Permutation Tests & False Detection Rate

Session 10

Module 1 Probability & Statistical Inference

The Summer Institutes

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

Permutation Tests

Computer-intensive methods for hypothesis testing

Used when distribution of the test statistic (under the null hypothesis) is unknown

Permutation tests maintain the Type I error level without any large sample approximations / assumptions

EXAMPLE

HPV Vaccine Trial

200 uninfected women are randomly assigned 1:1 to HPV vaccine or placebo (i.e., 100 to each group). After 1 year subjects are tested for HPV infection (yes/no).

Scientific Question
Is the risk of infection the same or different in the two groups?

Restate scientific question as statistical hypotheses:

H_0: $p_v = p_p$

H_A: $p_v < p_p$

where

p_V = probability of infection in the vaccine group

p_p = probability of infection in the placebo group

EXAMPLE

HPV Vaccine Trial

Scientific Question
Is the risk of infection the same or different in the two groups?

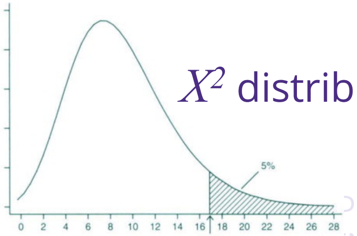
	Vaccine	Placebo	
HPV+	20	40	60
HPV-	80	60	140
	100	100	**200**

The overall infection rate is 30%, but we observe:

20% for vaccine
40% for placebo

How could we could test for differences in infection rate between the groups?

$$X^2 = \frac{N(ad - bc)^2}{n_1 \cdot n_2 \cdot m_1 \cdot m_2}$$



X^2 distribution for 1 df

$$RR = {}^{p_v}/_{p_p}$$

$$OR = p_v(1 - p_v)/(p_p(1 - p_p))$$

$$RD = p_v - p_p$$

EXAMPLE

HPV Vaccine Trial

Scientific Question
Is the risk of infection the same or different in the two groups?

	Vaccine	Placebo	
HPV+	20	40	60
HPV-	80	60	140
	100	100	**200**

The overall infection rate is 30%, but we observe:

20% for vaccine
40% for placebo

How could we could test for differences in infection rate between the groups?

But…

What if we repeated the experiment … would we see similar results?

Could a difference this large be due to chance alone?

EXAMPLE

HPV Vaccine Trial

Scientific Question
Is the risk of infection the same or different in the two groups?

We need to pick a way of **summarizing the difference** in infection probabilities between vaccine and placebo groups.

Let's use the risk difference:

Example ⇒ $p_v - p_p$

One particular value (in this case, 0) of the summary statistic corresponds to the null hypothesis being exactly true.

Example ⇒ $p_v – p_p = 0$

🔑 **We expect values near 0 if the null hypothesis is true.**
We expect values far from 0 if the null hypothesis is false.

EXAMPLE

HPV Vaccine Trial

Scientific Question
Is the risk of infection the same or different in the two groups?

What is the null distribution for this scenario?

Imagine a deck of 200 cards. Write HPV+ on 60 of them. Shuffle, then deal into two piles of 100.

How many HPV+ were in the first pile vs the second pile? Compute the ”risk difference” value.

Shuffle and re-deal many times. This gives us a null distribution!

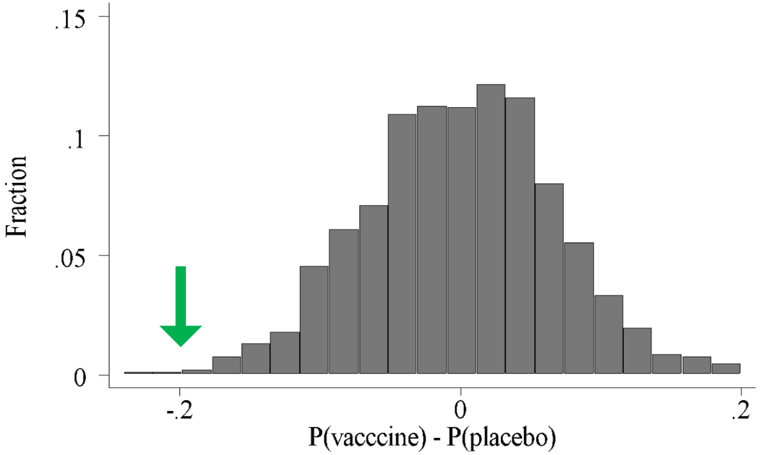
	Vaccine	Placebo	
HPV+	20	40	60
HPV-	80	60	140
	100	100	**200**

EXAMPLE

HPV Vaccine Trial

Scientific Question
Is the risk of infection the same or different in the two groups?

Here is a distribution of risk differences for the vaccine trial, permuted 2000 times.



What proportion of the simulated risk differences were greater than the observed risk difference? ***That's our p-value!***

P-value = probability of getting the observed outcome (or one more extreme given the direction of the alternative hypothesis) when the null hypothesis is true.

Here, only 3/2000 simulated differences were more extreme than the observed difference of -0.2. So p = 0.0015.

EXAMPLE

HPV Vaccine Trial

Scientific Question
Is the risk of infection the same or different in the two groups?

Summary We have answered our scientific question by using a permutation test.

1. Restate the scientific question as statistical hypotheses
2. Choose (any) reasonable summary statistic that quantifies deviations from the null hypothesis
3. Resample data assuming the null hypothesis is true and compute the summary statistic for each resampled data set.
4. Compare the observed value of the summary statistic to the null distribution generated in Step 3.

Summary: Permutation Tests

Useful when we can do resampling under the null hypothesis

Relatively few assumptions (i.e., no assumption about skewness or Normality of underlying distribution)

If the sample size is small, you can enumerate all possible permutations (permutation test)

If sample size is large, generate a random sample of permutations (randomization test)

Permutation samples are drawn without replacement

Many standard nonparametric methods (e.g., Wilcoxon Rank Sum Test) are permutation tests based on ranks.

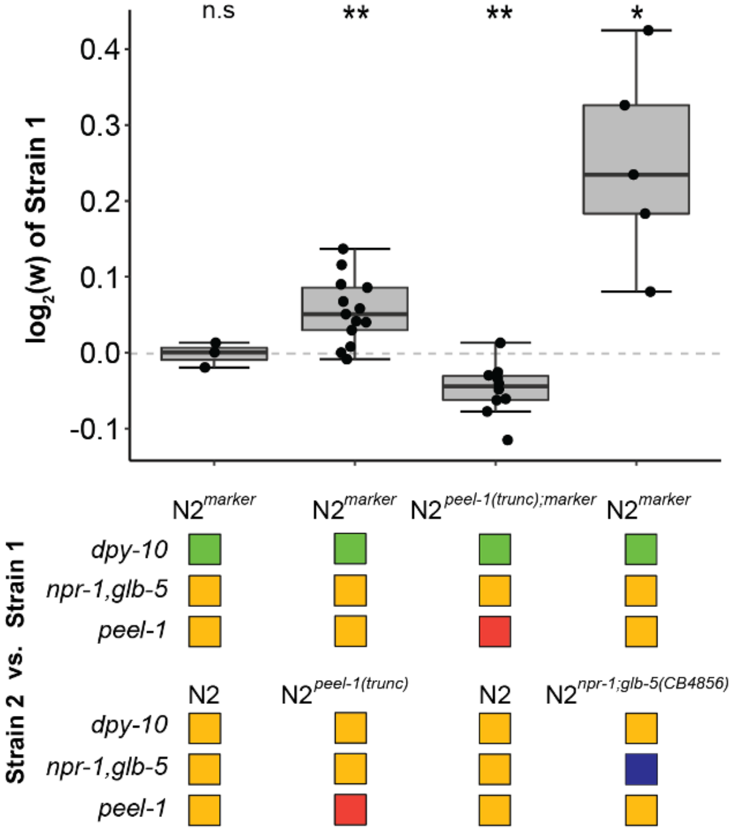
Good Reference: Manly (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman & Hall/CRC.

Correction for multiple tests

Here is an experiment my group recently conducted. We wanted to test hypotheses about whether mutations at particular genes affect fitness.

We created worm strains with our genotypes of interest. To test for fitness, we competed pairs of genotypes against each other.

On the plot, if the response variable is 0, the strains competed evenly; no difference in fitness.



	$N2^{marker}$	$N2^{marker}$	$N2^{peel\text{-}1(trunc);marker}$	$N2^{marker}$
dpy-10	🟩	🟩	🟩	🟩
npr-1,glb-5	🟨	🟨	🟨	🟨
peel-1	🟨	🟨	🟥	🟨

	N2	$N2^{peel\text{-}1(trunc)}$	N2	$N2^{npr\text{-}1;glb\text{-}5(CB4856)}$
dpy-10	🟨	🟨	🟨	🟨
npr-1,glb-5	🟨	🟨	🟨	🟦
peel-1	🟨	🟥	🟨	🟨

Correction for multiple tests

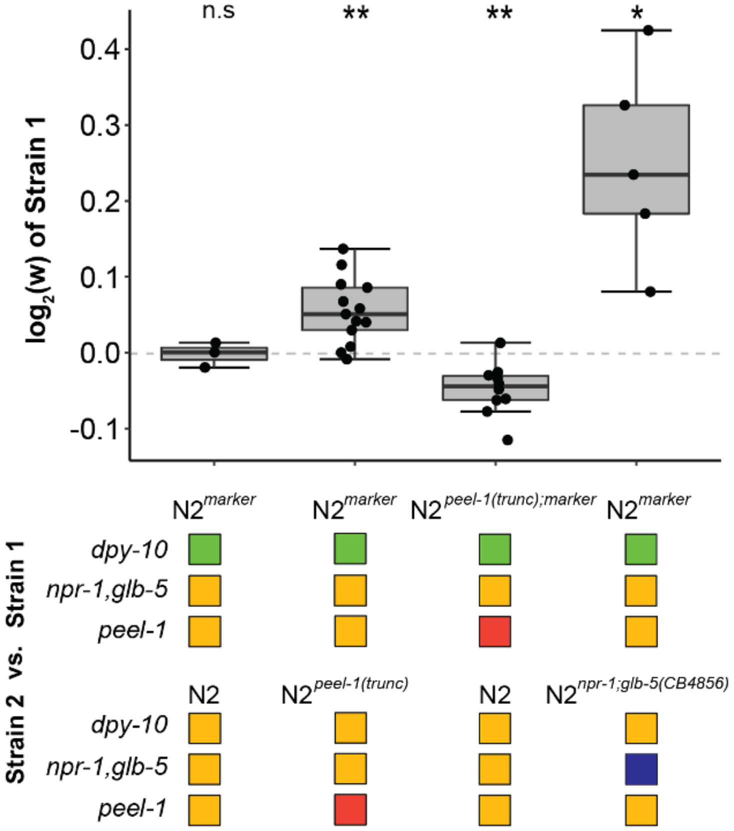
H_0: There is no difference in fitness between the strains.

$$log_2(w) of\ strain\ 1 = 0$$

H_A: There is a difference in fitness between the strains

$$log_2(w) of\ strain\ 1 \neq 0$$

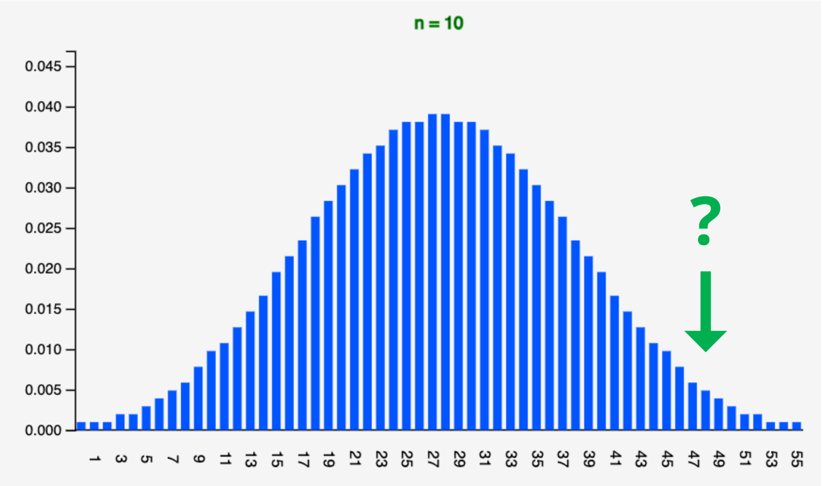
Each competition had a small number of replicates. We used a one-sample Wilcoxon non-parametric test to test the hypothesis for each of the four competitions.



Correction for multiple tests

Experiment	Observed test statistic
Control	W_1
WT vs *peel-1*	W_2
Peel-1 vs *WT*	W_3
WT vs *npr-1;glb-5*	W_4

For each experiment, we compare our observed test statistic to the Wilcoxon W distribution… is our observation sufficiently extreme to reject the null hypothesis?



Each experiment and analysis is conducted ***independently***.

Correction for multiple tests

So this scenario presents a problem of multiple tests. ***What is the problem?***

Suppose you are an unethical person and devise a get-rich-quick scheme to defraud people. You address 10,000 envelopes to 10,000 different people and include in each a unique claim.

> Our proprietary algorithm guarantees accurate stock market predictions! Purchase our service TODAY and SAVE!!! Send $100 by Aug 1 and your rate will be locked in for 12 MONTHS!!
>
> Don't believe us? This one time only, we are sharing with ONLY YOU the prediction that TOMORROW the stock of [x] will rise by [y]!

Each letter has different info for [x] and [y].

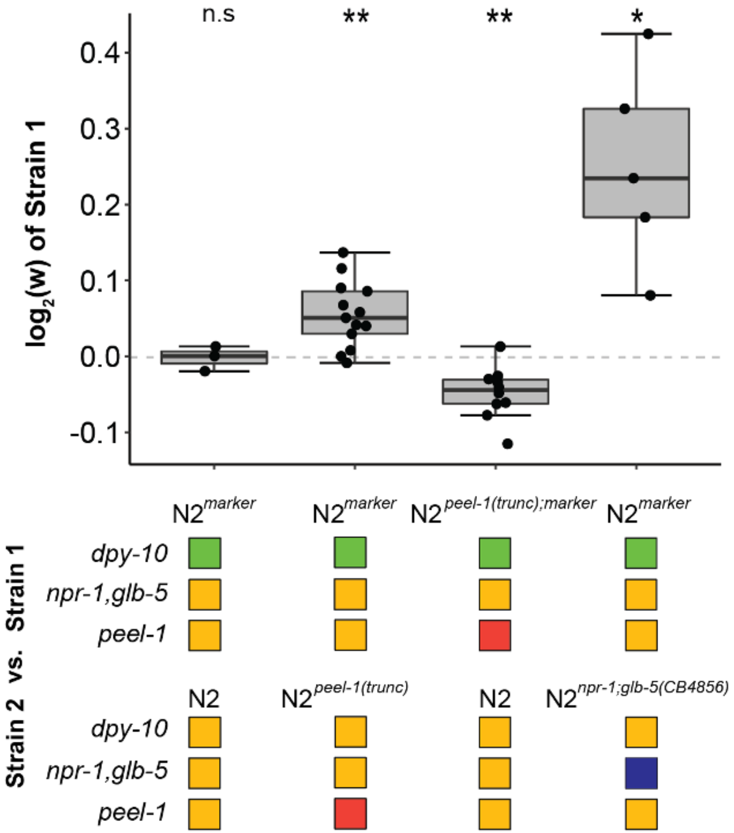
What do you think is going on here?

Correction for multiple tests

Although we conducted four ***independent*** experiments here, together they are exploring a central question.

With enough experiments, we'd most likely get a "significant" result in at least one of them ***even if the null hypotheses were always true***.

We need to control for multiple testing.



Correction for multiple tests

For each test, we compare the p-value to alpha.

$p < \alpha$ will lead us to reject H_0.

$p > \alpha$ will lead us to accept H_0.

For c tests, α_c gives the corrected alpha for committing at least one Type I error:

$$\alpha_c = 1 - (1 - \alpha)^c$$

What is α_c for 4 tests, if $\alpha = 0.05$? **0.1855**

What is α_c for 10 tests, if $\alpha = 0.05$? **0.4013**

How to correct for this problem?

We could simply adjust alpha by dividing by the total number of tests.

$$\alpha_{Bonferroni} = {}^{\alpha}/_{c}$$

But…

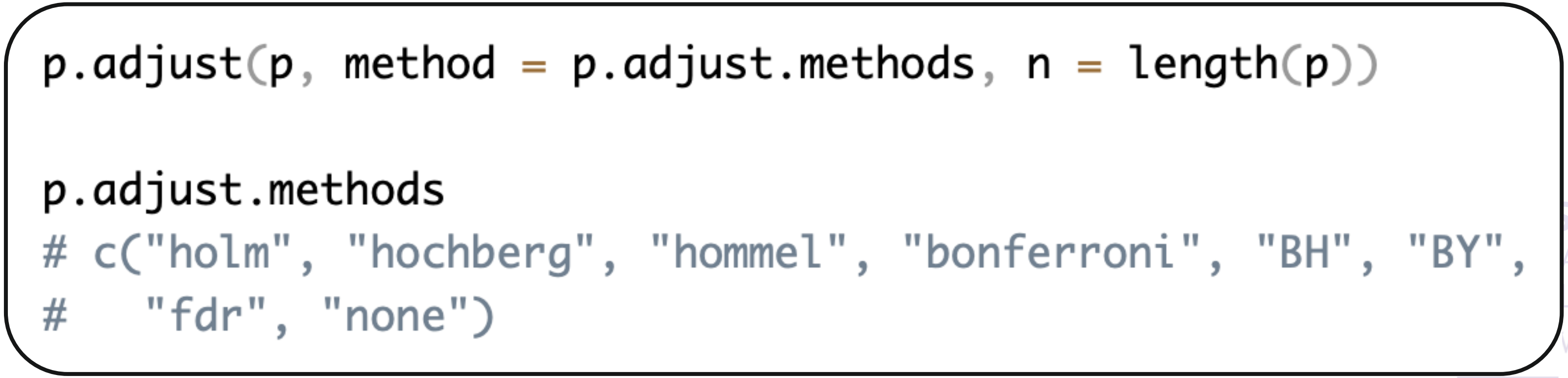
For $c = 4$, $\alpha_{Bonferroni} = {}^{0.05}/_{4} = 0.0125.$

For $c = 10$, $\alpha_{Bonferroni} = {}^{0.05}/_{10} = 0.005.$

Correction for multiple tests

How to correct for this problem?

The Bonferroni correction is very conservative. What are some other options?



```
p.adjust(p, method = p.adjust.methods, n = length(p))

p.adjust.methods
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
#   "fdr", "none")
```

The Holm method (or B-H or H-B) is a modification of the Bonferroni. Rank the p-values from smallest to largest. Does the lowest pass $^{\alpha}/_{c}$ threshold? If yes, proceed. Does next smallest p-value pass $^{\alpha}/_{c-1}$ threshold? If yes, proceed…

False Discovery Rate

For some studies, answering the scientific question of interest may require testing hundred, thousands, or millions of hypotheses. This is especially true of genetics.

Hedenfalk et al (2001) screened 3226 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.

The traditional solution for correcting for multiple tests, such as Bonferroni or Holm method, are far too conservative. It just doesn't work for high volume data.

New Solution: Don't eliminate false positives … control them.

False Discovery Rate

Hedenfalk et al (2001)
Screened 3226 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.

	Reject null	Fail to reject null	
Null true	F	m_0-F	m_0
Null not true (Alternative true)	T	m_1-T	m_1
	S	m-S	**m**

false positive rate = F/ m_0 = type I error rate = α

false discovery rate = F/S = q

Idea: Control the false discovery rate (q-value) instead of the false positive rate (related to the p-value)

EXAMPLE

False Discovery Rate

Hedenfalk *et al* (2001)

Hedenfalk et al (2001)
Screened ~~3226~~ 3170 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.

Order the 3170 p-values: $\mathbf{p_i}$ **,** i = 1,2, …, 3170
(56 genes were excluded from this analysis)

Pick a p-value cutoff, say α : reject H_o for all $p_i < \alpha$.

What is the false discovery rate (FDR) associated with this choice of α?

FDR = F / S

$S = \#\{p_i < \alpha\}$

$F = \alpha * m_0$

FDR = q-value = $\alpha * m_0 / \#\{p_i < \alpha\}$

I know S, I know α, but how do I know what is m_0?

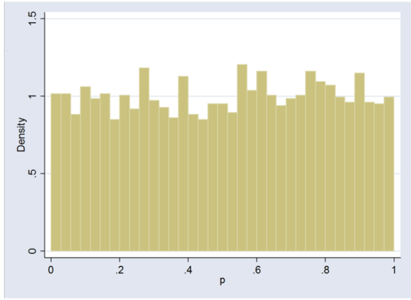
EXAMPLE

False Discovery Rate

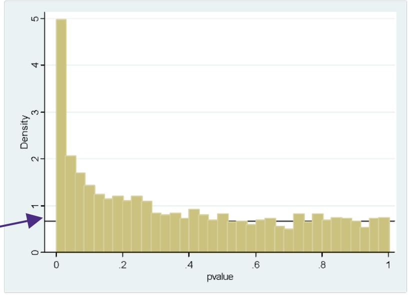
Hedenfalk *et al* (2001)

Hedenfalk et al (2001)

Screened ~~3226~~ 3170 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.



Distribution of 3170 p-values when all null hypotheses are true



0.676

Distribution of 3170 p-values from Hedenfalk *et al.*

Height of the line gives estimated proportion of true null hypotheses.

$$m_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1 \ldots m\}}{(1-\lambda)}$$

False Discovery Rate

Hedenfalk et al (2001)
Screened ~~3226~~ 3170 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.

	Reject null	**Fail to reject null**	
Null true	F	m_0-F	m_0 $=3170*0.676$ $=2143$
Null not true (Alternative true)	T	m_1-T	m_1
	S = $\#\{p_i<\alpha\}$	m-S	**m=3170**

false positive rate = F/m_0 = type I error rate = α (we set alpha)

false discovery rate = F/S = q → **F = q * S**

Idea: Control the false discovery rate (q-value) instead of the false positive rate (related to the p-value)

False Discovery Rate

Hedenfalk et al (2001)
Screened ~~3226~~ 3170 genes using microarrays to find differential expression between BRCA-1 and BRCA-2 mutation positive tumors.

$\mathbf{q(\alpha) = \alpha * m_0(\lambda) / \#\{p_i < \alpha\}}$
[technically $q(\alpha) = \min_{t\geq\alpha} q(t)$]

Package **`qvalue`** in R

<u>Example</u> : Analysis of data from Hedenfalk *et al* (using $m_0(0.5) = 2143$)

q false discovery rate	α false positive rate	$\#\{ p_i < \alpha \}$ expected # of positives	expected # of false positives
0.01	0.0000126	5	0
0.05	0.00373	160	8
0.10	0.0148	317	32

Compare: Using traditional methods Hedenfalk *et al* concluded 9-11 genes were differentially expressed