Statistics ST2334 Topic 7: Inferences Concerning a Mean (part b)

Hypothesis Testing Framework, Hypothesis Test of one mean, p-value, Relationship between Hypothesis tests and Confidence Intervals.

2011/2012 Semester 2

Hypothesis Testing

- Sometimes, we need to decide whether a statement concerning a parameter is true or false.
- Example: NUS students on average have higher IQ than the general population (100).
- It is difficult/expensive to ask every NUS student to take an IQ test. So we take a sample!
- Suppose the sample average is 110. Does that mean we're right?
- What if the sample average is 101? 100.1? Does sample size matter?

How to do a hypothesis test

There are five main steps to hypothesis testing.

1. Set your competing hypotheses: Null and Alternative.
2. Determine level of significance.
3. Identify test statistic, distribution and rejection criteria.
4. Compute the test statistic value based on your data.
5. Conclusion.

1. Null Hypothesis vs Alternative Hypothesis

- Our goal is to decide between two competing hypotheses.
- Often times, the two hypotheses are not on equal footing.
- For example, in a trial, we are trying to decide if the defendant is innocent or guilty. Here, we presuppose innocence unless proven otherwise.
- Similarly, in our above example, the onus is on the person making the statement to show that NUS students have higher IQ. Without evidence, we are inclined to believe NUS students, being part of the general population, to have the same IQ on average.
- In general, we assume the null hypothesis to be true unless there is overwhelming evidence against it.

1. Null Hypothesis vs Alternative Hypothesis

- We let the hypothesis that we want to prove be the alternative hypothesis.
- The idea is to collect evidence to show how improbable the null hypothesis is.
- We can then reject the null hypothesis, and claim the alternative hypothesis.
- We usually phrase the hypotheses in terms of population parameters. For example, let μ be the average IQ of NUS students.

$$H_0 : \mu = 100$$

$$H_1 : \mu > 100$$

1. One sided vs Two Sided Alternative

- The above is an example of a one-sided hypothesis test. We don't particularly care if $\mu < 100$: the goal here is just to show NUS students have higher IQ.
- In some other cases, it is more natural to do a two-sided hypothesis test. For example, let p be the probability of heads for a particular coin. You want to test if the coin is fair, it is equally problematic if p was larger or smaller. Hence our hypotheses would be

$$H_0 : p = 0.5$$

$$H_1 : p \neq 0.5$$

2. Significance Level

- For any test of hypothesis problem, there are two possible conclusions:
  - Reject H_0 and therefore conclude H_1
  - Do not reject H_0 and therefore conclude H_0
- Whatever decision is made, there is a probability of making an error.

	Not reject H_0	Reject H_0
H_0 is true	Correct Decision	Type I error
H_0 is false	Type II error	Correct Decision

- The probability of making a Type I error is called the level of significance α.

2. Type I vs Type II Error

- Note that the two types of error are at odds with each other.
- If you try to minimize Type I error, for example by rejecting H_0 only if there is very very strong evidence, then you naturally will be more prone to Type II error, that is to not reject H_0 when it is false.
- Similarly, If you were to minimize Type II error, that is to be reject H_0 easily, you will be more likely to commit Type I error if H_0 is true.
- To deal with this, we set a maximum probability of making one mistake and then try to minimize the other. Since making our point hinges on rejecting H_0, we would like to be clear on the probability of falsely rejecting H_0. Hence, we declare α.

3. Test Statistic, Distribution and Rejection Region

- Test statistic and its distribution are the main issue in the problem of tests of hypotheses
- They are derived based on the sample.
- The choice is often natural, but in some cases, we may need to choose between different test statistics (out of the scope of this class).
- The test statistic serves to quantify just how unlikely it is to observe the sample assuming the null hypothesis is true.
- We designate a "rejection region", which we deem to be too unlikely.

4 & 5. Calculation and Conclusion

- We now plug in our sample data into our test statistic.
- We check if it is within our rejection region.
  - If it is, our sample was too improbable assuming H_0 is true, hence we reject H_0 and conclude H_1 is true.
  - If it is not, we didn't accomplish anything. We failed to reject H_0 and hence fall back to our original assumption that H_0 is true.
- Note that in the latter case, we did not "prove" H_0 is true. Hence, it is prudent to use the term "fail to reject H_0" instead of "accept H_0."

Hypothesis test of one mean

- Let's apply our hypothesis steps to testing a population mean.
- For now, we further assume our data is from Case I: σ known, data normal.
- Step 1:
  - Null hypothesis is always H_0: $\mu = \mu_0$.
  - We use two-sided alternative hypothesis as an example: H_1: $\mu \neq \mu_0$.
- Step 2: set level of significance: α=0.05.
  - why 0.05?
  - because Fisher says so

Hypothesis test of one mean: Case I, two-sided

- Step 3: Test Statistic
  - Statistic & its distribution: with σ being known and population being Normal,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

When H_0 is true, i.e. $\mu = \mu_0$, the above statistic becomes

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

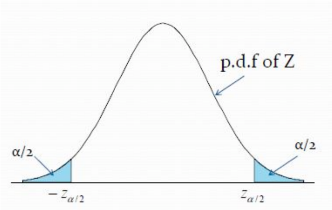
which serves as our test statistic.

Hypothesis test of one mean: Case I, two-sided

- Step 3 (continued)
- Rejection region: intuitively, we should reject H_0 when $\bar{X}$ is too large or too small compare with μ_0, that is, when Z is too large or too small. In theory, the probability $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$ is α i.e.

$$P(|Z| > z_{\alpha/2}) = \alpha$$

the rejection region is $|z| > z_{\alpha/2}$



Hypothesis test of one mean: Case I, two-sided

- Step 4, z should be computed from the statistic above based upon the observed sample.
- Step 5, the conclusion
  - Check whether z is located within rejection region. If so, reject H_0, otherwise do not reject H_0.

p-value

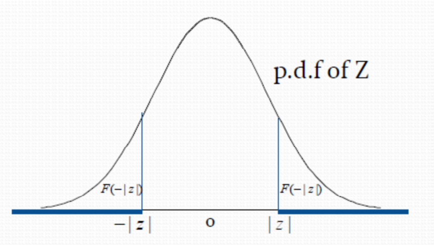
- The above technique introduced by Fisher is based on a pre-declared significance level α.
- Today, there is little reason to stick to the arbitrary $1\%, 5\%$ levels that Fisher suggested.
- We can instead use the idea of the p-value.
- The p-value is the "observed significance level". It is the probability of observing our data or worse, assuming H_0.
- p-value being lower than the significance level if and only if our test statistic is in the rejection region.

p-value in above case

- Suppose our computed test statistic was z. Then a "worse" result would be if $Z > |z|$ or $Z < -|z|$, i.e. $|Z| > |z|$.
- So the p-value is given by

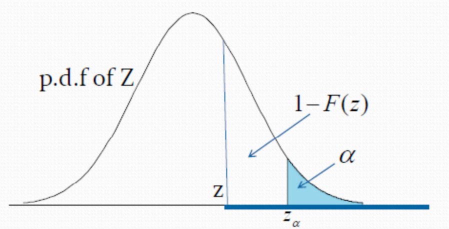
$$p\text{-value} = P(|Z| \geq |z|) = 2\Phi(-|z|).$$

- and our rejection criteria would be If p-value $> \alpha$, do not reject H_0, otherwise reject H_0.



Steps for one-sided test

- Suppose instead in H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$.
- Similar steps can be used to address this problem, we only need to do the following changes:
  - Step 1, H_1 is replace with H_1: $\mu > \mu_0$.
  - Step 3: test statistic and its distribution are kept the same. The rejection region should be replaced with $z > z_\alpha$, since now, we should reject only when $\bar{x}$ is large.



One sided vs Two sided and p-value

- Note that for the one sided test, p-value $= 1 - \Phi(z) = \Phi(-z)$, i.e. only the area in the right tail is used.
- Similarly, if the alternative hypothesis was $\mu < \mu_0$, only the area in the left tail is used.
- This is in contrast to the two-sided test where both tails are used. (Both sides are considered worse results than our sample).
- The effect is that the p-value is halved for one sided tests. That is all.
- In practice, it is not crucial if you use one or two sided tests. What is important is to specify carefully if you are doing a one or two sided test when reporting the p-value.

Example: Midterm Exam Score

- Recall the midterm exam scores example in our Topic 6.
- The data he obtained are 20, 19, 24, 22, 25.
- If the lecturer announced that the variance of the exam score over the class is 5 (just believe that this is the truth).
- Remember that the exam scores can be well approximately by a normal distribution.
- The student used the five steps to test at α significance level whether the average midterm score is not 16.

Example: Midterm Exam Score

- Step 1: H_0: μ=16 versus H_1: $\mu > 16$
- Step 2: choose $\alpha = 0.01$.
- Step 3: Now that $\sigma = \sqrt{5}$ is known, data are normal, n=5. Therefore test statistic and its distribution is

$$Z = \frac{\bar{X} - 16}{\sqrt{5}/\sqrt{n}} \sim N(0, 1)$$

  rejection region: $Z > z_\alpha$, that is $|z| > 2.33$.
- Step 4: $z = (22 - 16)/(\sqrt{5}/\sqrt{5}) = 6$.
- Step 5: 6 is located within rejection region, therefore H_0 is rejected. (Or p-value $= 0.0000003 < \alpha$, reject H_0)

Other Cases

- Again, we use our knowledge of the sampling distribution to determine the test statistic for other distribution.
- For example, let's say the lecturer didn't announce the variance. i.e. σ is unknown. We are now in Case III.
- Step 1 and Step 2 are not changed.
- Step 3: Now that σ is known, data are normal, n=5.

$$t = \frac{\bar{X} - 16}{S/\sqrt{n}} \sim t(n-1) = t(4)$$

  rejection region: $t > t_\alpha$.
- Step 4: $t = (22 - 16)/(2.55/\sqrt{5}) = 5.26$.
- Step 5: 5.26 is contained within rejection region, we reject H_0 (or p-value $= 0.00313 < \alpha$, reject H_0)

Example: Department Store

- A department store manager determines that a new billing system will be cost-effective only if the mean monthly account is more than \$170. It is known that the accounts has standard deviation \$65.
- A random sample of 400 monthly accounts is drawn, for which the sample mean is \$178.
- Can we conclude that the new system will be cost-effective?

Example: Department Store

- Step 1:

$$H_0 : \mu = 170$$

$$H_1 : \mu > 170$$

- Step 2: Choose $\alpha = 0.05$
- Step 3: We are in Case II. We use the following test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Under H_0, by CLT we have, $Z \sim N(0, 1)$
At a 5% significance level (i.e. $\alpha = 0.05$), we get

$$z_\alpha = z_{0.05} = 1.65$$

- Step 4:

$$n = 400, \quad \bar{x} = 178, \quad \sigma = 65, \quad \alpha = 0.05$$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{178 - 170}{65/\sqrt{400}} = 2.46 (> z_\alpha)$$

- Step 5: Therefore, we reject the null hypothesis and conclude that the mean monthly account is more than \$170.

The Relation between Tests and Confidence Intervals

- We use Case III as an example. Recall that in Case III, the (1-α) C.I. for μ is given by

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

- If we rearrange the above inequality, we obtain

$$-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2}$$

- On the other hand, the rejection region of two-sided test of hypothesis problem in case III is $|t| > t_{\alpha/2}$

The Relation between Tests and Confidence Intervals

- When the $(1-\alpha)$ C.I. contains μ_0, i.e.

$$-t_{\alpha/2} \leq \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \leq t_{\alpha/2}$$

  that is, the computed statistic value $-t_{\alpha/2} \leq t \leq t_{\alpha/2}$, which in turn means t is not located within the rejection region. H_0 will not be rejected at level α.
- Similarly, when C.I. does not contain μ_0 , then, $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$, indicating that t is contained by the rejection region. H_0 will be rejected.
- Therefore, C.I. can be used to perform a two-sided test.

Example: Use C.I. to Perform A Test

- Back to the midterm exam scores example. Assume that the lecturer didn't announce the variance. i.e. σ is unknown. Case III should be used. The student constructed 99% (i.e. $\alpha = 0.01$) C.I. for the average score of students for the midterm:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 22 \pm 4.604\frac{2.55}{\sqrt{5}} = (16.75, 27.25)$$

- Note that the interval does not contain 16, therefore, the following test of hypothesis problem should be rejected: H_0: μ=16 versus H_1: $\mu \neq$16.
- However, the following test of hypothesis problem should not be rejected: H_0: μ=17 versus H_1: $\mu \neq$17.