Advanced Statistics
DS2003 (BDS-4A)
Lecture 10

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

17 March, 2022

Previous Lecture

- Inference for a single proportion
- Difference of two proportions
  - Melting ice cap survey → bothered *a great deal* or not?
  - Pooled estimate of a proportion
  - CI and HT for proportions

Recap - comparing two proportions

- Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$

- Conditions:
  - independence within groups
    - random sample and 10% condition met for both groups
  - independence between groups
  - at least 10 successes and failures in each group
    - if not → randomization (Section 6.4)

$$SE_{(\hat{p}_1-\hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- for CI: use $\hat{p}_1$ and $\hat{p}_2$
- for HT:
  - when H_0: $p_1 = p_2$: use $\hat{p}_{pool} = \frac{\#\ suc_1 + \# suc_2}{n_1 + n_2}$
  - when H_0: $p_1 - p_2 =$ (some value other than 0): use $\hat{p}_1$ and $\hat{p}_2$
    - this is pretty rare

Reference - standard error calculations

	one sample	two samples
mean	$SE = \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
proportion	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- When working with means, it's very rare that σ is known, so we usually use *s*.
- When working with proportions,
  - if doing a hypothesis test, *p* comes from the null hypothesis
  - if constructing a confidence interval, use $\hat{p}$ instead

Comparing Two Proportions

- A study of births in Liverpool, England, investigated a possible association between parental smoking during pregnancy and the gender of the baby
- In a sample of 5045 babies born to non-smoking parents, 2685 were male
- In a sample of 363 babies born to heavy-smoking parents, 158 were male

2 Proportions: *Male babies & parents who smoke*

- A study of births in Liverpool, England, investigated a possible association between parental smoking during pregnancy and the gender of the baby

- In a sample of 5045 babies born to non-smoking parents, 2685 were male
  - $\hat{p}_1 = \frac{2685}{5045} = 0.532$

- In a sample of 363 babies born to heavy-smoking parents, 158 were male
  - $\hat{p}_2 = \frac{158}{363} = 0.435$

- $\hat{p}_1 - \hat{p}_2$ estimates $p_1 - p_2$ where:
  - p_1 is the true proportion for males born to non-smoking parents, and
  - p_2 is the true proportion for males born to heavy-smoking parents

2 Proportions: *Male babies & parents who smoke*

- Common points of interest:
  - Construct a CI for $p_1 - p_2$
  - Test the null hypothesis → $H_0: p_1 - p_2 = 0$
    - That is, no association between parental smoking and male birth rate

- Sampling distribution of $\hat{p}_1 - \hat{p}_2$:
  - Has a mean of $p_1 - p_2$
  - Has a standard deviation of $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
  - Is approximately normal if the sample sizes are large

2 Proportions: *Male babies & parents who smoke*

- The assumption of the 2 sample inference procedures on proportions:
  - We have independent simple random samples from the populations of interest
  - The sample sizes are large enough for the normal approximation to be reasonable
- A (1-α)100% CI for $p_1 - p_2$ is given by:
  - $\hat{p}_1 - \hat{p}_2 \pm\ z_{\frac{\alpha}{2}} * SE(\hat{p}_1 - \hat{p}_2)$
- Don't have p_1 or p_2 so we use estimators:
  - $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- Alternative hypothesis?
  - $H_A: p_1 < p_2$ or $H_A: p_1 > p_2$ or $H_A: p_1 \neq p_2$ (two-sided alternative)

2 Proportions: *Male babies & parents who smoke*

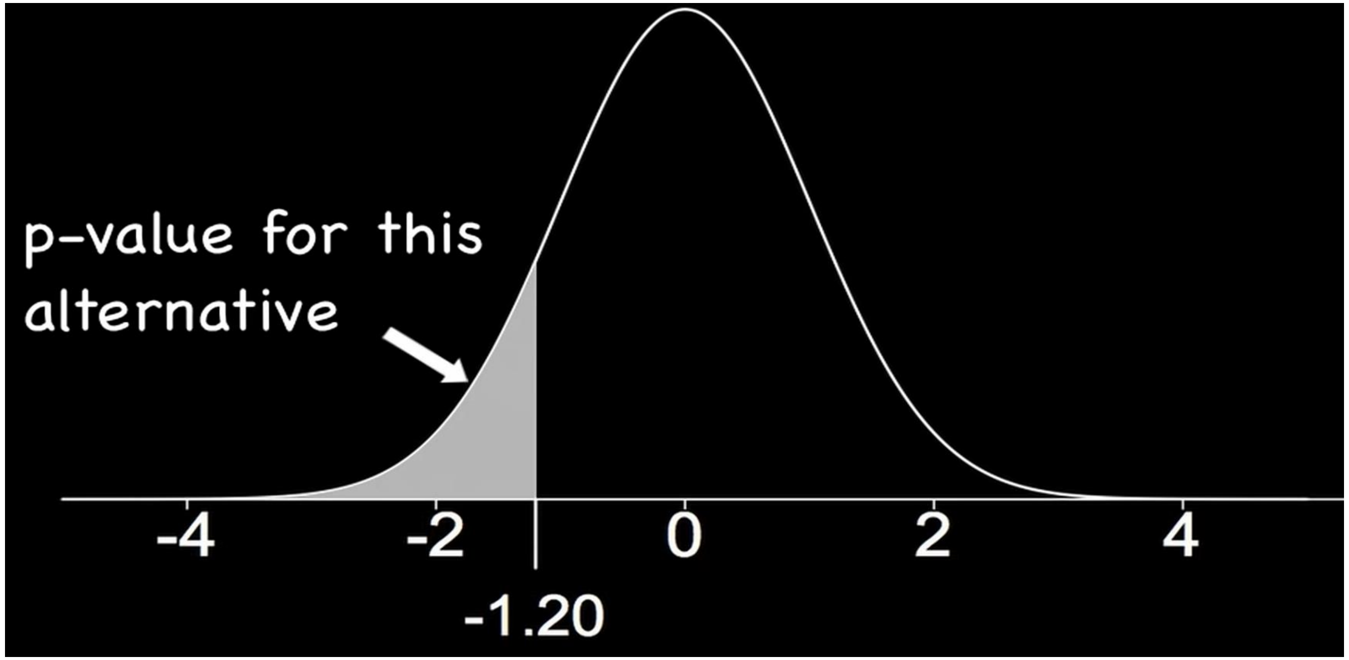
- To test null hypothesis $p_1 - p_2 = 0$ we calculate:
  - $z = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{p}_1 - \hat{p}_2)}$
- $\hat{p}$ is the pooled sample proportion:
  - $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ = # of individuals with the characteristic/# of total individuals

- If H_0 is true, the z-test statistic has approximately the standard normal distribution

Hypothesis testing for two proportions

- H_0: $p_1 = p_2$; H_A: $p_1 < p_2$

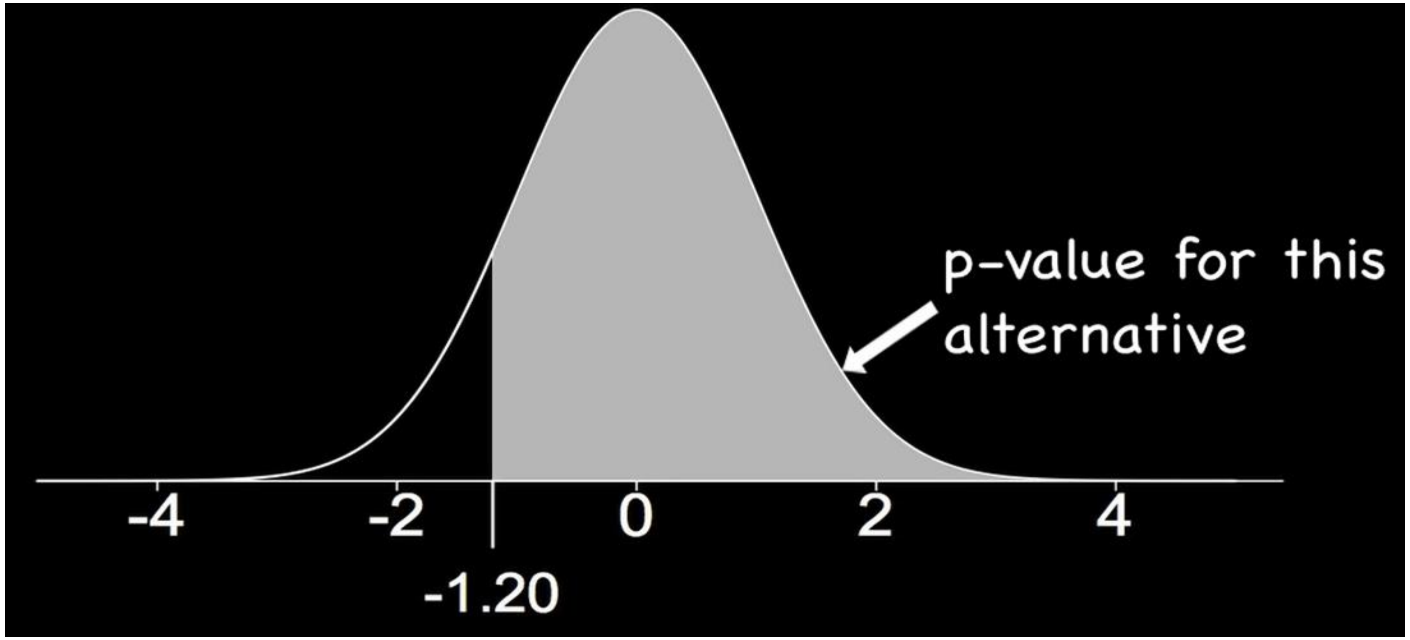
- Suppose Z = -1.2



Hypothesis testing for two proportions

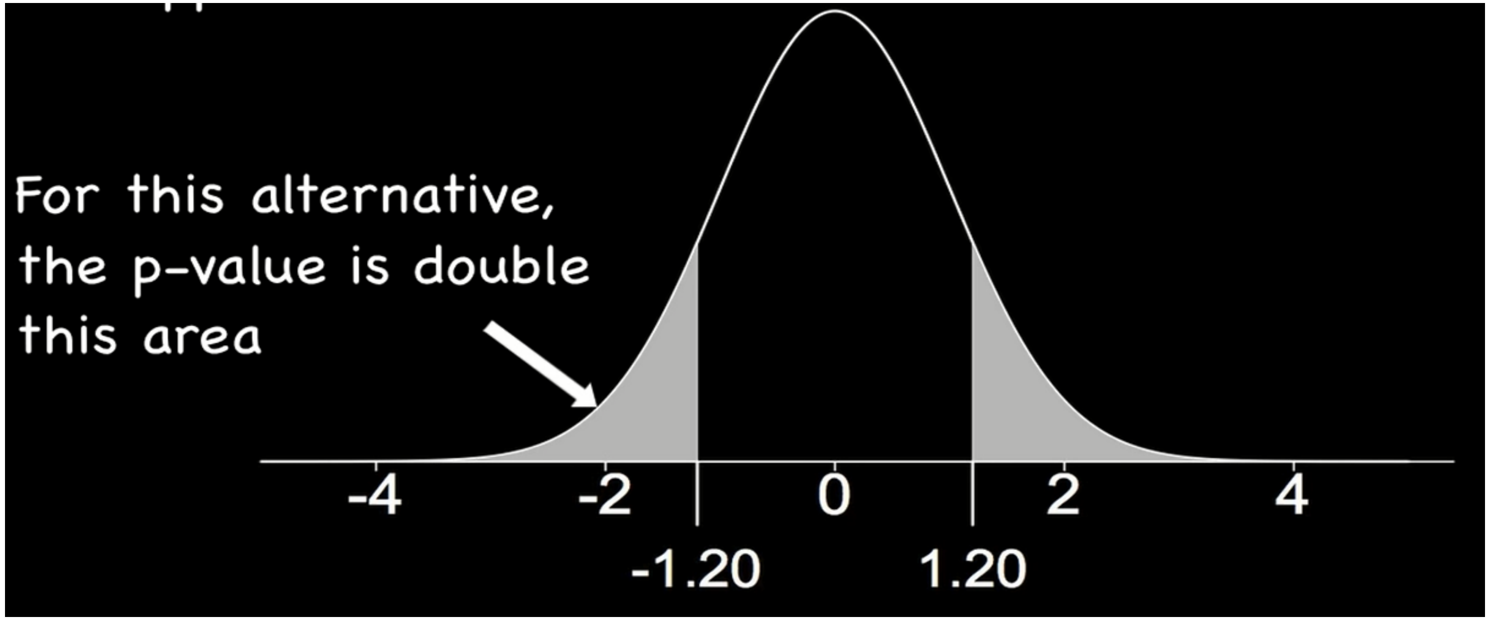
- H_0: $p_1 = p_2$; H_A: $p_1 > p_2$

- Suppose Z = -1.2



Hypothesis testing for two proportions

- H_0: $p_1 = p_2$; H_A: $p_1 \neq p_2$
- Suppose Z = -1.2



Drawing Conclusions

- Draw a conclusion in the usual ways:
  - A very small p-value gives very strong evidence against H_0
  - If we have a set significance level α, reject H_0 if p-value ≤ α

Example of *male babies* and *smoking parents*

$\hat{p}_1 = 0.532$ (non-smoking parents); $\hat{p}_2 = 0.435$ (heavy-smoking parents)

- H_0: $p_1 = p_2$ (true proportion of male births is same for both groups)
- H_A: $p_1 \neq p_2$

Z = 3.57, p-value = 0.00035 (strong evidence against H_0)

- A 95% Confidence Interval for : $p_1 - p_2$:

(0.044, 0.150)

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

$$\hat{p} = \frac{158 + 2685}{363 + 5045} = \frac{2843}{5408} = 0.5257$$

$$\frac{0.532 - 0.435}{\sqrt{\frac{0.5257(0.4743)}{5045} + \frac{(0.5257)(0.4743)}{363}}}$$

$$= \frac{0.097}{\sqrt{\frac{0.24934}{5045} + \frac{0.24934}{363}}} = \frac{0.097}{0.02713} = \underline{\underline{3.5747}}$$

95% C.I. $p_1 - p_2$:

$(0.044, 0.150)$

Sources

- openintro.org/os (Chapter 6)
- An Introduction to Inference for Two Proportions - YouTube