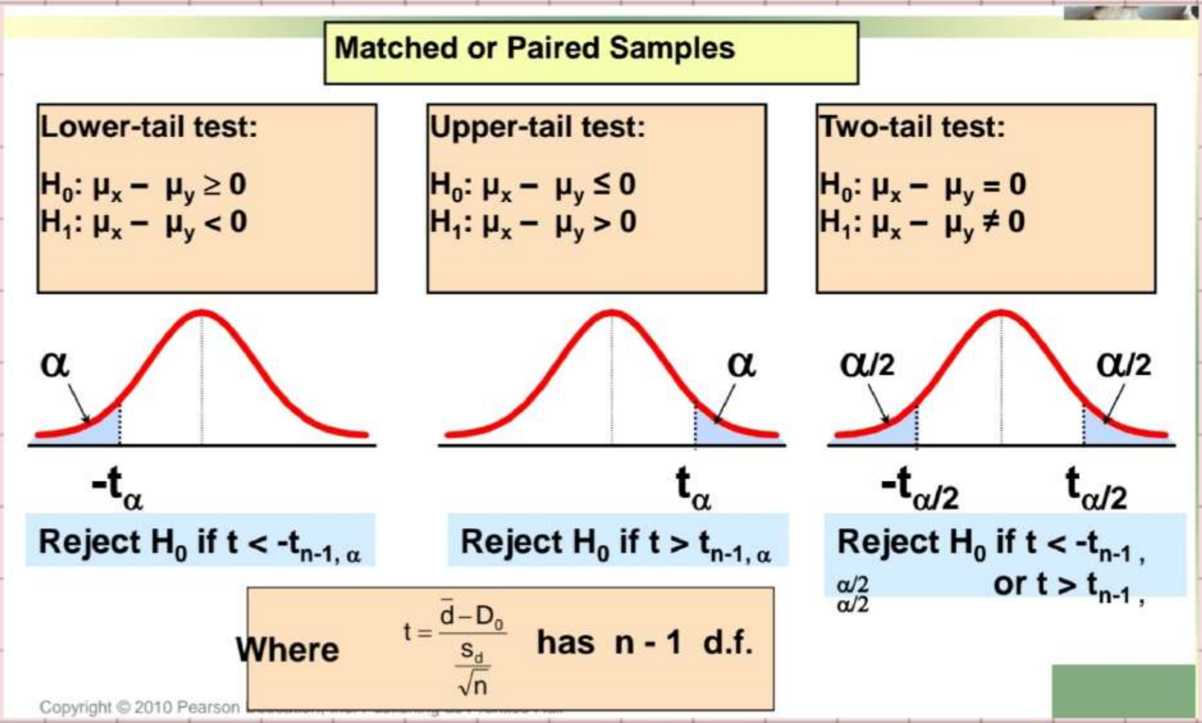
t value, with n-1 degrees of freedom.

$$\boxed{\bar{d} = \frac{\sum d_i}{n} = \bar{x} - \bar{y}} \longrightarrow \boxed{t = \frac{\bar{d} - D_0}{\frac{S_d}{\sqrt{n}}}}$$

$D_0$: hypothesized mean difference

$S_d$ : sample standard dev.

n : sample size.



**Matched or Paired Samples**

| Lower-tail test: | Upper-tail test: | Two-tail test: |
|---|---|---|
| $H_0: \mu_x - \mu_y \geq 0$<br>$H_1: \mu_x - \mu_y < 0$ | $H_0: \mu_x - \mu_y \leq 0$<br>$H_1: \mu_x - \mu_y > 0$ | $H_0: \mu_x - \mu_y = 0$<br>$H_1: \mu_x - \mu_y \neq 0$ |

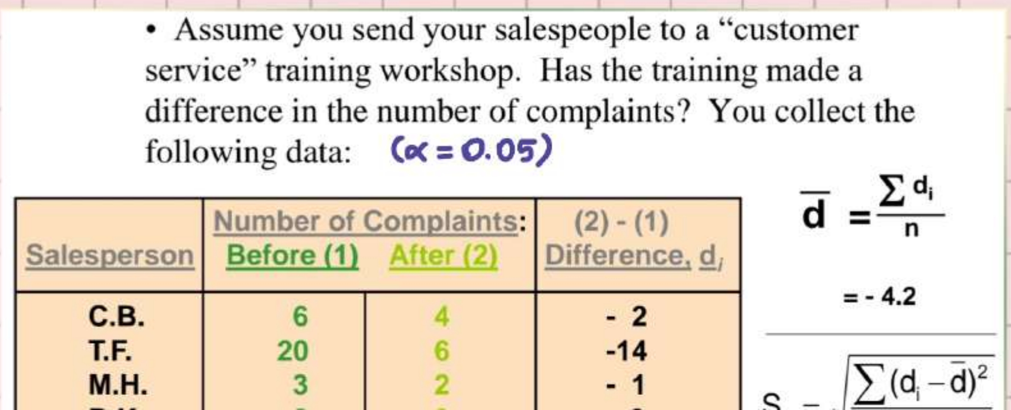
Reject $H_0$ if $t < -t_{n-1,\alpha}$

Reject $H_0$ if $t > t_{n-1,\alpha}$

Reject $H_0$ if $t < -t_{n-1}$, or $t > t_{n-1}$, $\alpha/2$ $\alpha/2$

Where $t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$ has n - 1 d.f.

Example



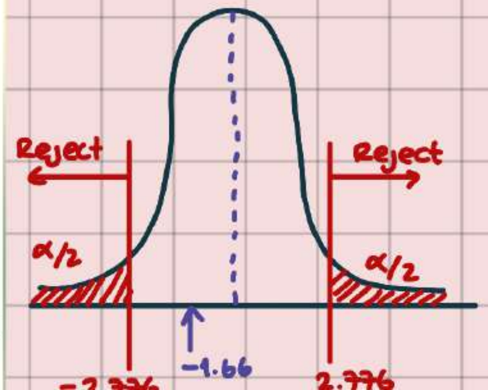
- Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? You collect the following data: ($\alpha = 0.05$)

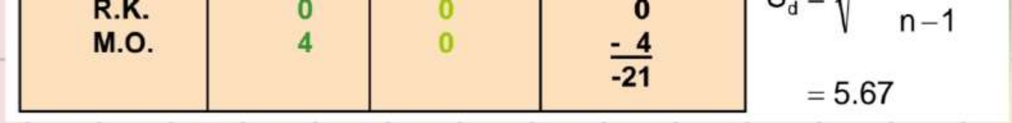
| Salesperson | Number of Complaints: Before (1) | After (2) | (2) - (1) Difference, $d_i$ |
|---|---|---|---|
| C.B. | 6 | 4 | - 2 |
| T.F. | 20 | 6 | -14 |
| M.H. | 3 | 2 | - 1 |

$$\bar{d} = \frac{\sum d_i}{n}$$

$$= -4.2$$

$$S_d = \sqrt{\sum(d_i - \bar{d})^2}$$





| | | | |
|---|---|---|---|
| R.K. | 0 | 0 | 0 |
| M.O. | 4 | 0 | $\underline{-4}$<br>-21 |

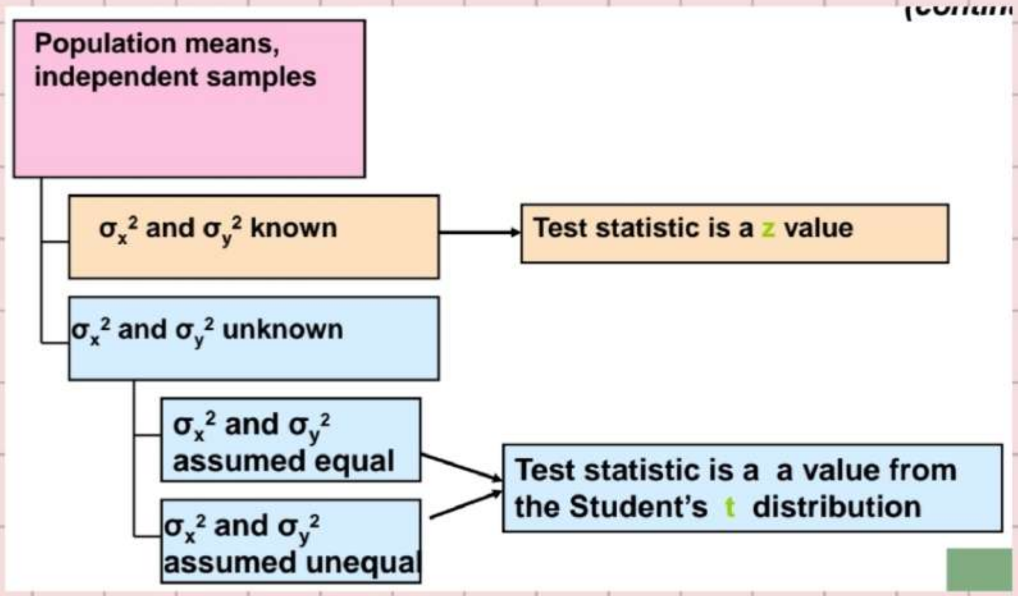
$\sigma_d = \sqrt{\dfrac{\quad}{n-1}}$

$= 5.67$

Do Not Reject Ho

$$t = \frac{\bar{d} - D_0}{S_d/\sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{5}} = -1.66$$

## → DIFFERENCE BETWEEN TWO MEANS



Population means, independent samples

$\sigma_x^2$ and $\sigma_y^2$ known → Test statistic is a z value

$\sigma_x^2$ and $\sigma_y^2$ unknown

- $\sigma_x^2$ and $\sigma_y^2$ assumed equal
- $\sigma_x^2$ and $\sigma_y^2$ assumed unequal

→ Test statistic is a a value from the Student's t distribution

### ⊛ $\sigma_x^2$ and $\sigma_y^2$ known

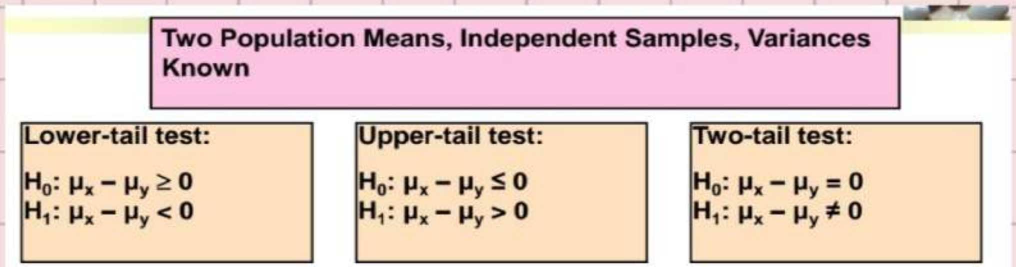
- Samples are randomly and independently drawn.
- Both population distributions are normal
- Population variances are known.

$$\sigma^2_{\bar{X}-\bar{Y}} = \frac{\sigma_x{}^2}{n_x} + \frac{\sigma_y{}^2}{n_y}$$

$$Z = \frac{(\bar{x}-\bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_x^2}{n_X} + \dfrac{\sigma_y^2}{n_Y}}}$$

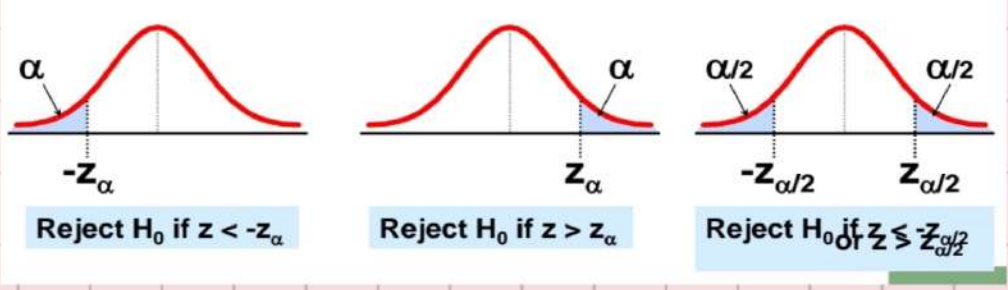
$$H_0 : \mu_x - \mu_y = D_0$$

$$z = \frac{(\bar{x}-\bar{y}) - D_0}{\sqrt{\dfrac{\sigma_x{}^2}{n_x} + \dfrac{\sigma_y{}^2}{n_y}}}$$



Two Population Means, Independent Samples, Variances Known

| Lower-tail test: | Upper-tail test: | Two-tail test: |
|---|---|---|
| $H_0: \mu_x - \mu_y \ge 0$<br>$H_1: \mu_x - \mu_y < 0$ | $H_0: \mu_x - \mu_y \le 0$<br>$H_1: \mu_x - \mu_y > 0$ | $H_0: \mu_x - \mu_y = 0$<br>$H_1: \mu_x - \mu_y \ne 0$ |



Reject $H_0$ if $z < -z_\alpha$

Reject $H_0$ if $z > z_\alpha$

Reject $H_0$ if $z \le -z_{\alpha/2}$ or $z > z_{\alpha/2}$

## ⊛ $\sigma_x^2$ and $\sigma_y^2$ Unknown (Equal)

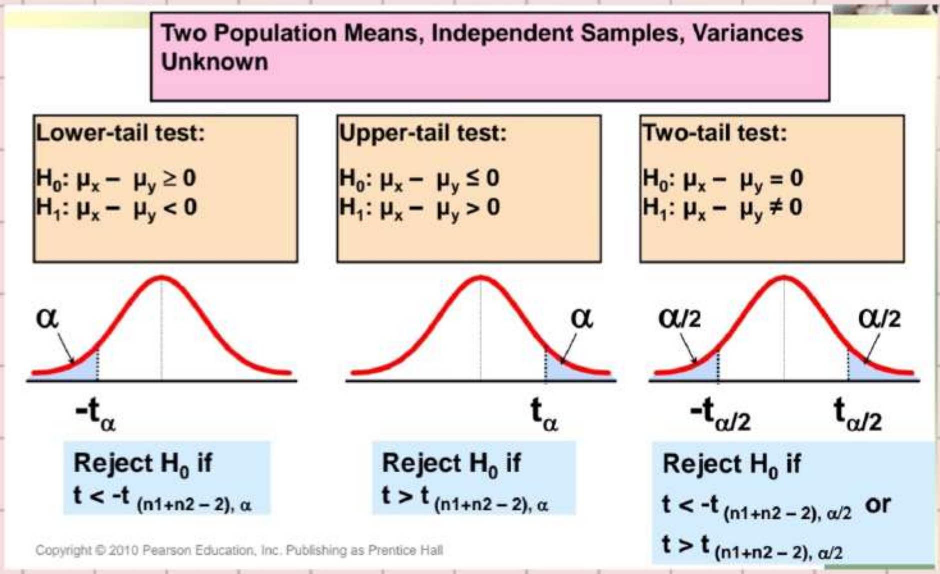
- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown but assumed equal.

Use a t value with $(n_x+n_y-2)$ degrees of freedom.

$$t = \frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}}$$

Where t has $(n_1 + n_2 - 2)$ d.f., and

$$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$$



**Two Population Means, Independent Samples, Variances Unknown**

| Lower-tail test: | Upper-tail test: | Two-tail test: |
|---|---|---|
| $H_0: \mu_x - \mu_y \ge 0$<br>$H_1: \mu_x - \mu_y < 0$ | $H_0: \mu_x - \mu_y \le 0$<br>$H_1: \mu_x - \mu_y > 0$ | $H_0: \mu_x - \mu_y = 0$<br>$H_1: \mu_x - \mu_y \ne 0$ |

Reject $H_0$ if $t < -t_{(n1+n2-2),\alpha}$

Reject $H_0$ if $t > t_{(n1+n2-2),\alpha}$

Reject $H_0$ if $t < -t_{(n1+n2-2),\alpha/2}$ or $t > t_{(n1+n2-2),\alpha/2}$

## Example

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

| | NYSE | NASDAQ |
|---|---|---|
| Number | 21 | 25 |
| Sample mean | 3.27 | 2.53 |
| Sample std dev | 1.30 | 1.16 |

Assuming both populations are approximately normal with equal variances, is there a difference in average yield ($\alpha = 0.05$)?

The test statistic is:

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(3.27-2.53)-0}{\sqrt{1.5021\left(\frac{1}{21}+\frac{1}{25}\right)}} = \boxed{2.040}$$

$$S^2 = (n_1-1)S_1^2+(n_2-1)S_2^2 = (21-1)1.30^2+(25-1)1.16^2 = 1.5021$$

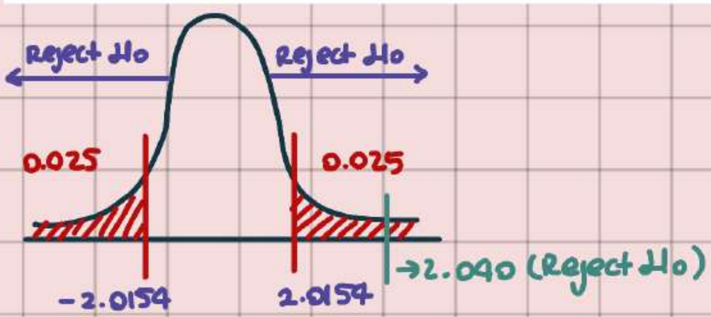
$(n_1 - 1) + (n_2 - 1) \qquad (21-1)+(25-1)$

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$ → critical values = $\mp 2.0154$

$df = 21 + 25 - 2 = 44$



## ④ $\sigma_x^2$ and $\sigma_y^2$ Unknown, unequal

- Samples are ramdomly and independently drawn.
  Populations are normally distributed.
  Population variances are unknown and assumed unequal.

$$v = \frac{\left[ \left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right) \right]^2}{\left(\frac{s_x^2}{n_x}\right)^2 / (n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2 / (n_y - 1)}$$

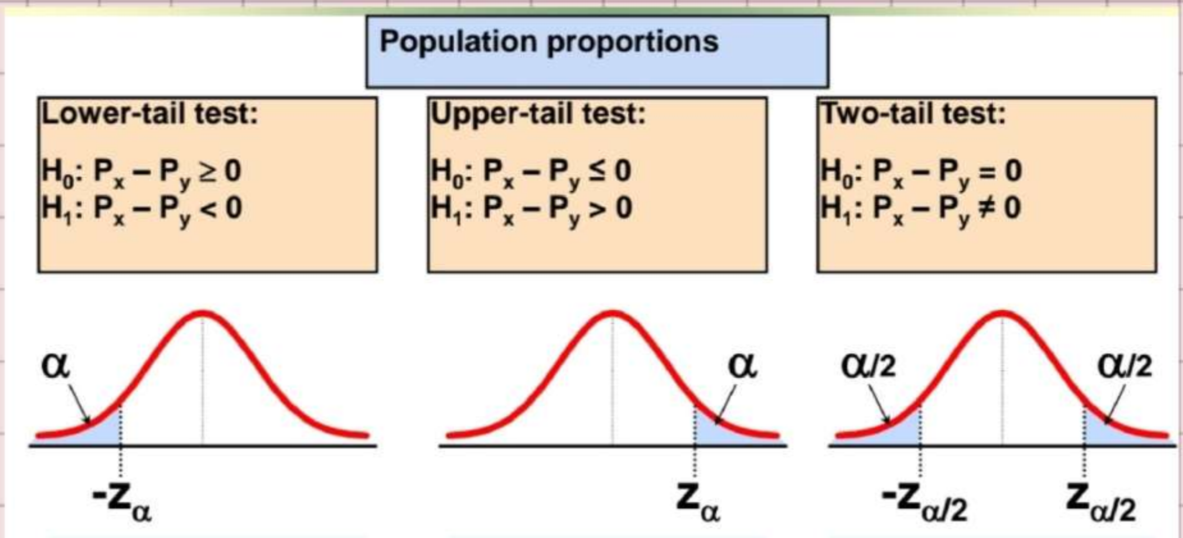
$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_x^2}{n_X} + \frac{s_y^2}{n_Y}}}$$

→ TEST STATISTIC for TWO POPULATION PROPORTIONS :

- The test statistic for $H_0 : P_x - P_y = 0$ is a z value :

$$z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}$$

$$\hat{p}_0 = \frac{n_x \hat{p}_x + n_y \hat{p}_y}{n_x + n_y}$$



**Population proportions**

| Lower-tail test: | Upper-tail test: | Two-tail test: |
|---|---|---|
| $H_0$: $P_x - P_y \geq 0$<br>$H_1$: $P_x - P_y < 0$ | $H_0$: $P_x - P_y \leq 0$<br>$H_1$: $P_x - P_y > 0$ | $H_0$: $P_x - P_y = 0$<br>$H_1$: $P_x - P_y \neq 0$ |

| | | |
|---|---|---|
| Reject $H_0$ if $z < -z_\alpha$ | Reject $H_0$ if $z > z_\alpha$ | Reject $H_0$ if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ |

## Example

Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A?

- In a random sample, 36 of 72 men and 31 of 50 women indicated they would vote Yes
- Test at the .05 level of significance

$$z = \frac{(\hat{p}_M - \hat{p}_W)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}}$$

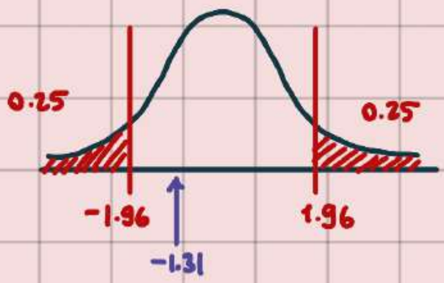
$$= \frac{(.50 - .62)}{\sqrt{\left(\frac{.549(1-.549)}{72} + \frac{.549(1-.549)}{50}\right)}}$$

$$= \boxed{-1.31}$$

$H_0: P_m - P_w = 0$ , $\hat{p}_m = 36/72 = 0.50$

$H_1: P_m - P_w \neq 0$ , $\hat{p}_w = 31/50 = 0.62$

$$\hat{p}_0 = \frac{n_M\hat{p}_M + n_W\hat{p}_W}{n_M + n_W} = \frac{72(36/72) + 50(31/50)}{72 + 50} = \frac{67}{122} = .549$$



→ Do not Reject Ho

## HYPOTHESIS TESTS for Two VARIANCES:

| | |
|---|---|
| $H_0: \sigma_x^2 \geq \sigma_y^2$<br>$H_1: \sigma_x^2 < \sigma_y^2$ | **Lower-tail test** |
| $H_0: \sigma_x^2 \leq \sigma_y^2$<br>$H_1: \sigma_x^2 > \sigma_y^2$ | **Upper-tail test** |
| $H_0: \sigma_x^2 = \sigma_y^2$<br>$H_1: \sigma_x^2 \neq \sigma_y^2$ | **Two-tail test** |

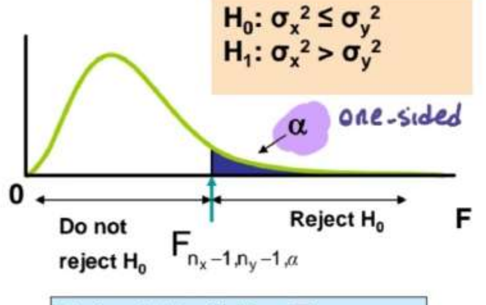
$$F = \frac{s_x^2}{s_y^2}$$

→ Has an F distribution with (nx−1) numerator degrees of freedom and (ny−1) denominator degrees of freedom.

Denote an F value with $v_1$ numerator and $v_2$ denominator degrees of freedom by $F_{v_1, v_2}$

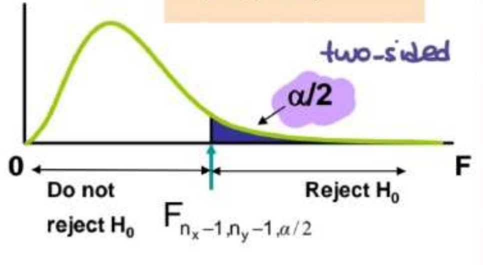
**Use $s_x^2$ to denote the larger variance.**

$H_0: \sigma_x^2 \leq \sigma_y^2$
$H_1: \sigma_x^2 > \sigma_y^2$

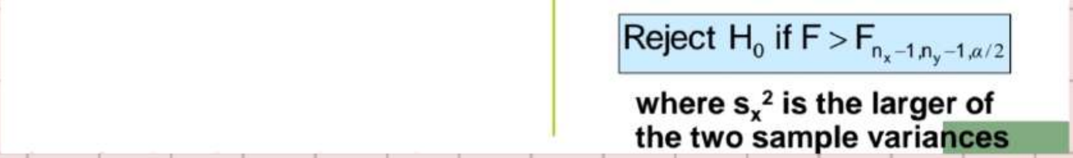


Reject $H_0$ if $F > F_{n_x-1, n_y-1, \alpha}$

$H_0: \sigma_x^2 = \sigma_y^2$
$H_1: \sigma_x^2 \neq \sigma_y^2$



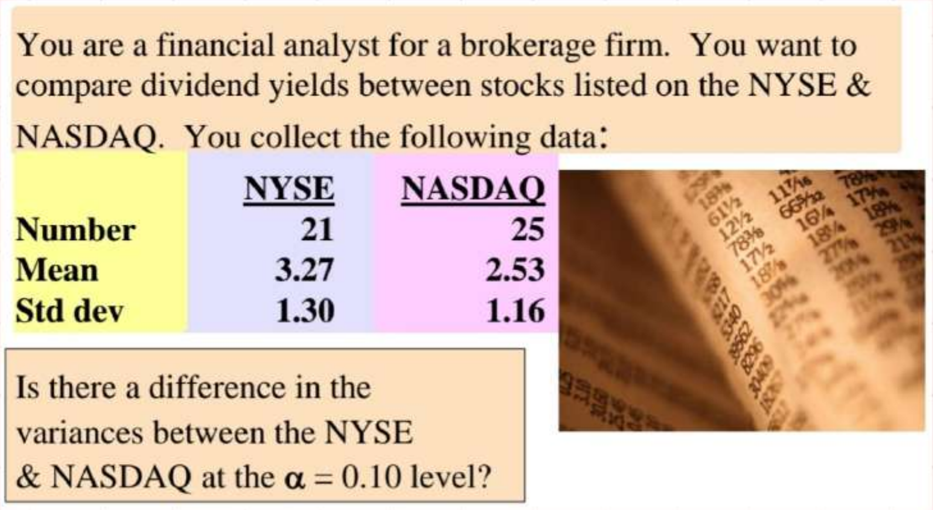
- **rejection region for a two-tail test is:**



Reject $H_0$ if $F > F_{n_x-1, n_y-1, \alpha/2}$

where $s_x^2$ is the larger of the two sample variances

## Example



You are a financial analyst for a brokerage firm. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

| | NYSE | NASDAQ |
|---|---|---|
| **Number** | 21 | 25 |
| **Mean** | 3.27 | 2.53 |
| **Std dev** | 1.30 | 1.16 |

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.10$ level?

$H_0: \sigma_x^2 = \sigma_y^2$

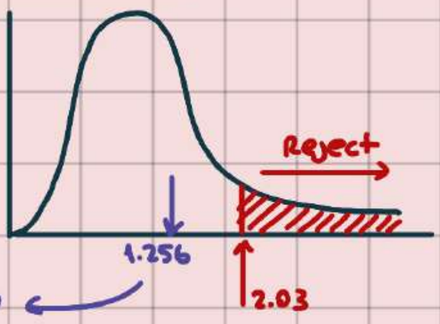
$H_1: \sigma_x^2 \neq \sigma_y^2$

$n_x - 1 = 21 - 1 = 20$ d.f.

$n_y - 1 = 25 - 1 = 24$ d.f.

$F_{n_x-1,\, n_y-1,\, \alpha/2} = F_{20,24,0.10/2} = 2.03$

$F = \dfrac{s_x^2}{s_y^2} = \dfrac{1.30^2}{1.16^2} = 1.256$


Reject
1.256
2.03
Do not Reject $H_0$