Estimation for the population mean

Recall the sampling distribution of the mean. From the CLT, this distribution (especially for large sample sizes) is normal with mean μ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ where n is the size of the collected sample. Then,

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

is distributed according to the standard normal distribution, N(0,1).

Thus, $P(-1.96 \le Z \le 1.96) = P(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96) = 0.95$ (i.e., Z is found between –1.96 and 1.96 about 95% of the time).

Thus, after some mathematical derivation (i.e., by multiplying all sides by $\sigma/\sqrt{n}$, then subtracting $\overline{X}$ and finally multiplying by –1.0), we have

$$P(\overline{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{X} + 1.96\sigma/\sqrt{n}) = 0.95$$

This essentially means that even though we do not know the exact value of μ, we *expect* it to be between $\overline{X} - 1.96\sigma/\sqrt{n}$ and $\overline{X} + 1.96\sigma/\sqrt{n}$ 95% of the time.

In this case, $\overline{X}$ is the *point estimate* of μ, while the interval $(\overline{X} - 1.96\sigma/\sqrt{n}, \overline{X} + 1.96\sigma/\sqrt{n})$ is the *95% confidence interval* for μ.

Characteristics of confidence intervals

In the previous simple example we saw how a 95% two-sided confidence interval is constructed. If 95% is not an acceptably high confidence, we may elect to construct a 99% confidence interval. Similarly to the last case, this interval will have the form interval $(\overline{X} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{\alpha/2}\sigma/\sqrt{n})$, where $z_{\alpha/2} = z_{0.005} = 2.58$, and consequently the 99% two-sided confidence interval of the population mean is

$$\left(\overline{X} - (2.58)\sigma/\sqrt{n}, \overline{X} + (2.58)\sigma/\sqrt{n}\right)$$

All else being equal therefore, higher confidence (99% versus 95%) gets translated to a *wider* confidence interval. This is intuitive, since the more

certain we are that the interval *covers* the unknown population mean, the more values we must allow this unknown quantity to take.

Example: Consider the distribution of cholesterol levels for all males in the United States who are hypertensive (have high systolic blood pressure) and smoke. This distribution has an unknown mean μ and standard deviation σ=46.

If we draw a sample of n=12 subjects from this group of hypertensive smokers and compute their (sample) mean cholesterol level as $\bar{x}_{12} = 217 mg/ml$, the 95% confidence interval based on information from this sample is

$$\left(217 - 1.96\frac{46}{\sqrt{12}}, 217 + 1.96\frac{46}{\sqrt{12}}\right) = (191, 243)$$

In other words we are 95% confident that the interval (191, 243) covers the unknown mean of the population of hypertensive smokers.

Note that approximately 5% of the time the confidence interval that we compute will not cover the unknown population mean.

One-sided confidence intervals

Just as in the case of one-sided hypothesis testing, there are occasions where we are only interested in an *upper* or *lower limit* of the range of values that we will consider for the estimated quantity.

In those cases we construct a *one-sided* confidence interval.

In the case where only an upper limit is sought, we consider only the upper tail of the normal distribution. Conversely, when a lower limit is considered, we are concentrating in the *lower* tail of the normal distribution.

For example, an *upper* one-sided 95% confidence interval (when the population standard deviation is known) is constructed as

$$\left(-\infty, \overline{X}_n + z_{0.05}\frac{\sigma}{\sqrt{n}}\right) = \left(-\infty, \overline{X}_n + 1.645\frac{\sigma}{\sqrt{n}}\right),$$

while a *lower* one-sided confidence interval is constructed as

$$\left(\overline{X}_n - z_{0.05}\frac{\sigma}{\sqrt{n}}, +\infty\right) = \left(\overline{X}_n - 1.645\frac{\sigma}{\sqrt{n}}, +\infty\right)$$

Example: Suppose that we select 74 children that have been exposed to high levels of lead, and we calculate their mean hemoglobin levels as $\overline{X}_{74} = 10.6 g/100 ml$.

Since there maybe some concern that exposure to lead is associated with lower levels of hemoglobin, we are interested only in an upper limit of this value in the group of lead-exposed children.

Based on this sample, and knowledge of the population standard deviation of 0.85g/ml, the 95% upper one-sided confidence interval is

$$\left(-\infty, \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right) = \left(-\infty, 10.6 + 1.645 \frac{0.85}{\sqrt{74}}\right)$$

The unknown population mean of hemoglobin level among lead-exposed children is at most 10.8g/ml.

When σ is unknown

In most cases knowledge of the true variability of the measurement will not be available.

In these cases, we proceed as before, substituting $s=\sqrt{\frac{1}{(n-1)}\Sigma_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}$ but now basing our inference on the *t* distribution with *n*-1 degrees of freedom (where *n* is again the size of the sample).

Two-sided confidence intervals:

$$\left(\overline{X} - t_{\alpha/2,n-1}\, s/\sqrt{n}, \overline{X} + t_{\alpha/2,n-1}\, s/\sqrt{n}\right)$$

One-sided confidence intervals:

$$\left(-\infty, \overline{X} + t_{\alpha,n-1}\, s/\sqrt{n}\right)$$

$$\left(\overline{X} - t_{\alpha,n-1}\, s/\sqrt{n}, +\infty\right)$$

for the (1-α)% upper- and lower one-sided confidence interval respectively.

Example:

In estimating the plasma aluminum level among infants that have taken antacids containing aluminum, a random sample of n=10 infants was collected. The sample mean plasma aluminum level in this sample is $\bar{x}_{10} = 37.2 \mu g/l$, while the sample standard deviation is $s = 7.13 \mu g/l$.

If the mean and standard deviation of the plasma aluminum level in the population is unknown, a 95% two-sided confidence interval is based on the t distribution with n-1=9 degrees of freedom

$$\left(\overline{X} - t_{\alpha/2,n-1} s/\sqrt{n}, \overline{X} + t_{\alpha/2,n-1} s/\sqrt{n}\right) = \left(37.2 - 2.262\frac{7.13}{\sqrt{10}}, 37.2 + 2.262\frac{7.13}{\sqrt{10}}\right) = (32.1, 42.3)$$

Example (continued):

Compare the previous interval to the 95% confidence interval based on the normal distribution derived if we pretend that the estimate of the standard deviation of 7.13 $\mu g/l$ is the true population standard deviation. This interval is (32.8, 41.6) and has length 8.8(=41.6-32.8) whereas the one based on the *t* distribution has length 10.2(=42.3-32.1).

This loss of accuracy (widening of the confidence interval) is the “penalty” we pay for the lack of knowledge of the true population standard deviation.

Example (continued):

In the example above, a 95% (two-sided) confidence interval is as follows:

```
. cii 10 37.2 7.13

Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
         |      10        37.2    2.254704        32.09951    42.30049
```

Which agrees with our hand calculations (i.e., the 95% CI is (32.1, 42.3).

Caution! STATA only produces two-sided confidence intervals. If you want to obtain one-sided confidence intervals for the aluminum example, you have to use the `level(#)` option as follows:

```
. cii 10 37.2 7.13, level(90)

Variable |     Obs        Mean    Std. Err.       [90% Conf. Interval]
---------+--------------------------------------------------------------
         |      10        37.2    2.254704        33.06687    41.33313
```

Thus, an *upper* **95%** confidence interval would be (-∞, 41.3), while a *lower* **95%** confidence interval would be (33.1, +∞).

Confidence intervals of a difference of two means

In similar fashion as with two-sample tests of hypothesis, inference is based on the statistic

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

where s_p is the pooled estimate of the population standard deviation. In this case, t is distributed according to a t distribution with n_1+n_2-2 degrees of freedom.

The two-sided confidence interval for the difference of two means is

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \left((\bar{x}_1 - \bar{x}_2\right) + t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right)$$

Serum iron levels and cystic fibrosis (continued)

In the iron levels in healthy children and among those with cystic fibrosis, recall that $\bar{x}_1 = 18.9\mu\text{mol/l}$ was the sample mean iron level in a sample of $n_1 = 9$ healthy children with standard deviation $s_1 = 5.9m\text{mol/l}$, while these figures collected from $n_2 = 13$ children with cystic fibrosis $\bar{x}_2 = 11.9\mu\text{mol/l}$ and $s_2 = 6.3\mu\text{mol/l}$ respectively.

With this information, a two-sided confidence interval of the true difference in iron levels between healthy children and children with cystic fibrosis can be calculated.

With the *pooled* estimate of the std. deviation $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} = 6.14$

we have,

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right)$$

$$\Downarrow$$

$$\left((18.9 - 11.9) \pm (2.086)(6.14) \sqrt{\left(\frac{1}{9} + \frac{1}{13}\right)} \right)$$

$$\Downarrow$$

$$(1.4, 12.6)$$

How does this compare to the result of the hypothesis test (which as you may recall *rejected* the null hypothesis at the 5% level)?

Performing hypothesis testing by using confidence intervals

To perform tests of hypotheses using confidence intervals we proceed as follows:

STEP 1. Formulate the null and alternative hypotheses as before

STEP 2. Choose the alpha level

STEP 3. Construct the $(1-\alpha)\%$ confidence interval as described above. Use a one-sided or two-sided confidence interval depending on the test you want to carry out.

STEP 4. **Rejection rule.** Reject the null hypothesis (as described in STEP 1) if the confidence interval does *not* cover the hypothesized value (of the null hypothesis).

In the example of the iron levels of children with cystic fibrosis versus healthy children, we carry out the test of no difference in the iron levels as follows:

STEP 1. H_o: $\mu_1 = \mu_2$ (or equivalently, $\mu_1 - \mu_2 = 0$)

H_a: $\mu_1 \neq \mu_2$ (or equivalently, $\mu_1 - \mu_2 \neq 0$)

STEP 2. The alpha level is 5%

STEP 3. The two-sided 95% confidence interval of the *difference* of the two means is (1.4, 12.6)

STEP 4. Since the hypothesized value of zero difference (equality of the two means) is not covered by this interval we *reject* the null hypothesis, in favor of the alternative. That is, children with cystic fibrosis do not have the same iron levels as healthy children (in fact they have lower levels).

Serum iron levels and cystic fibrosis (continued)

STATA output is as follows:

```
. ttesti 9 18.9 5.9 13 11.9 6.3

Two-sample t test with equal variances

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       9        18.9    1.966667         5.9    14.36486    23.43514
       y |      13        11.9    1.747306         6.3    8.092948    15.70705
---------+--------------------------------------------------------------------
combined |      22    14.76364    1.482474     6.95342    11.68066    17.84661
```

```
---------+--------------------------------------------------------------------
    diff |                   7    2.663838                1.443331    12.55667
---------+--------------------------------------------------------------------
Degrees of freedom: 20

                    Ho: mean(x) - mean(y) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   2.6278                  t =   2.6278                  t =   2.6278
   P < t =   0.9919          P > |t| =   0.0161          P > t =   0.0081
```

Thus, a two-sided 95% confidence interval for the difference in serum iron levels between healthy children and children who are suffering from cystic fibrosis is (1.4, 12.6) as we saw before.

One-sided tests

If a 95% one-sided confidence interval were required (corresponding to a one-sided hypothesis test), the computer solution would be as follows:

```
. ttesti 9 18.9 5.9 13 11.9 6.3, level(90)

Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       9        18.9    1.966667         5.9    15.24289    22.55711
       y |      13        11.9    1.747306         6.3    8.785799     15.0142
---------+--------------------------------------------------------------------
combined |      22    14.76364    1.482474     6.95342    12.21268    17.31459
---------+--------------------------------------------------------------------
    diff |                   7    2.663838                2.40563    11.59437
------------------------------------------------------------------------------
Degrees of freedom: 20
                     Ho: mean(x) - mean(y) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =   2.6278                  t =   2.6278                  t =   2.6278
   P < t =   0.9919              P > |t| =   0.0161              P > t =   0.0081
```

The **95%** lower one-sided confidence interval for the difference of the mean serum iron level is then $(2.4, +\infty)$