# Statistically Speaking: Multiple Testing Correction

DAVID L. TABB, PH.D.

NOVEMBER 23, 2017

# Overview

- Type I and Type II error
- Controlling Family-Wide Error Rate (FWE):

  Bonferroni
- False Discovery Rate (FDR): Benjamini-Hochberg
- Olive Jean Dunn

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Why do we need Multiple Testing Correction?

Each row is a test

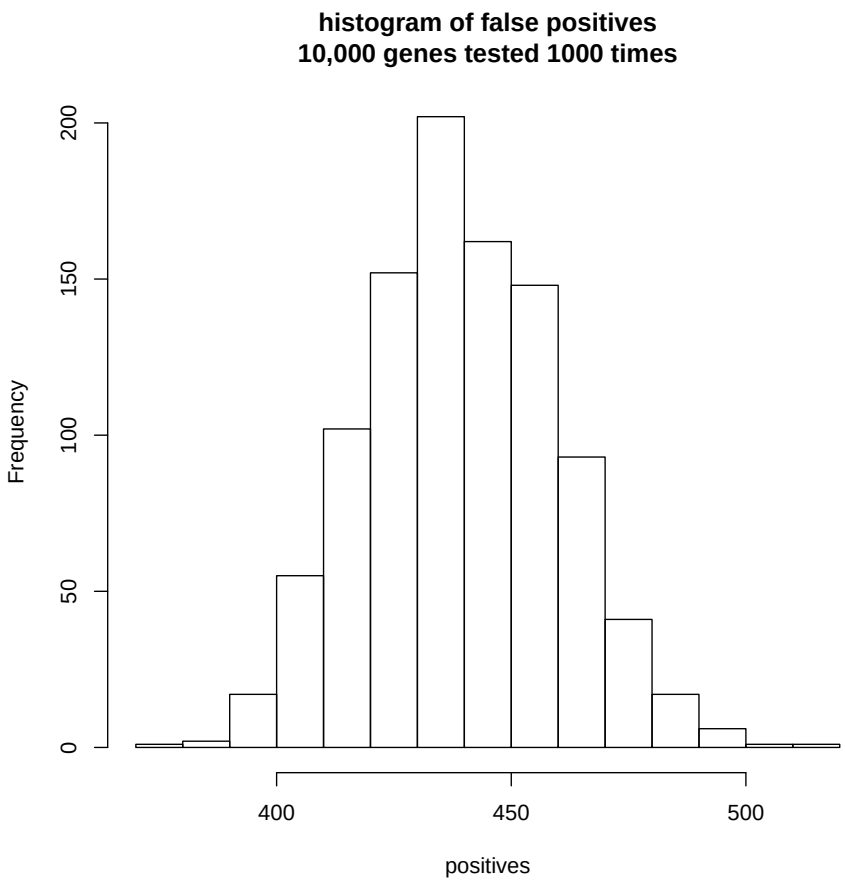
| probe_set_id | HNE0_1 | HNE0_2 | HNE0_3 | HNE60_1 | HNE60_2 | HNE60_3 |
|---|---|---|---|---|---|---|
| 1007_s_at | 8.6888 | 8.5025 | 8.5471 | 8.5412 | 8.5624 | 8.3073 |
| 1053_at | 9.1558 | 9.1835 | 9.4294 | 9.2111 | 9.1204 | 9.2494 |
| 117_at | 7.0700 | 7.0034 | 6.9047 | 9.0414 | 8.6382 | 9.2663 |
| 121_at | 9.7174 | 9.7440 | 9.6120 | 9.7581 | 9.7422 | 9.7345 |
| 1255_g_at | 4.2801 | 4.4669 | 4.2360 | 4.3700 | 4.4573 | 4.2979 |
| 1294_at | 6.3556 | 6.2381 | 6.2053 | 6.4290 | 6.5074 | 6.2771 |
| 1316_at | 6.5759 | 6.5330 | 6.4709 | 6.6636 | 6.6438 | 6.4688 |
| 1320_at | 6.5497 | 6.5388 | 6.5410 | 6.6605 | 6.5987 | 6.7236 |
| 1405_i_at | 4.3260 | 4.4640 | 4.1438 | 4.3462 | 4.3876 | 4.6849 |
| 1431_at | 5.2191 | 5.2070 | 5.2657 | 5.2823 | 5.2522 | 5.1808 |
| 1438_at | 7.0155 | 6.9359 | 6.9241 | 7.0248 | 7.0142 | 7.0971 |
| 1487_at | 8.6361 | 8.4879 | 8.4498 | 8.4470 | 8.5311 | 8.4225 |
| 1494_f_at | 7.3296 | 7.3901 | 7.0886 | 7.2648 | 7.6058 | 7.2949 |
| 1552256_a_at | 10.6245 | 10.5235 | 10.6522 | 10.4205 | 10.2344 | 10.3144 |
| 1552257_a_at | 10.3224 | 10.1749 | 10.1992 | 10.2464 | 10.2191 | 10.2405 |

Table by Bing Zhang

# Type I and Type II

- Each p-value from a t-test estimates how frequently data these differential or more would be, given no actual difference ($H_0$).
- To incorrectly reject the null hypothesis is to claim difference when none exists. This is a ***type I error***, with probability alpha (α).
- To incorrectly judge that no difference exists for a real difference is a ***type II error*** (β).

# With no real differences, do we still “detect” differences?




```
trials <- 1000
genes <- 10000
positives <- rep (0,trials)
for (counter in 1:trials) {
   tvals <- rep(0,genes)
   for (gcounter in 1:genes) {
            a<- rnorm(5)
            b<- rnorm(5)
            tvals[[gcounter]] <- t.test(a,b)$p.value
            }
   hits <- subset(tvals, tvals<0.05)
   positives[[counter]] <- length(hits);
   }
```

When no differences exist, T-test p-values are uniformly distributed.

# Expected values

- If you perform twenty t-tests and use 0.05 as a threshold, you should expect one false "hit" ($1.0 = 20 * 0.05$).
- If you perform 100 t-tests and use 0.01 as a threshold, you should expect one false "hit."
- You must choose your threshold *before* seeing how many "hits" you will get!

UNIVERSITEIT STELLENBOSCH UNIVERSITY

# Multiple testing correction: **Bonferroni** protects against *any* errors

- Family-Wide Error Rate (FWE) estimates the probability that *none* of the claimed differences are incorrect.
- Carlo Bonferroni, an Italian financial mathematician, framed a set of inequalities used in the “union bound” of probability.
- Olive Dunn, American biostatistician, adapted these inequalities to FWE control.

Olive Jean Dunn. *J. Amer. Statis. Assoc*. (1961) 56: 52-64.

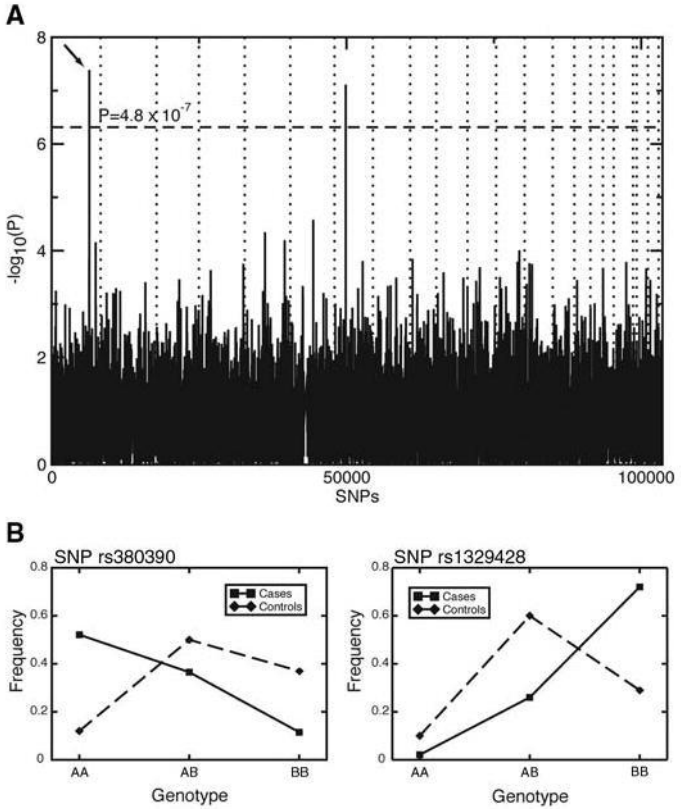
# The Bonferroni method

- $T_a = \frac{T_o}{n}$
- $T_o$ is original p-value threshold
- $T_a$ is adjusted p-value threshold
- $n$ is number of difference tests to perform
- If I compare 40 cytokines and want no false positives in my difference list to $\alpha < 0.05$,
- My adj. threshold = $0.00125 = \frac{0.05}{40}$

NOTE: Bonferroni keeps α down, but β may soar with large $n$!

# The Manhattan plot

- SNPs are evenly spaced on the x-axis.
- Height is the –log of p-value for one SNP.
- Correcting for multiple comparisons requires $p < \frac{0.05}{103{,}611}$ for a "hit" to protect against any false hits. (Bonferroni)




RJ Klein et al. *Science* (2005) 308: 385-389

# Multiple testing correction: **B-H FDR** limits *rate* of errors

www.tau.ac.il/~ybenja/

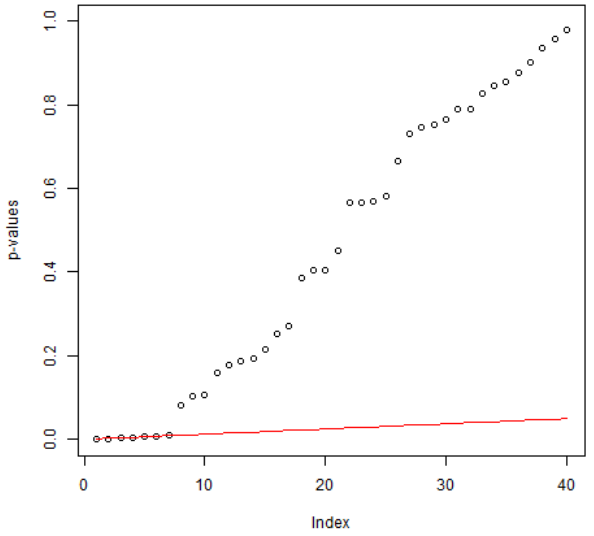
- Controlling against *any* errors with a large number of comparisons is too punitive.
- Limiting errors to be no greater than a defined fraction of all claimed differences is *much* less damaging to sensitivity.
- Instead of Family-Wide Error (FWE), B-H controls False Discovery Rate (FDR).

Y. Benjamini and Y. Hochberg. *J. Royal Statis. Soc*. (1995) 57: 289-300.

# Benjamini-Hochberg is a "step-up" procedure.

www.tau.ac.il/~ybenja/

- We return to our 40 cytokine example.
- Sort our 40 p-values from lowest to highest.
- Threshold for *each* p-value: $T_a = \frac{r}{n} T_o$, where $r$ is rank.
- Effective thresholds: 0.00125, 0.00250, 0.00375, 0.00500, 0.00625, 0.00750, 0.00875…




http://www.statisticshowto.com/benjamini-hochberg-procedure/

# Common multiple testing complaints

- "All my hits vanished after MTC."
  - Maybe the null distribution was correct!
  - Maybe your experiment was underpowered.
- "Dr. X published without MTC, so why must I?"
  - People publish bad statistics all the time. *Don't be one of them*.
- "I did five groups, so I just compared between each pair of groups."
  - Comparing between all pairs in five groups yields ten comparisons; you must correct!

$$\frac{n(n-1)}{2} = \frac{5*4}{2} = 10$$

# Olive Jean Dunn (1915-2008)

*Photo from collaborator, Dr. Ruth Mickey*

- Completed Mathematics B.A. and M.A. at UCLA in 1936 and 1951, respectively. Defended her Ph.D. in 1956.
- After three years as assistant professor at Iowa State College, she returned to UCLA in 1959, joining the department of biostatistics.
- In 1968, she was named “fellow” of American Statistical Association. She also served in the American Association for the Advancement of Science and the American Public Health Assoc.

http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470248793.html

# Humility

"The method given here is so simple and so general that I am sure it must have been used before this. I do not find it, however, so can only conclude that perhaps its very simplicity has kept statisticians from realizing that it is a very good method in some situations. In any case, the users of statistics in the main seem unaware of it, so I feel that it is worth presenting."

Olive Jean Dunn. *J. Amer. Statis. Assoc*. (1961) 56: 52-64.

# Takeaways

- Experiments producing many statistical tests require multiple testing correction.
- Bonferroni prevents making any wrong calls, but it requires very low p-values for a “hit.”
- BH is a more liberal correction, controlling rate with a threshold that “steps up.”
- While Bonferroni and Benjamini-Hochberg are dominant methods for MTC, others exist.