

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024
CONSTRUCTING A CONFIDENCE INTERVAL

We seek two random variables

$$L = L(X_1, \dots, X_n) \quad \text{and} \quad R = R(X_1, \dots, X_n)$$

which are functions of the random sample $X_1, \dots, X_n$ such that for a specified probability α

$$P(L < \theta < R) = 1 - \alpha.$$

Both random variables $L(X_1, \dots, X_n)$ and $R(X_1, \dots, X_n)$ have their own (sampling) probability distributions.

The interval (L, R) is called a $100(1 - \alpha)\%$ **confidence interval** (C. I.) for θ .

- The **random interval** (L, R) will ‘cover’ the true value θ with probability $1 - \alpha$.
- When we write $P(L < \theta < R)$, the probability calculation is done with respect to the sampling distribution of L and R .
- Recall that here we are assuming θ is unknown; we will only see the data, and make assumptions about the distribution of the data.

We interpret the interval (L, R) as follows: if we were to carry out the same study infinitely often under the same conditions, and each time compute a $100(1 - \alpha)\%$ C.I. using the same procedure, then $100(1 - \alpha)\%$ of these intervals will contain θ .

Procedure: We will study the case where θ is either

- population mean μ when the data are Normally distributed
- population probability p when the data are Binomially distributed.

The steps are

1. Identify the parameter of interest: either μ or p .
2. Specify the value of α : often, we assume $\alpha = 0.05$.
3. Identify the experimental setting: the formula for (L, R) depends on the setting.
4. Once we know the setting, we use the given formulas for L and R .
5. Substitute the data $x_1, \dots, x_n$ into the formulas to obtain the observed interval (l, r) , where

$$l = L(x_1, \dots, x_n) \quad r = R(x_1, \dots, x_n).$$

To determine the correct formulas, we need to identify

(A) the assumed probability model/parameter of interest

- Normal, μ
- Binomial, p

(B) the sample size assumption

- n small ($n < 30$)
- n large ($n \geq 30$)

(C) the variance assumption for the Normal case

- σ known
- σ unknown

We will base confidence intervals on the summary statistic random variables

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

that is, the *sample mean* and *sample variance* random variables. We will also refer to the *upper a point* of the Normal and (Student) t -distributions:

- The upper a point of a standard Normal distribution is the point on the horizontal axis, z_a such that $P(Z > z_a) = a$.
- The upper a point of a t -distribution with ν degrees of freedom is the point on the horizontal axis, $t_{a,\nu}$ such that $P(T > t_{a,\nu}) = a$, where the random variable T has a t -distribution with ν degrees of freedom; ν depends on the sample size, and here $\nu = n - 1$. The Table below gives the upper a points for a range of ν values.

Table 1: Upper a points of t_ν distribution

ν	a					ν	a				
	0.1	0.05	0.025	0.01	0.001		0.1	0.05	0.025	0.01	0.001
4	1.533	2.132	2.776	3.747	7.173	14	1.345	1.761	2.145	2.624	3.787
5	1.476	2.015	2.571	3.365	5.893	15	1.341	1.753	2.131	2.602	3.733
6	1.440	1.943	2.447	3.143	5.208	16	1.337	1.746	2.120	2.583	3.686
7	1.415	1.895	2.365	2.998	4.785	17	1.333	1.740	2.110	2.567	3.646
8	1.397	1.860	2.306	2.896	4.501	18	1.330	1.734	2.101	2.552	3.610
9	1.383	1.833	2.262	2.821	4.297	19	1.328	1.729	2.093	2.539	3.579
10	1.372	1.812	2.228	2.764	4.144	20	1.325	1.725	2.086	2.528	3.552
11	1.363	1.796	2.201	2.718	4.025	25	1.316	1.708	2.060	2.485	3.450
12	1.356	1.782	2.179	2.681	3.930	30	1.310	1.697	2.042	2.457	3.385
13	1.350	1.771	2.160	2.650	3.852	Inf	1.282	1.645	1.960	2.326	3.090

Case 1: Suppose that $X_1, \dots, X_n$ are a random sample from a **Normal** distribution with **unknown** μ and σ , and $n < 30$. A $100(1 - \alpha)\%$ confidence interval for μ is (L, R) , where:

$$L = \bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} \quad R = \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}.$$

Once we have recorded the data, we report the interval

$$(l, r) = \left(\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right).$$

Example: Suppose we have a random sample of $n = 15$ measurements from a Normal population, and observe that $\bar{x} = 4.68$ and $s = 2.31$. Here $n - 1 = 14$, so we can report the following intervals

	α	$t_{\alpha/2, 14}$	C.I.
90%	0.10	1.761	$4.68 \pm (1.761 \times 2.31)/\sqrt{15} = (3.629, 5.731)$
95%	0.05	2.145	$4.68 \pm (2.145 \times 2.31)/\sqrt{15} = (3.401, 5.959)$
99%	0.01	2.977	$4.68 \pm (2.977 \times 2.31)/\sqrt{15} = (2.904, 6.456)$

Case 2: Suppose that $X_1, \dots, X_n$ are a random sample from a Normal distribution with unknown μ , with σ **known**. A $100(1 - \alpha)\%$ confidence interval for μ is (L, R) , where

$$L = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad R = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Here we know that the sampling distribution of $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$, so therefore

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

We know that $z_{\alpha/2}$ is the upper $\alpha/2$ value for the standard Normal, so therefore, by symmetry of the Normal pdf

$$P(-z_{\alpha/2} < Z \leq z_{\alpha/2}) = 1 - \alpha \quad \text{i.e.} \quad P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Rearranging the inequalities, this means that

$$P(L < \mu \leq R) = P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha.$$

Example: Suppose we have a random sample of $n = 15$ measurements from a Normal population with $\sigma = 0.5$, and observe that $\bar{x} = 4.68$. We can report the following intervals.

	α	$z_{\alpha/2}$	C.I.
90%	0.10	1.645	$4.68 \pm (1.645 \times 0.5)/\sqrt{15} = (4.468, 4.892)$
95%	0.05	1.960	$4.68 \pm (1.960 \times 0.5)/\sqrt{15} = (4.427, 4.933)$
99%	0.01	2.576	$4.68 \pm (2.576 \times 0.5)/\sqrt{15} = (4.347, 5.013)$

Case 3: Suppose that $X_1, \dots, X_n$ are a random sample from a distribution with unknown μ and σ (normality of the X s is not required), where $n \geq 30$. A $100(1 - \alpha)\%$ confidence interval for μ is (L, R) , where:

$$L = \bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \quad R = \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}$$

Here by the Central Limit Theorem, we have the approximate result

$$\bar{X} \dot{\sim} Normal(\mu, \sigma^2/n),$$

and hence, approximately, as above

$$P(L < \mu \leq R) = P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha.$$

However, we do not know σ , so we use instead a point estimator of σ , namely S .

Example: Suppose we have a random sample of $n = 68$ measurements from a population and observe that $\bar{x} = 4.68$ and $s = 1.6$. We can report the following intervals.

	α	$z_{\alpha/2}$	C.I.
90%	0.10	1.645	$4.68 \pm (1.645 \times 1.6)/\sqrt{68} = (4.361, 4.999)$
95%	0.05	1.960	$4.68 \pm (1.960 \times 1.6)/\sqrt{68} = (4.300, 5.060)$
99%	0.01	2.576	$4.68 \pm (2.576 \times 1.6)/\sqrt{68} = (4.180, 5.180)$

Case 4: Let $X_1, \dots, X_n$ be a random sample of Bernoulli random variables with unknown probability of success p . A $100(1 - \alpha)\%$ confidence interval for p is (L, R) , where:

$$L = \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad R = \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where

$$\hat{p} = \frac{\text{Number of "successes" in the sample}}{n} = \frac{X}{n}$$

say. This applies when n is large enough, that is, when $n\hat{p} \geq 15$ and $n(1 - \hat{p}) \geq 15$. Here by the Central Limit Theorem, we have the approximate result

$$\bar{X} \dot{\sim} Normal(p, p(1 - p)/n)$$

so we can apply the formula for Case 3 with S replaced by the point estimator of $p(1 - p)$, that is, $\hat{p}(1 - \hat{p})$.

Example: Suppose we have a random sample of $n = 124$ measurements from a Bernoulli population and observe that the total number of successes is $x = 72$, so that the point estimate of p is $\hat{p} = \bar{x} = 0.5806$ and $\hat{p}(1 - \hat{p}) = 0.2435$. We can report the following intervals.

	α	$z_{\alpha/2}$	C.I.
90%	0.10	1.645	$0.5806 \pm (1.645 \times \sqrt{0.2435})/\sqrt{124} = (0.508, 0.653)$
95%	0.05	1.960	$0.5806 \pm (1.960 \times \sqrt{0.2435})/\sqrt{124} = (0.494, 0.667)$
99%	0.01	2.576	$0.5806 \pm (2.576 \times \sqrt{0.2435})/\sqrt{124} = (0.466, 0.695)$