

# Unit 5: Inference for Numerical Data

Statistics 102 Teaching Team

March 09, 2020

Introduction

Two-sample test for paired data

Two-sample test for independent group data

Statistical power and sample size

Comparing many means with ANOVA

The multiple testing problem

# Introduction

# COMPARING TWO POPULATION MEANS

Two-sample data can be paired or unpaired (independent).

- Paired measurements for each 'participant' or study unit
  - each observation can be logically matched to one other observation in the data
  - e.g., scores on a standardized test before taking a prep course versus scores after the prep course
- Two independent sets of measurements
  - observations cannot be matched on a one-to-one basis
  - e.g., scores on a standardized test of students who did take a prep course versus scores of students who did not

The nature of the data dictate which two-sample testing procedure is appropriate: the two-sample test for paired data, or the two-sample test for independent group data.

## Two-sample test for paired data

## WETSUITS AND SWIMMING VELOCITY

Did a new wetsuit design allow for increased swim velocities during the 2000 Olympics?

A study was conducted to assess the effects of wetsuits on swim velocity.

- 12 competitive swimmers were asked to swim 1500m at maximal velocity, once wearing a wetsuit and once wearing a standard swimsuit.
- Order of wetsuit versus swimsuit randomized.
- Investigators recorded mean velocity (m/sec) for each trial.

# VIEWING SWIM VELOCITIES

```
library(oibistat)
data("swim")
swim
```

##	swimmer.number	wet.suit.velocity	swim.suit.velocity	velocity.diff
## 1	1	1.57	1.49	0.08
## 2	2	1.47	1.37	0.10
## 3	3	1.42	1.35	0.07
## 4	4	1.35	1.27	0.08
## 5	5	1.22	1.12	0.10
## 6	6	1.75	1.64	0.11
## 7	7	1.64	1.59	0.05
## 8	8	1.57	1.52	0.05
## 9	9	1.56	1.50	0.06
## 10	10	1.53	1.45	0.08
## 11	11	1.49	1.44	0.05
## 12	12	1.51	1.41	0.10

## IDEA BEHIND THE PAIRED $t$ -TEST

The velocities are paired by swimmer: each swimmer has two velocity measurements.

Suppose that for each swimmer  $i$ , we have observations  $x_{i,WS}$  and  $x_{i,SS}$ .

- Let  $d_i$  be the difference between the measurements:

$$d_i = x_{i,WS} - x_{i,SS}$$

- $x_{i,WS}$  is the wetsuit velocity measurement for swimmer  $i$
- $x_{i,SS}$  is the swimsuit velocity measurement for swimmer  $i$
- Base inference on  $\bar{d}$ , the sample mean of the  $d_i$ :

$$\bar{d} = \frac{\sum d_i}{n}$$



## THE PAIRED $t$ -TEST

Let  $\delta$  be the population mean of the difference in velocities during a 1500m swim if all competitive swimmers recorded swim velocities with each suit type.

The null and alternative hypotheses are

- $H_0 : \delta = 0$ , there is no difference in mean swim velocities between swimming with a wetsuit versus a swimsuit
  - i.e., wetsuits do not change mean swim velocities
- $H_A : \delta \neq 0$ , there is a difference in mean swim velocities between swimming with a wetsuit versus a swimsuit
  - i.e., wetsuits do change mean swim velocities

## THE PAIRED $t$ -TEST . . .

The formula for the test statistic is

$$t = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}},$$

where  $\bar{d}$  is the mean of the differences,  $s_d$  is the standard deviation of the differences, and  $n$  is the number of differences (i.e., number of pairs).

Note how the formula is identical to the one introduced in Unit 4.

- A paired  $t$ -test is essentially a one-sample test of difference values.

The  $p$ -value is calculated as usual, from a  $t$  distribution with  $df = n - 1$ .

## CONFIDENCE INTERVALS FOR PAIRED DATA

A 95% confidence interval for paired data has the form

$$\bar{d} \pm \left( t^* \times \frac{s_d}{\sqrt{n}} \right),$$

where  $t^*$  is the point on a  $t$  distribution with  $df = n - 1$  that has area 0.025 to its right.

# LETTING R DO THE WORK

```
#two-sample syntax
t.test(swim$wet.suit.velocity, swim$swim.suit.velocity,
       alternative = "two.sided", paired = TRUE)

##
## Paired t-test
##
## data: swim$wet.suit.velocity and swim$swim.suit.velocity
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of the differences
##                0.0775
```

Note: `t.test(x, y, paired = TRUE)` returns results based on the differences  $x - y$ .

## LETTING R DO THE WORK...

```
#one-sample syntax
```

```
t.test(swim$velocity.diff, mu = 0, alternative = "two.sided")
```

```
##  
## One Sample t-test  
##  
## data:  swim$velocity.diff  
## t = 12.318, df = 11, p-value = 8.885e-08  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
##  0.06365244 0.09134756  
## sample estimates:  
## mean of x  
##      0.0775
```

## Two-sample test for independent group data

# FAMUSS: COMPARING NDRM.CH BY SEX

Does change in non-dominant arm strength after resistance training differ between men and women?

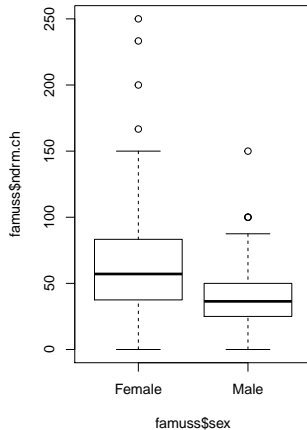
The *FAMuSS* study introduced in Unit 1 examined the change in non-dominant arm strength after resistance training.

```
data("famuss")
famuss[c(1, 2, 3, 595), c("sex", "age", "race", "height", "weight",
                           "actn3.r577x", "ndrm.ch")]
```

##	sex	age	race	height	weight	actn3.r577x	ndrm.ch
## 1	Female	27	Caucasian	65.0	199	CC	40.0
## 2	Male	36	Caucasian	71.7	189	CT	25.0
## 3	Female	24	Caucasian	65.0	134	CT	40.0
## 1348	Female	30	Caucasian	64.0	134	CC	43.8

# FAMuSS: COMPARING NDRM.CH BY SEX...

```
boxplot(famuss$ndrm.ch ~ famuss$sex)
```





# THE INDEPENDENT TWO-GROUP $t$ -TEST

The null and alternative hypotheses are

- $H_0 : \mu_F = \mu_M$ , the population mean change in arm strength for women is the same as the population mean change in arm strength for men
  - Equivalently,  $H_0 : \Delta = \mu_F - \mu_M = 0$
- $H_A : \mu_F \neq \mu_M$ , the population mean change in arm strength for women is different from the population mean change in arm strength for men

In general, the hypotheses are written in terms of  $\mu_1$  and  $\mu_2$ .<sup>1</sup>

- The parameter of interest is  $\mu_1 - \mu_2$ .
- The point estimate is  $\bar{x}_1 - \bar{x}_2$ .

---

<sup>1</sup>The numerical labels are arbitrary, so it is best to explicitly specify which group is considered group 1 versus group 2.

## THE INDEPENDENT TWO-GROUP $t$ -TEST...

The  $t$ -statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The  $p$ -value is calculated as usual, but the degrees of freedom for the distribution are different than for the paired data setting...

## DEGREES OF FREEDOM FOR THE INDEPENDENT TWO-GROUP $t$ -TEST

When doing the test by hand, use the following approximation:

$$df = \min(n_1 - 1, n_2 - 1)$$

R uses a better approximation, known as the Satterthwaite approximation:

$$df = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]}$$

## CONFIDENCE INTERVALS FOR INDEPENDENT TWO-GROUP DATA

The 95% confidence interval for the difference in population means has the form

$$(\bar{x}_1 - \bar{x}_2) \pm \left( t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right),$$

where  $t^*$  is the point on a  $t$  distribution that has area 0.025 to the right, with the same  $df$  as used for calculating the  $p$ -value of the associated test.

# LETTING R DO THE WORK

```
#define categories for sorting ndrm.ch
female = (famuss$sex == "Female"); male = (famuss$sex == "Male")

#comma syntax
t.test(famuss$ndrm.ch[female], famuss$ndrm.ch[male], mu = 0,
       alternative = "two.sided", paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  famuss$ndrm.ch[female] and famuss$ndrm.ch[male]
## t = 10.073, df = 574.01, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19.07240 28.31175
## sample estimates:
## mean of x mean of y
##  62.92720  39.23512
```

## LETTING R DO THE WORK...

Alternatively, take advantage of the way the data are structured and use the tilde ( $\sim$ ) syntax.

```
#tilde syntax
t.test(famuss$ndrm.ch ~ famuss$sex, mu = 0,
       alternative = "two.sided", paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  famuss$ndrm.ch by famuss$sex
## t = 10.073, df = 574.01, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19.07240 28.31175
## sample estimates:
## mean in group Female    mean in group Male
##          62.92720          39.23512
```

## PAIRED VS. INDEPENDENT DATA

What if the swimsuit data had been incorrectly analyzed with an independent two-group test?

```
t.test(swim$wet.suit.velocity, swim$swim.suit.velocity,  
       mu = 0, alternative = "two.sided", paired = FALSE)
```

```
##  
##  Welch Two Sample t-test  
##  
## data:  swim$wet.suit.velocity and swim$swim.suit.velocity  
## t = 1.3688, df = 21.974, p-value = 0.1849  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.03992937  0.19492937  
## sample estimates:  
## mean of x mean of y  
##  1.506667  1.429167
```

## Statistical power and sample size



## BACKGROUND

Most studies are done to establish evidence in favor of an alternative hypothesis.

The **power** of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true.

- In other words, power is the probability of correctly rejecting  $H_0$ .

Power depends on. . .

- the hypothesized difference between two population means, also known as the population **effect size** ( $|\mu_1 - \mu_2|$ )
- the population standard deviation of each group ( $\sigma_1, \sigma_2$ )
- the sample size of each group ( $n_1, n_2$ )

Usually, a study team can only control sample size.

## OUTCOMES AND ERRORS IN TESTING

	Result of test	
State of nature	Reject $H_0$	Fail to reject $H_0$
$H_0$ is true	Type I error, probability = $\alpha$ (false positive)	No error, probability = $1 - \alpha$ (true negative)
$H_A$ is true	No error, probability = $1 - \beta$ (true positive)	Type II error, probability = $\beta$ (false negative)

# ERROR PROBABILITIES IN TESTING

Lab 2 uses simulation to explore how Type I and Type II error are controlled, and examines factors influencing the power of a statistical test.

- Type I error is controlled via rejecting  $H_0$  only when a  $p$ -value is smaller than  $\alpha$ .
- Type II error (and power) is affected by sample size, standard deviation, and effect size.
  - As sample size increases, power increases.
  - As standard deviation decreases, power decreases.
  - As effect size increases, power increases.

# CHOOSING THE RIGHT SAMPLE SIZE

Study design often includes calculating a study size (sample size) such that the probability of rejecting a null hypothesis is acceptably large, typically between 0.80 and 0.90.

It is important to have a precise estimate of an appropriate study size.

- A study needs to be large enough to allow for sufficient power to detect a difference between groups.
- However, unnecessarily large studies are expensive, and can even be unethical.

## A TYPICAL SAMPLE SIZE QUESTION

A pharmaceutical company has developed a new drug to lower blood pressure and is planning a clinical trial to test its effectiveness.

- Individuals whose systolic blood pressure is between 140 and 180 mmHg will be recruited.
- Based on previous studies, blood pressures from such individuals will be approximately normally distributed with standard deviation of about 12 mmHg.
- Participants will be randomly assigned to the new drug or a placebo, and the company will measure the difference in mean blood pressure levels between the groups.

The company expects to receive FDA approval for the drug if there is evidence that the drug lowers blood pressure, on average, by at least 3 mmHg more than the standard drug.

How large should the study be if the company wants the power of the study to be 0.80 (80%)?

## A TYPICAL SAMPLE SIZE QUESTION . . .

*Of Biostat* Section 5.4 has an extended discussion of this example, with formulas for hand calculations.

Instead, we go right to the R function `power.t.test`. Details about syntax for this function are in the Unit 5 Lab Notes.

```
power.t.test(n = NULL, delta = 3, sd = 12, sig.level = 0.05, power = 0.80)
```

```
##  
##      Two-sample t test power calculation  
##  
##              n = 252.1281  
##            delta = 3  
##              sd = 12  
##      sig.level = 0.05  
##            power = 0.8  
##      alternative = two.sided  
##  
## NOTE: n is number in *each* group
```

## Comparing many means with ANOVA

# ANALYSIS OF VARIANCE (ANOVA)

Suppose we are interested in comparing means across more than two groups. Why not conduct several two-sample  $t$ -tests?

- If there are  $k$  groups, then  $\binom{k}{2} = \frac{k(k-1)}{2}$   $t$ -tests are needed.
- Conducting multiple tests on the same data increases the overall rate of Type I error.

ANOVA uses a single hypothesis test to assess whether means across many groups are equal:

- $H_0$ : mean outcome is same across all groups ( $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ )
- $H_A$ : at least one mean is different from the others (i.e., means are not all equal)



# IDEA BEHIND ANOVA

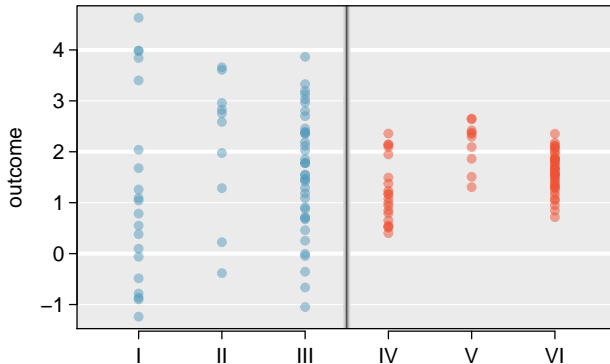
Is the variability in the sample means large enough that it seems unlikely to be from chance alone?

Compare two quantities:

- Variability between groups ( $MSG$ ): how different are the group means from each other, i.e., how much does each group mean vary from the overall mean?
- Variability within groups ( $MSE$ ): how variable are the data within each group?

$MSG$  denotes mean square between groups, while  $MSE$  denotes mean square error. Refer to *OI Biostat* Section 5.5.1 for details.

## IDEA BEHIND ANOVA...



- I, II, and III: difficult to discern differences in means, variability within each group is high
- IV, V, and VI: appears to be differences in means, these differences are large relative to variance within each group

## IDEA BEHIND ANOVA...

Under the null hypothesis, there is no real difference between the groups; thus, any observed variation in group means is due to chance.

- Think of all observations as belonging to a single group.
- Variability between group means should equal variability within groups

The *F-statistic* is the test statistic for ANOVA.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{MSG}{MSE}$$

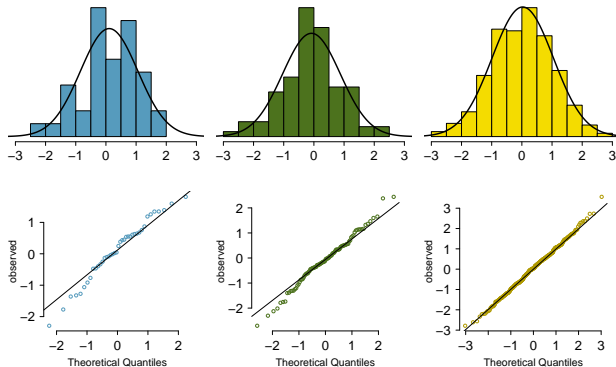
- When the population means are equal, the *F*-statistic is approximately 1.
- When the population means differ, *F* will be larger than 1. Larger values of *F* represent stronger evidence against the null.
- The *F* statistic follows an *F* distribution, with two degrees of freedom,  $df_1$  and  $df_2$ ;  $df_1 = n_{groups} - 1$ ,  $df_2 = n_{obs} - n_{groups}$ .
- The *p*-value for the *F*-statistic is the probability *F* is larger than the *F*-statistic.

# ASSUMPTIONS FOR ANOVA

It is important to check whether the assumptions for conducting ANOVA are reasonably satisfied:

1. Observations independent within and across groups
  - Think about study design/context
2. Data within each group are nearly normal
  - Look at the data graphically, such as with a histogram
  - Normal Q-Q plots can help. . .
3. Variability across groups is about equal
  - Look at the data graphically
  - Numerical rule of thumb: ratio of largest variance to smallest variance  $< 3$  is considered “about equal”

# NORMAL PROBABILITY PLOTS (Q-Q PLOTS)



If points fall on or near the line, data closely follow a normal distribution.

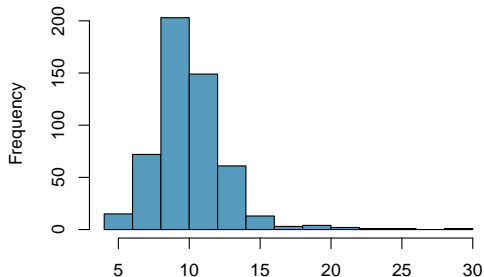
- Difficult to evaluate in small datasets
- Plots show three simulated normal datasets: from L to R,  $n = 40$ ,  $n = 100$ ,  $n = 400$

# NORMAL PROBABILITY PLOTS (Q-Q PLOTS)...

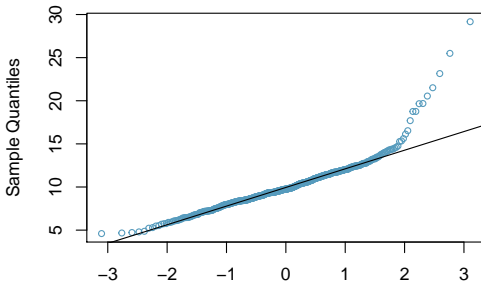
```
#simulate right-skewed distribution  
set.seed(2019)  
sim.data <- c(rnorm(500, 10, 2), rnorm(25, 15, 5))
```

```
#plots  
par(mfrow = c(1, 2))  
hist(sim.data, col = COL[1])  
qqnorm(sim.data, cex = 0.75, col = COL[1]); qqline(sim.data)
```

Histogram of sim.data



Normal Q-Q Plot



## PAIRWISE COMPARISONS

If the  $F$ -test indicates there is sufficient evidence that the group means are not all equal, proceed with pairwise comparisons to identify which group means are different.

Pairwise comparisons are made using the two-sample  $t$ -test for independent groups.

- To maintain the overall Type I error rate at  $\alpha$ , each pairwise comparison is conducted at an adjusted significance level referred to as  $\alpha^*$ .
- The Bonferroni correction is one method for adjusting  $\alpha$ .

$$\alpha^* = \alpha/K, \text{ where } K = \frac{k(k-1)}{2} \text{ for } k \text{ groups}$$

- Note that the Bonferroni correction is a very stringent (i.e., conservative) correction, made under the assumption that all tests are independent.

## FAMuSS: COMPARING NDRM.CH BY GENOTYPE

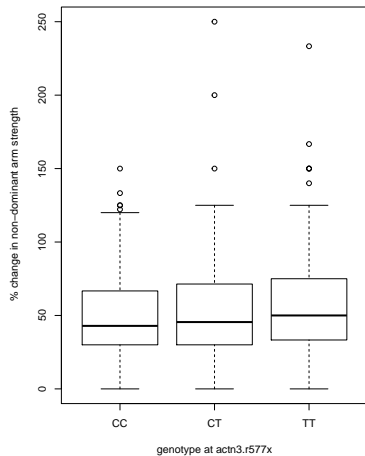
The main question of interest in the FAMuSS study can be approached with ANOVA.

*Is change in non-dominant arm strength after resistance training associated with genotype?*

Questions 1 - 3 in Lab 3 step through this analysis.



## FAMuSS: COMPARING NDRM.CH BY GENOTYPE...



## FAMuSS: COMPARING NDRM.CH BY GENOTYPE...

The null and alternative hypotheses are

- $H_0 : \mu_{CC} = \mu_{CT} = \mu_{TT}$ , the mean percent change in non-dominant arm strength is equal across the three genotypes
- $H_A$ : At least one group has mean percent change in non-dominant arm strength that is different from the other groups

Let  $\alpha = 0.05$ .

## LETTING R DO THE WORK

Formulas for hand calculations shown in *Ol Biostat* Section 5.5.1.

```
#use summary(aov())  
summary(aov(famuss$ndrm.ch ~ famuss$actn3.r577x))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)  
## famuss$actn3.r577x    2    7043     3522   3.231 0.0402 *  
## Residuals          592 645293     1090  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:  $p < \alpha$ , sufficient evidence to reject  $H_0$  in favor of  $H_A$ . There is at least one group with a mean different from the other groups.

- But which groups have different means?

## CONTROLLING TYPE I ERROR RATE

If the ANOVA  $F$ -test is significant, then it is appropriate to proceed to conducting pairwise comparisons; i.e., using two-sample  $t$ -tests to compare each possible pairing of the groups.<sup>2</sup>

- Each test should be conducted at the  $\alpha^*$  significance level so that the overall Type I error rate remains at  $\alpha$ .
- These tests are still conducted under the assumption that the variance between groups is equal; thus, the test statistics are calculated using the pooled estimate of standard deviation between groups. Details are in *Ol Biostat* Section 5.5.3.
- We will use `pairwise.t.test( )` to perform these *post hoc* two-sample  $t$ -tests. Refer to Unit 5, Lab 2 for an example; this function is also discussed in the Unit 5 Lab Notes.

---

<sup>2</sup>These  $t$ -tests are typically referred to as *post hoc* tests.

## CONTROLLING TYPE I ERROR RATE...

Pairwise comparisons using two-sample  $t$ -tests (CC to CT, CC to TT, CT to TT) can now be done if the Type I error rate is controlled.

- Apply the Bonferroni correction.
- In this setting,  $\alpha^* = 0.05/3 = 0.0167$ .

## LETTING R DO THE WORK

Only CC versus TT resulted in a  $p$ -value less than  $\alpha^*$  of 0.0167.

- Mean strength change in non-dominant arm for CT individuals not distinguishable from strength change for CC and TT.
- However, evidence at  $\alpha = 0.05$  level that mean strength change for individuals of genotype CC and TT are different.

```
pairwise.t.test(famuss$ndrm.ch, famuss$actn3.r577x, p.adj = "none")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: famuss$ndrm.ch and famuss$actn3.r577x  
##  
##      CC      CT  
## CT 0.179 -  
## TT 0.011 0.144  
##  
## P value adjustment method: none
```

## LETTING R DO THE WORK...

Alternatively, set `p.adj` to "bonf"; this instructs R to rescale the  $p$ -values (by multiplying by  $K$ ) so they can be compared to the original  $\alpha$  level of 0.05.

```
pairwise.t.test(famuss$ndrm.ch, famuss$actn3.r577x, p.adj = "bonf")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: famuss$ndrm.ch and famuss$actn3.r577x  
##  
##      CC      CT  
## CT 0.537 -  
## TT 0.034 0.433  
##  
## P value adjustment method: bonferroni
```

## The multiple testing problem



## TYPE I ERROR RATE FOR A SINGLE TEST

Hypothesis testing was originally intended for use in either controlled experiments or studies with a small number of comparisons, such as ANOVA.

Recall that making a Type I error (rejecting  $H_0$  when  $H_0$  is true) occurs with probability  $\alpha$ .

- Type I error rate is controlled by rejecting only when the  $p$ -value of a test is smaller than  $\alpha$ .
- $\alpha$  is typically kept low.
- With a single two-group comparison at  $\alpha = 0.05$ , there is a 5% chance of incorrectly identifying an association where none actually exists.

## WHAT ABOUT MANY TESTS?

What happens to Type I error when making several comparisons?

When conducting more than one  $t$ -test in an analysis. . .

- The significance level ( $\alpha$ ) used in each test controls the error rate for that test.
- The **experiment-wise error rate** is the chance that at least one test will incorrectly reject  $H_0$  when all tested null hypotheses are true.
- Controlling the experiment-wise error rate is one specific approach for controlling Type I error.

## SIMULATING ERROR RATE

Questions 1 - 3 in Lab 4 explore how experiment-wise error rate increases as the number of hypothesis tests increases.

## PROBABILITY OF EXPERIMENT-WISE ERROR

Suppose a scientist is using two  $t$ -tests to examine the possible association of each of two genes with a disease type. Assume the tests are independent and each are conducted at the  $\alpha = 0.05$  significance level.

Let  $A$  be the event of making a Type I error on the first test, and  $B$  be the event of making a Type I error on the second test, where  $P(A) = P(B) = 0.05$ .

The probability of making at least one error is equal to the complement of the event that a Type I error is not made with either test.

$$1 - [P(A^C)P(B^C)] = 1 - (1 - 0.05)^2 = 0.0975$$

Thus, when making two independent  $t$ -tests, there is about a 10% chance of making at least one Type I error; the experiment-wise error is 10%.

## PROBABILITY OF EXPERIMENT-WISE ERROR...

With 10 tests...

$$\text{experiment-wise error} = 1 - (1 - 0.05)^{10} = 0.401$$

With 25 tests...

$$\text{experiment-wise error} = 1 - (1 - 0.05)^{25} = 0.723$$

With 100 tests...

$$\text{experiment-wise error} = 1 - (1 - 0.05)^{100} = 0.994$$

With 100 independent tests, there is a 99.4% chance an investigator will make at least one Type I error!

# THE GOLUB LEUKEMIA DATA

Recall the Golub leukemia data from Unit 1.

- Expression level of 7,129 genes measured from children known to have either AML or ALL
- Goal: identify genes differentially expressed between AML vs. ALL

The analysis from Unit 1 used a “data-driven” approach:

- Calculate mean differences in expression levels (AML - ALL) for each gene
- Identify genes with mean differences that qualify as outliers, based on the distribution of differences in mean expression levels.

No claims made about whether observed differences more extreme than expected by chance alone.

## ANOTHER APPROACH TO THE GOLUB DATA

A hypothesis testing approach can be used to assess whether, for a particular gene  $i$ , there is significant evidence that the mean expression level among ALL patients is different from the mean expression level among AML patients.

- Questions 4 - 6 in Lab 4 walk through the details of this approach.

In this setting, it is unrealistic to assume that each test is independent.

- From a biological perspective, it is unlikely the expression of level of each gene is completely independent of the expression level of another.
- Question 7 in Lab 4 illustrates a simulation-based method for finding  $\alpha^*$  that maintains overall experiment-wise error at  $\alpha$  and does not assume independent tests.
- Stat 102 assignments will not test the technical details of conducting a simulation-based correction for correlated data, such as how to use `mvrnorm()`.