Introduction to Statistics
Part 2

List of Topics

- Review – Session 1
- Introduction to Estimation
- Introduction to Hypothesis Testing
- Inference about a Population
- Inference about comparing two populations
- Summary and Conclusion

Introduction to Statistics – Part 2

Introduction to Estimation

Binomial, Poisson, and Normal distributions are used in making probability statements about random variable X. The following population parameters are required:

- Binomial: p
- Poisson: µ
- Normal: µ and σ

However, these parameters are unknown in most scenarios. Therefore, use the sampling distribution to draw inferences about the unknown population parameters.

Statistical Inference

- Process by which we acquire information and draw conclusions about populations from samples
- The objective is to determine the approximate value of a population parameter on the basis of a sample statistic. For example, the sample mean ($\bar{x}$) is employed to estimate the population mean (μ).

Knowledge required for Statistical Inference

Descriptive statistics
Probability distributions
Sampling distributions

Types of Estimators

Point Estimator

- The sample statistics such as x bar or s, that provides the point estimate of the population parameter

Point Estimate

- The value of a point estimator used in a particular instance as an estimate of a population parameter

Interval Estimate or Confidence Interval

- An estimate of a population parameter that provides an interval believed to contain the value of the parameter

Confidence Level

- The confidence associated with an interval estimate.
- For example, if an interval estimation procedure provides intervals such that 95% of the intervals formed using the procedure will include the population parameter, the interval estimate is said to be constructed at the 95% confidence level.

Estimating a Population Proportion

- Estimate population proportion using the sample proportion.
- The sampling distribution of $\hat{p}$ is approximately normal with mean p and standard deviation $\sqrt{p(1-p)/n}$
- $Z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$ is (approximately) standard normally distributed.
- Lower confidence limit = $\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
- Upper confidence limit = $\hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

Prescription Drug Example

Goal

The parameter is p and its estimator is

Lower confidence limit

Upper confidence limit

Interpretation

Student t Distribution

- The Student ***t*** distribution is “mound” shaped and symmetrical about its mean of zero
- μ and σ define the normal distribution, ν, the degrees of freedom, defines the Student ***t*** Distribution
- As the number of degrees of freedom increases, the t distribution approaches the standard normal distribution

Determining Student *t* Values

Excel provides values of a Student ***t*** random variable with ν degrees of freedom such that:

$$P(t > t_{A,\nu}) = A$$

For example, if we want the value of t with 10 degrees of freedom such that the area under the Student t curve is .05: T.INV(0.95,10) = 1.812

Hypothesis Testing

Statistical Inference

- Hypothesis testing is one form of statistical inference

Nonstatistical Hypothesis Testing

Example

- In a trial a jury must decide between two hypotheses:
- Null Hypothesis
  - H_0: The defendant is innocent
- Alternative Hypothesis
  - H_1: The defendant is guilty

The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.

Rejecting the null hypothesis in favor of the alternative hypothesis

Convicting the defendant

- Enough evidence to conclude that the defendant is guilty (i.e., enough evidence to support the alternative hypothesis)
- If the jury acquits, it is stating that *there is not enough evidence to support the alternative hypothesis*.

  (Notice that the jury is not saying that the defendant is innocent, only that there is not enough evidence to support the alternative hypothesis. That is why we never say that we accept the null hypothesis)

Possible Errors

Type I error (α error)

- When we reject a **true** null hypothesis i.e. Type I error occurs when the jury convicts an innocent person

Type II error (β error)

- When we don’t reject a false null hypothesis. That occurs when a guilty defendant is acquitted

The probabilities of Type I and Type II error are inversely related. Decreasing one increases the other.

In our judicial system Type I errors are regarded as _____________. We try to avoid convicting innocent people. We are more willing to acquit guilty people. We arrange to make α small by requiring the prosecution to prove its case and instructing the jury to find the defendant guilty only if there is “evidence beyond a reasonable doubt.”

Summary

1. Two hypotheses: the null and the alternative hypotheses
2. The procedure begins with the assumption that the null hypothesis is true
3. Goal: Determine if there is enough evidence to infer that the alternative hypothesis is true
4. Two possible decisions:

   Conclude that there is enough evidence to support the alternative hypothesis

   Conclude that there is ***not*** enough evidence to support the alternative hypothesis
5. Two possible errors:

   Type I error: Reject a true null hypothesis

   Type II error: Do not reject a false null hypothesis

   P(Type I error) = α

   P(Type II error) = β

Example: Election Day Exit Poll

Goal:

Null Hypothesis Ho:

Alternative Hypothesis H_1:

Test Statistic:

p-Value of a Test:

Rejection Region:

Interpretation:

p value	Alternative Hypothesis Support
< 1%	Overwhelming evidence (Highly Significant)
between 1% and 5%	Strong evidence (Significant)
between 5% and 10%	Weak evidence (Not Significant)
>10%	No evidence (Not Significant)

Two- Tail Test

A two-tail test is conducted when the alternative hypothesis specifies that the parameter is *not equal* to the value indicated in the null hypothesis

H_0: p = ____
H_1: $p \neq$ ____

The rejection region is $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$. We need to find the left-tail and right-tail critical values.

Left-tail critical value:

Right-tail critical value:

Interpretation:

Inference about a Population

How to estimate and test two population parameters: µ and σ^2 and review inference about p.

Inference about a Population Mean

- The best estimator of an unknown population mean µ is the sample mean
- The sampling distribution of $\bar{x}$ is normal (if the population is normal) or approximately normal (if the population is nonnormal and the sample size is large)
- The standard error (the standard deviation of the sample mean) is$\sigma / \sqrt{n}$
- The test statistic is $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

Inference about a Population Mean (unknown population variance)

- Estimate population standard deviation value using the sample standard deviation s
- Substitute s for σ the sampling distribution becomes Student t distribution
- The test statistic is $t = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$ Student t distributed with υ = n – 1 degrees of freedom
- The confidence interval estimator is $\bar{x} \pm t_{\alpha/2} \dfrac{s}{\sqrt{n}}$

Example - Newspaper

Objective:

Parameter to be tested:

Hypothesis:

H_1:

H_0:

Test Statistic:

Rejection region:

Conclusion:

Example - IRS

Objective:

Parameter to be estimated:

Confidence Interval Estimator:

Mean additional tax collected lies between $_____ and $__________

Check Required Conditions

- Normality Assumption

  Histogram (Newspaper)

Inference about Population Variance

- The parameter to investigate is the population variance: σ^2
- Sample variance (**s^2**) is the unbiased point estimator for σ^2
- Statistic $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ has a chi-squared distribution, with n–1 degrees of freedom
- ***Confidence interval estimator for*** σ^2

$$LCL = \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

$$UCL = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Example: Container Filling Machines

Objective

Null Hypothesis

Alternative Hypothesis

Test Statistic

Excel Formulas

```
=CHISQ.DIST( )
```

```
=CHISQ.INV( )
```

Conclusion

Example

Estimate with 99% confidence the variance of fills in the previous example

Conclusion

Inference about comparing two populations

Comparing Two Populations

Previously we looked at techniques to estimate and test parameters for **one population**:

- Population Mean μ
- Population Variance σ^2
- Population Proportion p

We will still consider these parameters when we are looking at ***two populations***, however our interest will now be the ***difference*** between two means.

Difference between Two Means

- Draw random samples from each of two populations with the assumption that the samples drawn are completely unrelated to one another

Difference between Two Means $\bar{x}_1 - \bar{x}_2$

- Use the statistic $\bar{x}_1 - \bar{x}_2$ to compare two population means μ_1- μ_2.

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- $\bar{x}_1 - \bar{x}_2$ is normally distributed if the original populations are normal or approximately normal if the populations are nonnormal and the sample sizes are large (n_1, $n_2 > 30$)

- The expected value of $\bar{x}_1 - \bar{x}_2$ is μ_1- μ_2

- The variance of $\bar{x}_1 - \bar{x}_2$ is $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ and the standard error is: $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Making Inferences About μ_1-μ_2

- Since $\bar{x}_1 - \bar{x}_2$ is ***normally distributed*** if the original populations are normal or ***approximately normal*** if the populations are nonnormal and the sample sizes are large, then:

  $$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  is a standard normal (or approximately normal) random variable

- Use this statistics to build the test statistic and the confidence interval estimator for μ_1 - μ_2.

- In practice, the z statistic is rarely used since the population variances are unknown, instead use a t-statistic

Test Statistic for μ_1-μ_2 (equal variances)

Calculate t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \; \nu = n_1 + n_2 - 2$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

s_p^2 – the *pooled variance estimator*

Which test statistic do we use? Equal variance or unequal variance?

Testing the Population Variances

H_0: $\sigma_1^2 / \sigma_2^2 = 1$

H_1: $\sigma_1^2 / \sigma_2^2 \neq 1$

Test statistic: s_1^2 / s_2^2 (F-distributed with degrees of freedom $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 2$)

This is a two-tail test. Find the rejection region with the help of Excel

Required condition

- Same as that for the t-test of μ_1 - μ_2**,** which is both populations are normally distributed

Example – Mutual Funds

Problem objective

Hypothesis Testing

H_1: μ_1- $\mu_2 > 0$

and

H_0: μ_1- $\mu_2 = 0$

Statistics s_1^2 = ___ and s_2^2 = ___

Conclusion: There is ___________ enough evidence to infer that the population variances differ.

As a result we conclude that there is _________evidence to infer that on average directly-purchased mutual funds outperform broker-purchased mutual funds

Example: Family-run Business

Problem Objective

Population 1:
Population 2:

Data Type:

Parameter to be tested:

H_1:

H_0:

Determine whether to use the equal-variances or unequal-variances t –test of μ_1- $\mu_{2.}$

s_1^2 = _____ and s_2^2 = _____

Test statistic: F = ________

Rejection region:

Click Data, Data Analysis, and F-Test Two Sample for Variances

The p-value of the test we're conducting is 2 x __________ = ______

Thus, the correct technique is the unequal-variances t-test of μ_1- $\mu_{2.}$

Click Data, Data Analysis, t-Test: **Two-Sample Assuming Unequal Variances**

The t-statistic is _______ and its p-value is _______. Accordingly, we conclude that there is ___________ evidence to infer that the mean times differ.

Checking the Required Condition

Both the equal-variances and unequal-variances techniques require that the populations be normally distributed.

Summary and Conclusion

Reference: BSTAT (1st edition) by Gerald Keller. South-Western College Pub