**Session 9: Hypothesis testing**

1. **The concept of normal distribution**
2. **What is a Normal Distribution?**

* **Shape:** The normal distribution looks like a bell-shaped curve.
* **Symmetry:** It is perfectly symmetrical around the center.

A diagram of a normal distribution

Description automatically generated

1. **Key Characteristics:**

* **Mean (Average):** The center of the curve.
* **Standard Deviation:** Measures the spread of the data.
  + 68.2% of the data falls within 1 standard deviation of the mean.
  + 95.4% falls within 2 standard deviations.
  + 99.7% falls within 3 standard deviations.

1. **Why is it Important?**

* **Natural Occurrences:** Many natural phenomena follow this distribution (e.g., heights, test scores). For example, most students score around the average in a class, fewer scoring very high or very low.
* **Central Limit Theorem:** In large samples, the samples’ mean tend to be normally distributed. ([Video](https://www.youtube.com/shorts/TwctT3Ncm1w))
* **Statistical Inferences:** Helps in making predictions and decisions based on data.

1. **Hypothesis testing**
2. **What is Hypothesis Testing?**

* Hypothesis testing is a method used to decide whether there is enough evidence to support a particular claim about a population based on a sample of data.
* **Null Hypothesis (**H0**)**: This is the default statement that there is no effect or no difference. It assumes that any observed differences are due to random chance.

Example: "The average age is equal to 20."

* **Alternative Hypothesis (**H1**)**: This is what you want to prove, stating there is an effect or a difference.

Example: "The average age is not equal to 20."

1. **Procedure of hypothesis testing**

* State the null and alternative hypothesis. (e.g. , )
* Collect sample data.
* Calculate sample mean and stadard error ().
* Calculate t-statistics ( ).
* Compare absolute value of t-statistics |t| with critical values for given level of significance (). [1.65 (10% significance level), 1.96 (5%), 2.58 (1%)]

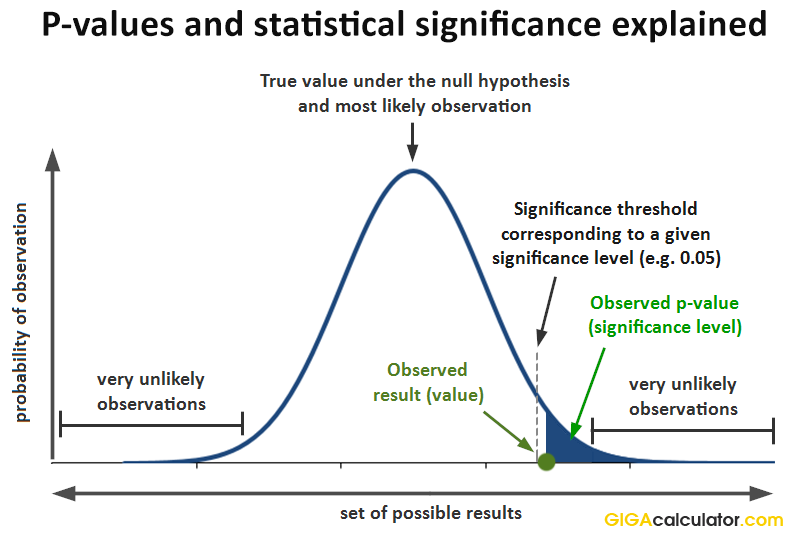
A diagram of a function

Description automatically generated

* Decision: reject null hypothesis if |t| exceeds critical value, otherwise fail to reject null hypothesis.

1. **Hypothesis testing with p-value**

* p-value : probability (area under normal distribution) beyond |t|.



* **Decision :** reject null hypothesis if p-value is lower than the significance level, otherwise fail to reject null hypothesis.
* Easier to conduct hypothesis testing with p-value. No need to calculate t-statistics and remember different critical values.

1. **Hypothesis testing in Stata**

\* Clear existing data

**clear**

\* Create a dummy dataset

**set** seed 12345

**set** obs 100

**gen** group = mod(\_n, 2)

**gen** **score** = 50 + group \* 10 + rnormal(0, 10)

\*conducting hypothesis testing

**ttest** **score** = 50 //H0: pop\_mean = 50

**ttest** **score** = 55 //H0: pop\_mean = 55

**ttest** **score** = 60 //H0: pop\_mean = 60

\* conducting two-sample t-test

**ttest** **score**, **by**(group) //H0: pop\_mean\_group1 = pop\_mean\_group2

//OR H0: pop\_mean\_group1 - pop\_mean\_group2 = 0

\*Same answer can be obtained from regression

**reg** **score** group

***Exercise:***

Using NMICS6 data (hl.sav), conduct a hypothesis test whether average age between male and female is statistically different.

import spss "https://gitlab.com/misc.a/referenced/-/raw/main/NMICS6/hl.sav", **clear**

\* HL6 -> Age, HL4 -> Sex

**sum** HL6 **if** HL4 == 1 //male : average age is 28.263

**sum** HL6 **if** HL4 == 2 //female : average age is 28.827

\*Looks like the population means for male and female are not statistically different.

\*Let's conduct the hypothesis testing

**ttest** HL6, **by**(HL4)

\*Alternatively

**reg** HL6 HL4

1. **Hypothesis testing using non-parametric approach (bootstraping)**

**Bootstrap :** generating distribution of statistics of interest by resampling the sample with replacement. Using Bootstrap, we can calculate standard errors, confidence intervals, and other statistical measures.

**clear**

**set** seed 1

**set** obs 100

**gen** **score** = round(runiform() \* 100)

\* Bootstrap the median and test against a specified value (e.g., 50)

bootstrap **r**(p50), reps(1000): **summarize** **score**, detail

\* Testing whether median is equal to 50 or not

**test** \_bs\_1 = 50

***Exercise:***

Using NMICS6 data (hl.sav), conduct a hypothesis test whether medeian age between male and female is statistically different.

import spss "https://gitlab.com/misc.a/referenced/-/raw/main/NMICS6/hl.sav", **clear**

**set** seed 12345

\* Define a program to calculate the difference in medians

**program define** diff\_medians, rclass

**summarize** HL6 **if** HL4 == 1, detail

**local** med0 = **r**(p50)

**summarize** HL6 **if** HL4 == 2, detail

**local** med1 = **r**(p50)

**return** **scalar** diff = `med1' - `med0'

**end**

\* Bootstrap the difference in medians

bootstrap **r**(diff), reps(100): diff\_medians

**Session 10: Regression analysis**

1. **Simple regression analysis**

**clear**

**set** seed 12345

**set** obs 100

**gen** study\_hours = round(runiform() \* 10)

**gen** **score** = 50 + 5 \* study\_hours + rnormal(0, 5)

**reg** **score** study\_hours

**A screenshot of a computer

Description automatically generated**

1. **Multiple regression and diagnostics**

**clear**

**set** obs 200

**gen** age = mod(\_n,52) + 18

**gen** educ\_year = mod(\_n,18)

\* Generate Income variable with a positive relationship with Age and Education

**gen** income = 20000 + 800 \* age + 3000 \* educ\_year + rnormal(0, 2000)

\* Regression with omitted variable

**reg** income age

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Residual diagnostics

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Residual visual inspection

**rvfplot**

\* Histogram plot for residual's distribution visualization

**predict** resid, residuals

**hist** resid

\*Formal test of residuals normality

**swilk** resid

**drop** resid

\* Multiple regression with correct specification

**reg** income age educ\_year

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* residual diagnostics

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Residual visual inspection

**rvfplot**

\* Histogram plot for residual's distribution visualization

**predict** resid, residuals

**hist** resid

\*Formal test of residuals normality

**swilk** resid

**Session 11: Advance regression with binary dependent variables (logit/probit)**

import spss "https://gitlab.com/misc.a/referenced/-/raw/main/NMICS6/hh.sav", **clear**

\* dropping missing values

**drop** **if** missing(HHSEX)

\* checking levels of HHSEX (Household Head Sex)

**codebook** HHSEX

**label** **list** labels410

**gen** hh\_size = HH48 //HH member size variable

**gen** urb\_rur = HH6 //1=Urban 2=Rural

**gen** province = HH7 //province number

\* generating binary dependent variable separately

**gen** hhsex\_male = 1

**replace** hhsex\_male = 0 **if** HHSEX == 2 //1=Male 2=Female

\*running logistic regression

**logit** hhsex\_male hh\_size ib1.urb\_rur ib3.province

margins, **dydx**(hh\_size urb\_rur province)

\* Similar results can be obtained using probit

\* Running probit regression

**probit** hhsex\_male hh\_size ib1.urb\_rur ib3.province

margins, **dydx**(hh\_size urb\_rur province)

**Session 12: Time series analysis**

1. **Stationarity concept**

* Stationarity refers to a time series whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time.
* Non-stationary series are prone to spurious relationships.

1. **Spurious relationship**

**clear**

**set** seed 1

**set** obs 100

**gen** **year** = 1900 + \_n

**tsset** **year**

**gen** ice\_cream\_sales = **year**\*10 + rnormal(0, 50)

**gen** shark\_attacks = **year**\*5 + rnormal(0, 20)

\* visual inspection for stationarity

twoway line ice\_cream\_sales **year**, name(ice\_cream\_sales, **replace**)

twoway line shark\_attacks **year**, name(shark\_attacks, **replace**)

**dfuller** ice\_cream\_sales //H0 : Non-stationary

**dfuller** shark\_attacks //H0 : Non-stationary

\* Run the initial regression (spurious relationship)

**reg** shark\_attacks ice\_cream\_sales

1. **Making series stationary to avoid spurious relationship**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Making Series Stationary

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Differencing variable makes series stationary

\* If a variable is stationary at first difference, then its called

\* I(1). I(0) means the variable is stationary at level.

twoway line D.ice\_cream\_sales **year**, name(ice\_cream\_sales, **replace**)

twoway line D.shark\_attacks **year**, name(shark\_attacks, **replace**)

**dfuller** D.ice\_cream\_sales //H0 : Non-stationary

**dfuller** D.shark\_attacks //H0 : Non-stationary

\*no relationship observed after differencing

**reg** D.shark\_attacks D.ice\_cream\_sales

\*\* log difference is preferred over simple difference as

\*\* interpretation of coefficient becomes easier.

**gen** lshark\_attacks = **log**(shark\_attacks)

**gen** lice\_cream\_sales = **log**(ice\_cream\_sales)

twoway line D.lice\_cream\_sales **year**, name(ice\_cream\_sales, **replace**)

twoway line D.lshark\_attacks **year**, name(shark\_attacks, **replace**)

**dfuller** D.lice\_cream\_sales //H0 : Non-stationary

**dfuller** D.lshark\_attacks //H0 : Non-stationary

**reg** D.lshark\_attacks D.lice\_cream\_sales

1. **Example of non-stationary series with actual relationship**

**clear**

**set** seed 1

**set** obs 100

**gen** **year** = 1900 + \_n

**tsset** **year**

**gen** income = **year**\*10 + rnormal(0, 50)

**gen** expenditure = income\*0.5 + rnormal(0, 20)

\* visual inspection for stationarity

twoway line income **year**, name(income, **replace**)

twoway line expenditure **year**, name(expenditure, **replace**)

**dfuller** income //H0 : Non-stationary

**dfuller** expenditure //H0 : Non-stationary

\* Run the initial regression

**reg** expenditure income

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Making Series Stationary

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**gen** lincome = **log**(income)

**gen** lexpenditure = **log**(expenditure)

\* visual inspection for stationarity

twoway line D.lincome **year**, name(income, **replace**)

twoway line D.lexpenditure **year**, name(expenditure, **replace**)

**dfuller** D.lincome //H0 : Non-stationary

**dfuller** D.lexpenditure //H0 : Non-stationary

\* Run the regression at first difference

**reg** D.lexpenditure D.lincome