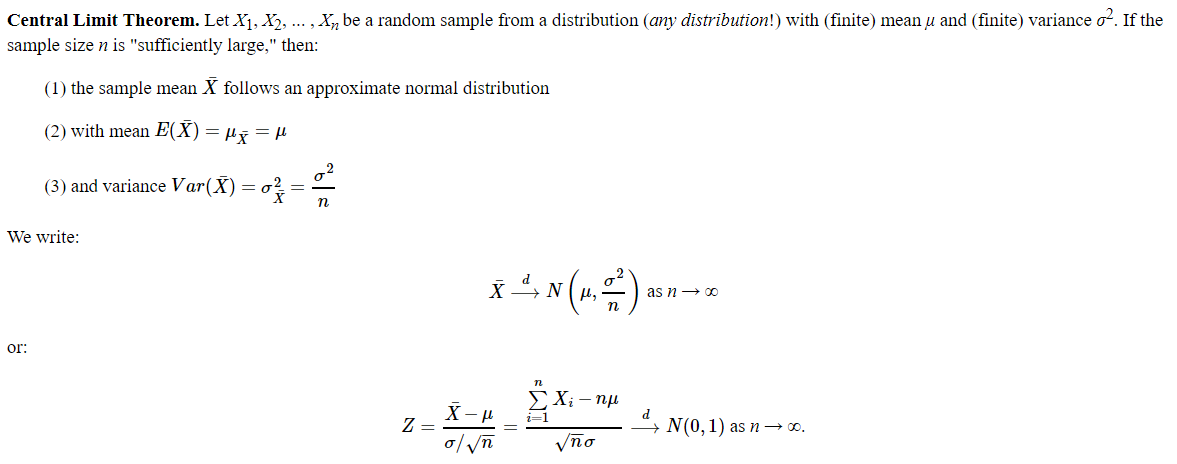
**Learning outcomes:**

After solving these exercises, you should be able to understand the following:

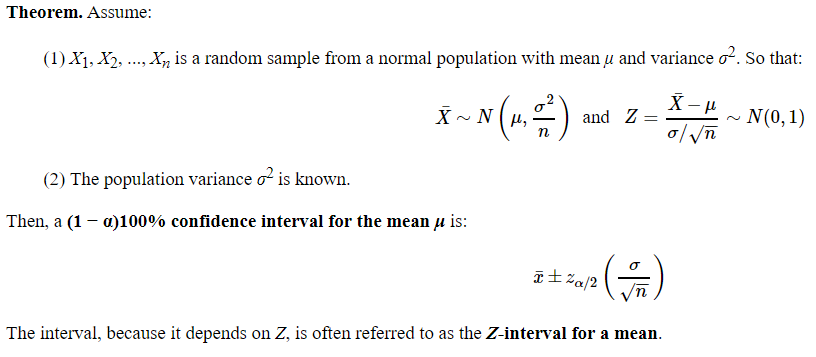
* **Central Limit Theorem**
* **Confidence Intervals**
* **Hypotheses Testing**

**Key takeaways:**

**Central Limit Theorem**



**Confidence Intervals**

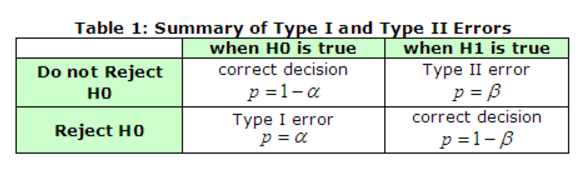


* **Margin of error:** Margin of error is the **maximum expected difference** between the true population parameter and a sample estimate of that parameter
* Remember:

|  |  |
| --- | --- |
| **Level of Confidence** | **Value of z** |
| 90% | 1.64 |
| 95% | 1.96 |
| 99% | 2.58 |

**Hypotheses Testing**

* **Statistical hypothesis**, or simply a hypothesis, is an assumption about a population parameter.
* **Hypothesis testing** is the procedure whereby we decide to “reject” or “fail to reject” a hypothesis.
* **Null hypothesis H0**: This is the hypothesis (assumption) under investigation or the statement being tested. The null hypothesis is a statement that “there is no effect,” “there is no difference,” or “there is no change.” The possible outcomes in testing a null hypothesis are ‘reject’ or ‘fail to reject.’
* **Alternate hypothesis H1**: This is a statement you will adopt if there is strong evidence (sample data) against the null hypothesis. A statistical test is designed to assess the strength of the evidence (data) against the null hypothesis.
* **Fail to Reject H0**: We never say we “accept H0” - we can only say we “fail to reject” it. Failing to reject H0 means there is NOT enough evidence in the data and in the test to justify rejecting H0. So, we retain the H0 knowing we have not proven it true beyond all doubt.
* **Rejecting H0:** This means there is significant evidence in the data and in the test to justify rejecting H0. When H0 is rejected the data is said to be statistically significant. We adopt H1 knowing we will occasionally be wrong.



1. **Central Limit Theorem**
   1. A population of 25-28 year-old males has a mean salary of $29,321 with a standard deviation of $2,120. If a sample of 100 men is taken, what is the probability their mean salaries will be less than $29,000?

pnorm(29000,29321,2120/sqrt(100))

0.06499378

* 1. The engines made by Ford for speedboats had an average power of 220 horsepower (HP) and standard deviation of 15 HP.  A potential buyer intends to take a sample of forty engines and will not place an order if the sample mean is less than 215 HP. What is the probability that the buyer will not place an order?

pnorm(215,220,15/sqrt(40))

0.01750749

1. **Confidence Intervals**
   1. Identify the situation in which you can construct the confidence intervals
      1. Construct the confidence interval for the mean score of 10 students

We don’t do confidence interval for sample but we do it for population.

We know mean of 10 students so we don’t need to construct confidence interval. It’s needed only when we do inferential statistics

* + 1. Construct the confidence interval for the mean salary of a city
  1. A random sample of 100 items is taken, producing a sample mean of 49. The population std. deviation is: 4. 49.
     1. Compute the margin of error and construct a 90% confidence interval to estimate the population mean.

Std. Error = 4.49/sqrt(100) = 0.449

Margin = Z \* SE

Z score for 90% confidence = 1.64

Margin = 1.64 \*0.449 = 0.73636

Confidence Interval = 49 – 0.736, 49 + 0.736

= 48.264, 49.736

marginError <- qnorm(0.05,lower.tail = F)\*4.49/10

49 - marginError

49 + marginError

* + 1. If the desired margin of error at 95% confidence is 1% then what should be the sample size?

M.E (Margin of error ) = 1%

Z \* SE = 1/100

1.96 \* S.D/sqrt(n) = 1/100

1.96 \* 4.49/sqrt(n) = 1/100

Sqrt(n) = 880.4

N = 774470

* 1. Click fraud has become a major concern as more and more companies advertise on the internet. When Google places an ad for a company with its search results, the company pays a fee to Google each time someone clicks on the link. That’s fine when it’s a person who’s interested in buying a product or service, but not so good when it’s a computer program pretending to be a customer. An analysis of 1200 clicks coming into a company’s site during a week identified that 175 of these clicks are fraudulent. Compute the confidence interval with 95% confidence for the proportion of fraudulent clicks

Z = 1.96

P = 175/1200 = 0.14583

q = 1 – 0.14583 = 0.854

Confidence Interval = P – z \* sqrt(p\*q/n) , P + z \* sqrt(p\*q/n)

0.14583 – 1.96 \* sqrt(0.14583 \* 0.854/1200), 0.14583 + 1.96 \* sqrt(0.14583 \* 0.854/1200)

0.1258628, 0.1657972

p <- 175/1200

q <- 1-p

lowerLimit <- p - 1.96\*sqrt(p\*q/1200)

UpperLimit <- p + 1.96\*sqrt(p\*q/1200)

* 1. A random sample of 35 items is taken, producing a sample mean of 2.364 with a sample variance of 0.81. Assume x is normally distributed and construct a 90% confidence interval for the population mean.

We have sample variance and not population variance. If sample variance is given using

T-distribution.

**Below is done using T-distribution:**

n <- 35

mean <- 2.364

var <- 0.81

90% confidence internval

tvalue <- qt(0.05,34,lower.tail = F) # we have 34 degrees of freedom

2.364 - tvalue \* 0.9/sqrt(35)

2.364 + tvalue \* 0.9/sqrt(35)

2.106763, 2.621237

* 1. The average zinc concentration recovered from a sample of zinc measurements in 36 different locations 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. Does the increase in the confidence level increases the confidence interval?

N = 36

95% Z – 1.96

ME for 95% = 1.96 \* 0.3/6 = 0.098

2.6 – 0.098, 2.6 + 0.098

2.502, 2.698

99% Z – 2.58

M.E for 99 = 2.58 \* 0.3/6 = 0.129

2.6 – 0.129, 2.6 + 0.129

2.471, 2.729

Yes increase in confidence level increases the confidence interval.

# 95%

n <- 36

z <- qnorm(0.025,lower.tail = F)

ME <- z \* 0.3/sqrt(n)

lowerLimit <- 2.6 - ME

upperLimit <- 2.6 + ME

c(lowerLimit,upperLimit)

# 99%

z <- qnorm(0.005,lower.tail = F)

ME <- z \* 0.3/sqrt(n)

lowerLimit <- 2.6 - ME

upperLimit <- 2.6 + ME

c(lowerLimit,upperLimit)

* 1. The life in hours of a 75- watt light bulb is known to be normally distributed with σ = 25 hours. A random sample of 100 bulbs has a mean life of x ̅ = 1014 hours. Construct a 95 % two-sided confidence interval on the mean life.

Z score 95% - 1.96

ME = 1.96 \* 25/10 = 4.9

1014 – 4.9, 1014 + 4.9

marginError <- qnorm(0.025,lower.tail = F)

lowerLimit <- 1014 - marginError\*25/10

upperLimit <- 1014 + marginError\*25/10

lowerLimit - 1009.1

upperLimit - 1018.9

1. **Hypothesis Testing**

3.1 State the null and alternative hypotheses to be used in testing the following claims and determine generally where the critical region is located:

a. The mean snowfall at Lake George during the month of February is 21.8 centimeters

H –null -> average snowfall = 21.8

H –alternative -> average snowfall != 21.8

Critical region – same as reason for rejection

Average > 21.8 or Average <21.8 we can reject, both sides we have critical region

b. No more than 20% of the faculty at the local university contributed to the annual giving fund

H –null -> mean contribution < 0.2

H –alternative -> mean contribution > 0.2

As it is greater than 0.2 , critical region will be on right side

3.2 Suppose a manufacturer claims that the mean lifetime of a light bulb is at least 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assuming the population standard deviation to be 120 hours, at 0.05 significance level, can we reject the claim by the manufacturer?

alpha <- 0.05

~~limitOnCriticalZone <- qnorm(0.05,lower.tail = F) # 1.644854~~

options(scipen = 2)

actual <- pnorm(9900,10000,120/sqrt(30)) # output 0.000002505166

actual

As actual < alpha, we can reject Manufacturer's claim

**Another way**

mean <- 10000

n = 30

smean = 9900

sd = 120

zval <- (9900-10000)/(120/sqrt(30))

zval # -4.56

limitOnCriticalZone <- qnorm(0.05,lower.tail = F) # 1.644854

-4.56 < 1.64 reject the claim

3.3 In a confidential meeting, a manager claims that the customer satisfaction can be at most 80%. A new analyst wants to test this claim. What would be his null and alternative hypotheses? If he surveyed 100 people of which 84 people were satisfied. With 0.05 significance level, what is his finding?

Null hypothesis -> atmost 80% customers are satisfied. (mean <= 80)

Alternative hypothesis -> satisfied customers can be more than 80% (mean >80)

If Pvalue > Alpha

n = 100, p =0.84, q = 0.16

mean = 84, SD = sqrt(100\*0.84\*0.16) = 3.67

1 - pnorm(80,84,3.67) # 0.86

P value = 0.86 < alpha 0.95, I will go with the claim(accept H0)

0.95 is 1 – 0.05 as it is right side

OR

If Z value > Z critical value , reject hypothesis

qnorm(0.95)= 1.644854

cal <- (84-80)/ sqrt(100\*0.84\*0.16) = 1.091

1.091 > 1.644854 which is not true, so going with managers claim.

1. Computing the Type 1 , Type 2 and Power of the Tests

Type I error : Reject Ho when Ho is true

Type II error : Not Reject Ho when Ho is false

Power : Reject Ho when H0 is false

Both power and Type II error depends on the n, alpha, mean and Standard deviation

Suppose we are about to randomly sample 16 values from a normally distributed population where sigma = 8 but mean is unknown. We are going test . Assume alpha = 0.05

Ho : Mean (Mu) = 75

H1 : Mean (Mu) != 75

For what values of Z will we reject Ho ?

It’s a two tailed test.

Critigal region is 0.05, so on each side it is 0.025

Z(0.025) = 1.96

So for two sides it is +/- 1.96

Z = (x – mean)/ ( sd/sqrt(n))

For +/- 1.96, x values are 71.08,78.92

For values not falling in between 71.08 and 78.92, we will reject H0.

If the sample mean is 77 , compute the power of the test and type 2 errors ?

Power(test) = 1 – beta

= power(reject H0 when H0 is false)

= prob(Z1 < Z < Z2 | Given mean = 77)

= 1 – pnorm(78.92,77,8/sqrt(16))

= 1 – 0.83

= 0.168

There type2 error is 0.83 and power of test is 0.168

5. Suppose we are about to randomly sample 36 values from normally distributed population, where sigma = 21 but mean is unknown. We are going to test

Ho : mean = 50 Halt : mean < 50 at alpha = 0.10

1. For what values of Z will we reject Ho

Given S.D = 21, mean is unknown.

By default 95% confidence interval, then alpha = 0.05

It’s a one tailed test as per hypothesis.

Z = Qnorm(0.05) = -1.64

Z = (x-mean )/(sd/sqrt(n))

-1.64 = (x – 50)/(21/6)

X = 44.29

For any value less than 44.29 we can reject H0.

II. Compute the Type II and power of the test if the sample mean =43?

Type-2 error = 1 – pnorm(44.3,43,21/6)

Type 2 error = 1 - 0.6448408

Type 2 error = 0.355

Power test = 1 – type2 error

= 1 – 0.355

= 0.645