Statistics II
Factorial ANOVA

Çağrı Çöltekin

ideas/examples/slides from
John Nerbonne & Hartmut Fitz

University of Groningen, Dept of Information Science

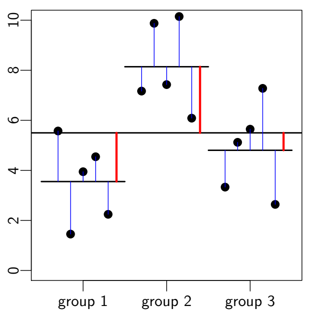
May 15, 2013

Single ANOVA: step by step

ANOVA is applicable when you have numeric observations on more than two independent groups.

- Collect your data: observations should be independent!
- Plot your data: typically, using box and whisker plots (box plots)
- Check for assumptions:
  - observations within each group should be approximately normal
  - the variances of the observations in each group should be approximately equal
- (optionally) set your prior contrasts
- calculate F and associated p-value (run ANOVA in a statistical software)
- (optionally) run pairwise comparisons between each group

Logic of ANOVA



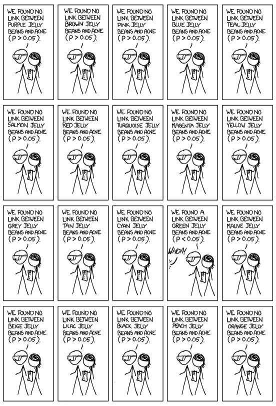
$$F = \frac{MS_{between}}{MS_{within}}$$

$$F = \frac{\frac{SS_{between}}{DF_{between}}}{\frac{SS_{within}}{DF_{within}}}$$

$$DF_{between} = k - 1$$
$$DF_{within} = n - k$$

where k is the number of groups, and n is the number of observations.

Why not multiple t-tests?



- Logic of hypothesis testing is based on obtaining a difference by chance.
- Finding a significant result at α-level 0.05 means that your result will be wrong with probability 0.05.
- 1 in 20 comparisons will cause you to get a significant result.
- If you need to do multiple comparisons, you need to adjust your α-level (for example using Bonferroni adjustment).

* From xkcd.com.

Single ANOVA: an example

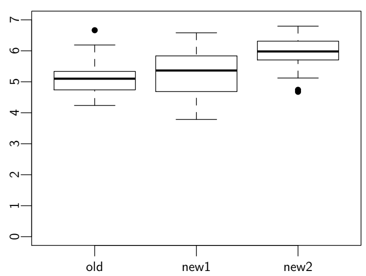
- We have two new web sites for a company.
- We want to know whether users prefer one of the new designs or the old design more.
- We do a survey on three groups of users, each answering questions on
  - Old web page.
  - New design 1.
  - New design 2.
- Our data is the average positive response between 1 and 7.

Example: data

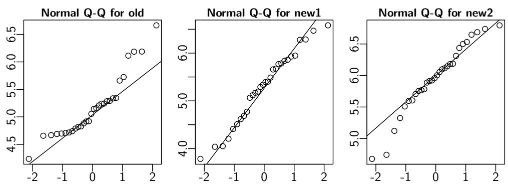
- The measurements we take are survey results 'opinion'.
- We have three groups which is represented with a categorical (factor) variable 'design' with three levels.

participant	design	opinion
1	old	5.1
2	old	4.7
⋮	⋮	⋮
21	new1	4.8
22	new1	5.8
⋮	⋮	⋮
59	new2	5.5
60	new2	6.1

Example: visualizing the data



Checking assumptions: normality



The plot for 'old' have deviances from the normal Q-Q line.

Alternatively you can also use a formal test like Shapiro–Wilk test or Kolmogorov–Smirnov test.

Checking assumptions: homogeneity of variance

- Box plots indicate that the variance of 'new 1' is larger than others. The variances are: $s^2_{old} = 0.31$, $s^2_{new1} = 0.57$, $s^2_{new2} = 0.29$.
- A common suggestion is to start worrying when ratio of any two variances are above 2.
- A few formal tests for equality of variance exists:

```
Levene's Test for Homogeneity of Variance (center = "mean")
       Df F value  Pr(>F)
group  2  2.6849 0.07388 .
       87
```

or

```
         Bartlett test of homogeneity of variances
Bartlett's K-squared = 4.218, df = 2, p-value = 0.1214
```


Example: results from software

```
Analysis of Variance Table
Response: ease
            Df Sum Sq Mean Sq F value    Pr(>F)
design     2 10.796  5.3978  13.955 5.541e-06 ***
Residuals 87 33.652  0.3868
```

- There is a significant effect (p-value is 0.0000055)
- but we do not know where the effect is.

Example: prior contrasts

- We have 3 groups, so we can specify 2 contrasts.
- Two interesting questions to ask:
  - Are new designs (on average) better than the old one?
  - Are the new designs different?

	Contrast 1	Contrast 2
old	−2	0
new1	1	-1
new2	1	1

- A contrast is orthogonal if columns sum to 0 and product of rows sum to 0.
- Orthogonal contrasts do not increase Type I errors.

Example: ANOVA with prior contrasts = linear regression

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.47094    0.06556  83.452  < 2e-16 ***
design1      0.15043    0.04636   3.245  0.00167 **
design2      0.33473    0.08029   4.169 7.21e-05 ***
Residual standard error: 0.6219 on 87 degrees of freedom
Multiple R-squared:  0.2429,    Adjusted R-squared:  0.2255
F-statistic: 13.95 on 2 and 87 DF,  p-value: 5.541e-06
```

- Main ANOVA result is the same ($p = 0.0000055$)
- First contrast 'desing1' indicates the difference between the old and the two new designs.
- Second contrast indicates the difference between the two new designs.

Example: post-hoc comparisons

If we do not have prior hypotheses, or our hypotheses cannot be expressed by planned contrasts, we can do post-hoc pairwise comparisons with corrections:

```
          old      new1
new1 1.00000  -
new2 1.3e-05  0.00022
P value adjustment method: bonferroni
```

or, a more powerful method (Bonferroni is too conservative):

```
old      new1
new1 0.46987 -
new2 1.3e-05 0.00014
P value adjustment method: holm
```


Factorial ANOVA

- Factorial ANOVA is used when there are more than one categorical variables (multiple factors, or grouping dimensions).
  - treatment and type of illness
  - impairment and gender
  - education and socio-economic status.
- Factorial (n-way) ANOVA follows essentially the same logic as single (one-way) ANOVA.

Example problems for Factorial ANOVA

- Compare time needed for lexical recognition in
  1. healthy adults
  2. patients with Wernicke's aphasia
  3. patients with Broca's aphasia

  and gender of the subject.
- Usability of an application based on different user interfaces and input methods.
- Language development of children based on their parent's education and socio-economic status.
- Compare Dutch proficiency scores of second language learners based on their native language and profession.

Why not multiple one-way ANOVAs?

- Efficiency: answer more questions with smaller sample size.
- Interactions: effects of different factors are not always additive.

Consider participants needed for a web site usability study. We want to choose between two designs, and two background colors

- Two one-way ANOVAs:

design 1	design 2
30	30

dark bg	light bg
30	30

Total participants needed: 120

- One two-way ANOVA:

	design 1	design 2
dark background	15	15
light background	15	15

Total participants needed: 60

Interactions

Interactions occur when change in one of the variables depends on the change in another.

- A particular treatment may have different effects on different illnesses.
- Living in big cities may increase life expectancy for people with low socio-economic status (SES), but may have no or reverse effect for people with higher SES.
- A new teaching method may be more effective with respect to the old one for girls but less effective for boys.

When there is an interaction, interpretation of main effects alone is incomplete and can be misleading.

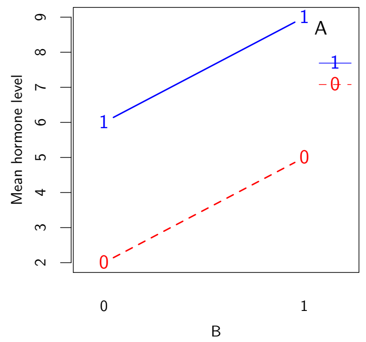
An example for interaction

Two drugs, A and B, are tested with a factorial design. Each drug is administered in doses 0 and 1.
In other words, four groups, receive none, A, B and A and B respectively.

		drug A	
		0	1
drug B	0	control	A only
	1	B only	A and B

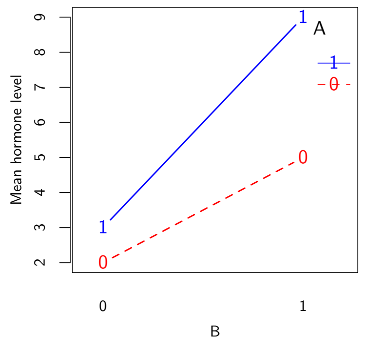
Response measure: blood level of some hormone.

Types of interaction (1)



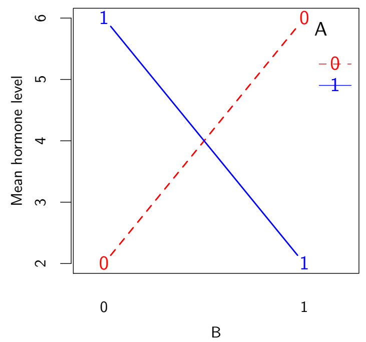
- both drugs have positive effects
- combined effect is additive
- no interaction

Types of interaction (2)



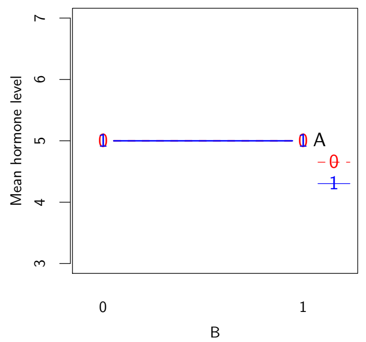
- both drugs have positive effects
- combined effect is stronger than sum of separate effects
- interaction

Types of interaction (3)



- both drugs have positive effects separately
- combination cancel out each other's effect
- interaction

Types of interaction (4)



- drugs show no effect
- either separately or in combination
- null hypothesis is true
- no interaction

ANOVA: partitioning the variance

In single ANOVA, we partition the total variance (SS_T) as variance due to group means (or, due to the groups, or the model SS_M) and the variance around the group means (or, residual variance, SS_R).

$$SS_T = SS_M + SS_R$$

The F-test used in single ANOVA is based on,

$$F = \frac{MS_M}{MS_R}$$

Associated degrees of freedom, for n observations, and k groups, are:

$$\begin{aligned} DF_T &= DF_M + DF_R \\ n-1 &= k-1 + n-k \end{aligned}$$

Factorial ANOVA: partitioning the variance

Factorial ANOVA partitions the SS_M further.

- For two-way ANOVA, with factors A and B, SS_M is partitioned as:

$$SS_M = \underbrace{SS_A + SS_B}_{\text{main effects}} + \underbrace{SS_{A\times B}}_{\text{interaction}}$$

- For three-way ANOVA, with factors A, B and C, SS_M is partitioned as:

$$SS_M = \underbrace{SS_A + SS_B + SS_C}_{\text{main effects}} + \underbrace{SS_{A\times B} + SS_{A\times C} + SS_{B\times C}}_{\text{2-way interctions}} + \underbrace{SS_{A\times B\times C}}_{\text{3-way inter.}}$$

Factorial ANOVA: degrees of freedom

As in single ANOVA:

$$\begin{array}{ccccc} DF_T & = & DF_M & + & DF_R \\ n-1 & = & k-1 & + & n-k \end{array}$$

If we have k_A levels due to factor A, and k_B levels due to factor B, total number of groups is $k = k_A \times k_B$. We can now further partition the DF_M as,

$$\begin{array}{ccccccc} DF_M & = & DF_A & + & DF_B & + & DF_{A\times B} \\ k-1 & = & k_A - 1 & + & k_B - 1 & + & (k_A - 1) \times (k_B - 1) \end{array}$$

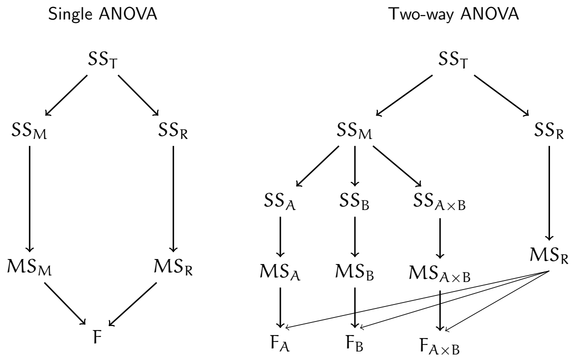
Factorial ANOVA: degrees of freedom

Once we have calculated sums of squares, and degrees of freedom values, we can calculate the estimated variance (mean squares) for each component as $MS = \frac{SS}{DF}$.
For two-way ANOVA we will get three F-tests:

$$\begin{array}{ccc} F_A & = & \frac{MS_A}{MS_R} \\ F_B & = & \frac{MS_B}{MS_R} \\ F_{A\times B} & = & \frac{MS_{A\times B}}{MS_R} \end{array}$$

For three-way ANOVA there will be 7 F-tests (three main effects, three two-way interactions and one three-way interaction).

Factorial ANOVA: the picture



Factorial ANOVA: an example

We return to our 'web design' example.

- We have two new web page designs.
- We also want know the effect of dark or light background.
- This is a two-way ANOVA with two levels at each dimension: commonly called 2×2 (experiment) design.
- If we also wanted to know the effect of age (young, middle aged, old), we would do a three-way, 2×2×3, ANOVA.

Example: participants

We gather a random sample of 60 people from our target audience, and randomly assign equal number of participants to one of the following groups (15 in each):

		Design	
		1	2
BG color	light	design 1, light BG	design 2, light BG
	dark	design 1, dark BG	design 2, dark BG

The response is the average opinion of each participant assessed through a 7-point questionnaire with multiple questions.

Example: data

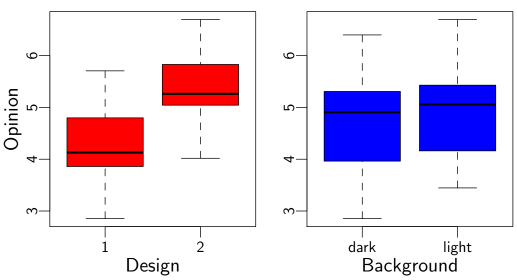
We have a numeric response variable (opinion) and two categorical variables (design and background color), both with two levels.

participant	opinion	design	background
1	6.2	1	light
2	5.8	1	dark
⋮	⋮	⋮	⋮
59	4.8	2	light
60	6.4	2	dark

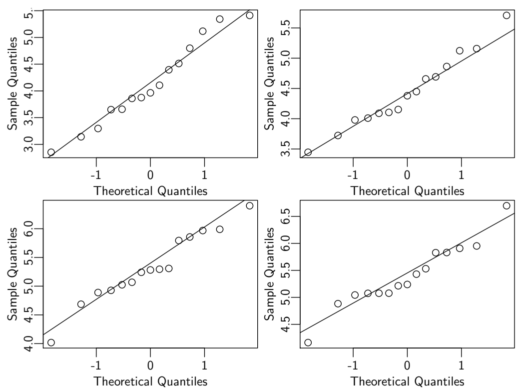
Important:

- participants are randomly selected and randomly assigned to a combination of design and background color
- each participant provides a single observation

Example: visualizing the data



Example: checking for normality



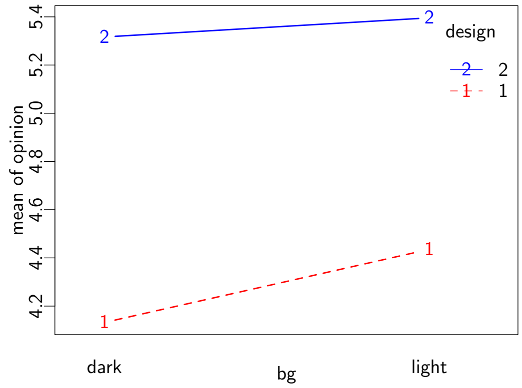
Example: checking homogeneity of variance

- Box plots already indicate that the variances are similar.
- Two possible tests for homogeneity of variance (output is from R, In SPSS enable the option in ANOVA dialog):

```
> leveneTest(opinion~design*bg)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.6342 0.5961
      56
> bartlett.test(opinion~design*bg)
        Bartlett test of homogeneity of variances
data:  opinion by design by bg
Bartlett's K-squared = 0.9009, df = 1, p-value = 0.3426
```

- Both tests support the visual inspection. No significant evidence of non-homogeneity.

Example: visualizing the interaction



Example: the two-way ANOVA

```
Analysis of Variance Table
Response: opinion
          Df  Sum Sq Mean Sq F value    Pr(>F)
design     1 17.2513 17.2513 40.5053 3.859e-08 ***
bg         1  0.5493  0.5493  1.2898    0.2609
design:bg  1  0.1880  0.1880  0.4414    0.5092
Residuals 56 23.8506  0.4259
```

- 'design' has a significant effect.
- the background color does not have significant effect.
- there is no evidence for interaction.

Example: post-hoc comparisons after factorial ANOVA

R output for multiple comparisons using "Tukey's Honest Significant Differences":

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
$design
        diff       lwr      upr p adj
2-1 1.072422 0.7348684 1.409976     0
$bg
                diff        lwr       upr     p adj
light-dark 0.1913667 -0.1461872 0.5289206 0.2609272
$`design:bg`
                       diff        lwr        upr     p adj
2:dark-1:dark    1.18437358  0.5533807  1.8153665 0.0000388
1:light-1:dark   0.30331806 -0.3276748  0.9343109 0.5838209
2:light-1:dark   1.26378895  0.6327961  1.8947818 0.0000117
1:light-2:dark  -0.88105552 -1.5120484 -0.2500626 0.0027282
2:light-2:dark   0.07941537 -0.5515775  0.7104083 0.9871044
2:light-1:light  0.96047089  0.3294780  1.5914638 0.0009520
```


(factorial) ANOVA and effect size

Simplest form of effect size for ANOVA is called η^2 (eta-squared). η^2 is equivalent to r^2 for regression.

$$\eta^2 = \frac{SS_M}{SS_T}$$

For factorial ANOVA, we can calculate partial-η^2 for each grouping variable.

$$\eta^2_A = \frac{SS_A}{SS_A + SS_R}$$

Like r^2, η^2 increases as number of levels/factors increase. An adjusted effect size measure, called ω^2 (omega-squared), corrects for chance increase caused by additional factor levels. Statistical software (typically) will give you both numbers.

(factorial) ANOVA and effect size (2)

Another alternative is using effect size measure for t-test (Cohen's d) for pairwise comparisons.

In general, for standardized effect size measures, the rule of thumb for interpretation is,

less than 0.1	weak effect
between 0.1 and 0.6	medium-size effect
greater than 0.6	large effect

Effect sizes are best interpreted with considering the particular problem at hand. For example, obtaining small effect sizes may be important in some problems.

Factorial ANOVA: summary

- Factorial ANOVA is a generalization of single ANOVA (or t-test).
- Compare groups along more than one dimension.
- Assumptions: the response variable in all groups
  - is (approximately) normally distributed
  - have (approximately) equal variances
- Efficient in use of subjects.
- Allows to investigate interaction.

Next week: Repeated-measures ANOVA (reading: chapter 13).