方差分析

ANOVA: ANalysis Of VAriance

吴志勇

清华大学深圳研究生院

ANOVA Examples

- 三种不同的教学方法对于学生的成绩是否有影响?
- 学校中各年级的同学智商是否有区别?
- 不同洗涤剂和不同水温的去污能力是否不同?

Head Movement Example

Compare the naturalness of head movement on a talking avatar generated by different methods

Neutral Talking Head
Expressive Head Movement
Random Head Movement

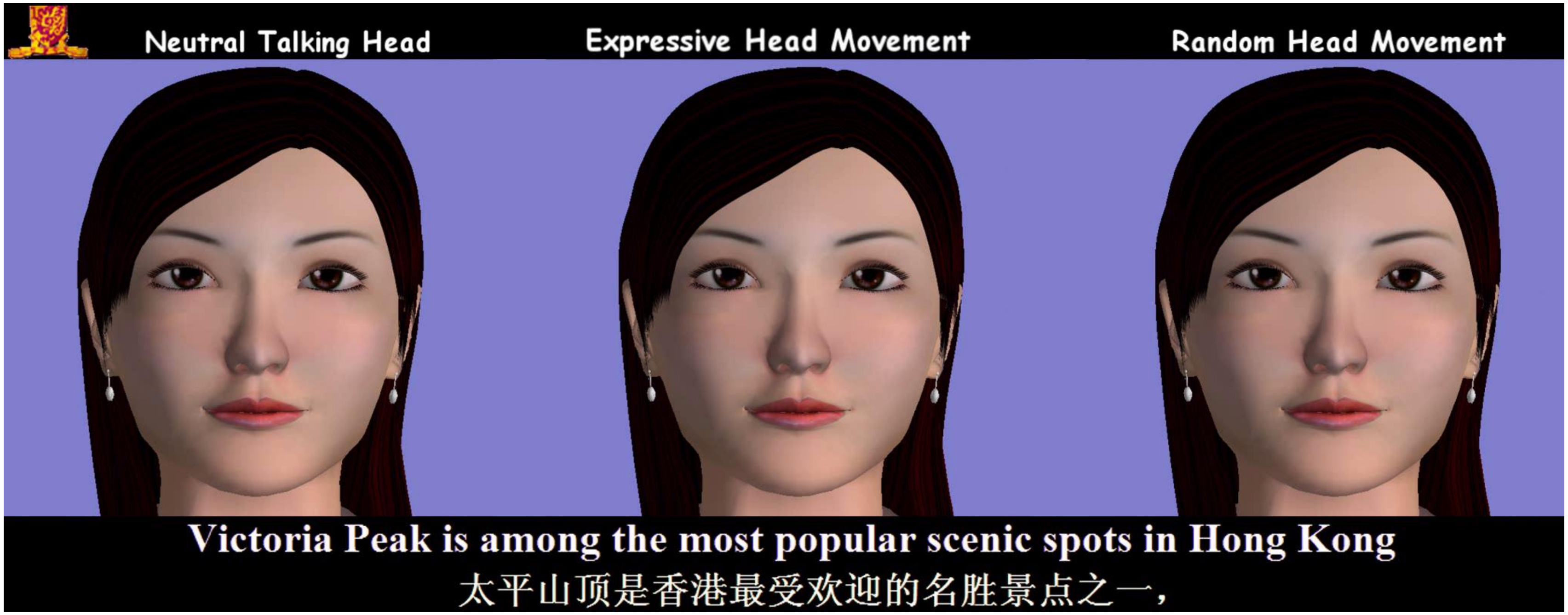
Victoria Peak is among the most popular scenic spots in Hong Kong
太平山顶是香港最受欢迎的名胜景点之一，

Head Movement

Head Movement Example

Compare the naturalness of head movement on a talking avatar generated by different methods

- Session I: without any head movements
- Session II: with random head movements
- Session III: with the proposed expressive head movements



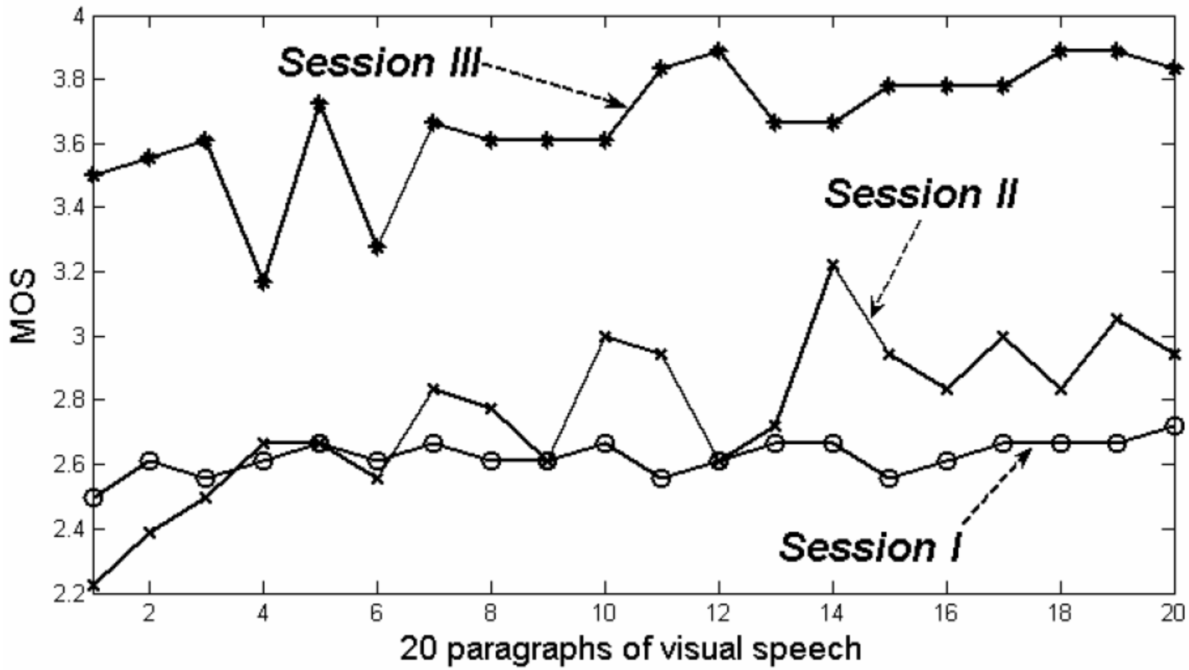
Victoria Peak is among the most popular scenic spots in Hong Kong
太平山顶是香港最受欢迎的名胜景点之一，

Head Movement Example

Ask subjects to score the naturalness of head movements on a five-point Likert scale:

- (5) expressive (4) natural (3) acceptable (2) unnatural (1) erratic

Calculate the mean opinion score (MOS):



avg. MOS = 3.7

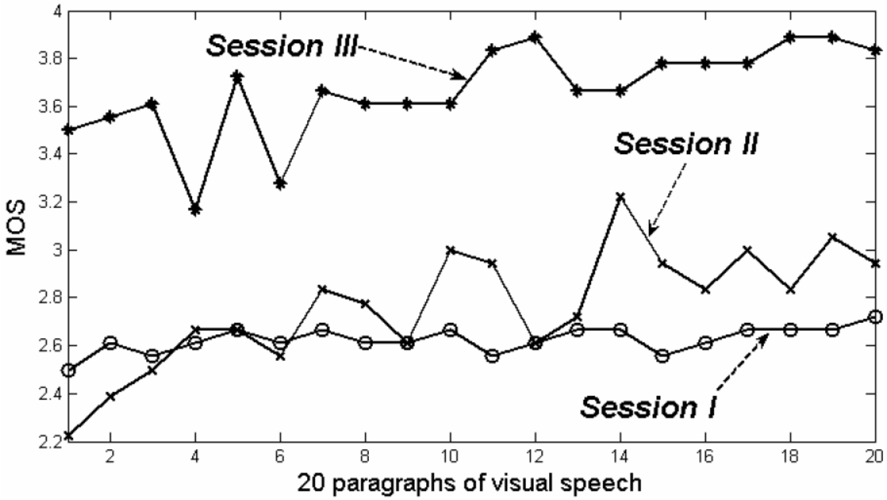
avg. MOS = 2.8

avg. MOS = 2.6

Head Movement Example

Do ANOVA test

- The average MOS are 2.6 (session I), 2.8 (session II) and 3.7 (session III).
- A one-way ANOVA test revealed a significant effect of head movement [F(2,57)=192.97, *p*=0].
- *Post-hoc* analysis (Tukey HSD) showed the result from session III is significantly better than the results from session I and session II.



Another Example

Subjects: 25 patients with blisters

Treatments: Treatment A, Treatment B, Placebo

Measurement: # of days until blisters heal

Data [and means]:

A: 5,6,6,7,7,8,9,10	[7.25]
B: 7,7,8,9,9,10,10,11	[8.875]
P: 7,9,9,10,10,10,11,12,13	[10.11]

Are these differences significant?

What Does ANOVA Do?

At its simplest (there are extensions), ANOVA tests the following hypotheses:

H_0: The means of all groups are equal.

$H_0: \mu_1 = \mu_2 = \cdots$

H_1: Not all the means are equal

- doesn’t say how or which ones differ.
- Can follow up with “multiple comparisons”

Note: we usually refer to the sub-populations as “groups” when doing ANOVA.

The ANOVA Definition

Analysis of variance (ANOVA) is also called “F Test”, which provides a statistical test of whether or not the means of several groups are equal.

ANOVA is useful for comparing (testing) three or more means for statistical significance.

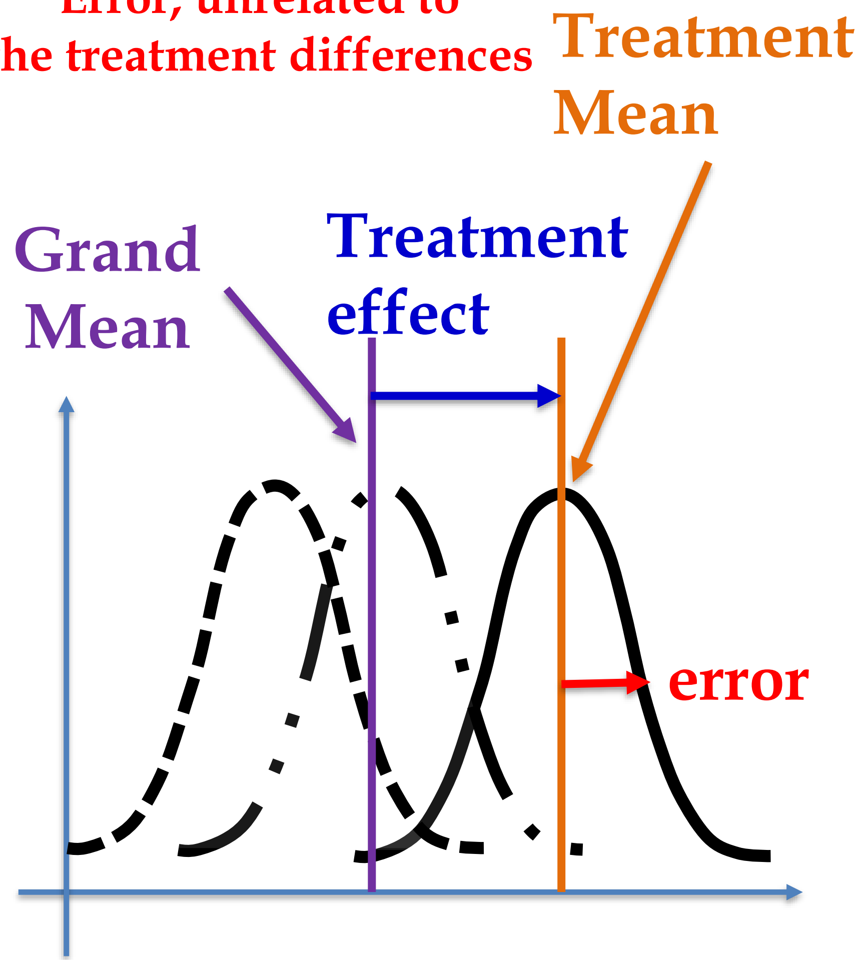
方差分析又称F检验, 用于三个及以上样本均值差别的显著性检验.

The ANOVA Model

$$X_{Ai} = \bar{\mu}_G + \underbrace{(\bar{\mu}_A - \bar{\mu}_G)}_{\text{The treatment effect}} + \underbrace{(X_{Ai} - \bar{\mu}_A)}_{\text{Error, unrelated to the treatment differences}}$$

The Grand Mean

- $\bar{\mu}_G$: The Grand Mean taken over all observations
- $\bar{\mu}_A$: The mean of Group A treatment
- X_{Ai}: The observation from the i$^{\text{th}}$ person in Group A.



Treatment Mean

Grand Mean

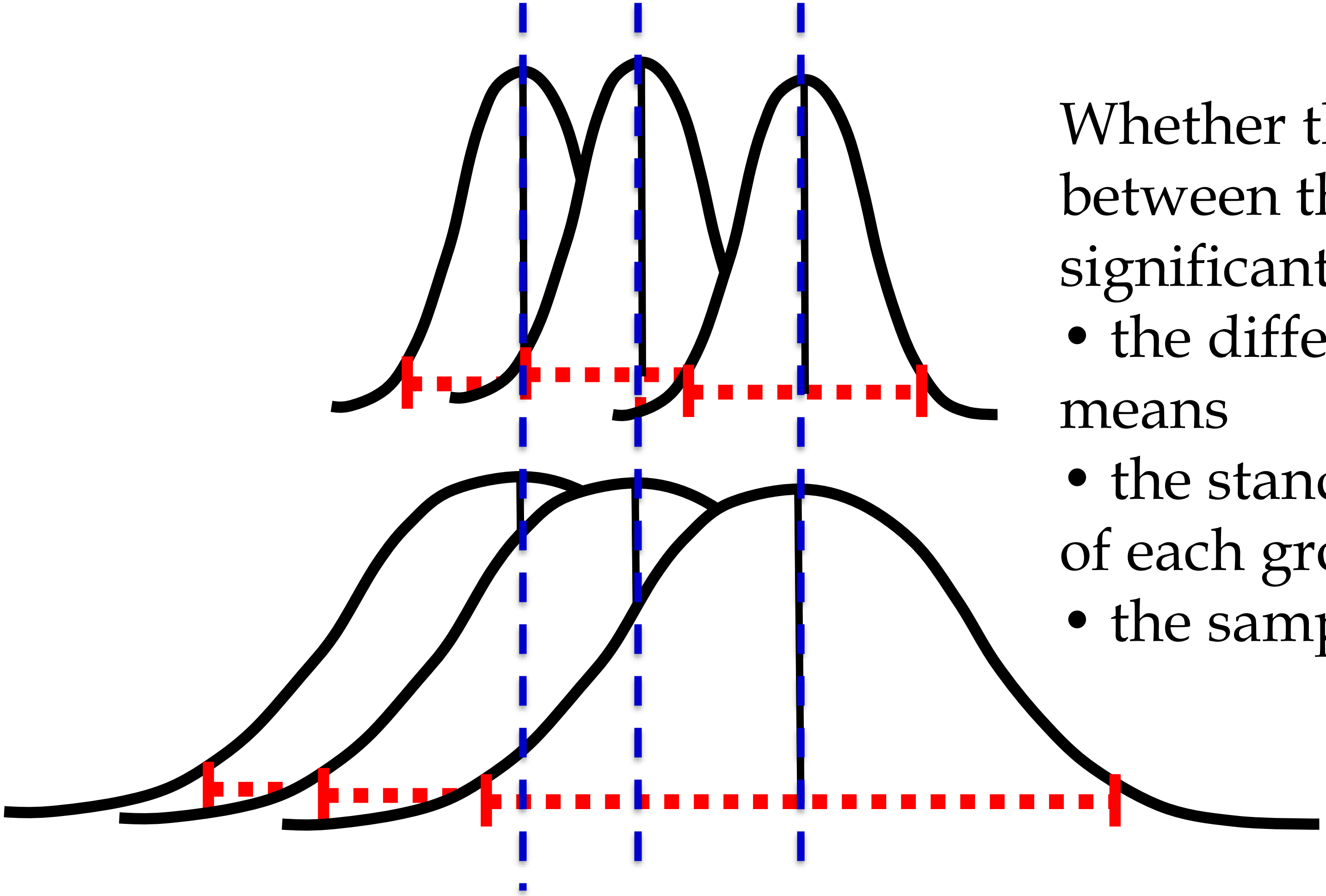
Treatment effect

error

Compare the observed variances among group means to what we would expect to get by chance.

Analysis of Variance

Comparing two different scenarios



Whether the differences between the groups are significant depends on

- the difference in the means
- the standard deviations of each group
- the sample sizes

Assumptions

- The data are randomly sampled
- The variances of each group are assumed equal
  - Equal variances is called Homoscedastic (Same + Scatter)
  - rule of thumb: ratio of largest to smallest group std. dev. must be less than 2:1
- The residuals are normally distributed (not skewed or partial)

Standard Deviation Check

Treatment	n	Sample Mean	Sample Std Dev	Sample Variance
A	8	7.25	1.6690	2.7857
B	8	8.875	1.4577	2.1250
P	9	10.111	1.7638	3.1111

Treatment A: 5,6,6,7,7,8,9,10, n_A=8 data points

$$\mu_A = \sum_{i=1}^{n_A} \frac{x_{Ai}}{n_A} = \frac{5+6+6+7+7+8+9+10}{8} = 7.25$$

$$var_A = \frac{\sum_{i=1}^{8}(x_{Ai} - \mu_A)^2}{n_A - 1} = 2.7857, s_A = \sqrt{var_A} = 1.6690$$

Compare largest and smallest standard deviations:

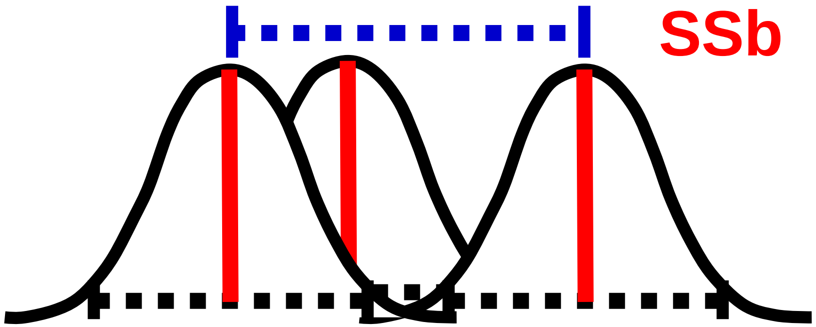
- largest: 1.764
- smallest: 1.458
- 1.458 x 2 = 2.916 > 1.764

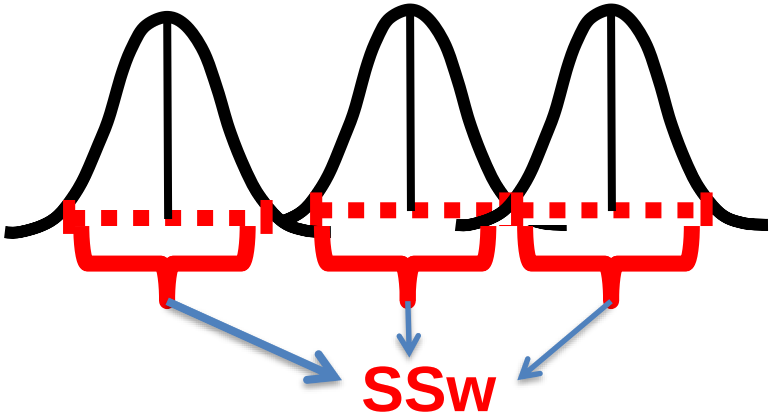
Variation

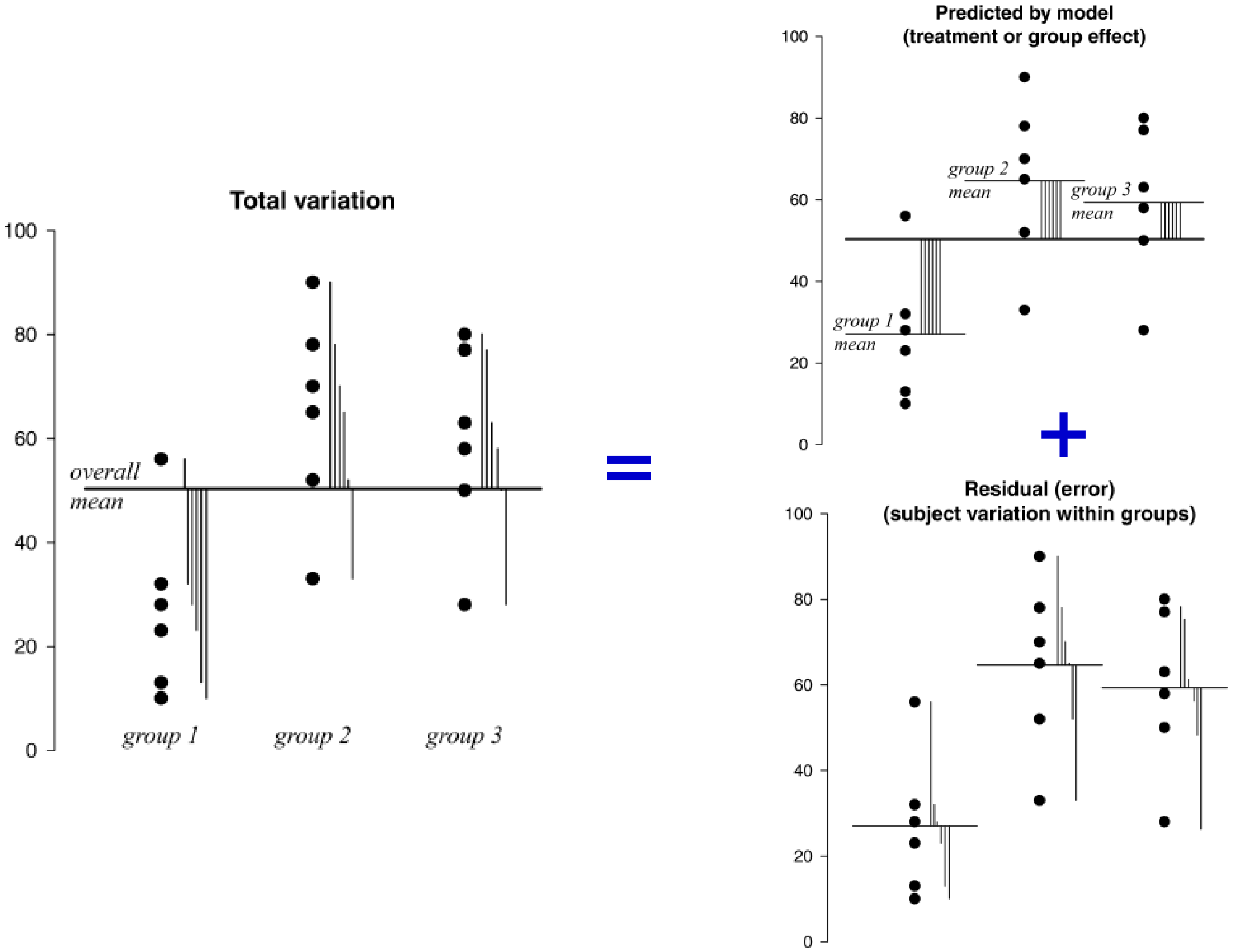
Variation is the sum of squares of the deviations between a value and the mean of the value

There are two types of variation:

- Within-group variation: variation of data within each group
- Between-group variation: variation of the sample means between different groups
- Total variation is the sum of these two.







Total variation

Predicted by model (treatment or group effect)

Residual (error) (subject variation within groups)

F Statistic

The ANOVA F-statistic is a ratio of the between-group and the within-group variation:

$$F = \frac{Between}{Within}$$

Observed variances among group means. $\approx \sigma^2$ only when H_0 is true

(avg) variances among data within each group. $\approx \sigma^2$ under both H_0 and H_1.

- If H_0 is true, F should be close to 1.
- A large F is evidence *against* H_0 (i.e., evidence shows that different groups have different means), since it indicates that there is more difference between groups than within groups.

One–Way ANOVA

How to do the calculation?

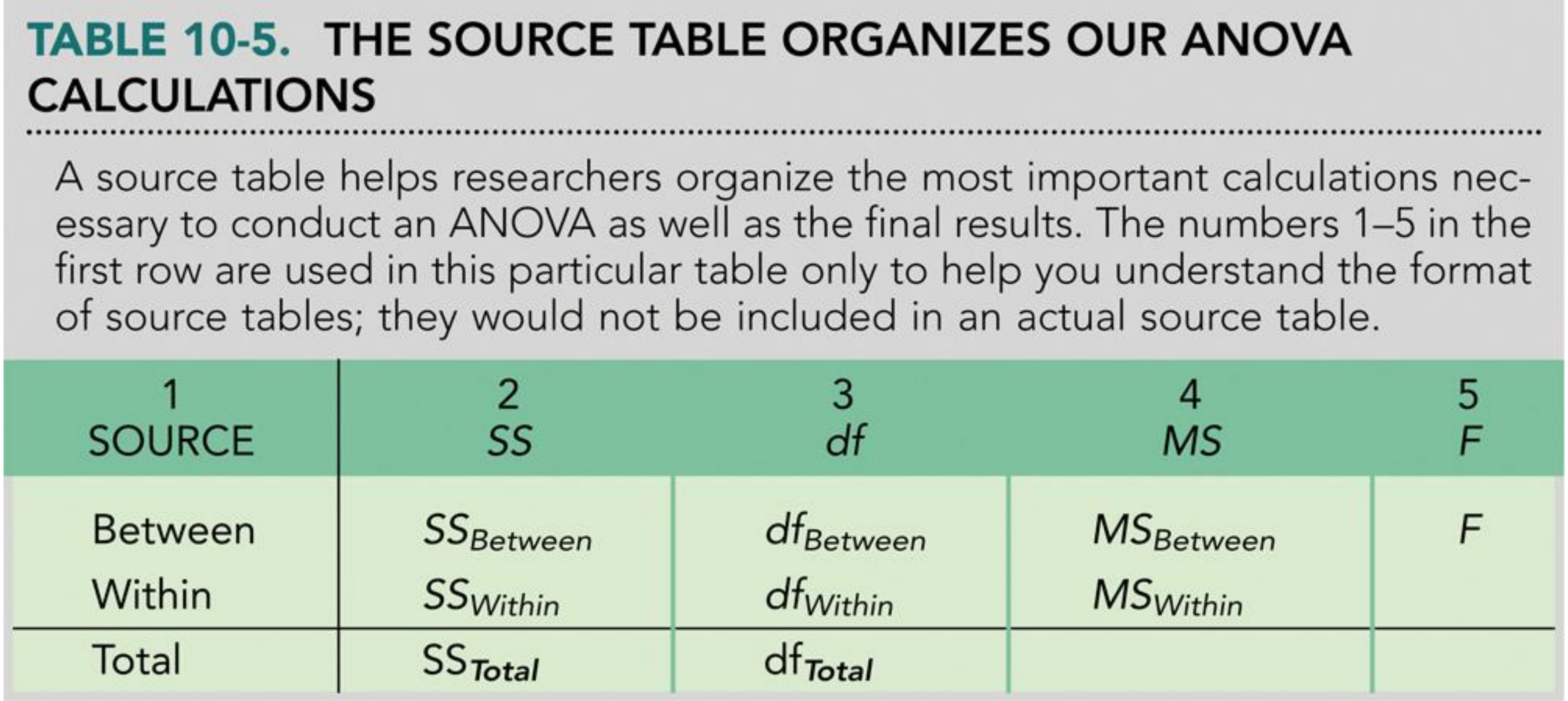


TABLE 10-5. THE SOURCE TABLE ORGANIZES OUR ANOVA CALCULATIONS

A source table helps researchers organize the most important calculations necessary to conduct an ANOVA as well as the final results. The numbers 1–5 in the first row are used in this particular table only to help you understand the format of source tables; they would not be included in an actual source table.

1	2	3	4	5
SOURCE	*SS*	*df*	*MS*	*F*
Between	$SS_{Between}$	$df_{Between}$	$MS_{Between}$	F
Within	SS_{Within}	df_{Within}	MS_{Within}	
Total	SS_{Total}	df_{Total}		

One–Way ANOVA

Here is the basic one-way ANOVA table

Source	**SS**	**df**	**MS**	**F**	**p**
Between					
Within				▨	▨
Total			▨	▨	▨

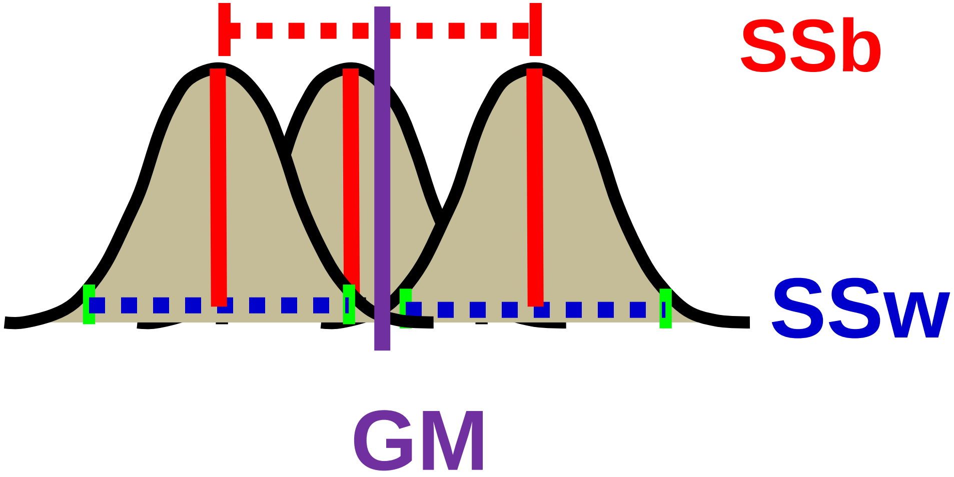
Grand Mean

- The grand mean is the average of all the values, or equivalently, a weighted average of the individual group's sample means

$$\bar{\mu} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}}{n}$$

In our example,

$$\bar{\mu} = \frac{7.25 * 8 + 8.875 * 8 + 10.111 * 9}{8 + 8 + 9}$$
$$= 8.8$$



Variation/Sum of Squares (SS)

- Between-Group Variation (SS$_b$)

Sum over all groups

Grand mean

$$SS_b = \sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2$$

# of data points in group i

Sample mean within group i

- Within-Group Variation (SSw)

Sum over all groups

$$SS_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \mu_i\right)^2$$

Sum over all data points in group i

Sample mean within group i

Our Example:

Source	SS	df	MS	F	p
Between	34.7361				
Within	59.2639				
Total	94.0000				

Degree of Freedom

Degree of freedom: # of variables that can vary before the rest of the values are predetermined

Example:

- We know the total of six numbers is 240.
- Five of the six numbers could be anything, but once the first five are known, the last one is fixed so the sum is 240.
- The df would be 6-1=5

Degree of Freedom

For group i with n_i data points

- $df_i = n_i - 1$

Consider the entire data set with k groups and $n = n_1 + n_2 + \cdots + n_k$ data points in total

- $df_b = k - 1$
- $df_w = df_1 + \cdots + df_k = n - k$

Our Example:

Given k=3 groups with $n_A = 8, n_B = 8, n_P = 9$:

Source	**SS**	**df**	**MS**	**F**	**p**
Between	34.7361	2			
Within	59.2639	22			
Total	94.0000	24			

Mean of Squares (MS) = SS/df

- In our example:

$$MS_b = \frac{SS_b}{df_b} = 17.3681, MS_w = \frac{SS_w}{db_w} = 2.6938$$

Source	**SS**	**df**	**MS**	**F**	**p**
Between	34.7361	2	17.3681		
Within	59.2639	22	2.6938		
Total	94.0000	24			

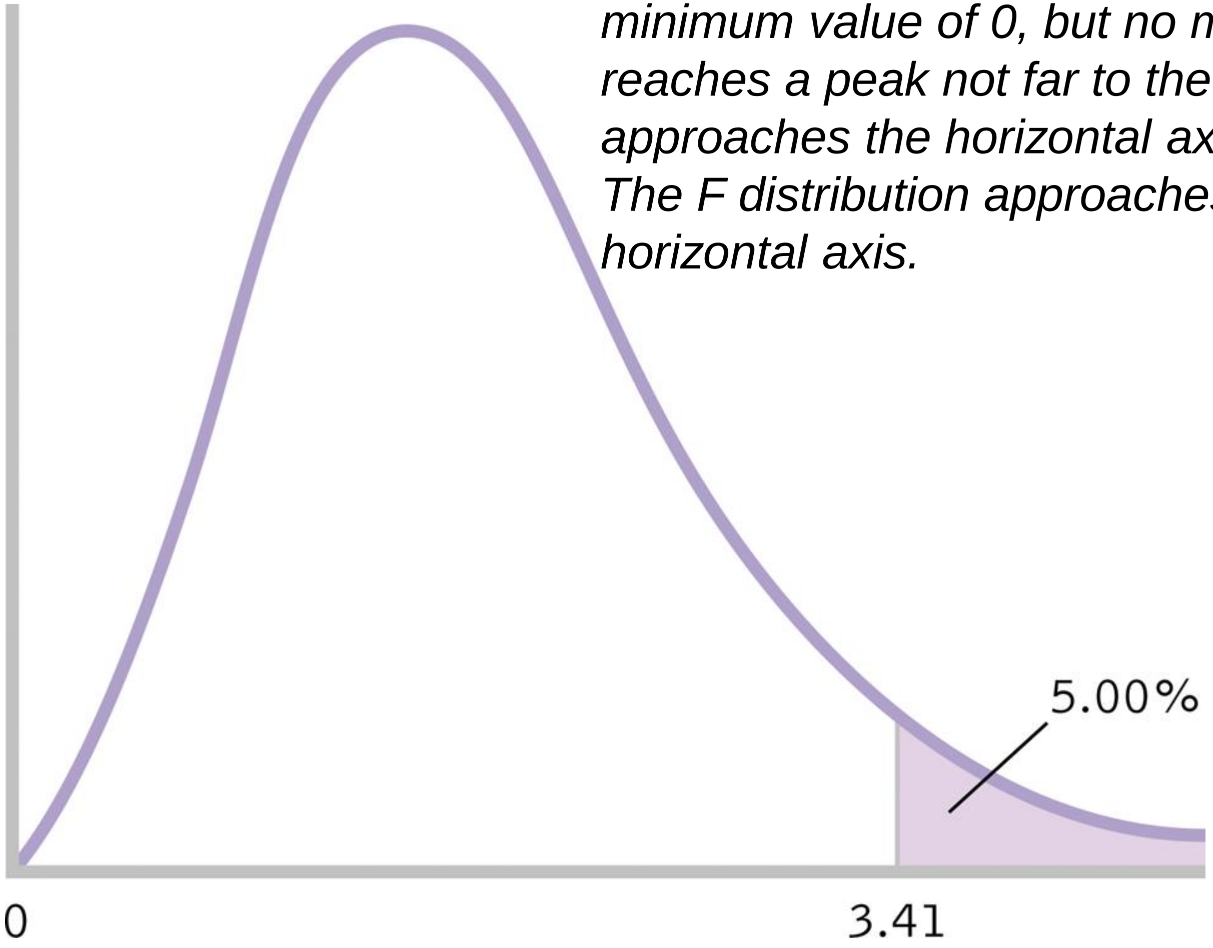
F Statistic

- $F = \frac{MS_b}{MS_w} = \frac{17.3681}{2.6938} = 6.4474$

Source	SS	df	MS	F	p
Between	34.7361	2	17.3681	6.4474	
Within	59.2639	22	2.6938		
Total	94.0000	24			

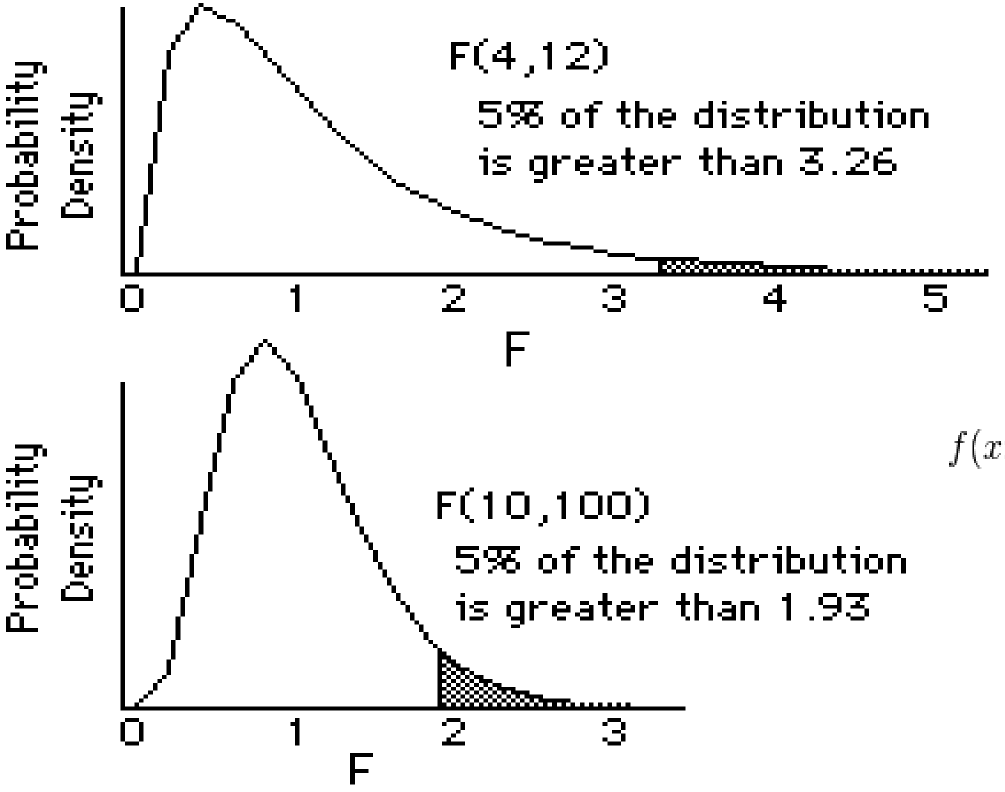
F Distribution

The F distribution is an asymmetric distribution that has a minimum value of 0, but no maximum value. The curve reaches a peak not far to the right of 0, and then gradually approaches the horizontal axis the larger the F value is. The F distribution approaches, but never quite touches the horizontal axis.



F Distribution

- F distribution changes shape depending on sample size and # of groups: $F(df_b, df_w)$



$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\,x)^{d_1}\,\,d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\!\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\!\left(\frac{d_1}{2},\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}}$$

P-value

Source	SS	df	MS	F	p
Between	34.7361	2	17.3681	6.4474	0.006
Within	59.2639	22	2.6938		
Total	94.0000	24			

With significance level of 5%, $P=0.006 < 0.01 < 0.05$

→ There is highly significant evidence to reject the null hypothesis

http://www.socr.ucla.edu/Applets.dir/F_table.html

https://www.danielsoper.com/statcalc/calculator.aspx?id=7

Do nematodes affect plant growth?

	Seedling growth				$\bar{x}_i$	s_i
0 nematode	10.8	9.1	13.5	9.2	10.65	2.053
1000 nematodes	11.1	11.1	8.2	11.3	10.425	1.486
5000 nematodes	5.4	4.6	7.4	5.0	5.6	1.244
10000 nematodes	5.8	5.3	3.2	7.5	5.45	1.771

Conditions required:

- equal variances: checking that largest s_i no more than twice smallest s_i

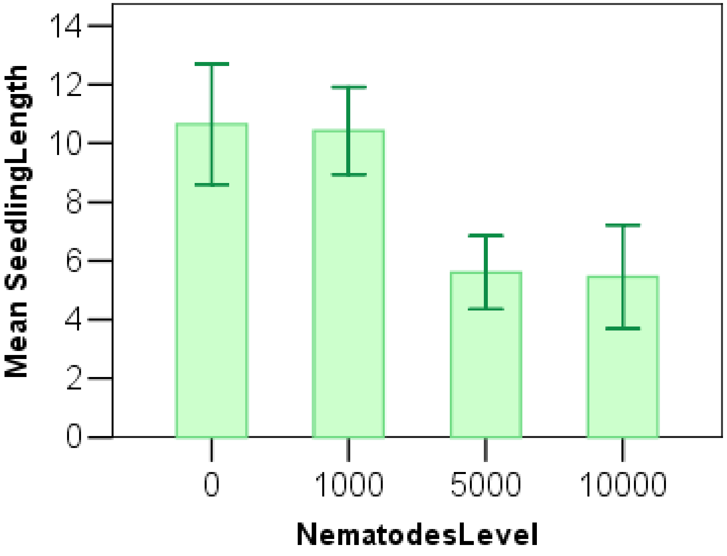
  Largest s_i = 2.053; smallest s_i = 1.244

- Independent SRSs

  Four groups obviously independent

- Distributions “roughly” normal

  It is hard to assess normality with only four points per condition. But the pots in each group are identical, and there is no reason to suspect skewed distributions.



Error bars: +/- 1.00 SD

Excel output for the one-way ANOVA

Menu/Tools/DataAnalysis/AnovaSingleFactor

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
0 nematode	4	42.6	10.65	4.21667
1000 nematodes	4	41.7	10.425	2.20917
5000 nematodes	4	22.4	5.6	1.54667
10000 nematodes	4	21.8	5.45	3.13667

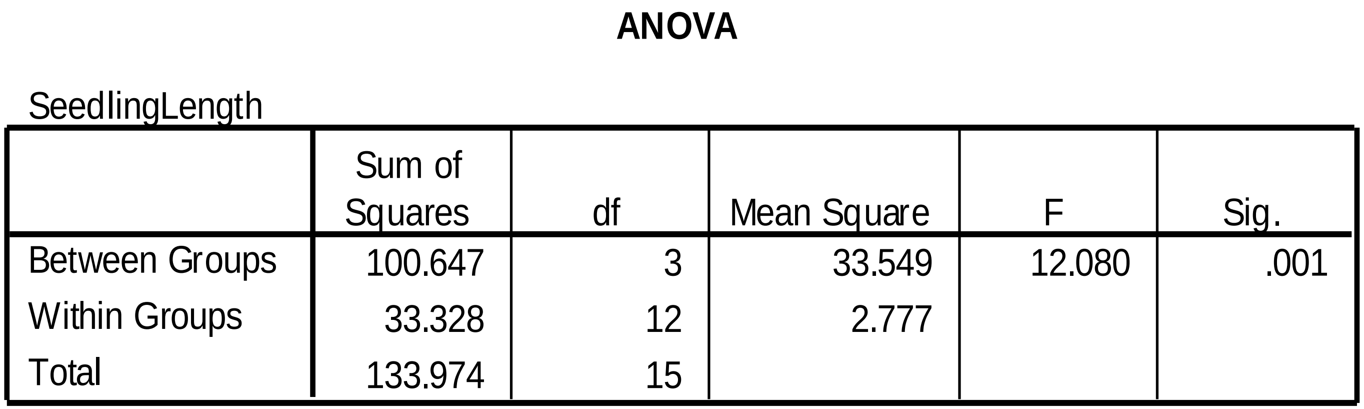
ANOVA

	Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
numerator	Between Groups	100.647	3	33.549	12.0797	0.00062	3.4902996
denominator	Within Groups	33.3275	12	2.77729			
	Total	133.974	15				

Here, the calculated F-value (12.08) is larger than $F_{critical}$ (3.49) for $\alpha = 0.05$.

Thus, the test is significant at $\alpha = 5\%$ ➔ Not all mean seedling lengths are the same; the number of nematodes is an influential factor.

SPSS output for the one-way ANOVA



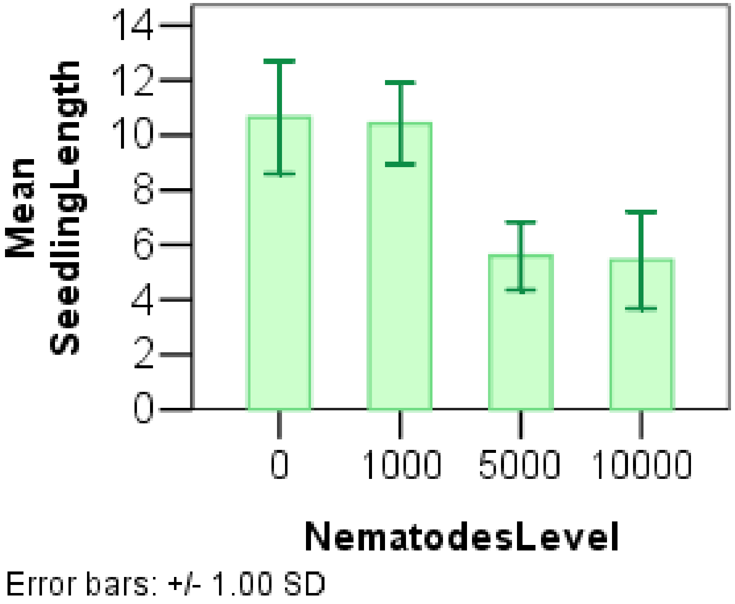
ANOVA

SeedlingLength

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	100.647	3	33.549	12.080	.001
Within Groups	33.328	12	2.777		
Total	133.974	15			

The **ANOVA** found that the amount of nematodes in pots significantly impacts seedling growth.

The **graph** suggests that nematode amounts above 1,000 per pot are detrimental to seedling growth.



Using Table E

The F distribution is asymmetrical and has two distinct degrees of freedom. This was discovered by Fisher, hence the label "F."

Once again, what we do is calculate the value of F for our sample data and then look up the corresponding area under the curve in Table E.

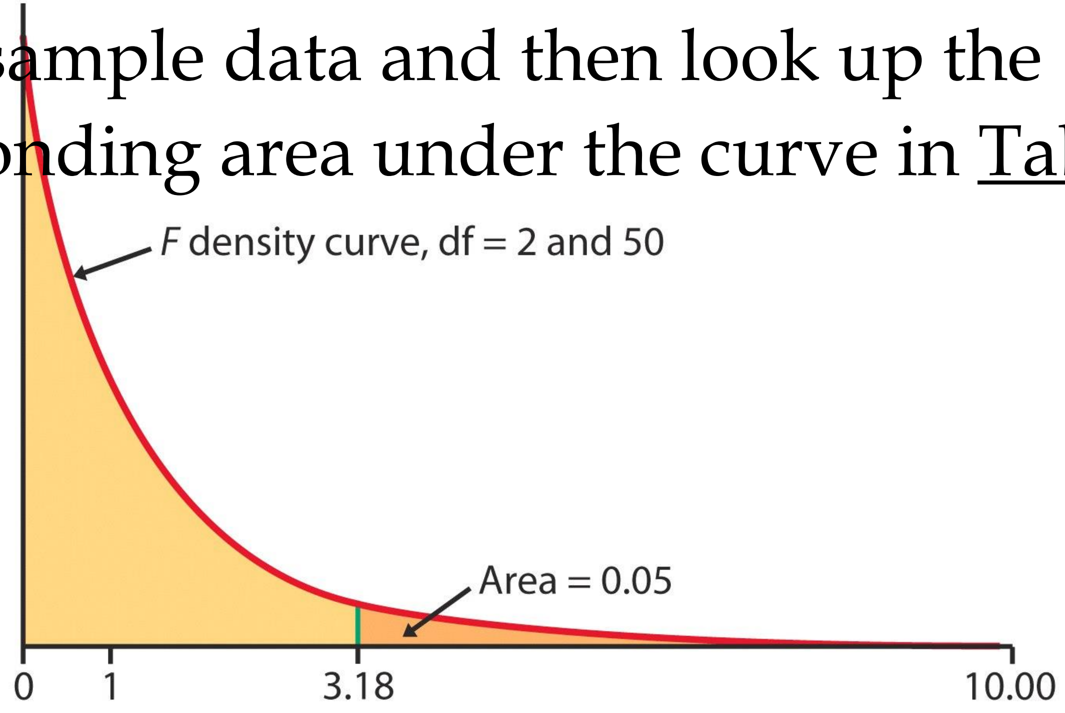


Table E *F* distribution critical values

$df_{num} = I - 1$

For df: 5,4

		Degrees of freedom in the numerator							
Degrees of freedom in the denominator	*p*	1	2	3	4	5	6	7	8
1	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66
	0.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859	5928.4	5981.1
	0.001	405284	500000	540379	562500	576405	585937	592873	598144
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98
	0.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
	0.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
	0.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03

p

F

$df_{den} = N - I$

One-Way vs Two-Way ANOVA

- One-way: only one factor is analyzed
  - For example, whether the brand of laundry detergent affects the amount of dirt removed from the laundry
- Two-way: 2 factors
  - Whether the brand of detergent and whether the water temperature affect the amount of dirt removed from the laundry
- Multi-way : more than 2 factors

Two-Way ANOVA

- Assumptions:
  - Samples are independent.
  - Data are (approximately) normally distributed.
  - The variances among different groups are equal.

- Balanced design: same sample size for all groups

http://statweb.stanford.edu/~susan/courses/s141/exanova.pdf
https://people.richland.edu/james/lecture/m170/ch13-2wy.html

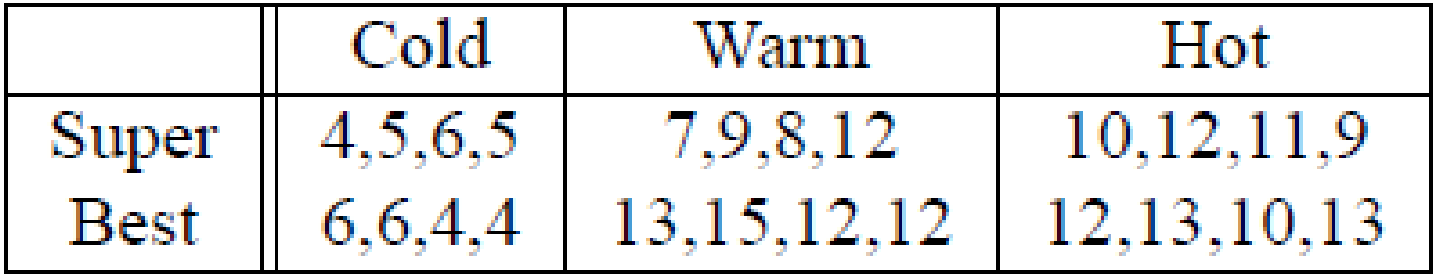
Null Hypotheses

Three sets of null hypotheses:

1. The amount of dirt removed does not depend on the type of detergent (Factor D)
2. The amount of dirt removed does not depend on the water temperature (Factor T)
3. There is no interaction between the two factors.

Treatment Groups

- Factor D: a = 2 (Super, Best)
- Factor T: b = 3 (cold, warm, hot)
- A total of k=a*b=6 treatment groups
- n=4 loads per group, N=nab=24 loads in total



	Cold	Warm	Hot
Super	4,5,6,5	7,9,8,12	10,12,11,9
Best	6,6,4,4	13,15,12,12	12,13,10,13

Variation (Error) Decomposition

Different T

Within group

Different D

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk}-\bar{X}\right)^2 \quad\longleftarrow\quad SS_{Total}$$

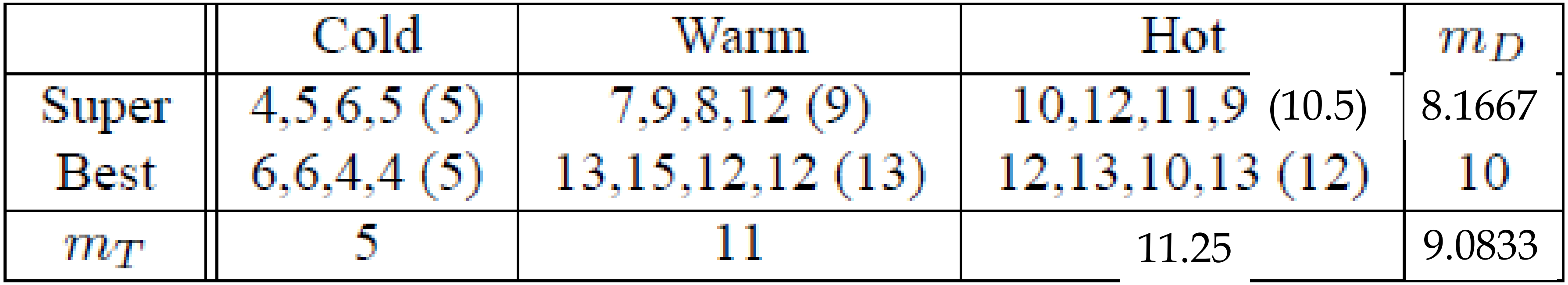
$$= n\cdot b\sum_{i=1}^{a}\left(\bar{X}_{i\cdot\cdot}-\bar{X}\right)^2 \quad\longleftarrow\quad SS_{D}$$

$$+n\cdot a\sum_{j=1}^{b}\left(\bar{X}_{\cdot j\cdot}-\bar{X}\right)^2 \quad\longleftarrow\quad SS_{T}$$

$$+n\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{X}_{ij\cdot}-\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot j\cdot}+\bar{X}\right)^2 \quad\longleftarrow\quad SS_{D\times T}$$

$$+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk}-\bar{X}_{ij\cdot}\right)^2 \quad\longleftarrow\quad SS_{W}$$

- Degree of freedom:
  - Factor D: $df_D = a - 1 = 1$
  - Factor T: $df_T = b - 1 = 2$
  - Interaction: $df_I = df_D \times df_T = 2$
  - Within group: $df_w = N - ab = 18$
- Sample means



	Cold	Warm	Hot	m_D
Super	4,5,6,5 (5)	7,9,8,12 (9)	10,12,11,9 (10.5)	8.1667
Best	6,6,4,4 (5)	13,15,12,12 (13)	12,13,10,13 (12)	10
m_T	5	11	11.25	9.0833

Two-Way ANOVA Table

Source	SS	df	MS	F	p
Factor D	20.167	1	20.167	9.8108	0.005758
Factor T	200.333	2	100.167	48.7297	5.44e-08
Inter. D x T	16.333	2	8.167	3.9790	0.037224
Within	37	18	2.056		
Total	266	23			

4 * { (5-8.1667-5+9.0833)^2 +
(9-8.1667-11+9.0833)^2 + … +
(12-11.25-10+9.0833)^2 }

Z-, t- and F-statistics

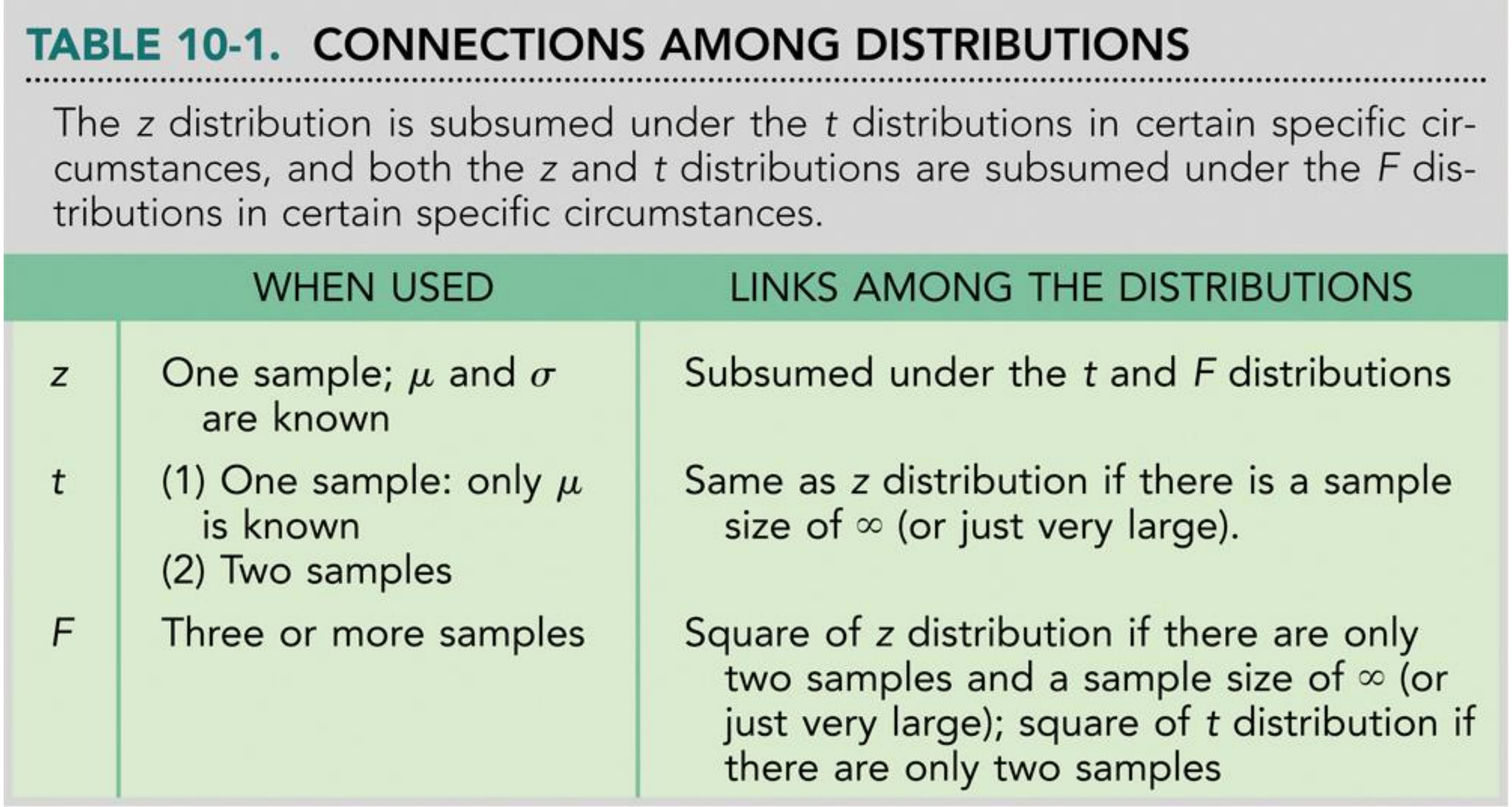


TABLE 10-1. CONNECTIONS AMONG DISTRIBUTIONS

The *z* distribution is subsumed under the *t* distributions in certain specific circumstances, and both the *z* and *t* distributions are subsumed under the *F* distributions in certain specific circumstances.

	WHEN USED	LINKS AMONG THE DISTRIBUTIONS
z	One sample; μ and σ are known	Subsumed under the *t* and *F* distributions
t	(1) One sample: only μ is known (2) Two samples	Same as *z* distribution if there is a sample size of ∞ (or just very large).
F	Three or more samples	Square of *z* distribution if there are only two samples and a sample size of ∞ (or just very large); square of *t* distribution if there are only two samples

Review: z- and t-Statistics

- One-sample hypothesis testing
  - Data collection: $\{X_1, \cdots, X_n\} \sim N(\mu, \sigma^2)$
  - Null hypothesis: $\mu = \mu_0$
  - Known var.: z $= \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ given H_0
  - Unknown var.: $t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}, s^2 = \frac{\sum_i (X_i-\bar{X})^2}{n-1}$

Paired Two-sample t-Test

- Two samples (distr.) of the same length n

  $\{X_1, X_2, \cdots X_n\}$ before drug treatment

  $\{Y_1, Y_2, \cdots, Y_n\}$ after drug treatment

- Determine whether the "after" member of the pair is different from the "before" one

  $\{d_1, d_2, \cdots d_n\}$ difference $d_i = Y_i - X_i$

- Hypothesis testing:
  - Null: the mean of the difference sample is 0
  - Alternative: the mean is not 0

Unpaired Two-Sample t Test

- Two indep. samples without natural pairs

$$X_1, X_2, \cdots X_n \sim N(\mu_x, \sigma_x^2)$$

$$Y_1, Y_2, \cdots Y_m \sim N(\mu_y, \sigma_y^2)$$

- Hypothesis testing
  - Null: $\mu_x = \mu_y$
  - Alternative: $\mu_x \neq \mu_y$

Theoretically…

- Sample mean difference $\bar{X} - \bar{Y}$

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

- Let’s think about how the t-value should be defined here

$$t = \frac{\bar{X} - \bar{Y} - \cancel{(\mu_x - \mu_y)}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$$

Under the null hypothesis, $\mu_x = \mu_y$

- Without knowing the true variances σ_x^2 and σ_y^2, we have to use the sample variances s_x^2 and s_y^2

Un-Pooled Variances

- Replace the true variances with sample var.

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_x^2/n + s_y^2/m}}$$

- It follows the Student's t-distribution with degree of freedom v
- It is complicated to figure out v here!
- A good approx. is given as $\approx \frac{2}{n^{-1}+m^{-1}}$ (harmonic mean of n and m)

Pooled Variance

- If we assume the variances are the same in both groups, we can pool all data to estimate a common variance.

- Redefined t-statistic

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_x^2/n + s_y^2/m}} = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{1/n + 1/m}}$$

Pooled Variance

- Pooling variance

$$s_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \Leftrightarrow (n-1)s_x^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$s_y^2 = \frac{\sum_{j=1}^{m}\left(Y_j - \bar{Y}\right)^2}{m-1} \Leftrightarrow (m-1)s_y^2 = \sum_{j=1}^{m}\left(Y_j - \bar{Y}\right)^2$$

$$s_p^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}\left(Y_j - \bar{Y}\right)^2}{n+m-2} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Degree of freedom

Comparing Two-Sample t- and F-stat

- t-statistic

$$t = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{1/n + 1/m}}, \text{ where } s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

- F-statistic

$$F = \frac{SS_b/1}{SS_w/(n+m-2)}, \text{ where}$$

$$SS_b = n(\bar{X} - \bar{\mu}_G)^2 + m(\bar{Y} - \bar{\mu}_G)^2, \bar{\mu}_G = \frac{n\bar{X}+m\bar{Y}}{n+m}$$

$$SS_w = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}\left(Y_j - \bar{Y}\right)^2 = (n-1)s_x^2 + (m-1)s_y^2$$

$F = t^2$ for two indep. groups

Post-Hoc Test

ANOVA can compare 3 or more groups.

- However, ANOVA just says at least one group has a different mean from the rest, but not which one.

A post-hoc test ("after the fact test") is a series of independent samples t-tests comparing each group's mean to each of the others' means.

- In our example of blister treatment, it will be k(k-1)/2=3 pairwise t-tests.
- Before that, be careful of the type-I error inflation!

Inflation of Type-I Error

- In general, if we perform m hypothesis tests, what is the probability of at least 1 false positive (type-I) error?
  - Assume all m null hypotheses are true.

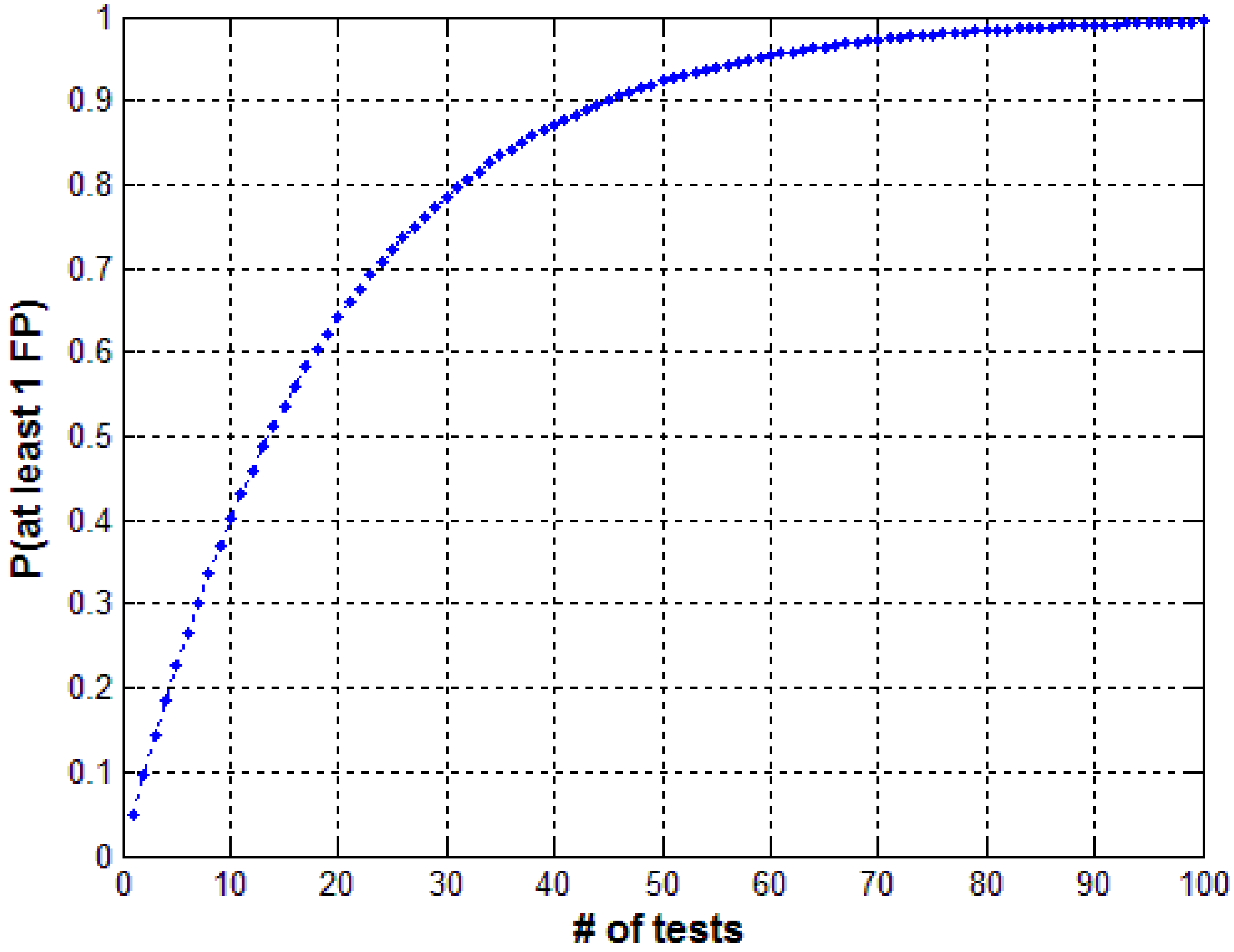
P(Making a f.p. error) = α

P(Not making a f.p. error) = 1 – α

P(Not making a f.p. error in m tests) = $(1 – \alpha)^m$

P(Making at least 1 f.p. error in m tests) = $1 - (1 – \alpha)^m \approx m\alpha$ for small α and $m's$

Probability of At Least 1 FP Error



$\alpha = 0.05$

Multiple Comparison Techniques

Because of the likelihood of multiple comparison errors, statisticians have created ways to reduce the multiple comparison error rate.

One of these is the Bonferroni, which adjusts the α-level for each comparison by the number of comparisons. This lowers the likelihood of rejection in each test, making the joint α-level equal to the original α-level.

For example with 6 comparisons, .05/6 = .008. So α-level for each comparison becomes .008. The combined likelihood of a type 1 error will be .05.

http://www2.hawaii.edu/~taylor/z631/multcomp.pdf

Multiple Comparison Techniques

There are other post-hoc tests out there.

- For example: Scheffe’s Test and Tukey Test

These will often allow you to not only compare each group to the others one at a time, but they will also allow you to combine groups to test each group against the combination of the others.

1. 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	.010	Yes
.011	.05	4	.013	Yes
.019	.05	3	.017	No
.032	.05	2	.025	No
.048	.05	1	.050	Yes

2/3. Tukey and Sheffé

- Both methods increase your *p*-values to account for the fact that you’ve done multiple comparisons, but are less conservative than Bonferroni (let computer calculate for you!).

- SAS options in PROC GLM:
    - adjust=tukey
    - adjust=scheffe

4/5. Holm and Hochberg

- Arrange all the resulting p-values (from the $T={}_kC_r$ pairwise comparisons) in order from smallest (most significant) to largest: p_1 to p_T

Holm

1. Start with p_1, and compare to Bonferroni p (=α/T).
2. If $p_1 < \alpha/T$, then p_1 is significant and continue to step 2. If not, then we have no significant p-values and stop here.
3. If $p_2 < \alpha/(T-1)$, then p_2 is significant and continue to step. If not, then p_2 thru p_T are not significant and stop here.
4. If $p_3 < \alpha/(T-2)$, then p_3 is significant and continue to step If not, then p_3 thru p_T are not significant and stop here.

Repeat the pattern…

Holm

- Let $H_1, \ldots, H_m$ be a family of hypotheses and $P_1, \ldots, P_m$ the corresponding P-values.
- Start by ordering the p-values (from lowest to highest) $P_{(1)} \ldots P_{(m)}$ and let the associated hypotheses be $H_{(1)} \ldots H_{(m)}$
- For a given significance level α, let k be the minimal index such that $$P_{(k)} > \frac{\alpha}{m+1-k}$$
- Reject the null hypotheses $H_{(1)} \ldots H_{(k-1)}$ and do not reject $H_{(k)} \ldots H_{(m)}$
- If $k = 1$ then do not reject any of the null hypotheses and if no such k exist then reject all of the null hypotheses.

Hochberg

1. Start with largest (least significant) p-value, p_T, and compare to α. If it's significant, so are all the remaining p-values and stop here. If it's not significant then go to step 2.
2. If $p_{T-1} < \alpha/(T-1)$, then p_{T-1} is significant, as are all remaining smaller p-vales and stop here. If not, then p_{T-1} is not significant and go to step 3.

Repeat the pattern…

Note: Holm and Hochberg should give you the same results. Use Holm if you anticipate few significant comparisons; use Hochberg if you anticipate many significant comparisons.

Practice Problem

A large randomized trial compared an experimental drug and 9 other standard drugs for treating motion sickness. An ANOVA test revealed significant differences between the groups. The investigators wanted to know if the experimental drug (“drug 1”) beat any of the standard drugs in reducing total minutes of nausea, and, if so, which ones. The p-values from the pairwise ttests (comparing drug 1 with drugs 2-10) are below.

Drug 1 vs. drug …	2	3	4	5	6	7	8	9	10
p-value	.05	.3	.25	.04	.001	.006	.08	.002	.01

a. Which differences would be considered statistically significant using a Bonferroni correction? A Holm correction? A Hochberg correction?

Answer

Bonferroni makes new α value = $\alpha/9 = .05/9 = .0056$; therefore, using Bonferroni, the new drug is only significantly different than standard drugs 6 and 9.

Arrange p-values:

6	9	7	10	5	2	8	4	3
.001	.002	.006	.01	.04	.05	.08	.25	.3

Holm: .001<.0056; .002<.05/8=.00625; .006<.05/7=.007; .01>.05/6=.0083; therefore, new drug only significantly different than standard drugs 6, 9, and 7.

Hochberg: .3>.05; .25>.05/2; .08>.05/3; .05>.05/4; .04>.05/5; .01>.05/6; .006<.05/7; therefore, drugs 7, 9, and 6 are significantly different.

Practice problem

- b. 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?

Answer

- The magnitude of the reduction in minutes of nausea.
- If large enough sample size, a 1-minute difference could be statistically significant, but it's obviously not clinically meaningful and you probably wouldn't recommend a switch.

References

- Casella, G. and Berger, R. (2002). *Statistical Inference*. United States: Duxbury.
- Cochran, W. G. (1947). Some Consequences When the Assumptions for the Analysis of Variances are not Satisfied. *Biometrics*. Vol. 3, 22-38.
- Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrics*. Vol. 3, 1-21.
- Ito, P. K. (1980). Robustness of ANOVA and MANOVA Test Procedures. *Handbook of Statistics 1: Analysis of Variance* (P. R. Krishnaiah, ed.), 199-236. Amsterdam: North-Holland.
- Kaskey, G., et al. (1980). Transformations to Normality. *Handbook of Statistics 1: Analysis of Variance* (P. R. Krishnaiah, ed.), 321-341. Amsterdam: North-Holland.
- Kuehl, R. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*, 2nd edition. United States: Duxbury.
- Kutner, M. H., et al. (2005). *Applied Linear Statistical Models*, 5th edition. New York: McGraw-Hill.
- Mardia, K. V. (1980). Tests of Univariate and Multivariate Normality. *Handbook of Statistics 1: Analysis of Variance* (P. R. Krishnaiah, ed.), 279-320. Amsterdam: North-Holland.
- Tabachnik, B. and Fidell, L. (2001). *Computer-Assisted Research Design and Analysis*. Boston: Allyn & Bacon.

Q&A?