0603 Slides

MA 116

June 2025

Quiz 2

Quiz 2 will only cover 9.1, 9.2, 10.1, 10.2.
There will be 6 questions.
Question 1. (9.1) Calculate an interval estimator for a population proportion p.
Question 2. (9.1) Given a margin of error E' we want to achieve, determine the sample size n needed for an interval estimator of p to have at most this error.
Question 3. (9.2) Calculate an interval estimator for a population mean $\overline{x}$.
Question 4. (10.1) Basic concepts of hypothesis testing: H_0, H_1, Type of errors, how to draw conclusion to a hypothesis test.
Question 5. (10.2) Classical method of hypothesis test for a population proportion: assume H_0 holds, calculate test statistics z_0 to see whether z_0 falls into critical region.
Question 6. (10.2) Confidence interval method of hypothesis test for a population proportion.

Review: 0602 lecture

9.2 Question type

To calculate an interval estimator of a population mean, first determine if the distribution of variable t can be approximated by the Student's t-distribution of some $df = n - 1$, then use the formulas to calculate E.

10.1

Determine H_0, H_1, test type, error type, state the conclusion of hypothesis test.

10.2 Classical approach

To do a hypothesis test about a population proportion, first determine H_0, H_1, test type. Assume H_0 is true. Verify that under this assumption the distribution of $\hat{p}$ is normal. Then get a distribution of $\hat{p}$. Change to a standard normal variable z. Determine the critical z value(s) and critical region under the standard normal curve. Calculate the test statistic z_0. If z_0 falls into a critical region, we reject H_0.

Review: 0602 lecture

10.2 Confidence interval approach

To do a hypothesis test about a population proportion, first determine H_0, H_1, test type. Assume H_0 is true. Verify that under this assumption the distribution of $\hat{p}$ is normal. Then use formulas to calculate E and get an interval estimator $\hat{p} \pm E$. If the p given by H_0 does not lie in this interval, reject H_0.

10.3

To do a hypothesis test about a population mean, first determine H_0, H_1, test type. Assume H_0 is true. Change variable $\overline{x}$ to a new variable t. Verify that under this assumption the distribution of t is approximately Student's t-distribution (two situations!). Determine the critical t value(s) and critical region under the Student's t-distribution curve. Calculate the test statistic t_0. If t_0 falls into a critical region, we reject H_0.

We want to know if Generaion Z has a higher average phone screen time than average Americans, given that the average phone screen time of Americans is 5 hours. Let's do a hypothesis test with a level of significance $\alpha = 0.05$.

Suppose we obtain a random sample of size 36 from Generaion Z Americans with a sample mean $\overline{x} = 6.5$, sample variance $s = 1.5$.

What are my H_0 and H_1?

10.3 Hypothesis test for a population mean

We want to know if Generaion Z has a higher average phone screen time than average Americans, given that the average phone screen time of Americans is 5 hours. Let's do a hypothesis test with a level of significance $\alpha = 0.05$.

Suppose we obtain a random sample of size 36 from Generaion Z Americans with a sample mean $\overline{x} = 6.5$, sample variance $s = 1.5$.

$H_0 : \mu = 5$; $H_1 : \mu > 5$. Right-tailed test.

How can we determine the critical region and whether our sample statistic falls into the critical region? We again asssume H_0 holds and change to variable t via

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}.$$

Our variable t is $\dfrac{\overline{x}-5}{s/6}$.

Since $36 > 30$, the distribution of $t = \dfrac{\overline{x}-5}{s/6}$ is approximately standard normal AND approximately Student's t-distribution with $df = 35$. Given $\alpha = 0.05$ and the test is right-tailed, the critical value is $t_{0.05}\,(df = 35)$. Look up the Student's t-distribution table, this value is 1.69. The critical region is then the area under the df=35 Student's t-distribution curve with $t \geq 1.69$.

It remains to calculate our test statistic and see if it falls into out critical region.

Today's new content

1. Set up background: Inference about two population proportions
2. Distribution of the difference between two proportions
3. Testing hypothesis regarding two population proportions
4. Interval estimator for the difference between two population proportions

In clinical trials of Nasonex, 3774 adult and adolescent allergy patients (patients 12 years and older) were randomly divided into two groups. The patients in group 1 (experimental group) received 200 µg of Nasonex, whereas the patients in group 2 (control group) received a placebo. Of the 2013 patients in the experimental group, 547 reported headaches as a side effect. Of the 1671 patients in the control group, 368 reported headaches as a side effect. It is known that more than 10 million Americans who are 12 years and older are allergy sufferers. Is there evidence to conclude that the proportion of Nasonex users who experienced headaches as a side effect is greater than the proportion in the control group at the $\alpha = 0.05$ level of significance?

Quantity we care about is p_1 vs. p_2.

Sampling distribution of the difference between two proportions

Population sizes: N_1, N_2.
Sample sizes: n_1, n_2.
Variable: $\hat{p}_1 - \hat{p}_2$.

The mean of this variable is

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2,$$

and the standard deviation of this variable is

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

These two formulas come mathematically from the mean and standard deviation formula of the variable $\hat{p}$, so $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ always holds, while $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ required $n_1 < 0.05N_1$ and $n_2 < 0.05N_2$.

Suppose $n_1 < 0.05N_1$, $n_2 < 0.05N_2$, $n_1p_1(1-p_1) \geq 10$, $n_2p_2(1-p_2) \geq 10$. **Then the variable $\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2$ has a distribution that is approximately normal.** In the case that we do not know p_1 and/or p_2 but we want to describe the distribution of $\hat{p}_1 - \hat{p}_2$, we may replace the condition that $n_1p_1(1-p_1) \geq 10$, $n_2p_2(1-p_2) \geq 10$ by $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$. (The reason is varying p a bit does not have big effect on the value of $np(1-p)$.)

Testing hypothesis regarding two population proportions

Condition. To do hypothesis test like this, we require that the two samples are independently obtained using random sampling. I.e. Choosing an individual from the second population does not depend on individual chose from the first population.

Usual set-up

In clinical trials of Nasonex, 3774 adult and adolescent allergy patients (patients 12 years and older) were randomly divided into two groups. The patients in group 1 (experimental group) received 200 μg of Nasonex, whereas the patients in group 2 (control group) received a placebo. Of the 2013 patients in the experimental group, 547 reported headaches as a side effect. Of the 1671 patients in the control group, 368 reported headaches as a side effect. It is known that more than 10 million Americans who are 12 years and older are allergy sufferers. Is there evidence to conclude that the proportion of Nasonex users who experienced headaches as a side effect is greater than the proportion in the control group at the $\alpha = 0.05$ level of significance?

1. Determine H_0, H_1, test type.
2. Assume H_0 is true. Check if the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.
3. If normal, change to a standard normal variable z using some approximations. (Keep assuming H_0 is true.)
4. Determine the critical value(s) and the critical region on the standard normal curve diagram. (Keep assuming H_0 is true.)
5. Calculate the test statistic z_0. If z_0 falls into the critical region, reject H_0.

Example.

In clinical trials of Nasonex, 3774 adult and adolescent allergy patients (patients 12 years and older) were randomly divided into two groups. The patients in group 1 (experimental group) received 200 µg of Nasonex, whereas the patients in group 2 (control group) received a placebo. Of the 2013 patients in the experimental group, 547 reported headaches as a side effect. Of the 1671 patients in the control group, 368 reported headaches as a side effect. It is known that more than 10 million Americans who are 12 years and older are allergy sufferers. Is there evidence to conclude that the proportion of Nasonex users who experienced headaches as a side effect is greater than the proportion in the control group at the $\alpha = 0.05$ level of significance?

1. $H_0 : p_1 = p_2$ vs. $H_1 : p_1 > p_2$. Test is two-tailed.
2. Assume H_0 is true. Check if the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal: $x_1 = 547$, $n_1 = 2103$, $x_2 = 368$, $n_2 = 1671$, so $\hat{p}_1 = 0.26$, $\hat{p}_2 = 0.22$. Indeed, $n_1\hat{p}_1(1 - \hat{p}_1) = 2103 \cdot 0.26 \cdot 0.74 = 404.1 \geq 10$ and $n_2\hat{p}_2(1 - \hat{p}_2) \geq 10$. We claim that the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal. Do we know the mean and standard deviation of this normal curve?

Example.

1. $H_0 : p_1 = p_2$ vs. $H_1 : p_1 > p_2$. Test is two-tailed.
2. Assume H_0 is true. We find out that the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal. By our assumption, $p_1 = p_2$, so we may let $p = p_1 = p_2$. Notice that

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = 0$$

and

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

Do we know the standard deviation $\sigma_{\hat{p}_1 - \hat{p}_2}$ of this normal curve now? No, because we do not know p.

Example.

1. $H_0 : p_1 = p_2$ vs. $H_1 : p_1 > p_2$. Test is two-tailed.
2. Assume H_0 is true. We find out that the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = 0$$

and

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

We approximate p by $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ (what's called a pooled estimate of p.) Now we can fully describe the normal curve that approximates the distribution of $\hat{p}_1 - \hat{p}_2$.

Example.

1. $H_0: p_1 = p_2$ vs. $H_1: p_1 > p_2$. Test is two-tailed.
2. Assume H_0 is true. We find out that the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = 0$$

and

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

3. Then we are finally able to change to a standard normal variable z by $z = \dfrac{\hat{p}_1 - \hat{p}_2 - 0}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$. (We know every value in the denominator.)

Example.

1. $H_0 : p_1 = p_2$ vs. $H_1 : p_1 > p_2$. Test is two-tailed.
2. Assume H_0 is true. We find out that the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.
3. $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$. In this example, we have a variable
$$z = \frac{\hat{p}_1 - \hat{p}_2}{0.014}$$
. Let's find critical value(s). Since this is a right-tailed test and $\alpha = 0.05$, $z_{0.05} = 1.645$, so the critical region is $z \geq 1.645$.
4. Let's calculate our test statistic.
$z_0 = \dfrac{\hat{p}_1 - \hat{p}_2}{0.014} = \dfrac{0.26 - 0.22}{0.014} = 2.85 > 1.645$.
5. **Draw conclusion.** There is sufficient evidence at $\alpha = 0.05$ level of significance to conclude...

Conclusion. There is sufficient evidence at $\alpha = 0.05$ level of significance to conclude the proportion of individuals 12 years and older taking 200 μg of Nasonex who experience headaches is greater than the proportion of individuals 12 years and older taking a placedo who experience headaches.

Interval estimator for the difference between two population proportions

If $\hat{p}_1$ and $\hat{p}_2$ are checked to have approximately normal distribution, then we may use the following formula

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