Statistics

Week 4: Confidence Intervals (Chapter 6)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Reminder

Homework 1 due on Tuesday 1pm. (Let me know if you have trouble submitting it on *eDimension*.)

This week's project recitation: Thursday 2pm, in **TT15** (1.510).

Outline

Motivation

So far we have studied *point* estimators for a statistic (such as μ), which give a single value estimate (such as $\bar{x}$) for that statistic.

However, it would be more useful to give an *interval* estimate, so that we could make statements such as: 'most' of the time, μ will lie between the values L and U. (We will quantify this soon.)

Consider the following problem: a new training technique is believed to improve running times. After a month of training with this technique, six runners from a team recorded times of 50.1, 50.3, 50.3, 51.2, 51.5, 51.6 (in s). Is this *significantly* lower than the team average before the technique was introduced, 52s?

We will look at two approaches to this problem: **confidence intervals**, and **hypothesis testing**.

Confidence interval

Confidence interval (CI)

A confidence interval is an interval estimate of a parameter θ, and can be used to indicate the reliability of an estimate.

It is an interval $[L, U]$ such that

$$P(L \leq \theta \leq U) = 1 - \alpha,$$

where $\alpha \in (0, 1)$, and L, U are functions of the sample X_i.

$(1 - \alpha)$ is called the *confidence level*.

Commonly used values for α include 0.1, 0.05 and 0.01. By convention, 0.05 is very common, though the value of α you pick should depend on the nature of the problem.

Two-sided confidence interval

Consider a random sample $X_1, X_2, \ldots, X_n$ drawn from a distribution with mean μ and variance σ^2. Suppose that n is large (e. g. $n > 30$), μ is unknown and to be estimated, but σ^2 is *known*.

From CLT, $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approximately standard normal, and hence (approximately)

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Rearranging, we obtain

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

This gives a 95%, *two-sided* confidence interval *for* the mean μ,

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

Interpretation

Note 1: if we know that the population distribution is normal, then Z is exactly normal for any n.

Note 2: once the confidence interval (CI) is calculated, the true mean μ either lies inside it, or it doesn't. So, technically speaking, it is *incorrect* to say that μ lies inside $\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ with 95% probability.

One acceptable interpretation for the CI: if we repeatedly draw samples of size n from the same population, and calculate the CI using the same method each time, then the proportion of CIs that contains μ will be 95%.

One-sided confidence intervals

We can also construct *one-sided* confidence intervals.

Lower 95% CI: $P\Big(\mu \geq \bar{X} - 1.645\dfrac{\sigma}{\sqrt{n}}\Big) = 0.95$,

so the interval is $\Big[\bar{X} - 1.645\dfrac{\sigma}{\sqrt{n}}, \infty\Big)$.

Upper 95% CI: $P\Big(\mu \leq \bar{X} + 1.645\dfrac{\sigma}{\sqrt{n}}\Big) = 0.95$,

so the interval is $\Big(-\infty, \bar{X} + 1.645\dfrac{\sigma}{\sqrt{n}}\Big]$.

Exercises: (1) Where does the constant 1.645 come from? (Draw a picture!)

(2) Find the expressions for the 99% one-sided CIs. (Use the *Excel* command `norm.s.inv`).

Outline

Unknown variance

In most applications, however, σ^2 is *unknown*, and is estimated using s^2.

Remember that for *normal* X_i, $\dfrac{\bar{X} - \mu}{s/\sqrt{n}}$ follows a t-distribution.

Therefore, to calculate a two-sided CI for μ, we use

$$P\left(\bar{X} - t_{n-1,\,1-\alpha/2}\,\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1,\,1-\alpha/2}\,\frac{s}{\sqrt{n}}\right) = 1 - \alpha,$$

where $t_{n-1,\,1-\alpha/2}$ is a critical point (inverse of the cdf) of the t-distribution with $(n-1)$ degrees of freedom.

In *Excel*: use $\texttt{t.inv}(1-\alpha/2,\ n-1)$.

Unknown variance, continued

$s/\sqrt{n}$ is called the standard error.

Using the t-distribution results in wider CIs (since the distribution has heavier tails).

One-sided confidence intervals can be similarly obtained.

Whether to use two-sided or one-sided CI depends on the context of the problem.

Example: suppose 14 sheets of rubber have sample mean strength $\bar{x} = 33.7$ and sd $s = 0.80$. If you wish to make a statement like 'with 95% confidence, the population mean strength of the rubber is at least L', then we require a one-sided confidence interval, with

$$L = \bar{x} - t_{13,\,0.95}\,\frac{s}{\sqrt{14}} \approx 33.3$$

t-distribution vs z-distribution

Short summary:

If σ^2 is **known**, and the population is **normal**, then we use the z-distribution (standard normal) to find the CI.

If σ^2 is **known**, the population is not normal but n is large, CLT allows us to also use the z-distribution.

If σ^2 is **unknown**, and the population is **normal**, then we use the t-distribution to find the CI.

If σ^2 is **unknown**, the population is not normal but n is large, we may again apply the CLT and observe that $s^2 \approx \sigma^2$. However, a *conservative* approach is to still use the t-distribution for this case.

Outline

Confidence interval for variance

Suppose $X_1, X_2, \ldots, X_n$ are random samples drawn from a normal population with variance σ^2.

Recall that $\dfrac{(n-1)s^2}{\sigma^2}$ is a χ^2_{n-1} random variable. Therefore, to find a $(1-\alpha)$ CI for σ^2:

$$P\left(\chi^2_{n-1,\,\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,\,1-\alpha/2}\right) = 1-\alpha,$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}\right) = 1-\alpha.$$

Here $\chi^2_{n-1,\,\alpha/2}$ is a critical point of the distribution. In *Excel*: use `chisq.inv`$(\alpha/2,\ n-1)$.

One-sided CIs can be similarly obtained.

Exercise

A bottling company uses a filling machine, and the amount filled is normally distributed. Based on 16 samples, the sample standard deviation of the amount filled is 0.700ml. Find a 95% two-sided confidence interval for σ.