4.2 Confidence intervals

4.2.1 Preliminaries

Definition: If for a given probability distribution, we have

$$\text{Prob}\{a \le x \le b\} = c,$$

the interval $x \in [a, b]$ is called a confidence interval with the confidence level c (measured in %).

Example: Confidence intervals for normal distribution with mean μ and variance σ^2:

$$\text{Prob}\{\mu - 1.96\sigma \le x \le \mu + 1.96\sigma\} = 0.95$$

$$\text{Prob}\{\mu - 2.58\sigma \le x \le \mu + 2.58\sigma\} = 0.99$$

Main problem: Given a sample of independent, normally distributed data $(x_1, ..., x_n)$ and a confidence level, estimate a confidence interval for the mean μ and variance σ^2.

4.2.2 Recipe # 14: How to estimate a confidence interval for mean μ if variance σ^2 is given

1. Compute sample mean $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$
2. Choose a confidence level c
3. Define a standardized normal variable $z = (\bar{x} - \mu)\sqrt{n}/\sigma$ and compute a from probability distribution:

$$\Phi(a) = \frac{1+c}{2}, \qquad \Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2} dz$$

4. Compute the confidence interval for μ:

$$\bar{x} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + a\frac{\sigma}{\sqrt{n}}$$

4.2.3 Recipe # 15: How to estimate a confidence interval for mean μ if variance σ^2 is also unknown

Theorem: Let $(x_1, ..., x_n)$ be distributed normally with mean μ and variance σ^2. Let $\bar{x}$ be sample mean and s^2 be sample variance. The standardized variable

$$z = \frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

has t-distribution of the $(n-1)$-th degree with the density:

$$p_{n-1}(z) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)} \frac{1}{\left(1 + \frac{z^2}{n-1}\right)^{n/2}}.$$

1. Compute sample mean and variance

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

2. Choose a confidence level c
3. Define a standardized variable $z = (\bar{x} - \mu)\sqrt{n}/s$ and compute a from probability distribution:

$$\Phi_{n-1}(a) = \frac{1+c}{2}, \qquad \Phi_{n-1}(z) = \int_{-\infty}^{z} p_{n-1}(z)dz$$

4. Compute the confidence interval for μ:

$$\bar{x} - a\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{s}{\sqrt{n}}$$

Remark: The confidence interval becomes wider if the variance is unknown and is estimated from the sample variance.

Remark: In the limit $n \to \infty$, confidence intervals shrink and the point estimates $\mu \approx \bar{x}$ and $\sigma^2 \approx s^2$ become more accurate.

4.2.4 Recipe # 16: How to estimate a confidence interval for variance σ^2

Theorem: Let $(x_1, ..., x_n)$ be distributed normally with mean μ and variance σ^2. Let $\bar{x}$ be sample mean and s^2 be sample variance. The standardized variable

$$y = \frac{(n-1)s^2}{\sigma^2}$$

has χ-square distribution of the $(n-1)$-th degree with the density:

$$p_{n-1}(y) = \frac{y^{(n-3)/2}e^{-y/2}}{2^{(n-1)/2}\Gamma((n-1)/2)}, \qquad y > 0.$$

1. Compute sample mean and variance

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

2. Choose a confidence level c

3. Define a standardized variable $y = (n-1)s^2/\sigma^2$ and compute a_- and a_+ from probability distribution:

$$\Phi_{n-1}(a_\pm) = \frac{1 \pm c}{2}, \qquad \Phi_{n-1}(y) = \int_0^y p_{n-1}(y)dy$$

4. Compute the confidence interval for σ^2:

$$\frac{(n-1)s^2}{a_+} \le \sigma^2 \le \frac{(n-1)s^2}{a_-}$$

Remark: If sample data $(x_1, ..., x_n)$ are not normally distributed, the sample mean $\bar{x}$ is still normally distributed in the limit $n \to \infty$.

4.2.5 Recipe # 17: How to estimate goodness of a fit

Theorem: Let $(x_1, ..., x_n)$ be distributed with the probability density $p(x)$. Let the interval for $x \in \mathbb{R}$ is divided into m subintervals for $x_j \le x \le x_{j+1}$, $j = 0, 1, ..., m$. Let ω_j be the relative frequency of data points in the interval $x_j \le x \le x_{j+1}$, while p_j be the theoretical probability

$$p_j = \text{Prob}\{x_j \le x \le x_{j+1}\} = \int_{x_j}^{x_{j+1}} p(x)dx.$$

The standardized variable

$$\chi^2 = n \sum_{j=1}^{m} \frac{(\omega_j - p_j)^2}{p_j}$$

has χ-square distribution of the $(m-1)$-th degree.

1. Compute the value of χ^2
2. Choose a confidence level c
3. Compute a from probability distribution:

$$\Phi_{m-1}(a) = c, \qquad \Phi_{m-1}(y) = \int_0^y p_{m-1}(y)dy$$

4. If $\chi^2 > a$, the probability distribution $p(x)$ is not a good fit to the data points. The opposite holds if $\chi^2 < a$.

Remark: The greater is the value of n, the closer the experimental probabilities $(\omega_1, ..., \omega_m)$ are expected to match with the theoretical probabilities $(p_1, ..., p_m)$. If they are not, the theoretical probability density $p(x)$ is not a good fit to probability distribution of the data points $(x_1, ..., x_n)$.