Hypothesis Testing

Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

Overview of Statistical Hypothesis Testing

- Probabilistic "Proof by contradiction": Assume the negation of the proposition, and show that it leads to the contradiction

1. Construct a null hypothesis (H_0) and its alternative (H_1)
2. Pick a test statistic T
3. Figure out the sampling distribution of T under H_0 (reference distribution)
4. Is the observed value of T likely to occur under H_0?
    - Yes – Retain H_0
    - No – Reject H_0

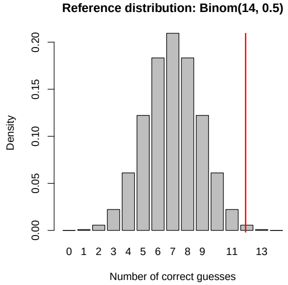
Paul the Octopus

- 2010 World Cup
  - Group: Germany vs Australia
  - Group: Germany vs Serbia
  - Group: Ghana vs Germany
  - Round of 16: Germany vs England
  - Quarter-final: Argentina vs Germany
  - Semi-final: Germany vs Spain
  - 3rd place: Uruguay vs Germany
  - Final: Netherlands vs Spain

- Question: Did Paul the Octopus get lucky?
- Null hypothesis: Paul is randomly choosing winner
- Test statistics: Number of correct answers
- Reference distribution: $\mathrm{Binomial}(8, 0.5)$
- The probability that Paul gets them all correct: $\frac{1}{2^8} \approx 0.004$
- Tie is possible in group rounds: $\frac{1}{3^3} \times \frac{1}{2^5} \approx 0.001$

More Data about Paul

- UEFA Euro 2008
  - Group: Germany vs Poland
  - Group: Croatia vs Germany
  - Group: Austria vs Germany
  - Quarter-final: Portugal vs Germany
  - Semi-final: Germany vs Turkey
  - Final: Germany vs Spain
- A total of 14 matches
- 12 correct guesses



- *p*-value: Probability that under the null you observe something at least as extreme as what you actually observed
- $\Pr(\{12, 13, 14\}) \approx 0.001$

```
pbinom(12, size = 14, prob = 0.5, lower.tail = FALSE)
## [1] 0.000916
```

Paul's Rival, Mani the Parakeet

- 2010 World Cup
  - Quarter-final: Netherlands vs Brazil
  - Quarter-final: Uruguay vs Ghana
  - Quarter-final: Argentina vs Germany
  - Quarter-final: Paraguay vs Spain
  - Semi-final: Uruguay vs Netherlands
  - Semi-final: Germany vs Spain
  - Final: Netherlands vs Spain

- Mani did pretty good: p-value is 0.063
- Danger of multiple testing
- Take 10 animals with no forecasting ability. What is the chance of getting p-value less than 0.05 at least once?

$$1 - 0.95^{10} \approx 0.4$$

- If you do this with enough animals, you will find another Paul

Hypothesis Testing for Proportions

1. Hypotheses – $H_0 : p = p_0$ and $H_1 : p \neq p_0$
2. Test statistic: $\overline{X}_n$
3. Under the null, by the central limit theorem

$$z\text{-score} \;=\; \frac{\overline{X} - p_0}{\text{standard deviation}} \;=\; \frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \overset{\text{approx.}}{\sim} \mathcal{N}(0,1)$$

4. Is Z_{obs} unusual under the null?
    - Reject the null when $|Z_{obs}| > z_{\alpha/2}$
    - Retain the null when $|Z_{obs}| \leq z_{\alpha/2}$

- The level (size) of the test: $\Pr(\text{rejection} \mid H_0) \;=\; \alpha$
- Duality with confidence intervals:
    - Reject the null $\Longleftrightarrow$ p_0 not in CI_α
    - Retain the null $\Longleftrightarrow$ p_0 in CI_α
- When X_i is normally distributed, use t-statistic and obtain the critical value using Student's t distribution

p-value

- (two-sided) p-value$=$ $\Pr(Z > |Z_{obs}|) + \Pr(Z < -|Z_{obs}|)$
- One sided alternative hypothesis: $H_1 : p > p_0$ or $p < p_0$
- one-sided p-value$=$ $\Pr(Z > Z_{obs})$ or $\Pr(Z < Z_{obs})$
- Use `pnorm()` or `pt()`

- p-value is the probability, computed under H_0, of observing a value of the test statistic at least as extreme as its observed value
- A smaller p-value presents stronger evidence against H_0
- p-value less than α indicates statistical significance $\leftrightarrow$ α-level test

- p-value is NOT the probability that H_0 (H_1) is true (false)
- The statistical significance indicated by the p-value does not necessarily imply scientific significance

Back to the Polling Examples

Obama's approval rate

- $H_0 : p = 0.5$ and $H_1 : p \neq 0.5$
- $\alpha = 0.05$ level test
- $\overline{X}_n = 0.45$ and $n = 1500$
- $Z_{obs} = (0.45 - 0.5)/\sqrt{0.5 \times 0.5/1500} = 3.87 > z_{0.025} = 1.96$
- p-value $= 0.00005 \times 2 = 0.0001$
- Reject the null

Error and Power of Hypothesis Test

- Two types of errors:

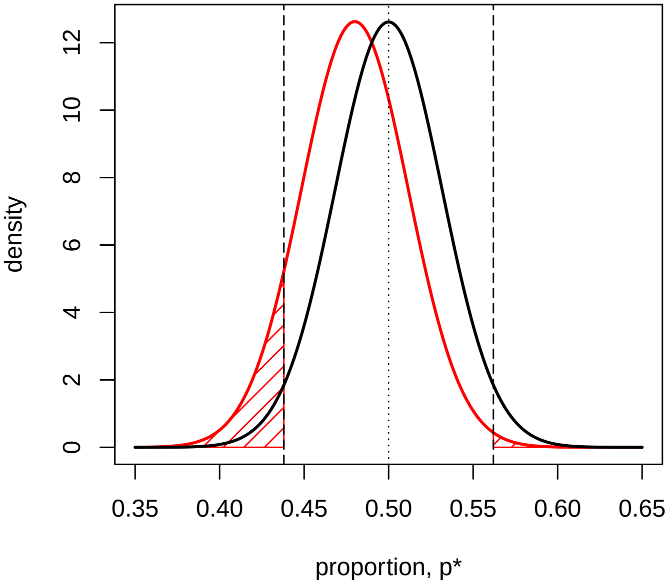
	Reject H_0	Retain H_0
H_0 is true	Type I error	Correct
H_0 is false	Correct	Type II error

- Hypothesis tests control the probability of Type I error, which is equal to the level of tests or α
- They do not control the probability of Type II error
- Tradeoff between the two types of error
- A large p-value can occur either because H_0 is true or because H_0 is false but the test is not powerful

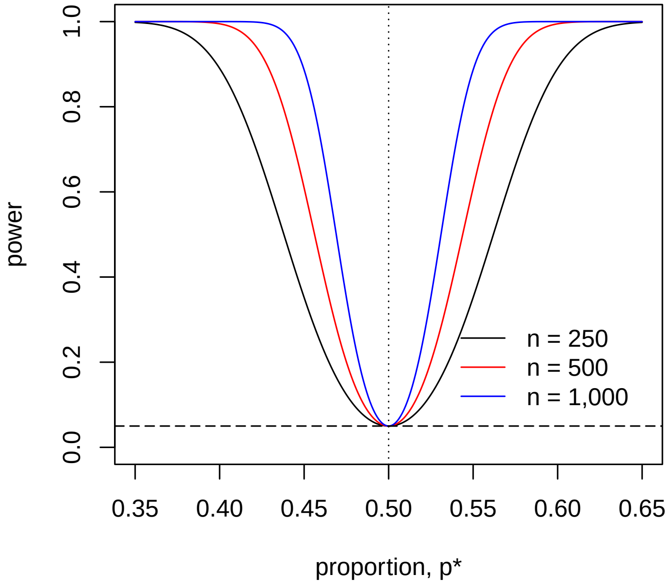
- Level of test: probability that the null is rejected when it is true
- Power of test: probability that a test rejects the null
- Typically, we want a most powerful test given the level

Power Analysis

- Null hypotheses are often uninteresting
- But, hypothesis testing may indicate the strength of evidence for or against your theory
- Power analysis: What sample size do I need in order to detect a certain departure from the null?
- Power $= 1 - \Pr(\text{Type II error})$
- Three steps
  1. Suppose $\mu = \mu^*$ which implies $\overline{X}_n \sim \mathcal{N}(\mu^*, \mathbb{V}(X)/n)$
  2. Calculate the rejection probability noting that we reject $H_0 : \mu = \mu_0$ if $|\overline{X}_n| > \mu_0 + z_{\alpha/2} \times$ standard error
  3. Find the smallest n such that this rejection probability equals a pre-specified level



12
10
8
6
4
2
0
density
0.35
0.40
0.45
0.50
0.55
0.60
0.65
proportion, p*


power
proportion, p*
0.0
0.2
0.4
0.6
0.8
1.0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
n = 250
n = 500
n = 1,000

Social Pressure Experiment (Review)

- Turnout rate: $\overline{X}_T = 0.37$, $\overline{X}_C = 0.30$,
- Sample size: $n_T = 360$, $n_C = 1890$
- Estimated average treatment effect:

$$\widehat{\text{ATE}} \; = \; \overline{X}_T - \overline{X}_C \; = \; 0.07$$

- Standard error:

$$\sqrt{\frac{\overline{X}_T(1 - \overline{X}_T)}{n_T} + \frac{\overline{X}_C(1 - \overline{X}_C)}{n_C}} \; = \; 0.028$$

- 95% Confidence intervals:

$$\begin{aligned} & [\widehat{\text{ATE}} - \text{standard error} \times z_{0.025}, \; \widehat{\text{ATE}} - \text{standard error} \times z_{0.025}] \\ = \; & [0.016, \; 0.124] \end{aligned}$$

Social Pressure Example (Continued)

- Two-sample test
    - $H_0 : p_T = p_C$ and $H_1 : p_T \neq p_C$.
    - Reference distribution: $\mathcal{N}\left(0, \ \frac{p(1-p)}{n_T} + \frac{p(1-p)}{n_C}\right)$
    - p-value: 0.010
- Power calculation:
    - $p_T = 0.37$ and $p_C = 0.30$
    - Two-sample test at the 5% significance level
    - Equal group size: $n_T = n_C$
    - If $n = 1000$, what is the power of the test?