Biostatistics

Chapter 6 Comparison of several groups: ANOVA II

Jing Li

jing.li@sjtu.edu.cn

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

Dept of Bioinformatics & Biostatistics, SJTU

Recall: Steps for ANOVA test

1) State null and alternative hypotheses
   - H_0: $m_1 = m_2 = \; .... = m_n$
   - H_A: at least one mean is different
2) Specify α level
3) Calculate test statistic: See ANOVA table
4) Calculate p-value: See ANOVA table
5) Reject or not reject null
6) Make conclusions

Review lecture 6A

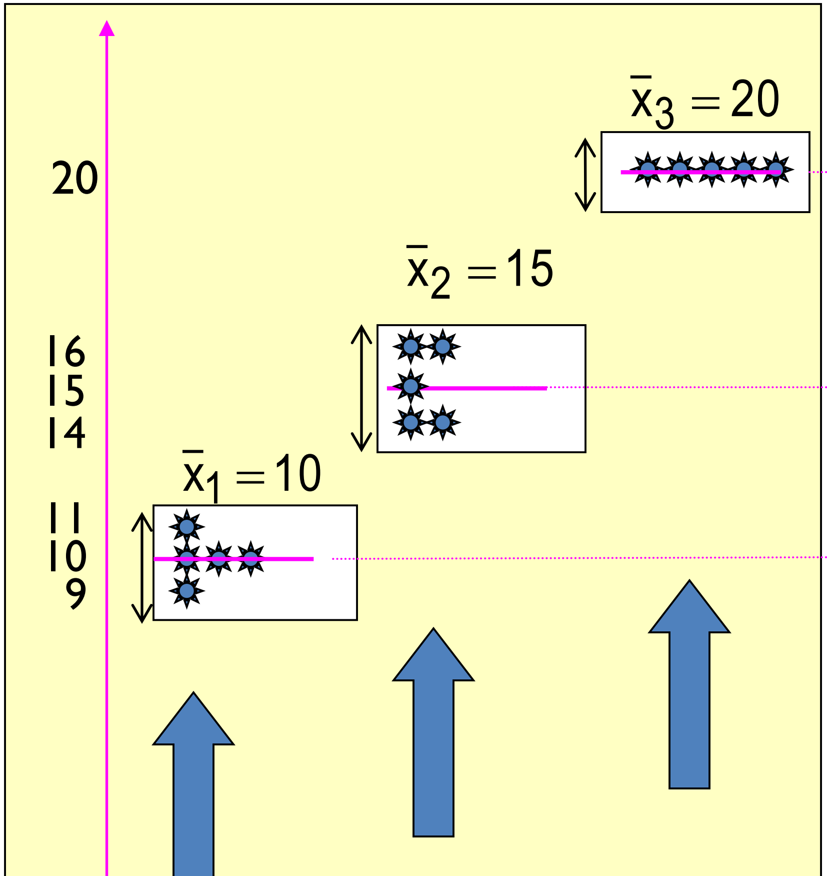
- ANOVA (one-way)

Summarizes the mean differences between all groups at once.

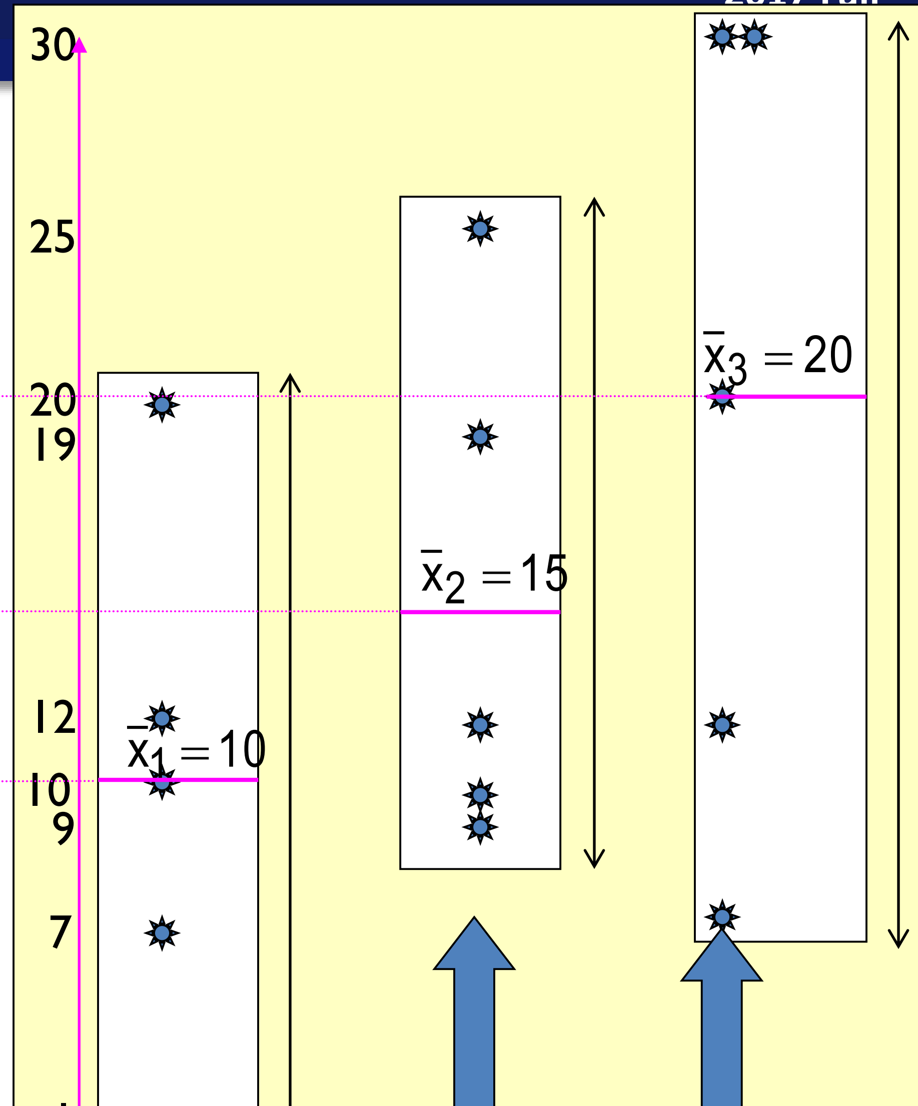
$$F = \frac{Variability\ between\ groups}{Variability\ within\ groups}$$

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

Analogous to pooled variance from a ttest.



A small variability within the samples makes it easier to draw a conclusion about the population means.



The sample means are the same as before, but the larger within-sample variability makes it harder to draw a conclusion about the population means.

Model Assumptions in ANOVA

- *Homoscedasticity* (common group variances).
- Normality of responses (or of residuals).
- Independence of responses (or of residuals).
  (Hopefully achieved through randomization…)
- Effect additivity. (Only an issue in multi-way AOV).

Recall: variances equality for t-test

- One way to test if the two variances are equal is to check if the ratio is equal to 1
- Under the null, the ratio simplifies to $\frac{s_1^2}{s_2^2}$
- The ratio of 2 chi-square random variables has an F-distribution
- The F-distribution is defined by the numerator and denominator degrees of freedom
- Here we have an F-distribution with n_1-1 and n_2-1 degrees of freedom
- This works better with $s_1^2 > s_2^2$

Checking the Equal Variance Assumption of ANOVA

$$H_0 : s_1^2 = s_2^2 = \cdots = s_t^2$$

H_A: some of the variances are different from each other

Little work but little power

Hartley's Test: A logical extension of the F test for t=2.

Requires equal replication, **n**, among **t** groups. Requires normality.

$$F_{max} = \frac{s^2_{max}}{s^2_{min}}$$

Reject if $F_{max} > F_{\alpha,t,n-1}$, tabulated in F Table.

Checking the Equal Variance Assumption of ANOVA

- Bartlett's Test

More work but better power

The Bartlett test is defined as:

H_0:	$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$
H_a:	$\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j).
Test Statistic:	The Bartlett test statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups.

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i - 1)\ln s_i^2}{1 + (1/(3(k-1)))((\sum_{i=1}^{k} 1/(N_i - 1)) - 1/(N-k))}$$

In the above, s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as:

$$s_p^2 = \sum_{i=1}^{k}(N_i - 1)s_i^2/(N-k)$$

Reject H_0 if $T > \chi^2_{(k-1),\alpha}$

One-way ANOVA

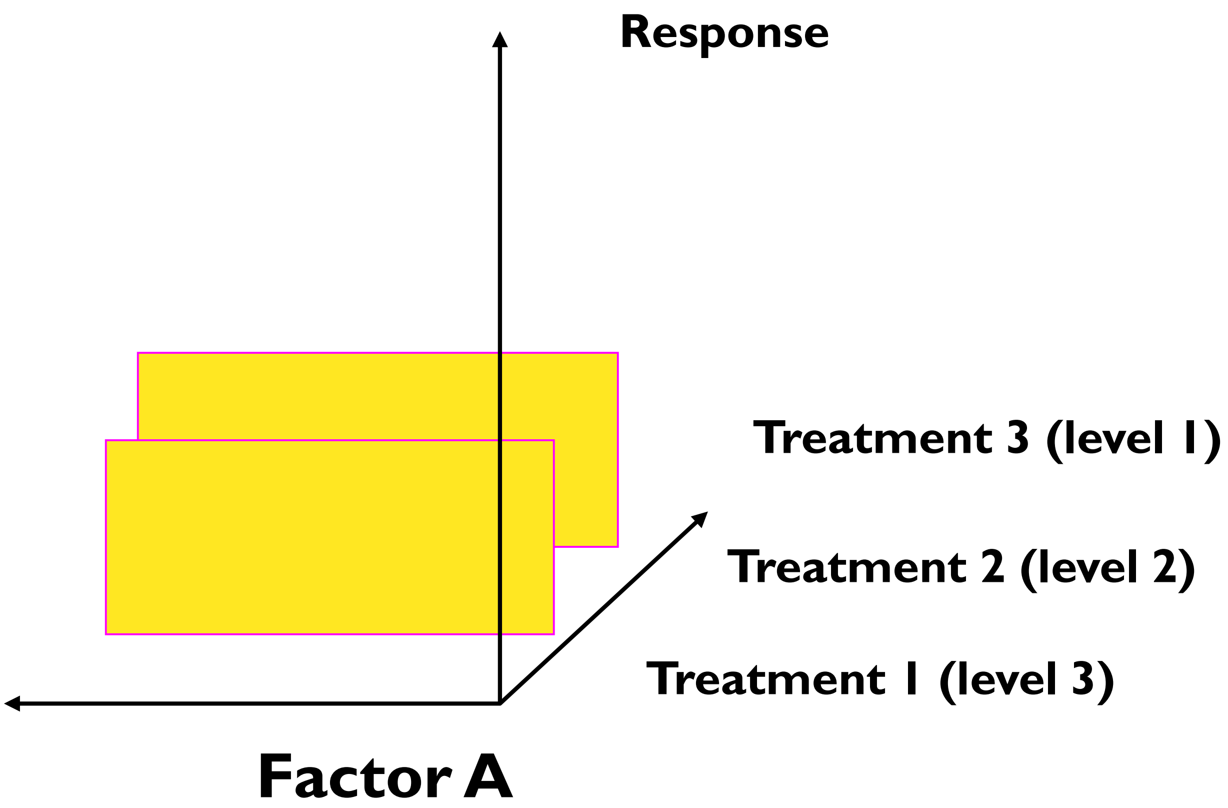
$$F = \frac{Variability\ between\ groups}{Variability\ within\ groups}$$

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

Analysis of Variance Experimental Designs

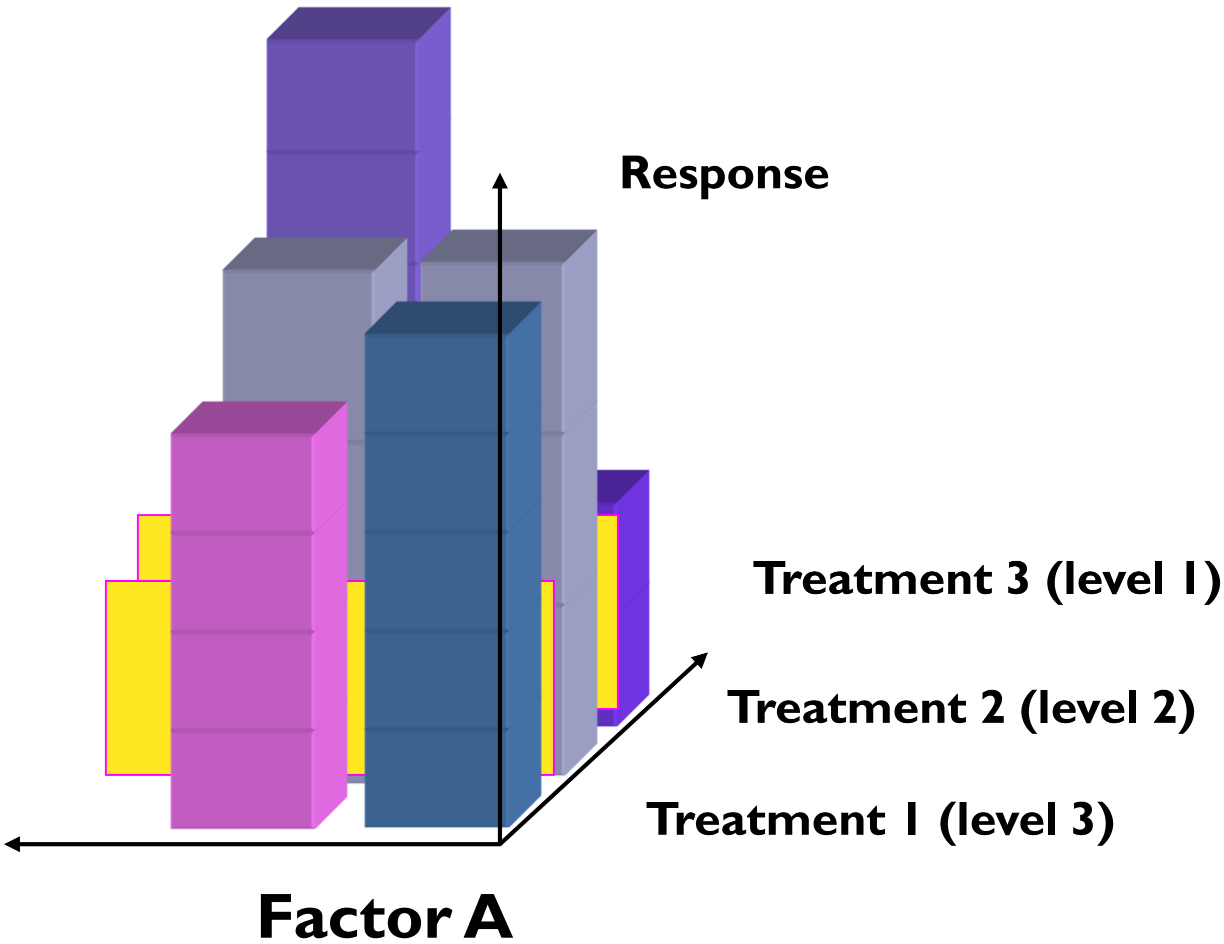
- Several elements may distinguish between one experimental design and others.
  - **The number of factors.**
    - Each characteristic investigated is called a **factor (因素)**.
    - Each factor has several **levels** (水平).

One - way ANOVA : single factor

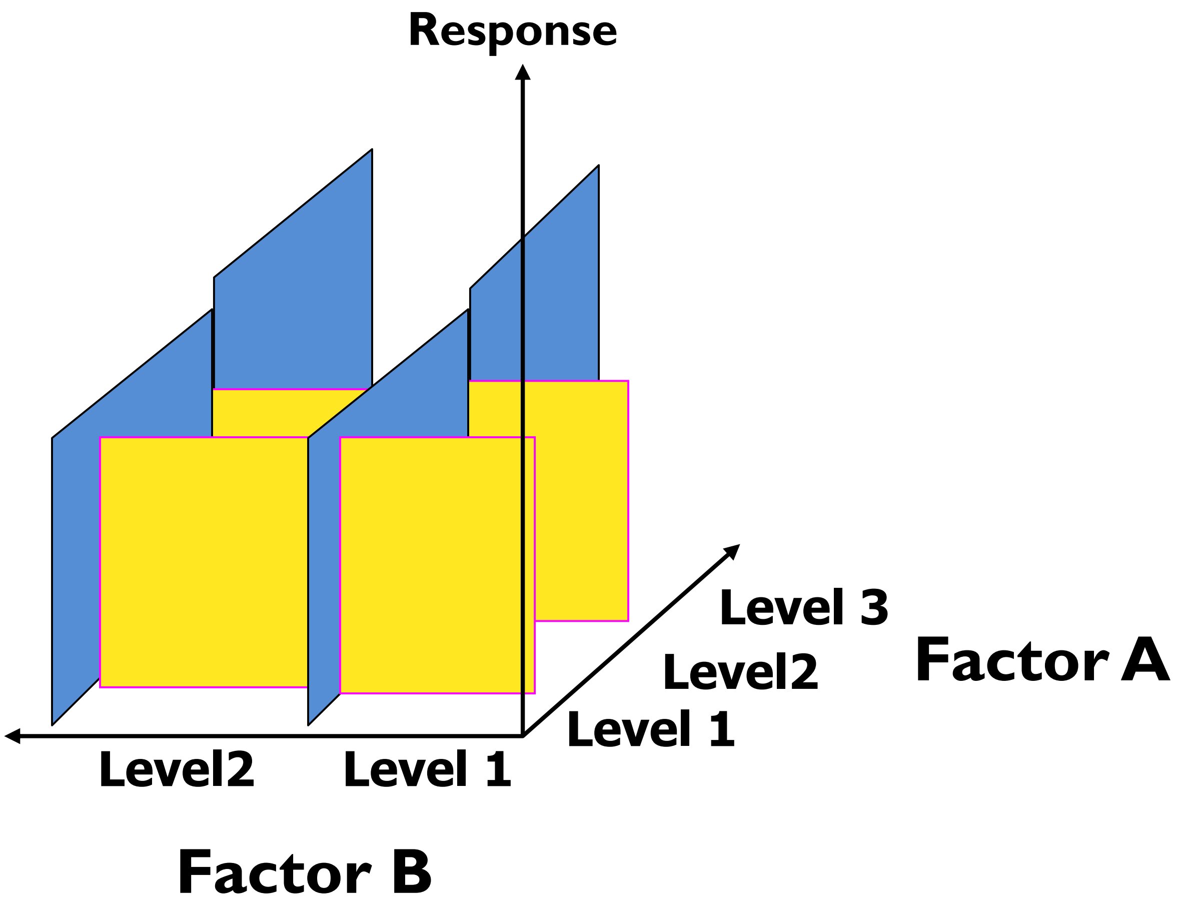


One - way ANOVA ：single factor

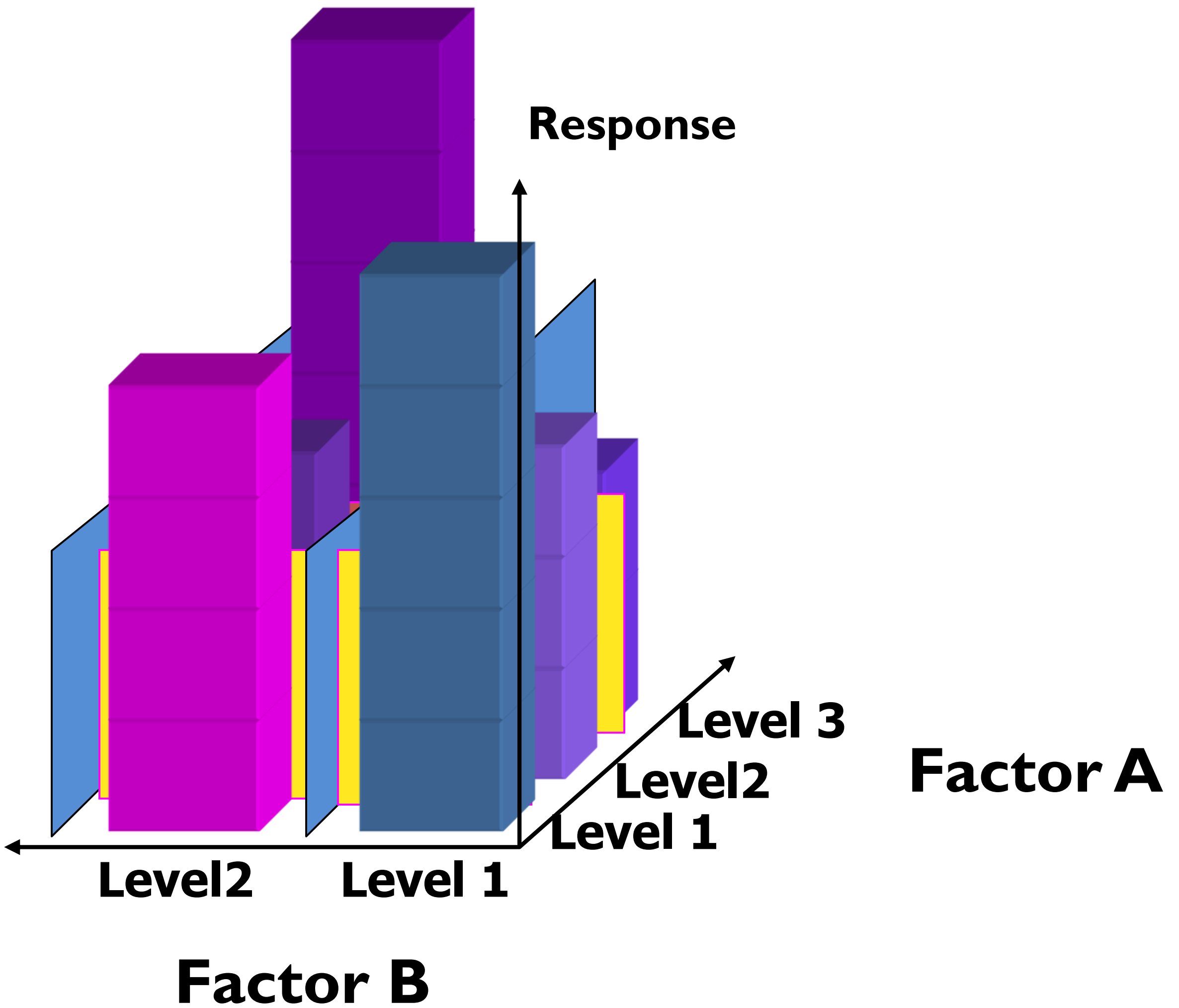
单因素三水平



Two - way ANOVA: two factors



Two - way ANOVA: two factors



Randomized Block Design (Two-way ANOVA without replication)

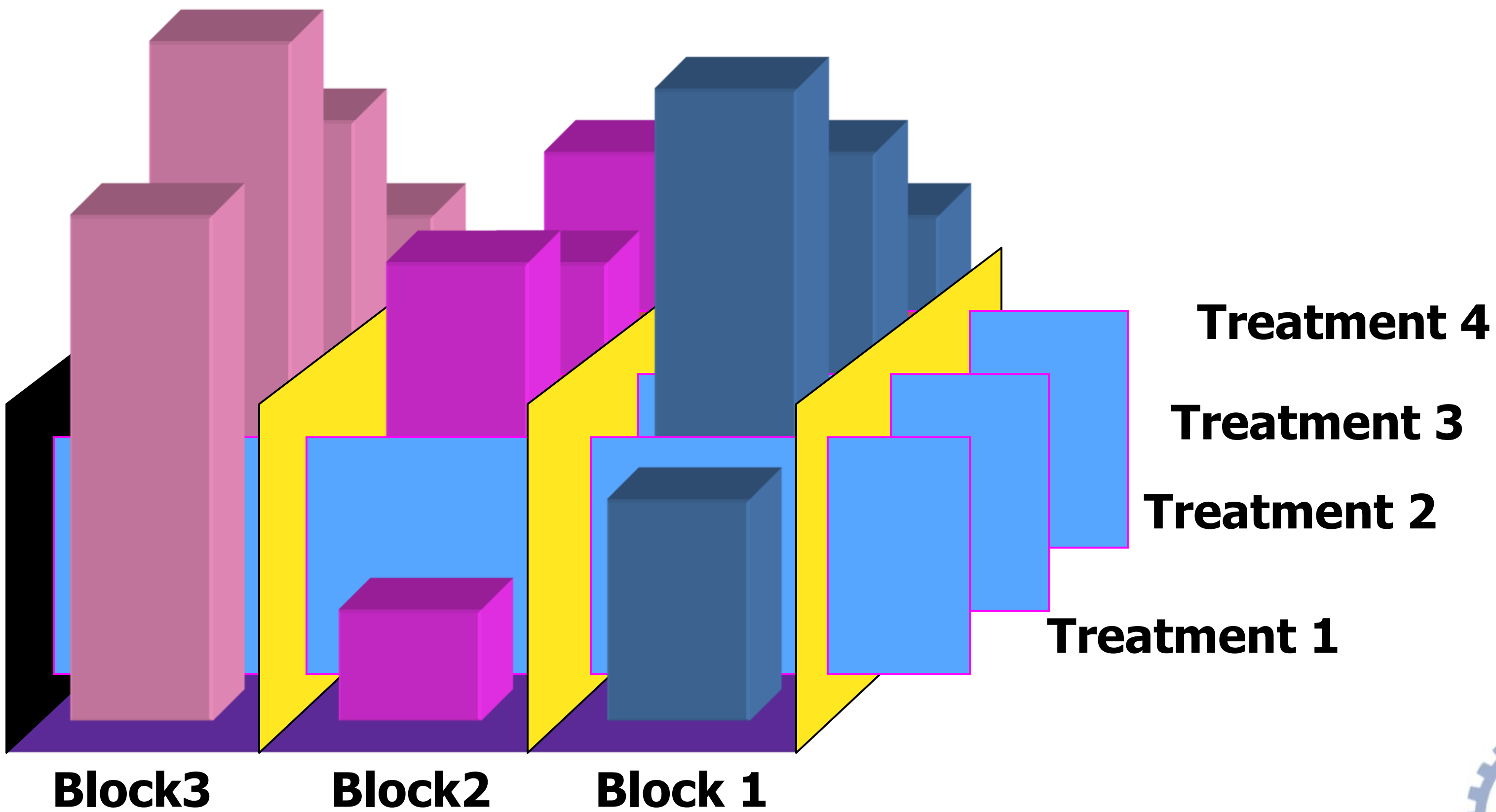
- The purpose of designing a randomized block experiment is to **reduce** the *within-treatments variation* thus increasing the *relative amount of between treatment variation.*
- This helps in detecting differences between the treatment means more easily.

Response =factor A + factor B + random error

随机区组

Randomized Block Design 随机区组方差分析

Block all the observations with some commonality across treatments



Randomized Blocks

Block all the observations with some commonality across treatments

	Treatment				
Block	1	2		k	**Block mean**
1	X11	X12	. . .	X1k	$\bar{x}[B]_1$
2	X21	X22		X2k	$\bar{x}[B]_2$
.					
.					
.					
b	Xb1	Xb2		Xbk	$\bar{x}[B]_b$
Treatment mean	$\bar{x}[T]_1$	$\bar{x}[T]_2$		$\bar{x}[T]_k$	

Partitioning the total variability

- The sum of square total is partitioned into three sources of variation
  - **Treatments**
  - **Blocks**
  - **Within samples (Error)**

$$SS(Total) = SST + SSB + SSE$$

Sum of square for treatments | **Sum of square for blocks** | **Sum of square for error**

Partitioning the total variability

- The sum of square total is partiti sources of variation
  - **Treatments**
  - **Blocks**
  - **Within samples (Error)**

For the independent samples design we have:
SS(Total) = SST + SSE

$$SS(Total) = SST + SSB + SSE$$

Sum of square for treatments	**Sum of square for blocks**	**Sum of square for error**

Calculating the sums of squares

- Formulai for the calculation of the sums of squares

	Treatment				
Block	1	2		k	**Block mean**
1	X11	X12	. . .	X1k	$\bar{x}[B]_1$
2	X21	X22		X2k	$\bar{x}[B]_2$
.					
.					
.					
b	Xb1	Xb2		Xbk	
Treatment mean	$\bar{x}[T]_1$	$\bar{x}[T]_2$		$\bar{x}[T]_k$	$\bar{\bar{X}}$

$$SSB = k\left((\bar{x}[B]_1) - \bar{\bar{X}}\right)^2 + k\left((\bar{x}[B]_2) - \bar{\bar{X}}\right)^2 + \cdots + k\left((\bar{x}[B]_k) - \bar{\bar{X}}\right)^2$$

$$SST = b\left((\bar{x}[T]_1) - \bar{\bar{X}}\right)^2 + b\left((\bar{x}[T]_2) - \bar{\bar{X}}\right)^2 + \cdots + b\left((\bar{x}[T]_k) - \bar{\bar{X}}\right)^2$$

Calculating the sums of squares

- Formulai for the calculation of the sums of squares

	Treatment				
Block	1	2		k	**Block mean**
1	X11	X12	. . .	X1k	$\bar{x}[B]_1$
2	X21	X22		X2k	$\bar{x}[B]_2$
b	Xb1	Xb2		Xbk	
Treatment mean	$\bar{x}[T]_1$	$\bar{x}[T]_2$		$\bar{x}[T]_k$	$\bar{\bar{X}}$

$$SS(Total) = (x_{11} - \bar{\bar{X}})^2 + (x_{21} - \bar{\bar{X}})^2 + \ldots + (x_{12} - \bar{\bar{X}})^2 + (x_{22} - \bar{\bar{X}})^2 + \ldots + (X_{1k} - \bar{\bar{X}})^2 + (x_{2k} - \bar{\bar{X}})^2 + \ldots =$$

$$SSB = k\left((\bar{x}[B]_1) - \bar{\bar{X}}\right)^2 + k\left((\bar{x}[B]_2) - \bar{\bar{X}}\right)^2 + \cdots + k\left((\bar{x}[B]_k) - \bar{\bar{X}}\right)^2$$

$$SST = b\left((\bar{x}[T]_1) - \bar{\bar{X}}\right)^2 + b\left((\bar{x}[T]_2) - \bar{\bar{X}}\right)^2 + \cdots + b\left((\bar{x}[T]_k) - \bar{\bar{X}}\right)^2$$

Mean Squares

To perform hypothesis tests for treatments and blocks we need

- **Mean square for treatments**
- **Mean square for blocks**
- **Mean square for error**

$$MST = \frac{SST}{k-1}$$

$$MSB = \frac{SSB}{b-1}$$

SSE=SS$_{total}$ **- SST-SSB**

$$MSE = \frac{SSE}{N-k-b+1}$$

Test statistics for the randomized block design ANOVA

Test statistic for treatments

$$F = \frac{MST}{MSE}$$

Test statistic for blocks

$$F = \frac{MSB}{MSE}$$

The F test rejection regions

- Testing the mean responses for treatments

$$F > F_{a,k-1,n-k-b+1}$$

- Testing the mean response for blocks

$$F > F_{a,b-1,n-k-b+1}$$

Randomized Blocks ANOVA - Example

– Are there differences in the effectiveness of cholesterol reduction drugs?

– To answer this question the following experiment was organized:

- 25 groups of men with high cholesterol were matched by age and weight. Each group consisted of 4 men.
- Each person in a group received a different drug.
- The cholesterol level reduction in two months was recorded.

Group	Drug 1	Drug 2	Drug 3	Drug 4
1	6.6	12.6	2.7	8.7
2	7.1	3.5	2.4	9.3
3	7.5	4.4	6.5	10
4	9.9	7.5	16.2	12.6
5	13.8	6.4	8.3	10.6
6	13.9	13.5	5.4	15.4
7	15.9	16.9	15.4	16.3
8	14.3	11.4	17.1	18.9
9	16	16.9	7.7	13.7
10	16.3	14.8	16.1	19.4
11	14.6	18.6	9	18.5
12	18.7	21.2	24.3	21.1
13	17.3	10	9.3	19.3
14	19.6	17	19.2	21.9
15	20.7	21	18.7	22.1
16	18.4	27.2	18.9	19.4
17	21.5	26.8	7.9	25.4
18	20.4	28	23.8	26.5
19	21.9	31.7	8.8	22.2
20	22.5	11.9	26.7	23.5
21	21.5	28.7	25.2	19.6
22	25.2	29.5	27.3	30.1
23	23	22.2	17.6	26.6
24	23.7	19.5	25.6	24.5
25	28.4	31.2	26.1	27.4

Can we infer from the data that there are differences in mean cholesterol reduction among the four drugs?

Randomized Blocks ANOVA - Example

- **Solution**
  - Each drug can be considered a treatment.
  - Each 4 records (per group) can be blocked, because they are matched by age and weight.
  - This procedure eliminates the variability in cholesterol reduction related to different combinations of age and weight.
  - This helps detect differences in the mean cholesterol reduction attributed to the different drugs.

Randomized Blocks ANOVA - Example

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	3848.7	24	160.36	10.11	0.0000	1.67
Columns	196.0	3	65.32	4.12	0.0094	2.73
Error	1142.6	72	15.87			
Total	5187.2	99				

Treatments

Blocks

b-1

K-1

MST / MSE

MSB / MSE

Conclusion: At 5% significance level there is sufficient evidence to infer that the mean "cholesterol reduction" gained by at least two drugs are different.

Two way ANOVA without replication

- A new fertilizer has been developed to increase the yield on crops, and the makers of the fertilizer want to better understand which of the three formulations (blends) of this fertilizer are most effective for wheat, corn, soy beans and rice (crops). They test each of the three blends on one sample of each of the four types of crops. The crop yields for the 12 combinations are as shown in Figure 1.

	Wheat	Corn	Soy	Rice
Blend X	123	138	110	151
Blend Y	145	165	140	167
Blend Z	156	176	185	175

Two way ANOVA without replication

- Repeat the analysis from Example 1 of Two Factor ANOVA without Replication, but this time with the data shown in Figure 1 where each combination of blend and crop has a sample of size 5.

	Crop			
Fertilizer	Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

Two Way ANOVA with replication

两因素交叉分组实验设计

Source of Variation	Sums of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)	F-statistic
Cells	$\sum_{i=1}^{a}\sum_{j=1}^{b} n(\bar{X}_{ij} - \bar{X})^2$	ab - 1		
Factor A	$bn\sum_{i=1}^{a}(\bar{X}_{i.} - \bar{X})^2$	a – 1	$\frac{SS(A)}{DF(A)}$	$F = \frac{MS(A)}{MSE}$
Factor B	$an\sum_{j=1}^{b}(\bar{X}_{.j} - \bar{X})^2$	b - 1	$\frac{SS(B)}{DF(B)}$	$F = \frac{MS(B)}{MSE}$
A x B	cells SS – factor A SS – factor B SS	(a – 1)(b –1)	$\frac{SS(AxB)}{DF(AxB)}$	$F = \frac{MS(AxB)}{MSE}$
Error	$\sum_{i=1}^{a}\sum_{j=1}^{b}\left[\sum_{l=1}^{n}(X_{ijl} - \bar{X}_{ij})^2\right]$	ab(n – 1)	$\frac{SS(Error)}{DF(Error)}$	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{l=1}^{n}(X_{ijl} - \bar{X})^2$	N – 1		

Where: a= the number of levels of factor A
b= the number of levels of factor B
n= the number of replicants

Model

$$SST = SSA + SSB + SS(AB) + SSE$$

Sums of squares

$$SS(A) = rb\sum_{i=1}^{a}(\bar{x}[A]_i - \bar{\bar{x}})^2$$

$$SS(B) = ra\sum_{j=1}^{b}(\bar{x}[B]_j - \bar{\bar{x}})^2$$

$$SS(AB) = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}[AB]_{ij} - \bar{x}[A]_i - \bar{x}[B]_j + \bar{\bar{x}})^2$$

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{x}[AB]_{ij})^2$$

F tests for the Two-way ANOVA

- Test for the difference between the levels of the main factors A and B

$$F = \frac{MS(A)}{MSE}$$

MS(A) = SS(A)/(a-1)

Rejection region: $F > F_{\alpha, a-1, n-ab}$

$$F = \frac{MS(B)}{MSE}$$

MS(B) = SS(B)/(b-1)

MSE = SSE/(n-ab)

$F > F_{\alpha, b-1, n-ab}$

- Test for interaction between factors A and B

$$F = \frac{MS(AB)}{MSE}$$

MS(AB) = SS(AB)/(a-1)(b-1)

Rejection region: $F > F_{\alpha,(a-1)(b-1),n-ab}$

Two way ANOVA without replication

- Repeat the analysis from Example 1 of Two Factor ANOVA without Replication, but this time with the data shown in Figure 1 where each combination of blend and crop has a sample of size 5.

	Crop			
Fertilizer	Wheat	Corn	Soy	Rice
Blend X	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Blend Y	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Blend Z	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169