Date: Thursday, March 21, 2019

Last Time:

* Key terminology:
  - Response, Predictors
  - Factors, levels
  - Conditions
  - Experimental units

features

values of features

individuals

set of levels

We want to know how a change in condition affects the response.

* Observational vs Experimental studies
* Randomization
* Causality
* How to answer a data-driven question: QPDAC

Question — Come up with a "binary" question about a response variable/metric you can measure.

* **Plan** — Developing an experiment to answer the question
  1) Come up with factors & levels conditions
  2) Randomly assign units across conditions,

Data — Collecting the data; ie running the experiment.

Considerations:
- Assumptions of the randomization (SUTVA)
- How long?
  Too long: Frustrating to users
  Too short: Not enough statistical power
- Sample size calculations

Analysis — Applying statistical testing to answer the question first posed.

Conclusion — What did we learn? How does it affect the service of a company? Do we need more experimentation?

Can we learn from significant and non-significant results.

Experiments with two conditions

↳ A/B testing!

Type of experiment used to determine which of two alternatives (conditions) leads to increased KPI (metric / response) performance.

↓ key performance indicators.

Examples:

- Video games { 1/2 users new item costs 10 diamonds
  1/2 users new item costs 100 diamonds

  ↳ How much should the new item cost?

- Twitter { 10% see ♡ and "Like"
  90% see ☆ and "Favorite"

Why use A/B testing?

- Some things cannot be determined by observation only.

  ↳ In the video game example, it would be confusing to expose all users to different costs.

Why is A/B testing difficult?

- Ensuring there is no interference (ie VG users don't talk to each other)
- Tracking users in "A" vs "B" is challenging.
- Multiple games per user, etc.
- Statistics is hard

Statistical tests are used to determine if there is enough evidence in a sample data set to infer that a certain conclusion is true. They are stated in terms of population parameters.

Components:

1) Null hypothesis (H_0) — What is generally believed to be true. Typically A & B are statistically the same.

2) Alternative hypothesis (H_A) — Complements H_0. States that there is some difference in A and B. We aim to support through evidence from our sample data.

Mantra for testing:

Comes from the US judicial system:

"INNOCENT UNTIL PROVEN GUILTY"

↕

"H_0 is true until data says otherwise".

3) Data — Assumed to be collected from random samples
↳ ie each observation is independent of other observations.

In the context of our experiment, the data is the metric/response variable in the question asked.

4) Population metric of interest — A population summary of the data from each condition.

Ex: (Video game)
Suppose metric is number of hours played in a week from start of test.

(*) - Question: Does seeing 100 diamonds lead to more game play?

- Randomly assign 100 users to version A (10 diamonds)
  " " " " B (100 diamonds)

- Run experiments for a week and calculate

x_j = # hours user j played the game.

- Population metric:

μ_A = true mean # hours per week for those who see 10 diamonds

μ_B = " " " " 100 diamonds.

(mean gets central tendency of gameplay. Could use median, etc.)

- Hypothesis test: $H_0: \mu_A = \mu_B$ (same game play)

vs

$H_A: \mu_B > \mu_A$

We now use data $x_1, \dots, x_N$ to resolve the above test.

Note: H_A can be stated in many ways depending on the question at hand.

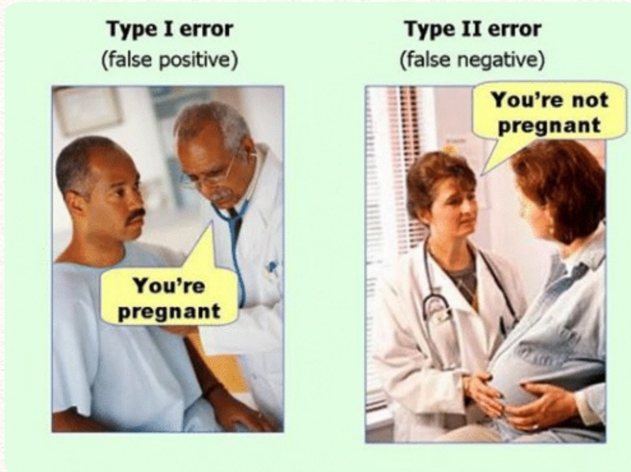
↳ Could be: $H_A: \mu_B < \mu_A$ (less play on B)

$H_A: \mu_B \neq \mu_A$ (different play on B)

Errors:

Truth

Decision from Data		H_0	H_A
	H_0	✓	Type II
	H_A	Type I	✓



Type I error: Incorrectly rejecting H_0.

Type II error: Failing to reject H_0 when it is false.

* In general, there is a trade-off between type I & II errors for a fixed sample size.

* when deciding whether or not H_0 is true, we hold the probability of type I error fixed and evaluate the probability of type II error.

Notation: $\underbrace{\alpha}_{\text{fix this!}} = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true}) = \mathbb{P}(\text{type I error}) =$ "Significance of test"

$\underbrace{\beta}_{\text{unfortunately overlooked :(}} = \mathbb{P}(\text{failing to reject } H_0 \mid H_A \text{ true}) = \mathbb{P}(\text{type II error})$

$\underbrace{\text{Power}}_{\text{evaluate this}} = 1-\beta = \mathbb{P}(\text{reject } H_0 \mid H_A \text{ true})$

we want a small α (close to 0 but not exactly 0) and a high power (as close to 1 as possible).

How do we make decisions?

1) Define a criteria for rejection of H_0

- p-value
- rejection region

} we want a cutoff to decide when to reject H_0.

Example: - Reject H_0 when p-value $< \alpha$

- Reject when data lies in rejection region.

2) Calculate a sample statistic (from data) which completely depends on H_0 & H_A.
   ↳ Calculate p-value - decide.

Note: The above is frequentist. For Bayesian analysis recall that we calculate posterior probabilities of H_0 being true and make decisions on that.