Chapter 7: Inference for numerical data

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

OpenIntro

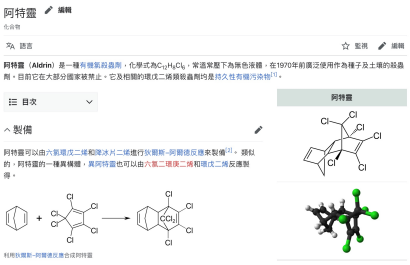
Outline

Comparing means with ANOVA

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides).
- These highly toxic organic compounds can cause various cancers and birth defects.
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth.
- But since these compounds are denser than water and their molecules tend to stick to particles of sediment, they are more likely to be found in higher concentrations near the bottom than near mid-depth.

Data

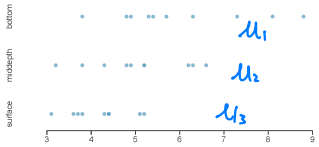
Aldrin concentration (nanograms per liter) at three levels of depth.



	aldrin	depth
1	3.80	bottom
2	4.80	bottom
...		
10	8.80	bottom
11	3.20	middepth
12	3.80	middepth
...		
20	6.60	middepth
21	3.10	surface
22	3.60	surface
...		
30	5.20	surface

Exploratory analysis

Aldrin concentration (nanograms per liter) at three levels of depth.



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

	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.20	0.66
overall	30	5.10	1.37

ANOVA table (except p-value)

Research question

Is there a difference between the mean aldrin concentrations among the three levels?

- To compare means of groups we use a new test called ANOVA and a new statistic called F.

ANalysis Of VAriance

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

H_0: The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different than others.

Conditions

1. The observations should be independent within and between groups
   - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
   - Carefully consider whether the data may be independent (e.g. no pairing).
   - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
   - Especially important when the sample sizes are small.
3. The variability across the groups should be about equal.
   - Especially important when the sample sizes differ between groups.

		Obs.			
Group	1	X_{11}	X_{12} ...	X_{1,n_1}	μ_1
	2	X_{21}	X_{22} ...	X_{2,n_2}	μ_2
	3	X_{31}	X_{32} ...	X_{3,n_3}	μ_3

$$X_{ij} \overset{\text{indep.}}{\sim} N(\mu_i, \sigma^2)$$

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

H_1: Not H_0

ANOVA - Method

ANOVA Compute a test statistic (a ratio).

between

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

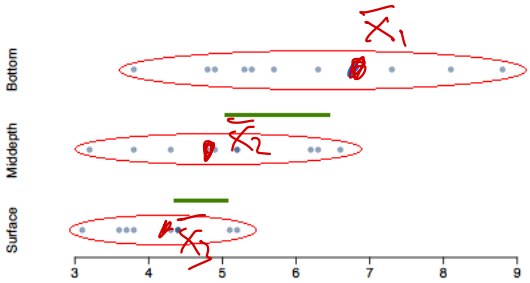
within

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.
- ANOVA compares the sample means to an overall grand mean.

Test statistic

Does there appear to be a lot of variability within groups? How about between groups?

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



F distribution and p-value

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

F - dist.

- $X \sim \chi^2(r)$
- $Y \sim \chi^2(m)$
- $X \perp Y$

indep.

$$\frac{X/r}{Y/m} \sim F \quad df_1 = r, \quad df_2 = m$$

$$\text{dist}(df_1, df_2)$$

ANOVA Table

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Within Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_B = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_B$

$n - k$

- $df_B = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$
- $df_E = 29 - 2 = 27$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares between groups, SSB Measures the variability between groups

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSB &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \\ &+ \left(10 \times (4.2 - 5.1)^2\right) \\ &= 16.96 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares total, SST Measures the variability between groups

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

where x_{ij} represent each observation in the dataset.

$\frac{SST}{n-1}$ = sample variance of all obs. data

$(x_{11}, \ldots, x_{2j}, \ldots, x_{3,n_3})$

pool all data together

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.13	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Sum of squares error, SSE Measures the variability within groups:

$$SSE = SST - SSB$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$SSE = 54.29 - 16.96 = 37.33$$

Mean square error Mean square error is calculated as sum of squares divided by the degrees of freedom.

$$MSB = 16.96/2 = 8.48$$
$$MSE = 37.33/27 = 1.38$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

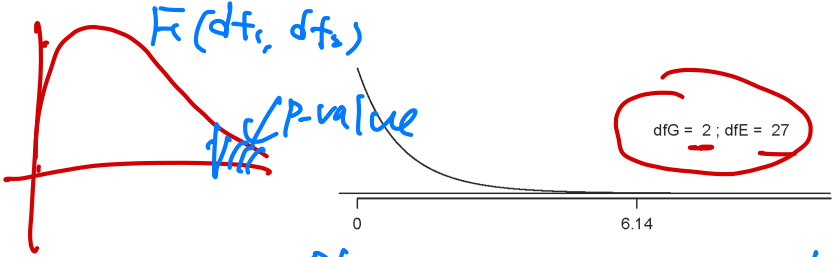
Test statistic, F value As we discussed before, the F statistic is the ratio of the between group and within group variability.

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

$$F = \frac{8.48}{1.38} = 6.14$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between)	depth	2	16.96	8.48	6.14	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

p-value p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



$P(F(3,5) > c) = \alpha \Leftrightarrow P(F(5,3) < \frac{1}{c}) = 1-\alpha$

Conclusion

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

(a) is different for all groups.
(b) on the surface is lower than the other levels.
(c) is different for at least one group.
(d) is the same for all groups.

Example 1

- A college administrator claims that there is no difference in first-year grade-point averages for students entering college from any of three different local high schools. The following data give the first-year grade-point averages of 15 randomly chosen students 5 from each of the three high schools. Are they strong enough, at the 5 percent level, to disprove the claim of the administrator?

School A	School B	School C
3.2	2.8	2.5
2.7	3.0	2.8
3.0	3.3	2.4
3.3	2.5	2.2
2.6	3.1	3.0

- This type of problem can be analyzed by one-factor ANOVA (or one-way ANOVA).

One-Way ANOVA

Notations:

- k groups to be compared
- n_i: the sample size for the i-th group
- x_{ij}: j-th observation in the i-th group
- $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ is the sample mean for the i-th group
- $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$ is the sample variance for the i-th group
- total sample size: $n \equiv \sum_{i=1}^{k} n_i$
- overall sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$
- overall sample variance:

$$\frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

SS total

- pooled variance estimate:

$$s_p^2 \equiv \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

→ SSE

n-k

(With-in) SS

Analysis Of Variance

Decomposition of total variation (SSTO):

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{x})\right)^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + 2\sum_{i=1}^{k}(\bar{x}_i-\bar{x})\underbrace{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)}_{0}$$

$$= \underbrace{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2}_{\text{between group variation}} + \underbrace{\sum_{i=1}^{k}(n_i-1)s_i^2}_{\text{within group variation}}$$

- SSTO = between-group variation (SSB) + within-group variation (SSW)
- SSW is also SSE (sum of squares for errors)

School A	School B	School C
3.2	2.8	2.5
2.7	3.0	2.8
3.0	3.3	2.4
3.3	2.5	2.2
2.6	3.1	3.0

Data x

GPA School

3.2 A

2.7 A

⋮

2.6 A

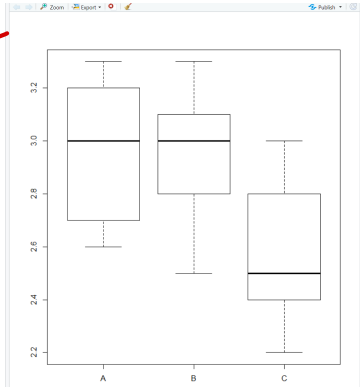
2.8 B

⋮

3.0 C

$n_1 = n_2 = n_3 = 5$

```
> aggregate(GPA~School, data=x, FUN=mean)
  School  GPA
1      A 2.96
2      B 2.94
3      C 2.58
> aggregate(GPA~School, data=x, FUN=var)
  School   GPA
1      A 0.093
2      B 0.093
3      C 0.102
> fit = lm(GPA~School, data=x)
> #summary(fit)
> anova(fit)   #ANOVA table
Analysis of Variance Table

Response: GPA
          Df  Sum Sq Mean Sq F value Pr(>F)
School     2 0.45733 0.22867  2.3819 0.1345
Residuals 12 1.15200 0.09600
> sum((x$GPA-mean(x$GPA))^2) #total variation
[1] 1.609333
```



ANOVA Table for GPA Data

Source of Variation	SS	df	MS	F value	p-value
Between group variation	0.4573	$3-1$	0.2287	2.3819	0.1345
Within group variation	1.1520	$15-3$	0.0960		
Total (corrected)	1.6093	$15-1$			

Residuals, Errors

- Given the data, SSTO is fixed. Therefore, SSB and SSW are compromised with each other.
- Larger SSB indicates larger deviation among group means and simultaneously smaller SSW.
- In GPA data, the F-test is not significant indicating that insufficient evidence to claim that the group means are not equal!

One-way ANOVA Model

- Model: $$X_{ij} \sim N(\mu_i, \sigma^2), \quad j = 1, 2, ..., n_i, \;\; i = 1, 2, ..., k,$$
  and X_{ij}'s are mutually independent.

$E X_{ij} = \mu_i$
$i = 1, 2, 3$

- The model can be equivalently expressed as
$$X_{ij} = \beta_0 + \beta_1 \cdot I(i = 2) + \beta_2 \cdot I(i = 3) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2),$$
  where the dummy variables are defined as
$$I(i = 2) = \begin{cases} 1, & \text{if } i = 2, \\ 0, & \text{otherwise,} \end{cases} \qquad I(i = 3) = \begin{cases} 1, & \text{if } i = 3, \\ 0, & \text{otherwise.} \end{cases}$$

$E(X_{ij}) = \beta_0 + \beta_1 I(i=2) + \beta_2 I(i=3)$

$E(X_{ij}) = \begin{cases} \beta_0, & i=1 \\ \beta_0 + \beta_1, & i=2 \\ \beta_0 + \beta_2, & i=3 \end{cases}$

- The intercept β_0 indicates the mean for the 1st group (reference group).
- relationship between 2 sets of parameters:
$$(\beta_0, \beta_1, \beta_2) = (\mu_1, \mu_2 - \mu_1, \mu_3 - \mu_1)$$

$(\mu_1 \; \mu_2 \; \mu_3) = (\beta_0, \beta_0 + \beta_1, \beta_0 + \beta_2)$

R Output for GPA Data

```
> fit = lm(GPA~School, data=x)
> summary(fit)

Call:
lm(formula = GPA ~ School, data = x)

Residuals:
   Min     1Q Median     3Q    Max
 -0.44  -0.22   0.04   0.23   0.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9600     0.1386  21.362 6.45e-11 ***
SchoolB      -0.0200     0.1960  -0.102    0.920
SchoolC      -0.3800     0.1960  -1.939    0.076 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3098 on 12 degrees of freedom
Multiple R-squared:  0.2842,    Adjusted R-squared:  0.1649
F-statistic: 2.382 on 2 and 12 DF,  p-value: 0.1345

> anova(fit)    #ANOVA table
Analysis of Variance Table

Response: GPA
          Df  Sum Sq Mean Sq F value Pr(>F)
School     2 0.45733 0.22867  2.3819 0.1345
Residuals 12 1.15200 0.09600
```

$H_0: \mu_1 = \mu_2$

$\beta_1 = 0$

2.96 A

2.9x B

2.58 C

μ_1

$\mu_2 - \mu_1$

$\mu_3 - \mu_1$

$H_0: \mu_1 = \mu_3$

$\beta_2 = 0$

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

From Summarized Statistics to ANOVA

An experiment aims to compare the effectiveness of three different fertilizers (A, B, C) on plant yield. The results are summarized as follows:

Fertilizer	Sample Size	Mean	Variance
A	5	10	12
B	6	8	15
C	7	12	20
Overall	18	10.11	17.34

Table: Summary of yield data for different fertilizers.

Source	Df	Sum Sq	Mean Sq	F value
Between				
Residuals				
Total				

Table: Analysis of Variance (ANOVA)

From Summarized Statistics to ANOVA

An experiment aims to compare the effectiveness of three different fertilizers (A, B, C) on plant yield. The results are summarized as follows:

Fertilizer	Sample Size	Mean	Variance
A	5	10	12
B	6	8	15
C	7	12	20
Overall	18	10.11	17.34

Table: Summary of yield data for different fertilizers.

Source	Df	Sum Sq	Mean Sq	F value
Between	2	51.8	25.9	1.6
Residuals	15	243	16.2	
Total	17	294.8		

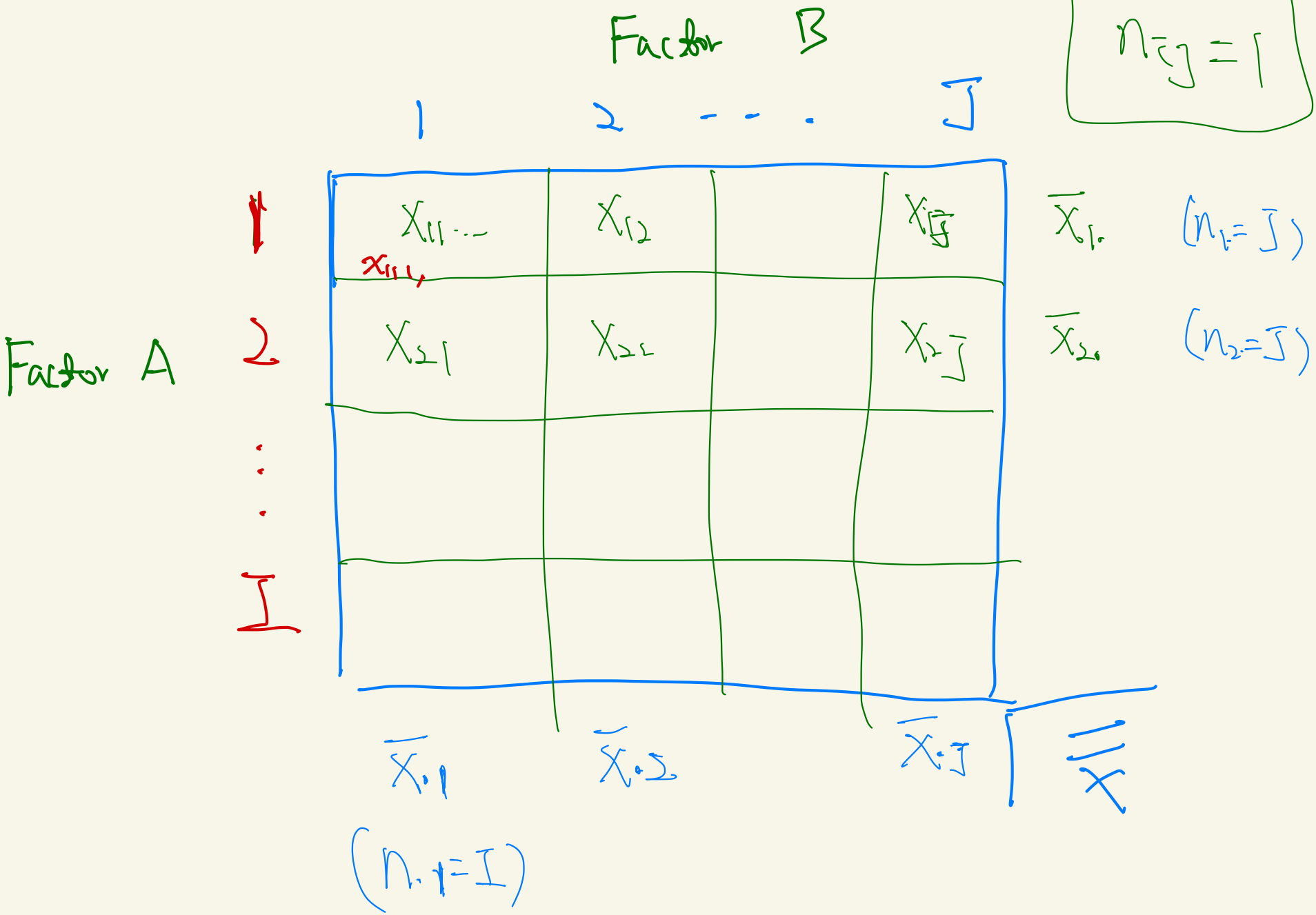
Table: Analysis of Variance (ANOVA)

Two-way ANOVA

Notations for Two-Way ANOVA

- groups are determined by 2 factors, indicated by (i, j), $i = 1, 2, ..., I$ and $J = 1, 2, ..., J$
- n_{ij}: the replication for the (i, j)-th group
- X_{ijk}: k-th observation in the (i, j)-th group, $k = 1, 2, ..., n_{ij}$
- $\bar{X}_{ij\cdot} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$ is the sample mean for the (i, j)-th group
- $n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$
- $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$
- $n_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$
- $\bar{X}_{i\cdot\cdot} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} X_{ijk}$ is the sample average over j and k, for each i
- $\bar{X}_{\cdot j\cdot} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{I} \sum_{k=1}^{n_{ij}} X_{ijk}$ is the sample average over i and k, for each j
- $\bar{X}_{\cdot\cdot\cdot} = \frac{1}{n_{\cdot\cdot}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} X_{ijk}$ is the sample average over all data
- Very often $n_{ij} = 1$, in this case k and the third "dot" is dropped.

$n_{ij} = 1$



$n_{ij} > 1$: $x_{111}, x_{112}, \ldots x_{11,n_{11}}$

Variance Decomposition for 2-way ANOVA

For simplicity, the following decomposition is for $n_{ij} = 1$ (k is dropped).

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{..})^2$$

overall mean

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\Big((X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}) + (\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..})\Big)^2$$

$$= \underbrace{\sum_{i=1}^{I} n_{i.}(\bar{X}_{i.} - \bar{X}_{..})^2}_{\text{SS due to Factor 1}} + \underbrace{\sum_{j=1}^{J} n_{.j}(\bar{X}_{.j} - \bar{X}_{..})^2}_{\text{SS due to Factor 2}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2}_{\text{SSE}}$$

- SSTO = SS(Factor 1) + SS(Factor 2) + SSE
- Model: $X_{ij} = \underbrace{\alpha_i + \beta_j}_{\mu_{ij}} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \iff X_{ij} \sim N(\mu_{ij}, \sigma^2)$

Example 2

- An important consideration in deciding which database management system to employ is the mean time required to learn how to use the system. A test was designed involving three systems and four users. Each user took the following amount of time (in hours) in training with each system:

	User			
	1	2	3	4
System 1	20	23	18	17
System 2	20	21	17	16
System 3	28	26	23	22

The study goal is to infer the differences between systems and between users.

- This type of problem can be analyzed by 2-factor ANOVA (or 2-way ANOVA).

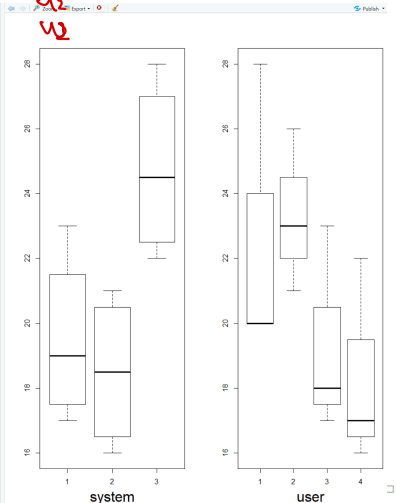
	User			
	1	2	3	4
System 1	20	23	18	17
System 2	20	21	17	16
System 3	28	26	23	22

time sys user

20 s1 u1
20 s2 u1
28 s3 u1
23 s1 u2
21 s2 u2
s3 u2

```
> par(mfcol=c(1,2))
> boxplot(time~system, data=x, xlab="system", cex.lab=2)
> boxplot(time~user, data=x, xlab="user", cex.lab=2)
> fit1 = lm(time~system+user, data=x)
> anova(fit1)   #2-way ANOVA table
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value   Pr(>F)
system     2 90.167  45.083  41.615 0.000304 ***
user       3 54.250  18.083  16.692 0.002570 **
Residuals  6  6.500   1.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(fit1)

Call:
lm(formula = time ~ system + user, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25000 -0.18750  0.08333  0.08333  1.50000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.2500     0.7360  28.873 1.14e-07 ***
system2     -1.0000     0.7360  -1.359 0.223089
system3      5.2500     0.7360   7.133 0.000382 ***
user2        0.6667     0.8498   0.784 0.462605
user3       -3.3333     0.8498  -3.922 0.007781 **
user4       -4.3333     0.8498  -5.099 0.002224 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.041 on 6 degrees of freedom
Multiple R-squared:  0.9569,    Adjusted R-squared:  0.921
F-statistic: 26.66 on 5 and 6 DF,  p-value: 0.0004991
```

F(2,6) F(3,6)



```
> fit1<-lm(time~System,data=x)#one-way ANOVA table
> summary(fit1)

Call:
lm(formula = time ~ System, data = x)

Residuals:
   Min     1Q Median     3Q    Max
-2.750 -1.938 -0.500  1.750  3.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.500      1.299  15.011 1.12e-07 ***
System2       -1.000      1.837  -0.544   0.5994
System3        5.250      1.837   2.858   0.0188 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.598 on 9 degrees of freedom
Multiple R-squared:  0.5975,    Adjusted R-squared:  0.508
F-statistic: 6.679 on 2 and 9 DF,  p-value: 0.01666

> anova(fit1)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value  Pr(>F)
System     2 90.167  45.083   6.679 0.01666 *
Residuals  9 60.750   6.750
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

54.25+6.5

```
> fit2<-lm(time~User,data=x)#one-way ANOVA table
> summary(fit2)

Call:
lm(formula = time ~ User, data = x)

Residuals:
   Min     1Q Median     3Q    Max
-2.667 -2.333 -1.333  2.917  5.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.6667     2.0069  11.294  3.4e-06 ***
User2         0.6667     2.8382   0.235    0.820
User3        -3.3333     2.8382  -1.174    0.274
User4        -4.3333     2.8382  -1.527    0.165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.476 on 8 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.1193
F-statistic: 1.497 on 3 and 8 DF,  p-value: 0.2877

> anova(fit2)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value Pr(>F)
User       3 54.250  18.083  1.4966 0.2877
Residuals  8 96.667  12.083
```

90.167+6.5

```
Response: time
          Df Sum Sq Mean Sq F value   Pr(>F)
system     2 90.167  45.083  41.615 0.000304 ***
user       3 54.250  18.083  16.692 0.002570 **
Residuals  6  6.500   1.083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way ANOVA (for data with $n_{ij} = 1$)

Source of Variation	SS	df	MS	F value	p-value
Factor 1 (row)	SS_r	$I-1$	$MS_r = \frac{SS_r}{I-1}$	$F_r = \frac{MS_r}{MSE}$	$P(\cdot > F_r)$
Factor 2 (col)	SS_c	$J-1$	$MS_c = \frac{SS_c}{J-1}$	$F_c = \frac{MS_c}{MSE}$	$P(\cdot > F_c)$
Error	SSE	$(I-1)(J-1)$	$MSE = \frac{SSE}{(I-1)(J-1)}$		
Total (corrected)	SSTO	$IJ-1$			

$IJ-1-(I-1)-(J-1) = (I-1)(J-1)$

- Test for row effects with $H_0 : \alpha_1 = \cdots = \alpha_I$,

$$F_r \sim \mathcal{F}(I-1, (I-1)(J-1)), \quad \text{under } H_0$$

- Test for column effects with $H_0 : \beta_1 = \cdots = \beta_J$,

$$F_c \sim \mathcal{F}(J-1, (I-1)(J-1)), \quad \text{under } H_0$$

Assumptions on ANOVA Models

- 1-factor model: $X_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.
- 2-factor model: $X_{ij} = \mu_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$.
- Key assumptions:
  - data are independent
  - normal distributed
  - equal variance accross groups
- Can we check these assumptions?

Check Equal Variance Assumption Among k Groups

- A general model: $X_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_i^2)$
- Test for $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
- Define the deviation measure associated with group dispersion:

$$Y_{ij} \equiv X_{ij} - \bar{X}_i, \quad \text{for all } i, j.$$

- Levene's test: apply one-way ANOVA to the absoluate deviation Y_{ij}:

$$\begin{aligned} SSB &= \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2, \quad SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2, \\ F &= \frac{SSB/(k-1)}{SSE/(n-k)} = \frac{MSB}{MSE} \overset{H_0}{\sim} \mathcal{F}(k-1, n-k). \end{aligned}$$

- Equal variance assumption is rejected if $F > \mathcal{F}_{1-\alpha}(k-1, n-k)$.

Check Normality Assumption

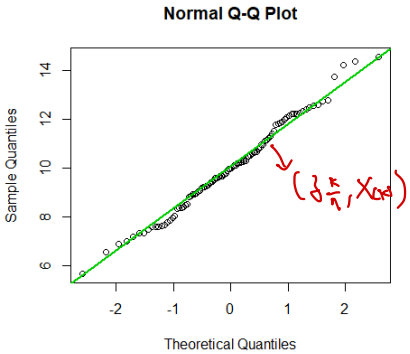
QQ (quantile-quantile) plot: a visualization method

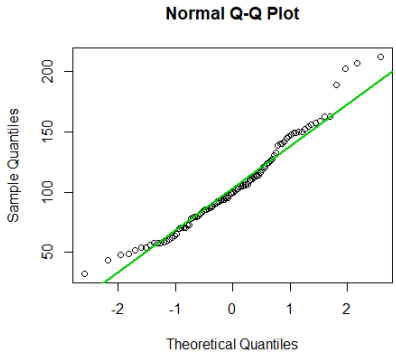
$\frac{x_q - \mu}{\sigma} = z_q$

- q-th quantile of $N(\mu, \sigma^2)$ satisfies $x_q = \mu + \sigma z_q$, where z_q is q-th quantile of N(0,1).
- Under normality, x_q should be a linear function of z_q.
- Ordered data: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
- Sample quantile estimated from the ordered data:
  $X_{(k)}$ corresponds to $\frac{k-0.5}{n}$-th sample quantile
- If data X_i's follow a normal distribution, we expect that the plot of $X_{(k)}$ v.s. N(0,1) quantiles $z_{(k-0.5)/n}$ behaves like a straight line.
- View this quantile-quantile plot (sample quantiles vs normal quantiles) to determine if the normality is a reasonable assumption.

QQ Plot for Normality Checking

Q: Which data set is more normal-like?





- A straight line indicates normality assumption is reasonable!

GoF

$\chi^2 = 14$ $(df = 10)$

If $\chi^2 \geq df$, cannot reject H_0

- F-stat $(df_1 = n_1,\ df_2 = n_2)$

$F = 4$. reject H_0 ?