Basic Inference

K. Gibert

Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group

Universitat Politècnica de Catalunya, Barcelona

karina.gibert@upc.edu

www.eio.upc.edu/homepages/karina

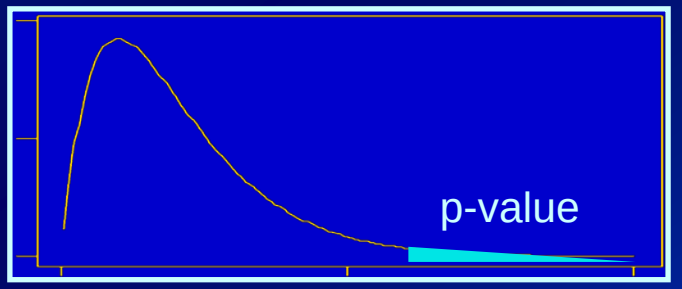
Statistical Significance test

1. Hypothesis

   H0: (null hypothesis)

   H1: (alternative hypothesis)

f(s), assuming H0 is true



2. Statistics (S):

   Measure of proximity between sample and H0

   Known probability distribution under H0 (f)

   CARE: Technical conditions might be required

3. Observed value of statistics (s0):

   Compute statistics over sample data

4. p-value: $P_f(|S|>s0)$

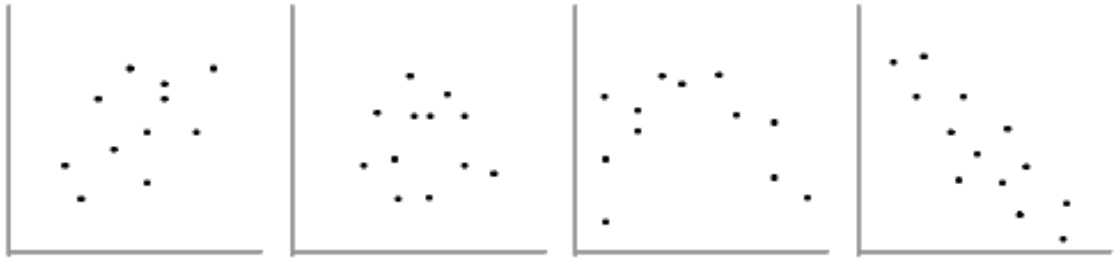
5. Decision rule: if p-value < alpha then Reject H0

p-value depends on H_1

Association between numerical variables

Correlation coefficient

$$r_{x_j y} = \frac{\sum_i (x_{ij} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_{ij} - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{s_{x_j y}}{s_x s_y}$$



Test : H_0: cor(X,Y)=0

H_1: cor(X,Y) ≠ 0

$$\sqrt{(n-2)}\,\frac{R}{\sqrt{(1-R^2)}} \sim t_{(n-2)}$$

Only linear relationships

Sheffer Generalized coefficient

Assessing association between categorical variables

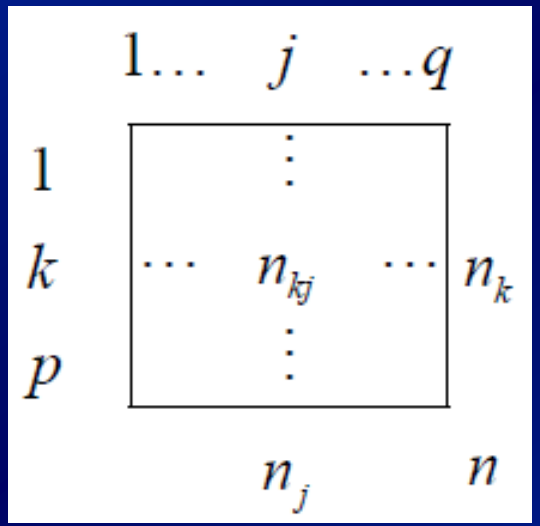
The chi2 independence Test

Test: H_0: X,Y are independent ($n_{kj}=np_kp_j\ \forall kj$)
H_1: X,Y are associated

Missperformance if nkj<5

Statistics:

$$X^2=\sum_{k=1}^{p}\sum_{j=1}^{q}\frac{(n_{kj}-n_kn_j/n)^2}{n_kn_j/n}\sim\chi^2_{(p-1)(q-1)}$$



Care with Simpson's Paradox

Assessing association between categorical variables

The Simpson's Paradox

Apparently independent

or

Apparently dependent

The Simpson's Paradox

Apparently independent

1978: Warren McClesky (black man) sentenced to death in Georgia for killing a police (white man)

He appealed to the US Supreme Court argueing racial bias of death penalty in Georgia

326 defendants in homicide in 20 Florida counties in between 1976-1977

[Radlet 1981] [Agresti 1990]

		YES	NO	Total	(%Y
Race of Suspect	White	39	308	347	(
	Black	32	345	377	(
	Total	71	653	724	(9.8%

Apparently Independent

When color of victim considered…..

White victim

		YES	NO	Total	(%Yes)
Race of Suspect	White	39	279	318	(12.3%)
	Black	29	121	150	(19.3%)
	Total	68	400	468	(14.5%)

Black victim

		YES	NO	Total	(%Yes)
Race of Suspect	White	0	29	29	(0.0%)
	Black	3	224	227	(1.3%)
	Total	3	253	256	(1.2%)

Assessing association between one categorical variable and one numerical

The F Test

Requires Normality

Test: H_0: $\mu_{Y|X=x_1} = \mu_{Y|X=x_2} = \ldots = \mu_{Y|X=x_s} = \mu$ (X, Y independent)

H_1: $\exists\, x \in \{x_1 \ldots x_s\}$: $\mu_{Y|X=x} \neq \mu$ (X,Y associated)

Statistics:

$$F = \frac{S_B^2/(q-1)}{S_W^2/(n-q)} \sim F_{q-1,n-q}$$

$$S_W^2 = \sum_{k=1}^{q} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2$$

$$S_B^2 = \sum_{k=1}^{q} n_k (\bar{x}_k - \bar{x})^2$$

Kruskall-Wallis test

Basic Inference

Karina Gibert

Dpt. Statistics and Operation Research

Knowledge Engineering and Machine Learning Research group

Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

karina.gibert@upc.edu

www.eio.upc.edu/homepages/karina

Are there any questions?...