STAT 22000 Lecture Slides
Hypothesis Testing About Population Means

Yibi Huang
Department of Statistics
University of Chicago

Outline

- Hypothesis Testing About Population Means (Section 4.3)
- Relationships Between Confidence Intervals and Hypothesis Tests (Section 4.3.2)
- Common Misunderstandings About Hypothesis Testing (Not in the textbook)

Hypothesis Tests about Population Means

Example: Number of College Applications

To know how many colleges students applied to, the dean of a certain university took a random sample of size 106 from their newly admitted students. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all freshmen in this university apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

Example: Number of College Applications – Hypotheses

- *Population*: all freshmen in this university
- The *parameter of interest* μ is the average number of schools applied to by all freshmen in this university
- There are two explanations why the sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- H_0 : $\mu = 8$ (the average number of colleges freshmen in this university have applied to is 8, as recommended)
- H_A : $\mu > 8$ (the average number of colleges freshmen in this university have applied to is > 8)

Wrong Ways to State H_0 and H_A

H_0 and H_A are **ALWAYS** stated in terms of population parameters, not sample statistics

Neither

$$H_0: \bar{x} = 8, \quad H_A: \bar{x} > 8$$

or

H_0 : average number of colleges applied in the sample is 8
H_A : average number of colleges applied in the sample is 9.7

is correct. The correct statements should be

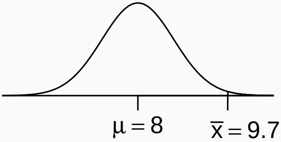
$$H_0: \mu = 8, \quad H_A: \mu > 8$$

Also please **clearly specify what is** μ.
e.g., μ is the average number of colleges freshmen in this university have applied to.

Number of College Applications — Test Statistic

By CLT, under H_0: $\mu = 8$, the sampling distribution of the sample mean is

$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{106}} = 0.68\right)$$

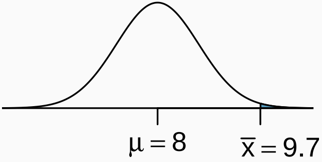


To gauge how unusual the observed sample mean $\bar{x} = 9.7$ is relative to its the hypothesized sampling distribution above, the *test statistic* we used is the z-statistic, which is the z-score of the sample mean relative to the distribution above

$$z\text{-statistic} = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{9.7 - 8}{7/\sqrt{106}} \approx 2.5$$

Number of College Applications — P-Value

Recall *p-value* is the probability of observing data such that the evidence for the H_A is at least as strong as our current data set ($\bar{x} > 9.7$), if in fact H_0: $\mu = 8$ were true.



$$p\text{-value} = P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 2.50) = 0.0062$$

- Since p-value is *low* (lower than 5%) we *reject H_0*.
- The data provide convincing evidence that freshmen in this university have applied to more than 8 schools on average.
- The diff. between the null value of 8 schools and observed sample mean of 9.7 schools is *not due to chance* or sampling variability.

Example: Number of College Applications – Conditions

As CLT is used in the hypothesis test above, we need to check the same conditions as we construct confidence intervals for the population mean.

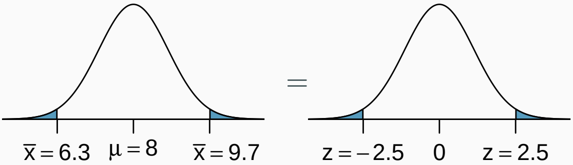
- Observations must be *independent*
  - Use your knowledge to judge if the data might be dependent
- The population distribution of the number of colleges students apply to should not be extremely skewed.
- In the z-statistic $= \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$, if the unknown population SD σ is replaced with the sample SD s, we need to further check that
  - sample size cannot be too small (at least 30)
  - no outliers & not too skewed $\Rightarrow$ Check the histogram of data!

Two-Sided Hypothesis Test

If the dean wanted to know whether the data provide convincing evidence that the average number of colleges applied is *different* than the recommended 8 schools, the alternative hypothesis would be different.

$$H_0 : \mu = 8$$
$$H_A : \mu \neq 8$$

In this case, a sample mean $\bar{x}$ far below 8 would also be evidence in favor of H_A. Hence the *p*-value would be the *two-tail* probability



p-value
$= 0.0062 \times 2$
$= 0.0124$

Recap: Hypothesis Testing for a Population Mean

1. Set the hypotheses
   - $H_0 : \mu = \mu_0$
   - $H_A : \mu <$ or $>$ or $\neq \mu_0$
2. Check assumptions and conditions
   - Independence
   - Normality: nearly normal population or $n \geq 30$, no extreme skew – or use the t distribution (Section 5.1)
3. Calculate a *test statistic* and a *p-value* (draw a picture!)

$$Z = \frac{\bar{x} - \mu_0}{SE}, \ where\ SE = \frac{s}{\sqrt{n}}$$

4. (Optional) Make a decision
   - If p-value $< \alpha$, reject H_0
   - If p-value $> \alpha$, do not reject H_0

Relationship Between Confidence Intervals and Two-Sided Hypothesis Tests

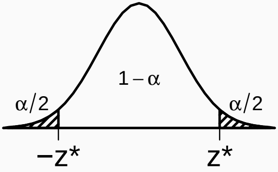
Confidence Intervals and Two-Sided Hypothesis Tests

For a two-sided test:

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu \neq \mu_0$$

the following are equivalent:

- p-value $> \alpha$ (and hence $\mathrm{H}_0 : \mu = \mu_0$ is not rejected at level α)
- $|z\text{-statistic}| = |(\bar{x} - \mu_0)/SE| < z^*$, where z^* is a value such that



- μ_0 is in the $100(1-\alpha)\%$ confidence interval for μ

$$\bar{x} - z^* SE < \mu_0 < \bar{x} + z^* SE$$

Example

Suppose in a study,

- 90% CI for μ is $(4.81, 11.39)$;
- 95% CI for μ: $(4.18, 12.02)$;
- 99% CI for μ: $(2.95, 13.25)$.

Then

- $H_0 : \mu = 4$ is rejected at 5% level but not at 1% level
  (2-sided *p*-value is between 1% and 5%)
  because 4 is in the 99% CI but not in the 95% CI
- $H_0 : \mu = 4.5$ is rejected at 10% level but not at 5% level
  because 4.5 is in the 95% CI but not in the 90% CI

Common Misunderstandings About Hypothesis Testing

In the lecture slide "General Framework of Hypothesis Testing", we have introduced a number of common misunderstanding about hypothesis testing

- Rejecting H_0 doesn't means we are 100% that H_0 is false. We might make Type 1 errors. Setting a significance level just guarantee we won't make Type 1 error too often
- P-value is not $P(H_0 \textit{ is true} \mid \textit{data})$ but it is $P(\textit{data} \mid H_0 \textit{ is true})$.

We are going to talk about more common misunderstanding about hypothesis testing here.

Failing to Reject H_0 Does Not Prove H_0 to Be True

Another mistake is to conclude from a high *p*-value that the H_0 is probably true

- We have said that if our *p*-value is low, then this is evidence that the H_0 is not true
- If our *p*-value is high, can we conclude that H_0 is true?
  - No, we could make a type 2 error when failing to reject H_0
  - Moreover, unlike type 1 error rate is controlled at a low level, type 2 error rate is usually quite high. It is quite often that H_0 is not true but the data fail to reject it.
- When we fail to reject H_0, often it just means the data are not able to distinguish between H_0 and H_A (because the data are to noisy, etc)

Real Example

- As an example, the Women's Health Initiative found that low-fat diets reduce the risk of breast cancer with a *p*-value of 0.07
- The *New York Times* headline: "*Study finds low-fat diets won't stop cancer*"
- The lead editorial claimed that the trial represented "*strong evidence that the war against fats was mostly in vain*" and sounded "*the death knell for the belief that reducing the percentage of total fat in the diet is important for health*"
- Failing to prove the effect of low-fat diets doesn't prove that low-fat diets have no effect

http://www.nytimes.com/2006/02/07/health/study-finds-lowfat-diet-wont-stop-cancer-or-heart-disease.html

Don't Take the 0.05 Significance Level Too Seriously

- A *p*-value of 0.049 and a *p*-value of 0.051 give nearly the same strength of evidence against H_0
- For example, in the highly publicized 2009 study involving a vaccine that may protect against HIV infection, the two-sided *p*-value is 0.08, and the one-sided *p*-value of is 0.04
- Much debate and controversy ensued, partially because the two ways of analyzing the data produce *p*-values on either side of 0.05
- Much of this debate and controversy is fairly pointless; both *p*-values tell you essentially the same thing — that the vaccine holds promise, but that the results are not yet conclusive

Hypothesis Testing Cannot Tell Us...

Hypothesis testing cannot tell us

- whether the design of a study is flawed
- whether the data is appropriately collected

So we cannot conclude from a small P-value about whether one variable has a causal effect on another variable or whether the conclusion can be generalized to a bigger population.

Garbage In → Garbage Out

Statistical Significance Does Not Mean Practical Importance

Another mistake is reading too much into the term "statistically significant"

- Saying that results are statistically significant informs the reader that the findings are unlikely to be due to chance alone
- However, it says nothing about the practical importance of the finding.
- E.g., rejecting the H_0: $\mu_1 = \mu_2$ just tells us $\mu_1 \neq \mu_2$, but not how big and how important $\mu_1 - \mu_2$ is. It is possible that the difference is too small to be relevant even if it is significant.
- Remedy: *Attach a confidence interval* for the parameter so that people can decide whether the difference is big enough to be relevant.

Example

A 95% CI for the average number of colleges freshmen in this university have applied is

$$\overline{x} \pm 1.96\frac{s}{\sqrt{n}} = 9.7 \pm 1.96\frac{7}{\sqrt{106}} \approx 9.7 \pm 1.3 = (8.4, 11.0).$$

from which one can decide whether the difference from 8 is big enough to be relevant.

Recap: Common Misunderstandings about Hypothesis Testing

- Rejecting H_0 doesn't mean we are 100% sure that H_0 is false. We might make Type 1 errors
- *P*-value is not the probability that the H_0 is true
- Failing to reject H_0 does not prove H_0 to be true
- Don't take the 0.05 significance level too seriously
- Hypothesis testing cannot tell us if data were collected properly or if the design of a study was flawed
- Statistical significance does not mean practical importance