4.3 Confidence Intervals for the Mean and Variance of One Population

Let X be a population characteristic, with mean $\mu = E(X)$ and variance $V(X) = \sigma^2$, whose pdf depends on a parameter θ, $f(x;\theta)$. Let $X_1, X_2, \ldots, X_n$ be a sample drawn from the pdf of X.
The formulas for finding confidence intervals for the mean μ and variance σ^2 are based on the following results (which were discussed in Chapter 3).

Proposition 4.1. *Assume $X \in N(\mu, \sigma)$. Then*

a) $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \in N(0,1)$;

b) $T = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}} \in T(n-1)$;

c) $V = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{(n-1)\,s^2}{\sigma^2} \in \chi^2(n-1)$.

Proposition 4.2. *If the sample size is large enough ($n > 30$), then*

a) $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \in N(0,1)$;

b) $T = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}} \in T(n-1)$.

CI for the mean, known variance

If either $X \in N(\mu, \sigma)$ or the sample is large enough ($n > 30$) and σ is known, then by Propositions 4.1 and 4.2, we can use the pivot

$$Z = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \in N(0,1).$$

The procedure will go *exactly* as described in the previous section, with $\theta = \mu, \overline{\theta} = \overline{X}, \sigma_{\overline{\theta}} = \dfrac{\sigma}{\sqrt{n}}$.

The $100(1-\alpha)\%$ CI for the mean is given by

$$\mu \in \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]. \tag{4.1}$$

Since $N(0,1)$ is symmetric (and one quantile is the negative of the other), we can write it in short as

$$\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \overline{X} \mp z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}. \tag{4.2}$$

CI for the mean, unknown variance

In practice, it is somewhat unreasonable to expect to know the value of σ, if the value of μ is unknown. We can find CI's for the mean, without knowing the variance. If either $X \in N(\mu, \sigma)$ or the sample is large enough ($n > 30$), then by Propositions 4.1 and 4.2, we can use the pivot

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1).$$

The same computations as before will lead to the $100(1-\alpha)\%$ CI for the mean:

$$\mu \in \left[\overline{X} - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}},\ \overline{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right]. \tag{4.3}$$

Notice that we change the notations for the quantiles, according to the pdf of the pivot (z for $N(0,1)$, t for $T(n-1)$, etc.). The Student $T(n-1)$ is also symmetric (see Figure 1), so again, we can write the CI in short as

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad \text{or} \quad \overline{X} \mp t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}. \tag{4.4}$$

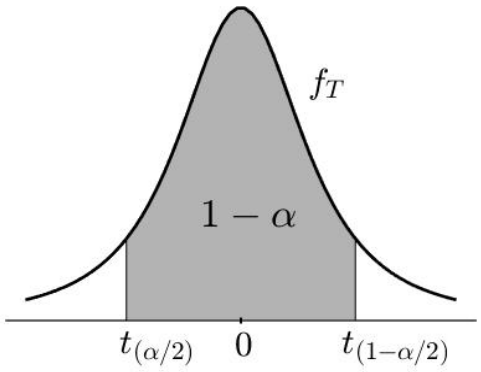

Fig. 1: Confidence Interval for the T distribution

Remark 4.3. The parameter of a Student T distribution, ν, is generally called *number of degrees of freedom*. One might wonder why in estimating the mean, this parameter is $\nu = n - 1$ and not $\nu = n$, the sample size. The sample variables $X_1, \ldots, X_n$ are independent, so it would seem that there are $\nu = n$ degrees of freedom. But its meaning is the dimension of the vector used to estimate the sample variance

$$s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2,$$

where we use the vector $X_1 - \overline{X}, \ldots, X_n - \overline{X}$. Notice that by subtracting the sample mean $\overline{X}$ from each observation, there exists a linear relation among the elements, namely

$$\sum_{k=1}^{n} (X_k - \overline{X}) = 0,$$

so we lose 1 degree of freedom due to this constraint. In general, the number of degrees of freedom can be computed by

$$\nu \quad = \quad \text{sample size} \quad - \quad \text{number of estimated parameters.}$$

However, it should be noted that this issue is important only when the sample size is *small* ($n < 30$), when there is significant difference in the values of the quantiles. When n is large, we may use the quantiles for $T(\nu)$ with $\nu = n$ or $\nu = n - 1$, since both distributions

$$T(n),\ T(n-1) \overset{n \to \infty}{\longrightarrow} N(0,1),$$

so both quantiles are approximately equal to the z quantiles.

CI for the variance

By Proposition 4.1, if $X \in N(\mu, \sigma)$, then we can use the pivot

$$V = \frac{(n-1)\, s^2}{\sigma^2} \in \chi^2(n-1).$$

Let us see how to do that. Even though the $\chi^2(n-1)$ is not symmetric (see Figure 2), so we cannot really talk about the "middle" for the area, we can still use the quantiles as before. So, we have:

$$1 - \alpha \quad = \quad P\left(\chi^2_{\frac{\alpha}{2}} \leq V \leq \chi^2_{1-\frac{\alpha}{2}}\right)$$

$$
\begin{aligned}
&= P\left(\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)\,s^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\right) \\
&= P\left(\frac{1}{\chi^2_{1-\frac{\alpha}{2}}} \leq \frac{\sigma^2}{(n-1)\,s^2} \leq \frac{1}{\chi^2_{\frac{\alpha}{2}}}\right) \\
&= P\left(\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}}\right).
\end{aligned}
$$

Thus, a $100(1-\alpha)\%$ CI for the variance is

$$
\sigma^2 \in \left[\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}},\ \frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}}\right] \tag{4.5}
$$

and one for the standard deviation is

$$
\sigma \in \left[\sqrt{\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}}},\ \sqrt{\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}}}\right] \tag{4.6}
$$

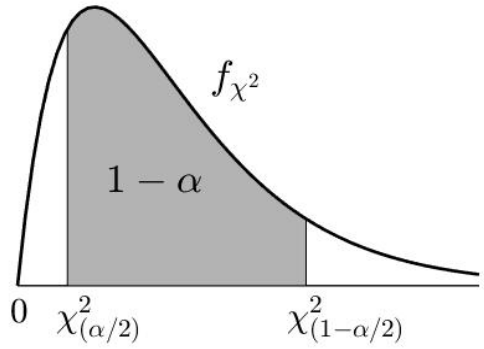

Fig. 2: Confidence Interval for the χ^2 distribution

Remark 4.4.

1. Remember, "χ^2_α" is just a notation for the quantile of order α for the $\chi^2(n-1)$ distribution, it *does not* mean you have to take the square of it!

2. Since the $\chi^2(n-1)$ is no longer symmetric, there is no relationship between the two quantiles, we have to use *both* and there is no shorter writing for the CI for the variance than the one in (4.5) (or (4.6) for the standard deviation).

3. Again, the parameter of the χ^2 distribution is number of degrees of freedom and it is $n-1$ (and not n) for the same reason as the one mentioned in Remark 4.3.

Selecting the sample size

Notice that in the case of a Normal distribution of the pivot, the CI we find is symmetric and the length of the CI is

$$2\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}.$$

We can revert the problem and ask a very practical question: How large a sample should be collected to provide a certain desired precision of our estimator? In other words, what sample size n guarantees that the margin of a $(1-\alpha)100\%$ CI does not exceed a specified limit Δ? To answer this question, we only need to solve the inequality

$$2\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} \quad \leq \quad \Delta \tag{4.7}$$

in terms of n. Typically, parameters are estimated more accurately based on larger samples, so that the standard error $\sigma_{\overline{\theta}}$ and the margin are decreasing functions of the sample size n. Then, (4.7) will be satisfied for sufficiently large n.

For example, when estimating the mean in the case of known variance, inequality (4.7) comes down to

$$2\frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \quad \leq \quad \Delta,$$

so we require

$$n \quad \geq \quad \left(\frac{2\sigma z_{1-\frac{\alpha}{2}}}{\Delta}\right)^2 \tag{4.8}$$

Example 4.5. Consider a sample of measurements

$$2.5,\ 7.4,\ 8.0,\ 4.5,\ 7.4,\ 9.2,$$

drawn from an approximately Normal distribution.
a) Find a 95% confidence interval for the population mean, if the measurement device guarantees a standard deviation of $\sigma = 2.2$.
b) How many measurements should be taken in order for the length of the 95% confidence interval

for the mean to not exceed 1?

Solution. This sample has size $n = 6$ and sample mean $\overline{X} = 6.5$. To attain a confidence level of $1 - \alpha = 0.95$, we need $\alpha = 0.05$ and $\alpha/2 = 0.025$.
a) If $\sigma = 2.2$ is known, we use formula (4.1). Hence, we need quantiles

$$z_{0.025} \ = \ -1.96, \ z_{0.975} \ = \ 1.96.$$

We find the 95% CI for the mean

$$\left[\overline{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] \ = \ [4.74, \ 8.26].$$

That means that the mean μ of the population from which the sample was drawn is between 4.74 and 8.26 with probability 0.95.

d) Notice that the length of the CI found in part a) is ≈ 3.52, quite large (not much precision). If we want to improve the accuracy of our estimate (shorten the length of the interval), we need to enlarge the sample, take more measurements.
With $\sigma = 2.2$, $z_{0.975} = 1.96$ and $\Delta = 1$, we find from (4.8),

$$n \ \geq \ \left(\frac{2\sigma z_{1-\frac{\alpha}{2}}}{\Delta}\right)^2 \ = \ 74.37,$$

so, a sample of size at least 75 will ensure the fact that the length of the 95% CI for the mean does not exceed 1.

■

4.4 Confidence Intervals for Comparing Means and Variances of Two Populations

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations, with means $\mu_1 = E(X_{(1)}), \mu_2 = E(X_{(2)})$ and variances $\sigma_1^2 = V(X_{(1)}), \sigma_2^2 = V(X_{(2)})$, respectively.
We draw from both populations random samples of sizes n_1 and n_2, respectively, that are **independent**. Denote the two sets of random variables by

$$X_{11}, \ldots, X_{1n_1} \ \text{ and } \ X_{21}, \ldots, X_{2n_2}.$$

Then we have two sample means and two sample variances, given by

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

and

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2, \;\; s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2,$$

respectively. In addition, denote by

$$s_p^2 = \frac{\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the *pooled variance* of the two samples, i.e. a variance that considers ("pools") the sample data from both samples.

Recall that when comparing the means of two populations, we estimate their difference and when comparing the variances, we estimate their ratio.

The formulas for finding confidence intervals for the difference of means $\mu_1 - \mu_2$ and for the ratio of variances $\dfrac{\sigma_1^2}{\sigma_2^2}$ are based on the following results.

Proposition 4.6. *Assume $X_{(1)} \in N(\mu_1, \sigma_1)$ and $X_{(2)} \in N(\mu_2, \sigma_2)$. Then*

a) $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \in N(0, 1)$;

b) $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \in T(n_1 + n_2 - 2)$;

c) $T^* = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \in T(n)$, *where* $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1 - c)^2}{n_2 - 1}$ *and* $c = \dfrac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$;

d) $F = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F(n_1 - 1, n_2 - 1)$.

Proposition 4.7. *If the samples are large enough ($n_1 + n_2 > 40$), then parts a), b) and c) of Proposition 4.6 still hold.*

CI for the difference of means

Case σ_1, σ_2 known

If either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$) and σ_1, σ_2 are known, then by Propositions 4.1 and 4.2, we can use the pivot

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0, 1).$$

With the same line of computations as before, we find a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ as

$$\mu_1 - \mu_2 \ \in \ \left[\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right], \tag{4.9}$$

or, using symmetry,

$$\left[\overline{X}_1 - \overline{X}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]. \tag{4.10}$$

Case $\sigma_1 = \sigma_2$ unknown

Assume that either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$). The population variances are *not* known anymore, but they are known to be equal. Then each is approximated by the pooled variance s_p^2. Then by Propositions 4.6 and 4.7, we use the pivot

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in T(n_1 + n_2 - 2).$$

A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$\mu_1 - \mu_2 \ \in \ \left[\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right], \tag{4.11}$$

where the quantiles $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$ refer to the $T(n_1 + n_2 - 2)$ distribution. Again, by symmetry we can

write the CI in short as

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{4.12}$$

Case σ_1, σ_2 unknown

Assuming that either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$), by Propositions 4.6 and 4.7, we use the pivot

$$T^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \in T(n),$$

where $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}$ and $c = \dfrac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

We find a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ as

$$\mu_1 - \mu_2 \ \in \ \left[\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right], \tag{4.13}$$

or, by symmetry,

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] \tag{4.14}$$

where the quantile $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$ refer to the $T(n)$ distribution, with n given above.

CI for the ratio of variances

Assume the two independent samples were drawn from approximately Normal distributions $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively. By Proposition 4.7, we use the pivot

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F(n_1 - 1, n_2 - 1).$$

A $100(1-\alpha)\%$ CI for $\dfrac{\sigma_1^2}{\sigma_2^2}$ is given by

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{1}{f_{1-\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2},\ \frac{1}{f_{\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2}\right] \tag{4.15}$$

and, from here, a $100(1-\alpha)\%$ CI for $\dfrac{\sigma_1}{\sigma_2}$ is

$$\frac{\sigma_1}{\sigma_2} \in \left[\sqrt{\frac{1}{f_{1-\frac{\alpha}{2}}}}\cdot\frac{s_1}{s_2},\ \sqrt{\frac{1}{f_{\frac{\alpha}{2}}}}\cdot\frac{s_1}{s_2}\right], \tag{4.16}$$

where the quantiles $f_{\frac{\alpha}{2}}, f_{1-\frac{\alpha}{2}}$ refer to the $F(n_1-1, n_2-1)$ distribution.

Example 4.8. An account on server A is more expensive than an account on server B. However, server A is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results:

Server A			Server B		
n_1	$=$	30	n_2	$=$	20
$\overline{X}_1$	$=$	6.7 min	$\overline{X}_2$	$=$	7.5 min
s_1	$=$	0.6 min	s_2	$=$	1.2 min

a) Construct a 95% confidence interval for the difference $\mu_1-\mu_2$ between the mean execution times on server A and server B.
b) Assuming that the observed times are approximately Normal, find a 95% confidence interval for the ratio of the two population standard deviations.

Solution.
a) The samples are large enough ($n_1+n_2=50$), that we can use Proposition 4.7. Nothing is said about the population variances (that they might be known, or known to be equal). Also, the second sample standard deviation is twice as large as the first one, therefore, equality of population variances can hardly be assumed. We use the general case for unknown, unequal variances and use formula (4.14).
We want confidence level $1-\alpha=0.95$, so $\alpha=0.05$ and $\alpha/2=0.025$.

The parameter n in (4.6) is found to be

$$n \;=\; 25.3989 \;\approx\; 25.$$

For the $T(25)$ distribution, we find the quantile

$$t_{0.025} \;=\; -2.0595.$$

Them the 95% CI for the difference of means is

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] \;=\; \left[6.7 - 7.5 \pm 2.06\sqrt{\frac{0.6^2}{30} + \frac{1.2^2}{20}}\right] \;=\; [-0.8 \pm 0.505],$$

so,

$$\mu_1 - \mu_2 \;\in\; [-1.305,\; -0.295]$$

with probability 0.95. Since *all* values in the CI are negative, with high probability, it seems that $\mu_1 - \mu_2 < 0$, so indeed the first server seems to be faster, on average.

b) Since now the times are assumed to be approximately Normal, we can use formula (4.15). For the $F(29, 19)$ distribution, the quantiles are

$$f_{0.025} \;=\; 0.4482,\;\; f_{0.975} \;=\; 2.4019.$$

Now,

$$\begin{aligned}\frac{s_1}{s_2} \;&=\; \frac{0.6}{1.2} \;=\; 0.5,\\ \frac{s_1^2}{s_2^2} \;&=\; \frac{0.36}{1.44} \;=\; 0.25.\end{aligned}$$

Then, the 95% CI for the ratio of variances is

$$\left[\frac{1}{f_{1-\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2},\;\; \frac{1}{f_{\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2}\right] \;=\; \left[\frac{1}{2.4019}\cdot 0.25,\;\; \frac{1}{0.4482}\cdot 0.25\right] = [0.104,\;\; 0.558]$$

and the 95% CI for the ratio of standard deviations is

$$\left[\sqrt{\frac{1}{f_{1-\frac{\alpha}{2}}}}\cdot\frac{s_1}{s_2},\ \sqrt{\frac{1}{f_{\frac{\alpha}{2}}}}\cdot\frac{s_1}{s_2}\right] \ = \ \left[\sqrt{\frac{1}{2.4019}}\cdot 0.5,\ \sqrt{\frac{1}{0.4482}}\cdot 0.5\right] = [0.323,\ 0.747].$$

■

4.5 Confidence Intervals for Proportions

Recall (from Lecture 4) that a *population proportion* is

$$p \ = \ P(i \in A),$$

where A is a subpopulation.
Based on a random sample $X_1, \ldots, X_n$, we define the *sample proportion* as

$$\overline{p} \ = \ \frac{\text{number of sampled items from } A}{n}.$$

Then

$$\begin{aligned} E(\overline{p}) &= p, \\ V(\overline{p}) &= \frac{p(1-p)}{n} = \frac{pq}{n}. \end{aligned} \tag{4.17}$$

So $\overline{p}$ is an absolutely correct estimator for p and by a CLT,

$$Z \ = \ \frac{\overline{p}-p}{\sqrt{\frac{pq}{n}}} \tag{4.18}$$

converges in distribution to a Standard Normal $N(0,1)$ variable, as $n \to \infty$.
Now, as p is unknown, we estimate the standard error $\sigma_{\overline{p}} = \sqrt{V(\overline{p})} = \sqrt{\frac{p(1-p)}{n}}$ by

$$s_{\overline{p}} \ = \ \frac{\overline{p}(1-\overline{p})}{n}.$$

So, again, for large samples ($n > 30$), we can use

$$Z = \frac{\overline{p} - p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \in N(0,1)$$

as a pivot to construct a confidence interval for p.
For a given confidence level $1-\alpha$, with the same computations as before, we obtain a $100(1-\alpha)\%$ CI for the population proportion p as

$$\left[\overline{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right]. \tag{4.19}$$

Selecting the sample size

Just as we did for the population mean (in the case of known variance), we can derive a formula for the sample size that will provide a certain precision of our interval estimator. The length of the CI in (4.19) is

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

Notice that for any $\overline{p} \in (0,1)$, we have

$$\overline{p}(1-\overline{p}) \ \leq \ \frac{1}{4}.$$

Then to get a desired precision

$$2\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} z_{1-\frac{\alpha}{2}} \ \leq \ \Delta,$$

we solve

$$2 \cdot \frac{1}{2} \frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \ \leq \ \Delta,$$

for n. We get

$$n \ \geq \ \left(\frac{z_{1-\frac{\alpha}{2}}}{\Delta}\right)^2. \tag{4.20}$$

CI for the difference of proportions

To estimate the difference of two population proportions $p_1 - p_2$, based on two independent samples of sizes n_1 and n_2, respectively, we use the estimator $\overline{p}_1 - \overline{p}_2$ for which we know (again, from Lecture 4) that

$$\begin{aligned} E(\overline{p}_1 - \overline{p}_2) &= p_1 - p_2, \\ V(\overline{p}_1 - \overline{p}_2) &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}, \\ s^2(\overline{p}_1 - \overline{p}_2) &= \frac{\overline{p}_1 \overline{q}_1}{n_1} + \frac{\overline{p}_2 \overline{q}_2}{n_2} \end{aligned} \tag{4.21}$$

with $q_i = 1 - p_i, \overline{q}_i = 1 - \overline{p}_i, i = 1, 2$. Also, for large samples ($n_1 + n_2 > 40$), by a CLT,

$$Z = \frac{\overline{p}_1 - \overline{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\overline{p}_1 \overline{q}_1}{n_1} + \frac{\overline{p}_2 \overline{q}_2}{n_2}}} \in N(0, 1). \tag{4.22}$$

Using Z as a pivot, we construct a $100(1-\alpha)\%$ CI for the difference of population proportions $p_1 - p_2$ as

$$\left[\overline{p}_1 - \overline{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}} \right]. \tag{4.23}$$

Example 4.9. A company has to accept or reject a large shipment of items. For quality control purposes, they collect a sample of 200 items and find 12 defective items in it.
a) Find a 99% confidence interval for the proportion of defective items in the whole shipment.
b) How many items should be tested to ensure a 99% confidence interval of length at most 0.05?

Solution. The sample is large enough and we have

$$\overline{p} = \frac{12}{200} = 0.06.$$

For $1 - \alpha = 0.99$, $\alpha = 0.01$, $\alpha/2 = 0.005$, the quantile is

$$z_{0.005} = -2.576.$$

Then the 99% confidence interval for the proportion of defective items is

$$\left[\overline{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \right] = \left[0.06 \pm 2.576 \sqrt{\frac{0.06 \cdot 0.94}{200}} \right] = [0.017,\ 0.103].$$

So, with 99% confidence, the percentage of defective items is between 1.7% and 10.3%.

b) The length of the 99% CI we found is 0.086. For a margin of $\Delta \leq 0.05$ of the 99% CI, we need a sample size of

$$n \;\geq\; \left(\frac{z_{0.995}}{\Delta}\right)^2 \;=\; \left(\frac{2.576}{0.05}\right)^2 \;=\; 2653.898 \;\approx\; 2654.$$

■

Example 4.10. Two candidates prepare for the local elections. During a phone poll, 42 out of 70 randomly selected people said they would vote for candidate A and 59 out of 100 randomly selected people said they preferred candidate B and would vote for him. Estimate the difference in support for the two candidates with 95% confidence. Can we state affirmatively that candidate A gets a stronger support than candidate B?

Solution. We have

$$\begin{aligned} n_1 &= 70,\ n_2 = 100, \\ \overline{p}_1 &= 42/70 = 0.6, \\ \overline{p}_2 &= 59/100 = 0.59. \end{aligned}$$

For the confidence interval, we want $1-\alpha = 0.95$, so we compute the quantile

$$z_{0.025} \;=\; -1.96.$$

We find the 95% CI for the difference of proportions,

$$\left[0.6 - 0.59 \pm 1.96\sqrt{\frac{0.6\cdot 0.4}{70} + \frac{0.59\cdot 0.41}{100}}\right] \;=\; [0.01 \;\pm\; 0.15] \;=\; [-0.14,\ 0.16].$$

So, is the support stronger for candidate A? On one hand, the estimator $\overline{p}_1 - \overline{p}_2 = 0.01$ suggests that the support is 1% higher for candidate A than for B. On the other hand, the difference could appear positive just because of a sampling error. As we see, the 95% confidence interval includes a large range of negative values too. Therefore, the obtained data does not indicate affirmatively that the support for candidate A is stronger.

In the following sections, we will learn how to *test* if there is any significant difference between the two candidates, so that we can conclude for it or against it. ■