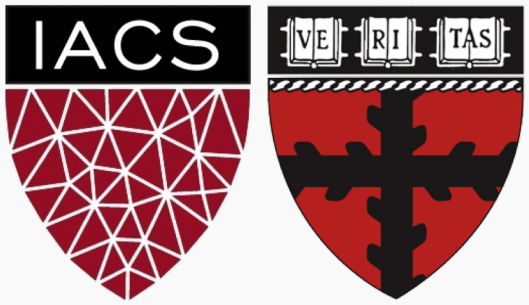
Lecture 23: AB Testing

CS109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader



Outline

- Causal Effects
- Experiments and *AB*-testing
- *t*-tests, binomial *z*-test, fisher exact test, oh my!
- Adaptive Experimental Design

Association vs. Causation

In many of our methods (regression, for example) we often want to measure the association between two variables: the response, *Y*, and the predictor, *X*. For example, this association is modeled by a β coefficient in regression, or amount of increase in R^2 in a regression tree associated with a predictor, etc...

If β is *significantly different* from zero (or amount of R^2 is greater than by chance alone), then there is evidence that the response is associated with the predictor.

How can we determine if β is *significantly different* from zero in a model?

Association vs. Causation (cont.)

But what can we say about a *causal association*? That is, can we manipulate *X* in order to influence *Y*?

Not necessarily. Why not?

There is potential for confounding factors to be the driving force for the observed association.

Controlling for confounding

How can we fix this issue of confounding variables?

There are 2 main approaches:

1. Model all possible confounders by including them into the model (multiple regression, for example).
2. An *experiment* can be performed where the scientist manipulates the levels of the predictor (now called the *treatment*) to see how this leads to changes in values of the response.

What are the advantages and disadvantages of each approach?

Controlling for confounding: advantages/disadvantages

1. Modeling the confounders

- Advantages: cheap
- Disadvantages: not all confounders may be measured.

2. Performing an experiment

- Advantages: confounders will be *balanced*, on average, across treatment groups
- Disadvantages: expensive, can be an artificial environment

Experiments and *AB*-testing

Completely Randomized Design

There are many ways to design an experiment, depending on the number of treatment types, number of treatment groups, how the treatment effect may vary across subgroups, etc...

The simplest type of experiment is called a Completely Randomized Design (CRD). If two treatments, call them treatment *A* and treatment *B*, are to be compared across *n* subjects, then *n*/2 subject are randomly assigned to each group.

- If $n = 100$, this is equivalent to putting all 100 names in a hat, and pulling 50 names out and assigning them to treatment *A*.

Experiments and *AB*-testing

In the world of Data Science, performing experiments to determine causation, like the completely randomized design, is called *AB*-testing.

AB-testing is often used in the tech industry to determine which form of website design (the treatment) leads to more ad clicks, purchases, etc... (the response). Or to determine the effect of a new app rollout (treatment) on revenue or usage (the response).

Assigning subject to treatments

In order to balance confounders, the subjects must be properly randomly assigned to the treatment groups, and sufficient enough sample sizes need to be used.

For a CRD with 2 treatment arms, how can this randomization be performed via a computer?

You can just sample $n/2$ numbers from the values 1, 2, ..., n without replacement and assign those individuals (in a list) to treatment group *A*, and the rest to treatments group *B*. This is equivalent to sorting the list of numbers, with the first half going to treatment *A* and the rest going to treatment *B*.

This is just like a 50-50 test-train split!

t-tests, binomial *z*-test, fisher exact test, oh my!

Analyzing the results

Just like in statistical/machine learning, the analysis of results for any experiment depends on the form of the response variable (categorical vs. quantitative), but also depends on the design of the experiment.

For *AB*-testing (classically called a 2-arm CRD), this ends up just being a 2-group comparison procedure, and depends on the form of the response variable (aka, if *Y* is binary, categorical, or quantitative).

Analyzing the results (cont.)

For those of you who have taken Stat 100/101/102/104/111/139:

If the response is quantitative, what is the classical approach to determining if the means are different in 2 independent groups?

- a 2-sample *t*-test for means

If the proportions of successes are different in 2 independent groups?

- a 2-sample *z*-test for proportions

2-sample *t*-test

Formally, the 2-sample *t*-test for the mean difference between 2 treatment groups is:

$H_0: \mu_A = \mu_B$ vs. $H_0: \mu_A \neq \mu_B$

$$t = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

The *p*-value can then be calculated based on a $t_{\min(n_A, n_B)-1}$ distribution.

The assumptions for this test include (i) independent observations and (ii) normally distributed responses within each group (or sufficiently large sample size).

2-sample *z*-test for proportions

Formally, the 2-sample z test for the difference in proportions between 2 treatment groups is:

$H_0: p_A = p_B$ vs. $H_0: p_A \neq p_B$

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_p(1-\hat{p}_p)\frac{1}{n_A} + \frac{1}{n_B}}}$$

where $\hat{p}_p = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A + n_B}$ is the overall 'pooled' proportion of successes.

The *p*-value can then be calculated based on a standard normal distribution.

Normal approximation to the binomial

The use of the standard normal here is based on the fact that the binomial distribution can be approximated by a normal, which is reliable when $np \geq 10$ and $n(1 - p) \geq 10$.

What is a Binomial distribution? Why can it be approximated well with a Normal distribution?

Summary of analyses for CRD Experiments

Variable Type	# Trt's	Classic Approach	Alternative Approach
Quantitative	2	t-test	Randomization test
	3+	ANOVA	
Binary	2	z-test	Fisher's exact test
	3+	χ^2 test	
Categorical (3+)	2+	χ^2 test	Fisher's exact test

The classical approaches are typically *parametric*, based on some underlying distributional assumptions of the individual data, and work well for large *n* (or if those assumptions are actually true). The alternative approaches are *nonparameteric* in that there is no assumptions of an underlying distribution, but they have slightly less power if assumptions are true and may take more time & care to calculate.

Analyses for CRD Experiments in Python

- *t*-test:
  `scipy.stats.ttest_ind`
- proportion *z*-test:
  `statsmodels.stats.proportion.proportions_ztest`
- ANOVA *F*-test:
  `scipy.stats.f_oneway`
- χ^2 test for independence:
  `scipy.stats.chi2_contingency`
- Fisher's exact test:
  `scipy.stats.fisher_exact`
- Randomization test: ???

ANOVA procedure

The classic approach to compare 3+ means is through the Analysis of Variance procedure (aka, ANOVA).

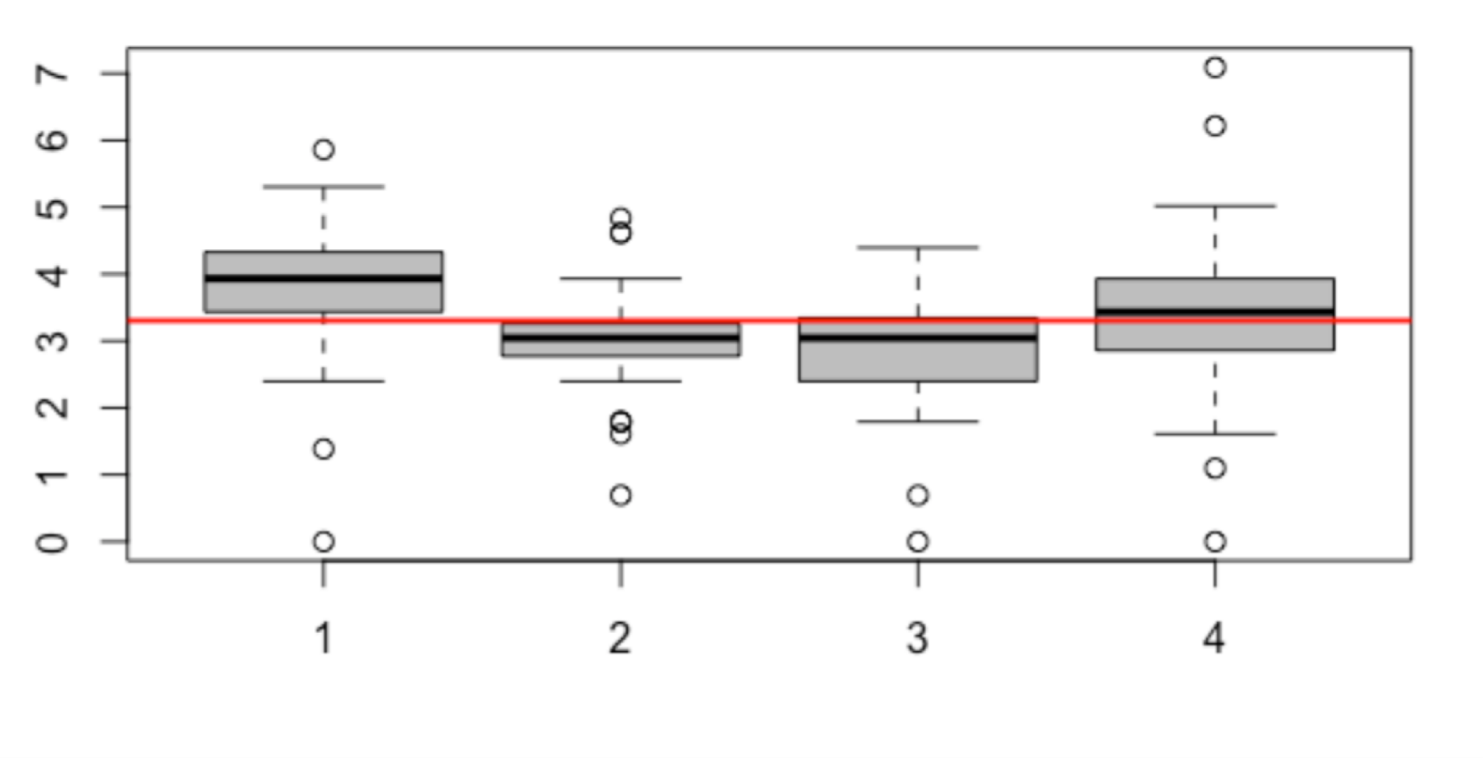
The ANOVA procedure's *F*-test is based on the decomposition of sums of squares in the response variable (which we have indirectly used before when calculating R^2).

$$SST = SSM + SSE$$

In this multi-group problem, it boils down to comparing how far the group means are from the overall grand mean (*SSM*) in comparison to how spread out the observations are from their respective group means (*SSE*).

A picture is worth a thousand words...

Boxplot to illustrate ANOVA



ANOVA *F*-test

Formally, the ANOVA F test for differences in means among 3+ groups can be calculated as follows:

H_0: the mean response is equal in all K treatment groups.
H_A: there is a difference in mean response somewhere among the treatment group.

$$F = \frac{\sum_{k=1}^{K} \frac{n_k(\bar{Y}_k - \bar{Y})^2}{(K-1)}}{\sum_{k=1}^{K} \frac{(n_k - 1)S_k^2}{(n-K)}}$$

where n_k is the sample size in treatment group k, $\bar{Y}_k$ is the mean response in treatment group k, S_k^2 is the variance of responses in treatment group k, $\bar{Y}$ is the overall mean response, and $n = \sum n_k$ is the total sample size.

The *p*-value can then be calculated based on a $F_{df_1=(K-1),df_2=(n-K)}$ distribution.

Comparing categorical variables

The classic approach to see if a categorical response variable is different between 2 or more groups is the χ^2 test for independence. A contingency table (we called it a confusion matrix) illustrates the idea:

Abortion Should be	Republican	Democrat	total
Legal	166	430	596
Illegal	366	345	711
Total	532	775	1307

If the two variables were independent, then:
$P(Y = 1 \cap X = 1) = P(Y = 1)P(X = 1)$.

How far the inner cell counts are from what they are expected to be under this condition is the basis for the test.

χ^2 test for independence

Formally, the χ^2 test for independence can be calculated as follows:

H_0: the 2 categorical variables are independent

H_A: the 2 categorical variables are not independent (response depends on the treatment).

$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* is the observed cell count and *Exp* is the expected cell count:

$Exp = \frac{(row\ total) \times (column\ total)}{n}$.

The *p*-value can then be calculated based on a $\chi^2_{df=(r-1)\times(c-1)}$ distribution (*r* is the # categories for the row var., *c* is the # categories for the column var.).

Randomization test

A randomization test is the non-parametric approach to analyzing quantitative data in an experiment. It is an example of a *resampling* approach (the bootstrap is another resampling approach).

The basic assumption of the randomization test is that if the treatments are truly the same, then the measured response variable, Y_i, for subject *i* would not change if that subject was instead randomly assigned to a different treatment. This is sometimes called *exchangeability*.

Randomization test (cont.)

So to analyze the results, we re-randomize the individuals to treatment through simulation (keeping the sample sizes the same). We then re-calculate the statistic of interest (difference in 2 sample means or sums of squares between 3+ groups) many-many times and build a histogram of the results. This histogram is then used as the reference distribution to determine how extreme our actual observed result is.

This approach is also called a permutation test, since we are re-permuting each of the subjects into the treatment groups (and then assume this has no bearing on the response).

Fisher's exact test

R.A. Fisher also came up with what is known as Fisher's exact test.

This analysis approach is useful for a contingency table, and does not need to rely on large sample size.

It fixes the row and column totals, and then determines all the ways in which the inner cells can be calculated given those row and column totals.

The probability of any of these filled out tables, given the row and column totals is fixed, is then based on a hypergeometric distribution.

Then the possible filled out tables that are less likely to occur than what was actually observed contribute to the *p*-value (by adding up their probabilities).

Fisher's exact test

Abortion Should be	Republican	Democrat	total
Legal	166	430	596
Illegal	366	345	711
Total	532	775	1307

$$P(X_1 = 166) = \frac{\binom{596}{166}\binom{711}{366}}{\binom{1307}{532}} = 1.33\times10^{-18}$$

Then a similar calculation is done for all possible values of X_1, and these probabilities are summed up for those cases of X_1 that are not more likely to occur.

The app update roll-out problem

A company is interested in updating their app/program, so they start a 'pilot program' to test the waters to see how this update will affect some important measure (like revenue or usage). How should they do this?

They select a sample of users and ask them to voluntarily update the app on their phones in order to estimate the affect of this update.

Any issues with this design?

Volunteers will always be the most excited, dedicated users: a biased sample from all of their users.

We can potentially check for this bias via a χ^2 test for goodness-of-fit.

χ^2 test for goodness-of-fit

Formally, the χ^2 test for goodness-of-fit can be calculated as follows:

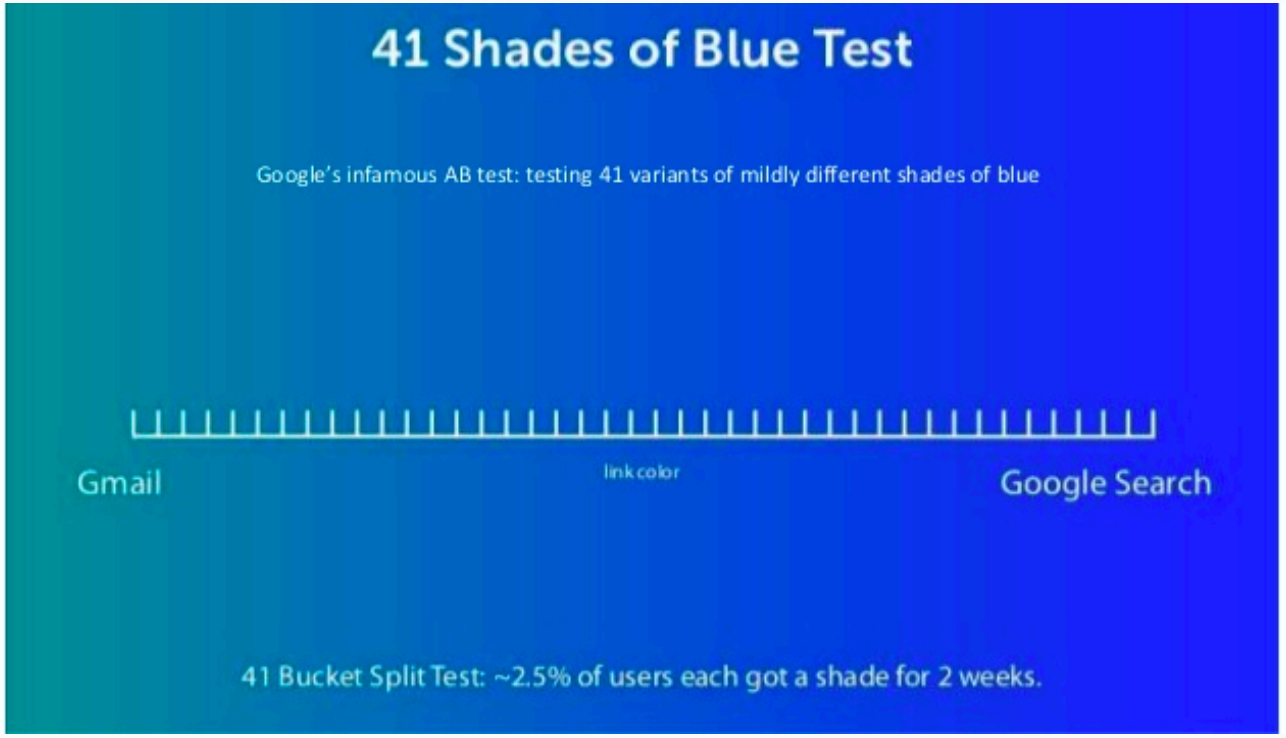
H_0: the variable follows some known distribution in the population
H_A: the variable does not follow this distribution

$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* is the observed cell count and *Exp* is the expected cell count: $Exp_i = n\pi_i$ (π_i is the theoretical probability of being in category/bucket *i*).

The *p*-value can then be calculated based on a $\chi^2_{df=(k-1)}$ distribution (*k* is the # categories in the population).

An infamous AB Test: 41 Shades of Blue



How should the study proceed? How should the data be analyzed?

Adaptive Experimental Design

Beyond CRD designs

The approaches we have seen to experiments all rely on the completely randomized design (CRD) approach. There are many extensions to the CRD approach depending on the setting. For example:

- If there are more than two types of treatments (for example: (i) font type and (ii) old vs. new layout), then a *factorial* approach can be used to test both types of treatments at the same time.
- If the treatment effect is expected to be different across different subgroups (for example possibly different for men vs. women), then a stratified/cluster randomized design should be used.

Beyond CRD designs (cont.)

These different experimental designs will need to have adjusted analysis approaches to analyze them appropriately.

Examples:

1. **factorial design:** a multi-way ANOVA when the response variable is quantitative.
2. **stratified analysis**: the Mantel-Haenszel test for cluster randomized design with a categorical response variable.

Beyond CRD designs (cont.)

But all of these procedures rely on the fact that there is a fixed sample size for the experiment. This has a glaring limitation: you have to wait to analyze until *n* is recruited/reached.

If you peak at the results before *n* is reached, then this is a form of *multiple comparisons* and thus overall Type I error rate is inflated.

Bandit Designs

A *sequential* or *adaptive* procedure can be used if you would like to intermittently check the results as subjects are recruited (or want to look at the results after each and every new subject is enrolled).

One example of a sequential test/procedure is a *bandit-armed* design. In this design, after a burn-in period based on a CRD, then the treatment that is performing better is chosen more often to be administered to the subjects.

Bandit Design Example

For example, in the *play the winner* approach for a binary outcome, if treatment *A* is successful for a subject, then you continue to administer this treatment to the next subject until it fails, and then you skip to treatment *B*, and vice versa.

The advantage to this approach is that if one treatment is truly better, then the number of subjects exposed to the worse treatment is lessened.

What is a major disadvantage?

Bayesian Bandit Designs

Our friend Bayes' theorem comes into play again if we would like to have a bandit design for a quantitative outcome.

The randomization to treatment for each subject is based on a biased coin, where the probability of being assigned to treatment *A* is based on the poster probability that treatment *A* is a better treatment.

Bayesian Bandit Designs (cont.)

This probability can be calculated based on the Bayes theorem as follows:

$$P\left(\mu_{Y|trt_A} > \mu_{Y|trt_B} \middle| Data\right)$$
$$\propto P\left(Data \middle| \mu_{Y|trt_A} > \mu_{Y|trt_B}\right) P\left(\mu_{Y|trt_A} > \mu_{Y|trt_B}\right)$$

where $P(\mu_{Y|trt_A} > \mu_{Y|trt_B})$ is the prior belief (can be set to 0.5).

This can easily extend to more than just 2 treatment groups.

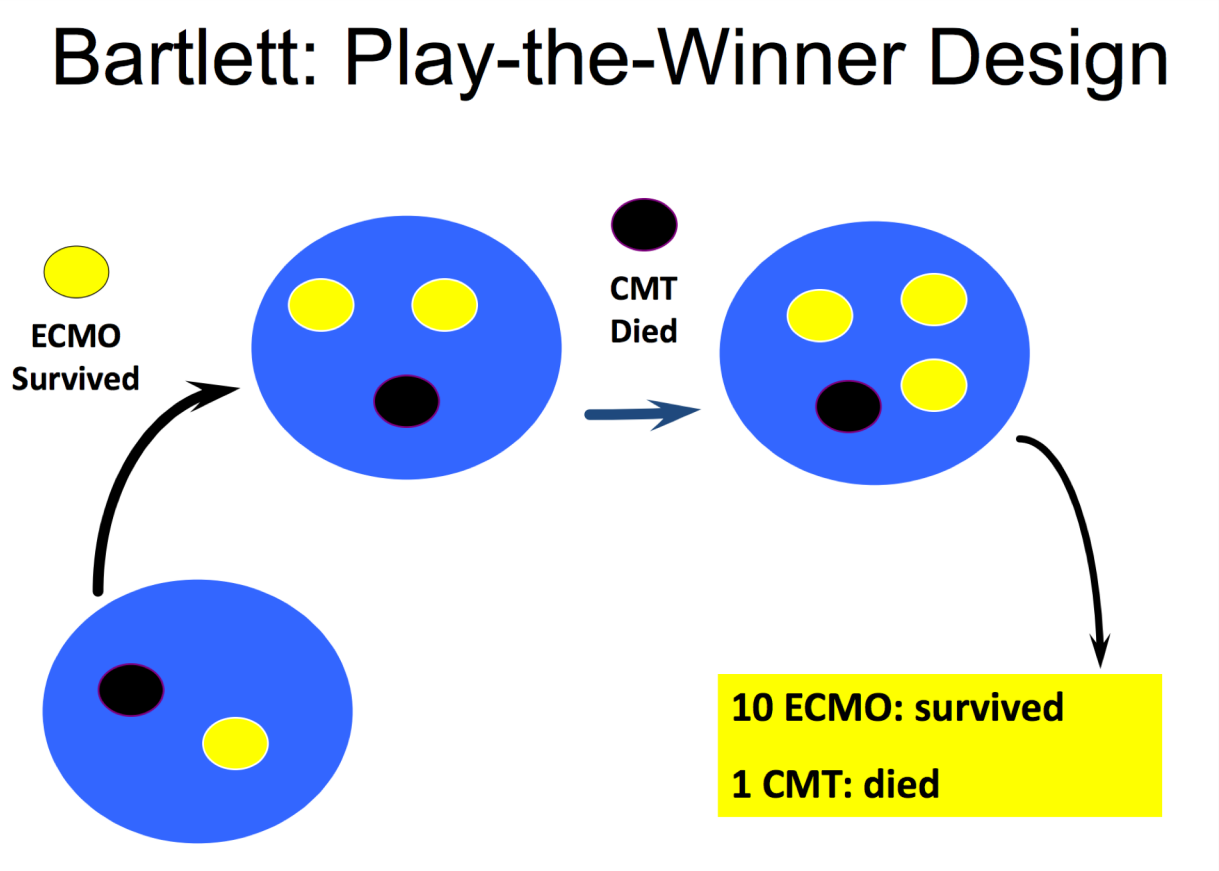
ECMO Trial: Bayesian Bandit Trial Example

In the 80's a bandit-armed design (Bartlett, et al.) was used to determine whether or not Extracorporeal Membrane Oxygenation (ECMO) would improve survival (compared to 'standard of care') of neonatal patients (premature babies) experiencing respiratory failure.

In the end, only 11 patients were enrolled before "statistical significance" was achieved.

What is an issue with these results?

ECMO Trial


Bartlett: Play-the-Winner Design
ECMO
Survived
CMT
Died
10 ECMO: survived
1 CMT: died

Analysis of Bayesian

So when should you stop an adaptively designed trial?

You could continue the trial until a *p*-value of less than 0.05 is achieved (or until a large sample size is taken without coming to a statistically significant result)?

What is an issue with this "stopping criterion"?

If our *p*-value is determined from a classical method, then this is an example of multiple comparisons: you have looked at the data at many points along the timeline, so a significant result is more likely to occur than 0.05 if there is not a *true* difference in the treatments.

We need to adjust how the 'statistical significance' is determined!