1 Confidence intervals (interval estimates) contd

1.1 Two independent samples

Assume that we have given
observations $x_1, \ldots, x_{n_1}$ of independent random variables $X_1, \ldots, X_{n_1}$ where $X_i \in N(\mu_1, \sigma_1)$, $i = 1, \ldots, n_1$ and
observations $y_1, \ldots, y_{n_2}$ of independent random variables $Y_1, \ldots, Y_{n_2}$ where $Y_j \in N(\mu_2, \sigma_2)$, $j = 1, \ldots, n_2$.
Furthermore assume that the X_i's nd Y_j's are independent. This setup is termed **two independent samples**.

Example 1 Suppose that we would like to compare the suger content in two large batches of beets.

The observations $x_1, \ldots, x_{n_1}$ come from beets from batch 1 and

the observations $y_1, \ldots, y_{n_2}$ come from beets from batch 2

What we would like to know is if the (expected) sugar content is higher in one of the batches.
$\square$

One way to investigate if there is a difference between the expected suger content in the two batches, or in general between the expectations in two groups of random variables is to construct a confidence interval for the difference in expectations i.e. a confidence interval for

$$\mu_1 - \mu_2.$$

If $\mu_1 > \mu_2$ then $\mu_1 - \mu_2 > 0$, if $\mu_1 = \mu_2$ then $\mu_1 - \mu_2 = 0$, and if $\mu_1 < \mu_2$ then $\mu_1 - \mu_2 < 0$.

- We estimate $\mu_1 - \mu_2$ by
$$(\mu_1 - \mu_2)^*_{obs} = \bar{x} - \bar{y}.$$

We will now consider three different cases based on the assumptions regarding σ_1 and σ_2.

1.1.1 Confidence interval for $\mu_1 - \mu_2$ when σ_1 and σ_1 are known

First we will assume that both σ_1 and σ_2 are known.

- For the sample variable $\bar{X} - \bar{Y}$ we have that
$$\bar{X} - \bar{Y} \in N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$
since it is a linear combination of independent normally distributed random variables (this can also be found in the compiled formulae, Section 11.3).

- A confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha$ is, under the current assumptions, given by (see compiled formulae, Section 12.1 λ-method)

$$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} - \lambda_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x} - \bar{y} + \lambda_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \qquad (1 - \alpha)$$

1.1.2 Confidence interval for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2 = \sigma$ but σ is unknown

Now we will assume that $\sigma_1 = \sigma_2 = \sigma$, but that the common standard deviation σ is unknown.

- For the sample variable $\bar{X} - \bar{Y}$ it still holds that

$$\bar{X} - \bar{Y} \in N\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

  or

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in N(0,1)$$

  but since σ is unknown this is not of much use to us. Instead we estimate σ by

$$s = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

  (see the compiled formulae 11.2 b)). Then we have that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in t(n_1 + n_2 - 2).$$

  (see he compiled formulae 11.2 d)).

- A confidence interval for $\mu_1 - \mu_2$ with confidence level $1 - \alpha$ is, under the current assumptions, given by (see compiled formulae, Section 12.2 t-method)

$$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} - t_{\alpha/2}(f) \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x} - \bar{y} + t_{\alpha/2}(f) \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \qquad (1 - \alpha)$$

  where $f = n_1 + n_2 - 2$.

1.1.3 Confidence interval for $\mu_1 - \mu_2$ whenσ_1 and σ_2 are unknown

Finally we will assume that both σ_1 and σ_2 are unknown and not neccesarily the same.

- For the sample variable $\bar{X} - \bar{Y}$ it still holds that

$$\bar{X} - \bar{Y} \in N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

  but since σ_1 and σ_2 are both unknown this does not help. Instead we estimate σ_1 and σ_2 by s_1, and s_2, respectively, where

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2}, \qquad s_2 = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2}.$$

- A confidence interval for $\mu_1 - \mu_2$ with **approximate** confidence level $1 - \alpha$ is, under the given assumptions, given by (the compiled formulae, Section 12.3 Approximate method)

$$I_{\mu_1 - \mu_2} = \left(\bar{x} - \bar{y} - \lambda_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x} - \bar{y} + \lambda_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \qquad (\approx 1 - \alpha)$$

To obtain the last interval we have used:

1.2 The Approximate method

According to the compiled formulae the following holds: if θ^* is approximately $N(\theta, D)$ then

$$I_\theta = (\theta^*_{obs} - \lambda_{\alpha/2} \cdot D^*_{obs}, \theta^*_{obs} + \lambda_{\alpha/2} \cdot D^*_{obs}) \qquad (\approx 1-\alpha)$$

is a confidence interval for θ with approimate confidence level $1-\alpha$. Here D^*_{obs} denotes a suitable point estimate of D (i.e. the standard error).
There are numerous ways to end up with an approximately normal distibution, for instance you may use the Central limit theorem (problem 1314), or you may use a normal approximation of the Poisson distribution (problem 1320), or a normal approximation of the binomial distribution (see below).

Example 2 Opinion poll

Suppose that there have been two opinion polls one in May and one in November and that the results were the following:

	May	November
Number of interviews	n=3002	m=3273
Number of *-suppoters	x=120	y=141

We can model x as an outcome of $X \in B(n, p_1) = Bin(3002, p_1)$ and y as an outcome of $Y \in B(m, p_2) = Bin(3273, p_2)$.

In the news what is usually reported is if there has been a change in the proportion of *-suppoters during the six months period between the two polls. To start analysing this you can construct a confidence interval for the difference in proportions $p_1 - p_2$. If the prorportion has decreased, i.e if $p_1 > p_2$ then $p_1 - p_2 > 0$, if the proportion has stayed the same, i.e. $p_1 = p_2$ then $p_1 - p_2 = 0$ and if the proportion has increased i.e. $p_1 < p_2$ then $p_1 - p_2 < 0$.

1.Estimate $p_1 - p_2$ by

$$(p_1)^*_{obs} - (p_2)^*_{obs} = p^*_{obs} - \hat{p}_{obs} = \frac{x}{n} - \frac{y}{m} = \frac{120}{3002} - \frac{141}{3002} \approx 0.003106$$

2.We model x as an outcome of $X \in Bin(n, p_1)$ and since

$$np^*_{obs}(1 - p^*_{obs}) = 3002 \cdot \frac{120}{3002}\left(1 - \frac{120}{3002}\right) \approx 115 > 10$$

a normal approximation is feasible, i.e.

$$X \in Bin(n, p_1) \sim N\left(np_1, \sqrt{np_1(1-p_1)}\right)$$

(see the compiled formulae Section 6). In the same way we have that

$$Y \in Bin(m, p_2) \sim N\left(mp_2, \sqrt{mp_2(1-p_2)}\right),$$

since

$$m\hat{p}_{obs}(1 - \hat{p}_{obs}) = 3273 \cdot \frac{141}{3273}\left(1 - \frac{141}{3273}\right) \approx 135 > 10.$$

Finally we obtain that

$$(p_1 - p_2)^* = \frac{X}{n} - \frac{Y}{m} \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}\right)$$

since a linear combination of independent approximately normally distributed random variables.

3. A confidence interval for $p_1 - p_2$ with approximate confidence level 95% is given by (use 12.3 Approximate method in the compiled formulae)

$$\begin{aligned} I_{p_1-p_2} &= \left(p^*_{obs} - \hat{p}_{obs} \pm \lambda_{\alpha/2}\sqrt{\frac{p^*_{obs}(1-p^*_{obs})}{n} + \frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{m}}\right) \\ &= (-0.0031 \pm 0.01) = (-0.0131, 0.0069) \qquad (\approx 95\%) \end{aligned}$$

In this case you would not be able to claim that there has been a change in the number of supporters since 0 lies within the confidence interval.

□

1.3 Paired samples

Assume that we have given
observations $x_1, \ldots, x_n$ of independent random variables $X_1, \ldots, X_n$ where $X_i \in N(\mu_i, \sigma_1)$, $i = 1, \ldots, n$ and
observations $y_1, \ldots, y_n$ of independent random variables $Y_1, \ldots, Y_n$ where $Y_i \in N(\mu_i + \Delta, \sigma_2)$, $i = 1, \ldots, n$.
Furthermore assume that the X_i's and the Y_i's are independent. In this situation we say that we have **paired samples**.

Example 3 Suppose that you would like to evaluate the effect a certain type of medicine has on blood pressure. You therefore measure the blood pressure of a number of patients before and after treatment with the medicine. If the observations before treatment are denoted by $x_1, \ldots, x_n$, and the observations after treatment by $y_1, \ldots, y_n$, then Δ would capture the effect of the treatment with the drug and $\Delta > 0$ would mean that the medicine raises the bloodpressure (on average), $\Delta = 0$ would mean that the medicine (on average) has no effect on the blood pressure, and $\Delta < 0$ would mean that the medicine (on average) lowers the blood pressure. □

One way to examine if there is difference in expectation or on average between the two groups of variables is to construct a confidence interval for the parameter Δ.
Trick: Form the differences $z_i = y_i - x_i$. Then $z_1, \ldots, z_n$ are observations of independent random variables $Z_1, \ldots, Z_n$ such that $Z_i \in N(\Delta, \sigma)$, $i = 1, \ldots, n$ (it holds that $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$, but this does not help since they are unknown and can not be estimated in an easy way). They problem has now been reduced to constructing a confidence interval for the expectation given observations from a normal distribution with unknown variance.

Remark 1 When estimating σ one should use $\sigma^*_{obs} = s$, where

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

This is because estimating σ_1 (or σ_2) is not possible since the X_i's have different and unknown expectations.

Remark 2 It is not necessary to assume that the X_i's and Y_i's are normally distributed, but the differences between them, $Y_i - X_i$, have to be.

Example 4 Do problem 1313 from [2]. $\square$

Referenser

[1] Blom, G., Enger, J., Englund, G., Grandell, J., och Holst, L., (2005). Sannolikhetsteori och statistikteori med tillämpningar.

[2] Blom, Gunnar, (1989). Probability and Statistics. Theory and Applications.