

# Hypotheses

Macartan Humphreys

Feb 2017

# Intro

# Roadmap

- What a hypothesis is
- What makes for a good hypothesis
- Accepting or rejecting hypotheses
- The  $p$  value
- Calculating  $p$  values using randomization
- Making decisions using  $p$  values
- The dangers of fishing
- Estimation

# Take home ideas

- Stating expectations in terms of hypotheses provides *discipline* to a research project.
- Hypotheses are statements about the world that you seek to *reject*
- A good hypothesis is simple and falsifiable
- A  $p$  value is the probability of data like what you see under some particular hypothesis
- You can calculate a  $p$  value from the randomization and observed data without heroic assumptions
- Warning: no fishing
- Warning: do not fetishize hypothesis tests. Often the real interest is in *estimating* the size of an effect not rejecting any particular null.

# Hypotheses

# Characteristics of good hypotheses

- They are possibly TRUE or FALSE
- They are *falsifiable*
- They are statements about the world, not your analysis.
- They are simple (not double barreled)
- They involve clear concepts
- They are **few**, and they are motivated
- They are **contested**: You will learn something whether the data supports them or rejects them. Most importantly: you are not sure if they are true or false
- They are numbered, and maybe even named

# Some hypotheses

Consider these:

- Education is very important
- Education increases your income
- Education either increases, decreases, or has no effect on your income
- Education is good for you because it strengthens your character in very fundamental ways that you could never measure

Now:

- Just one of these is not a hypothesis. Which one?
- Just one of these is a good hypothesis. Which one?

## Some more bad ones

**H1:** The election of Donald Trump will have a big effect on world security because it will make many world leaders uncertain about whether the US will support them in case of attack and by increasing competition over natural resources

**H2** The election of Donald Trump will have no statistically significant effect on world security

What is wrong with these?



## Nulls: A point of confusion

Because of an unusual convention, social scientists often describe hypotheses in terms of what they **expect** but then *test* the null hypothesis of no effect

eg:

- H1: Competition reduces prices
- H-null: Competition has no effect on prices

**Test:** how likely is the data given the null

# Tests

# Tests: Hypotheses often rejected, sometimes maintained, but rarely accepted

In the classical approach to testing a hypothesis we ask:

**How likely are we to see data like this if the hypothesis is true?**

- If the answer is “not very likely” then we treat the hypothesis as suspect.
- If the answer is *not* “not very likely” then the hypothesis is maintained (some say “accepted” but this is tricky as you may want to “maintain” multiple incompatible hypotheses)

How unlikely is “not very likely”

# Weighing Evidence

When we test a hypothesis we decide first on what sort of evidence we need to see in order to decide that the hypothesis is not reliable.

**Othello** has a hypothesis that Desdemona is innocent. **Iago** confronts him with evidence:

- See how she looks at him: would she look at him like that if she were innocent?
- See how she defends him: would she defend him like that if she were innocent?
- See he carries her handkerchief: would he have her handkerchief if she were innocent?
- Othello, the chances of all of these things arising if she were innocent is surely less than 5%

# Hypotheses are often rejected, sometimes maintained, but rarely accepted

Note that Othello is focused on the probability of the events if she were innocent but not the probability of the events if Iago were trying to trick him.

He is not directly assessing his belief in whether she is faithful, but rather how likely the data would be if she were faithful.

That is, he assesses:

$$\Pr(\text{Data} | \text{Hypothesis is TRUE})$$

not

$$\Pr(\text{Hypothesis is TRUE} | \text{Data})$$

# Not Bayes

Note:  $Pr(\text{Data}|\text{Hypothesis is TRUE})$  and  $Pr(\text{Hypothesis is TRUE}|\text{Data})$  are connected but in a slightly complex way (Bayes Rule):

$$Pr(H|D) = \frac{Pr(D|H) Pr(H)}{Pr(D|H) Pr(H) + Pr(D|NOT H) Pr(NOT H)}$$

So your belief about the hypothesis should depend not just on the likelihood of seeing the data given the hypothesis but also on your prior belief about how plausible the hypothesis is. But this second part is ignored in classical tests.

p

# The $p$ value

The famous  $p$  value reports the **probability of observing this type of evidence**: eg the probability of getting such a large estimated effect

Consider this: “I estimate the treatment increased income by \$10, my standard error is 4, and my  $p$  value is 0.05.”

Which of these statements is correct (just one!):

- 1 The probability that treatment increased income by \$10 is just 5%
- 2 The probability that treatment increased income by \$10 is 95%
- 3 The probability that treatment increases income is 95%
- 4 The probability that treatment does **not** increase income is just 5%
- 5 The probability that we would treatment increases income is 95%
- 6 The probability that we would estimate an effect of \$10 if the true effect were 0 is 5%
- 7 The probability that we would estimate an effect of \$10 if the true effect were positive is 95%



## The $p$ value: One sided and two sided

- $p$  values are sometimes based on “one sided” or “two sided” tests.
- These are very similar ideas, the key difference is:
- For a one sided test ask: What is the probability that you would get such a large estimate if there were no true effect. e.g. If I estimate “5” then: what is the probability that I would get 5 or larger by chance?”
- For a two sided test ask: What is the probability that you would get such a large estimate in absolute magnitude if there were no true effect. e.g. If I estimate “5” then: what is the probability that I would get 5 or larger OR -5 or smaller by chance?”

## Calculate some $p$ values

I have a coin. You are not sure if it is a fair coin (ie has a heads and a tails), or if in fact there are heads on both sides.

**Null hypothesis:** It is a fair coin: you are equally likely to get heads or tails

**Evidence:**

- I toss the coin once. It comes up heads. What are the chances that it would come up heads if it were a fair coin? Should we reject the null?
- I toss again. It comes up heads again. What are the chances that it would come up heads twice if it were a fair coin?
- I toss again. It comes up heads again. What are the chances that it would come up heads three times if it were a fair coin?
- I toss two more times. It comes up heads *both* times. What's the chance it would come up heads *five* times if it were a fair coin?

## Calculate some $p$ values

- What's the chance it would come up heads *five* times if it were a fair coin?

```
.5^(1:5)
```

```
[1] 0.50000 0.25000 0.12500 0.06250 0.03125
```

### More Evidence:

- I keep going. In the end I get 99 heads and 1 tail.
- What should I conclude?
- Should I reject the null hypothesis that this is a fair coin?

# The rejection decision

It is up to you to decide how low low is in order to reject a hypothesis.

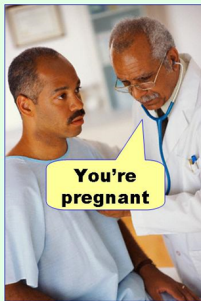
The most common standard in social science is  $p \leq 0.05$ . That means that you reject the hypothesis if you get a  $p$  value below 0.05.

Here 0.05 is a cutoff, sometimes called the **alpha level** of the test. There are advantages and disadvantages of choosing different alpha levels.

- Say we set  $\alpha = 1$ . This means that we will ALWAYS say that effects are significant. This has the really great advantage of guaranteeing that if there is a real effect we will always reject the null of no effect. What is the disadvantage?
- Say we set  $\alpha = 0$ . This means that we will NEVER say that effects are significant. This has the really great advantage of guaranteeing that if really there is no effect we will not mistakenly say that the null of no effect is incorrect. (Apologies for four (six?) negatives in a row). What is the disadvantage?

# Test types

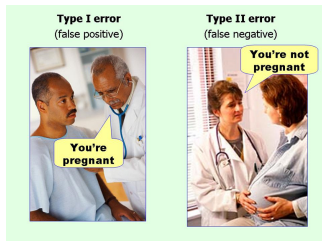
**Type I error**  
(false positive)



**Type II error**  
(false negative)



# Test types



- If you have a high alpha, eg 0.1 and your null is “Not pregnant” then you are more likely to say “Pregnant” even if you see a moderate belly.
  - High alpha: More Type I Error (False Positives)
- If you have a low alpha, eg 0.01 and your null is “Not pregnant” then you are not likely to say “Pregnant” unless you see a really big belly.
  - Low alpha: More Type II Error (False Negatives)

# Tests and Decisions

Decision resulting from data analysis	True condition	
	$H_0$ false ("Change has occurred")	$H_0$ true ("No change")
Reject $H_0$ ("Change has occurred")	Correct decision ( $1-\beta$ : Power of the test)	Error (Type I) ( $\alpha$ )
Fail to reject $H_0$ ("No change")	Error (Type II) ( $\beta$ )	Correct decision

# Odd terminology

We often say a strategy is “**conservative**” if it demands a lot of evidence to reject a maintained null.

- if you start off assuming that people are *not* pregnant then you really need to see a lot of evidence to make you change your mind
- in statistical analysis the analogue is that you want to see a very small  $p$  to reject the null – that is you have a low alpha value
- in general researchers tend to prefer conservative strategies that guard against false new claims. But they also prevent some true new claims from being recognized.



# Randomization Inference

# Randomization Inference

Introducing an entirely new way to think about statistical significance. . .

- You can calculate the  $p$  value using information about the randomization
- Say you randomly assigned one unit to treatment and your data looked like this.

Unit	1	2	3	4	5	6	7	8	9	10
Treatment	0	0	0	0	0	0	0	1	0	0
Healthy?	3	2	4	6	7	2	4	9	8	2

- Does the treatment improve your health?
- $p = ?$

# Randomization Inference

- Introducing an entirely new way to think about statistical significance. . .
- Say you randomly assigned one unit to treatment and your data looked like this.

Unit	1	2	3	4	5	6	7	8	9	10
Treatment	0	0	0	0	0	0	0	1	0	0
Healthy?	3	2	4	6	7	2	4	8	9	2

- Does the treatment improve your health?
- $p = ?$

## Extra Example: A Very Tiny Experiment Fully Analyzed

- Consider an experiment with 4 units assigned
- 2 are assigned to treatment, 2 to control
- Outcomes in the treatment group: 45, 50
- Outcomes in the control group: 30, 35
- Estimated effect?

[Credit to Gareth Nellis for this example]

# Tiny Experiment : Outcomes

Actual outcomes are in the top left panels. Other panels show what we *might* have seen if the null were true.

	T	C
Voter 1	45	
Voter 2	50	
Voter 3		35
Voter 4		30
Diff-in-means = $[(45+50)/2] - [(35+30)/2] = 15$		

	T	C
Voter 1	45	
Voter 2		50
Voter 3	35	
Voter 4	30	
Diff-in-means = $[(35+30)/2] - [(45+50)/2] = -15$		

	T	C
Voter 1	45	
Voter 2		50
Voter 3	35	
Voter 4		30
Diff-in-means = $[(45+35)/2] - [(50+30)/2] = 0$		

	T	C
Voter 1		45
Voter 2	50	
Voter 3		35
Voter 4	30	
Diff-in-means = $[(50+30)/2] - [(45+35)/2] = 0$		

	T	C
Voter 1	45	
Voter 2		50
Voter 3	35	
Voter 4	30	
Diff-in-means = $[(45+30)/2] - [(50+35)/2] = -5$		

	T	C
Voter 1		45
Voter 2	50	
Voter 3	35	
Voter 4		30
Diff-in-means = $[(50+35)/2] - [(45+30)/2] = 5$		

# Tiny Experiment: Inference

So we estimate an effect of 15. We now want to figure out if we can reject the null of NO true effect.

Procedure:

- Hold onto the observed outcomes, but imagine that T and C had been assigned differently
- Q: How many permutations? i.e. how many ways of assigning 2 out of 4 subjects to treatment?
- A: 4-choose-2:  $\frac{4!}{2!2!} = 6$  (binomial coefficient)
- Now: recalculate the average outcome under all possible permutations
- End up with all 6 possible average outcomes:

−15, −5, 0, 0, 5, 15

# Tiny Experiment: Permutation Test

- Two-sided  $p$ -value: sum of the absolute value of the mass at or greater than the observed value
- There are two values as big as the absolute value of 15 (that is, 15 and -15)
- Probability of observing an outcome as big or bigger than 15 is  $\frac{2}{6} = 0.33$ , assuming strict null hypothesis is true
- This is a large  $p$ -value; we fail to reject the null at conventional levels

It gets worse: no matter what the data looked like we could *never* get a  $p$  value below 0.33 with only four units. . . .

# Multiple Comparisons



# Multiple comparisons

Say you have a design which lets you correctly calculate the  $p$  values for the effect of treatment  $X$  on outcome  $Y$ .

- Say in truth that there is no true effect.
- What are the chances that you will conclude that there *is* a treatment effect?

Say now that look at 10 independent outcomes.

- Say in truth that there is no true effect on any of them.
- What are the chances that you will conclude that there *is* a treatment effect on at least one of them?

# Multiple comparisons

Say you have a design which lets you correctly calculate the  $p$  values for the effect of treatment  $X$  on outcome  $Y$ .

- Say in truth that there is no true effect.
- What are the chances that you will conclude that there *is* a treatment effect?

Say now that look at 10 independent outcomes.

- Say in truth that there is no true effect on any of them.
- What are the chances that you will conclude that there *is* a treatment effect on at least one of them?

$$1 - .95^{20}$$

$$[1] 0.6415141$$

# Go fishing

See exact fishy test app on egap website (run locally)

# Estimation and Testing

- Very often, especially when hypotheses are contested, it is hard to make a good argument for why you expect there to be a positive effect or a negative effect or no effect
- In such cases it can be appropriate to resist the pressure to state a particular hypothesis, instead state your estimand very clearly:
- Not: I expect education increases income
- But: I want to see how large the effect of income on education is
- In fact even when you do estimation, once you generate confidence intervals you are implicitly conducting a whole series of hypotheses tests: your confidence interval is the set of hypotheses that you do not reject

End

End

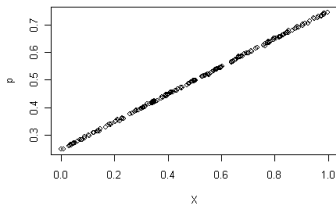
# Extra Slides

# Randomization Inference Comes to the Rescue

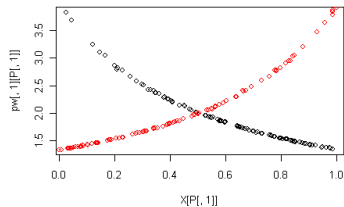
- Say you had a silly randomization procedure and forgot to take account of it in your estimates.
- eg you assigned richer people to treatment with a higher propensity than poor people
- If you are aware of this then you could adjust for these weights and you will be fine in your analysis
- But if you are not aware of it then you risk false estimates of treatment effects. eg are you more likely to think that an intervention had a positive or a negative effect on people's wealth?
- The good news: even if you mess up the assignment AND the estimation, you might still get the right  $p$  value. . .

# Randomization Inference Comes to the Rescue

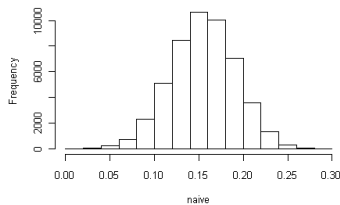
Propensities correlated with some covariate



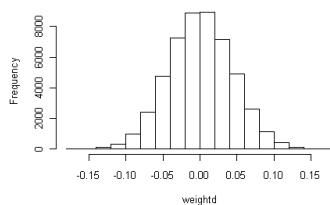
Inverse propensity weights (Red=Control)



Distribution of possible estimates from naive analysis



Distribution of estimates from weighted analysis





# Randomization Inference in the Lab

- Randomization procedures are sometimes funky in lab experiments
- Using randomization inference would force a focus on the true assignment of individuals to treatments
- Fake (but believable) example follows

# Randomization Inference in the Lab

Table 1: Optimal assignment to treatment given constraints due to facilities

		Capacity	T1	T2	T3
Session	Thursday	40	10	30	0
	Friday	40	10	0	30
	Saturday	10	10	0	0
		90	30	30	30

Table 2: Constraints due to subjects

Subject Type	N	Available
A	30	Thurs, Fri
B	30	Thurs, Sat
C	30	Fri, Sat

# Randomization Inference in the Lab

If you think hard about assignment you might come up with an allocation like this.

Table 3: Assignment of people to days

Subject Type	N	Available	Allocation		
			Thurs	Fri	Sat
A	30	Thurs, Fri	15	15	
B	30	Thurs, Sat	25		5
C	30	Fri, Sat		25	5

That allocation balances as much as possible. Given the allocation you might randomly assign individuals to different days as well as randomly assigning them to treatments within days. If you then figure out assignment propensities, this is what you would get:

Subject Type	N	Available	Assignment Probabilities		
			T1	T2	T3
A	30	Thurs, Fri	0.25	0.375	0.375
B	30	Thurs, Sat	0.375	0.625	0
C	30	Fri, Sat	0.375		0.625

# Randomization Inference in the Lab

Even under the assumption that the day of measurement does not matter, these assignment probabilities have big implications for analysis.

Subject Type	N	Available	Assignment Probabilities		
			T1	T2	T3
A	30	Thurs, Fri	0.25	0.375	0.375
B	30	Thurs, Sat	0.375	0.625	0
C	30	Fri, Sat	0.375		0.625

- Only the type  $A$  subjects could have received any of the three treatments.
- There are no two treatments for which it is possible to compare outcomes for subpopulations  $B$  and  $C$
- A comparison of  $T1$  versus  $T2$  can only be made for population  $A \cup B$
- However subpopulation  $A$  is assigned to  $A$  (versus  $B$ ) with probability  $4/5$ ; while population  $B$  is assigned with probability  $3/8$
- **Implications for design:** need to uncluster treatment delivery
- **Implications for analysis:** need to take account of propensities

# Randomization Inference Generalized

- Randomization inference can get quite a bit more complicated when you want to test a null other than the sharp null of no effect.
- Say you want to test the null that the effect is 2 for all units. How to do it?
- Say you want to test the null that an *interaction effect* is zero. How to do it?
- In both cases by filling in a potential outcomes schedule given the hypothesis in question and then generating a test statistic

Observed		Under null that effect is 0		Under null that effect is 2	
Y(0)	Y(1)	Y(0)	Y(1)	Y(0)	Y(1)
1	?	1	1	1	3
2	?	2	2	2	4
?	4	4	4	2	4
?	3	3	3	1	3

## Randomization Inference: Some code

- In principle it is very easy.
- These few lines generate data, produce the regression estimate and then an RI estimate of  $p$ :

```
X <- rep(c(FALSE,TRUE),50)
Y <- .5*X + rnorm(100)           # DATA

b = matrix(NA,10000)           # RI
for(i in 1:length(b)){
  Z <- sample(X)
  b[i] <- mean(Y[Z]) - mean(Y[!Z])
}
mean(b >= mean(Y[X]) - mean(Y[!X])) # One sided p value
```

```
[1] 0
```

# Preparing for Randomization Inference

- In practice it is a good idea to create a  $P$  matrix when you do your randomization (although note: if the null is about one treatment, then you are interested only in the randomization of that treatment, not the joint randomization of all)
- Then you can draw permutations from the original assignment code