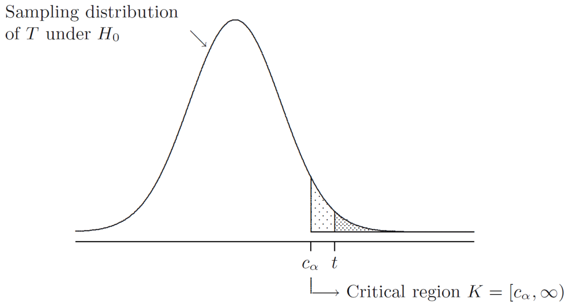
Statistical Methods for Data Science

Lesson 22 - Multiple comparisons. Fitting distributions.

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

Critical values and p-values



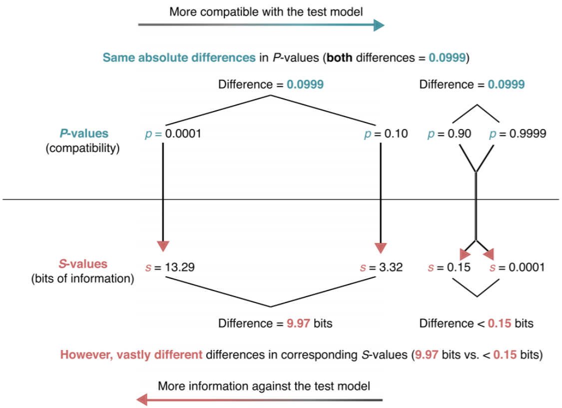
- *Critical region* K: the set of values that reject H_0 in favor of H_1 at significance level α
- *Critical values*: values on the boundary of the critical region
- *p-value*: the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that H_0 is true
- $t \in K$ iff p-value $\leq \alpha$

Misues of p-values

Misinterpretations of p-values, Greenland et al., 2016

- ~~The p-value is the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.~~ A p-value indicates the degree of compatibility between a dataset and a particular hypothetical explanation
- ~~The 0.05 significance level is the one to be used:~~ No, it is merely a convention. There is no reason to consider results on opposite sides of any threshold as qualitatively different.
- ~~A large p-value is evidence in favor of the test hypothesis:~~ A p-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller p-values
- ~~If you reject the test hypothesis because $p \leq 0.05$, the chance you are in error is 5%:~~ No, the chance is either 100% or 0%. The 5% refers only to how often you would reject it, and therefore be in error.

s-values



- Shannon information value or surprisal value (s-value) is $-\log_2 p$
- $p = 0.05 \Rightarrow s = 4.3$ - no more surprising than getting all heads on 4 fair coin tosses.
- $p = 0.005 \Rightarrow p = 7.64$ - no more surprising than getting all heads on 8 fair coin tosses.

The multiple comparisons problem

- Single test $H_0 : \theta = v$, with significance level $\alpha = 0.05$ *[false positive rate]*
  - test is called *significant* when we reject H_0
  - α is Type I error, probability of rejecting H_0 when it is true
- Multiple tests, say $m = 20$
  - E.g., $H_0^i : \theta_i = 0$ for $i = 1, \ldots, m$ where θ_i is the **expectation of a subpopulation**
- What is the probability of rejecting at least one H_0^i when all of them are true?

$$P(\text{at least one reject}) = P(\cup_{i=1}^{m}\{p_i \leq \alpha\}) = 1 - P(\cap_{i=1}^{m}\{p_i > \alpha\}) = 1 - (1-\alpha)^m$$

and then $1 - (0.95)^{20} \approx 0.64$

Family-wise error rate (FWER)

The FWER is the probability of making at least one Type I error in a family of n tests. If the tests are independent:

$$\alpha_{FWER} = 1 - (1-\alpha)^m$$

If the test are dependent: $\alpha_{FWER} \leq m \cdot \alpha$

Multiple comparisons: corrections

- *Bonferroni correction* (most conservative one):

$$\alpha = \frac{\alpha_{FWER}}{m}$$

  Hence, $p < \alpha$ iff $p \cdot m < \alpha_{FWER}$

- *Šidák correction* (exact for independent tests):

$$\alpha = 1 - (1 - \alpha_{FWER})^{1/m}$$

  Hence, $p < \alpha$ iff $1 - (1 - p)^m < \alpha_{FWER}$

See R script

False Discovery Rate and q-values

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	*False Positive*	*True Positive*
	Not reject H_0	*True Negative*	*False Negative*

- False Positive Rate: $FPR = FP/(FP + TN)$
  - Corrections control for FPR since $FWER = P(FP > 0|H_0^i\ i = 1, \ldots, m)$
- Drawback: acting on α increases $FNR = FN/(FN + TP)$
- False Discovery Rate: $FDR = FP/(FP + TP)$
  - $FDR = 0.05$ means 5% of rejected H_0's are actually true
- q-value is $P(H_0|T \geq t)$ $[p = P(T \geq t|H_0)]$
  - FDR can be controlled by requiring $q \leq$ threshold

See R script

Distribution fitting

- Dataset $x_1, \ldots, x_n$ realization of $X_1, \ldots, X_n \sim F$
- What is a plausible F?
- Parametric approaches:
  - Assume $F = F(\lambda)$ for some family F, and estimate λ as $\hat{\lambda}$
    - Maximum Likelihood Estimation (point estimate):

$$\hat{\lambda} = argmax_{\lambda} L(\lambda)$$

    - Parametric bootstrap (p-value):

$$T_{ks} = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\Lambda}^*}(a)|$$

- Non-parametric approaches:
  - Empirical distribution
  - Kernel Density Estimation
- Goodness of fit: how good is F in fitting the data?

Goodness of fit

- Loss functions (to be minimized)
  - Akaike information criterion (AIC), balances model fit against model simplicity
$$AIC(F(\lambda)) = 2|\lambda| - 2\ell(\lambda)$$
  - Bayesian information criterion (BIC), stronger balances over model simplicity
$$BIC(F(\lambda)) = |\lambda| \log n - 2\ell(\lambda)$$
- Statistics (continuous data):
  - **KS test** $H_0 : X \sim F \quad H_1 : X \nsim F$ with Kolmogorov-Smirnov (KS) statistic:
$$D = \sup_{a \in \mathbb{R}} |F_n(a) - F_\lambda(a)| \sim K$$
  - **LR test** $H_0 : X \sim F_1 \quad H_1 : X \sim F_2$ with the likelihood-ratio test:
$$\lambda_{LR} = \log \frac{L(F_1(\lambda_1))}{L(F_2(\lambda_2))} = \ell(F_1(\lambda_1)) - \ell(F_2(\lambda_2)) \quad \text{with} \quad -2\lambda_{LR} \sim \chi^2$$

See R script

Goodness of fit

Chi-square distribution

The Chi-square distribution with k degrees of freedom $\chi^2(k)$ has density:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Let $X_1, \ldots, X_k \sim N(0,1)$. Then $Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$

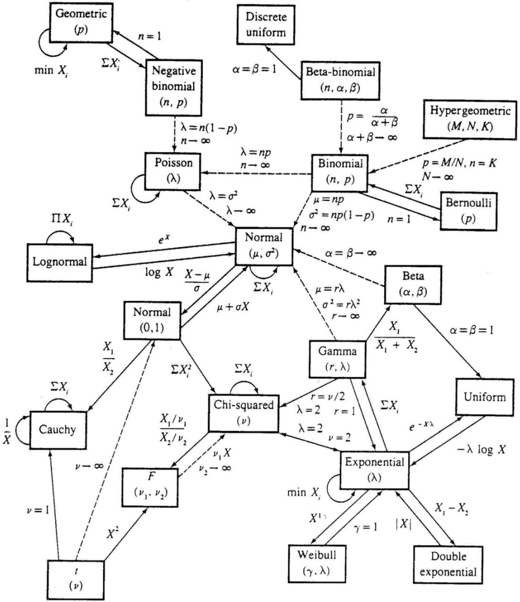
- Statistics (discrete data):
  - **Pearson's Chi-Square test** $H_0 : X \sim F(\gamma) \quad H_1 : X \not\sim F(\gamma)$ with χ^2 statistic:

$$\chi^2 = \sum_{N_i > 0 \vee n_i > 0} \frac{(N_i - n_i)^2}{n_i} = n \cdot \sum_{N_i > 0 \vee p(i) > 0} \frac{(N_i/n - p(i))^2}{p(i)} \sim \chi^2(df)$$

where N_i number of observations of value i, $n_i = n \cdot p(i)$ expected number of observations, and $df = |\{i \mid N_i > 0\}| - |\gamma|$ is the number of observed values minus the number of estimated parameters. $\chi^2 = \infty$ if for some i: $n_i = 0$ and $N_i > 0$

See R script

Common distributions



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Comparing two datasets

- Dataset $x_1, \dots, x_n$ realization of $X_1, \dots, X_n \sim F_1$
- Dataset $y_1, \dots, y_m$ realization of $Y_1, \dots, Y_n \sim F_2$
- $H_0 : F_1 = F_2 \qquad H_1 : F_1 \neq F_2$
- Continuous data: KS statistics

$$D = \sup_{a \in \mathbb{R}} |F_1(a) - F_2(a)| \sim K$$

- Discrete data: χ^2 statistics

$$\chi^2 = \sum_{R_i > 0 \vee S_i > 0} \frac{(\sqrt{\frac{m}{n}} R_i - \sqrt{\frac{n}{m}} S_i)^2}{R_i + S_i} \sim \chi^2(df)$$

where R_i (resp., S_i) is the number of observations in $x_1, \dots, x_n$ (resp., $y_1, \dots, y_m$) of value i, $df = |\{i \,|R_i > 0 \vee S_i > 0\}| - 1$

See R script