Classnote 1, Statistical tests and multiple hypothesis tests.

Mateusz Chojnacki

Definitions:

Statistic – value computed with any method from one or more random variables, used for statistical purpose, e.g. mean, variation, estimator of mean, test statistics etc.

Null hypothesis (H_0) – statement that some relation between data is true, which cannot be conformed, but only rejected in favour of chosen **alternative hypothesis (H_1)** using **statistical significance test**, e.g. samples are from distribution with mean $\mu = \mu_0$ (H_0) versus $\mu \neq \mu_0$ (H_1) or versus $\mu < \mu_0$ (H_2). Set of tested parameters is marked θ_i, for hypothesis from previous example: $\theta_0 = \{\mu_0\}$, $\theta_1 = \mathbb{R}\backslash\{\mu_0\}$, $\theta_2 = (-\infty, \mu_0)$.

Simple hypothesis – hypothesis which specify distribution completely, set θ_i contain only one element; **complex hypothesis** – set θ_i contains more than one element (e.g. previously mentioned θ_1). **Parametric hypothesis** - hypothesis about one or more parameters of chosen distribution, **nonparametric hypothesis** – says nothing about parameters of distribution.

Statistical (hypothesis) test – set of rules, which decide upon chosen data set in favour of or against rejection of particular null hypothesis. It includes: formulating a null hypothesis, formulating correct alternative hypothesis, choosing correct test statistic T, selecting **significance level (α)** and **critical region (W)**, computing value t of test statistic T from dataset, deciding whether reject null hypothesis or not (usually null hypothesis is rejected, when t is in the critical region).

Significance level of the test (α, probability of type I error, size) - probability of rejecting true H_0, $\boldsymbol{\alpha = P(t \in W|H_0)}$**.** Type I error is also called false positive.

β, probability of type II error - probability of not rejecting false H_0, $\boldsymbol{\beta = P(t \in (\mathbb{R}\backslash W)|H_1)}$**.** Type II error is also called false negative. **Power of the test: $1 - \beta$.**

Critical region (region of rejection, W) – set of values for which null hypothesis is rejected; it is computing based on chosen significance level α, type of alternative hypothesis (one-tailed or two-tailed) and test statistic.

p-value – "the probability under the assumption of null hypothesis, of obtaining a result equal to or more extreme than what was actually observed" [3]. Because it is probability, it can only have values from 0 to 1. The smaller p-value is, the more certainly null hypothesis should be rejected. With given significance level $\boldsymbol{\alpha}$ null hypothesis is rejected, when its p-value is smaller than $\boldsymbol{\alpha}$.

Possible results of single hypothesis testing shows table below.

	Not rejected H_0, positive, not significant	**Rejected H_0, negative, significant**
True H_0	TP – True positive, $\boldsymbol{1 - \alpha}$	FN – False negative, type I error, $\boldsymbol{\alpha}$
False H_0	FP – False positive, type II error, $\boldsymbol{\beta}$	TN – True negative, $\boldsymbol{1 - \beta}$

There are many test statistics, same for continuous data and others for discrete data. List below contains the most commonly used test statistics with short descriptions.

n, n_1, n_2 – size of the tested population; s, s_1, s_2 – specific estimator of sample standard deviation; df – degrees of freedom; t: uses t-statistic, testing H_0 under Student's t-distribution; z: z-statistic, testing H_0 under normal distribution; $d_0 = \mu_1 - \mu_2$; F – uses F-distribution; χ^2 – uses chi-square distribution;

Significance tests of mean value μ, all under assumption of normal population(s) distribution or population(s) is(are) large enough:

- **one-sample z-test** (H$_0$: μ = μ$_0$, known σ, $z = (\bar{x} - \mu_0)/\sigma * \sqrt{n}$),
- **one-sample t-test** (H$_0$: μ = μ$_0$, unknown σ, $t = \frac{\bar{x}-\mu_0}{s} * \sqrt{n}, df = n - 1$),
- **two-sample z-test** (H$_0$: $\mu_1 - \mu_2 = d_0$, independent observations, $\sigma_1\ and\ \sigma_2$ are known, $z = (\overline{x_1} - \overline{x_2} - d_0)/\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$),
- **paired t-test** (H$_0$: $\mu_1 - \mu_2 = d_0$, pairs of independent observations, each observation in pair is taken from the same object, unknown σ, d$_0$: $\mu_1 - \mu_2$, $t = \frac{\bar{d}-d_0}{s} * \sqrt{n}, df = n - 1$),
- **two-sample pooled t-test** (H$_0$: $\mu_1 - \mu_2 = d_0$, independent observations, $\sigma_1 = \sigma_2$ unknown, $t = (\overline{x_1} - \overline{x_2} - d_0)/(s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$)), df= $n_1 + n_2 - 2$, $s = ((n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2)/(n_1 + n_2 - 2))$),
- **two-sample unpooled t-test** (H$_0$: $\mu_1 - \mu_2 = d_0$, unequal variances, modified **Welch's t-test,** independent observations, $\sigma_1 \neq \sigma_2$ both unknown, $t = (\overline{x_1} - \overline{x_2} - d_0)/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, $df = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 / \left(\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)\right)$). [4]

Significance tests of variance value σ all under assumption of normal population and that population is big enough:

- **chi-squared test for variance** (H$_0$: $\sigma^2 = \sigma_0^2, \chi^2 = (n - 1) * s^2/\sigma_0^2), df = n - 1$)
- **two-sample F test for equality of variances** (H$_0$: $\sigma_1^2 = \sigma_2^2,\ F = s_1^2/s_2^2,\ df_1 = n_1 - 1,\ df_2 = n_2 - 1$, is under assumption that all samples from set have the same variance)

One-way analysis of variance (one-way **ANOVA**, used to check if there is any statistically significant difference means between groups of quantitative variables, under H$_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$, k-number of groups, assumptions: independent observations, normal or close to normal distribution, variations between groups are homogeneous, groups sizes are equal, F=(between-group variability)/(within-group variability), $df_1 = k - 1$, $df_2 = n - k$. This test cannot specify which means differ between groups.) [5]

Pearson's chi-squared test for fit of distribution (for categorical or semi categorical data, H$_0$: observations from chosen distribution E$_0$, independent observations, dataset big enough, random samples from population, $\chi^2 = \sum_{i=1}^{k}(n_i - n * p_i)^2/(n * p_i)$, $df = k - number\ of\ distribution\ parameeers - 1$)

Pearson's chi-squared test for independence (for categorical or semicategorical data of pairs of variables, uses s x r contingency table, H$_0$: variables are independent – $p_{ij} = p_{i.} * p_{.j}$, where p_{ij} – observed probability of cell i,j, $p_{i.}, p_{.j}$ – observes probabilities of column i/row j, under assumptions: independent observations, dataset big enough, random samples from population, $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(n_{ij} - n_{i.} * \frac{n_{.j}}{n}\right)^2/(n_{i.} * n_{.j}/n)$, $df = (r - 1) * (s - 1)$)

Kolmogorov–Smirnov test (for continuous quantitative variables, nonparametric test of equality of one-dimensional probability distributions: one-sample K–S test checks equality of experimental probability distribution to selected probability distribution, two-sample K–S test check if two dataset came from the same probability distribution; used to check if two one-dimensional distributions are statistically different; uses Kolmogorov–Smirnov statistic with cumulative distribution functions F(x) and $F_n(x)$, $Dn = \sup\ |F_n(x) - F(x)|$) [6]

Fisher's exact test (assumptions the same as in chi-squared tests, uses 2x2 contingency table as below to check if true category from two possible was given to set of samples, valid for small samples, $p = (a+b)! * (c+d)! * (a+c)! * (b+d)!\ /(a! * b! * c! * d! * n!)$, $n = a + b + c + d$)

	Is True	Is False
Given True	a	b
Given False	c	d

Mann–Whitney U test (nonparametric test comparing two independent populations of samples X,Y, which are ordinal meaning that for each pair of observations can tell, which is bigger, H$_0$: $P(X < Y) = P(X > Y)$, uses U-statistic)

Wilcoxon signed-rank test (nonparametric test uses to compare two population using pairs of matched samples or to check if difference between median is not zero. Observations from two populations are paired based of their ranks.)

Multiple hypothesis testing:

When we test **m** multiple and independent hypothesis with significance level $\boldsymbol{\alpha}$ each, probability of making at least one type I error grows with m and probability of not making any decrease as in following expression: $(1-\alpha)^m < (1-\alpha)^{m-1} < \cdots < (1-\alpha)$, because m>1, $0 < \alpha < 1$. Therefore many different methods to control type I error were invented, but before we shortly discuss them, let's specify more common used types of error, those mentioned in [1] and terminology associated with multiple hypothesis testing.

	positive, not significant	**negative, significant**	**Total**
True H$_0$	TP or U	FN or V	m0
False H$_0$	FP or T	TN or S	m1
Total	P or W	N or R	m

FWER (Family-wise error rate) –"probability of at least one type I error, $FEWR = P(V \geq 1)$" [1]

FDR (False Discovery Rate) $FDR = \frac{TN}{TN+FN} = \frac{V}{R}$, for $R > 0$, $FDR = 0$ for $R = 0$;

pFDR (positive FDR) = $E(V/R|R\ >\ 0)$;

cFDR (conditional FDR) $=\ E(V/R\ |R\ =\ r)$ for E(x) – expected value of x, r – observed R (rejections);

mFDR (marginal FDR) = $\frac{E(FN)}{E(N)} = \frac{E(V)}{E(R)}$;

eFDR (empirical FDR) = $E(V)/\,r$;

Accuracy (ACC) $= \frac{TP+TN}{m}$ for $m > 0$;

FPR (False Positive Rate) $= \frac{FP}{FP+TN} = \frac{FP}{m1} = 1 - TNR$ for $m1 > 0$;

FNR (False Negative Rate) $= \frac{FN}{FN+TP} = \frac{FN}{m0} = 1 - TPR$ for $m0 > 0$;

Sensitivity (True Positive Rate, TPR) $= \frac{TP}{TP+FP} = \frac{TP}{m0} = 1 - FNR$ for $m0 > 0$;

Specificity (True Negative Rate, TNR) = $\frac{TN}{TN+FP} = \frac{TN}{m1} = 1 - FPR$ for $m1 > 0$;

q-value – statistic measure similar to p-value. If p-value can be defined as "minimum possible false positive rate (**FPR**) when calling that feature significant" [2], then q-value can be defined as "the minimum **FDR** that can be attained when calling that feature significant"[2] when R>0, so more technically it should be pFDR. Because FPR can be defined as "Pr(feature i is significant | feature i is truly null)" and pFDR as "Pr(feature i is truly null | feature i is significant) for any i = 1, …, m" so they are strongly connected. Usually q-value is bigger than corresponding p-value, so it is more fitting (stronger) to use for multiple hypothesis testing.

Let's mark by $p_1, p_2, \ldots, p_m$ p-values corresponding to null hypotheses $H_{01}, H_{02}, \ldots, H_{0m}$ and $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, ordered p-values corresponding to $H_{0(1)}, H_{0(2)}, \ldots, H_{0(m)}$.

In all methods of controlling type I error we give up on controlling set significance level $\boldsymbol{\alpha}$ for each tested hypothesis and choose to control overall FWER or FDR at level q. Because FWER and FDR are equal (when all null hypotheses are true) and FDR $\leq FWER$ (when number of true null hypotheses is smaller than m), methods that control the FWER control also FDR.

Methods of controlling FWER:

- Bonferroni – significant are hypotheses with $p_i \leq q\backslash m$, which ensure that overall $\mathrm{FWER} \leq q$;
- Holm-Bonferroni – sequential algorithm, stronger than previous method, uses ordered p-values, starting from the smallest $p_{(1)}$ for each $p_{(i)}$ checks whether $p_{(i)} < \frac{q}{m+1-i}$ and if it's true rejects $H_{0(i)}$ and goes to i+1, otherwise stops. At the result FWER $\leq q$ in strong sense;
- Hochberg – very similar to previous method and even more powerful, starts from the biggest $p_{(m)}$ and doesn't reject any hypothesis until $p_{(i)} < \frac{q}{m+1-i}$, than it rejects $H_{0(1)}$, …, $H_{0(i)}$, overall FWER $\leq q$;
- Hommel – similar to Hochberg's, start with finding $k = max\{i \in \{1, \ldots, m\}: p_{(m-i+j)} > \frac{j\alpha}{i} for\ j = 1, .., i\}$, than it rejects null hypotheses with $p_i \leq \alpha/k$ or all when no maximum exists;

There are different methods used to control FDR, same of them use formula for density of p-values: $f(p) = \pi_0 * f_0(p) + (1 - \pi_0) * f_1(p)$, where $f_0(p)$ – density under null hypothesis, which uniform for continuous tests, $f_1(p)$ – density under alternative hypothesis, usually unknown, π_0- estimated parameter.

Methods of controlling FDR (continuous tests):

- Benjamini-Hochberg (BH) – rejects hypotheses $H_{0(1)}$, …, $H_{0(k)}$ for $k = \max\{i : p_{(i)} \leq \frac{i}{m} * q\}$;
- Benjamini-Liu (BL) – similar to BH, first calculates critical values for each i: $\delta_i = 1 - [1 - \min\left(1, m * \frac{q}{m-i+1}\right)]^{1/(m-i+1)}$, then finds $k = \min\{i : p_{(i)} > \delta_i\}$ and rejects $H_{0(1)}$, …, $H_{0(k-1)}$;
- Storey – different from BH and BL, first fixes rejection regions $\{[0, \gamma_{(i)}]\}$ for each i (e.g. for $\gamma_{(i)} = p_{(i)}$, rejected region is: $\{p_{(1)}, \ldots, p_{(i)}\}$), then for each region estimates FDR and then chooses good enough estimate of FDR with corresponding set of p-values and so null hypothesis to reject;

Controlling FDR of discrete test is less discussed in literature and has same problems, e.g. p-values density under null hypothesis usually is not uniform, moreover p-values depends on ancillary statistic, so as a result for each test density functions are usually different. Suggested solution is using modified p-values: midP-values, which are average of $p_{(i)}$ and the next smallest possible, then using method BH.

There are same methods, which enable future improve control over FDR called adaptive procedures, which use different approaches to estimate $m_0 = \pi_0 * m$ and then replace m with $\widehat{m}_0$.

Methods of estimating $\hat{\pi}_0$ (and thus $\widehat{m}_0$) for continuous tests:

- Storey: $\hat{\pi}_0 = \frac{\#(p_i>\lambda)}{m*(1-\lambda)}, for\ \lambda \in [0,1]$ – tuning parameter, #(S) number of elements in S;
- Pounds and Cheng: $\hat{\pi}_0 = \begin{cases} \min(1,2\overline{p})\, for\ two-sided\ tests \\ \min(1,2\overline{t})\ \ for\ one-sided\ tests \end{cases}$, where $\overline{p}$ -mean of p_i, $\overline{t} = \frac{1}{m} * \sum_{i=1}^{m}[2 * \min{(p_i, 1-p_i)}]$. It is biased upwardl;
- Location Based Estimator: $\hat{\pi}_0 = (\frac{1}{m} * \sum_{i=1}^{m}[-\log(1-p_i)]^k)/k!$, where $k \geq 0$ – tuning parameter. It provides bias-variance balance and often works better than two abow;
- Nettleton: computes $\widehat{m}_0$ in following steps: 1) partitioning [0,1] in B equal-width bins, 2) assuming all null hypotheses are true, set $m_0 = \pi_0 * m = m$, 3) calculating "the expected number of p-values for each bin given the current estimate of the number of true null hypotheses, 4) beginning with the leftmost bin, sum the number of p-values in excess of the expected until a bin with no excess is reached. 5) use the excess sum as an updated estimate of m_1" then update $m_0 = m - m_1$, 6) return to 3) and repeat until m_0 stops changes;

Methods of estimating $\hat{\pi}_0$ (and thus $\widehat{m}_0$) for discrete tests:

- Pounds and Cheng: $\hat{\pi}_0 = \begin{cases} \min(1,2\overline{p})\, for\ two-sided\ tests \\ \min(1,\mathbf{8}\overline{t})\ \ for\ one-sided\ tests \end{cases}$, rest the same as in for continuous tests;
- Regression Method: used when mixture distribution can be estimated from data. $\hat{\pi}_0$ is estimated by slope regression equation of distribution function $f(p) = \pi_0 * f_0(p) + (1-\pi_0) * f_1(p)$, with statistically estimated from data f(p), $f_0(p)$ and $f_1(p)$, and with assumption that $(1-\pi_0) * f_1(p)$ is constant;
- Bancroft: adaptation of Nettleton method for continuous tests to discrete tests;
- T-Methods: removing part of hypotheses with no power before any analysis, e.g. tests with p-values which cannot be smaller than set significance level $\boldsymbol{\alpha}$**.** Then proceeding with one of the previous method;

Gilberts Procedure – analogous to T-Methods, it proceeds with BH algorithm only on those tests with power.

Same of the previous mention methods of controlling FWER/FDR work under certain circumstances when assumption of independence among the tests is not fulfilled and tests are **positive dependent** (so $P(X \cap Y) > P(X) + P(Y)$ for X, Y positive dependent random variables). These methods are: BH, BL, Hochberg and Hommel (both use Simes inequality), adaptive Holm and Hochberg and same modification. To more complicated hypotheses we can use pairwise competition of part or all of their parameters, e.g. methods described by Turkey, Krammer and Games.

There are also other problems and corresponding methods in this field, e.g. working with multiple hypotheses or using weighted p–values to incorporate same known information about hypotheses;

Bibliography:

[1] Austin et al. (2014) - Stefanie R. Austin, Isaac Dialsingh, and Naomi S. Altman, *Multiple Hypothesis Testing: A Review* June 4, 2014

[2] Storey, John D., Tibshirani, Robert, *Statistical significance for genomewide studies*, Proceedings of the National Academy of Sciences 2003 May, doi: 10.1073/pnas.1530509100

[3] Dahiru T. P - value, a true test of statistical significance? A cautionary note. Ann Ib Postgrad Med. 2008 Jun;6(1):21-6. doi: 10.4314/aipm.v6i1.64038. PMID: 25161440; PMCID: PMC4111019.

[4] https://en.wikipedia.org/wiki/Test_statistic

[5] https://en.wikipedia.org/wiki/F-test

[6] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test