James Green-Armytage
Econ 229, fall 2015

Confidence Intervals: Part 2

Here we are interested in constructing a confidence interval, i.e. a range where we believe the value we are trying to estimate is located with probability k. We will suppose that the sample size is sufficiently large, given the underlying distribution(s), that we can invoke the central limit theorem and approximate the distribution of our point estimate with a corresponding normal distribution.

In this case, each value of k implies a value of z^*, according to $\boxed{z^* = \Phi^{-1}\left(\frac{k+1}{2}\right)}$, where $\Phi^{-1}(F)$ is the inverse function of the standard normal CDF. We can find the desired z^* value using a normal probability table, or find it in Excel with the code `=normsinv( (k+1)/2 )`. Once we know the mean and standard deviation of our point estimate, we will be able to make use of this z^* value to construct a confidence interval.

1. Confidence interval for a mean

Suppose that we are investigating a distribution of some random variable with mean μ, variance σ^2, and standard deviation σ. We want to construct a confidence interval within which we believe μ should be located with probability k.

Let $\boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$ be the sample mean, i.e. the simple average of n randomly selected draws from the distribution, x_i.

The expected value of the sample mean is $\boxed{\mu_{\bar{x}} = \mathrm{E}(\bar{x}) = \mu}$. Therefore it serves as our point estimate for μ. The variance and standard deviation of the sample mean are $\boxed{\sigma_{\bar{x}}^2 = \mathrm{var}(\bar{x}) = \frac{\sigma^2}{n}}$ and $\boxed{\sigma_{\bar{x}} = \mathrm{sd}(\bar{x}) = \frac{\sigma}{\sqrt{n}}}$, respectively. (We derived these in the previous installment.)

Therefore, we believe that $\bar{x}$ is distributed approximately according to a normal distribution with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Therefore, with probability $\approx k$, $\bar{x}$ has fallen within the interval $\mu \pm z^* \frac{\sigma}{\sqrt{n}}$.

Therefore, with probability $\approx k$, μ is located within the interval $\boxed{\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}}$.

We don't directly observe σ, the true standard deviation of x, but we can estimate it with s, our sample standard deviation. Using Bessel's correction, the formula for the sample standard deviation is $\boxed{s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$.

So, we estimate that with probability $\approx k$, μ is within the interval $\boxed{\bar{x} \pm z^* \frac{s}{\sqrt{n}}}$.

2. Confidence interval for a difference between two means

Now suppose that we are investigating two distributions, A and B. Define μ_A, σ_A^2, σ_A, μ_B, σ_B^2, and σ_B as the means, variances, and standard deviations of distributions A and B, respectively. We are interested in the difference between the means of the two distributions, $\mu_A - \mu_B$, and we want to construct a confidence interval within which we believe $\mu_A - \mu_B$ should be located with probability k.

Let $\boxed{\bar{x}_A - \bar{x}_B = \left(\frac{1}{n_A}\sum_{i=1}^{n_A} x_{Ai}\right) - \left(\frac{1}{n_B}\sum_{j=1}^{n_B} x_{Bi}\right)}$ be the difference between our sample means for distributions A and B. The expected value of $\bar{x}_A - \bar{x}_B$ is $\boxed{\mu_{\bar{x}_A - \bar{x}_B} = \mathrm{E}(\bar{x}_A - \bar{x}_B) = \mu_A - \mu_B}$; this can be derived quickly: $\mathrm{E}(\bar{x}_A - \bar{x}_B) = \mathrm{E}(\bar{x}_A) - \mathrm{E}(\bar{x}_B) = \mu_A - \mu_B$.

The variance and standard deviation of $\bar{x}_A - \bar{x}_B$ are $\boxed{\sigma_{\bar{x}_A - \bar{x}_B}^2 = \mathrm{var}(\bar{x}_A - \bar{x}_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$, and $\boxed{\sigma_{\bar{x}_A - \bar{x}_B} = \mathrm{sd}(\bar{x}_A - \bar{x}_B) = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$, respectively. We derive this as follows: $\mathrm{var}(\bar{x}_A - \bar{x}_B) = \mathrm{var}\big(\bar{x}_A + (-\bar{x}_B)\big) = \mathrm{var}(\bar{x}_A) + \mathrm{var}(\bar{x}_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$.

Therefore, we believe that $\bar{x}_A - \bar{x}_B$ is distributed approximately according to a normal distribution with mean $\mu_A - \mu_B$ and standard deviation $\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$.

Therefore, with probability $\approx k$, $\bar{x}_A - \bar{x}_B$ has fallen in the interval $\mu_A - \mu_B \pm z^* \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$.

Therefore, with probability $\approx k$, $\mu_A - \mu_B$ is located in the interval $\boxed{(\bar{x}_A - \bar{x}_B) \pm z^* \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$.

We can estimate σ_A^2 and σ_B^2 with $s_A^2 = \frac{1}{n_A - 1}\sum_{i=1}^{n_A}(x_{Ai} - \bar{x}_A)^2$ and $s_B^2 = \frac{1}{n_B - 1}\sum_{j=1}^{n_B}(x_{Bi} - \bar{x}_B)^2$.

So, we estimate that with probability $\approx k$, $\mu_A - \mu_B$ is in the interval $\boxed{(\bar{x}_A - \bar{x}_B) \pm z^* \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$.

3. Confidence interval for a proportion

Suppose that we are investigating a distribution of a Bernoulli random variable, i.e. a random variable x such that $x = 1$ with probability p, and $x = 0$ with probability $1 - p$. In this case we can write the mean, variance, and standard deviation of x as $\boxed{\mu = p}$, $\boxed{\sigma^2 = p(1-p)}$, and $\boxed{\sigma = \sqrt{p(1-p)}}$, respectively. These formulae can be derived as follows:

$$\mu = \mathrm{E}(x) = (1 \cdot p) + \big(0 \cdot (1-p)\big) = p$$

$$\begin{aligned}\sigma^2 = \mathrm{var}(x) = \mathrm{E}((x-\mu)^2) &= (1-p)^2 p + (0-p)^2(1-p) = p(1-p)^2 + p^2(1-p) \\ &= (1 - 2p + p^2)p + p^2(1-p) = p - 2p^2 + p^3 + p^2 - p^3 = p - p^2 = p(1-p)\end{aligned}$$

Suppose that we take n draws from the distribution, and observe π successes (where $x_i = 1$) and $n - \pi$ failures (where $x_i = 0$). In this case, our sample mean $\bar{x}$ is equivalent to the number of successes divided by the number of trials. We call this $\hat{p}$, the sample proportion. That is, $\boxed{\hat{p} = \bar{x} = \frac{\pi}{n}}$.

The expected value of $\hat{p}$ is $\boxed{\mu_{\hat{p}} = E(\hat{p}) = p}$; therefore it is our point estimate for $\mu = p$.

The variance and standard deviation of the sample proportion are $\boxed{\sigma_{\hat{p}}^2 = \operatorname{var}(\hat{p}) = \frac{p(1-p)}{n}}$ and $\boxed{\sigma_{\hat{p}} = \operatorname{sd}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}}$, respectively. The variance can be derived by combining the identities $\sigma^2 = p(1-p)$ and $\operatorname{var}(\bar{x}) = \frac{\sigma^2}{n}$ as follows: $\operatorname{var}(\hat{p}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$.

Therefore, we believe that $\hat{p}$ is distributed approximately according to a normal distribution with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Therefore, with probability $\approx k$, $\hat{p}$ has fallen in the interval $p \pm z^* \sqrt{\frac{p(1-p)}{n}}$.

Therefore, with probability $\approx k$, p is located in the interval $\boxed{\hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}}}$.

We can estimate $\sigma = \sqrt{p(1-p)}$ with $\boxed{s = \sqrt{\frac{n}{n-1}\hat{p}(1-\hat{p})}}$. We can derive this via $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - p)^2 = \frac{1}{n-1}[\pi(1-\hat{p})^2 + (n-\pi)(0-\hat{p})^2] = \frac{1}{n-1}[n\hat{p}(1-\hat{p})^2 + n(1-\hat{p})\hat{p}^2] = \frac{n}{n-1}[\hat{p}(1-\hat{p})^2 + (1-\hat{p})\hat{p}^2] = \cdots = \frac{n}{n-1}\hat{p}(1-\hat{p})$, with the steps represented by the ellipsis following logic similar to the derivation of $\sigma^2 = p(1-p)$ above.

So, we estimate that with probability $\approx k$, p is in the interval $\boxed{\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}}$.

Alternatively, we could take a more cautious approach by assuming the highest possible value of σ, which is $\sigma_{\max} = 1/2$. In this case, we estimate that with probability $\approx k$ (or slightly greater), p is in the interval $\boxed{\hat{p} \pm z^* \frac{1}{2\sqrt{n}}}$.