Significance and multiple testing

Yurii Aulchenko

yurii [dot] aulchenko [at] gmail [dot] com

August 21, 2012

Outline

1. **Significance**
   - Definitions
   - Significance of the score test
2. **Multiple testing**
   - Examples
   - Dealing with multiple testing

Contents

Definitions

- An *experiment* is a planned process of data collection
- Let our experiment consist of measuring outcome y and predictor x in a set of people
- Let *"null hypothesis"* be the hypothesis of no association between x and y
- Data gathered in an experiment can be characterized by a *test statistics*, which measures how much the data "deviate" from expected under the null. Here we will use the score test $T^2 = \hat{\rho}^2 \cdot n$
- The p-value is defined as the probability to obtain the value of test statistic at least as big as the observed one, given null hypothesis is true

Definitions (continued)

- In an experiment ...
  - We will reject the null hypothesis if the test statistic is greater or equal to some pre-defined threshold termed *critical value*
  - *Type 1 error* occurs when the null hypothesis is rejected when it is true
  - The p-value can be interpreted as the probability that null hypothesis is rejected while it is true
  - Thus the smaller is the p-value, the greater is our confidence that the null hypothesis is not true
- If experiments are repeated (and test statistic is proper!), the type 1 error rate should converge to the *p-value*

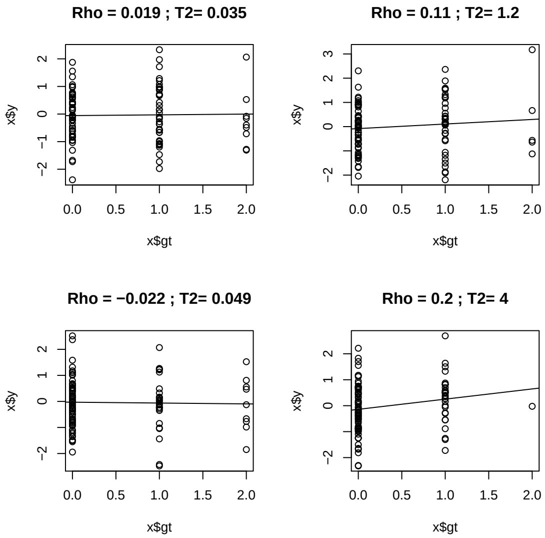
Picking up critical value

- The score test T^2 is known to be distributed as χ^2_1 if null hypothesis is true
- For the χ^2_1 distribution, we know that 5% of the distribution is behind the point of 3.84
- Let us use 3.84 as critical value: we will reject null hypothesis if observed $T^2 \geq 3.84$
- If experiments are repeated over and over again, the type one error rate (proportion of times we rejected null hypothesis) should converge to 5%
- Let us check this!

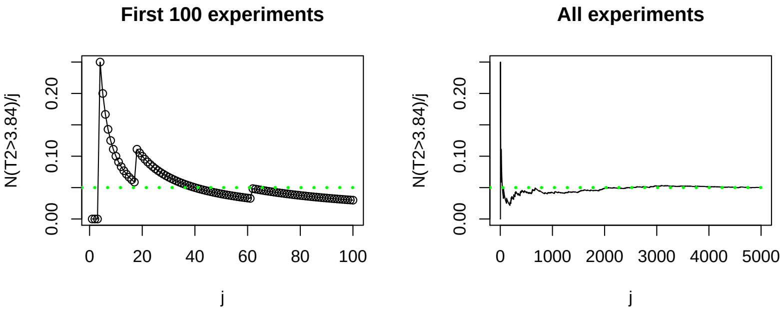
Checking the score test – outline

- I will simulate a phenotype for 100 people at random, and then independently simulate a genotype (so, null hypothesis holds!)
- For each simulated data set, I will estimate the coefficient of determination $\hat{\rho}^2$ and obtain the score test value $T^2 = n \cdot \hat{\rho}^2 = 100 \cdot \hat{\rho}^2$
- I will record this value, and repeat the procedure over and over again
- At the end, I will have a vector $\{T_1^2, T_2^2, ..., T_j^2, ..., T_m^2\}$ of obtained test values. The type one error rate is the proportion of T^2's ≥ 3.84
- We will check is this error rate converges to the *p*-value, corresponding to the chosen critical threshold of 3.84 (which is 5%)

Example of simulated data analysis - first 4 experiments



Convergence of type 1 error rate to the *p*-value



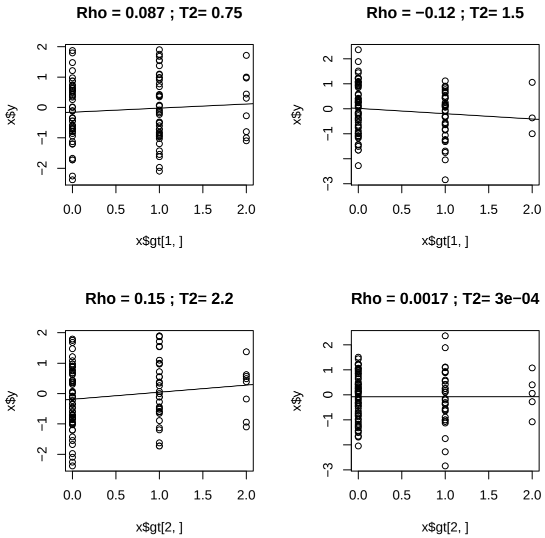
- The figures show how type 1 error rate converges to the *p*-value when we repeat experiment
- Type 1 error rate at point j is defined as the proportion of T^2's ≥ 3.84 among in experiments from 1 to j

Contents

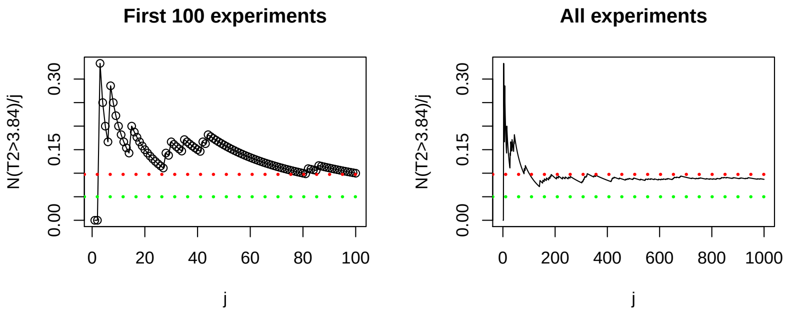
Experiment with two tests

- Let our experiment consist of testing TWO SNPs for association with an outcome of interest
- We will claim significant association (reject the null) if we observe the $T^2 \geq 3.84$ for either of the two SNPs
- What will by the type 1 error in this scenario? Will it still be the same as the p-value (5%)?
- Let us check this!

Example of simulated data analysis - first 2 experiments

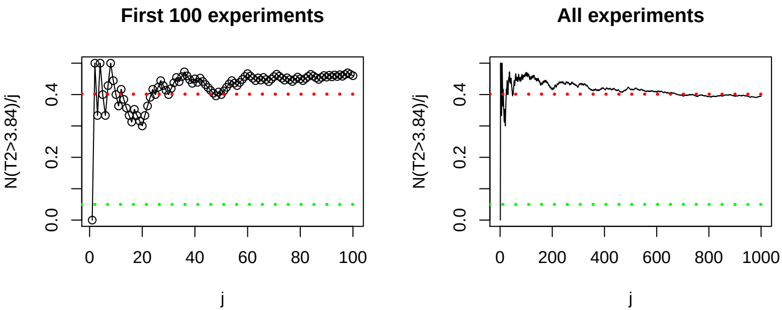


Type 1 error rate when 2 SNPs are tested



- Null hypothesis is rejected if $T^2 \geq 3.84$ for any of SNPs tested
- The figures show how type 1 error rate DOES NOT converge to the p-value of 5%
- Instead, it converges to the value of 0.0975

Type 1 error rate when 10 SNPs are tested



- Null hypothesis is rejected if $T^2 \geq 3.84$ for any of SNPs tested
- Type 1 error rate converges to the value of 0.401

Dealing with multiple testing

- Clearly, when doing multiple tests in an experiment, experiment-wise type 1 error rate is not the same as the p-value for a single test
- For example, when experiments consist of 2 tests, and we claim significance if in any of them we obtain p-value ≤ 0.05 ($T^2 \geq 3.84$), experiment-wise type 1 error rate is 0.0975

Estimation of Type 1 error rate

Experiment-wise type 1 error rate can be estimated from critical p-value level used and the number of tests per experiment:

$$
\begin{aligned}
&P(T_1^2 \geq 3.84 \cup T_2^2 \geq 3.84) = \\
&\quad P(T_1^2 \geq 3.84) \cdot P(T_2^2 < 3.84) + \\
&\quad P(T_1^2 < 3.84) \cdot P(T_2^2 \geq 3.84) + \\
&\quad P(T_1^2 \geq 3.84) \cdot P(T_2^2 \geq 3.84) = \\
&\quad p(1-p) + (1-p)p + p^2 = \\
&\quad 0.05 \cdot 0.95 + 0.95 \cdot 0.05 + 0.05 \cdot 0.05 = 0.0975
\end{aligned}
$$

Patching p to keep T1E

- Actully, we can figure out what the nominal p-value should be so that type 1 error is acceptable (e.g. is 5%):

$$0.025 \cdot 0.975 + 0.975 \cdot 0.025 + 0.025 \cdot 0.025 \approx 0.05$$

- Thus, if we used p-value of 0.025 as critical level to claim significance, type 1 error rate would have been 5%!

Bonferroni correction

This is the basic idea underlying the *Bonferroni correction*: in an experiment including N independent tests, to keep experiment-wise type 1 error rate of α, the critical p-value is chosen

$$p_{crit} = \frac{\alpha}{N}$$

and results of an expedient are claimed significant if p-value less than p_{crit} was reached in any of the tests