

# Biostatistics

## Chapter 6 Comparison of several groups: ANOVA

Jing Li

[jing.li@sjtu.edu.cn](mailto:jing.li@sjtu.edu.cn)

<http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/>

*Dept of Bioinformatics & Biostatistics, SJTU*



## Group Discussion(10 min)

- If we are going to have a cancer proteome project in large scale population, which experiment design will you recommend between paired and unpaired t-test ?  
Share your points.

# Review lecture 5

- t-test

- Paired t-test

$n \rightarrow$  the number of pairs,  $df=n-1$

$$t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}} \quad s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

- Independent samples

- Equal variance

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

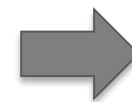
- Unequal variance

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad v = \frac{\left[ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left[ \frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)} \right]}$$

# Health Benefits of drinking tea



**Drinker** vs. **Non-drinker**  
(heart, skin, allergies, ... )



**Two Groups**

# How much we should take



0 cup, 1~4 cups, and more



**Several Groups**

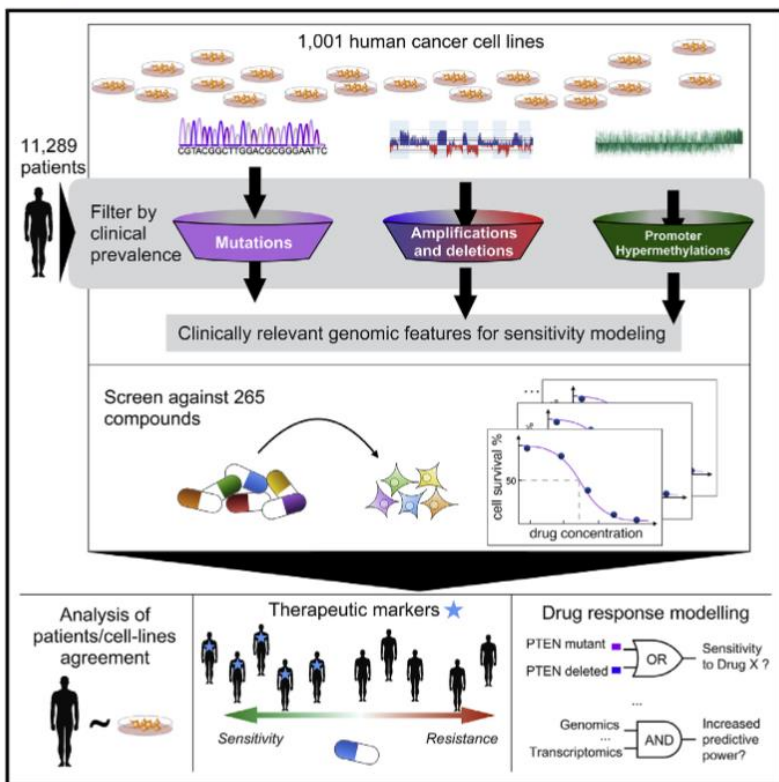
# Example: anti-cancer drugs

## Cell

### A Landscape of Pharmacogenomic Interactions in Cancer

Resource

- ANOVA Analysis Defines a Landscape of Pharmacogenomic Interactions



For pan-cancer **ANOVA**, the set of CFEs included 267 CGs, 407 RACs, and three gene fusions (BCR-ABL, EWSRI-FLI1, and EWSRI-X). Overall, for the 265 compounds, we identified 688 statistically significant interactions between unique CFE-drug pairs ( $p$  value  $< 10^{-3}$  at a false discovery rate [FDR]  $< .25$ ).

Effect ~

Different cancer cellline (CFE-drug pair)

Cell, 166, 740–754, 2016

CFEs: Cancer functional events

CGs: cancer genes

RACs: focal recurrently aberrant copy number segments

# One-way ANOVA

## Comparing more than two groups...

You have a group of individuals randomly split into smaller groups and completing different tasks.

For example, you might be studying the effects of tea on weight loss and form three groups: much tea, some tea and no tea.

# ANOVA (ANalysis Of VAriance)

- **Idea:** For two or more groups, test difference between means, for quantitative normally distributed variables.
- Just an extension of the t-test.



# One-Way Analysis of Variance (一元方差分析)

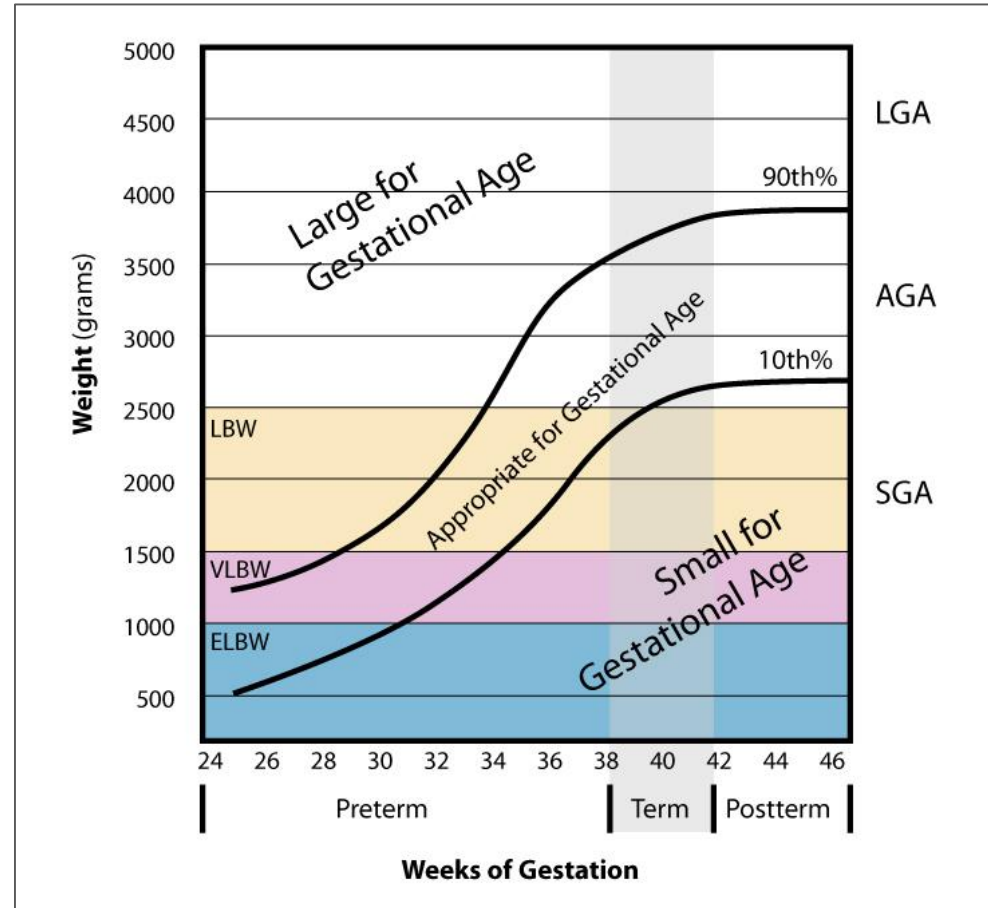
- Assumptions, same as t-test
  - Normally distributed outcome
  - Equal variances between the groups
  - Groups are independent

# Hypotheses of One-Way ANOVA

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

**$H_1$  : Not all of the population means are the same**

# Example 2- gestational age



## Example 2- gestational age

- A project at Shanghai and Women's Hospital has been investigating the differences of gestational age (孕龄) between singleton, twin, triplet, and more than 3 babies births, in order to limit poor outcomes from neonatal and fetal complications (新生儿并发症).
- The first step in this analysis was to determine **if there was a difference in the gestational age at which babies of these different types are delivered.**

## Example 2

- The null hypothesis is that all of the groups have the same gestational age on average
- We would like to test this null hypothesis at the 0.05 level
- You could compare each of the groups to each of the other groups which would be 6 pair wise comparisons at the 0.05 level, but what happens to the overall alpha level?
- Remember  $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$  so in this case  $\alpha = P(\text{one difference} \mid \text{all are equal})$

# Overall $\alpha$ level

- Now, if we completed each of the 6 pair wise tests at the 0.05 level and all of the tests were independent, we know that  $P(\text{fail to reject all 6 hypotheses} \mid H_0 \text{ is true}) = (1-0.05)^6 = 0.735$
- Therefore,  $P(\text{reject at least 1} \mid H_0 \text{ is true}) = 1-0.735 = 0.265 = \alpha$   
= type I error
- Notice that the type I error is now much worse than 0.05 and you can imagine how this would get even worse as the number of pair wise increases
- What can we do?
  - **ANOVA**

# The “F-test”

Is the difference in the means of the groups more than background noise (=variability within groups)?

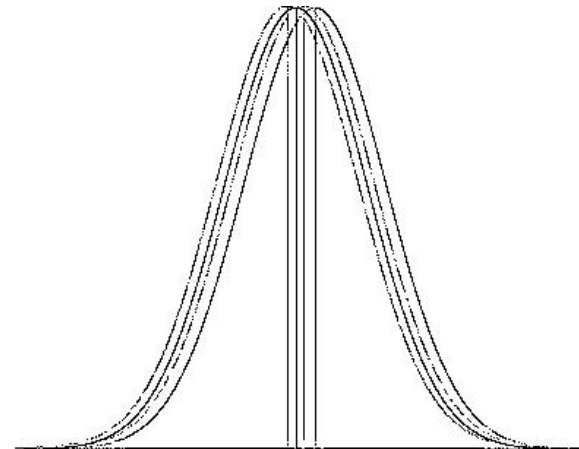
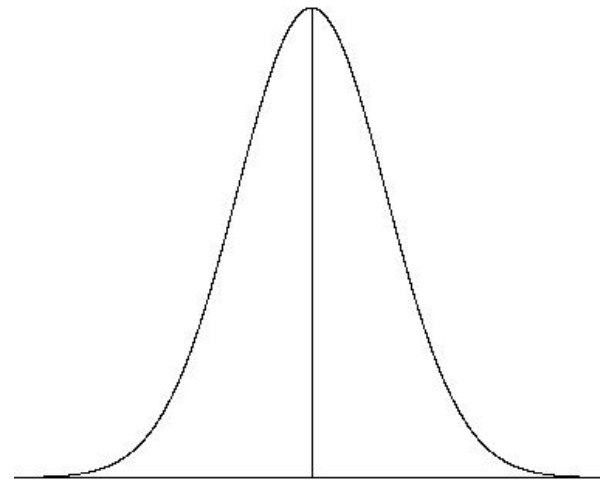
**Summarizes the mean differences between all groups at once.**

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}}$$

**Analogous to pooled variance from a t-test.**

# Picture I

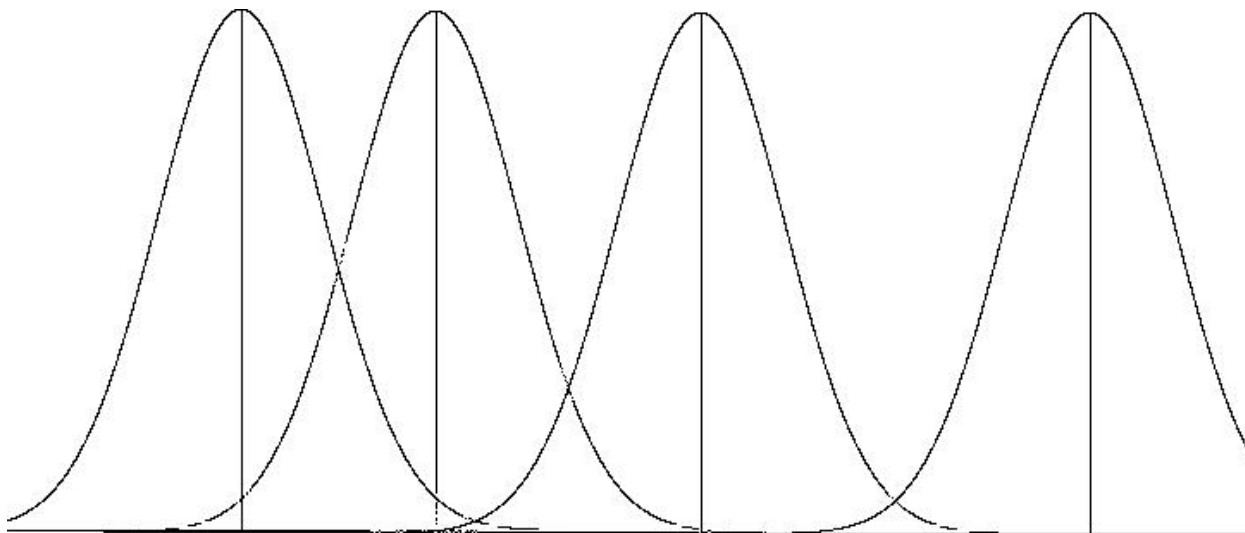
- Now, back to our example.
- Let's assume for the rest of the class that the normality.
- If all of the groups had the same means, the distributions for all of the populations would look exactly the same (overlaid graphs)





# Picture II

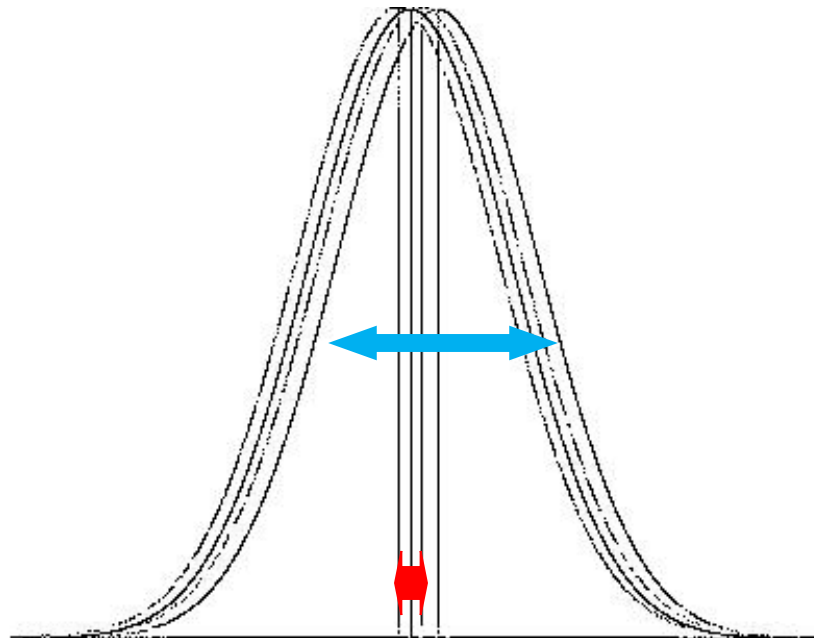
- Now, if the means of the populations were different, the picture would look like this. Notice that the variability between the groups is much greater than within a group



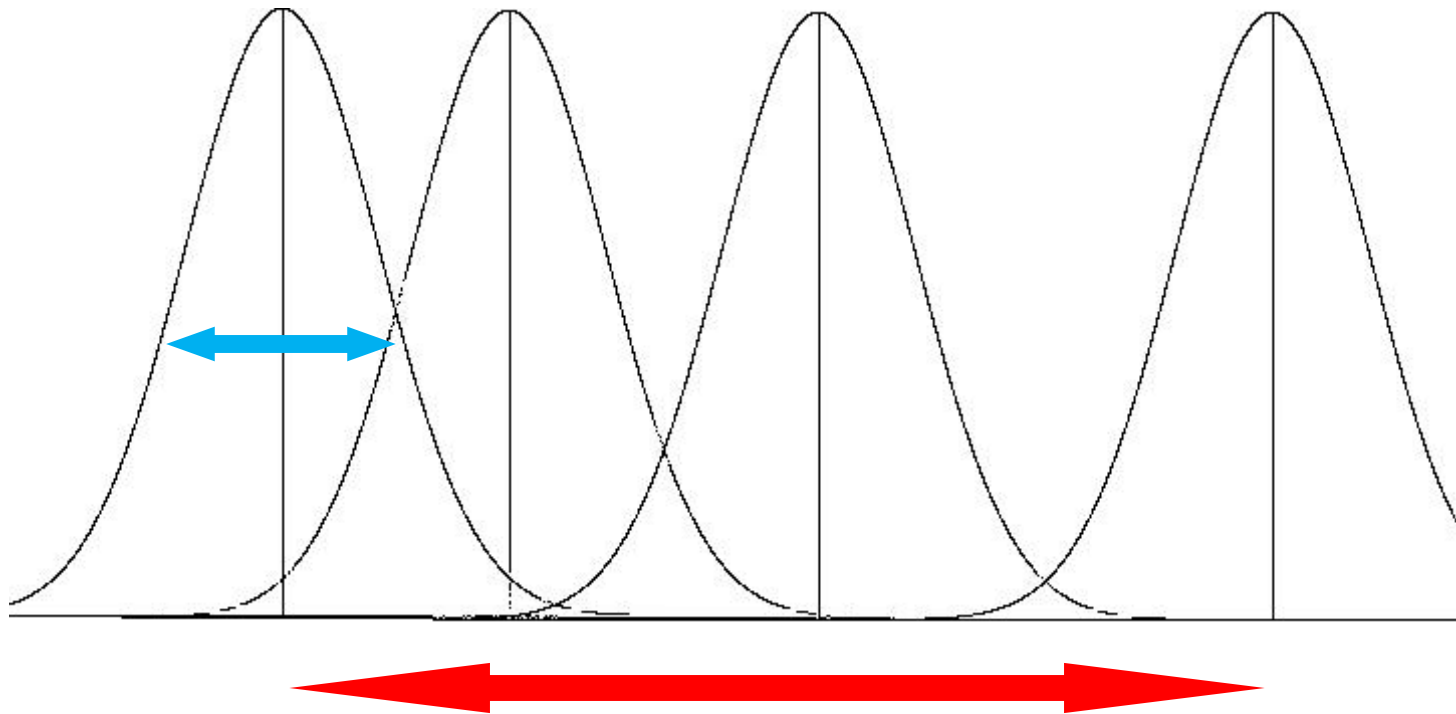
# Sources of variance

- When we take samples from each population, there will be two sources of variability
  - **Within group variability** : when we sample from a group there will be variability from person to person in the same group
    - We will always have this form of variability because it is sampling variability
  - **Between group variability** : the difference from group to group
    - This form of variability will only exist if the groups are different
    - If the between group variability is large, the means of the two groups is likely not the same

- We can use the two types of variability to determine if the means are likely different
- How can we do this?
- Look again at the picture
- Blue arrow: within group, red arrow: between group



- Blue arrow: within group, red arrow: between group
- Notice that when the distribution are separate, the between group variability is much greater than the within group



# F-statistic

- In the comparison of variance from the two sample t-test, we compared the ratio of the two variances to an F-distribution
- ANOVA uses a similar method of comparison to an F-distribution

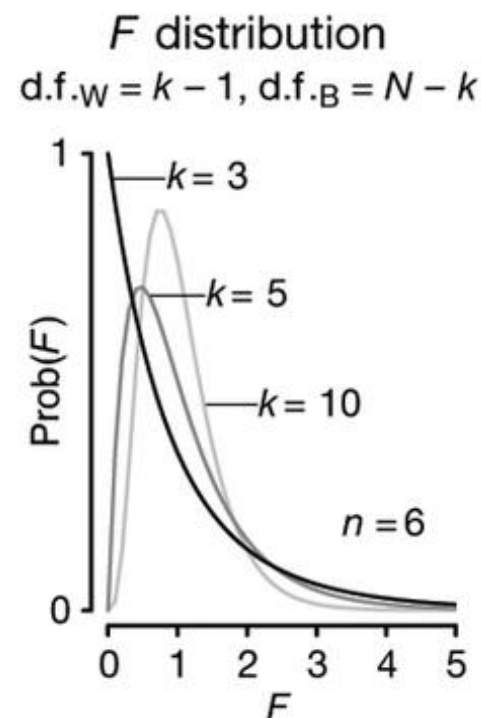
# The F-distribution

- The F-distribution is a continuous probability distribution that depends on two parameters  $d_1$  and  $d_2$  (numerator and denominator degrees of freedom, respectively):
- A random variate of the F-distribution with parameters  $d_1$  and  $d_2$  arises as the ratio of two appropriately scaled chi-squared variates:

Where 
$$X = \frac{U_1/d_1}{U_2/d_2}$$

•  $U_1$  and  $U_2$  have chi-squared distributions with  $d_1$  and  $d_2$  degrees of freedom respectively

•  $U_1$  and  $U_2$  are independent.



# The F-distribution

- A ratio of variances follows an F-distribution:

$$\frac{\sigma_{between}^2}{\sigma_{within}^2} \sim F_{n,m}$$

- The F-test tests the hypothesis that two variances are equal.
- F will be close to 1 if sample variances are equal.

# Notation

- First we will define

$x_{ij}$  = observation from student  $i$  from group  $j$

$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$  mean of group  $j$

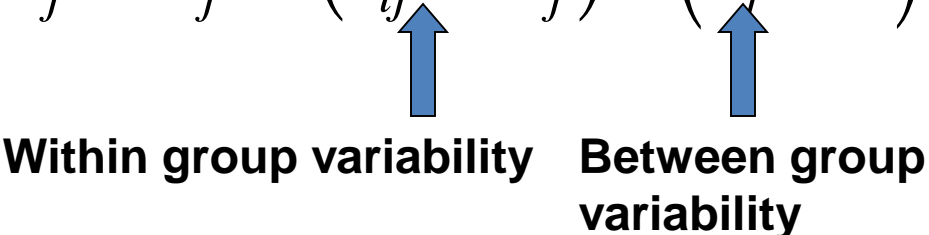
$\bar{x} = \frac{\sum_j n_j \bar{x}_j}{\sum_j n_j}$  grand mean over all of the groups

- How could we express the different forms of variability?



# Sources of variability

- The deviation of each observation from the grand mean can be broken into two pieces

$$x_{ij} - \bar{x} = x_{ij} - \bar{x} + \bar{x}_j - \bar{x}_j = \left( x_{ij} - \bar{x}_j \right) + \left( \bar{x}_j - \bar{x} \right)$$


**Within group variability**
**Between group variability**

- Like the calculation of the variance, we are interested in the square of the deviation
- What does the squared deviation look like?

- The equation of the squared deviation summed over group

$$\begin{aligned}
 \sum_{j=1}^4 \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 &= \sum_{j=1}^4 \sum_{i=1}^{n_j} \left( (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}) \right)^2 \\
 &= \sum_{j=1}^4 \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^4 \sum_{i=1}^{n_j} 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) \\
 &\quad + \sum_{j=1}^4 \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2
 \end{aligned}$$

- Now the second term in the final equation equals 0 (show on your own if you are interested)
- The final squared deviation simplifies to

$$\sum_{j=1}^4 \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{j=1}^4 \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{j=1}^4 \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2$$



**Total sum of squares  
(SS<sub>T</sub>)**

总变异



**Within group sum of  
squares (SS<sub>w</sub>)**

组内变异



**Between group sum of  
squares (SS<sub>B</sub>)**

组间变异

- As we discussed earlier, we are going to compare the two errors to determine if the group means are equal

# The within group variability

- It can be written in terms of the individual group standard deviations,  $s_j$ , which are often given summary statistics

$$SS_W = \sum_{j=1}^4 \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^4 (n_j - 1) s_j^2$$

- Remember that we are under the assumption that all of the  $s_j$ 's are equal, so we can pool the estimates of the  $s_j$ 's (just like the pooled variance estimate of two sample t-test). The result is called the within group mean square error, which is the overall estimate of the within group variance

$$MS_W = \frac{\sum_{j=1}^4 (n_j - 1) s_j^2}{\sum_{j=1}^4 (n_j - 1)}$$

- Note the denominator is the total sample size minus the number of groups

# The between group variability

- The between group variability can be broken into pieces from the summary statistics as well

$$SS_B = \sum_{j=1}^4 \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^4 n_j (\bar{x}_j - \bar{x})^2$$

- The between group mean square error can be written as

$$MS_B = \frac{\sum_{j=1}^4 n_j (\bar{x}_j - \bar{x})^2}{4 - 1}$$

- The denominator of the  $MS_B$  is the number of groups minus 1 because we are considering the group means as the observations and the grand mean as the mean

# F-statistic

- Now that we have estimates of the between group and within group variation, we can use our F-statistic

$$F_{k-1, n-k} = \frac{MS_B}{MS_W} = \frac{SS_B / (k - 1)}{SS_W / (n - k)}$$

where  $k$  is the number of groups and  $n$  is the total sample size

- This test statistic is compared to an F-statistic with  $k-1$  and  $n-k$  degrees of freedom

# ANOVA Table

- To complete the analysis, we need to calculate the  $SS$ 's,  $MS$ 's and the F-statistic
- A specific display of this data is often used called the ANOVA table
- Standard software may provide results in this form

Source of variation	SS	df	MS	F	p-value
Between	$SS_B$	$k-1$	$MS_B$	$MS_B/MS_W$	
Within	$SS_W$	$n-k$	$MS_W$		
Total	$SS_T$				

## Example \* \* \*

- Let's perform an ANOVA test for the test score data
- Here are the summary statistics

	Singleton	Twin	Triplet	> 3 babies
Mean	39.3	38.6	37.2	36.2
Standard deviation	1.67	1.93	1.55	1.17
Sample size	30	16	10	6



# Steps for the hypothesis test

- 1) State null and alternative hypotheses
  - $H_0: m_1 = m_2 = \dots = m_n$
  - $H_A$ : at least one mean is different
- 2) Specify  $\alpha$  level
- 3) Calculate test statistic: See ANOVA table
- 4) Calculate p-value: See ANOVA table
- 5) Reject null or not reject null
- 6) Conclusions: There is at least one difference we reject null

# ANOVA table

- Here is the ANOVA table for this data

Source of variation	SS	df	MS	F	p-value
Between	94.629	3	31.542	11.08	<0.001
Within	165.038	58	2.845		
Total					

- **Conclusions:** We conclude that at least one group is significantly different from the others.

# Example 3: Vitamins ~ height

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inch	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

# Example 3

**Step 1:** calculate the sum of squares between groups:

Mean for group 1 = 62.0

Mean for group 2 = 59.7

Mean for group 3 = 56.3

Mean for group 4 = 61.4

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

Grand mean= 59.85

$$SSB = [(62-59.85)^2 + (59.7-59.85)^2 + (56.3-59.85)^2 + (61.4-59.85)^2] \times (n \text{ per group}) = 19.65 \times 10 = 196.5$$

# Example 3

**Step 2:** calculate the sum of squares within groups:

$(60-62)^2 + (67-62)^2 + (42-62)^2 + (67-62)^2 + (56-62)^2 +$   
 $(62-62)^2 + (64-62)^2 + (59-62)^2 + (72-62)^2 + (71-62)^2 +$   
 $(50-59.7)^2 + (52-59.7)^2 + (43-59.7)^2 + (67-59.7)^2 + (67-59.7)^2 + (69-59.7)^2$   
 $^2 \dots + \dots$  (sum of 40 squared deviations) = **2060.6**

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

## Step 3: Fill in the ANOVA table

<u>Source of variation</u>	<u>d.f.</u>	<u>Sum of squares</u>	<u>Mean Sum of Squares</u>	<u>F-statistic</u>	<u>p-value</u>
Between	3	196.5	65.5	1.14	.344
Within	36	2060.6	57.2	-	-
Total	39	2257.1	-	-	-

### INTERPRETATION of ANOVA:

How much of the variance in height is explained by treatment group?

## Step 3: Fill in the ANOVA table

<u>Source of variation</u>	<u>d.f.</u>	<u>Sum of squares</u>	<u>Mean Sum of Squares</u>	<u>F-statistic</u>	<u>p-value</u>
Between	3	196.5	65.5	1.14	.344
Within	36	2060.6	57.2	-	-
Total	39	2257.1	-	-	-

### INTERPRETATION of ANOVA:

How much of the variance in height is explained by treatment group?

$$R^2 = \text{“Coefficient of Determination”} = \text{SSB/TSS} = 196.5/2275.1 = 9\%$$

# Coefficient of Determination

$$R^2 = \frac{SSB}{SSB + SSE} = \frac{SSB}{SST}$$

The amount of variation in the outcome variable (dependent variable) that is explained by the predictor (independent variable).



# ANOVA example

Table Mean micronutrient intake from the school lunch by school

		S1 <sup>a</sup> , n=25	S2 <sup>b</sup> , n=25	S3 <sup>c</sup> , n=25	P-value <sup>d</sup>
Calcium (mg)	Mean	117.8	158.7	206.5	0.000
	SD <sup>e</sup>	62.4	70.5	86.2	
Iron (mg)	Mean	2.0	2.0	2.0	0.854
	SD	0.6	0.6	0.6	
Folate (µg)	Mean	26.6	38.7	42.6	0.000
	SD	13.1	14.5	15.1	
Zinc (mg)	Mean	1.9	1.5	1.3	0.055
	SD	1.0	1.2	0.4	

<sup>a</sup> School 1 (most deprived; 40% subsidized lunches).

<sup>b</sup> School 2 (medium deprived; <10% subsidized).

<sup>c</sup> School 3 (least deprived; no subsidization, private school).

<sup>d</sup> ANOVA; significant differences are highlighted in bold ( $P < 0.05$ ).

FROM: Gould R, Russell J, Barker ME. School lunch menus and 11 to 12 year old children's food choice in three secondary schools in England-are the nutritional standards being met? *Appetite*. 2006 Jan;46(1):86-92.

# Answer

**Step 1:** calculate the sum of squares between groups:

Mean for School 1 = 117.8

Mean for School 2 = 158.7

Mean for School 3 = 206.5

Grand mean: 161

$$\text{SSB} = [(117.8-161)^2 + (158.7-161)^2 + (206.5-161)^2] \times 25 \text{ per group} = 98,113$$

# Answer

**Step 2:** calculate the sum of squares within groups:

S.D. for S1 = 62.4

S.D. for S2 = 70.5

S.D. for S3 = 86.2

Therefore, sum of squares within is:

$$(24)[62.4^2 + 70.5^2 + 86.2^2] = 391,066$$

# Answer

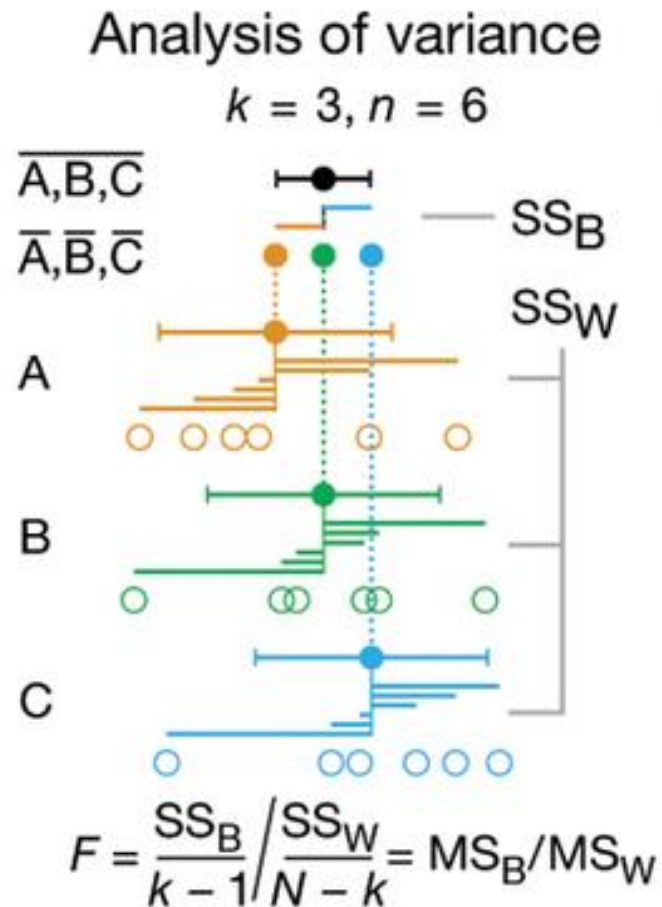
## Step 3: Fill in your ANOVA table

<u>Source of variation</u>	<u>d.f.</u>	<u>Sum of squares</u>	<u>Mean Sum of Squares</u>	<u>F-statistic</u>	<u>p-value</u>
Between	2	98,113	49.056	9	<.05
Within	72	391,066	5.431		
Total	74	489,179			

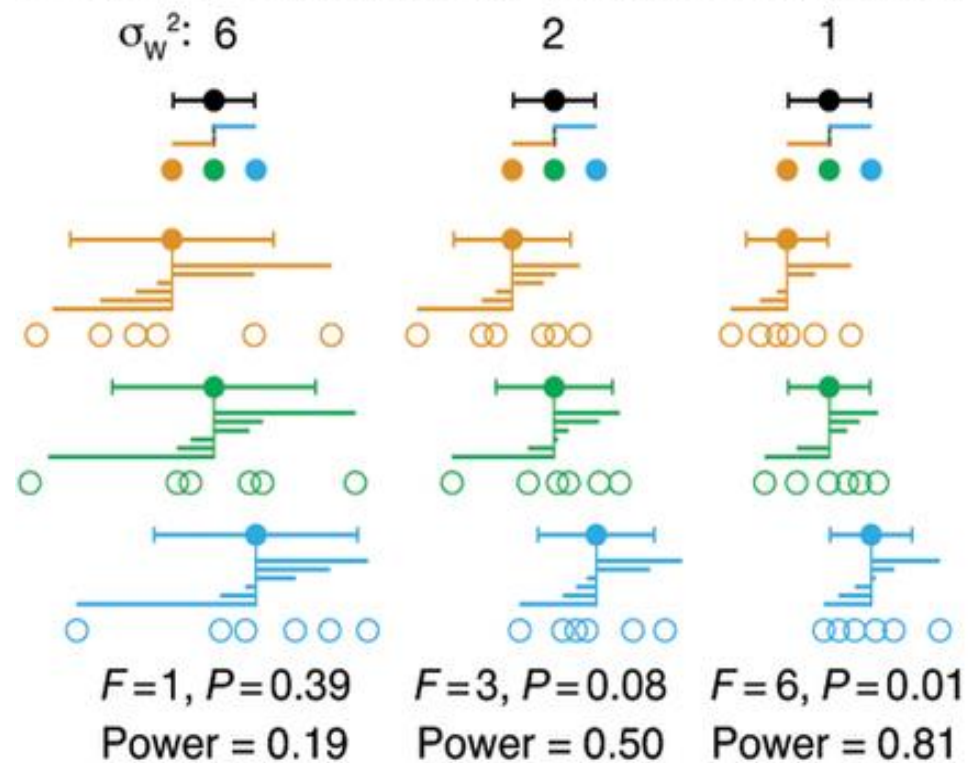
$$**R^2 = 98113 / 489179 = 20\%$$

School explains 20% of the variance in lunchtime calcium intake in these kids.

# ANOVA summary



## Impact of within-group variance on power



# Notes

- Remember the assumption of equal variance across groups is required
- We were able to conclude that one of the means is different, but we do not know which of the means is different. ANOVA is often considered a first step because it gives evidence if there are any differences and further testing is required to determine which are the significant differences
- We must do pair wise comparisons to determine which specific means are different, but we must still take into account the problem with multiple comparisons?
- How could we do this?

# ANOVA (ANalysis Of VAriance)

- ANOVA: just an extension of the t-test.

**Q: can we do multiple t-test here ?**

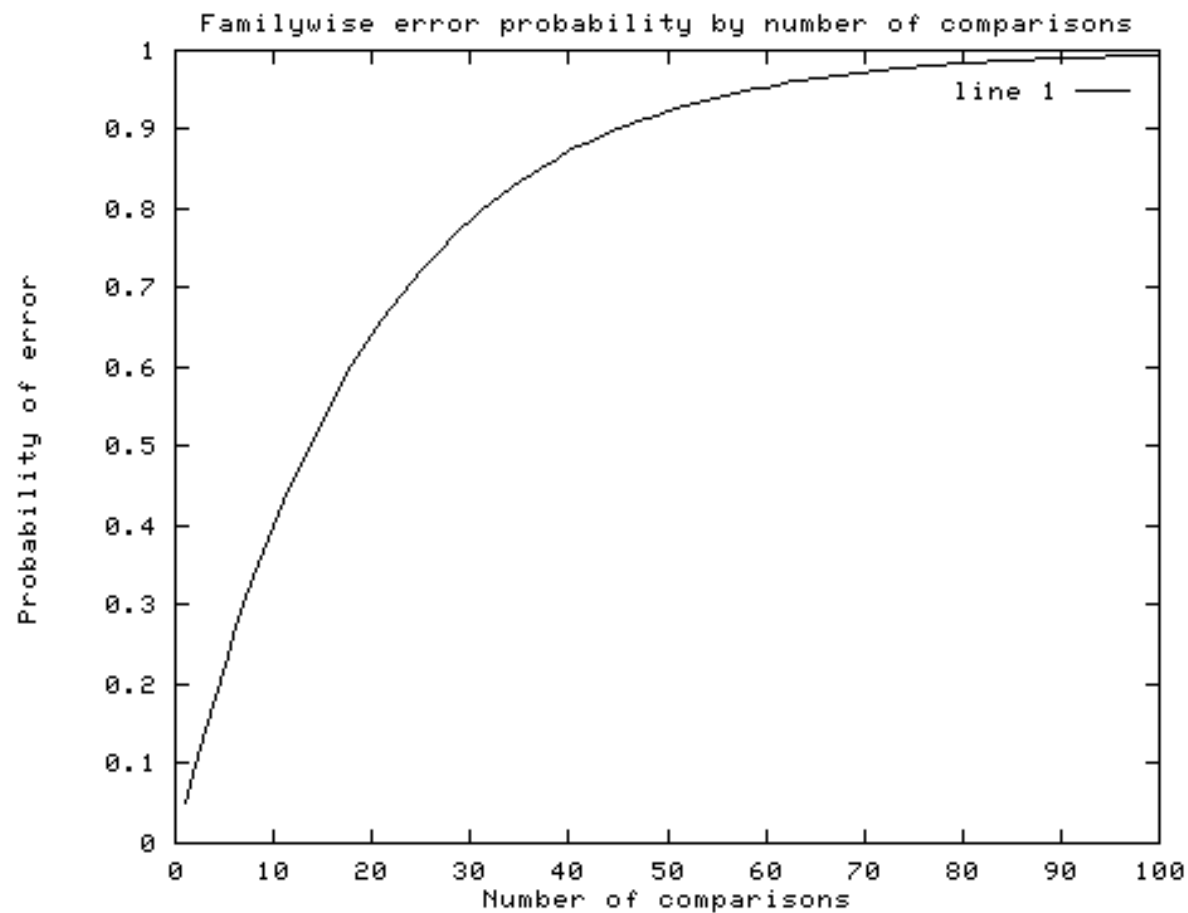
***Why not just do 3 pairwise ttests?***

## Question: *Why not just do 3 pairwise ttests?*

- Answer: because, at an error rate of 5% each test, this means you have an overall chance of up to  $1 - (.95)^3 = 14\%$  of making a type-I error (if all 3 comparisons were independent)
- If you wanted to compare 6 groups, you'd have to do  ${}_6C_2 = 15$  pairwise ttests; which would give you a high chance of finding something significant just by chance (if all tests were independent with a type-I error rate of 5% each); probability of at least one type-I error =  $1 - (.95)^{15} = 54\%$ .



# Recall: Multiple comparisons



# Correction for multiple comparisons

## How to correct for multiple comparisons *post-hoc*...

- Bonferroni correction (adjusts  $p$  by most conservative amount; assuming all tests independent, divide  $p$  by the number of tests)
- BH
- Fisher's LSD
- ... ..

# Procedures for Post Hoc Comparisons

If your ANOVA test identifies a difference between group means, then you must identify which of your  $k$  groups differ.

If you did not specify the comparisons of interest (“contrasts”) ahead of time, then you have to pay a price for making all pairwise comparisons to keep overall type-I error rate to  $\alpha$ .

Alternately, run a limited number of planned comparisons (making only those comparisons that are most important to your research question). (Limits the number of tests you make).

# I. Bonferroni

For example, to make a Bonferroni correction, divide your desired alpha cut-off level (usually .05) by the number of comparisons you are making. Assumes complete independence between comparisons, which is way too conservative.

Obtained P-value	Original Alpha	# tests	New Alpha	Significant?
.001	.05	5	0.01	Yes
.011	.05	4	.01	No
.019	.05	3	.01	No
.032	.05	2	.01	No
.048	.05	1	.01	No

## 2. Benjamini–Hochberg (FDR)

- Put the individual P values in order, from smallest to largest. The smallest P value has a rank of  $i=1$ , then next smallest has  $i=2$ , etc. The adjusted P value (FDR) for a test is the raw P value times  $m/i$ , whichever is smaller ( **$m$  is the number of tests and  $i$  is the rank of each test**). If the adjusted P value is smaller than the false discovery rate you choose, the test is significant

## 3. Fisher's LSD

- We replace  $s_p^2$  in our 2-sample t-test formula with  $MS_{error}$  and we get:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_{error} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- We then test this using a critical t, using our t-table and  $df_{error}$  as our df
- You can use either a one-tailed or two-tailed test, depending on whether or not you think one mean is higher or lower (one-tailed) or possibly either (two-tailed) than the other

# Discussion

- Can we do multiple t-tests for comparison of more than two groups.
- Can we do ANOVA for comparison of two groups

# Practice problem

- Your patient is taking one of the standard drugs that was shown to be statistically less effective in minimizing motion sickness (i.e., **significant p-value** for the comparison with the experimental drug). Assuming that none of these drugs have side effects but that the experimental drug is slightly more costly than your patient's current drug-of-choice, what (if any) other information would you want to know before you start recommending that patients switch to the new drug?