Master Program in *Data Science and Business Informatics*

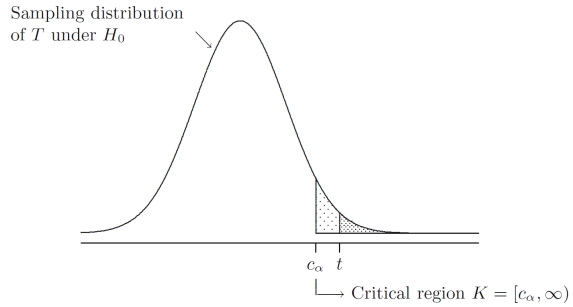
Statistics for Data Science

Lesson 31 - Multiple comparisons. Fitting distributions

Salvatore Ruggieri

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

Critical values and p-values



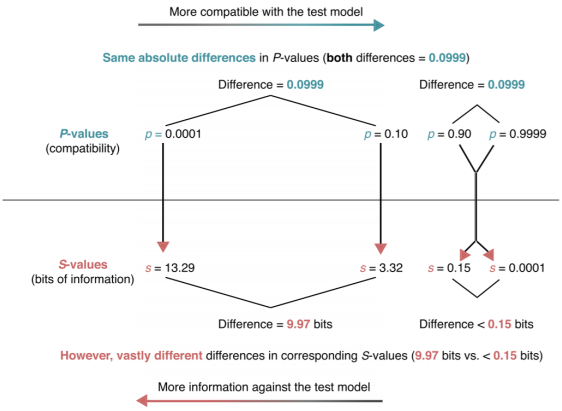
- *Critical region* K: the set of values that reject H_0 in favor of H_1 at significance level α
- *Critical values*: values on the boundary of the critical region
- *p-value*: the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that H_0 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 $-\log_2 p$ (unit: bit)
  - $p = 0.5 \Rightarrow s = 1$ surprising as getting one heads on 1 fair coin toss
  - $p = 0.10 \Rightarrow s = 3.32$ surprising as getting all heads on 3 fair coin tosses
  - $p = 0.0001 \Rightarrow s = 13.29$ surprising as getting all heads on 13 fair coin tosses

The multiple comparisons problem

- Single test $H_0 : \theta = 0$, 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?
  - For independent tests: $P(\cup_{i=1}^m \{p_i \leq \alpha\}) = 1 - P(\cap_{i=1}^m \{p_i > \alpha\}) = 1 - (1-\alpha)^m$ and then $1 - (0.95)^{20} \approx 0.64$
  - For dependent tests: $P(\cup_{i=1}^m \{p_i \leq \alpha\}) \leq \sum_i P(\{p_i \leq \alpha\}) = m \cdot \alpha$, and then $\leq 20 \cdot 0.05 = 1$

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

Objective: achieve significant tests ($p \leq \alpha'$) such that $\alpha_{FWER} \leq \alpha$

- *Bonferroni correction* (most conservative one):
  - scale significance level $\alpha' = \alpha/m$ *[invert $\alpha = m \cdot \alpha'$]*
  - Notice: $p \leq \alpha'$ is equivalent to scale p-values and test $p \cdot m \leq \alpha$

  Thus $\alpha_{FWER} \leq m \cdot \alpha' = \alpha$
- *Šidák correction* (exact for independent tests):
  - scale significance level $\alpha' = 1 - (1 - \alpha)^{1/m}$ *[invert $\alpha = 1 - (1 - \alpha')^m$]*
  - Notice: $p \leq \alpha'$ is equivalent to scale p-values and test $1 - (1 - p)^m \leq \alpha$

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

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: $FDR = FP/(FP + TP)$ *[Korthauer et al, 2019]*
  - $FDR = 0.05$ means 5% of rejected H_0's are actually true
- q-**value** is $P(H_0 | T \geq t)$ *[vs. $p = P(T \geq t | H_0)$]*
  - FDR can be controlled by requiring $q \leq$ threshold

See R script

Distribution fitting and quality of fitting

- Dataset $x_1, \ldots, x_n$ realization of $X_1, \ldots, X_n \sim F$
- **Distribution fitting**: What is a plausible F?
  - Useful in Data Science for understanding the data generation process, for checking assumptions (e.g., normality of noise in LR), for checking data distribution changes, etc.
  - 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
- **Quality of fitting**: Among several fits $F_1, \ldots, F_k$, which one is the best?
  - Goodness of fit: measure of how good/bad is F_i in fitting the data?
  - Comparison: which one between two F_1 and F_2 is better?

Quality of fitting

- 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** $H_0 : X \sim F \quad H_1 : X \not\sim F$ with Kolmogorov-Smirnov (KS) statistic:

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

  - **LR test** $H_0 : X \sim F_1 \quad H_1 : X \sim F_2$ with the likelihood-ratio test:

$$\lambda_{LR} = \log \frac{L(F_1(\lambda_1))}{L(F_2(\lambda_2))} = \ell(F_1(\lambda_1)) - \ell(F_2(\lambda_2)) \quad \text{with} \quad -2\lambda_{LR} \sim \chi^2(1)$$

See R script

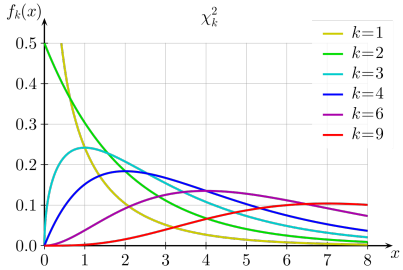
Chi-square distribution

Chi-square distribution

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

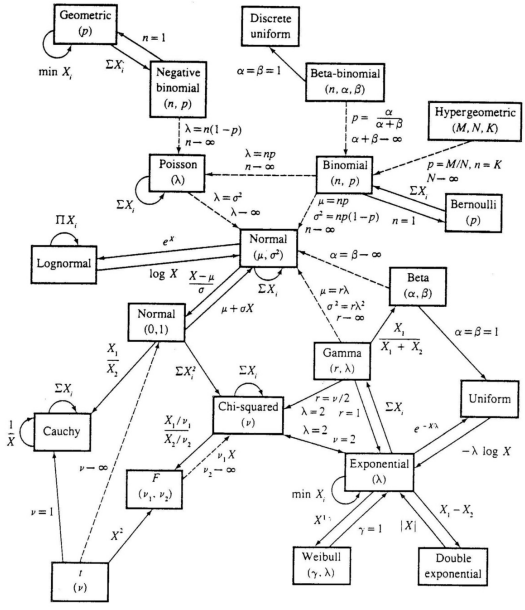
$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

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



Common distributions

- **Probability distributions at Wikipedia**
- **Probability distributions in R**
- C. Forbes, M. Evans,
  N. Hastings, B. Peacock (2010)
  Statistical Distributions, 4th Edition
  Wiley



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

Quality of fitting

- Statistics (discrete data):
  - **Pearson's Chi-Square test**

    $H_0 : X \sim F \quad H_1 : X \not\sim F$ with χ^2 statistic:

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

    where N_i number of observations of value i, $n_i = n \cdot p(i)$ expected number of observations (rescaled), and $df = |\{i \mid N_i > 0\}| - 1$ is the number of observed values minus 1.
    $\chi^2 = \infty$ if for some i: $n_i = 0$
  - **Yates's correction for continuity**

    It corrects for approximating the discrete probability of observed frequencies by the continuous chi-squared distribution

    $$\chi^2 = \sum_{N_i>0} \frac{(|N_i - n_i| - 0.5)^2}{n_i}$$

    It increases Type II error, so do not use it!

See R script

Comparing two datasets

- Dataset $x_1, \ldots, x_n$ realization of $X_1, \ldots, X_n \sim F_1$
- Dataset $y_1, \ldots, y_m$ realization of $Y_1, \ldots, Y_n \sim F_2$
- $H_0 : F_1 = F_2 \quad H_1 : F_1 \neq F_2$
- 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 variables in $X_1, \ldots, X_n$ (resp., $Y_1, \ldots, Y_m$) which are equal to i, $df = |\{i \,|R_i > 0 \vee S_i > 0\}| - 1$

- Useful to detect **covariate drift** (data stability) from source to target datasets (training set vs deployment set) *[See also Lessons 16 and 35 for association measures]*

See R script

Optional references

Keegan Korthauer, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm, and Stephanie C. Hicks (2019)

A practical guide to methods controlling false discoveries in computational biology.

Genome Biology 20, article 118

Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman (2016)

Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.

European Journal of Epidemiology 31, pages 337–350