Recap

- (Ch 7) Assessing the significance of evidence against a hypothesis
  - Hypothesis testing
  - P-values

Today

- (Ch 7) Assessing the significance of evidence against a hypothesis
  - More on p-values
  - A complex hypothesis: do two populations have the same mean?

The use and misuse of p-values

- P-values in scientific practice
  - Scientists are usually trying to reject the null hypothesis, the hypothesis that there is no special phenomenon and the data are just random noise
  - So $p < 0.05$ means there may be an interesting phenomenon and it has been the standard for publication in many fields

- What's wrong with using p-values in this way?
  - Rejecting the null hypothesis doesn't mean that the proposed alternative hypothesis is true
  - $p < 0.05$ is arbitrary and gives a 1-in-20 chance of false positives
  - It encourages p-value hacking and has contributed to the replication crisis

Hypothesis testing: election polling example

Hypothesis: Pritzker's vote percentage is 53%

Experiment

		DATES	POLLSTER	SAMPLE	RESULT			NET RESULT
U.S. House	• IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	44%	46% Bost	Bost +2
	• IL-12	SEP 26-27	DCCC Targeting Team*	574 LV	Kelly	41%	42% Bost	Bost +1
Governor	• Ill.	SEP 24-29	Southern Illinois University	715 LV	Pritzke	49%	37% Rauner	Pritzker +22

Should we reject the hypothesis based on this data?

Fraction of “less extreme” samples

- Assuming that the hypothesis is true, what fraction of samples would have had sample means less extreme than what we observed?
- Define a test statistic $g = \frac{(\text{sample mean}) - (\text{hypothesized value})}{\text{standard error}}$
- If $N \geq 30$, we can say g comes from a standard normal distribution
- So, the fraction of “less extreme” samples $f = \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(-\frac{x^2}{2}\right) dx$

P-value: fraction of “more extreme” samples

- It is conventional in science to report the p-value of an experiment

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(-\frac{x^2}{2}\right) dx$$

- So, a p-value is the fraction of samples that would have had sample means **more** extreme than what we observed, assuming that the hypothesis is true

- By convention, if the p-value < 0.05, we should reject the hypothesis

P-value: election polling example

- Hypothesis: Pritzker's vote percentage is 53%
- Recall that we calculated sample mean 49% and standard error 1.9%
- So the test statistic $g = \frac{49-53}{1.9} = -2.11$
- And the p-value tells us to reject the hypothesis

$$p = 1 - \frac{1}{\sqrt{2\pi}} \int_{-2.11}^{2.11} \exp\left(-\frac{x^2}{2}\right) dx = 0.035 < 0.05$$

Our p-value is two-tailed

- Another way to write the p-value on the previous slide

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2.11} \exp\left(-\frac{x^2}{2}\right) dx + \frac{1}{\sqrt{2\pi}} \int_{2.11}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx$$

- This p-value is the sum of two probability tails
  - Probability that the sample mean is $\leq 49\%$ (the realized sample mean)
  - Probability that the sample mean is $\geq 57\%$

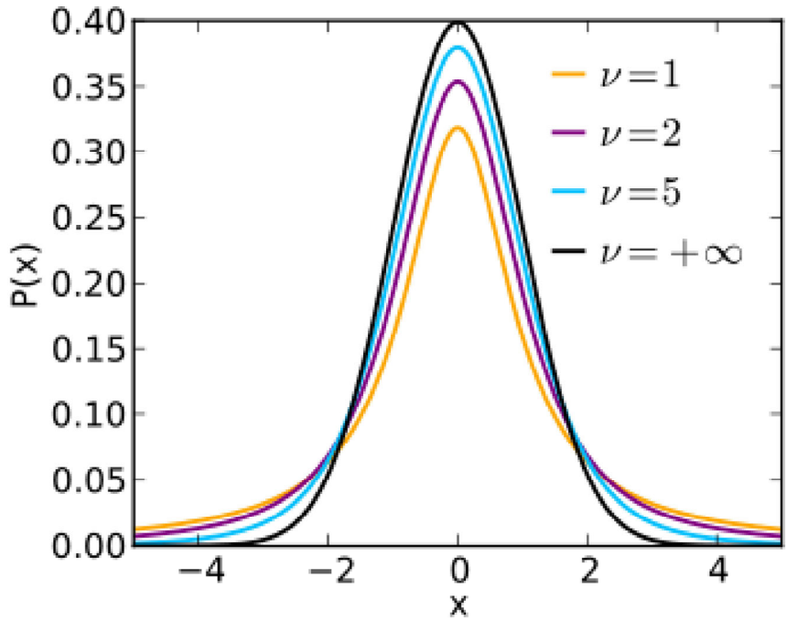
Be wary of one-tailed p-values

- It may be appropriate to calculate a one-tailed p-value if you are absolutely sure that the realized sample mean can only fall on one side of the hypothesized population mean

- Sometimes experimenters are tempted to use one-tailed p-values because they are smaller, especially if they are hoping to reject the null hypothesis. This is bad statistics!

P-values when $N < 30$

- If the sample size $N < 30$, we should not use the standard normal distribution to calculate p-values
- Instead we use Student's t-distribution with its parameter (called degrees of freedom) set to $N - 1$
- Student's t-distribution has heavier tails than the standard normal distribution, so the p-values are greater



Source: Wikipedia

Do two populations have the same mean?

Governor	Ill.	SEP 27-OCT 7	B+	Ipsos	2,000 A	Pritzker	42%	24%	Rauner	Pritzker +18
	Ill.	SEP 27-OCT 7	B+	Ipsos	968 LV	Pritzker	50%	30%	Rauner	Pritzker +20

Source: fivethirtyeight.com

- Ipsos surveyed 2000 adults, of whom 968 are likely voters
- So there are 1032 non-likely voters (NLV) in the sample
- We will later calculate Pritzker's vote margin among NLV as +16%.
- Hypothesis: the actual vote margins are the same for LV and NLV

Background: normal distribution properties

- Let X be a normal random variable with
  - mean μ
  - standard deviation σ

Notation

$X \sim N(\mu, \sigma)$

- Scaling property

If $k \neq 0$, $kX \sim N(k\mu, |k|\sigma)$

- Translation property

$X + c \sim N(\mu + c, \sigma)$

Background: sums and differences of normals

If $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$

and X_1 & X_2 are independent

- Sum property

$$X_1 + X_2 \sim N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

- Difference property

$$X_1 - X_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

The difference of two sample means

- Suppose we obtain samples from two populations $\{x\}$ and $\{y\}$
  - From a sample of size k_x from $\{x\}$, we get the sample mean $X^{(k_x)}$
  - From a sample of size k_y from $\{y\}$, we get the sample mean $Y^{(k_y)}$

- Define random variable $D = X^{(k_x)} - Y^{(k_y)}$ as the difference between the sample means

- If we hypothesize that $\text{popmean}(\{x\}) = \text{popmean}(\{y\})$, then

$$E[D] = E\left[X^{(k_x)}\right] - E\left[Y^{(k_y)}\right] = 0$$

Standard error of the difference

$\text{stderr} = \frac{\text{stdunbiased}}{\sqrt{N}}$

- Recall that standard error is the standard deviation of a sample mean
- By the property on the difference of normal random variables

$$\text{stderr}(D) = \sqrt{\text{stderr}(\{x\})^2 + \text{stderr}(\{y\})^2}$$

- Equivalently

$$\text{stderr}(D) = \sqrt{\frac{\text{stdunbiased}(\{x\})^2}{k_x} + \frac{\text{stdunbiased}(\{y\})^2}{k_y}}$$

P-value for comparing two means

- Define the test statistic

standard normal distribution

$$g = \frac{\text{mean}(\{x\}) - \text{mean}(\{y\})}{\text{stderr}(D)}$$

- If $k_x \geq 30$ and $k_y \geq 30$

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(-\frac{x^2}{2}\right) dx$$

Comparing two means: vote margin example

Governor	Ill.	SEP 27-OCT 7	B+	Ipsos	2,000 A	Pritzker	42%	24%	Rauner	Pritzker +18
	Ill.	SEP 27-OCT 7	B+	Ipsos	968 LV	Pritzker	50%	30%	Rauner	Pritzker +20

- For the 1032 non-likely voters (NLV)

Pritzker 34% Rauner 18% Vote margin: Pritzker +16

$$\frac{2000(0.42) - 968(0.50)}{2000 - 968} = 0.34$$

- Hypothesis: the actual vote margins are the same for LV and NLV

Standard error: vote margin example

- $(\text{stdunbiased}_{NLV})^2 =$

$$\frac{1032(0.34)(1-0.16)^2 + 1032(0.18)(-1-0.16)^2 + 1032(0.48)(0-0.16)^2}{1032-1} = 0.49$$

- $(\text{stdunbiased}_{LV})^2 =$ 0.76 (similar to above)

- $\text{stderr}(D) = \sqrt{\frac{0.49}{1032} + \frac{0.76}{968}} = 0.036 = 3.6\ \%$

P-value: vote margin example

- Hypothesis: the actual vote margins are the same for LV and NLV
- The test statistic $g = \frac{20-16}{3.6} = 1.11$
- The p-value tells us **not** to reject the hypothesis

$$p = 1 - \frac{1}{\sqrt{2\pi}} \int_{-1.11}^{1.11} \exp\left(-\frac{x^2}{2}\right) dx = 0.27 > 0.05$$