Chapter 12

Inference About A Population (Variance unknown)

Inference With Variance Unknown…

Previously, we looked at estimating and testing the population mean when the population standard deviation σ was known or given:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

But how often do we know the actual population variance? Generally, never.

Instead, we use the ***Student t-statistic***, given by:

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

Testing μ when σ is unknown…

When the population standard deviation is unknown and the population is normal, the test statistic for testing hypotheses about μ is:

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

which is Student ***t*** distributed with ν = n–1 degrees of freedom. The confidence interval estimator of μ is given by:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Example 12.1

It is likely that in the near future nations will have to do more to save the environment.

Possible actions include reducing energy use and recycling.

Currently (2007) most products manufactured from recycled material are considerably more expensive than those manufactured from material found in the earth.

Example 12.1

Newspapers are an exception.

It can be profitable to recycle newspaper.

A major expense is the collection from homes. In recent years a number of companies have gone into the business of collecting used newspapers from households and recycling them.

A financial analyst for one such company has recently computed that the firm would make a profit if the mean weekly newspaper collection from each household exceeded 2.0 pounds.

Example 12.1

In a study to determine the feasibility of a recycling plant, a random sample of 148 households was drawn from a large community, and the weekly weight of newspapers discarded for recycling for each household was recorded. Xm12-01*

Do these data provide sufficient evidence to allow the analyst to conclude that a recycling plant would be profitable?

Example 12.1

IDENTIFY

Our objective is to ***describe*** the population of the amount of newspaper discarded per household, which is an interval variable. Thus the parameter to be tested is the population mean μ.

We want to know if there is enough evidence to conclude that the mean is greater than 2. Thus,

$$H_1: \mu > 2$$

Therefore we set our usual null hypothesis to:

$$H_0: \mu = 2$$

Example 12.1

IDENTIFY

The test statistic is:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad \nu = n - 1$$

Because the alternative hypothesis is:

$$H_1\text{: } \mu \quad > 2$$

the rejection region becomes:

$$t > t_{\alpha,\nu} = t_{.01,148} \approx t_{.01,150} = 2.351$$

Example 12.1

INTERPRET

The value of the test statistic is $t = 2.24$ and its p-value is .0134.

There is not enough evidence to infer that the mean weight of discarded newspapers is greater than 2.0.

Note that there is some evidence; the p-value is .0134. However, because we wanted the Type I error to be small we insisted on a 1% significance level.

Thus, we cannot conclude that the recycling plant would be profitable.

Mercury Contamination

A sample of fish was selected from 53 Florida lakes, and mercury concentration in the muscle tissue was measured (ppm). The mercury concentration values were

1.230	1.330	0.040	0.044	1.200	0.270
0.490	0.190	0.830	0.810	0.710	0.500
0.490	1.160	0.050	0.150	0.190	0.770
1.080	0.980	0.630	0.560	0.410	0.730
0.590	0.340	0.340	0.840	0.500	0.340
0.280	0.340	0.750	0.870	0.560	0.170
0.180	0.190	0.040	0.490	1.100	0.160
0.100	0.210	0.860	0.520	0.650	0.270
0.940	0.400	0.430	0.250	0.270	

Find an approximate 95% CI on μ.

The *t* distribution

Let X_1, X_2, …, X_n be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. The random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \qquad (8\text{-}6)$$

has a *t* distribution with $n - 1$ degrees of freedom.

Example:Alloy Adhesion

Construct a 95% CI on μ to the following data.

19.8	10.1	14.9	7.5	15.4	15.4
15.4	18.5	7.9	12.7	11.9	11.4
11.4	14.1	17.6	16.7	15.8	
19.5	8.8	13.6	11.9	11.4	

The sample mean is $\bar{x} = 13.71$ and sample standard deviation is $s = 3.55$.

Since $n = 22$, we have $n - 1 = 21$ degrees of freedom for t, so $t_{0.025,21} = 2.080$.

The resulting CI is

$$\bar{x} - t_{\alpha/2,n-1} s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2,n-1} s/\sqrt{n}$$

$$13.71 - 2.080(3.55)/\sqrt{22} \le \mu \le 13.71 + 2.080(3.55)/\sqrt{22}$$

$$13.71 - 1.57 \le \mu \le 13.71 + 1.57$$

$$12.14 \le \mu \le 15.28$$

Interpretation: The CI is fairly wide because there is a lot of variability in the measurements. A larger sample size would have led to a shorter interval.

Golf Club Design

An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. It is of interest to determine if there is evidence (with $\alpha = 0.05$) to support a claim that the mean coefficient of restitution exceeds 0.82.

The observations are:

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

The sample mean and sample standard deviation are $\bar{x} = 0.83725$ and $s = 0.02456$. The objective of the experimenter is to demonstrate that the mean coefficient of restitution exceeds 0.82, hence a one-sided alternative hypothesis is appropriate.

The seven-step procedure for hypothesis testing is as follows:

1. **Parameter of interest:** The parameter of interest is the mean coefficient of restitution, μ.

2. **Null hypothesis:** H_0: $\mu = 0.82$

3. **Alternative hypothesis:** H_1: $\mu > 0.82$

Golf Club Design - Continued

4. Test Statistic: The test statistic is

5. Reject H_0 if: Reject H_0 if the P-value is less than 0.05.

6. Computations: Since $\bar{x} = 0.83725$, $s = 0.02456$, $\mu = 0.82$, and $n = 15$, we have

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad t_0 = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72$$

7. Conclusions: From Appendix A Table II, for a t distribution with 14 degrees of freedom, $t_0 = 2.72$ falls between two values: 2.624, for which $\alpha = 0.01$, and 2.977, for which $\alpha = 0.005$. Since, this is a one-tailed test the P-value is between those two values, that is, $0.005 < P < 0.01$. Therefore, since $P < 0.05$, we reject H_0 and conclude that the mean coefficient of restitution exceeds 0.82.

Interpretation: There is strong evidence to conclude that the mean coefficient of restitution exceeds 0.82.

Example 12.2

In 2007 (the latest year reported) 134,543,000 tax returns were filed in the United States

The Internal Revenue Service (IRS) examined 1.03% or 1,385,000 of them to determine if they were correctly done.

To determine how well the auditors are performing, a random sample of these returns was drawn and the additional tax was reported. Xm12-02

Estimate with 95% confidence the mean additional income tax collected from the 1,385,000 files audited.

Example 12.2

IDENTIFY

The objective is to describe the population of additional tax collected.

The data are interval.

The parameter to be estimated is the mean additional tax.

The confidence interval estimator is

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Example 12.2…

INTERPRET

We estimate that the mean additional tax collected lies between $10,703 and $11,983 .

We can use this estimate to help decide whether the IRS is auditing the individuals who should be audited.

Inference: Population Proportion…

When data are nominal, we count the number of occurrences of each value and calculate proportions. Thus, the parameter of interest in describing a population of nominal data is the population proportion ***p***.

This parameter was based on the binomial experiment.

Recall the use of this statistic: $\hat{p} = \frac{x}{n}$

where p-hat ($\hat{p}$) is the sample proportion: **x** successes in a sample size of **n** items.

Inference: Population Proportion…

When ***np*** and ***n(1–p)*** are both greater than 5, the sampling distribution of $\hat{p}$ is approximately normal with

mean: $\mu = p$

standard deviation: $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$

Hence: $z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$

Inference: Population Proportion

Test statistic for ***p***:

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

The confidence interval estimator for ***p*** is given by:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

(both of which require that ***np***>5 and ***n(1–p)***>5)

Example 12.5

After the polls close on election day networks compete to be the first to predict which candidate will win.

The predictions are based on counts in certain precincts and on exit polls.

Exit polls are conducted by asking random samples of voters who have just exited from the polling booth (hence the name) for which candidate they voted.

Example 12.5

In American presidential elections the candidate who receives the most votes in a state receives the state's entire Electoral College vote.

In practice, this means that either the Democrat or the Republican candidate will win.

Suppose that the results of an exit poll in one state were recorded where 1 = Democrat and 2 = Republican.

Xm12-05*

Example 12.5

The polls close at 8:00 P.M.

Can the networks conclude from these data that the Republican candidate will win the state?

Should the network announce at 8:01 P.M. that the Republican candidate will win?

Example 12.5

IDENTIFY

The problem objective is to describe the population of votes in the state. The data are nominal because the values are "Democrat" and "Republican." Thus the parameter to be tested is the proportion of votes in the entire state that are for the Republican candidate. Because we want to determine whether the network can declare the Republican to be the winner at 8:01 P.M., the alternative hypothesis is

$$H_1: p > .50$$

And hence our null hypothesis becomes:

$$H_0: p = .50$$

Example 12.5

IDENTIFY

The test statistic is

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

Example 12.5

INTERPRET

At the 5% significance level we reject the null hypothesis and conclude that there is enough evidence to infer that the Republican candidate will win the state.

However, is this the right decision?

Example 12.5

INTERPRET

One of the key issues to consider here is the cost of Type I and Type II errors.

A Type I error occurs if we conclude that the Republican will win when in fact he has lost.

Example 12.5

INTERPRET

Such an error would mean that a network would announce at 8:01 P.M. that the Republican has won and then later in the evening would have to admit to a mistake.

If a particular network were the only one that made this error it would cast doubt on their integrity and possibly affect the number of viewers.

Example 12.5

INTERPRET

This is exactly what happened on the evening of the U. S. presidential elections in November 2000.

Shortly after the polls closed at 8:00 P.M. all the networks declared that the Democratic candidate Albert Gore would win in the state of Florida.

A couple of hours later, the networks admitted that a mistake had been made and the Republican candidate George W. Bush had won.

Example 12.5

INTERPRET

Several hours later they again admitted a mistake and finally declared the race too close to call.

Fortunately for each network all the networks made the same mistake.

However, if one network had not done this it would have developed a better track record, which could have been used in future advertisements for news shows and likely drawn more viewers.

Considering the costs of Type I and II errors it would have been better to use a 1%significance level.