LECTURE 6: HYPOTHESIS TESTING

ECEN 321

Engineering Statistics

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

Introduction

- Recall: We discussed an example in Chapter 5 about microdrills.
- Our sample had a mean of 12.68 holes drilled and standard deviation of 6.83.
- Let us assume that the main question is whether or not the population mean lifetime μ is greater than 11.
- We can address this by examining the value of the sample mean. We see that our sample mean is larger than 11, but because of uncertainty in the means, this does not guarantee that $\mu > 11$.

Hypothesis

- We would like to know just how certain we can be that $\mu > 11$.
- A confidence interval is not quite what we need.
- The statement "$\mu > 11$" is a **hypothesis** about the population mean μ.
- To determine just how certain we can be that a hypothesis is true, we must perform a **hypothesis test**.

Section 6.1: Large-Sample Test for a Population Mean

- The **null hypothesis** says that the effect indicated by the sample is due only to random variation between the sample and the population. We denote this with H_0.
- The **alternative hypothesis** says that the effect indicated by the sample is real, in that it accurately represents the whole population. We denote this with H_1.
- In performing a hypothesis test, we essentially put the null hypothesis on trial.

Procedure

- We begin by assuming that H_0 is true, just as we begin a trial by assuming a defendant to be innocent.
- The random sample provides the evidence.
- The hypothesis test involves measuring the strength of the disagreement between the sample and H_0 to produce a number between 0 and 1, called a ***P-value***.

P-Value

- The *P*-value measures the plausibility of H_0.
- The smaller the *P*-value, the stronger the evidence is against H_0.
- If the *P*-value is sufficiently small, we may be willing to abandon our assumption that H_0 is true and believe H_1 instead.
- This is referred to as **rejecting** the null hypothesis.

Steps in Performing a Hypothesis Test

1. Define H_0 and H_1.
2. Assume H_0 to be true.
3. Compute a **test statistic**. A test statistic is a statistic that is used to assess the strength of the evidence against H_0.

Steps in Performing a Hypothesis Test

4. Compute the *P*-value of the test statistic. The *P*-value is the probability, assuming H_0 to be true, that the test statistic would have a value whose disagreement with H_0 is as great as or greater than what was actually observed. The *P*-value is also called the **observed significance level**.
5. State a conclusion about the strength of the evidence against H_0.

Example 1

A scale is to be calibrated by weighing a 1000 g test weight 60 times. The 60 scale readings have mean 1000.6 g and standard deviation 2 g. Find the *P*-value for testing $H_0 = 1000$ versus $H_1 \neq 1000$.

Answer

The null hypothesis specifies that μ (the population mean reading) is equal to a specific value. For this reason, values of the sample mean that are either much larger or much smaller than μ will provide evidence against H_0. We assume that H_0 is true, and that therefore the sample readings were drawn from a population with mean $\mu = 1000$. We approximate the population standard deviation with $s = 2$.

Answer (cont.)

- The null distribution of $\overline{X}$ is normal with mean 1000 and standard deviation of $2/\sqrt{60} = 0.258.$ The *z*-score of the observed $\overline{X} = 1000.6$ is

$$z = \frac{1000.6 - 1000}{0.258} = 2.32$$

- Since H_0 specifies $\mu = 1000$, regions in both tails of the curve are in greater disagreement with H_0 that the observed value of 1000.6. The *P*-value is the sum of the areas in both tails, which is 0.0204.

Answer (cont.)

- Therefore if H_0 is true, the probability of a result as extreme as or more extreme than that observed is only 0.0204.

- The evidence against H_0 is pretty strong. It would be prudent to reject H_0 and to recalibrate the scale.

One and Two-Tailed Tests

- When H_0 specifies a single value for μ, both tails contribute to the *P*-value, and the test is said to be a **two-sided** or **two-tailed** test.

- When H_0 specifies only that μ is greater than or equal to, or less than or equal to a value, only one tail contributes to the *P*-value, and the test is called a **one-sided** or **one-tailed** test.

Summary

- Let $X_1, \ldots, X_n$ be a *large* (e.g., $n > 30$) sample from a population with mean μ and standard deviation σ. To test a null hypothesis of the form

  $H_0: \mu \leq \mu_0$, $H_0: \mu \geq \mu_0$, or $H_0: \mu = \mu_0$.

- Compute the *z*-score:

$$z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

If σ is unknown it may be approximated by *s*.

Summary (cont.)

Compute the *P*-value. The *P*-value is an area under the normal curve, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is $H_1: \mu > \mu_0$, then the *P*-value is the area to the right of *z*.
- If the alternative hypothesis is $H_1: \mu < \mu_0$, then the *P*-value is the area to the left of *z*.
- If the alternative hypothesis is $H_1: \mu \neq \mu_0$, then the *P*-value is the sum of the areas in the tails cut off by *z* and $-z$.

Section 6.2: Drawing Conclusions from the Results of Hypothesis Tests

- There are two conclusions that we draw when we are finished with a hypothesis test,
  - We reject H_0. In other words, we conclude that H_0 is false.
  - We do not reject H_0. In other words, H_0 is plausible.
- One can never conclude that H_0 is true. We can just conclude that H_0 might be plausible.
- We need to know what level of disagreement, measured with the P-value, is great enough to render the null hypothesis implausible.

More on the *P*-value

- The smaller the *P*-value, the more certain we can be that H_0 is false.
- The larger the *P*-value, the more plausible H_0 becomes, but we can never be certain that H_0 is true.
- A rule of thumb suggests to reject H_0 whenever $P \leq 0.05$. While this rule is convenient, it has no scientific basis.

Statistical Significance

- Whenever the *P*-value is less than a particular threshold, the result is said to be "statistically significant" at that level.
- So, if $P \leq 0.05$, the result is statistically significant at the 5% level. And if $P \leq 0.01$, the result is statistically significant at the 1% level.
- If the test is statistically significant at the $100\alpha\%$ level, we can also say that the null hypothesis is "rejected at level $100\alpha\%$."

Example 2

A hypothesis test is performed of the null hypothesis H_0: $\mu = 0$. The *P*-value turns out to be 0.03.

Is the result statistically significant at the 10% level? The 5% level? The 1% level?

Is the null hypothesis rejected at the 10% level? The 5% level? The 1% level?

Comments

- Sometimes people report only that a test significant at a certain level, without giving the *P*-value. It is common, for example, to read that a result is "statistically significant at the 5% level" or "statistically significant ($P < 0.05$)."

- This is a poor practice, for three reasons.

Comments

- First, it provides no way to tell whether the *P*-value was just barely less than 0.05, or whether it was a lot less.
- Second, reporting that a result was statistically significant at the 5% level implies that there is a big difference between a *P*-value just under 0.05 and one just above 0.05, when in fact there is little difference.
- Third, a report like this does not allow readers to decide for themselves whether the *P*-value is small enough to reject the null hypothesis.

Comments

- Reporting the *P*-value gives more information about the strength of the evidence against the null hypothesis and allows each reader to decide for himself or herself whether to reject the null hypothesis.

- The bottom line: *P*-values should always be included whenever the results of a hypothesis test are reported.

Summary

Let α be any value between 0 and 1. Then, if $P \leq \alpha$,

- The result of the test is said to be statistically significant at the $100\alpha\%$ level.
- The null hypothesis is rejected at the $100\alpha\%$ level.
- When reporting the result of the hypothesis test, report the *P*-value, rather than just comparing it to 5% or 1%.

What the *P*-Value Is Not

- Since the *P*-value is a probability, and since small *P*-values indicate that H_0 is unlikely to be true, it is tempting to think that the *P*-value represents the probability that H_0 is true.
- This is not the case!
- The concept of probability discussed here is useful only when applied to outcomes that can turn out in different ways when experiments are repeated.
- This is a frequentist approach to probability, but there are other kinds of probability used in Statistics.

Example 3

Specifications for a water pipe call for a mean breaking strength μ of more than 2000 lb per linear foot. Engineers will perform a hypothesis test to decide whether or not to use a certain kind of pipe. They will select a random sample of 1 ft sections of pipe, measure their breaking strengths, and perform a hypothesis test. The pipe will not be used unless the engineers can conclude that $\mu > 2000$. Assume they test H_0: $\mu \leq 2000$ versus H_1: $\mu > 2000$. Will the engineers decide to use the pipe if H_0 is rejected? What if H_0 is not rejected?

Example 3 (cont.)

Suppose the hypotheses tested were H_0: $\mu \geq 2000$ versus H_1: $\mu < 2000$. Will the engineers decide to use the pipe if H_0 is rejected? What if H_0 is not rejected?

Significance

- When a result has a small *P*-value, we say that it is "statistically significant."
- In common usage, the word *significant* means "important."
- It is therefore tempting to think that statistically significant results must always be important. This is not the case.
- Sometimes statistically significant results do not have any scientific or practical importance.

Hypothesis Tests and CIs

- Both confidence intervals and hypothesis tests are concerned with determining plausible values for a quantity such as a population mean μ.

- In a hypothesis test for a population mean μ, we specify a particular value of μ (the null hypothesis) and determine if that value is plausible.

Hypothesis Tests and CIs

- In contrast, a confidence interval for a population mean μ can be thought of as a collection of all values for μ that meet a certain criterion of plausibility, specified by the confidence level $100(1 - \alpha)\%$.

- The relationship between confidence intervals and hypothesis tests is very close:
  - The values contained within a two-sided level $100(1 - \alpha)\%$ confidence intervals are precisely those values for which the *P*-value of a two-tailed hypothesis test will be greater than α.

Section 6.3: Tests for a Population Proportion

- A population proportion is simply a population mean for a population of 0s and 1s: a Bernoulli population.

- An example: A supplier of semiconductor wafers claims that of all the wafers he supplies, no more than 10% are defective. A sample of 400 wafers is tested, and 50 of them, or 12.5%, are defective. Can we conclude that the claim is false?

Section 6.3: Tests for a Population Proportion

- Our hypothesis test is similar to the one we saw before. But now we have a sample that consists of successes and failures, with "success" indicating a defective wafer.
- If the population proportion of defective wafers is denoted by p, the supplier's claim is that $p \leq 0.1$.
- Since our hypothesis concerns a population proportion, it is natural to base the test on the sample proportion.

Hypothesis Test

- Let X be the number of successes in n independent Bernoulli trials, each with success probability p; in other words, let $X \sim \text{Bin}(n, p)$.
- To test a null hypothesis of the form H_0: $p \leq p_0$, H_0: $p \geq p_0$, or H_0: $p = p_0$, assuming that both np_0 and $n(1 - p_0)$ are greater than 10:
- Compute the z-score:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

P-value for the Hypothesis Test

Compute the *P*-value. The *P*-value is an area under the normal curve, which depends on the alternate hypothesis as follows:

- If the alternative hypothesis is $H_1: p > p_0$, the *P*-value is the area to the right of *z*.
- If the alternative hypothesis is $H_1: p < p_0$, the *P*-value is the area to the left of *z*.
- If the alternative hypothesis is $H_1: p \neq p_0$, the *P*-value is the sum of the areas in the tails cut off by *z* and $-z$.

Example 4

An article presents a method for measuring orthometric heights above sea level. For a sample of 1225 baselines, 926 gave results that were within the class C spirit leveling tolerance limits. Can we conclude that this method produces results within the tolerance limits more than 75% of the time?

Section 6.4: Small Sample Test for a Population Mean

- When we had a large sample we used the sample standard deviation s to approximate the population deviation σ.
- When the sample size is small, s may not be close to σ, which invalidates this large-sample method.
- However, when the population is approximately normal, the Student's t distribution can be used.
- The only time that we don't use the Student's t distribution for this situation is when the population standard deviation σ is known. Then we are no longer approximating σ and we should use the z-test.

Hypothesis Test

- Let $X_1, \ldots, X_n$ be a sample from a normal population with mean μ and standard deviation σ, where σ is unknown.
- To test a null hypothesis of the form H_0: $\mu \leq \mu_0$, H_0: $\mu \geq \mu_0$, or H_0: $\mu = \mu_0$:
- Compute the test statistic

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}.$$

P-value

Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with $n - 1$ degrees of freedom, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is H_1: $\mu > \mu_0$, then the *P*-value is the area to the right of *t*.
- If the alternative hypothesis is H_1: $\mu < \mu_0$, then the *P*-value is the area to the left of *t*.
- If the alternative hypothesis is H_1: $\mu \neq \mu_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and $-t$.

Example 5

Before a substance can be deemed safe for landfilling, its chemical properties must be characterized. An article reports that in a sample of six replicates of sludge from a New Hampshire wastewater treatment plant, the mean pH was 6.68 with a standard deviation of 0.20. Can we conclude that the mean pH is less than 7.0?

Section 6.5: Large Sample Tests for the Difference Between Two Means

- Now, we are interested in determining whether or not the means of two populations are equal.
- The data will consist of two samples, one from each population.
- We will compute the difference of the sample means.
  - If the difference is far from 0, we will conclude that the population means are different.
  - If the difference is close to 0, we will conclude that the population means might be the same.

Hypothesis Test

- Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be *large* (e.g., $n_X > 30$ and $n_Y > 30$) samples from populations with mean μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Assume the samples are drawn independently of each other.

Hypothesis Test

- To test a null hypothesis of the form
  H_0: $\mu_X - \mu_Y \leq \Delta_0$, H_0: $\mu_X - \mu_Y \geq \Delta_0$, or
  H_0: $\mu_X - \mu_Y = \Delta_0$:
- Compute the *z*-score:

$$z = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}}.$$

  - If σ_X and σ_Y are unknown they may be approximated by s_X and s_Y, respectively.

P-value

Compute the *P*-value. The *P*-value is an area under the normal curve, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is H_1: $\mu_X - \mu_Y > \Delta_0$, then the *P*-value is the area to the right of z.
- If the alternative hypothesis is H_1: $\mu_X - \mu_Y < \Delta_0$, then the *P*-value is the area to the left of z.
- If the alternative hypothesis is H_1: $\mu_X - \mu_Y \neq \Delta_0$, then the *P*-value is the sum of the areas in the tails cut off by z and $-z$.

Example 6

An article compares properties of welds made using carbon dioxide as a shielding gas with those of welds made using a mixture of argon and carbon dioxide. One property studied was the diameter of inclusions, which are particles embedded in the weld. A sample of 544 inclusions in welds made using argon shielding averaged 0.37 μm in diameter, with a standard deviation of 0.25 μm. A sample of 581 inclusions in welds made using carbon dioxide shielding averaged 0.40 μm in diameter, with a standard deviation of 0.26 μm. Can you conclude that the mean diameters of inclusions differ between the two shielding gases?

Section 6.6: Tests for the Difference Between Two Proportions

- The procedure for testing the difference between two populations is similar to the procedure for testing the difference between two means.
- We have random variables X and Y each with binomial distributions: $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$.
- One set of hypotheses we might consider is $H_0: p_X - p_Y \geq 0$ versus $H_1: p_X - p_Y < 0$.

Comments

- The test is based on the statistic $\hat{p}_X - \hat{p}_Y$.
- We must determine the null distribution of this statistic.
- By the Central Limit Theorem, since n_X and n_Y are both large, we know that the sample proportions for X and Y have an approximately normal distribution.

More on Proportions

- Thus, the difference between the two is also normally distributed:

$$\hat{p}_X - \hat{p}_Y \sim N\left(p_X - p_Y,\ \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}\right)$$

- To obtain the null distribution, we must substitute values for the mean and variance.

More on Proportions

- The null hypothesis specifies that the mean is 0.
- Finding the variance is a bit trickier.
  - The null hypothesis specifies that the proportions are equal, so we should not just substitute the sample proportions in for their respective population proportions.
  - Instead, we examine the **pooled proportion** and use that in our variance.

More on Proportions

- The pooled proportion is obtained by dividing the total number of successes in both samples by the total sample size:

$$\hat{p} = \frac{X + Y}{n_X + n_Y}$$

- Our null distribution for the difference is thus

$$\hat{p}_X - \hat{p}_Y \sim N\left(0, \hat{p}(1-\hat{p})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$$

Hypothesis Test

- Let $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$. Assume there are at least 10 successes and 10 failures in each sample, and that X and Y are independent.
- To test a null hypothesis of the form $H_0: p_X - p_Y \leq 0$, $H_0: p_X - p_Y \geq 0$, and $H_0: p_X - p_Y = 0$.
- Compute

$$\hat{p}_X = \frac{X}{n_X}, \hat{p}_Y = \frac{Y}{n_Y}, \text{and } \hat{p} = \frac{X + Y}{n_X + n_Y}.$$

- Compute the z-score:

$$z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})(1/n_X + 1/n_Y)}}$$

P-value

Compute the *P*-value. The *P*-value is an area under the normal curve, which depends on the alternative hypothesis as follows:

- If the alternative hypothesis is $H_1: p_X - p_Y > 0$, then the *P*-value is the area to the right of *z*.
- If the alternative hypothesis is $H_1: p_X - p_Y < 0$, then the *P*-value is the area to the left of *z*.
- If the alternative hypothesis is $H_1: p_X - p_Y \neq 0$, then the *P*-value is the sum of the areas in the tails cut off by *z* and –*z*.

Example 7

Industrial firms often employ methods of "risk transfer," such as insurance or indemnity clauses in contracts, as a technique of risk management. An article reports the results of a survey in which managers were asked which methods played a major role in the risk management strategy of their firms. In a sample of 43 oil companies, 22 indicated that risk transfer played a major role, while in a sample of 93 construction companies, 55 reported that risk transfer played a major role. Can we conclude that the proportion of oil companies that employ the method of risk transfer is less than the proportion of construction companies that do?

Section 6.7: Small-Sample Tests for the Difference Between Two Means

- The *t* test can be used in some cases where samples are small, and thus the Central Limit Theorem does not apply.
- If both populations are approximately normal, the Student's *t* distribution can be used to construct a hypothesis test.

Hypothesis Test for the Difference in Means Assuming Unequal Variance

- Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be samples from *normal* populations with mean μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Assume the samples are drawn independently of each other.
- Assume that σ_X and σ_Y are not known to be equal.

Hypothesis Test for the Difference in Means Assuming Unequal Variance

- To test a null hypothesis of the form H_0: $\mu_X - \mu_Y \leq \Delta_0$, H_0: $\mu_X - \mu_Y \geq \Delta_0$, or H_0: $\mu_X - \mu_Y = \Delta_0$:
- Compute

$$\nu = \frac{\left[(s_X^2/n_X) + (s_Y^2/n_Y)\right]^2}{\left[(s_X^2/n_X)^2/(n_X - 1)\right] + \left[(s_Y^2/n_Y)^2/(n_Y - 1)\right]}$$

  rounded to the nearest integer.

- Compute the test statistic

$$z = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}}.$$

P-value

Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with *v* degrees of freedom, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is H_1: $\mu_X - \mu_Y > \Delta_0$, then the *P*-value is the area to the right of *t*.
- If the alternative hypothesis is H_1: $\mu_X - \mu_Y < \Delta_0$, then the *P*-value is the area to the left of *t*.
- If the alternative hypothesis is H_1: $\mu_X - \mu_Y \neq \Delta_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and $-t$.

Example 8

Good website design can make Web navigation easier. An article presents a comparison of item recognition between two designs. A sample of 10 users using a conventional Web design averaged 32.3 items identified, with a standard deviation of 8.56. A sample of 10 users using a new structured Web design averaged 44.1 items identified, with a standard deviation of 10.09. Can we conclude that the mean number of items identified is greater with the new structured design?

Hypothesis Test for the Difference in Means Assuming Equal Variance

- Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be samples from *normal* populations with mean μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Assume the samples are drawn independently of each other.
- Assume that σ_X and σ_Y are known to be equal.

Hypothesis Test for the Difference in Means Assuming Equal Variance

- To test a null hypothesis of the form H_0: $\mu_X - \mu_Y \leq \Delta_0$, H_0: $\mu_X - \mu_Y \geq \Delta_0$, or H_0: $\mu_X - \mu_Y = \Delta_0$.
- Compute

$$s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}$$

- Compute the test statistic

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{s_p \sqrt{1/n_X + 1/n_Y}}.$$

P-value

Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with *v* degrees of freedom, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is H_1: $\mu_X - \mu_Y > \Delta_0$, then the *P*-value is the area to the right of *t*.
- If the alternative hypothesis is H_1: $\mu_X - \mu_Y < \Delta_0$, then the *P*-value is the area to the left of *t*.
- If the alternative hypothesis is H_1: $\mu_X - \mu_Y \neq \Delta_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and $-t$.

Example 9

Two methods have been developed to determine the nickel content of steel. In a sample of five replications of the first method, X, on a certain kind of steel, the average measurement (in percent) was 3.16 with a standard deviation of 0.042. The average of seven replications of the second method, Y, was 3.24, and the standard deviation was 0.048. Assume that it is known that the population variances are nearly equal. Can we conclude that there is a difference in the mean measurements between the two methods?

Section 6.8: Tests with Paired Data

- We saw in Chapter 5 that sometimes it is better to design a two-sample experiment so that each item in one sample is paired with an item in the other.
- In this section, we present a method for testing hypotheses involving the difference between two population means on the basis of such paired data.

Hypothesis Test

- Let $(X_1, Y_1),\ldots, (X_n, Y_n)$ be sample of ordered pairs whose differences $D_1,\ldots, D_n$ are a sample from a *normal* population with mean μ_D.
- To test a null hypothesis of the form H_0: $\mu_D \leq \mu_0$, H_0: $\mu_D \geq \mu_0$, or H_0: $\mu_D = \mu_0$.
- Compute the test statistic

$$t = \frac{\overline{D} - \mu_0}{s_D / \sqrt{n}} .$$

P-value

Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with $n - 1$ degrees of freedom, which depends on the alternate hypothesis as follows.

- If the alternative hypothesis is H_1: $\mu_D > \mu_0$, then the *P*-value is the area to the right of *t*.
- If the alternative hypothesis is H_1: $\mu_D < \mu_0$, then the *P*-value is the area to the left of *t*.
- If the alternative hypothesis is H_1: $\mu_D \neq \mu_0$, then the *P*-value is the sum of the areas in the tails cut off by t and $-t$.

Note on the Test for Paired Data

- If the sample is large, the D_i need not be normally distributed, the test statistic is

$$z = \frac{\bar{D} - \mu_0}{s_D / \sqrt{n}}$$

and a z test should be performed.

Section 6.9: Distribution-Free Tests

- The Student's *t* distribution described in the previous sections require that samples come from normal populations.
- Distribution-free tests get their name from the fact that the samples are not required to come from any specific distribution.
- While distribution-free tests do require assumptions for their validity, these assumptions are somewhat less restrictive than the assumptions needed for the *t* test.
- Distribution-free tests are sometimes called **nonparametric tests**.

Wilcoxon Signed-Rank Test

- Under H_0, the population mean is $\mu = \mu_0$.
- To compute the rank-sum statistic, we begin by subtracting μ_0 from each sample observation to obtain differences.
- The difference closest to zero, ignoring the sign, is given the rank 1. The difference next closer to 0, ignoring the sign, is given the rank 2, and so on.
- Assign the ranks corresponding to negative differences, negative signs.

More on the Test

- Denote the sum of the positive ranks S_+. This value may be used as a test statistic.
- Larger values of S_+ will provide evidence against a null hypothesis of the form H_0: $\mu \leq \mu_0$, while small values of S_+ will provide evidence against a null hypothesis of the form H_0: $\mu \geq \mu_0$.
- Table A.5 presents certain probabilities for the null distribution of S_+.

Ties

- Sometimes two or more of the quantities to be ranked have exactly the same value. Such quantities are said to be tied.
- The standard method for dealing with ties is to assign each tied observation the average of the ranks they would have received if they had differed slightly.

Differences of Zero

- If the mean under H_0 is μ_0 and one of the observations is equal to μ_0, then its difference is 0, which is neither positive or negative.
- The appropriate procedure is to drop such observations from the sample altogether, and to consider the sample size to be reduced by the number of these observations.

Example 10

The nickel content for six welds was measured to be 9.3, 0.9, 9.0, 21.7, 11.5, and 13.9. Use these data to test H_0: $\mu \leq 5$ versus H_1: $\mu > 5$.

Large-Sample Approximation

- When the sample size n is large, the test statistic S_+ is approximately normally distributed.
- A rule of thumb is that the normal approximation is good if $n > 20$.
- It can be shown that that under H_0, S_+ has mean $n(n + 1)/4$ and variance $n(n + 1)(2n + 1)/24$.
- The Wilcoxon signed-rank test is performed by computing the z-score of S_+, and then using the normal table to find the P-value.

The Wilcoxon Rank-Sum Test

- The Wilcoxon rank-sum test, also called the Mann–Whitney test, can be used to test the difference in population means in certain cases where the populations are not normal.
- Two assumptions are necessary:
  - The populations must be continuous.
  - The probability density functions must be identical in shape and size; the only possible difference between them being their location.

The Wilcoxon Rank-Sum Test

- Let $X_1,\ldots, X_m$ be a random sample from one population and let $Y_1,\ldots, Y_n$ be a random sample from the other.
- We adopt the notational convention that when the sample sizes are unequal, the smaller sample will be denoted by $X_1,\ldots, X_m$.
- Thus the sample sizes are m and n, with $m \leq n$.

More on Wilcoxon Rank-Sum Test

- Denote the population means by μ_X and μ_Y, respectively.
- The test is performed by ordering the $m + n$ values by combining the two samples, and assigning ranks $1, 2, \ldots, m + n$ to them.
- The test statistic, denoted by W, is the sum of the ranks corresponding to $X_1, \ldots, X_m$.

More on Wilcoxon Rank-Sum Test

- Since the populations are identical with the possible exception of location, it follows that if $\mu_X < \mu_Y$, the values in the *X* sample will tend to be smaller than the ones in the *Y* sample, so the rank sum *W* will tend to be smaller as well.
- By similar reasoning, if $\mu_X > \mu_Y$, *W* will tend to be larger.
- Table A.6 presents *P*-values for the *W* statistic.

Example 11

Resistances, in mΩ, are measured for five wires of one type and six wires of another type. The results are as follows:

X:	36	28	29	20	38	
Y:	34	41	35	47	49	46

Use the Wilcoxon rank-sum test to test H_0: $\mu_X \geq \mu_Y$ versus H_1: $\mu_X < \mu_Y$.

Large-Sample Approximation

- When both sample sizes m and n are greater than 8, it can be shown that the null distribution of the test statistic W is approximately normal with mean $m(m + n + 1)/2$ and variance $mn(m + n + 1)/12$.
- In these cases the test is performed by computing the z-score of W, and then using the normal table to find the P-value.

Distribution Free, Not Assumption Free

- The distribution-free methods require certain assumptions for their validity.
- The necessary assumptions of symmetry for the signed-rank test and of identical shapes and spreads for the rank-sum test are actually rather restrictive.
- While these tests perform reasonably well under moderate violations of these assumptions, they are not universally applicable.

Section 6.10: The Chi-Square Test

- A generalization of the Bernoulli trial is the **multinomial trial,** which is an experiment that can result in any one of k outcomes, where

  $k \geq 2$. (See Section 4.4.)
- Example: Suppose we roll a die 600 times.
- The results obtained are called the **observed values**.
- To test the null hypothesis that $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$, we calculate the **expected values** for the given outcome.

The Test

- The idea behind the hypothesis test is that if H_0 is true, then the observed and expected values are likely to be close to each other.
- Therefore we will construct a test statistic that measures the closeness of the observed to the expected values.

The Test

- The statistic is called the **chi-square statistic**.
- Let k be the number of possible outcomes and let O_i and E_i be the observed and expected number of trials that result in outcome i.
- The chi-square statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Decision for Test

- The larger the value of χ^2, the stronger the evidence against H_0.
- To determine the *P*-value for the test, we must know the null distribution of this test statistic.
- When the expected values are all sufficiently large, a good approximation is available. It is called the **chi-square distribution** with $k - 1$ degrees of freedom.

Decision for Test

- Use of the chi-square distribution is appropriate whenever all the expected values are greater than or equal to 5.
- A table for the chi-square distribution is provided in Appendix A, Table A.7.

Chi-Square Tests

- Test for homogeneity:
  - Sometimes several multinomial trials are conducted, each with the same set of possible outcomes.
  - The null hypothesis is that the probabilities of the outcomes are the same for each experiment.
- Test for independence:
  - Here we are testing that both row and column totals are random. That is, the two variables are independent of one another.

Section 6.11: Tests for Variances of Normal Populations

- Sometimes it is desirable to test a null hypothesis that a population variance has a certain value.
- Sometimes it is desirable to test a null hypothesis that two populations have equal variances.
- In general, there is no good way to test these hypotheses.
- In the special case where both populations are normal, methods are available.

Testing the Variance of a Normal Population

- Let $X_1, \ldots, X_n$ be a simple random sample from a $N(\mu_1, \sigma_1^2)$ population. Let s^2 be the sample variance.
- The test statistic is $\dfrac{(n-1)s^2}{\sigma_0^2}$

- The test statistic has a chi-square distribution with $n-1$ degrees of freedom.

$$H_0 : \sigma^2 \leq \sigma_0^2 \qquad H_0 : \sigma^2 \geq \sigma_0^2 \qquad H_0 : \sigma^2 = \sigma_0^2$$

Example 12

- To check the reliability of a scale in a butcher shop, a test weight known to weigh 400 grams was weighed 16 times. For the scale to be considered reliable, the variance of repeated measurements must be less than 1. The sample variance of the 16 measured weights was $s^2 = 0.81$. Assume that the measured weights are independent and follow a normal distribution. Can we conclude that the population variance of the measurements is less than 1?

Test For Equality of Variances: Set-Up

- Let $X_1,\ldots, X_m$ be a simple random sample from a $N(\mu_1, \sigma_1^2)$ population, and let $Y_1,\ldots, Y_n$ be a simple random sample from a $N(\mu_2, \sigma_2^2)$ population.
- Assume that the samples are chosen independently.
- The values of the means are irrelevant, we are only concerned with the variances.

Hypotheses

- Let s_1^2 and s_2^2 be the sample variances.
- Any of three null hypothesis may be tested. They are

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq 1, \text{ or equivalently, } \sigma_1^2 \leq \sigma_2^2$$

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} \geq 1, \text{ or equivalently, } \sigma_1^2 \geq \sigma_2^2$$

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1, \text{ or equivalently, } \sigma_1^2 = \sigma_2^2$$

Test For Equality of Variances

- The test statistic is the ratio of the two sample variances:

$$F = \frac{s_1^2}{s_2^2}$$

- When H_0 is true, we assume that $\sigma_1^2/\sigma_2^2 = 1$, or equivalently $\sigma_1^2 = \sigma_2^2$. When s_1^2 and s_2^2 are, on average, the same size, F is likely to be near 1.
- When H_0 is false, we assume that $\sigma_1^2 > \sigma_2^2$. When s_1^2 is likely to be larger than s_2^2, and F is likely to be greater than 1.

F Distribution

- Statistics that have an *F* distribution are ratios of quantities, such as the ratio of two variances.
- The *F* distribution has two values for the degrees of freedom: one associated with the numerator, and one associated with the denominator.
- The degrees of freedom are indicated with subscripts under the letter *F*.
- Note that the numerator degrees of freedom are always listed first.
- A table for the *F* distribution is provided (Table A.8 in Appendix A).

Back to the Test

- The null distribution of the F statistic is $F_{m-1,n-1}$.

- The number of degrees of freedom for the numerator is one less than the sample size used to compute s_1^2, and the number of degrees of freedom for the denominator is one less than the sample size used to compute s_2^2.

Back to the Test

- Note that the *F* test is sensitive to the assumption that the samples come from normal populations.
- If the shapes of the populations differ much from the normal curve, the *F* test may give misleading results.
- The *F* test does not prove that two variances are equal. The basic reason for the failure to reject the null hypothesis does not justify the assumption that the null hypothesis is true.

Example 13

In a series of experiments to determine the absorption rate of certain pesticides into skin, measured amounts of two pesticides were applied to several skin specimens. After a time, the amounts absorbed (in μg) were measured. For pesticide A, the variance of the amounts absorbed in 6 specimens was 2.3, while for pesticide B, the variance of the amounts absorbed in 10 specimens was 0.6. Assume that for each pesticide, the amounts absorbed are a simple random sample from a normal population. Can we conclude that the variance in the amount absorbed is greater for pesticide A than for pesticide B?

Section 6.12: Fixed-Level Testing

- A hypothesis test measures the plausibility of the null hypothesis by producing a *P*-value.
- The smaller the *P*-value, the less plausible the null.
- We have pointed out that there is no scientifically valid dividing line between plausibility and implausibility, so it is impossible to specify a "correct" *P*-value below which we should reject H_0.

Fixed-Level Testing

- If a decision is going to be made on the basis of a hypothesis test, there is no choice but to pick a cut-off point for the *P*-value.
- When this is done, the test is referred to as a **fixed-level** test.

Conducting the Test

To conduct a fixed-level test:

- Choose a number α, where $0 < \alpha < 1$. This is called the significance level, or the level, of the test.
- Compute the *P*-value in the usual way.
- If $P \leq \alpha$, reject H_0. If $P > \alpha$, do not reject H_0.

Comments

- In a fixed-level test, a **critical point** is a value of the test statistic that produces a *P*-value exactly equal to α.
- A critical point is a dividing line for the test statistic just as the significance level is a dividing line for the *P*-value.
- If the test statistic is on one side of the critical point, the *P*-value will be less than α, and H_0 will be rejected.

Comments

- If the test statistic is on the other side of the critical point, the *P*-value will be more than α, and H_0 will not be rejected.
- The region on the side of the critical point that leads to rejection is called the **rejection region**.
- The critical point itself is also in the rejection region.

Example 14

A new concrete mix is being evaluated. The plan is to sample 100 concrete blocks made with the new mix, compute the sample mean compressive strength (X), and then test H_0: $\mu \leq 1350$ versus H_0: $\mu > 1350$, where the units are MPa. It is assumed that previous tests of this sort that the population standard deviation σ will be close to 70 MPa. Find the critical point and the rejection region if the test will be conducted at a significance level of 5%.

Type I and Type II Errors

When conducting a fixed-level test at significance level α, there are two types of errors that can be made. These are

- Type I error: Reject H_0 when it is true.
- Type II error: Fail to reject H_0 when it is false.

The probability of Type I error is never greater than α.

Section 6.13: Power

- A hypothesis test results in Type I error if H_0 is not rejected when it is false.
- The **power** of the test is the probability of *rejecting* H_0 when it is false. Therefore,

  Power = 1 – P(Type II error).
- To be useful, a test must have reasonable small probabilities of both type I and type II errors.

Power

- The type I error is kept small by choosing a small value of α as the significance level.
- If the power is large, then the probability of type II error is small as well, and the test is a useful one.
- The purpose of a power calculation is to determine whether or not a hypothesis test, when performed, is likely to reject H_0 in the event that H_0 is false.

Computing the Power

Computing the power involves two steps:

1. Compute the rejection region.
2. Compute the probability that the test statistic falls in the rejection region if the alternate hypothesis is true. This is the power.

When power is not large enough, it can be increased by increasing the sample size.

Example 15

Find the power of the 5% level test of H_0: $\mu \leq 80$ versus H_1: $\mu > 80$ for the mean yield of the new process under the alternative $\mu = 82$, assuming $n = 50$ and $\sigma = 5$.

Section 6.14: Multiple Tests

- Sometimes a situation occurs in which it is necessary to perform many hypothesis tests.
- The basic rule governing this situation is that as more tests are performed, the confidence that we can place in our results decreases.

The Bonferroni Method

- The Bonferroni method provides a way to adjust *P*-values upward when several hypothesis tests are performed.
- If a *P*-value remains small after the adjustment, the null hypothesis may be rejected.
- To make the Bonferroni adjustment, simply multiply the *P*-value by the number of test performed.

Example 16

Four different coating formulations are tested to see if they reduce the wear on cam gears to a value below 100 μm. The null hypothesis H_0: $\mu \geq 100$ μm is tested for each formulation and the results are

Formulation A: $P = 0.37$

Formulation B: $P = 0.41$

Formulation C: $P = 0.005$

Formulation D: $P = 0.21$

Example 16 (cont.)

The operator suspects that formulation C may be effective, but he knows that the *P*-value of 0.005 is unreliable, because several tests have been performed. Use the Bonferroni adjustment to produce a reliable *P*-value.

Section 6.15: Using Simulation to Perform Hypothesis Tests

- Simulation can be used to check normality assumptions that are needed for hypothesis tests.
- We can also use simulation to test the difference between two population means.
- Estimating the power of a test can be done with simulation as well.

Summary

We learned about:

- Large sample tests for a population mean
- Drawing conclusions from the results of hypothesis tests
- Tests for a population proportion and differences in two proportions
- Small sample tests for a population mean
- Large and small sample tests for the difference between two means
- Tests with paired data
- Distribution-free test
- Chi-Square test
- *F* test