Stat 23400 Lecture 13: CI for Proportions and General Framework of Hypothesis Testing

Outline for Today

- Topics:
  - C.I. for Proportions (5.2 in OIS and 8.2 in MMSA).
  - Sample Size Calculation (8.1 in MMSA p.387).
  - Hypothesis Testing: General Framework (9.1 in MMSA).
- Hw7 will be posted on Canvas.

C.I. for Proportions

C.I. for Proportions

Suppose we are interested in the proportion p, percentage of individuals with some characteristic of a certain population (e.g., IL residents infected with covid-19).

We may

- Draw simple random sample of size n from the population.
- Let $X_i, i = 1, \ldots, n$ be binary Bernoulli variable in the sample, with 1 means success and 0 otherwise. (Here a "success" is an observation with the characteristic of interest)
- Estimate the unknown true population proportion p with the sample proportion $\widehat{p} = \bar{X} = (\sum_{i=1}^{n} X_i)/n$.

Sampling Distribution of $\hat{p}$

What is the **sampling distribution** of $\widehat{p} = (\sum_{i=1}^{n} X_i)/n$?

- $\sum_{i=1}^{n} X_i \sim Binom(n, p)$, so we have

$$\mathrm{E}[\hat{p}] = p\,, \text{ and } \mathrm{Var}(\hat{p}) = p(1-p)/n\,.$$

- *Normal approximation to Binomial distribution* tells us that, when n is *sufficiently large*,

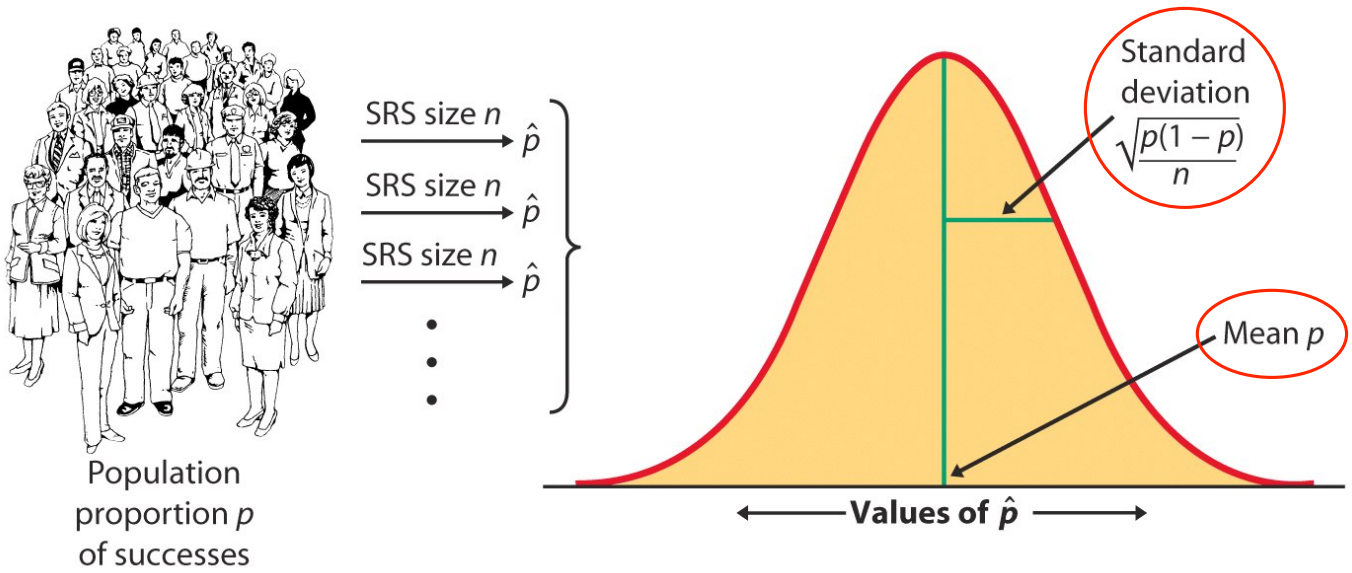
$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)\,.$$

Sampling Distribution of $\hat{p}$

What does this mean?

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right).$$

This means $\hat{p}$ approximately follows the normal distribution, if we repeatedly draw many SRSs of the same size n from the population.



Large-Sample Confidence Interval for p

An approximate $(1-\alpha)$ CI for the population proportion p is

$$\widehat{p} \pm z_{\alpha/2} SE \quad \text{where} \quad SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

- For various α, we can obtain the value $z_{\alpha/2}$ from the normal table, as follows.

Confidence level	90%	95%	99%
$z_{\alpha/2}$	1.645	1.960	2.576

- Remark: The exact SE should be $\sqrt{p(1-p)/n}$, but the unknown p is replaced with the estimate $\widehat{p}$. This large-sample CI is not very accurate, meaning the actual confidence level often falls below the nominal level $(1-\alpha)$.

Example: Side Effects of Pain Relievers

- Arthritis is a painful, chronic inflammation of the joints, so many arthritis patients rely on pain relievers, like Ibuprofen.
- However, Ibuprofen may induce side effects (like dizziness, muscle cramp, allergy, or even seizure) on some patients.
- A study interviewed 440 arthritis patients taking Ibuprofen, and found 23 had experienced side effects. Suppose the 440 patients is a SRS from the population of arthritis patients taking Ibuprofen.
- Find a 90% confidence interval for the population proportion p of arthritis patients who suffer some adverse symptoms.

Example: Side Effects of Pain Relievers

- The sample proportion $\hat{p}$ is
- The z^* for a 90% CI is $z_{\alpha/2} = 1.645$. So a 90% CI for p is
- Conclusion: With 90% confidence, between 3.5% and 6.9% of arthritis patients taking Ibuprofen will experience some adverse symptoms.

Discussion: Using the Large-Sample CI for Proportion

- **Conditions:**
  - The observations are (nearly) i.i.d. from the population studied.
    - If SRS, the sample size is at most 10% of the population size.
  - To use large sample CI, the sample size n is large enough. A rule of thumb is that
    - $n\widehat{p}$ and $n(1-\widehat{p})$ need to be both ≥ 10
- Why Not Using t-interval for Proportions?
  - If the sample size n is large enough to apply normal approximation to binomial, it is also large enough to apply CLT.
  - Even if we use the t-based C.I., the t_{n-1} distribution is very close to normal for large n,and therefore it is justified to do normal-based inference for proportions.

Sample Size Calculation

Sample Size Calculation: General Idea

- Sample size calculation is a very important aspect of any study at the stage of planning a study.
- How many people shall we test in order to estimate the percentage of the IL residents that are infected by covid-19, with margin of error $\pm 0.1\%$?
- How many patients shall we include in a clinical trial in order to estimate the drug effects with margin of error ± 0.01 unit?
- Sample size calculation tries to answer these questions.

Sample Size Calculation: Methods

How large the sample size n should be such that the margin of error m for a 100(1 - α)% CI is not larger than the prescribed margin of error m_0?

- (Z-based) Recall that the margin of error of our point estimate $\bar{X}$ is previously defined as

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- Then we can set

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq m_0.$$

- Solving for n, we get the sample size

$$\boxed{n \geq (z_{\alpha/2}\frac{\sigma}{m_0})^2.}$$

Example: Sample Size Calculation

E.x. The bursting strength of a cardboard has an unknown mean μ, and $\sigma = 0.8$. How large must the sample size be so that $\bar{X}$ estimate μ within 0.2 units with 99% confidence?

Sample Size Calculation for Proportions

Given a prescribed margin of error m_0, let us derive the sample size calculation for proportions.

$$n \geq \left(\frac{z^*}{m_0}\right)^2 \sigma^2 = \left(\frac{z^*}{m_0}\right)^2 p(1-p)\,.$$

But $\sigma^2 = p(1-p)$ is UNKNOWN. We need to make a guess.

Sample Size Calculation for Proportions

How to guess?

- We may conduct a small pilot study, or use prior studies or knowledge to get a range for possible values.
  - If we know the possible range of σ, choose the upper bound.
  - If we know the possible range of p, choose the bound that is closer to 0.5.
    E.g., if possible range of p is [0.1, 0.2], choose $p^* = 0.2$.
    if possible range of p is [0.85, 0.95], choose $p^* = 0.85$.
- For p, the most **conservative** approach is to choose $p = 0.5$ since $\sigma^2 = p(1-p)$ is the largest when $p = 0.5$.

Example: Sample Size Calculation for a Proportion

In a 1993 national survey, it was reported that 72.1% of freshmen responding to the survey were attending the college of their first choice. Suppose that $n = 500$ students responded to the survey.

- Find a 95% CI for the proportion p of college freshmen attending their first choice college.

- Suppose that given the CI, we want to conduct a survey which has a margin of error of 1% (i.e. $m_0 = 0.01$) with 95% confidence? How many people should we interview?

Hypothesis Testing: General Framework

Motivation

- Often, the decision making process requires testing for binary decisions.
- E.g. Does this gene impact height (Yes/No)?
- E.g. Does broccoli cause cancer (Yes/No)?
- E.g. Is Trump's phone source associated with negative words (Yes/No)?
- Today, we will talk about the general framework for **Hypothesis Testing**, a statistical tool for making binary decisions.

Hypothesis test

- A hypothesis test is an assessment of the evidence provided by the data in favor of (or against) some claim about the population.
- For example, suppose we perform a randomized experiment, take a random sample and calculate some sample statistic, say the sample mean.
- We want to decide if the observed value of the sample statistic is consistent with some hypothesized value of the corresponding population parameter.
- If the observed and hypothesized value differ (as they almost certainly will), is the difference due to an incorrect hypothesis or merely due to chance variation?

Can Dogs Smell Cancer?

Dogs Can Smell Cancer | Secret Life of Dogs | BBC

https://www.youtube.com/watch?v=e0UK6kkS0_M

These Dogs Can Detect Breast Cancer By Sniffing

https://www.youtube.com/watch?v=jlrgtWWwqZo

Data: Can Dogs Smell Bladder Cancer?

- The experiment[1] is **double blinded**, i.e., researchers blinded both dog handlers and experimental observers to the identity of urine samples.
- Each of the 6 dogs was tested with 9 trials. In each trial, one urine sample from a bladder cancer patient was randomly placed among 6 control urine samples.
- Outcome: In the total of 54 trials, the dogs made the correct selection 22 times.
  - The dogs were correct for $22/54 \approx 41\%$ of the time.
  - If the dogs just guessed at random, they were expected to be correct for $1/7 \approx 14\%$ of the time.
  - Is this difference (41% v.s. 14%) surprising?

[1] Olfactory detection of human bladder cancer by dogs: proof of principle study, *British Medical Journal*, vol. 329, September 25, 2004.

Two Competing Hypotheses

Let p be the probability that a dog makes the correct selection on a given trial.

- **Null hypothesis (H_0)**: $p = 1/7$
  "There is nothing going on."
  The dogs just guessed at random.
  - "null" means "nothing surprising is going on".
  - The dogs were just lucky to make more correct selections than expected.
- **Alternative hypothesis (H_A)**: $p > 1/7$
  "There is something going on."
  Dogs can do better than random guessing.

Weighing Evidence: Test Statistic

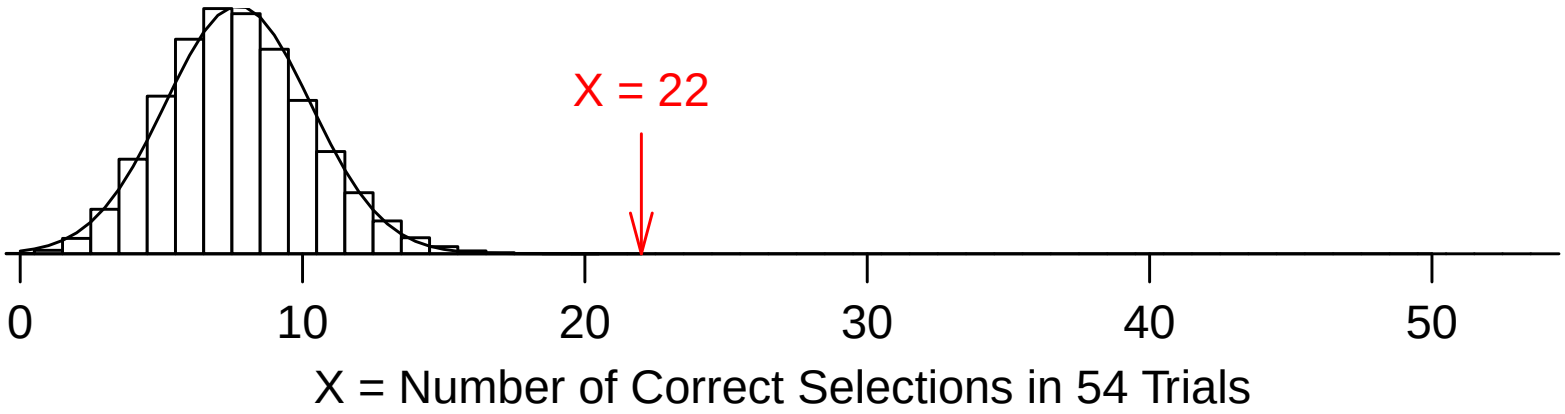
The next step of hypothesis testing is to weigh the evidence —

- *Could these data plausibly have happened by chance if the H_0 was true?*
- **Test Statistic:**
  - A summary of the data that best reflect the evidence for or against the hypotheses.
  - For this study, the test statistics we choose is X = the total number of correct selections in the 54 trials.
  - The larger X is, the stronger the evidence for H_A and against H_0 is.
  - The smaller X is, the stronger the evidence for H_0 and against H_A is.

Test Statistic

X = the total number of correct selections in the 54 trials.

If H_0 is true, then $X \sim Bin(n = 54, p = 1/7)$ (Why?)



Under H_0

$$P(X \geq 22) = \sum_{k=22}^{54} \binom{54}{k} (1/7)^k (6/7)^{54-k} \approx 1.86 \times 10^{-6}$$

```
> sum(dbinom(22:54,54,1/7))
[1] 1.861522e-06
```

If the dogs just guessed at random, they could be correct in 22 or more of the 54 trials for no more than 2 out of 1 million of the time.

The observed result was very unlikely to have occurred under the H_0 — strong evidence to against H_0, thus the data supports H_A.

p-value

p-value

Recall $H_0 : p = 1/7$ in the previous example. The probability $P(X \geq 22 \mid H_0) \approx 1.86 \times 10^{-6}$ is called the **p-value** of the test.

What is a **p-value**?

p-value

The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, *if the null hypothesis were true*.

- We know the distribution of $\bar{X}$ under H_0, so we can calculate the probability of seeing data as extreme or more extreme than $\bar{x}$ under H_0 using the sampling distribution of $\bar{X}$ under H_0.
- The p-value for the dog study is $P(X \geq 22)$ not $P(X = 22)$.

How do we interpret *p*-value?

- A small *p*-value (close to 0) means that the data would be very unlikely under H_0, providing evidence for H_A.
- A large *p*-value (not close to 0) means that the data would be likely under H_0, *not* providing evidence for H_A.
- The smaller the *p*-value is, the stronger the evidence against H_0.
- Generally, we reject H_0 if the *p*-value is below some level α. In this case, α is called the significance level of a test.
  - If the *P*-value $< \alpha$, we reject H_0.
  - If the *P*-value $> \alpha$, we don't reject H_0.
- Commonly used significance levels: $\alpha = 0.05$ and $\alpha = 0.01$
  - A test with *P*-value < 0.05 is said to be **(statistically) significant**
  - A test with *P*-value < 0.01 is said to be **highly significant**

Type 1 and Type 2 Errors

Type 1 and Type 2 Errors

In a hypothesis test, we make a decision about which of H_0 or H_A might be true, but our decision might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Nature State	H_0 true	✓	Type 1 Error
	H_A true	Type 2 Error	✓

- A **Type 1 Error** (often referred to as α) is rejecting the H_0 when it is true.
- A **Type 2 Error** (often referred to as β) is failing to reject the H_0 when it is false.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Consequences of Type 1 error and Type 2 error

Type 1 and type 2 errors are different sorts of mistakes and have different consequences

- Usually H_0 is the existing state, a thing we generally believe to be true.
- If H_0 is not rejected, usually it means the existing state is fine. No action needs to be taken.
- Rejecting H_0 means something we use to believe is overturned. For example, in criminal trial, one is not criminal unless proved guilty.
- A type 1 error introduces a false conclusion which could lead to an extremely high cost, e.g., *death penalty*.
- Thus we would like to control the Type 1 error at α level.
- A type 2 error — failing to recognize a scientific breakthrough — represents a missed opportunity for scientific progress.

Significance Level = Type 1 Error Rate

- **When H_0 is true**, there is only 5% chance to obtain a p-value $< 5\%$. $[\Pr(P \leq 0.05) = 0.05]$
- This means that, for those cases where H_0 is actually true, we won't incorrectly reject it more than 5% of those times **in the long run**.
- In other words, when using a 5% significance level, there is about 5% chance of making a Type 1 error if the H_0 is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha\,.$$

- This is why we prefer small values of α — increasing α increases the Type 1 error rate.
- However, significance level doesn't control Type 2 error rate

Failing to Reject $H_0 \neq H_0$ Is True

- When the evidence is not strong enough to reject the H_0, we don't say "we accept the H_0", but say "**we fail to reject the H_0.**"
- This is because the Type 2 error rate is usually not controlled and is usually quite big.
- There are philosophical reasons for this: lack of evidence against a hypothesis is not the same as evidence for a hypothesis — e.g. a "not guilty" verdict in court does not mean "innocent".
- There are practical reasons for this: If a scientist wanted to publish a result, he could make his desired hypothesis H_0 and then collect a very small sample size. He would usually fail to reject H_0 and could publish a lot of bad papers.

True or False

- If H_0 is rejected, then we can be certain that H_0 is false.
- If H_0 is rejected at 5% level, there is less than a 5% chance for H_0 to be true.

Reporting the p-Value

Reporting the p-Value

Don't simply report the conclusion of whether H_0 is rejected.
Attach the p-value.

- A p-value of 0.04 and a p-value of 0.000001 are not at all the same thing, even though H_0 will be rejected in both cases, but the strength of evidence are very different
- Simply reporting whether H_0 is rejected without p-value is like reporting the temperature as "cold" or "hot"
- It's much better to report the p-value and let people choose their own significance level, just like telling someone the temperature and let them decide for themselves whether they want to wear a coat

What's wrong?

Can you point out the mistakes in the following statement?

- A significance test rejected the null hypothesis that the sample mean is equal to 500.
- A test preparation company wants to test that the average score of its students on the ACT is better than the national average score of 21.2. The company states its null hypothesis to be $H_0 : \mu > 21.2$.
- A study summary says that the results are statistically significant and the p-value is 0.98.
- The z test statistic is equal to 0.018. Because this is less than $\alpha = 0.05$, the null hypothesis was rejected.

Conclusion of the Dogs Smell Bladder Cancer Study

- Recall that p-value is 1.86×10^{-6}. There is strong evidence that dogs have some ability to smell bladder cancer,
- However, the dogs were only correct 40% of the time, too low for practical application.
- Another study (M. McCulloch et al., Integrative Cancer Therapies, vol 5, p. 30, 2006.) considered whether dogs could be trained to detect whether a person has lung cancer by smelling the subjects' breath. In one test with 83 Stage I lung cancer samples, the dogs correctly identified the cancer sample 81 times.

Recap: Hypothesis Testing Framework

- We start with a **null hypothesis (H_0)**, the status quo.
- We also have an **alternative hypothesis (H_A)** that represents our research question, i.e. what we're testing for.
- We then collect data and often summarize the data as a **test statistic**, which is usually a measure gauging whether H_0 or H_A is more plausible.
- We then predict what the **test statistic** would be around under the assumption that the H_0 is true.
- If the **test statistic** is too far away from what the H_0 predicts, we then reject the H_0 in favor of the H_A.
  - We often computed a *p*-**value** based on the test statistic, which is the probability to obtain a test statistic at least as extreme as the one actually observed, assuming the H_0 is true.
  - If the *p*-value is too small, we then reject the H_0 in favor of the H_A.