4 Confidence intervals

Statisticians do it with 95% confidence.

4.1 Interval estimation

Let $a(X)$ and $b(X)$ be two statistics satisfying $a(X) \le b(X)$ for all X. Suppose that on seeing the data $X = x$ we make the inference $a(x) \le \theta \le b(x)$. Here $[a(x), b(x)]$ is called an **interval estimate** and $[a(X), b(X)]$ is called an **interval estimator**.

Previous lectures were concerned with making a **point estimate** for θ. Now we are being less precise. By giving up precision in our assertion about the value of θ we gain confidence that our assertion is correct. Suppose

$$P_\theta\Big(a(X) \le \theta \le b(X)\Big) = \gamma,$$

where γ does not depend on θ. Then the random interval $\big[a(X), b(X)\big]$ is called a $100\gamma\%$ **confidence interval** for θ. Typically γ is 0.95 or 0.99, so that the probability the interval contains θ is close to 1.

Given data x, we would call $[a(x), b(x)]$ a '$100\gamma\%$ confidence interval for θ'. Notice however, that θ is fixed, and therefore the interval either does or does not contain the true θ. However, if we repeat the procedure of sampling and constructing a a confidence interval many times, then our confidence interval will contain the true θ $100\gamma\%$ of the time. The point to understand here is that *it is the endpoints of the confidence interval that are random variables, not the parameter* θ.

Examples 4.1

(a) If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ independently, with μ unknown and σ^2 known, then

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \text{and hence} \quad \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)\,.$$

So if ξ and η are such that $\mathbb{P}(\xi \le N(0,1) \le \eta) = \gamma$, we have

$$P_{(\mu,\sigma^2)}\left(\xi \le \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \le \eta\right) = \gamma\,,$$

which can be rewritten as

$$P_{(\mu,\sigma^2)}\left(\bar{X} - \frac{\eta\sigma}{\sqrt{n}} \le \mu \le \bar{X} - \frac{\xi\sigma}{\sqrt{n}}\right)\,.$$

Note that the choice of ξ and η is not unique. However, it is natural to try to make the length of the confidence interval as small as possible, so the symmetry of the normal distribution implies that we should take ξ and η symmetric about 0.

Hence for a 95% confidence interval we would take $-\xi = \eta = 1.96$, as $\Phi(1.96) = 0.975$. The 95% confidence interval is

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}\,,\ \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right]$$

For a 99% confidence interval, 1.96 would be replaced by 2.58, as $\Phi(2.58) = 0.995$.

(b) If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ independently, with μ and σ^2 both unknown, then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1},$$

where t_{n-1} denotes the '**Student's** $\boldsymbol{t}$**-distribution** on $n-1$ degrees of freedom' which will be studied later. So if ξ and η are such that $\mathbb{P}(\xi \le t_{n-1} \le \eta) = \gamma$, we have

$$P_{(\mu,\sigma^2)}\left(\xi \le \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\{S_{XX}/(n-1)\}}} \le \eta\right) = \gamma,$$

which can be rewritten as

$$P_{(\mu,\sigma^2)}\left(\bar{X} - \eta\sqrt{S_{XX}/n(n-1)} \le \mu \le \bar{X} - \xi\sqrt{S_{XX}/n(n-1)}\right) = \gamma.$$

Again the choice of ξ and η is not unique, but it is natural to try to make the length of the confidence interval as small as possible. The symmetry of the t-distribution implies that we should choose ξ and η symmetrically about 0.

4.2 Opinion polls

Opinion polls are typically quoted as being accurate to $\pm 3\%$. What does this mean and how many people must be polled to attain this accuracy?

Suppose we are trying to estimate p, the proportion of people who support the Labour party amongst a very large population. We interview n people and estimate p from $\hat{p} = \frac{1}{n}(X_1 + \cdots + X_n)$, where $X_i = 1$ if the ith person supports Labour and $X_i = 0$ otherwise. Then

$$\mathbb{E}\hat{p} = p \qquad \text{and} \qquad \operatorname{var}\hat{p} = \frac{p(1-p)}{n} \le \frac{1}{4n}\,,$$

where the inequality follows from the fact that $p(1-p)$ is maximized by $p = \frac{1}{2}$.

Let us approximate the distribution of $\hat{p}(X)$ by $N\big(p, p(1-p)/n\big)$. This is very good for n more than about 20. Then we have that approximately

$$(\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0,1).$$

So

$$
\begin{aligned}
\mathbb{P}(\hat{p} - 0.03 \le p \le \hat{p} + 0.03) &= \mathbb{P}\left(-\frac{0.03}{\sqrt{p(1-p)/n}} \le \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le \frac{0.03}{\sqrt{p(1-p)/n}} \right) \\
&\approx \Phi\left(0.03\sqrt{n/p(1-p)}\right) - \Phi\left(-0.03\sqrt{n/p(1-p)}\right) \\
&\ge \Phi(0.03\sqrt{4n}) - \Phi(-0.03\sqrt{4n})
\end{aligned}
$$

For this to be at least 0.95, we need $0.03\sqrt{4n} \ge 1.96$, or $n \ge 1068$.

Opinion polls typically use a sample size of about 1,100.

Example 4.2 *U.S. News and World Report* (Dec 19, 1994) reported on a telephone survey of 1,000 Americans, in which 59% said they believed the world would come to an end, and of these 33% believed it would happen within a few years or decades.

Let us find a confidence interval for the proportion of Americans who believe the end of the world in imminent. Firstly, $\hat{p} = 0.59(0.33) = 0.195$. The variance of $\hat{p}$ is $p(1-p)/590$ which we estimate by $(0.195)(0.805)/590 = 0.000266$. Thus an approximate 95% confidence interval is $0.195 \pm \sqrt{0.00266}(1.96)$, or $[0.163, 0.226]$.

Note that this is only approximately a 95% confidence interval. We have used the normal approximation, and we have approximated $p(1-p)$ by $\hat{p}(1-\hat{p})$. These are both good approximations and this is therefore a very commonly used analysis.

Sampling from a small population*

For small populations the formula for the variance of $\hat{p}$ depends on the total population size N. E.g., if we are trying to estimate the proportion p of $N = 200$ students in a lecture who support the Labour party and we take $n = 200$, so we sample them all, then clearly $\mathrm{var}(\hat{p}) = 0$. If $n = 190$ the variance will be close to 0. In fact,

$$
\mathrm{var}(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}\,.
$$

4.3 Constructing confidence intervals

The technique we have used in these examples is based upon finding some statistic whose distribution does not depend on the unknown parameter θ. This can be done when θ is a **location parameter** or **scale parameter**. In section 4.1 μ is an example of a location parameter and σ is an example of a scale parameter. We saw that the distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$ does not depend on μ or σ.

In the following example we consider a scale parameter.

Example 4.3 *Suppose that $X_1, \dots, X_n$ are IID $\mathcal{E}(\theta)$.* Then

$$
f(x \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_i x_i}
$$

so $T(X) = \sum_i X_i$ is sufficient for θ. Also, $T \sim \mathrm{gamma}(n, \theta)$ with pdf

$$
f_T(t) = \theta^n t^{n-1} e^{-\theta t}/(n-1)!, \quad t > 0.
$$

Consider $S = 2\theta T$. Now $\mathbb{P}(S \le s) = \mathbb{P}(T \le s/2\theta)$, so by differentiation with respect to s, we find the density of S to be

$$
f_S(s) = f_T(s/2\theta)\frac{1}{2\theta} = \frac{\theta^n (s/2\theta)^{n-1} e^{-\theta(s/2\theta)}}{(n-1)!}\frac{1}{2\theta} = \frac{s^{n-1}(1/2)^n e^{-s/2}}{(n-1)!}, \qquad s > 0\,.
$$

So $S = 2\theta T \sim \mathrm{gamma}\left(n, \frac{1}{2}\right) \equiv \chi^2_{2n}$.

Suppose we want a 95% confidence interval for the mean, $1/\theta$. We can write

$$
\mathbb{P}(\xi \le 2T\theta \le \eta) = \mathbb{P}\left(2T/\eta \le 1/\theta \le 2T/\xi\right) = F_{2n}(\xi) - F_{2n}(\eta)\,,
$$

where F_{2n} is the cdf of a χ^2_{2n} RV.

For example, if $n = 10$ we refer to tables for the χ^2_{20} distribution and pick $\xi = 34.17$ and $\eta = 9.59$, so that $F_{20}(\xi) = 0.975$, $F_{20}(\eta) = 0.025$ and $F_{20}(\xi) - F_{20}(\eta) = 0.95$. Then a 95% confidence interval for $1/\theta$ is

$$
[2t/34.17\,,\ 2t/9.59\,]\,.
$$

Along the same lines, a confidence interval for σ can be constructed in the circumstances of Example 4.1 (b) by using fact that $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$. E.g., if $n = 21$ a 95% confidence interval would be

$$
\left[\sqrt{S_{xx}/34.17}\,,\ \sqrt{S_{xx}/9.59}\,\right]\,.
$$

4.4 A shortcoming of confidence intervals*

Confidence intervals are widely used, e..g, in reporting the results of surveys and medical experiments. However, the procedure has the problem that it sometimes fails to make the best interpretation of the data.

Example 4.4 *Suppose X_1, X_2 are two IID samples from $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$.* Then

$$
\mathbb{P}(\min_i x_i \le \theta \le \max_i x_i) = \mathbb{P}(X_1 \le \theta \le X_2) + \mathbb{P}(X_2 \le \theta \le X_1) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}\,.
$$

So $(\min_i x_i, \max_i x_i)$ is a 50% confidence interval for θ.

But suppose the data is $x = (7.4, 8.0)$. Then we know $\theta > 8.0 - 0.5 = 7.5$ and $\theta < 7.4 + 0.5 = 7.9$. Thus with certainty, $\theta \in (7.5, 7.9) \subset (7.4, 8.0)$, so we can be 100% certain, not 50% certain, that our confidence interval has captured θ. This happens whenever $\max_i x_i - \min_i x_i > \frac{1}{2}$.