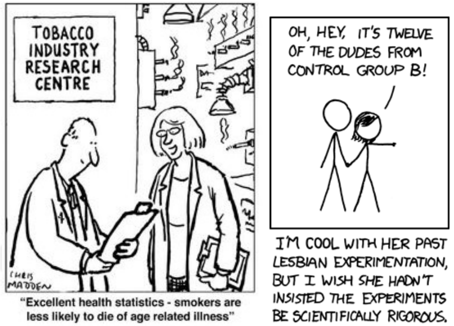
EMAT30007 Applied Statistics Lecture 9: Experimental Design and ANOVA

Ksenia Shalonova & Nikolai Bode

Observational versus experimental data

- Recap: statistical analysis with linear models. Selecting the right model is a key challenge. Ideally, data collection and model building go hand in hand.
- We can analyse *observational data* (data observed in natural setting) or *experimental data* (we control the explanatory variables). The former finds correlations, the latter can establish causal links.



Designing an experiment

Fact

The study of experimental design originated with R.A. Fisher's work in the UK in the 1900s.

- In an *experiment*, we collect data in a structured way. We decide on:
  - what to measure — *response*
  - what to measure the response on — *experimental unit*
  - the independent variables whose effect we study — *factors*
  - the combinations of factors we want to study — *treatments*
  - how many data points to collect — *sample size*
  - how to assign treatments to experimental units.
- Example: Compare the range of three electric car models. Range is the response, car model is a factor with three levels. With only one factor, there are no further treatments. Experimental units are cars and we can decide how many of each model we want to test.

Considerations in statistical experimental design

- The more data we collect, the more certain we can be about trends we find. However, collecting data is often expensive.
- Measurement errors, environmental variation or other effects lead to *noise* in data. *Noise-reducing* experimental designs can be used to counter this.
- Once we have identified factors we want to investigate, we have to think carefully about which treatments to test, to get the most out of our experiment (*volume-increasing design*).

Noise-reducing design

- *Noise-reducing designs* assign treatments to experimental units in such a way that extraneous noise is reduced.
- The simplest approach: *completely randomised design* — treatments assigned randomly to experimental units.
- Example: length of time to assemble a watch using three different methods A, B and C. Select 15 workers and assign them randomly to A, B or C (completely randomised design).
  But: assembly times could vary substantially between workers. This could skew our findings. To avoid this, we could get 5 workers to each use A, B and C in turn (*randomised block design*).
- In *randomised block designs*, we compare p treatments by using b blocks. Each block contains p relatively homogeneous (or identical) experimental units. The p treatments are assigned randomly to experimental units in each block (one experimental unit assigned per treatment).

Noise-reducing design

These experimental designs can be captured in linear models to investigate differences in the mean response across treatments. For the watch example:

- Model for completely randomised design: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$,

$$x_{1,i} = \begin{cases} 1 & \text{if worker } i \text{ uses method A} \\ 0 & \text{if not} \end{cases} \qquad x_{2,i} = \begin{cases} 1 & \text{if } i \text{ uses B} \\ 0 & \text{if not} \end{cases}$$

  ... selected method C as the base level.
- Model for randomised block design:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{1,i} + \beta_2 x_{2,i}}_{\text{treatment effects}} + \underbrace{\beta_3 x_{3,i} + \beta_4 x_{4,i} + \beta_5 x_{5,i} + \beta_6 x_{6,i}}_{\text{block effects}} + \epsilon_i$$

  ... $x_{3,i}$ up to $x_{6,i}$ are dummy variables for which worker assembles.

We can use the usual methods to test hypothesis on our data (e.g. T-test for individual parameters; F-test, Likelihood-ratio test for nested models).

Volume-increasing design

- *Volume-increasing designs* combine factors in experiments into treatments that are maximally informative.
- Example: an electricity company wants to measure customer satisfaction for two levels of peak time price increase, x_1, and two different peak period lengths, x_2 (2 levels). How should the levels of factors x_1 and x_2 be combined into treatments?
  - Option 1: keep one factor fixed and vary the other ("one-at-a-time"). This is consistent with block designs. However, it misses interactions between the factors.
  - Option 2: consider all possible combinations of factor levels (*complete factorial design*). For this example, we call it a 2×2 factorial design.
- **Warning:** if many factors are tested, complete factorial designs require A LOT of treatments.

Volume-increasing design

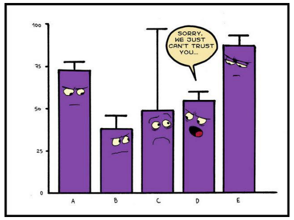
- *Complete factorial designs* can be captured in linear models with interaction terms. For the example on the previous slide, the model is:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{1,i} + \beta_2 x_{2,i}}_{\text{main effects}} + \underbrace{\beta_3 x_{1,i} x_{2,i}}_{\text{interaction}} + \epsilon_i$$

  ... $x_{1,i}$ and $x_{2,i}$ are dummy variables for peak time price increase levels and peak periond length levels, respectively.
- The number of parameters is the same as the number of treatments. This is always the case in complete factorial designs. Thus, we need *replicate measurements* for each treatment.
- Aside: this also works for quantitative predictors.

Selecting the sample size

- Deciding how many data points to collect is important:
  - On the one hand, the more data we have, the more certain we can be about observed trends (e.g. standard errors for parameter estimates in lecture 6).



  - On the other hand, collecting data is expensive, so we only want to collect what's necessary.
- *Power analysis* allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence.
- In general, the smaller the effect and the more confident we want to be, the more replicates we need.

University of BRISTOL
Applied Statistics 9
Spring Semester 2018

Typical steps in statistical experimental design

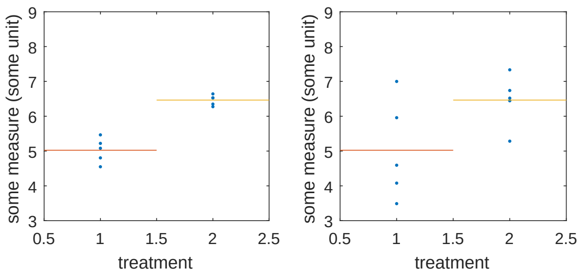
1. Select the factors.
2. Choose the treatments (factor level combinations; *Volume-increasing*).
3. Determine the sample size for each treatment.
4. Assign the treatments to the experimental units (*Noise-reducing*)

Often experiments combine volume-increasing and noise-reducing measures.

University of BRISTOL
Applied Statistics 9
Spring Semester 2018

Introduction to ANOVA

- *Analysis of variance (ANOVA)* is a statistical analysis for comparing means in experiments across different treatments.
- ANOVA is equivalent to analysing linear models. Before computers, ANOVA simplified calculations and it is still commonly used and referred to.
- Intuition for ANOVA: consider the variation within and between treatments.



- In ANOVA, we consider $F = \frac{\text{Between-treatment variation}}{\text{Within-treatment variation}}$

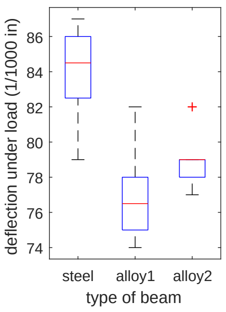
University of BRISTOL
Applied Statistics 9
Spring Semester 2018

One-way ANOVA

- Consider a one-factor completely randomised design, i.e. a number of p factor levels and experimental units assigned randomly to them.
- We have seen that this can be modelled as:
  $Y = \beta_0 + \beta_1 x_{1,i} + ... + \beta_{p-1} x_{p-1,i} + \epsilon_i$, where the $x_{j,i}$s are dummy variables for the factor levels.
- We wish to compare the means of the response, μ_j, across treatments j and test the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_p$.
- Suppose in the model above, treatment 1 is the base level, then: $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1, ... , \beta_{p-1} = \mu_p - \mu_1$
- So the H_0 above is equivalent to $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$.
- This is a *one-way ANOVA*. It is the same as an F-test on the corresponding linear model (lecture 7). It shows that at least two treatment means differ.
- The F-statistic is computed from sums of squared errors (no model fitting required).

One-way ANOVA

Example: a study on the strength of different structural beams (Hogg, 1987). The Matlab command is `anova1`.



ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	184.8	2	92.4	15.4	0.0002
Error	102	17	6		
Total	286.8	19			

...this suggests that at least two beams differ in strength.

University of BRISTOL
Applied Statistics 9
Spring Semester 2018

Two-way ANOVA

- Consider a complete factorial design with two factors, one of which has three and the other has two levels.
- We have seen that this can be modelled as: $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 w_{1,i} + \beta_4 x_{1,i} w_{1,i} + \beta_5 x_{2,i} w_{1,i} + \epsilon_i$, where $x_{1,i}$ and $x_{2,i}$ are dummy variables for the first factor and $w_{1,i}$ is a dummy variable for the second factor.
- Analogously to the one-way ANOVA, in *two-way ANOVA*, we perform a number of F-tests to compare the mean of the response across treatments.
- E.g., to test for interactions, we test $H_0 : \beta_4 = \beta_5 = 0$. Testing if the mean response for levels of the first factor are equal, requires $H_0 : \beta_1 = \beta_2 = 0$.
- Essentially, we use F-tests to compare *nested models* (a *restricted model* is obtained from a *full model* by setting some parameters to zero).
- These are tests on multiple parameters simultaneously (e.g. if factors have more than 2 levels). They can show that at least two treatment means differ.

University of BRISTOL
Applied Statistics 9
Spring Semester 2018

Two-way ANOVA

Example: a study on popcorn brands (columns, 3 levels: Gourment, National and Generic) and popcorn machine types (rows, 2 levels: oil- and air-based). A 3×2 factorial design. Measure cups of popped corn from a fixed amount (Hogg, 1987). In Matlab, use `anova2`.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	15.75	2	7.875	56.7	0
Rows	4.5	1	4.5	32.4	0.0001
Interaction	0.0833	2	0.04167	0.3	0.7462
Error	1.6667	12	0.13889		
Total	22	17			

...no evidence for an interaction, but separately popcorn machines and at least two popcorn brands differ. `anova2` includes interactions by default.

University of BRISTOL
Applied Statistics 9
Spring Semester 2018

ANOVA summary

- More complex models are possible — extend the models covered.
- Checking ANOVA assumptions is the same as for linear models (they are linear models!)
- ANOVA tells us that there is a difference in at least two means. Once an effect is detected, additional tests (e.g. t-tests) can be used for pairwise comparisons of means across treatments.
- **Warning:** if we compare A LOT of treatments, some p-values may be significant *by chance*. To account for such *multiple comparisons*:
  - Tukey test (v. brief description)
  - Bonferroni correction (re-scale p-values to account for number of comparisons).
- All ANOVA analysis is equivalent to using the linear model framewok and appropriate F-tests.

Analysis suggestion for a two-factor factorial experiment

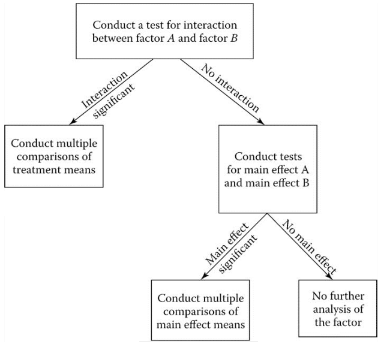


Diagram from: W.Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*, CRC Press, 2016