MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

Confidence Intervals

7.1 Introduction – $N(\mu, \sigma^2)$ – Confidence Interval for μ when σ^2 is known

Consider the simple case where we have a random sample of size n from a Normal $N(\mu, \sigma^2)$ distribution, where the population mean μ is an unknown parameter which we wish to estimate and (unrealistically) the population variance σ^2 is known (say $\sigma^2 = \sigma_0^2$).

The natural estimator of μ is $\bar{X}$ – it is the maximum likelihood estimator, the method of moments estimator and the estimator given by the sample mean. However, even for this 'best' estimator, different samples would give different estimates, so we know our estimate cannot be 'exactly' correct.

In such cases, it may be more informative to report the value of the estimate, together with some measure of the accuracy or the margin of error of the estimate. This leads to procedures which report their results in the form of an *interval* of values we have some *confidence* contains the true value of the unknown parameter.

95% Confidence Interval

Say we wanted to know what margin of error to report to be 95% confident that the true value of μ was within the declared margin of error of the estimate $\bar{x}$. One way of proceeding is as follows:

We know $\quad \bar{X} \sim N(\mu, \sigma_0^2/n)$ and $\dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \sim N(0,1)$ from 6.4

so $\quad P\left(-1.96 \le \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \le 1.96\right) = 0.95$ since $z_{0.025} = 1.96$.

But $\quad \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \le 1.96 \Longleftrightarrow \bar{X}-\mu \le \dfrac{1.96\,\sigma_0}{\sqrt{n}} \Longleftrightarrow \bar{X} - \dfrac{1.96\,\sigma_0}{\sqrt{n}} \le \mu$

and $\quad -1.96 \le \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \Longleftrightarrow -\dfrac{1.96\,\sigma_0}{\sqrt{n}} \le \bar{X}-\mu \Longleftrightarrow \mu \le \bar{X} + \dfrac{1.96\,\sigma_0}{\sqrt{n}}$

so the event $\left\{-1.96 \le \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \le 1.96\right\}$ occurs if and only if

the event $\left\{\bar{X} - \dfrac{1.96\,\sigma_0}{\sqrt{n}} \le \mu \le \bar{X} + \dfrac{1.96\,\sigma_0}{\sqrt{n}}\right\}$ also occurs.

Thus if we take a large number of simple random samples from the $N(\mu, \sigma_0^2)$ distribution, each of fixed size n, then in 95% of the samples the interval $\bar{X} \pm 1.96\sigma_0/\sqrt{n}$ will contain the true parameter value μ (and in 5% it will not). Thus we report a *95% confidence interval* with Lower end points c_L and Upper end point c_U given by

$$c_L = \bar{X} - \frac{1.96\,\sigma_0}{\sqrt{n}} \quad \text{and} \quad c_U = \bar{X} + \frac{1.96\,\sigma_0}{\sqrt{n}}$$

General $100(1-\alpha)\%$ confidence interval

More generally for a $100(1-\alpha)\%$ confidence interval :

we know $\quad P\left(-z_{\alpha/2} \leq \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$

but $\quad \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2} \Longleftrightarrow \bar{X}-\mu \leq \dfrac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}} \Longleftrightarrow \bar{X} - \dfrac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}} \leq \mu$

and $\quad -z_{\alpha/2} \leq \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \Longleftrightarrow -\dfrac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}} \leq \bar{X}-\mu \Longleftrightarrow \mu \leq \bar{X} + \dfrac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}}$

so the event $\left\{-z_{\alpha/2} \leq \dfrac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right\}$ occurs if and only if

the event $\quad \left\{\bar{X} - \dfrac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + \dfrac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}}\right\}$ also occurs.

Thus if we take a large number of simple random samples from the $N(\mu, \sigma_0^2)$ distribution, each of fixed size n, then in $100(1-\alpha)\%$ of the samples the interval $\bar{X} \pm z_{\alpha/2}\sigma_0/\sqrt{n}$ will contain the true parameter value μ (and in $100\alpha\%$ it will not). We can therefore report a $100(1-\alpha)\%$ *confidence interval* with Lower end points c_L and Upper end point c_U given by

$$c_L = \bar{X} - \frac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}} \quad \text{and} \quad c_U = \bar{X} + \frac{z_{\alpha/2}\,\sigma_0}{\sqrt{n}}$$

Notes

- For each sample the interval either will or will not contain the true value μ. In 95% of cases (or more generally in $100(1-\alpha)\%$ of cases) it will and in the remainder it will not – but it is impossible to tell for each sample whether the interval does or does not contain μ.

- The length of the confidence interval is $2 \times z_{\alpha/2}\,\sigma_0/\sqrt{n}$. Here it is clear that:
  – the length of the interval DECREASES as the sample size n INCREASES
  – the length of the interval INCREASES as the population variance σ_0^2 INCREASES
  – the length of the interval INCREASES as the confidence level $100(1-\alpha)$ INCREASES (since this means α DECREASES and so $z_{\alpha/2}$ INCREASES).
  These relationships between the (average) length of the interval and the sample size, the population variance and the confidence level are intuitively reasonable and are hold more generally for other confidence intervals we will meet.

- The interval is of the form (c_L, c_U), where the end points

  $$c_L(X_1, \ldots, X_n) = \bar{X} - z_{\alpha/2}\,\sigma_0/\sqrt{n} \quad \text{and} \quad c_U(X_1, \ldots, X_n) = \bar{X} + z_{\alpha/2}\,\sigma_0/\sqrt{n}$$

  depend on the data as well as on the value of n and the value of the known parameter σ_0. Thus the end points of the confidence interval, and (usually) the length of the confidence interval, are themselves random variables whose values will vary from sample to sample.

- The confidence statement defining the interval has the form

  $$P\{c_L \leq \mu \leq c_U\} = 1-\alpha.$$

  It is important to understand that in this statement, μ is fixed; it is c_L and c_U that depend on the data, so vary from sample to sample. The confidence statement is an assertion about the joint distribution of c_L and c_U.

7.9 Confidence interval for θ – Exp(θ) population

Say we have a simple random sample $X_1, \ldots, X_n$ from a population with Exp(θ) distribution, and we want to construct a $100(1-\alpha)\%$ confidence interval for the single unknown parameter θ.

The standard estimate (mom and mle) for θ is $\hat{\theta} = n / \sum_{j=1}^{n} X_j$,
where (from § 6.13) $\sum_{j=1}^{n} X_j \sim$ Gamma (n, θ) and hence $2\theta \sum_{j=1}^{n} X_j \sim \chi^2_{2n}$.

To construct an 'equal tailed' confidence interval we start by noting that

$$P\left(\chi^2_{2n\,;\,1-\alpha/2} \leq 2\theta \sum_{j=1}^{n} X_j \leq \chi^2_{2n\,;\,\alpha/2}\right) = 1-\alpha$$

so that

$$P\left(\frac{\chi^2_{2n\,;\,1-\alpha/2}}{2\sum_{j=1}^{n} X_j} \leq \theta \leq \frac{\chi^2_{2n\,;\,\alpha/2}}{2\sum_{j=1}^{n} X_j}\right) = 1-\alpha.$$

Thus, if we take

$$\boxed{c_L = \chi^2_{2n\,;\,1-\alpha/2}/(2\textstyle\sum_{j=1}^{n} X_j) \quad \text{and} \quad c_U = \chi^2_{2n\,;\,\alpha/2}/(2\textstyle\sum_{j=1}^{n} X_j)}$$

then (c_L, c_U) is a $100(1-\alpha)$% confidence interval for θ.

7.10 Confidence intervals by simulation in R

Given a simple random sample $X_1, \ldots, X_n$ of size n from a distribution in a parametric family with a single unknown parameter θ, and we can construct an approximate $100(1-\alpha)\%$ confidence interval by simulation as follows:

1. Calculate an estimate $\hat{\theta}$ for θ.
2. Simulate B simple random samples, each of the same size n as the original sample, from the distribution in the parametric family corresponding to the value $\hat{\theta}$.
3. Calculate the B estimates, $\theta^*_1, \ldots, \theta^*_B$, one from each simulated sample, using the same estimation method as in step 1 above.
4. Calculate the correponding B values of $\theta^*_k - \hat{\theta}$, $k = 1, \ldots, B$. If $\hat{\theta}$ is close to θ, then the distribution of the values of $\theta^* - \hat{\theta}$ for samples from the distribution with parameter $\hat{\theta}$ will be close to the distribution of $\hat{\theta} - \theta$ for samples from the distribution with parameter θ.
5. Identify values k_L and k_U such that $100\alpha/2$ of the B values of $\theta^*_k - \hat{\theta}$ are $< k_L$ and $100\alpha/2$ of the B values of $\theta^*_k - \hat{\theta}$ are $> k_U$. Then from step 4 above we have that
$$P(k_L \leq \hat{\theta} - \theta \leq k_U) \simeq P(k_L \leq \theta^* - \hat{\theta} \leq k_U) \simeq 1-\alpha.$$
6. The event $\{k_L \leq \hat{\theta} - \theta \leq k_U\}$ is equivalent to the event $\{\hat{\theta} - k_U \leq \theta \leq \hat{\theta} - k_L\}$, so for B large the interval (c_L, c_U) is an approximate $100(1-\alpha)\%$ confidence interval for θ, where
$$c_L = \hat{\theta} - k_U \quad \text{and} \quad c_U = \hat{\theta} - k_L.$$

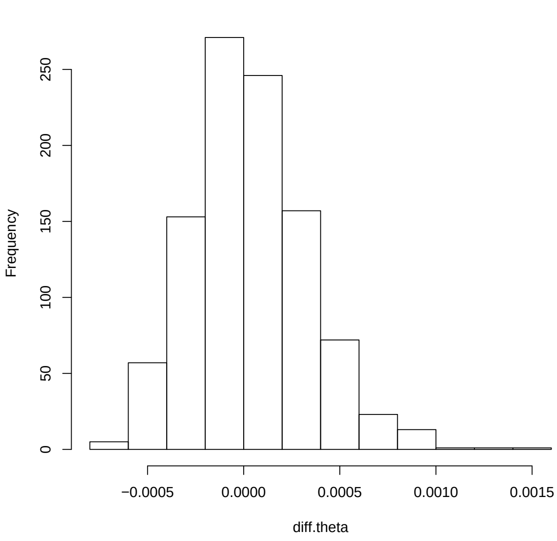
7.11 Example - Earthquakes - 90% Confidence Interval

Consider again the `quakes` data set (§1.6) with $n = 62$ observations. In line with the graphical plots in §1.6 and the assessment of fit in §2.11, we assume the data comes from a distribution in the Exp(θ) family. For this family, we found that the method of moments estimate for θ (§2.6) and the maximum likelihood estimate for θ (§3.6) both had the form $\hat{\theta} = 1/\bar{x}$.

The sequence of **R** commands below first calculates $\hat{\theta}$ for the `quakes` data; then generates $62,000$ observations from the Exponential distribution with this value of θ, and arranges the observations into a matrix of $B = 1000$ samples each with $n = 62$ observations. Next it calculates a vector of means for each of the $k = 1, \ldots, 1000$ samples; calculates the vector of estimates θ^*_k for the samples; calculates the vector of differences $\theta^*_k - \hat{\theta}$; sorts these differences in order of increasing value and puts the sorted values in a vector `sort.diff`. Finally, we want a 90% confidence interval, so $100(1 - \alpha) = 90$ giving $\alpha = 0.1$ and $\alpha/2 = 0.05$. Thus the last three commands output the 50th and the 950th of the 1000 ordered values of $\theta^*_k - \hat{\theta}$ (i.e. the 5th and the 95th quantiles of the ordered differences); calculate $c_L = \hat{\theta} - k_U$; and calculate $c_U = \hat{\theta} - k_L$. Other intervals can be calculated similarly - e.g. for a 95% confidence interval you would need the 2.5th and the 97.5th quantiles, which here would roughly correspond to the 25th and the 975th values in the set of ordered differences.

```
> theta.hat <- 1/mean(quakes)
> xsamples <- matrix(rexp(62000,theta.hat), nrow=1000)
> xmean <- apply(xsamples,1,mean)
> theta.star <- 1/xmean
> diff.theta <- (theta.star - theta.hat)
> sort.diff <- sort(diff.theta)
> sort.diff[c(50,950)]
> cl <- theta.hat - sort.diff[950]
> cu <- theta.hat - sort.diff[50]
```

A histogram of the 1000 differences for a particular simulation is given below. Recall that the estimate of θ here is $\hat{\theta} = 0.00229$. For this simulation the 5th and 95th quantiles were $k_L = -0.00039$ and $k_U = 0.00058$ respectively, so the 90% confidence interval calculated from the simulation had end points $c_L = \hat{\theta} - k_U = 0.00171$ and $c_U = \hat{\theta} - k_L = 0.00268$. This compares well with the exact 90% confidence interval, which has end points (§7.8) $c_L = \chi^2_{2n;1-\alpha/2}/2\sum_1^n x_i = 0.00183$ and $c_U = \chi^2_{2n;\alpha/2}/2\sum_1^n x_i = 0.00279$, calculated using **R**.


Histogram of diff.theta
Frequency
diff.theta