Lecture 5. Confidence Intervals

We now consider interval estimation for θ.

Definition 5.1

A $100\gamma\%$ $(0 < \gamma < 1)$ **confidence interval** (CI) for θ is a random interval $\big(A(\mathbf{X}), B(\mathbf{X})\big)$ such that $\mathbb{P}\big(A(\mathbf{X}) < \theta < B(\mathbf{X})\big) = \gamma$, no matter what the true value of θ may be.

Notice that it is the endpoints of the interval that are random quantities (not θ).

We can interpret this in terms of repeat sampling: if we calculate $\big(A(\mathbf{x}), B(\mathbf{x})\big)$ for a large number of samples $\mathbf{x}$, then approximately $100\gamma\%$ of them will cover the true value of θ.

IMPORTANT: having observed some data $\mathbf{x}$ and calculated a 95% interval $\big(A(\mathbf{x}), B(\mathbf{x})\big)$ we *cannot* say there is now a 95% probability that θ lies in this interval.

Example 5.2

Suppose $X_1, \ldots, X_n$ are iid $N(\theta, 1)$. Find a 95% confidence interval for θ.

- We know $\bar{X} \sim N(\theta, \frac{1}{n}\sigma^2)$, so that $\sqrt{n}(\bar{X} - \theta) \sim N(0,1)$, no matter what θ is.
- Let z_1, z_2 be such that $\Phi(z_2) - \Phi(z_1) = 0.95$, where Φ is the standard normal distribution function.
- We have $\mathbb{P}\left[z_1 < \sqrt{n}(\bar{X} - \theta) < z_2\right] = 0.95$, which can be rearranged to give

$$\mathbb{P}\left[\bar{X} - \frac{z_2}{\sqrt{n}} < \theta < \bar{X} - \frac{z_1}{\sqrt{n}}\right] = 0.95.$$

  so that

$$\left(\bar{X} - \frac{z_2}{\sqrt{n}}, \bar{X} - \frac{z_1}{\sqrt{n}}\right)$$

  is a 95% confidence interval for θ.
- There are many possible choices for z_1 and z_2. Since the $N(0,1)$ density is symmetric, the shortest such interval is obtained by $z_2 = z_{0.025} = -z_1$ (where recall that z_α is the upper $100\alpha\%$ point of $N(0,1)$).
- From tables, $z_{0.025} = 1.96$ so a 95% confidence interval is $\left(\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right)$. $\square$

The above example illustrates a common procedure for findings CIs.

1. Find a quantity $R(\mathbf{X}, \theta)$ such that the $\mathbb{P}_\theta$- distribution of $R(\mathbf{X}, \theta)$ does not depend on θ. This is called a *pivot*.

   In Example 5.2, $R(\mathbf{X}, \theta) = \sqrt{n}(\bar{X} - \theta)$.
2. Write down a probability statement of the form $\mathbb{P}_\theta \left(c_1 < R(\mathbf{X}, \theta) < c_2 \right) = \gamma$.
3. Rearrange the inequalities inside $\mathbb{P}(\ldots)$ to find the interval.

Notes:

- Usually c_1, c_2 are percentage points from a known standardised distribution, often equitailed so that use, say, 2.5% and 97.5% points for a 95% CI. Could use 0% and 95%, but interval would generally be wider.
- Can have confidence intervals for vector parameters
- If $\big(A(\mathbf{x}), B(\mathbf{x})\big)$ is a $100\gamma\%$ CI for θ, and $T(\theta)$ is a monotone increasing function of θ, then $\big(T(A(\mathbf{x})), T(B(\mathbf{x}))\big)$ is a $100\gamma\%$ CI for $T(\theta)$.

  If T is monotone decreasing, then $\big(T(B(\mathbf{x})), T(A(\mathbf{x}))\big)$ is a $100\gamma\%$ CI for $T(\theta)$.

Example 5.3

Suppose $X_1, \ldots, X_{50}$ are iid $N(0, \sigma^2)$. Find a 99% confidence interval for σ^2.

- Thus $X_i/\sigma \sim \mathrm{N}(0,1)$. So, from the Probability review, $\frac{1}{\sigma^2}\sum_{i=1}^n X_i^2 \sim \chi^2_{50}$.
- So $R(\mathbf{X}, \sigma^2) = \sum_{i=1}^n X_i^2/\sigma^2$ is a pivot.
- Recall that $\chi^2_n(\alpha)$ is the upper $100\alpha\%$ point of χ^2_n, i.e. $\mathbb{P}(\chi^2_n \leq \chi^2_n(\alpha)) = 1 - \alpha$.
- From χ^2-tables, we can find c_1, c_2 such that $F_{\chi^2_{50}}(c_2) - F_{\chi^2_{50}}(c_1) = 0.99$.
- An equi-tailed region is given by $c_1 = \chi^2_{50}(0.995) = 27.99$ and $c_2 = \chi^2_{50}(0.005) = 79.49$.
- In R,

```
qchisq(0.005,50) = 27.99075,  qchisq(0.995,50) = 79.48998
```

- Then $\mathbb{P}_{\sigma^2}\left(c_1 < \frac{\sum X_i^2}{\sigma^2} < c_2\right) = 0.99$, and so $\mathbb{P}_{\sigma^2}\left(\frac{\sum X_i^2}{c_2} < \sigma^2 < \frac{\sum X_i^2}{c_1}\right) = 0.99$ which gives a confidence interval $\left(\frac{\sum X_i^2}{79.49}, \frac{\sum X_i^2}{27.99}\right)$.
- Further, a 99% confidence interval for σ is then $\left(\sqrt{\frac{\sum X_i^2}{79.49}}, \sqrt{\frac{\sum X_i^2}{27.99}}\right)$. $\square$

Example 5.4

Suppose $X_1, \ldots, X_n$ are iid Bernoulli(p). Find an approximate confidence interval for p.

- The mle of p is $\hat{p} = \sum X_i / n$.
- By the Central Limit Theorem, $\hat{p}$ is approximately $N(p, p(1-p)/n)$ for large n.
- So $\sqrt{n}(\hat{p} - p)/\sqrt{p(1-p)}$ is approximately $N(0,1)$ for large n.
- So we have

$$\mathbb{P}\left(\hat{p} - z_{(1-\gamma)/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{(1-\gamma)/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx \gamma.$$

- But p is unknown, so we approximate it by $\hat{p}$, to get an approximate $100\gamma\%$ confidence interval for p when n is large:

$$\left(\hat{p} - z_{(1-\gamma)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{(1-\gamma)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

$\square$

NB. There are many possible approximate confidence intervals for a Bernoulli/Binomial parameter.

Example 5.5

Suppose an opinion poll says 20% are going to vote UKIP, based on a random sample of 1,000 people. What might the true proportion be?

- We assume we have an observation of $x = 200$ from a Binomial(n, p) distribution with $n = 1,000$.
- Then $\hat{p} = x/n = 0.2$ is an unbiased estimate, also the mle.
- Now $\text{var}\left(\frac{X}{n}\right) = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.2\times 0.8}{1000} = 0.00016$.
- So a 95% CI is $\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.20 \pm 1.96 \times 0.013 = (0.175, 0.225)$, or around 17% to 23%.
- Special case of common procedure for an unbiased estimator T: 95% CI $\approx T \pm 2\sqrt{\text{var}\, T} = T \pm 2\,\text{SE}$, where SE = 'standard error' $= \sqrt{\text{var}\, T}$
- NB: Since $p(1-p) \leq 1/4$ for all $0 \leq p \leq 1$, then a conservative 95% interval (i.e. might be a bit wide) is $\hat{p} \pm 1.96\sqrt{\frac{1}{4n}} \approx \hat{p} \pm \sqrt{\frac{1}{n}}$.
- So whatever proportion is reported, it will be 'accurate' to $\pm 1/\sqrt{n}$.
- Opinion polls almost invariably use $n = 1000$, so they are assured of $\pm 3\%$ 'accuracy'

(Slightly contrived) confidence interval problem*

Example 5.6

Suppose X_1 and X_2 are iid from Uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. What is a sensible 50% CI for θ?

- Consider the probability of getting one observation each side of θ,

$$\begin{aligned}\mathbb{P}_\theta\left(\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)\right) &= \mathbb{P}_\theta(X_1 \leq \theta \leq X_2) + \mathbb{P}_\theta(X_2 \leq \theta \leq X_1) \\ &= \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = \frac{1}{2}.\end{aligned}$$

  So $(\min(X_1, X_2), \max(X_1, X_2))$ is a 50% CI for θ.

- But suppose $\mid X_1 - X_2 \mid \geq \frac{1}{2}$, e.g. $x_1 = 0.2, x_2 = 0.9$. Then we *know* that, in this particular case, θ *must* lie in $(\min(X_1, X_2), \max(X_1, X_2))$.
- So guaranteed sampling properties does not necessarily mean a sensible conclusion in all cases.