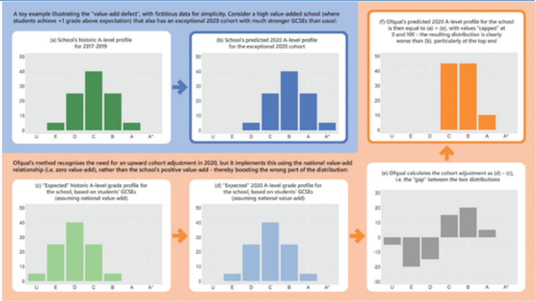
Methods of Applied Statistics I

STA2101H F LEC9101

Week 5

October 8 2020



Today

Start Recording

1. Second syllabus update; Two editions of Faraway
2. In the News: A-levels; Excel; 538
3. Designed Experiments SM 9.1, 9.2; FLM-2 Ch 16, 17; FLM-1 Ch 15, 16
4. Preliminary Analysis – CD Ch. 5
5. (2–3pm) HW1; Reading Statistical Models

Happy Thanksgiving!!!

Monday, Oct 12

No office hours

http://www.utstat.utoronto.ca/reid/sta2101f/syllabus20Update-2.pdf

Syllabus **updated Oct 5** STA 2101F: Methods of Applied Statistics I

Week	Date	Methods	References	Computing
1	Sept 10	Review of Linear Regression	SM Ch.8.2.1, 8.3; FLM-2 Ch.2-4; FLM-1 Ch.2-3; CD Ch.1	RStudio and RMarkdown
2	Sept 17	~~Model Selection~~ Comparing models; factors; model checking; diagnostics; collinearity	SM Ch.8.5,6; FLM Ch.3; FLM-2 14.1, 14.2, 2.11, 2.6; FLM-1 4,13; CD Ch.6	~~tidyverse~~
3→HW1	Sept 24	~~Random and Mixed Effects Models~~Model selection; Types of studies	SM 8.7.1; FLM-2 Ch. 10; FLM-1 Ch.8; CD Ch.1,2	~~ggplot~~ HW 1 Qs
4←HW1	Oct 1	~~Designed Experiments~~ Factor variables; Random and Mixed Effects; Principles of Measurement	SM Ch. ~~9.1~~,9.2.1; FLM-2 Ch.14-17;FLM-1 Ch.14-16; CD Ch.4	as.factor, is.factor, ggplot, anova, fruitfly data
5	Oct 8	~~Binary Responses~~Designed Experiments; Preliminary Analysis	SM Ch.9.1,2; FLM-2 Ch.14, 15 FLM-1 Ch.13, 14; ~~Ch.2~~; CD Ch.5, FLM-2 Ch.5	
6	Oct 15	Logistic Regression	SM 10.6.1; FELM Ch.3	
7→HW2	Oct 22	Generalized Linear Models	FELM Ch.6,7; SM 10.3	

In the News: A-level grades (England)

A-level and GCSE results: Pressure mounts on ministers to solve exam crisis

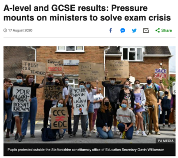
17 August 2020

Share

In the News: A-level grades (England)

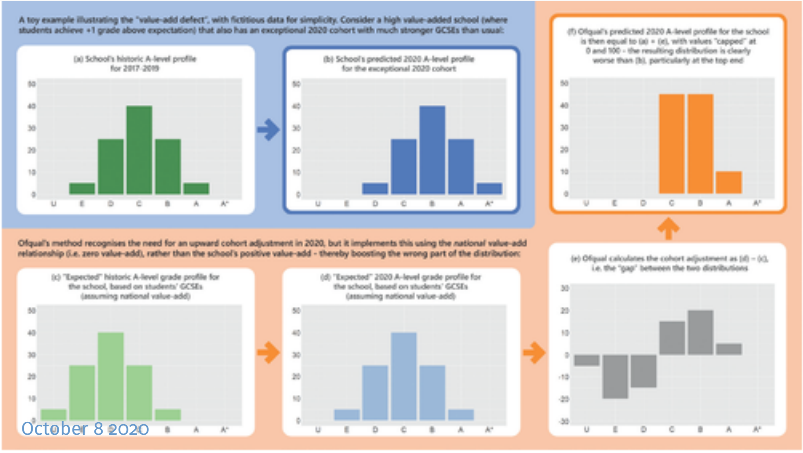
Signficance, Oct issue

- assignment of final grades in high school, used for university admissions Introduction
- in the absence of written exams Full article
- “exam boards would be asking teachers in schools and colleges to submit expected grades and rankings of their students in lieu of exams”
- “These assessments would then go through ‘a process of standardisation using a model’, or algorithm, that Ofqual had developed”
- “Ultimately, when grades were issued on 13 August, some students found the results to be anything but fair, with many receiving lower marks than expected”
- “on 17 August, after student protests, Ofqual abandoned the calculated grades in favour of teacher-assessed grades.”



… A-level grades

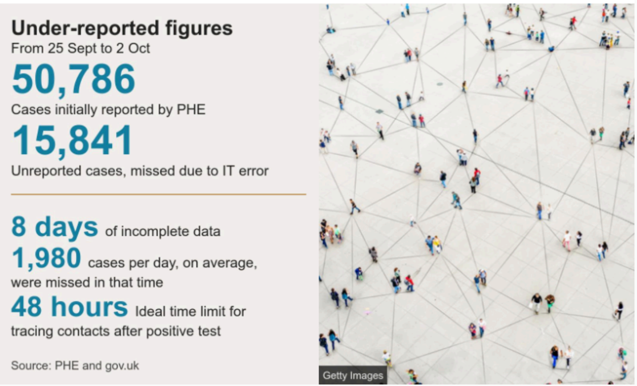
"the algorithm developed by England's exam regulator, Ofqual, contained no elements of machine learning, or indeed any artificial intelligence – it was simply a sequence of hand-coded procedural steps designed by an indisputably human team."



Also in the news: Covid testing

- The health secretary said that a technical glitch that saw nearly 16,000 Covid-19 cases go unreported in England “should never have happened”
- Excess rows in the database ignored by Excel software

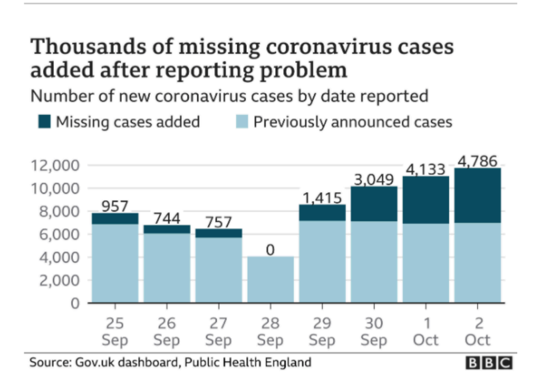
BBC News



Also in the news: Covid testing

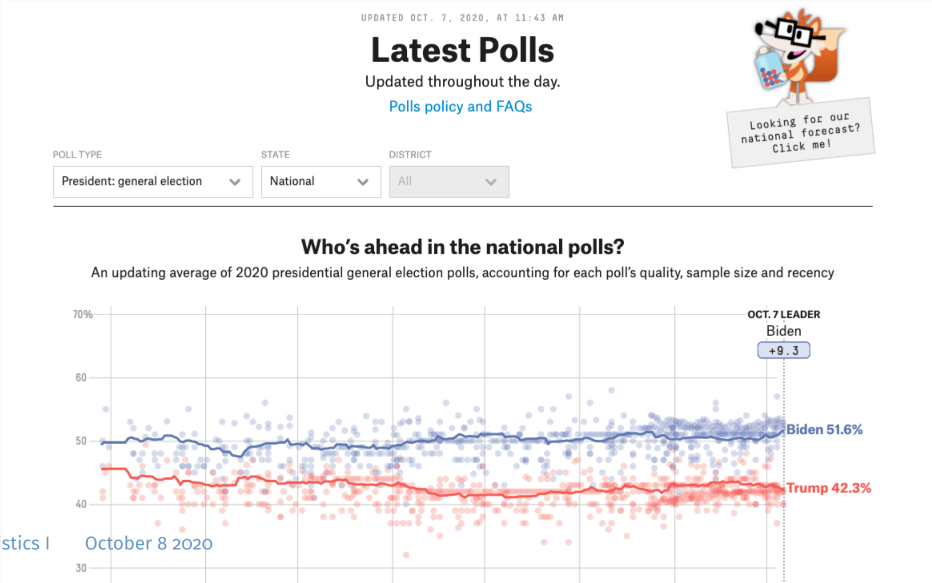
- The health secretary said that a technical glitch that saw nearly 16,000 Covid-19 cases go unreported in England “should never have happened”
- Excess rows in the databased ignored by Excel software

BBC News



Also in the news

538.com

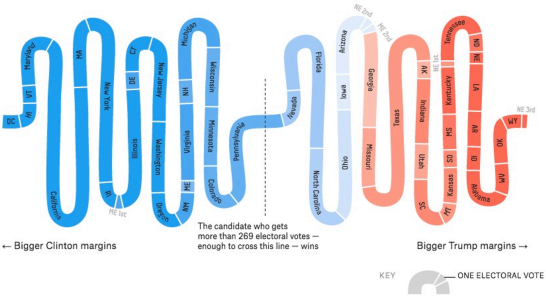


Visualization

fivethirtyeight.com

The winding path to 270 electoral votes

A candidate needs at least 270 electoral votes to clinch the White House. Here's where the race stands, with the states ordered by the projected margin between the candidates — Clinton's strongest states are farthest left, Trump's farthest right — and sized by the number of electoral votes they will award.



Pause

Recap of Linear Regression Part 4

- factor variables vs continuous variables
- analysis of variance, F-tests
- fruitfly example: $y_{ij} = \mu + \alpha_i + \beta x_{ij}$

- why use special techniques?
- one-way analysis of variance $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
  - parameters
  - analysis of variance table
  - partitioning of sums of squares
  - random effects modelling for factor/grouping variable

- principles of measurement
- phases of analysis

CD Ch.1, also Ch.4, p.54,55

Recall: one-way analysis of variance

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots R; i = 1, \ldots, T$$

- R observations in each group
- groups are defined by a factor variable
  - groups could be treatments, conditions, ... assigned by the investigator
  - groups could be families, litters, classrooms, ... sampled by the investigator
  - groups could be created by insisting that some measured covariate is treated as a factor
- the number of levels in the factor == number of groups
- in a completely randomized design, groups are created by random assignment of treatments to experimental units
- parameters α_i can be fixed or random depends on the application
- see FLM-1 13.1,2; FLM-2 14.1,2 for example with one continuous predictor and one two-level factor also HW1 Q3
- fruitflies (last week) on continuous predictor and one 5 level factor FLM-2 14.4; FLM-1 13.3

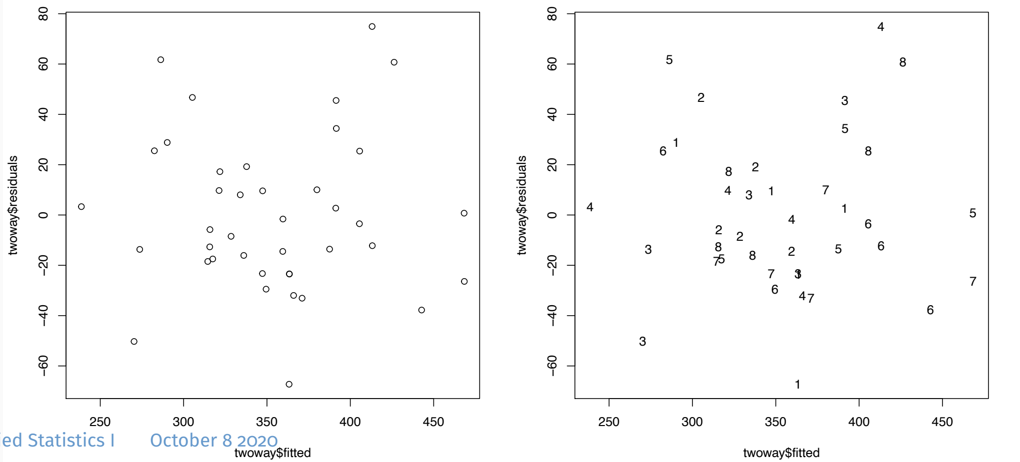
- two factor variables, treatment and block
- design: treatments assigned at random within blocks
- model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, \ldots, T; j = 1, \ldots R$$

- parameters:
  - $\mu = \mathbb{E}(y_{ij}$ if all $\alpha_i \equiv 0; \beta_j \equiv 0;$
  - α_i is change in $\mathbb{E}(y)$ from μ due to treatment i
  - β_j is change in $\mathbb{E}(y)$ due to effect of block j
  - ϵ_{ij} unexplained variation
- analysis:

$$\begin{aligned}
\sum_{ij}(y_{ij} - \bar{y}_{..})^2 &= \sum_{ij}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{.j} - \bar{y}_{..})^2 \\
&= \sum_{ij}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + \sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij}(\bar{y}_{.j} - \bar{y}_{..})^2
\end{aligned}$$

→ **oatvar.Rmd**


twoway$residuals
twoway$fitted
80
60
40
20
0
−20
−40
−60
250
300
350
400
450
twoway$residuals
twoway$fitted

RBD anova

$$
\begin{aligned}
\sum_{ij}(y_{ij} - \bar{y}_{..})^2 &= \sum_{ij}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{.j} - \bar{y}_{..})^2 \\
&= \sum_{ij}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + \sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij}(\bar{y}_{.j} - \bar{y}_{..})^2
\end{aligned}
$$

Table 9.5 Analysis of variance table for two-way layout model.

Term	df	Sum of squares
Treatments	$T-1$	$\sum_{t,b}(\bar{y}_{t.} - \bar{y}_{..})^2$
Blocks	$B-1$	$\sum_{t,b}(\bar{y}_{.b} - \bar{y}_{..})^2$
Residual	$(T-1)(B-1)$	$\sum_{t,b}(y_{tb} - \bar{y}_{t.} - \bar{y}_{.b} + \bar{y}_{..})^2$

Estimation of σ^2

```
     Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
variety    7  77524 11074.8  8.2839 1.804e-05 ***
block      4  33396  8348.9  6.2449  0.001008 **
Residuals 28  37433  1336.9
---

Residual standard error: 36.56 on 28 degrees of freedom
```

The interaction between blocks and treatments is used to estimate error. This is sometimes justified by assuming the block effects β_j are random.

Designed Experiments

- Completely randomized design:
  - can be used with more than one factor variable of interest
  - with two or more factors, often of interest to examine main effects and interactions
  - Example SM 9.6 and 8.10

- Randomized block design:
  - can also be used with two or more treatment factors
  - but sometimes it is hard to ensure the blocks are big enough to accommodate all combinations
  - leading to clever incomplete block designs FLM-2 17.2,3; FLM-1 16.2,3; SM 9.2.3 and p.432

  - A randomized block design with just two treatments in each block is a paired comparison paired t-test

Table 8.10 Poison data (Box and Cox, 1964). Survival times in 10-hour units of animals in a 3×4 factorial experiment with four replicates. The table underneath gives average (standard deviation) for the poison $\times$ treatment combinations.

Treatment	Poison 1	Poison 2	Poison 3
A	0.31, 0.45, 0.46, 0.43	0.36, 0.29, 0.40, 0.23	0.22, 0.21, 0.18, 0.23
B	0.82, 1.10, 0.88, 0.72	0.92, 0.61, 0.49, 1.24	0.30, 0.37, 0.38, 0.29
C	0.43, 0.45, 0.63, 0.76	0.44, 0.35, 0.31, 0.40	0.23, 0.25, 0.24, 0.22
D	0.45, 0.71, 0.66, 0.62	0.56, 1.02, 0.71, 0.38	0.30, 0.36, 0.31, 0.33

Treatment	Poison 1	Poison 2	Poison 3	Average
A	0.41 (0.07)	0.32 (0.08)	0.21 (0.02)	0.31
B	0.88 (0.16)	0.82 (0.34)	0.34 (0.05)	0.68
C	0.57 (0.16)	0.38 (0.06)	0.24 (0.01)	0.39
D	0.61 (0.11)	0.67 (0.27)	0.33 (0.03)	0.53
Average	0.62	0.55	0.28	0.48

Completely Randomized Design with 12 ‘treatments’

poison (3 levels) x Treatment (4 levels), 4 observations for each combination

Randomization

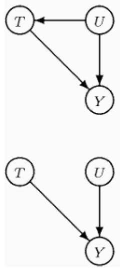
- Why do we randomize assignment of treatments to units? — if we can
- leads to approximate balance on potential confounding variables

- can't we adjust for confounding variables using regression?
- yes, if we know what they are — e.g. fruitflies (thorax length)

- can't adjust for unknown confounders — "unknown unknowns"

- depending on the sample size, there's a limit to how many variables can be included in the regression model

- randomization breaks the link from Unmeasured to T
  e.g. disease severity — causal diagram



Blocking

- comparison of treatments is more precise if the units are more homogeneous

- e.g. for agricultural trials, if properties of the soil are similar
- e.g. for clinical trials, if patients have similar levels of important measures, e.g. overall health

- putting experimental units into homogeneous (alike) subgroups before randomizing can give more precise estimates of treatment effects
- compare the two analysis of variance tables for one-way and two-way layouts
- 2-way has separate term for, e.g., blocks so residual SS is smaller

- "Block on what you can measure, randomize over what you can't measure"

- randomization helps to eliminate systematic error
- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run" CD

- treatment differences might be confounded by differences among patients, or by the time of day the treatment is applied, or by spatial differences among plots of land

- for example, units might be treated in space, or in time
- systematic error can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding

adjustment for measured covariate

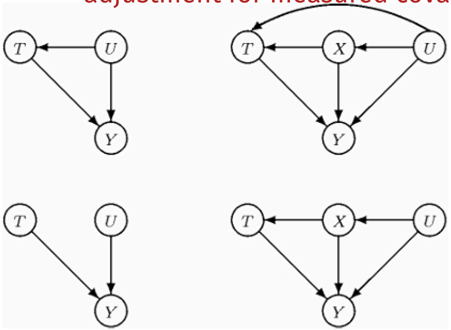
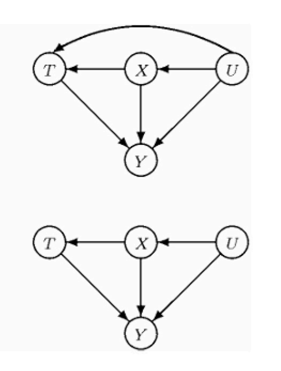


Figure 9.1 Directed acyclic graphs showing consequences of randomization. An arrow from T to Y indicates dependence of Y on T, and so forth. In general both response Y and treatment T may depend on properties U of units (upper left). Randomization (lower left) makes treatments and units independent, so any observed dependence of Y on T cannot be ascribed to joint dependence on U. The upper right graph shows the general dependence of Y, T, and covariates X on U. Randomization makes T and U independent, conditional on X (lower right), so any influence of U on T is mediated through X, for which adjustment is possible in

the control group. The response is to be the blood pressure of an individual measured a fixed time after the drug has first been administered. We calculate the average changes for the treated and control groups, $\overline{y}_1$ and $\overline{y}_0$, observe that $\overline{y}_1 - \overline{y}_0$ is significantly less than zero, and declare that the drug plays an effect in reducing blood pressure. Is this headline news? No!

Randomization and blocking



treatment allocation depends on measured covariate X, similar to blocking

effect of Unmeasured confounder mediated through X