

Inferences for a Single Population

OPRE 6301

Introduction . . .

In this chapter, we will introduce further procedures for making statistical inferences about a *single* population. As discussed in the previous two chapters, the basic steps we follow are:

- Identify a population parameter to be estimated or tested.
- Develop an estimator for the parameter and describe the sampling distribution of the estimator.
- Construct a confidence interval for the parameter or conduct a hypothesis test.

We will consider three parameters that are most important in applications: population mean μ , population variance σ^2 , and population proportion p .

Since most of the relevant concepts have been carefully explained in the previous three chapters, we will be brief.

Inferring μ When σ^2 is Unknown ...

Inferences for μ when σ^2 is known was discussed previously; the rationale at that time was that we may have reasonable “historical” information regarding the population variance. We now consider the same problem *without* assuming that σ^2 is known. This scenario is much more realistic in most applications.

Consider the example of a mid-scale department store. Suppose the store manager wishes to infer about the mean amount (in dollars) of purchase made by customers per visit. A random sample of $n = 25$ customers is taken and the sample mean and the sample variance of the sample are computed as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = 35.00$$

and

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = 900.00.$$

Recall that if we knew σ^2 , then

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

would be normally distributed for large n . Since we don't know σ , a natural question is: What can we say about the above ratio when σ is replaced by s ? This important question was answered by William Gossett (under the pseudonym "Student") in 1908, and the answer is that *if the population is normally distributed*, then

$$t \equiv \frac{\bar{X} - \mu}{s/\sqrt{n}}, \quad (1)$$

the so-called (Student) t -statistic, follows the (Student) t distribution with $n - 1$ "degrees of freedom."

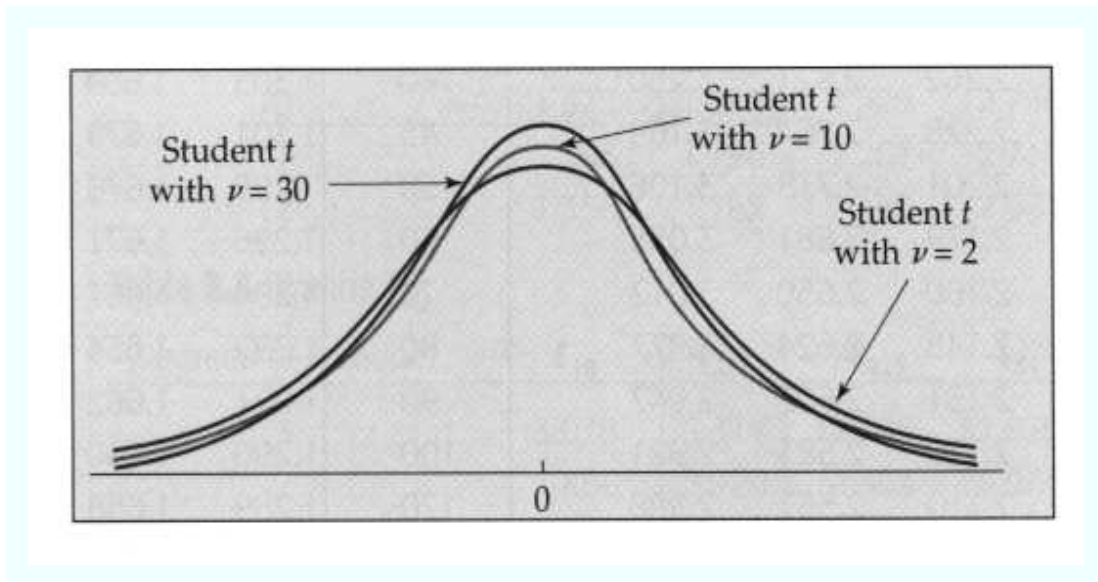
If the population is not extremely nonnormal, the above statement has been shown to be a good approximation to the "true" sampling distribution of the t statistic.

An informal way to understand the notion of degrees of freedom is to note that the presence of \bar{X} in the formula for s^2 introduces a “constraint” on the otherwise “freely-varying” X_1, X_2, \dots, X_n and therefore a loss of 1 degree of freedom for the X_i s.

The t distribution has a density that is symmetric and bell shaped. Its single parameter, degrees of freedom, is denoted by ν . Moreover, it can be shown that $E(t) = 0$ and, for $\nu > 2$, $V(t) = \nu/(\nu - 2)$. (A discussion of the t distribution can be found on pp. 260–266 of the text.)

The t density is flatter than the standard normal density ($\nu/(\nu - 2) > 1$ for $\nu > 2$) and as $\nu \rightarrow \infty$, the t densities converge to the standard normal density. This latter fact is not surprising, as it is known that s^2 is an unbiased and consistent estimator for σ^2 .

Pictorially, we have



It can be shown that when $\nu = 30$, the t density is almost indistinguishable from the standard normal density. Therefore, when (1) is used for statistical inference, the convention is to work with the t distribution if ν is less than 30, and to work with the standard normal otherwise.

Since $\nu = n - 1 = 24 < 30$ in our example, the $100(1 - \alpha)\%$ confidence interval would then be:

$$\left(\bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right), \quad (2)$$

where $t_{\alpha/2}$ denotes the two-sided critical value for the t density curve.

For $\alpha = 0.05$, the value $t_{\alpha/2}$ can be found using the Excel function `TINV()`. This function is by default **two sided**, and it takes two arguments: α and ν . In our case,

$$\begin{aligned} t_{\alpha/2} &= \text{TINV}(\alpha, n - 1) \\ &= \text{TINV}(0.05, 24) \\ &= 2.0639. \end{aligned}$$

Substitution of $\bar{X} = 35$, $t_{\alpha/2} = 2.0639$, $s = \sqrt{900}$, and $n = 25$ into (2) then yields the 95% confidence-interval estimate for μ :

$$(22.62, 47.38).$$

Once the confidence interval is constructed, it is very easy to test hypotheses of the form:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

As an example, suppose the proposed μ_0 is 45. Then, since 45 is contained in the above confidence interval, H_0 should be accepted. In fact, any H_0 with the specified μ_0 inside the interval (22.62, 47.38) would be accepted.

Of course, one could also use the two-sided p -value to conduct this test. From (1), we have

$$t = \frac{35 - 45}{30/\sqrt{25}} = -1.6667.$$

Now, using the Excel function TDIST(), we obtain the p -value as:

$$\text{TDIST}(\text{abs}(-1.6667), 25 - 1, 2) = 0.1086,$$

where the first argument (which, unlike NORMSDIST(), is required to be nonnegative) is the absolute value of our t statistic, the second is ν , and the third indicates that we need a two-tailed probability. Since 0.1086 is greater than $\alpha = 0.05$, we again would accept H_0 .

If the confidence interval constructed above is considered too wide, it is natural to attempt to pick n such that

$$n = \left(\frac{t_{\alpha/2} s}{w} \right)^2,$$

for a target interval of the form $\mu \pm w$. However, there are two issues with this formula:

- The value of $t_{\alpha/2}$ depends on n , which is what we are trying to determine. To circumvent this, note that we expect n to be greater than 30. Therefore, $t_{\alpha/2}$ can be well approximated by $z_{\alpha/2}$, which does *not* depend on n .
- The value of s also depends on n . To circumvent this, note that we already have a “pre-sample” of size 25. The idea is then to use $s = 30$ from the pre-sample.

Doing these with $w = 5$ yields that

$$n = \left(\frac{1.96 \cdot 30}{5} \right)^2 = 138,$$

where 1.96 comes from $\text{NORMSINV}(0.975)$.

Therefore, we would need to take a further $138 - 25 = 113$ observations from the population. A new confidence interval would then be constructed from the new sample mean and the new sample variance.

As a second example, suppose the production manager in a company claims that new workers could achieve 90% of the productivity level of experienced workers within one week of being hired and trained. The validity of this claim is considered important by his boss as it would have an impact on the rest of the operations in the company.

Suppose further that experienced workers can, on average, process 500 units of a product per hour. Thus, if the manager's claim is correct, then new workers should be able to process $0.9 \cdot 500 = 450$ products per hour, on average.

Our research hypothesis therefore is $\mu > 450$, implying that the null hypothesis should be $\mu = 450$:

$$H_0 : \mu = 450$$

$$H_1 : \mu > 450$$

A sample of $n = 50$ workers are taken and the productivity levels of these workers are measured. The sample mean turns out to be 460.38, and the sample standard deviation 38.83.

The t statistic is easily computed as:

$$t = \frac{460.38 - 450}{38.83/\sqrt{50}} = 1.89.$$

Let $\alpha = 0.05$. Noting that we are doing a **one-tailed** test, we have

$$t_{0.05} = \text{TINV}(0.1, 50 - 1) = 1.676$$

(note that $\text{NORMSINV}(0.95)$ yields 1.645, which is close; the t density is slightly flatter).

Since $1.89 > 1.68$, we reject H_0 in favor of H_1 , meaning that there is sufficient evidence to support the manager's claim.

Note that the **one-tailed** p -value for the observed \bar{X} is:

$$\text{TDIST}(1.89, 50 - 1, 1) = 0.0323,$$

which, as expected, is less than 0.05.

Inferring σ^2 ...

The extent of variability is an important attribute of a population. For example, it is important to understand:

- The *consistency* of a production process for quality-control purposes.
- The variance of investment return as a measure of risk.

The parameter of interest in such scenarios is σ^2 .

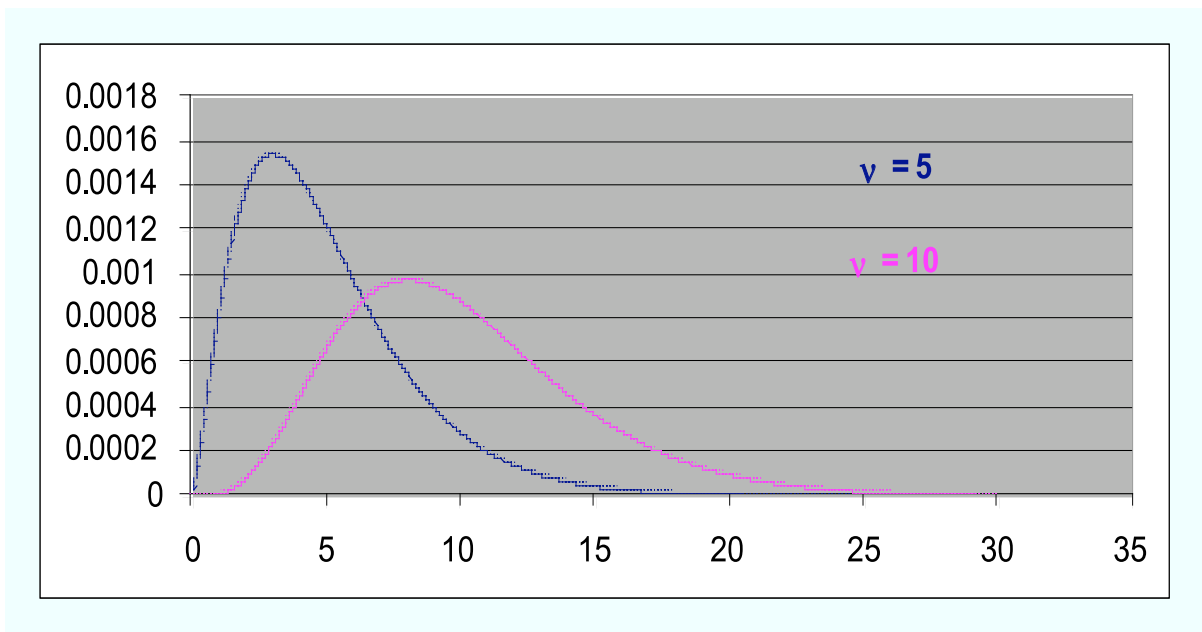
As noted earlier, the sample variance s^2 is an unbiased (this is why we required a division by $n - 1$, rather than by n , in the formula for sample variance) and consistent *point* estimator for σ^2 . How does one perform *interval* estimation and/or conduct hypotheses tests for σ^2 ? These require knowledge of the *sampling distribution* of s^2 ...

Fortunately, it can be shown that *if the population is normally distributed*, then

$$\chi^2 \equiv \frac{(n - 1)s^2}{\sigma^2}, \quad (3)$$

the so-called χ^2 (chi-squared) statistic, follows the chi-squared distribution with $\nu = n - 1$ degrees of freedom.

Unlike the “ z ” or t statistic, the χ^2 statistic is *not* symmetric, and is always nonnegative. It can be shown that $E(\chi^2) = \nu$ and $V(\chi^2) = 2\nu$. The shapes of the χ^2 density for $\nu = 5$ and 10 are shown below:



The Excel functions CHIDIST() and CHIINV() can be used in our calculations. Specifically, for a given degrees of freedom ν ,

$$\text{CHIDIST}(x, \nu) \equiv P(\chi^2 > x),$$

i.e., the probability of the **right** tail at x is returned; and

$$\text{CHIINV}(A, \nu) \equiv \chi_A^2,$$

which is the **right** critical value χ_A^2 such that

$$P(\chi^2 > \chi_A^2) = A.$$

A related discussion of the χ^2 distribution can be found on pp. 266–269 of the text.

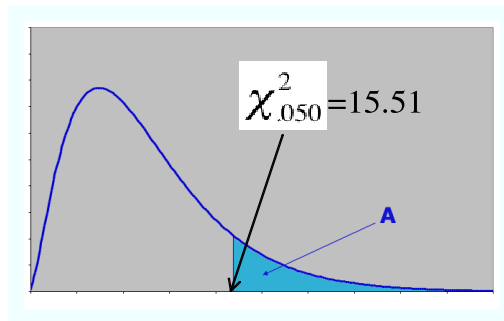
As an example, suppose $A = 0.05$ and $\nu = 8$. Then,

$$\chi_{0.05}^2 = \text{CHIINV}(0.05, 8) = 15.5073;$$

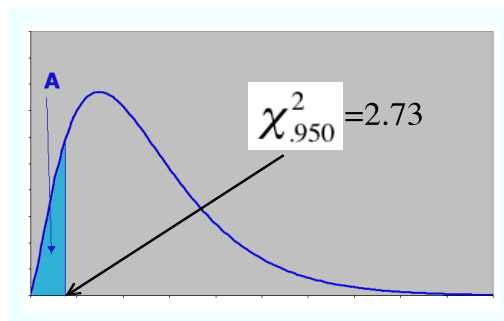
and conversely,

$$\text{CHIDIST}(15.5073, 8) = 0.05.$$

Pictorially, this means:



Note that if a *left* critical value is desired, one would need to work with $1 - A$, i.e., $P(\chi^2 \leq \chi_{1-A}^2) = A$. For $A = 0.05$, this means:



Armed with the applicable sampling distribution and the above two Excel functions, we can now perform statistical inferences regarding σ^2 .

The idea is to start with the fact that

$$P\left(\chi_{1-\alpha/2}^2 < \chi^2 \leq \chi_{\alpha/2}^2\right) = 1 - \alpha. \quad (4)$$

Upon substitution of (3), this rearranges into

$$P\left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2} < \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha. \quad (5)$$

It follows that the $100(1 - \alpha)\%$ confidence interval for σ^2 is given by:

$$\left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}\right). \quad (6)$$

Example 1: Filling Machine

A machine is used to fill liquid (milk or soft drink) into 1-liter (or 1,000 cc) containers. Ideally, the amount of liquid filled should vary only slightly. The manager of this operation would like to estimate the variance of the amount of liquid filled in a container.

Suppose a sample of 25 containers is taken. For this sample, it is found that $s^2 = 0.8088$ cc.

Analysis: The *point* estimate for σ^2 is simply 0.8088. The 95% confidence interval for σ^2 is given by (6) with $n = 25$ and $\alpha = 0.05$. That is,

$$\left(\frac{(25 - 1) \cdot 0.8088}{\chi_{0.025}^2}, \frac{(25 - 1) \cdot 0.8088}{\chi_{1-0.025}^2} \right),$$

where

$$\chi_{0.025}^2 = \text{CHIINV}(0.025, 25 - 1) = 39.364$$

and

$$\chi_{0.975}^2 = \text{CHIINV}(0.975, 25 - 1) = 12.401.$$

The above simplifies to the interval $(0.493, 1.565)$, and this is our interval estimate for σ^2 .

As noted before, it is now very easy to test hypotheses of the form:

$$\begin{aligned}H_0 &: \sigma^2 = \sigma_0^2 \\H_1 &: \sigma^2 \neq \sigma_0^2\end{aligned}$$

As an example, suppose the proposed σ_0^2 is 1. Then, since 1 is contained in the above confidence interval, H_0 should be accepted.

Example 2: Filling Machine — Continued

Suppose the operations manager “boasts” that the filling machine is so consistent that the variance of fills does not exceed 1 cc. Is this justified? Use the same data.

Analysis: Our research hypothesis is $\sigma^2 < 1$. Therefore,

$$\begin{aligned}H_0 &: \sigma^2 = 1 \\H_1 &: \sigma^2 < 1\end{aligned}$$

For this test, we need a *left* critical value for the χ^2 density. Recall that (similar to (4))

$$P(\chi^2 \leq \chi_{1-\alpha}^2) = \alpha.$$

For $\alpha = 0.05$ and $n = 25$, we have

$$\chi_{1-0.05}^2 = \text{CHIINV}(0.95, 25 - 1) = 13.848.$$

Now, if H_0 is true, i.e., $\sigma^2 = 1$, then the χ^2 statistic for our sample is:

$$\begin{aligned}\chi^2 &= \frac{(n-1)s^2}{\sigma^2} \\ &= \frac{(25-1) \cdot 0.8088}{1} \\ &= 19.41.\end{aligned}$$

Since 19.41 is greater than 13.848, we should accept H_0 . That is, the operations manager is overstating his case.

Inferring Population Proportion p ...

When the attribute of interest in a population is **nominal**, data collected from a sample will yield frequency counts. In such scenarios, it is natural to use the sample proportion to infer about the corresponding population proportion.

Recall that a sample proportion can be viewed as the fraction of “successes” in n independent Bernoulli trials with the same success probability p . Formally, we have

$$\hat{p} \equiv \frac{1}{n} \sum_{i=1}^n X_i = \frac{X}{n}, \quad (7)$$

where X_i equals 1 if the i th trial is a success and 0 otherwise, and where X equals the total number of successes in n trial.

Recall also from Chapter 9 that the central limit theorem implies that for large n , the sample proportion defined in (7) is normally distributed with mean p and standard deviation $\sqrt{p(1-p)/n}$. It follows that the statistic

$$z \equiv \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \quad (8)$$

follows the standard normal distribution; and this serves as the basis for statistical inference regarding population proportions.

Example 1: Election

Voters are asked by a certain network to participate in an exit poll in order to predict the winner in a state. Based on the collected data (see Xm12-05.xls, where “1” means Democrat and “2” means Republican), can the network project that the Republican candidate will win that state (and hence the state’s entire Electoral College vote)?

Analysis: Let p be the proportion of votes for the Republican candidate. Our research hypothesis is $p > 0.5$. Therefore,

$$H_0 : p = 0.5$$

$$H_1 : p > 0.5$$

Let $\alpha = 0.05$; then, $z_\alpha = \text{NORMSINV}(0.95) = 1.645$. We need to compare the z statistic against this (one-sided) critical value.

From the data file, we obtain $X = 407$ and $n = 765$. Hence, $\hat{p} = 407/765 = 0.532$. Assuming that the null hypothesis (i.e., $p = 0.5$) is true, we have

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \\ &= \frac{0.532 - 0.5}{\sqrt{0.5 \cdot (1 - 0.5)/765}} \\ &= 1.77. \end{aligned}$$

Since $1.77 > 1.645$, we conclude that there is sufficient evidence to project a win for the Republican candidate.

Example 2: Election — Confidence Interval

Construct a 95% interval estimate for p .

Analysis: If we *knew* p , then (8) implies that the confidence interval would be

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

This, however, is a circular argument, since p is actually unknown. To circumvent this problem, a natural approach is to **replace p by \hat{p}** in the above formula. It can be shown that this is a good approximation. Therefore, the approximate confidence interval is:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \quad (9)$$

Substituting $\hat{p} = 0.532$, $z_{0.025} = 1.96$, and $n = 765$ into (9) yields the interval (0.497, 0.567).

Example 3: Election — Sample Size

What sample size will allow us to reduce the width of the above confidence interval to $\hat{p} \pm w$, where $w = 0.03$ (i.e., within 3 percentage points)?

Analysis: From (9), we see that n should be chosen such that

$$n = \left(\frac{z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})}}{w} \right)^2. \quad (10)$$

Again, this is circular, since we don't yet have \hat{p} . Two approaches can be taken.

Method 1: Take a pre-sample, as we did in a previous example. Then, plug \hat{p} from the pre-sample into (10) to estimate the necessary sample size n .

Method 2: A *conservative* (i.e., larger than necessary) bound for n can be obtained by the simple observation that $\hat{p}(1 - \hat{p}) \leq 0.5(1 - 0.5) = 0.25$ for *all* \hat{p} in the interval $(0, 1)$. That is, an upper bound for n is:

$$n_u = \left(\frac{z_{\alpha/2}}{2w} \right)^2,$$

which is obtained from the right-hand side of (10) by setting $\hat{p} = 0.5$. In our case, we have $n_u = (1.96/(2 \cdot 0.03))^2 = 1068$. Note that the actual w obtained from a sample of size 1,068 will always be narrower than the specified 0.03.