Statistical Inference

Damien Benveniste

October 15, 2017

1 Confidence Interval

The confidence interval is the interval such that the true population statistics is contained with a certain level of confidence between a lower and upper bounds learned from a sample of data. Let's consider $\{X_1, X_2, \dots, X_n\}$ i.i.d. normal distributed random variables. The sample mean and variance are:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{2}$$

Then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{3}$$

has a Student's t-distribution with $n-1$ degrees of freedom. Let's a pause a second on the variance of $\bar{X}$

$$\begin{aligned} Var[\bar{X}] &= Var[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{4}$$

Therefore S^2/n is the unbiased estimate of $Var[\bar{X}]$. We now want to find a upper bound c and lower bound $-c$ such that

$$Pr(-c \leq T \leq c) = 1 - \alpha \tag{5}$$

where it is common to take $\alpha = 0.05$ (95% confidence) and we are going to make this assumption from now on. We have

$$\begin{aligned} & Pr(-c \le T \le c) = 0.95 \\ \Rightarrow \quad & Pr(-c \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le c) = 0.95 \\ \Rightarrow \quad & Pr(\bar{X} - c\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + c\frac{S}{\sqrt{n}}) = 0.95 \end{aligned} \tag{6}$$

This is a theoretical confidence interval. For an experiment, there are no longer probabilistic concepts attached to the problem and the confidence interval inferred from the experiment is $\left[\bar{X} - c\frac{S}{\sqrt{n}}, \bar{X} + c\frac{S}{\sqrt{n}}\right]$

If $n \to \infty$ then $c \to 1.96$. For a decent number of samples it is usually a good approximation to use $c \simeq 2$.

2 Hypothesis testing

Hypothesis testing is a framework within statistics theory to infer from the computation of a sample statistics if a population statistics has a certain value. For example, if we sample two populations and estimate the means $\hat{\mu}_1$ and $\hat{\mu}_2$, can we conclude $\mu_1 \neq \mu_2$ or not? We establish a null hypothesis:

$$H_0 : \mu_1 = \mu_2 \tag{7}$$

and an alternative hypothesis

$$H_a : \mu_1 \neq \mu_2 \tag{8}$$

and we test a hypothesis versus the other.

More generally we can establish different sets of hypotheses to test:

Null hypothesis	Alternative hypothesis
$\mu_1 - \mu_2 = d$	$\mu_1 - \mu_2 \neq d$
$\mu_1 - \mu_2 \le d$	$\mu_1 - \mu_2 > d$
$\mu_1 - \mu_2 \ge d$	$\mu_1 - \mu_2 < d$

2.1 Two-Tailed Two-sample t-Test

Two-tailed refers to

$$H_0 : (\mu_1 - \mu_2 \le d) \cap (\mu_1 - \mu_2 \ge d) \text{ or equivalently } H_0 : \mu_1 - \mu_2 = d \tag{9}$$

and

$$H_a : (\mu_1 - \mu_2 > d) \cup (\mu_1 - \mu_2 < d) \text{ or equivalently } H_a : \mu_1 - \mu_2 \neq d \tag{10}$$

and Two-sample refers to the estimates $\hat{\mu}_1$ and $\hat{\mu}_2$.

Once again let's assume that the two samples $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ are composed of i.i.d normal random variables. We have

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \hat{\mu}_2 = \frac{1}{m}\sum_{i=1}^{m} Y_i. \tag{11}$$

Let's consider the variable $\delta\mu = \mu_1 - \mu_2$. An unbiased estimate for $\delta\mu$ is

$$\widehat{\delta\mu} = \hat{\mu}_1 - \hat{\mu}_2. \tag{12}$$

We have also

$$\begin{aligned} Var[\widehat{\delta\mu}] &= Var[\hat{\mu}_1] + Var[\hat{\mu}_2] \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m} \end{aligned} \tag{13}$$

Therefore

$$SE^2 \;=\; \frac{S_1^2}{n} + \frac{S_2^2}{m} \tag{14}$$

is an unbiased estimate of $Var[\widehat{\delta\mu}]$ and SE is the standard error. Because with have a normal assumption

$$T = \frac{\widehat{\delta\mu} - d}{SE} \tag{15}$$

follows a Student's t-distribution. A good approximation for the number of degree of freedom for this distribution is given by the Welch-Satterthwaite equation:

$$d.f. = \frac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\left(\frac{S_1^2}{n}\right)^2 \frac{1}{n-1} + \left(\frac{S_2^2}{m}\right)^2 \frac{1}{m-1}}. \tag{16}$$

What are the chance to draw $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ if $\mu_1 - \mu_2 = d$?

$$\begin{aligned} P &= Pr\left(\left(t > \left|\frac{\widehat{\delta\mu} - d}{SE}\right|\right) \cup \left(t < -\left|\frac{\widehat{\delta\mu} - d}{SE}\right|\right) \middle| H_0\right) \\ &= Pr\left(t > \left|\frac{\widehat{\delta\mu} - d}{SE}\right| \middle| H_0\right) + Pr\left(t < -\left|\frac{\widehat{\delta\mu} - d}{SE}\right| \middle| H_0\right) \end{aligned} \tag{17}$$

P is the P-value and captures how likely an sample estimate $\widehat{\delta\mu}$ can depart from d. A small P-value highlights how unlikely it is to draw the samples we had to estimate $\widehat{\delta\mu}$ if $\mu_1 - \mu_2 = d$. A small P-value pushes us to reject that $\mu_1 - \mu_2 = d$ is true and therefore to accept $\mu_1 - \mu_2 \neq d$.

2.2 One-Tailed t-Test

The one-tailed test is simpler as we want to understand the probability to draw $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ if $\mu_1 - \mu_2 \geq d$:

$$P = Pr\left(t < \frac{\widehat{\delta\mu} - d}{SE} \middle| H_0 \right) \tag{18}$$

or if $\mu_1 - \mu_2 \leq d$

$$P = Pr\left(t > \frac{\widehat{\delta\mu} - d}{SE} \middle| H_0 \right) \tag{19}$$

2.3 Types of errors

Accepting or rejecting hypotheses based on some experimental evidences can lead to false claims due to the lack of statistics. Those errors have the following nomenclature

	H_0 is true	H_a is true
Accept H_0	Good decision	Type II Error
Reject H_0	Type I Error	Good decision