Statistics & Explanatory Data Analysis

Statistcal inference

dr Marcin Chlebus, dr Ewa Cukrowska - Torzewska

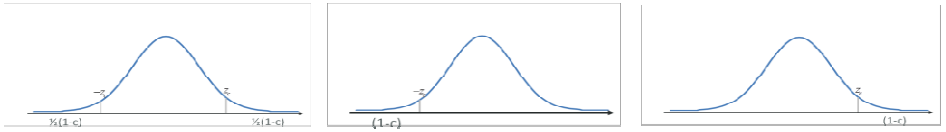
Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Introduction to hypothesis testing

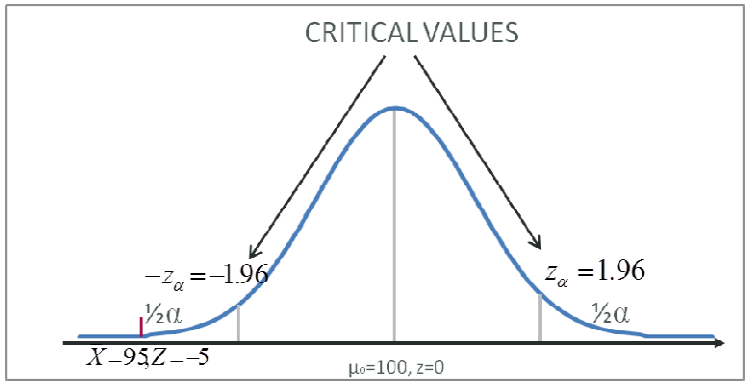
Hypothesis testing

- **Hypothesis test**
  - Process that uses sample statistics to test a claim about the value of a population parameter
  - Process of hypothesis testing
    - Stating a claim
      - Mean value of number of clients that would come to the shop within a day is equal to 100
    - Appropriate test choice (according to population parameter and to available data)
      - We have information from 100 days and average numer of clients per day is equal to 95, and standard deviation equal to 10 (z statistics)
    - Null and Alternative hypothesis definition

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$



    - Choice Level of significance α
      - Level of significance α defines how big part of the distribution which is true under the null hypothesis should be out of acceptance level – I type error definition



$$z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{95-100}{\frac{10}{\sqrt{100}}} = -5 \rightarrow N(0,1)$$

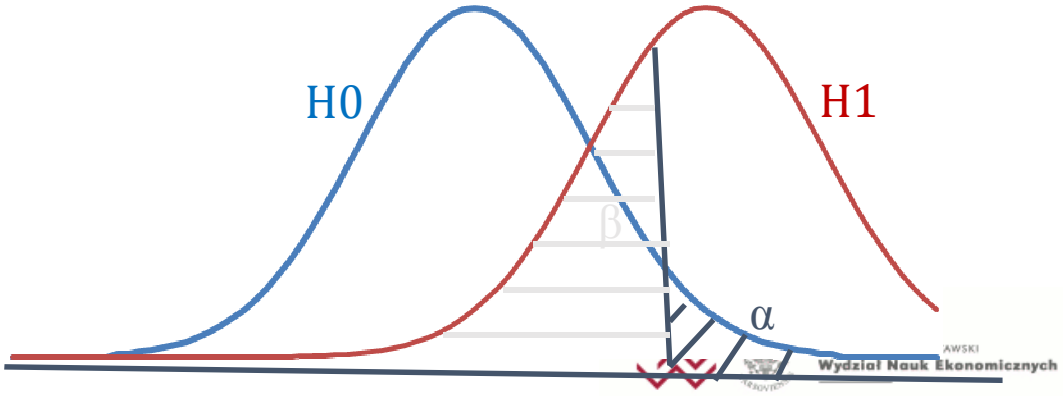
WE MAY:

- **REJECT NULL HYPOTHESIS**
- **FAIL TO REJECT NULL HYPOTHESIS**

Type I & II error

- **TYPE I: Null hypothesis is rejected when its true**
- **TYPE II: Null hypothesis is not rejected when it is false**

TEST RESULT	TRUTH OF H0	
	H0 IS TRUTH	H0 IS FALSE
DO NOT REJECT H0	CORRECT DECISION	TYPE II ERROR (β)
REJECT H0	TYPE I ERROR (α)	CORRECT DECISION



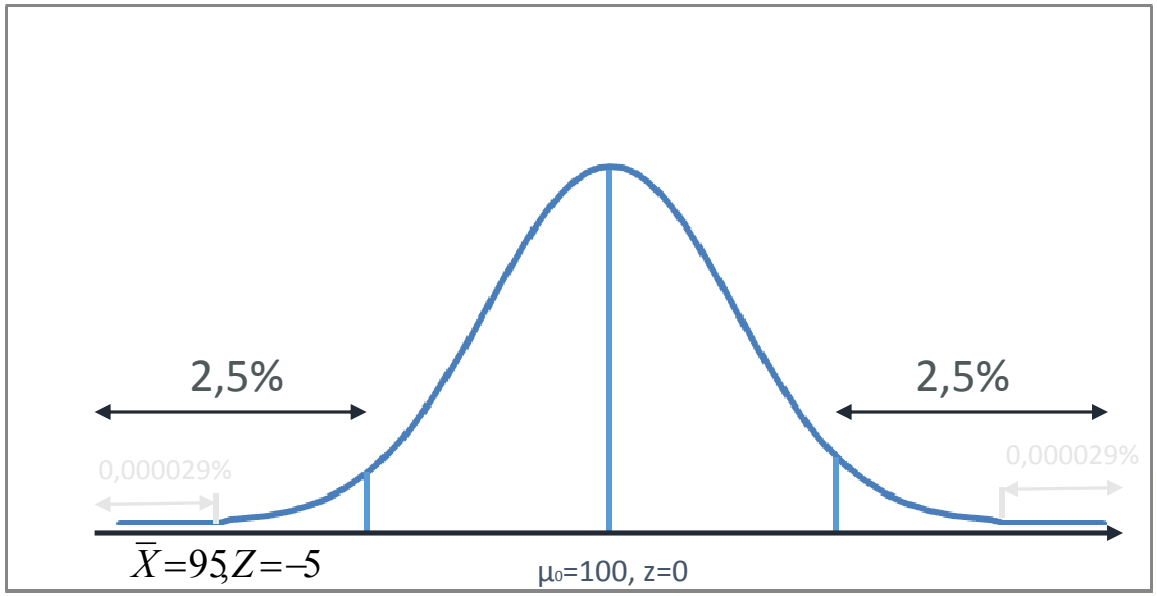
P-value concept

- **STANDARD HYPOTHESIS CONCEPT:**
  - Sample statistics value versus critical values

- **BUT: THERE IS ANOTHER APPROACH:**
  - p-value analysis
    - P-value → probability of obtaining sample statistics as extreme as or even more extreme than observed sample statistics (in absolute term) under HO
      - If p-value <= significance level → reject H0
      - If p-value > significance level → fail to reject H0
    - BENEFITS: We do not have to know precisely value of test statistics and critical value (simplicity of interpretation)

P-value example

$$p - value = P(-5 >= z \cup z >= 5)$$
$$= P(-5 >= z) + P(z >= 5) =$$
$$= P(-5 >= z) + 1 - P(z <= 5)$$
$$= \Phi(-5) + 1 - \Phi(5) =$$
$$= 0{,}00000029 + 1 - 0{,}99999971\ 3$$
$$= 0{,}00000057$$



Which test to choose?

- **What kind of dependent variable do I have?**
  - Is it **interval** or **ratio** type? → Parametric or nonparametric
  - **Ordinal** variable? → Nonparametric tests, or tests for ordinal data
  - **Nominal** variables? → Test for nominal variables
  - **Count** data? **Proportions**?
- What **kind of independent** and **how many** of them do I have?
- Are they **paired or not**?
- We will focus on problems with **one dependent variable**
- For **more dependent** variables analysis, in example:
  - MANOVA
  - Canonical analysis
  - Discriminant function analysis

Tests

Parametric vs nonparametric tests

- Parametric tests
  - When to use?
    - Dependent variable is an interval/ratio data variable.
  - Advantages?
    - Well known (techniques and interpretation of the results).
    - Often more flexible (than their nonparametric analogues)
    - More powerful than (their nonparametric analogues)
  - Disadvantages?
    - Assume something about the distribution of the underlying data. If these assumptions are violated, the resulting test statistics will not be valid, and the tests will not be as powerful as for cases when assumptions are met.
- Nonparametric tests
  - When to use?
    - Dependent variable is either an ordinal or interval/ratio data variable.
    - They are appropriate when interval/ratio data is: non-normal, skewed, leptokurtic/platokurtic, censored
  - Advantages?
    - Lack of assumption that the underlying data have any specific distribution.
  - Disadvantages?
    - Interpretation of the results: often results are incorrectly interpreted as a difference in medians when they are really describing a difference in distributions.
    - Lack of flexibility in designs.
    - Lack of power relative to their parametric equivalents.
- Note: For nominal data specific tests not related to aforementioned

One-sample tests

One-sample t Test - parametric

DATA TYPE:
- One-sample data
- Data are interval/ratio & continuous
- Data are normally distributed (CENTRAL LIMIT THEOREM)
- Moderate skewness is permissible if the data distribution is unimodal without outliers

HYPOTHESIS:
H0: The mean is equal to the MU
H1 (2 sided): The mean is not equal to the MU

INTERPRETATION:
H0: Fail to reject that mean is not significantly different from MU
H1 (2 sided): Mean is significantly different from MU

$$t = \frac{\overline{X} - MU}{S/\sqrt{n}} \sim t_{n-1}$$

Normality assumption:
- n>30 (rule of thumb) CLT holds and t test may be performed
- fail to reject H0 about normality in normality tests (Jarque-Bera; Shapiro-Wilk etc.)

$$JB = \frac{n-k+1}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right) \quad S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}, C = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2},$$

One-sample Wilcoxon Signed Rank Test

DATA TYPE:
- One-sample data
- Ordinal, interval, or ratio
- Relatively symmetrical about their median

HYPOTHESIS:
H0: The distribution of the data set is symmetric around ME
H1 (2 sided): The distribution of the data set is not symmetric around ME

INTERPRETATION:
H0: Fail to reject that distribution of the data set is symmetric around ME (ME is median)
H1 (2 sided): Values of variable are not symmetrically distributed around ME (ME is not median)

In R function *wilcox.test()* it is implemented differently:
1. $V = \sum_{i=1}^{N_r} [R_i] \, \mathbf{1}(X_i - ME > \mathbf{0})$
2. If ties then normal approximation

In R function *wilcox.exact()*:
1. $V = \sum_{i=1}^{N_r} [R_i] \, \mathbf{1}(X_i - ME > \mathbf{0})$
2. Exact distibution even if ties

- **Median instead of mean (however, if distribution is symmetric ME=MU)**
- **Procedure:**
  - For each observation (from N) calculate absolute difference from assumed ME: $|X_i - ME|$
  - Drop observations with absolute difference equal to 0 (N_r observations left)
  - Order the rest in ascending order and assign ranks R_i. For tied ranks assign an average rank.
  - Calculate test statistics: $W = \sum_{i=1}^{N_r} [sgn(X_i - ME) R_i]$
  - Take critical values from reference table (specific distribution with $E(W) = 0$ & $VAR(W) = \frac{N_r(N_r+1)(2N_r+1)}{6}$ or use normal approximation (N_r>10)

One-sample Sign Test - nonparametric

DATA TYPE:
- One-sample data
- Data are ordinal, interval, or ratio

HYPOTHESIS:
H0: The median of the data set is equal to ME
H1 (2 sided): The median of the data set is not equal to the ME

INTERPRETATION:
H0: Fail to reject that median is equal the ME (ME is median)
H1 (2 sided): Media is not equal to ME (ME is not median)

- Median instead of mean
- **Data does not have to be symmetric in distribution**
- Has **smaller power** than Wilcoxon one sample test
- **Procedure:**
  - For each observation (from N) calculate difference from assumed ME:$X_i - ME$
  - Drop observations with absolute difference equal to 0 (N_r observations left)
  - Calculate sum of positive difference (S) and sum of negative difference (F).
  - P-Value: $P(s \leq S); S \sim binomial(n = N_r, p = 0.5)$