COMP229: Introduction to Data Science
Lecture 17: confidence intervals

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

Statistical inference

In the past lecture we assumed that a random process had a specific probability density,

e.g. the normal probability density with a given mean μ and a standard deviation σ.

In practice we have sample data and wish to make conclusions about a population, e.g. about parameters (μ and σ) of a normal density.

Drawing such conclusions is *statistical inference*.

Point estimates vs interval estimates

A *point* estimate is a single value estimated from a sample, e.g. the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a point estimate of the mean of a normal density.

An *interval* estimate specifies a range containing the estimated parameter, e.g. the mean μ is within $[\bar{x} - \text{margin of error}, \bar{x} + \text{margin of error}]$. The last conclusion will come with a confidence, e.g. 95%.

Under what conditions can we do this?

Conditions for inference about a mean

1. We have a simple random sample (SRS) from the larger population (theoretically infinite).

2. The variable we measure has the normal density $N(\mu, \sigma^2)$ in the very large population (not in the much smaller sample).

3. We know the standard deviation σ of the variable in question, but not the mean μ.

To make an estimate, we need sample values.

A typical problem to estimate μ

Assume that we know only 9 exam marks $\{60, 70, 80, 50, 90, 40, 30, 45, 75\}$ and the standard deviation $\sigma = 15$. Assuming that the exam marks have a normal density $N(\mu, \sigma^2)$, estimate an interval for the mean μ with confidence 95%.

Step 1. Compute the sample mean $\bar{x} =$

$$\frac{60 + 70 + 80 + 50 + 90 + 40 + 30 + 45 + 75}{9} = 60.$$

Step 2. Considering $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ as the random variable, estimate its standard deviation below.

The deviation of the average variable

Claim 17.1. [no proof needed] If variables X_i have a normal density $N(\mu, \sigma^2)$ for $i = 1, \dots, n$, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has the normal density $N(\mu, \sigma^2/n)$.

The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$.

Step 3. For the required 95% confidence, the 68–95–99.7 rule says that the random variable $\bar{X}$ is within $2\sigma = 10$ of the mean μ with probability 95%, i.e. $P(|\bar{X} - \mu| < 10) = 0.95$. We'll rewrite the conclusion for μ in terms of $\bar{x}$ and an error margin.

Obtaining the error margin

Step 4. $P(|\bar{X} - \mu| < 10) = 0.95$ means that for 95% of all samples (of only 9 marks from a large class), μ is within 10 from the mean $\bar{x} = 60$.

Hence μ is within 60 ± 10 with confidence 95%.

Here $2\sigma = 10$ is the margin of the estimate error.

The confidence level 95% is the probability that the found interval will contain the true μ in repeated samples (of 9 marks from a large class).

General critical values z^*

For a given probability P and normal variable $Z \sim N(0,1)$, the equation $P(|Z| < z^*) = P$ has numerical solutions in a "standard normal table".

The approximate 68–95–99.7 rule said that $P(|Z| < 2) \approx 0.95$, the better value is 1.96.

confidence level	90%	95%	99%
critical value z^*	$1.645 \approx 1.6$	$1.96 \approx 2$	$2.576 \approx 2.6$

You could remember these values for the exam.

Confidence interval claim

Claim 17.2. Let z^* be the critical value satisfying $P(|Z| < z^*) = P$ for a given confidence level P and a normal variable $Z \sim N(0,1)$. Assume that n samples are independently drawn from a normal distribution whose standard deviation is σ. Then the mean μ has the confidence interval $\bar{x} \pm z^* \dfrac{\sigma}{\sqrt{n}}$.

Problem. The sample 60, 70, 65, 80, 85, 35, 55, 50, 100, 20, 90, 60, 40, 30, 45, 75 is from a normal distribution with the standard deviation $\sigma = 20$. Estimate the mean with confidence 95%.

One more typical solution

The 16 given values have the simple average $(60 + 70 + 65 + 80 + 85 + 35 + 55 + 50 + 100$ $+20 + 90 + 60 + 40 + 30 + 45 + 75)/16 = 60.$

For confidence 95%, the critical value is $z^* = 1.96$. The margin of error is $z^* \dfrac{\sigma}{\sqrt{n}} = 1.96 \times \dfrac{20}{\sqrt{16}} = 9.8$ and the mean μ is estimated between 60 ± 9.8.

μ is between 60 ± 12.88 with confidence 99%.

μ is between 60 ± 8.225 with confidence 90%.

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2–3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. For a normal variable $Z \sim N(0, 1)$, what is z^* satisfying $P(|Z| < z^*) = 0.95$?

Answer to the quiz and summary

Answer.

confidence level	90%	95%	99%
critical value z^*	1.645	1.96	2.576

To estimate the mean μ for a confidence P using a known deviation σ and a sample of n values, do

- find the simple average $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
- compute the standard deviation $\frac{\sigma}{\sqrt{n}}$ of $\bar{X}$;
- find the value z^* from $P(|Z| < z^*) = P$;
- write the interval for μ between $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$.