Stats Group: Last Lecture

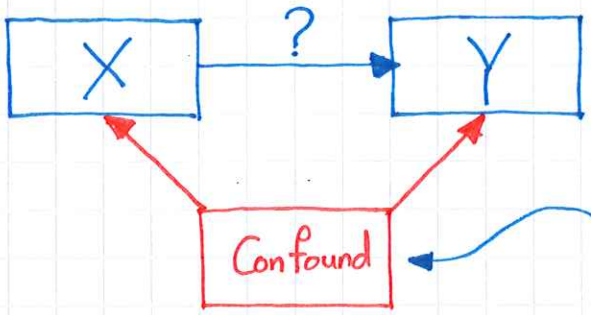
Objective:
- High level overview of course, drill principles

Experiments

Why perform an experiment?
- We have questions and want answers!

What form do our questions take on?

Does treatment (X) cause outcome (Y)?



Confounds can cause us to draw the wrong conclusions!

eg. Alcohol $\xrightarrow{?}$ Lung Cancer

Smoking

How does a researcher control for confounds?
- By randomly assigning treatment to experimental units

randomly: Spread confounds between groups

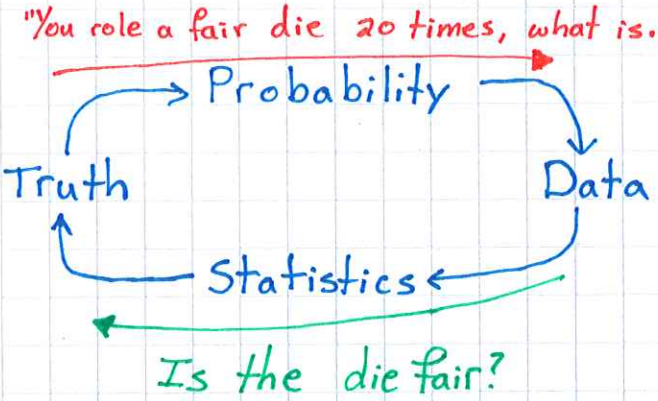
assigning: Control assignment of treatment to group

Given a question, how does an experiment help?
- Collect data from the natural world.
- Ask Mother Nature!
-

Statistics

How can we answer our question given data?

- Build a probability (mathematical) model!

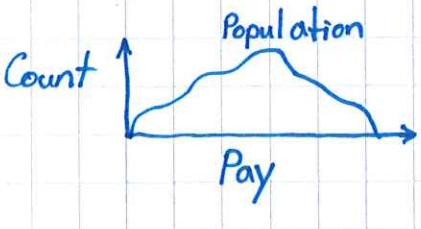


Statistics produces truths from data!

eg. Hypothesis: MSc are paid more than BSc.

The ask: H_0: $Y_{MSc} \stackrel{?}{=} Y_{BSc}$

How do you answer a question like that?



sample

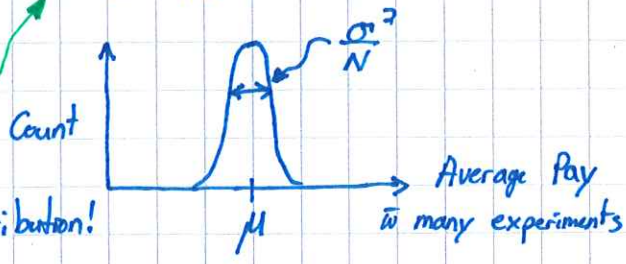
$X_1 = 20\,000$
$X_2 = 30\,000$
$\vdots$
$X_N = 100\,000$

Average $\rightarrow$ $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Can we say anything about $\bar{X}$?

Yes! As $n \to \infty$, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ for any population distribution!

(# samples)



We have a probability now! We can ask questions like:
"What is the probability $Y_{MSc} = \$100\,000$?"

$$p\text{-value} = P(Y_{msc} > \$100\,000)$$

Notice: We are building models of the <u>mean</u>

Quickly: What is the difference between a statistic and a parameter?

- Parameter is a value in our model (μ)
- Statistic is an estimate of the parameter ($\bar{x}$)

T-test

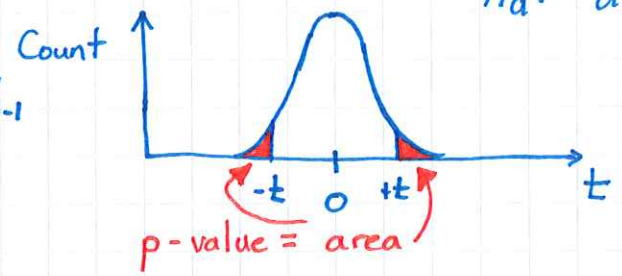
eg. MSc are paid more than BSc. How do we test this?

- Ask 20 MSc their pay, ask 20 BSc their pay
- Compute the averages: $\bar{X}_{MSc}$, $\bar{X}_{BSc}$
- Run a t-test!

$H_0: \bar{X}_{MSc} = \bar{X}_{BSc}$ $\quad H_a: \bar{X}_{MSc} \neq \bar{X}_{BSc}$ $\longrightarrow$ $d = \bar{X}_{MSc} - \bar{X}_{BSc}$ $\longrightarrow$ $H_0: d = 0$ $\quad H_a: d \neq 0$

$$t = \frac{d - 0}{SE(d)} \rightarrow T_{n-1}$$

$$SE(d) = \frac{\sigma}{\sqrt{n}}$$

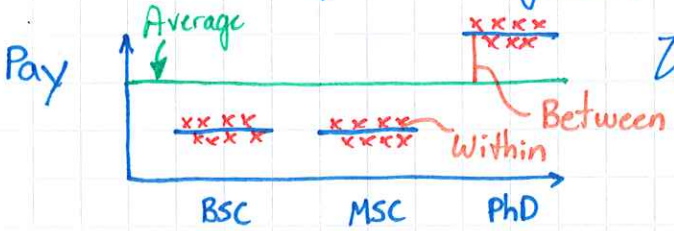


Note: Slightly more complicated because we don't have σ, we have $\hat{\sigma}$

ANOVA

eg. PhDs, MSCs, and BSCs are all paid the same.
Why can't we just run multiple tests?

- We get multiple comparisons
- The probability of finding a difference when one isn't really there grows with every test.



We now ask a different question:
Is the variance between the groups different from the variance within the groups

What is $\sigma^2_{between}$ relative to σ^2_{within}?

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}} \sim F_{K-1, n-K}$$

$$\sigma^2_{between} = MS_{between} = \frac{1}{K-1} \sum (\bar{Y}_K - \bar{Y})^2$$

$$\sigma^2_{within} = MS_{within} = \frac{1}{n-K} \sum (Y_{iK} - \bar{Y}_K)^2$$

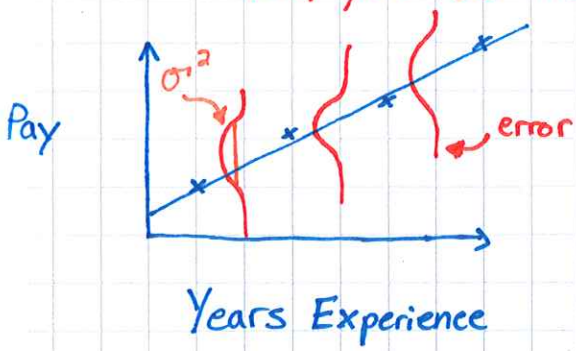
In General, $\sigma^2 = MS = \frac{SS}{df}$

~~General Linear Model~~ (I jumped the gun!)

Regression

ex. Pay increases with work experience. How do I test this?

- Do I run an ANOVA with 0, 1, 5, 10, 20, 30 years experience?
- No, you do regression!



$$Pay = 10\,000 + 15\,000 \cdot (Years\ of\ Experience)$$

$$Y = \beta_0 + \beta_1 X + N(0, \sigma^2)$$

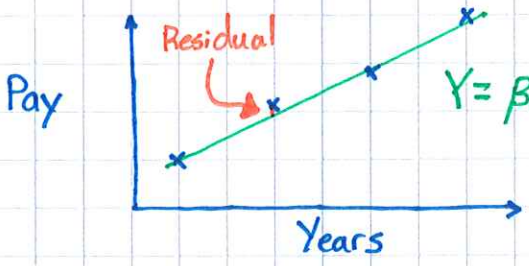
Parameters: β_0, β_1 Statistics: $\hat{\beta}_0, \hat{\beta}_1$

We built a model of the mean! Can ask questions like: $\beta_0 \stackrel{?}{=} 0$

How do we know if this is a good model?

Is $Y = \beta_0 + \beta_1 X$ (Model 1)
better than $Y = \beta_0$ (Model 2)

Notice: They are nested!
Model 1 = Model 2 + $\beta_1 X$



We can compair their residuals like ANOVA!

$$F = \frac{\text{Additional Variance Explain}}{\text{Unexplained Variance}}$$

$$F = \frac{\left(\frac{SS_{model\,2} - SS_{model\,1}}{df_1 - df_2}\right)}{MS_{model\,1}}$$

Additional Vari Sum of Squares

Additional predictors

Variance in new model

Wait, maybe t-tests and ANOVA ain't so different...

General Linear Model

It turns out that t-tests, ANOVA, and regression are the same thing...

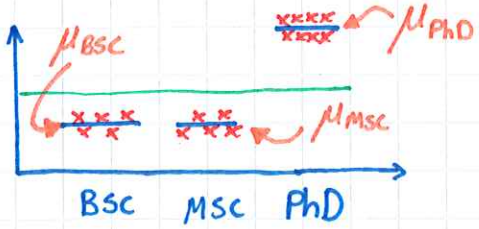
We have two basic principles: means and residuals

Means: All our models are predicting averages
Residuals: What our model cannot explain is variance

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + N(0, \sigma^2)$$

- Y: outcome
- $\beta_1 X_1$, $\beta_2 X_2$: treatment
- $N(0, \sigma^2)$: Unexplained
  - error, noise, sampling.

eg. PhD, MSc, BSc are all payed the same.



Can we make this a regression?

$$Y = \beta_0 + \beta_1 MSc + \beta_2 PhD$$

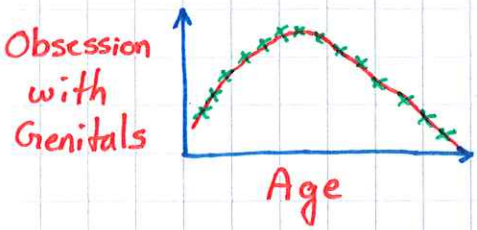
$$MSc = \begin{cases} 1, & \text{if MSc} \\ 0, & \text{otherwise} \end{cases} \qquad PhD = \begin{cases} 1, & \text{if PhD} \\ 0, & \text{otherwise} \end{cases}$$

Candidate	Indicator MSc	Indicator PhD	Model	
BSc	0	0	$Y = \beta_0$	$\mu_{BSc} = \beta_0$
MSc	1	0	$Y = \beta_0 + \beta_1$	$\mu_{MSc} = \beta_0 + \beta_1$
PhD	0	1	$Y = \beta_0 + \beta_2$	$\mu_{PhD} = \beta_0 + \beta_2$

They're all the same!

More GLM

ex. Can a linear model explain this data?



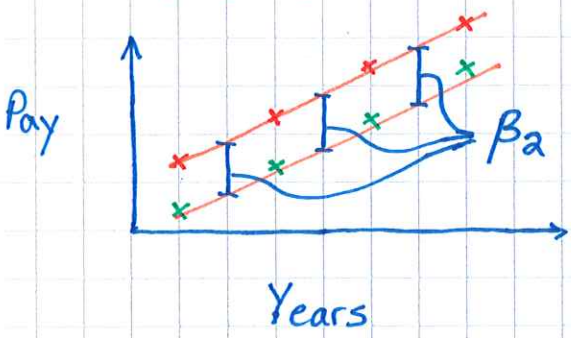
Well of course!

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

The model is linear in the <u>parameters</u> not the <u>treatment</u>!

~~ex~~. What if my treatment is categorical and continuous?

eg. Pay increase with work experience is different for men and women.

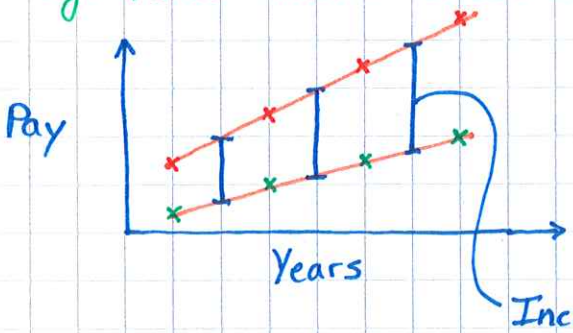


$$Y = \beta_0 + \beta_1 \text{Years} + \beta_2 \cdot \text{Men}$$

~~Men: $Y = \beta_0 + \beta_1$ Years~~
~~Women: $Y = (\beta_0 + \beta_2) + \beta_1$ Years~~

Men: $Y = (\beta_0 + \beta_2) + \beta_1$ Years
Women: $Y = \beta_0 + \beta_1$ Years

eg. AND there is an experience effect!

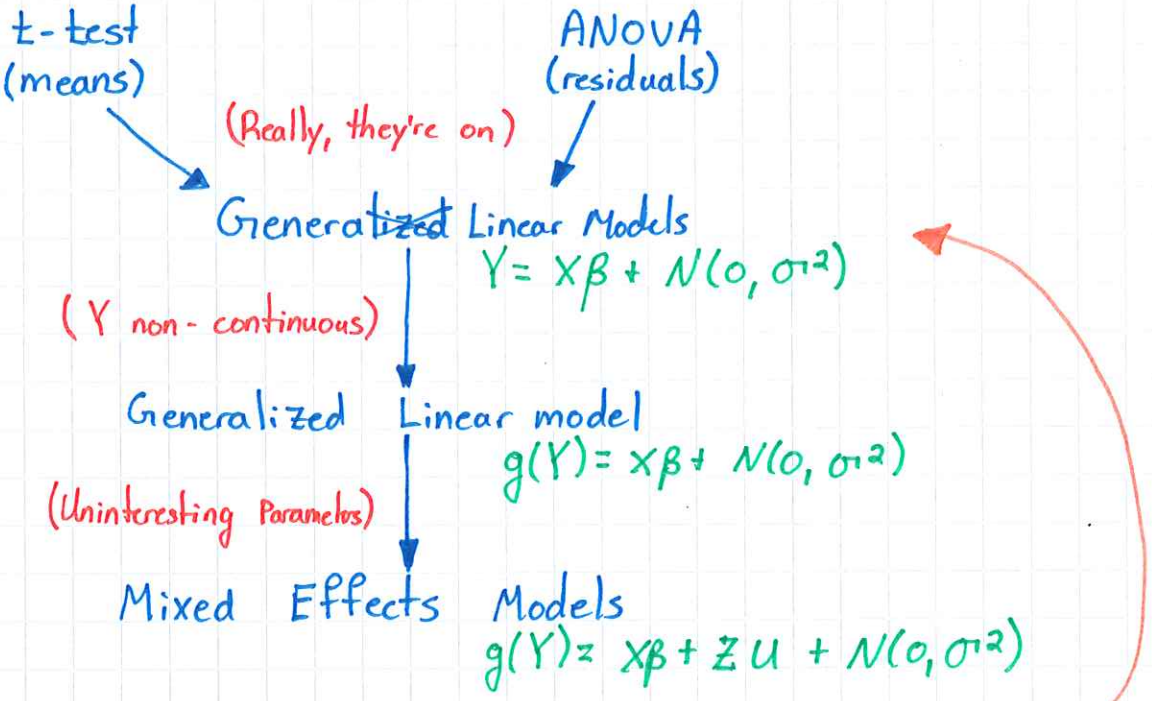


$$Y = \beta_0 + \beta_1 \text{Years} + \beta_2 \text{Men} + \beta_3 \text{Men Years}$$

Women: $Y = \beta_0 + \beta_1$ Years
Men: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)$ Years

A Model of Everything?

What does it model? — Means
What if Y isn't continuous? — Generalized Linear Models
What if we have uninteresting effects? — Mixed Effects Models



I hope after 10 weeks you understand up to here

- For assumptions, see online
- For understanding the models, I hope this has helped