Mathematical statistics
Week 5/a: Interval estimation

Péter Elek and Ádám Reiff

October 15, 2013

1. Concept of confidence intervals

Confidence intervals.

- Estimators so far: point estimates (i.e. give a single value as an estimate)
  - no information about how close these point estimates can be to the true parameter value
  - (although variance of estimators is related to this, that is only "indirect" information about goodness)
- Interval estimation: the estimate is an interval (not a single number or point)

Example.

- The weights of 16 boxes of cereal (in grams) are

$$506, 508, 499, 503, 504, 510, 497, 512$$
$$514, 505, 493, 496, 506, 502, 509, 496$$

- Assume that the weight is a normally distributed random variable with unknown mean (μ) and $\sigma = 5$ grams. Find the 90% confidence interval of μ (i.e. of the population expected value).

Solution.

- We know that $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, or $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Therefore

$$\Pr\left(z_{0.05} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{0.95}\right) = 0.9.$$

- This expression is for the random variable $\overline{X}$ (or more precisely, for $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$).

- Equivalent formulation (since $z_{0.05} = -z_{0.95}$):

$$\Pr\left(\overline{X} - z_{0.95}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{0.95}\frac{\sigma}{\sqrt{n}}\right) = 0.9.$$

- That is, a random interval $\left[\overline{X} - z_{0.95}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{0.95}\frac{\sigma}{\sqrt{n}}\right]$ contains parameter μ with 90% probability.

Solution (cont.).

- In our example $\overline{x} = 503.75$, $\sigma = 5$, $n = 16$, $z_{0.05} = -1.645$, $z_{0.95} = 1.645$,
- so $\overline{x} - z_{0.95}\frac{\sigma}{\sqrt{n}} = 501.694$ and $\overline{x} + z_{0.95}\frac{\sigma}{\sqrt{n}} = 505.806$.
- Thus we can be 90% "confident" that the interval $[501.694; 505.806]$ contains the true mean weight.

Remarks.

- In classical statistics we tend not to say that the probability that μ is in $[501.694; 505.806]$ is 90%, as μ is a number, and $[501.694; 505.806]$ are also numbers, so there is no random variable for which the probability could be interpreted.
  - In Bayesian statistics a more proper interpretation of confidence intervals can be given.
- Precise interpretation: *before* drawing the random sample $\{X_i\}$, there is a 90% probability that the random interval contains μ.
- We based the argument on $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Any other distribution can be used to construct similar arguments.

2. Estimation of population mean

2.1. Large-sample estimation of population mean

Confidence interval for μ in large samples.

- For *any* i.i.d. sample: $\overline{X} \overset{A}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$, or $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{A}{\sim} N(0,1)$. Also, $\frac{\overline{X}-\mu}{s^*/\sqrt{n}} \overset{A}{\sim} N(0,1)$.
- Similarly to previous calculations:

$$\Pr\left(z_{\alpha/2} < \frac{\overline{X} - \mu}{s^*/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

- or equivalently (using $z_{\alpha/2} = -z_{1-\alpha/2}$),

$$\Pr\left(\overline{X} - z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}}\right) = 1 - \alpha.$$

- Hence the $1-\alpha$ percent confidence interval of μ is $\left[\overline{x} - z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}}; \overline{x} + z_{1-\alpha/2} * \frac{s^*}{\sqrt{n}}\right]$.
- Rule of thumb for large sample approximation: $n > 30$ (although this is obviously not a theorem)

Example.

- X is the random variable of unoccupied seats on flights. To determine a tolerable level of "overbooking", airlines want to estimate the mean of X.
- Therefore they select $n = 225$ flights randomly, record $x_1, x_2, \ldots, x_{225}$, and find $\overline{x} = 11.6$ seats, $s^* = 4.1$ seats.
- Find the 90%, 95% and 99% confidence interval of $\mu = E(X)$.

Solution.

- The $1-\alpha$ percent confidence interval of μ is $\left[\overline{x} - z_{1-\alpha/2}\frac{s^*}{\sqrt{n}}; \overline{x} + z_{1-\alpha/2}\frac{s^*}{\sqrt{n}}\right]$, with $\overline{x} = 11.6$ and $s^* = 4.1$.
  - For the 90% confidence interval: $\alpha = 0.1$, then $z_{\alpha/2} = z_{0.05} = -1.645$, $z_{1-\alpha/2} = z_{0.95} = 1.645$, so the confidence interval is $11.6 \pm 1.645 * \frac{4.1}{\sqrt{225}} = 11.6 \pm 0.4496$.
  - For the 95% confidence interval: $\alpha = 0.05$, then $z_{\alpha/2} = z_{0.025} = -1.96$, $z_{1-\alpha/2} = z_{0.975} = 1.96$, so the confidence interval is $11.6 \pm 1.96 * \frac{4.1}{\sqrt{225}} = 11.6 \pm 0.5357$.
  - For the 99% confidence interval: $\alpha = 0.01$, then $z_{\alpha/2} = z_{0.005} = -2.575$, $z_{1-\alpha/2} = z_{0.995} = 2.575$, so the confidence interval is $11.6 \pm 2.575 * \frac{4.1}{\sqrt{225}} = 11.6 \pm 0.7038$.

2.2. Small-sample estimation of population mean

Confidence interval in normal samples: known σ.

- Assume $X \sim N\left(\mu, \sigma^2\right)$.
- Then, if σ is known, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

– (If X is not normally distributed, the distribution of $\overline{X}$ is undetermined.)

- Hence

$$\begin{aligned} \Pr\left(z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) &= 1-\alpha \\ \Pr\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha. \end{aligned}$$

- So the $1-\alpha$ percent confidence interval of μ is $\left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$.

Confidence interval in normal samples: unknown σ.

- Assume $X \sim N\left(\mu, \sigma^2\right)$.
- Then, if σ is unknown, $\frac{\overline{X}-\mu}{s^*/\sqrt{n}} \sim t_{n-1}$.
- Hence

$$\begin{aligned} \Pr\left(t_{\alpha/2;n-1} < \frac{\overline{X}-\mu}{s^*/\sqrt{n}} < t_{1-\alpha/2;n-1}\right) &= 1-\alpha \\ \Pr\left(\overline{X} - t_{1-\alpha/2;n-1}\frac{s^*}{\sqrt{n}} < \mu < \overline{X} + t_{1-\alpha/2;n-1}\frac{s^*}{\sqrt{n}}\right) &= 1-\alpha. \end{aligned}$$

- So the $1-\alpha$ percent confidence interval of μ is $\left[\overline{x} - t_{1-\alpha/2;n-1} * \frac{s^*}{\sqrt{n}}; \overline{x} + t_{1-\alpha/2;n-1} * \frac{s^*}{\sqrt{n}}\right]$.

Example.

- Estimate the average effect of a new drug on blood pressure. We have 6 patients, and their blood pressure decreased by $1.7, 3, 0.8, 3.4, 2.7, 2.1$. Based on previous studies we may assume that the change in blood pressure is approximately normal.
- What is the 95% confidence interval of the average effect of the drug on blood pressure? What would it be if we *knew* that $\sigma = 0.95$?

Solution.

- We have $\overline{x} = \frac{13.7}{6} = 2.28$ and $s^{*2} = 0.90$, $s^* = 0.95$. Then the 95% confidence interval is

$$\left[\overline{x} - t_{1-\alpha/2;n-1} * \frac{s^*}{\sqrt{n}}; \overline{x} + t_{1-\alpha/2;n-1} * \frac{s^*}{\sqrt{n}}\right],$$

- where $\alpha = 0.05$, $t_{0.975;5} = 2.571$ and $n = 6$. So the interval is $2.28 \pm 2.571 * \frac{0.95}{\sqrt{6}} = 2.28 \pm 1.00$.
- If we *know* that $\sigma = 0.95$, then the 95% confidence interval is

$$\left[\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

- where $z_{0.975} = 1.96$, $\sigma = 0.95$. So the interval is $2.28 \pm 1.96 * \frac{0.95}{\sqrt{6}} = 2.28 \pm 0.76$ (much narrower than in case of unknown σ).

3. Large-sample estimation of population proportion

Confidence interval for p in large samples.

- Let $X \sim Bernoulli(p) : \Pr(X = 1) = p, \Pr(X = 0) = 1 - p$.
- In large samples $\frac{\overline{X}-\mu}{s^*/\sqrt{n}} \overset{A}{\sim} N(0,1)$ and also $\frac{\overline{X}-\mu}{s/\sqrt{n}} \overset{A}{\sim} N(0,1)$.
- Then as $s^2 = \overline{X}(1-\overline{X})$, we have $\frac{\overline{X}-p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \overset{A}{\sim} N(0,1)$.
- Therefore

$$\Pr\left(z_{\alpha/2} < \frac{\overline{X} - p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} < z_{1-\alpha/2}\right) = 1 - \alpha.$$

- Thus the (approximate) $1-\alpha$ percent confidence interval is $\left[\overline{x} - z_{1-\alpha/2} * \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}; \overline{x} + z_{1-\alpha/2} * \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right]$.
- Rule of thumb for large sample approximation: $np(1-p) > 10$ (although there are still "unlucky" parameter values above that)

Example.

- The Labor Force Survey (LFS) is designed to determine the employment figures of a particular country. In the first quarter of 2013, based on surveying 54000 people, the employment rate of the working-age population (i.e. the population between 15-74 years) was reported as 51.5% by the Central Statistical Office in Hungary.
- Find the 95% confidence interval of the employment rate in Hungary.
- We know that the Hungarian working-age population is around 7.630 million people. Find the 95% confidence interval for the number of employed people in the second quarter of 2013.

Solution.

- Here $n = 54000$, $\alpha = 0.95$, $z_{1-\alpha/2} = 1.96$, hence the 95% confidence interval for the employment rate is $0.515 \pm 1.96 * \sqrt{\frac{0.515*(1-0.515)}{54000}} = 0.515 \pm 0.0042 = (0.511; 0.519)$.
- To determine aggregate employment, we have to estimate $N * p$, where $N = 7630000$ is the (fixed) size of the working-age population. The point estimate is $N * \widehat{p} = 7630000 * 0.515 = 3929450$ (or 3.929 million). Its confidence interval is $7630000 * (0.511 \pm 0.0042) = 3929450 \pm 32000$, or between 3.898 million and 3.962 million people.
- (The actual confidence interval is slightly different because of the features of the stratified sampling scheme used in the Labor Force Survey.)

4. Estimation of population variance

Confidence interval for σ^2 in normal samples.

- For normally distributed populations $\frac{(n-1)s^{*2}}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$.
- Then

$$\begin{aligned} \Pr\left(k_{\alpha/2;n-1} < \frac{ns^2}{\sigma^2} < k_{1-\alpha/2;n-1}\right) &= 1-\alpha \\ \Pr\left(\frac{ns^2}{k_{1-\alpha/2;n-1}} < \sigma^2 < \frac{ns^2}{k_{\alpha/2;n-1}}\right) &= 1-\alpha. \end{aligned}$$

- So the $1-\alpha$ percent confidence interval of σ^2 is $\left[\frac{ns^2}{k_{1-\alpha/2;n-1}}; \frac{ns^2}{k_{\alpha/2;n-1}}\right]$.

Example.

- Assume that the height of CEU female students is normally distributed. Find the 95% confidence interval of the true variance of the height distribution if in a sample of $n = 100$ we find that $s^2 = 169$.

Solution 1.

- The 95% confidence interval is $\left[\frac{ns^2}{k_{0.975;n-1}}; \frac{ns^2}{k_{0.025;n-1}}\right]$,
- where $n = 100$, $s^2 = 169$, $k_{0.025;99} = 73.36$ and $k_{0.975;99} = 128.42$ (by using a quantile function of a software package).
- So the 95% confidence interval is $\left[\frac{100*169}{128.42}; \frac{100*169}{73.36}\right] = [131.60; 230.37]$.

Solution 2.

- Since $n = 100$ is sufficiently large, we can also use the result that in case of a normal sample $\frac{s^2-\sigma^2}{\sqrt{2/n*\sigma^2}} \overset{A}{\sim} N(0,1)$.
- Hence $\Pr\left(z_{0.025} < \frac{s^2-\sigma^2}{\sqrt{2/n}\sigma^2} < z_{0.975}\right) \approx 0.95$,
- so $\Pr\left(\frac{s^2}{1+z_{0.975}\sqrt{2/n}} < \sigma^2 < \frac{s^2}{1+z_{0.025}\sqrt{2/n}}\right) \approx 0.95$.
- So the (approximate) 95% confidence interval of σ^2 is $\left[\frac{s^2}{1+z_{0.975}\sqrt{2/n}}; \frac{s^2}{1+z_{0.025}\sqrt{2/n}}\right] = [132.32; 233.81]$.

Material

Material.

- W Appendix C.5
- CB 9.1, 10.4.2 (except for Examples 10.4.8-10.4.9)
  - only to the extent covered in the course