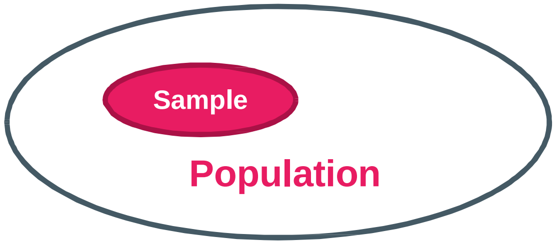
Inferential Statistics & Hypothesis Testing

V. Sridhar
IIIT Bangalore

Inferential Statistics

Sampling and Sampling Distributions

- In many instances, one cannot study an entire population. Why?
- In such a case, one selects a subset of the population, called a sample, and inspects the attributes of each item in the sample
- Based on the findings from the sample, one makes conclusion about the entire population.
    - For example, if one finds 3 percent of the items in the sample as defective, the conclusion is made that 3 percent of the items in the population is defective



Sample and Point Estimation

- Now that we know how to select a sample, let's use the sample to estimate population characteristics (mean, and proportion)
  - Using sample data to estimate a population mean or proportion is known as Point Estimation
- Central Limit Theorem:
  - The relationship between the shape of the population distribution and the shape of the sampling distribution of the sample statistic

Poll

- Suppose heights of all women have a standard deviation of 2.7 inches, and a random sample of 100 women's heights yields a standard deviation of 4 inches.
- Which one is the population parameter?
  - 2.7″ or 4″

Difference between SD and SE

- standard deviation measures the variability in the data, while standard error measures the variability in point estimates from different samples of the same size and from the same population, i.e. measures the sampling variability.

Sampling and Sampling Distribution

- MRF Tyres wants to know the mean (or average) life of its new brand of ZLX tyres
  - One way is testing and wearing out each tire manufactured. Obviously, this does not make sense.
- MRF in their tyres stock, takes a sample of tyres, tests and wears out each of these tyres and then calculates the mean (or average) life of the sampled tyres
- Suppose, the mean life is calculated as 42,000 kms
  - Based on this sample, it is concluded that the mean life all new brand of tyres (that is population) is 42,000 kms.

Central Limit Theorem

CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$\text{Mean} = \mu \qquad\qquad \text{Standard Error } (SE) = \frac{\sigma}{\sqrt{n}}$$

- **Independence:** The sample observations must be independent
  - The most common way to satisfy this condition is when the sample is a simple random sample from the population
- **Normality:** When a sample is small, we also require that the sample observations come from a normally distributed population
  - We can relax this condition more and more for larger and larger sample sizes.

Sample and Population Point Estimates

Parameter	Formula
Population Mean	μ
Population Variance, Standard Deviation	$\sigma^2, \sigma = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{(n-1)}}$
Sample Mean	$\bar{x}$
Sample Variance, Standard Deviation	s^2, s
Standard Error of the Mean	$^{a}/_{\sqrt{n}}\ ^{c}/_{\sqrt{n}}\ \sqrt{\frac{p(1-p)}{n}}$
Test Statistic (Z/t)	$\frac{Point\ Estimate - Null\ Value}{SE}$

Rule of Thumb

RULES OF THUMB: HOW TO PERFORM THE NORMALITY CHECK

There is no perfect way to check the normality condition, so instead we use two rules of thumb:

$n < 30$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

$n \geq 30$: If the sample size n is at least 30 and there are no *particularly extreme* outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Estimation

- Any sample statistic that is used to estimate a population parameter is called an estimator. An estimate is a specific observed value of a statistic.

Determining the Sample Size in Estimation

- If the sample size is too small, we may fail to achieve the objective of our analysis, but if it is too large, that will be waste of resources.
  - Some sampling error will arise because we have not studied the whole population. Sampling error is controlled by selecting a sample that is adequate in size.
- If we want a high level of precision, we have to sample enough of the population to provide the required information.

Confidence Intervals

Interval Estimates

- An interval estimate describes a range of values within which a population parameter is likely to appear.
  - Interval estimate is constructed using point estimate, standard error and corresponding probability
- The probability that we associate with an interval estimate is called the *confidence level*
  - The *confidence interval* is the range of the estimate we are making. *Confidence limits* are the upper and lower limits of the confidence interval
- 95% confidence interval means: "That if we select many random samples of the same size and calculate a confidence interval for each of these samples, then in about 95 percent of these cases, the population parameter will lie within that interval."

Poll

- If the confidence interval of sample mean is large, then the probability of population means falling within the interval is:
  a) High
  b) Low
- If the confidence interval of sample mean is large, then the accuracy of sample mean relative to population mean is:
  a) High
  b) Low

Interval Estimates

- **General formula for interval estimate (Large sample):**

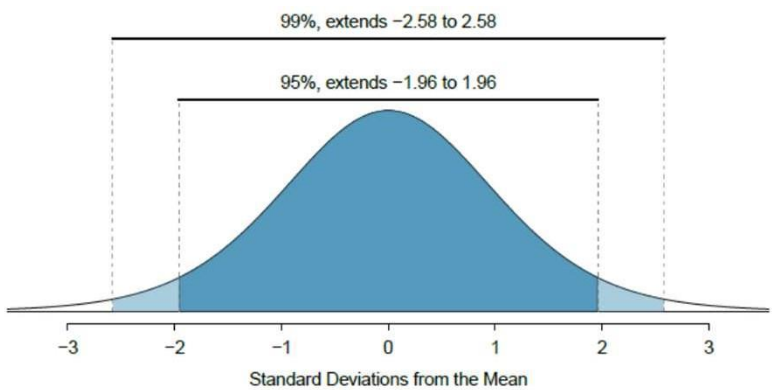
  Statistic + Z. Standard error (Statistic): Upper confidence limit

  Statistic – Z. Standard error (Statistic): Lower confidence limit

- **Interval estimate for mean of large sample (Population s.d. is known):**

  $$\bar{x} + z\left(\times \frac{\sigma}{\sqrt{n}}\right),\ \bar{x} - z\left(\times \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Interval from Point Estimate



- 99% CI -> $\bar{x} \pm 1.96 \times SE = \bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{\ }}$
- Margin of Error: z $\times SE$
- Confidence Intervals are always about population estimates
  - They provide plausible range of population parameters

Poll

- Which of the following is false about confidence intervals? All else held constant.
  - as the confidence level increases, the width decreases.
  - as the standard deviation of the sample increases, the width increases.
  - as the sample mean increases, the margin of error stays constant.
  - As the sample size increases, margin of error decreases
- If a given value (for example, the null hypothesized value of a parameter) is within a 95% confidence interval, it will also be within a 99% confidence interval. Yes/ No

Quiz

- In Bangalore on February 23rd, 2020, a doctor who had recently been treating Influenza patients went to the hospital with a high-grade fever and was subsequently diagnosed with Covid-19
    - Soon thereafter, the Republic TV conducted a survey and found that 82% of Bangaloreans favored a mandatory 14-day quarantine for anyone who has come in contact with a Covid patient
    - This poll included responses of 1,042 Bangaloreans between Feb 26th and 28th
- What is the point estimate in this case?
- If the SE of this opinion is 0.012, what is the 95% CI for the point estimate?
    - And what can you say about the general opinion on quarantine by Bangaloreans?
- Can we say with 95% confidence that 80% of the Bangaloreans are likely to opine in favour of quarantine?
    - Can we say with 99% confidence that 80% of the Bangaloreans are likely to opine in favour of quarantine?

Confidence Interval Estimation

(1)

~~Test Statistic~~

Point estimate $\bar{p} = 0.82$

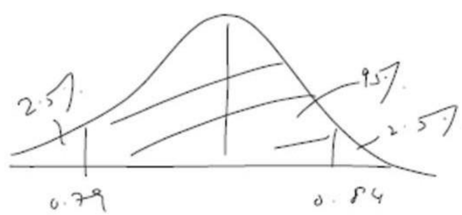
SE of the estimate : 0.012

95% CI for $\bar{p}$

$\bar{p} \pm 1.96 \times SE$

$0.82 \pm 1.96 \times 0.012$

$\{0.8435, 0.7965\}$.



Population mean $\bar{\mu}$ = ?

With 95% confidence level we can say that μ lies in $\{0.7965, 0.8435\}$

Hypothesis Testing

Testing Hypothesis Procedure

- Hypothesis testing begins with an assumption, called a hypothesis, that we make about a population parameter.
  - Hypothesis testing is about making inferences about a population from only a small sample
- In hypothesis testing, we must state the assumed or hypothesized value of the population parameter before we begin sampling
  - The assumption we wish to test is called the null hypothesis.
  - Whenever we reject the hypothesis, the conclusion we do accept is called alternative hypothesis.

Hypthesis Testing

- Type of Tests
  - One-Tailed
  - Two Tailed
- Test Statistic
  - Sample > 30 — Z
  - Sample <= 30 — T
- Level of Significance
  - 95%
  - 99%

Hypothesis Testing

Decision on the Test	In Reality (The population)	
	H_0	H_1
Accept: H_0		β :Type II Error
Reject: H_0	α : Type I Error	$(1 - \beta)$ Power

- Ideally α & β should both be small
  - α & β are inversely related
- Power of the test indicate how well your test works
  - Large sample size increases power of the test

Which one should we reduce α or β?

Decision on Conviction	Reality of the Defendant	
	Innocent	Perpetrator
Accept H_0 Let Defendant Free		β :Type II Error
Reject: H_0 Convict the Defendant	α : Type I Error	$(1-\beta)$ Power

- H0: Accused is Innocent
- H1: Accused is a Guilty

Should $\alpha > \beta$; or $\alpha < \beta$

The **presumption of innocence** is a legal principle that **every person accused of any crime is considered innocent until proven guilty**

What is the Null and Alternate Hypothesis

- A tutoring company would like to understand if most students tend to improve their grades (or not) after they use their services.
- They sample 200 of the students who used their service in the past year and ask them if their grades have improved or declined from the previous year

What is the Null and Alternate Hypothesis

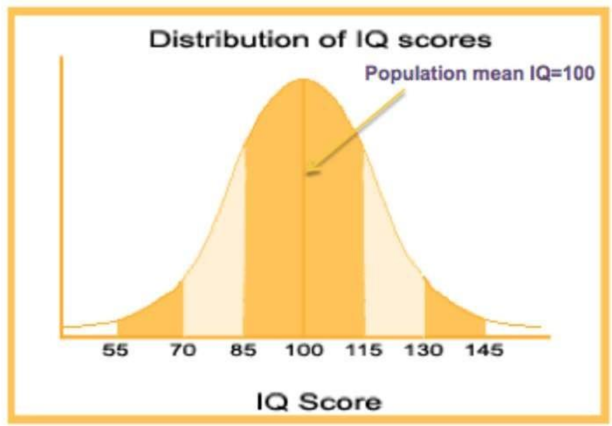
- **A study suggests that 60% of college student spend 10 or more hours per week communicating with others online.**
- **You believe that this is incorrect and decide to collect your own sample for a hypothesis test.**
- **You randomly sample 160 students from your hostel and find that 70% spent 10 or more hours a week communicating with others online.**

Tests of Significance

Tests of Significance

Population Mean: IQ=100

Is the sample mean significantly different than the population mean?



Population mean: IQ=100

Population st dev=16

Sample mean: IQ=108

Sample size: N=16

- In order to determine if two numbers are *significantly different*, a statistical test must be conducted to provide evidence
    - Researchers must collect statistical evidence to make a claim, and this is done by conducting a test of statistical significance.

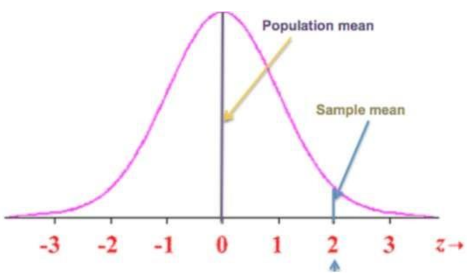
Tests of Significance

1. State the Hypothesis
   - $H_0: \mu = 100$
   - $H_1: \mu \neq 100$
2. Which Test to perform?
   - Should we do one-sided test? Or two-sided?
   - If $H_1: \bar{x} > \mu (> 100)$, which test?
3. What test statistic to use
   1. z or t?
4. Compute the test statistic

Population mean: IQ=100
Population st dev=16
Sample mean: IQ=108
Sample size: N=16

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

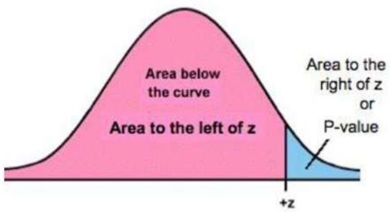
Test Statistic



$$\frac{(108 - 100)}{\left(16 / \sqrt{16}\right)} = 2$$

Tests of Significance

$H_0: \mu = 100)$
$H_1: \mu > 100)$



- Assuming H_0 is TRUE
  - Find out the probability of obtaining this score when the null hypothesis is true
- Area to the left of z = the probability of obtaining scores lower than z
  - Area to the right of z (p-value) = the probability of obtaining scores higher than z
  - The smaller the p-value, the stronger the evidence against H_0 provided by the data.

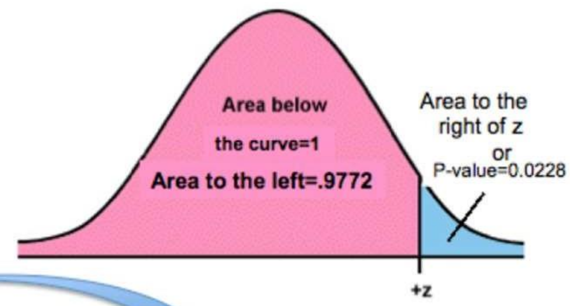
The p-value

- The p-value represents the probability of obtaining scores that are at the z level or higher when the null hypothesis is true
  - In other words, what percent chance exists of getting this specific sample mean score if it is actually no different from the population mean.
  - If z is far away from the mean, the p-value is small. The larger the test statistic (the farther from the mean), the smaller the p-value.
- When the p-value is very small, researchers can say they have strong evidence that the null hypothesis is FALSE
  - This is because if the p-value is very small, it means that the probability of obtaining a score that is so extreme or even higher is very small, if indeed the Null Hypothesis is TRUE
    - Hence decision is to Reject the Null Hypothesis in favour of Alternative Hypothesis

Tests of Significance

Example:

Table A provides the area to the left of z:



Z	.0	.01	.02
1.9	13	.9719	9726
2.0	.9772	.9778	.9783
2.1	.9821	.9826	.9830

z=2.00 => Area to the left=0.9772

P_value=1-0.9772 => P_value=0.0228

=> “If the population mean IQ is equal to 100, there is a 2% probability of recording a sample mean of 108.

- This means that there is only a 2% chance that the null hypothesis is true.
- In other words, if the population mean is 100, then there is only a 2% chance of having a sample mean equal to 108.

Relationship between p-value and Significance level

- Choosing a significance level α for a test is important in many contexts, and the traditional level is α = 0.05
- If making a Type 1 Error is dangerous or especially costly, we should choose a small α level (e.g. 0.01)
  - Under this scenario we want to be very cautious about rejecting the null hypothesis,
    - so we demand very strong evidence favoring HA before we would reject H0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we might choose a higher α (e.g. 0.10)
  - Here we want to be cautious about failing to reject H0 when the alternative hypothesis is actually true

Summary

1. State the null and alternative hypotheses.
2. Calculate the test statistic.
3. Find the *P*-value (using a table or statistical software).
4. Compare P-value with the threshold Significance Level: α and decide whether the null hypothesis should be rejected or accepted
    1. If the p-value is less than α, then there is a strong evidence for rejecting the Null Hypothesis
    2. Otherwise, there is not a stronger evidence and hence cannot reject the Null Hypothesis

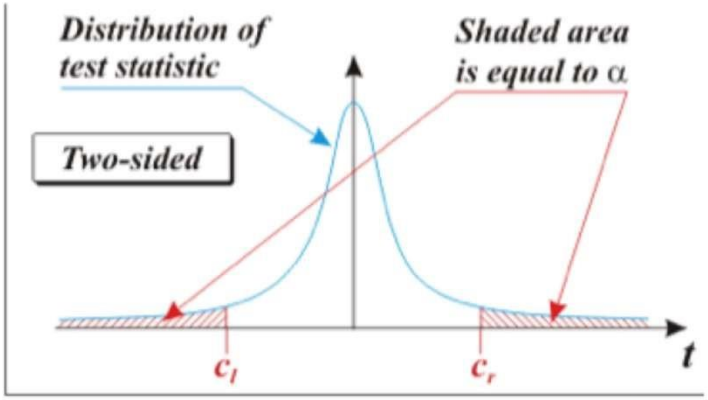
Two-Sided Test

When the alternative hypothesis is two sided:

P-value=Area to the right of the curve (from Table A)*2

Example:
Null Hypothesis: Sample IQ=100
Alternative Hypothesis: Sample IQ≠100

Z=2
Area to the right=.0228
P_Value=.0228*2= .0456



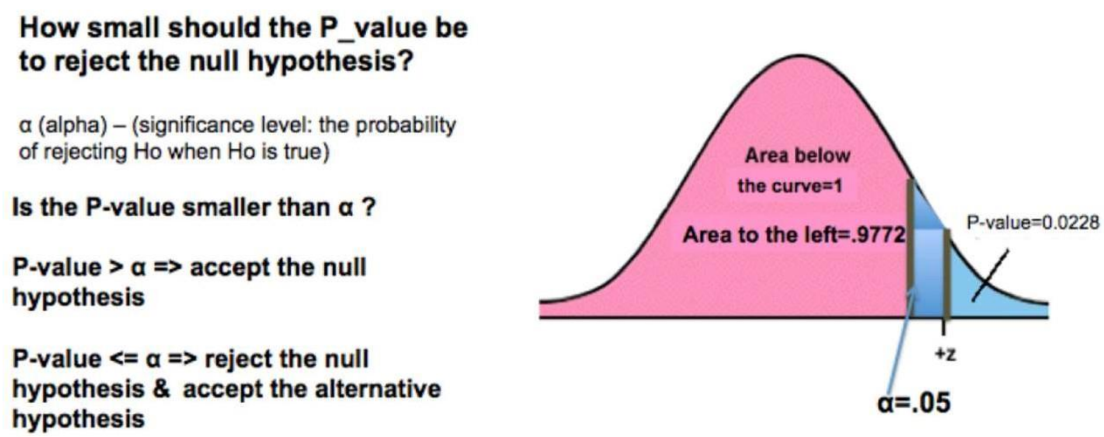
Example:

α =.05
p-value=0.04

p-value<α => Decision: Reject the null hypothesis & accept the alternative

Conclusion: "The test of significance provided evidence that the sample IQ is significantly different than 100."

Relationship between *p* value and α



- **Poll: If your p-value is greater than α then what should you do:**
  - **Accept the null hypothesis**
  - **Reject H_0 and accept H_1**

Polls

- Decreasing the significance level α will increase the probability of making a Type 1 Error Yes/ No

Quiz

- In Bangalore on February 23rd, 2020, a doctor who had recently been treating Influenza patients went to the hospital with a high-grade fever and was subsequently diagnosed with Covid-19
  - Soon thereafter, the Republic TV conducted a survey and found that 82% of Bangaloreans favored a mandatory 14-day quarantine for anyone who has come in contact with a Covid patient
  - This poll included responses of 1,042 Bangaloreans between Feb 26th and 28th
- However Republic TV announces that 90% of the Bangaloreans are in favour of quarantine
  - What is the null and alternative hypothesis?
  - What is the p-value?
  - With $\alpha = 0.05$, should you reject or not reject the null hypothesis?
  - How is it related to the 95% CI estimate that we calculated previously?
- What is Republic TV is conservative and say that 80% of Bangaloreans are in favour of quarantine?

https://www.ztable.net/

Play the Veena

- Meera claims that in Karaikudi - a small city renowned for its music school, the average child takes less than 5 years of Veena lessons.
- We have a random sample of 20 children from the city, with a mean of 4.6 years of Veena lessons and a standard deviation of 2.2 years.

a) Evaluate Meera's claim (or that the opposite might be true) using a hypothesis test.

b) Construct a 95% confidence interval for the number of years students in this city take piano lessons, and interpret it in context of the data.

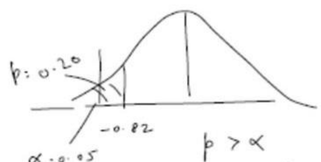
c) Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

Solution (7.11 of Open Stats)

(a) $H_0: \mu = 5$
$H_1: \mu < 5$ { one sided test)

(b)
$\bar{x} = 4.6$
$s = 2.2$
$n = 20$
$SE = \frac{s}{\sqrt{n}} = \frac{2.2}{\sqrt{20}} = 0.49$

Test statistic $t = \frac{(4.6-5)}{0.49}$
$= -0.82$
$df = (20-1) = 19$



$p > \alpha$
Do not reject H_0.

(c) CI estimation:
t^* for $\alpha = 0.05 = 1.729$

90% CI : $\bar{x} \pm t^* \times SE$
$4.6 \pm 1.729 \times 0.49$ $\mu = 5$ lies in the CI.
{3.76, 5.44} Do Not reject

References

- Diez, D. M., Barr, C. D., & Çetinkaya-Rundel, M. (2015). OpenIntro Statistics. OpenIntro. Retrieved April, 18, 2019.
- Levin R., and Rubin, D. (1998). Statistics for management (7th Edition), Pearson.
- Gujarati, D. (2012). Econometrics by example. Macmillan
- **R-Tutorial on Elementary Statistics. Available at: http://www.r-tutor.com/elementary-statistics**
- Coursera: Inferential Statistics course offered by Mine Çetinkaya-Rundel, Duke University
- Distribution App: https://gallery.shinyapps.io/dist_calc/