Hypothesis testing

When we are concerned with a real situation in which observations may be made and described by a probabilistic model, *a scientific hypothesis is a statement about the probabilistic structure describing the inherent variability in the observational situation*.

For instance, suppose that a large population is classified according to 2 factors A y B. There are r A categories A_1, A_2,….. A_r y s B categories B_1, B_2,….. B_s. Each individual of the population belongs to one and only one of th *rs* cells A_iB_j , and the proportion θ_{ij} of the population in cell A_iB_j is unknown. An individual chosen at random has probability θ_{ij} of falling in the cell A_iB_j.If we observe the numbers in a random sample of n individuals belonging to the different cells, then a typical observation x takes the form of x = (n_{11}, n_{12}, ………n_{rs}) being n_{ij} the number of individuals in the cell A_iB_j. The appropriate family of possible distributions on the sample space is the multinomial family, parametrized by θ=(θ_{11}, θ_{12},……θ_{rs}).

The parameter space $\Theta = \{\theta_{ij}: 0 \leq \theta_{ij} \leq 1; \Sigma_{ij}\ \theta_{ij} = 1\}$

Hypothesis testing

Let our hypothesis be : “factors A and B are nor related”

Going back to the probabilistic multinomial model it means $\forall$ i, j $\theta_{ij} = \theta_{i.}\ \theta_{.j}$, being $\theta_{i.} = \Sigma_j\ \theta_{ij}$ and $\theta_{.j} = \Sigma_i\ \theta_{ij}$

So, our hypothesis implies a restriction on the set of possible distribution explained the observed variability.

Now $\Theta = \{\theta: 0 \leq \theta \leq 1; \Sigma_{ij}\ \theta_{ij} = 1 \text{ and } \theta_{ij} = \theta_{i.}\ \theta_{.j}\ \}$

So, we can generally represent an hypothesis through a proper subset of the parameter space, Θ.

We can say “Hypothesis ω” being $\omega \subset \Theta$

Hypothesis testing

The theory and practice of hypothesis testing is related to the question: ***"Is a given observation consistent with some stated hypothesis or not?"***

We will split the set of all possible observations, X, the sample space, in two regions:

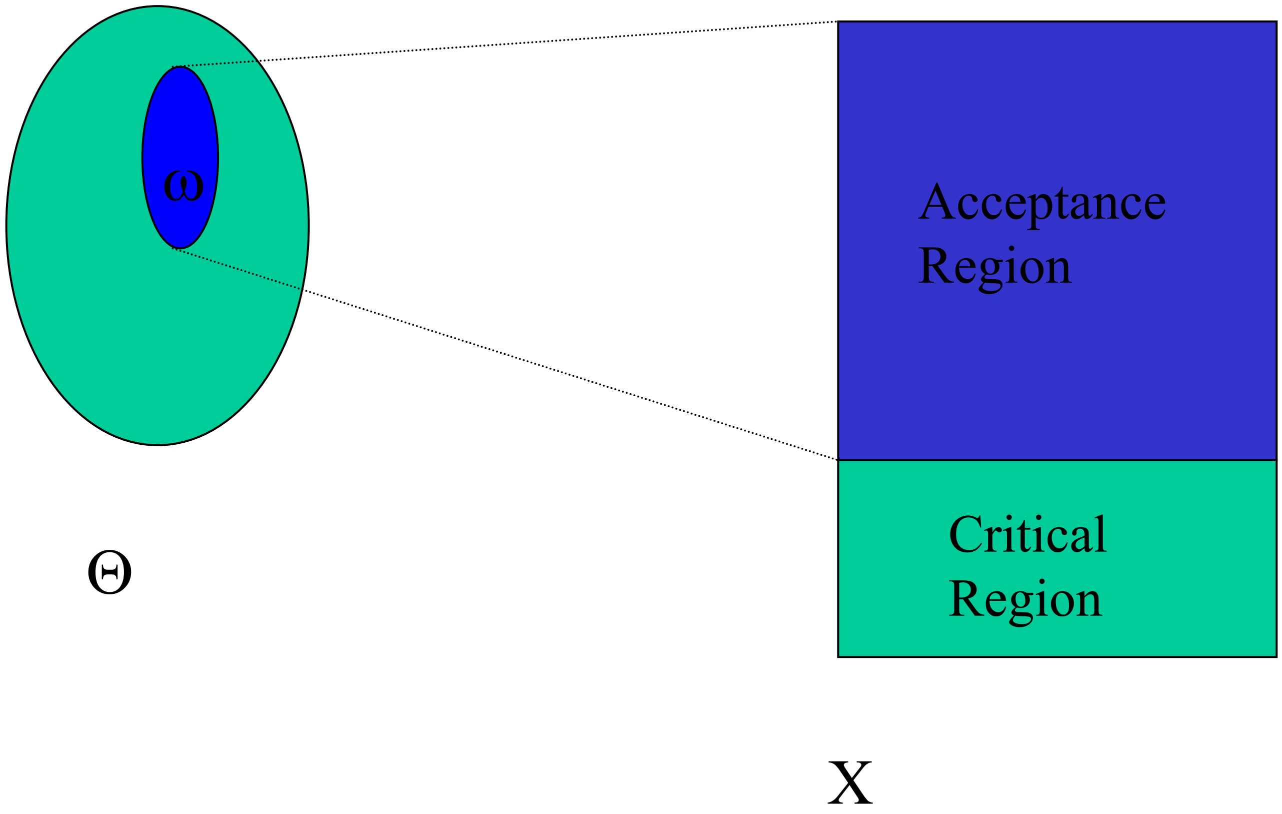
Those observations consistent with the hypothesis ω, *called the region of acceptance*

Those observations not consistent with the hypothesis ω, *called the region of rejection or Critical Region*

A statistical test of a hypothesis is a rule which assigns each possible observation to one of these exclusive regions.

For a given hypothesis, there are as many testes as there are subsets of X. The problem is to choose a test which is good in some sense.

Hypothesis testing: Example



Hypothesis testing: Example

Hypothesis statement: The proportion of smokers in a given population is less than 50%. The observation consist in n randomly chosen persons. X, the sample space is {0,1,2,…n}. The family of distributions is the family of binomial distributions with parameter θ; $0 \leq \theta \leq 1$.

The hypothesis can be written as $\omega = [0, 0.5)$

The class of test consistent with the hypothesis are of the form $\{x: x \leq k\}$ being x the number of smokers in n, and k some value between 0 and n. We could also refine our Critical Region to be as: $\{x: x \leq \frac{1}{2} n\}$ or, ‘less than half the sample smokes’.

Let n = 50 -> 24 or less smokers in 50 is C.

Hypothesis testing: Example

Let n = 50 -> 24 or less smokers in 50 is C.

It could be θ < 50% and more than 24 smokers

It could be θ > 50% and less than 24 smokers

Definitions:

Null Hypothesis

Alternative Hypothesis

$\alpha(\theta)$ = P(TI E)

$\beta(\theta)$ = P(TII E)

If $\alpha(\theta) \leq \alpha$

α *Significance Level of the Test*

		Hypothesis is	
		True	**False**
Action	**Reject:**	Error (TI E)	
	No Reject:		Error (TII E)

Hypothesis testing: Recap & Summary

Choosing an optimal C is a theoretical problem.

Classical approach fixes weights as more important TI E so works primarily with α, provided the null hypothesis has some theoretical support.

Classical Significance Hypothesis testing works with a measure of “discrepancy”, “D”, between Hypothesis and Evidence (given by sample) which Probability Distribution is known in advance, and set the critical Region by imposing the condition:

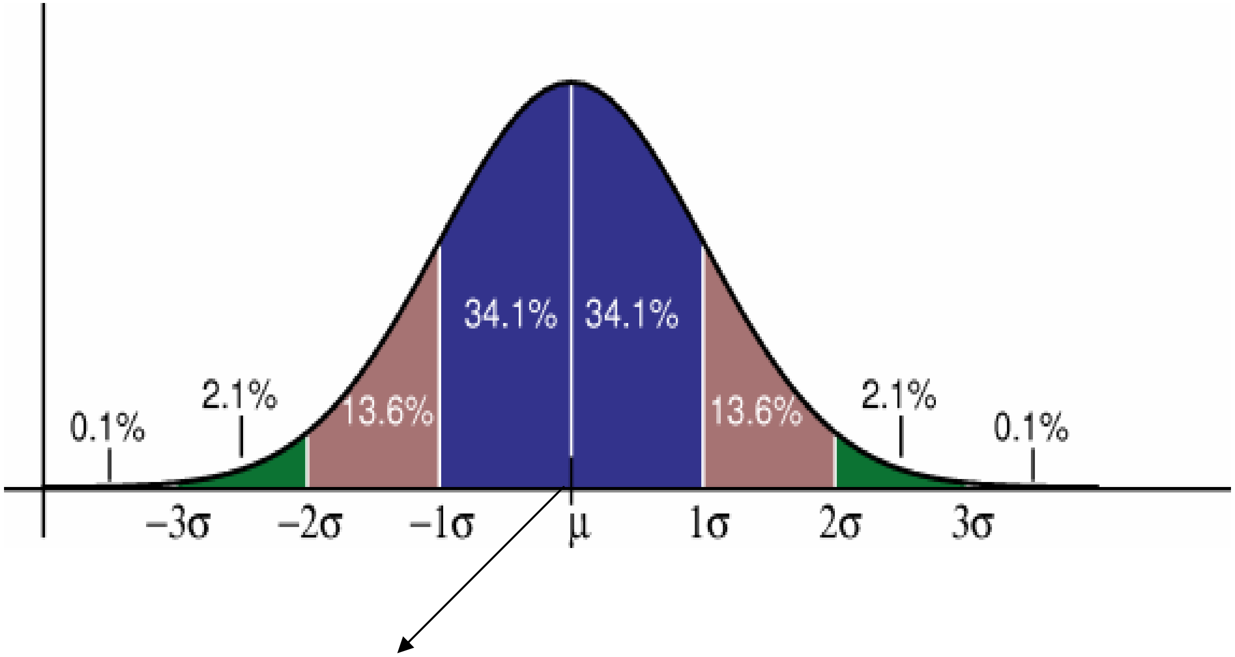
$$P[D>d_\alpha /H0 \text{ true}] = \alpha$$

Where D can take different forms and the level of significance has to be set in advance.

If $D>d_\alpha$ or equivalently $P[D>d_\alpha /H0 \text{ true}] < \alpha$ -> H0 is rejected

Discrepancy based on Normal Distribution

- Testing simple hypothesis on μ with n= 1



Hypothetic μ value

Types of data: *The way we observe affects the way we infer*

- **Nominal**: 2 or more categories, mutually exclusive with no order. Lowest level of measure.
  - Marital status, religion, etc
- **Ordinal**: Categories that can be ordered:
  - Non smoker/ ex-smoker/light smoker/heavy smoker
  - The difference between consecutive categories is not measurable.
- **Scale**: Variables with intrinsic metric: age, income, weight, etc. Can be numerically transformed: aditions, substraction, etc

FREQUENCY TABLES

A frequency table is a table where each cell corresponds to a particular combination of characteristics relating to 2 or more classifications. We will deal only with two way tables, which apply to two categorical variables. Frequency tables are also known as contingency tables.

The method for analysing frequency tables varies according to:

- Number of categories.
- Whether categories are ordered or not.
- Number of independent groups of subjects.
- The nature of the question being asked.

FREQUENCY TABLES

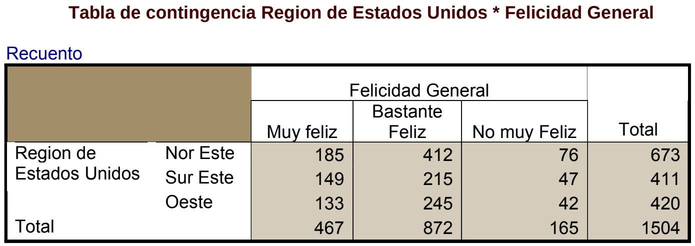


Tabla de contingencia Region de Estados Unidos * Felicidad General

Recuento

		Felicidad General			Total
		Muy feliz	Bastante Feliz	No muy Feliz	
Region de Estados Unidos	Nor Este	185	412	76	673
	Sur Este	149	215	47	411
	Oeste	133	245	42	420
Total		467	872	165	1504

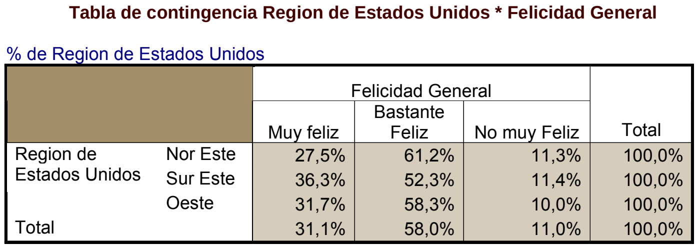


Tabla de contingencia Region de Estados Unidos * Felicidad General

% de Region de Estados Unidos

		Felicidad General			Total
		Muy feliz	Bastante Feliz	No muy Feliz	
Region de Estados Unidos	Nor Este	27,5%	61,2%	11,3%	100,0%
	Sur Este	36,3%	52,3%	11,4%	100,0%
	Oeste	31,7%	58,3%	10,0%	100,0%
Total		31,1%	58,0%	11,0%	100,0%

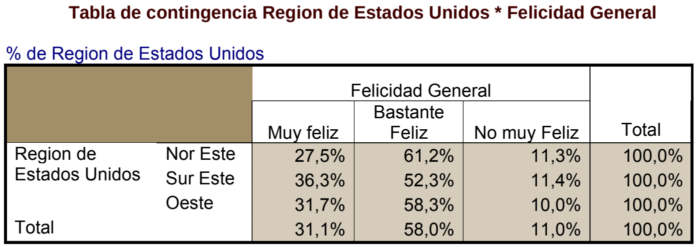


Tabla de contingencia Region de Estados Unidos * Felicidad General

% de Region de Estados Unidos

		Felicidad General			Total
		Muy feliz	Bastante Feliz	No muy Feliz	
Region de Estados Unidos	Nor Este	27,5%	61,2%	11,3%	100,0%
	Sur Este	36,3%	52,3%	11,4%	100,0%
	Oeste	31,7%	58,3%	10,0%	100,0%
Total		31,1%	58,0%	11,0%	100,0%

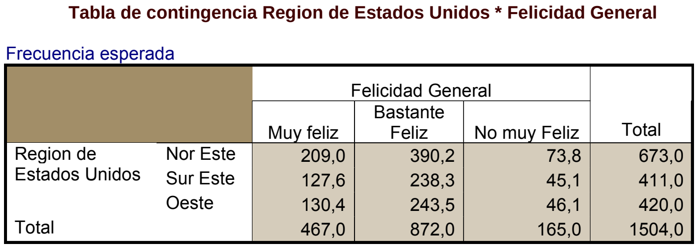


Tabla de contingencia Region de Estados Unidos * Felicidad General

Frecuencia esperada

		Felicidad General			Total
		Muy feliz	Bastante Feliz	No muy Feliz	
Region de Estados Unidos	Nor Este	209,0	390,2	73,8	673,0
	Sur Este	127,6	238,3	45,1	411,0
	Oeste	130,4	243,5	46,1	420,0
Total		467,0	872,0	165,0	1504,0

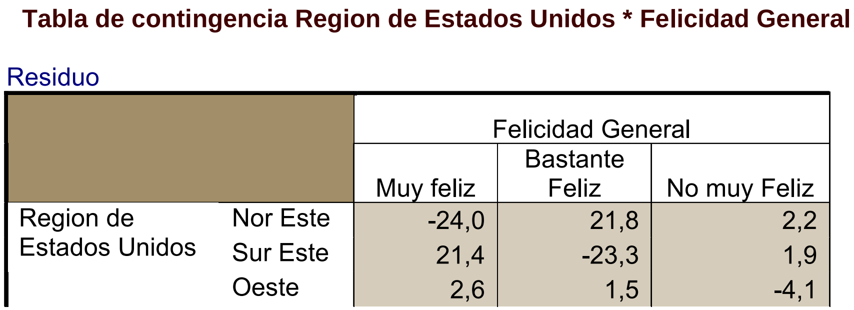


Tabla de contingencia Region de Estados Unidos * Felicidad General

Residuo

		Felicidad General		
		Muy feliz	Bastante Feliz	No muy Feliz
Region de Estados Unidos	Nor Este	-24,0	21,8	2,2
	Sur Este	21,4	-23,3	1,9
	Oeste	2,6	1,5	-4,1

Chi-Square Significance Tests

Chi-square is a family of distributions commonly used for significance testing.

Pearson's chi-square is by far the most common type of chi-square significance test. If simply "chi-square" is mentioned, it is probably Pearson's chi-square. This statistic is used to test the hypothesis of no association of columns and rows in tabular data. It can be used even with nominal data.

Note that chi square is more likely to establish significance to the extent that (1) the relationship is strong, (2) the sample size is large, and/or (3) the number of values of the two associated variables is large. A chi-square probability of .05 or less is commonly interpreted by social scientists as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable.

$$X^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$