Confidence intervals for mean/proportion estimation

- Mean estimation when σ is known. Suppose that we have a random sample $(X_1,\ldots,X_n)$. Let $\mu = E(X_1)$ and $\sigma = \sqrt{Var(X_1)}$. Consider the problem of estimating μ based on the sample when σ is known. Then the sample mean $\bar{X}$ is a reasonable estimator of μ. For $a \in (0,1)$, if we can find $d \geq 0$ such that d can be computed based on the sample and

$$P(|\bar{X} - \mu| \leq d) \geq 1 - a,$$

then

$$P(\mu \in [\bar{X} - d, \bar{X} + d]) \geq 1 - a \tag{1}$$

then we say that $[\bar{X} - d, \bar{X} + d]$ is a $(1 - a)$ confidence interval (信賴區間) for μ. Here $(1 - a)$ is called the confidence level (信心水準) or the coverage probability. A common choice for $(1 - a)$ is 0.95.

- Suppose that $(X_1,\cdots,X_n)$ is a sample and θ is a quantity of interest. Suppose that we can find a range $[L,U]$ such that

  - $P(\theta \in [L,U]) \geq 1 - a$ for some $a \in (0,1)$ and
  - $[L,U]$ can be computed based on the sample.

  Then $[L,U]$ is called a $(1 - a)$ C.I. (confidence interval) for θ.

- Example 1. Suppose that $(X_1,\ldots,X_n)$ is a random sample with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$, and $\sigma > 0$ is known. Then for $k > 0$,

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

is a $(1 - 1/k^2)$ C.I. for μ. Here $\dfrac{k\sigma}{\sqrt{n}}$ is called the margin of error of the confidence interval.

- If $(X_1,\ldots,X_n)$ is a random sample from $N(\mu,\sigma^2)$, where $\sigma > 0$ is known, then a tighter C.I. for μ can be constructed based on the following fact

  Fact 1 Suppose that two random variables X and Y are independent and both are normally distributed. Then $(X + Y)$ is normally distributed.

  Note that Fact 1 implies that $\bar{X} \sim N(\mu,\sigma^2/n)$ when $(X_1,\ldots,X_n)$ is a random sample from $N(\mu,\sigma^2)$. In such case,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1), \tag{2}$$

so

$$\left[\bar{X} - z_{a/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{a/2}\frac{\sigma}{\sqrt{n}}\right] \tag{3}$$

is a $(1-a)$ confidence interval of μ, where z_a is defined so that $P(N(0,1) > z_a) = a$ for $a \in (0,1)$.

- Example 2. Check Fact 1 by generating two independent random samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ from $N(0,9)$ and $N(0,16)$ respectively with $n = 10^6$, drawing a normalized histogram based on the random sample $(X_1 + Y_1, \ldots, X_n + Y_n)$ and comparing the shape of the normalized histogram with a PDF of $N(0,25)$.

  R commands

```
n <- 10^6
x <- rnorm(n, mean=0, sd=3)
y <- rnorm(n, mean=0, sd=4)
f <- function(x){ exp(-x^2/50)/sqrt(2*pi*25) }
hist(x+y, freq=FALSE, nclass="scott")
curve(f, add=T, col=2)

#normal PDF can also be computed using dnorm
f <- function(x){ exp(-x^2/50)/sqrt(2*pi*25) }
g <- function(x){ dnorm(x, mean=0, sd=5) }
curve(f,-15,15)
curve(g,-15,15, add=T, col=2)
```

- Quantile (分位數). Suppose that $\mathcal{D}$ is a distribution with CDF F and the inverse function of F is F^{-1}. For $b \in (0,1)$, the b quantile of $\mathcal{D}$ is the number t that

$$P(\mathcal{D} \leq t) = F(t) = b,$$

  which implies that $t = F^{-1}(b)$.

  - The $(1-a)$ quantile of $N(0,1)$ is denoted by z_a, so $P(N(0,1) > z_a) = a$.
  - In R, running the command `qnorm(1-a, mean=0, sd=1)` or `qnorm(1-a)` gives z_a. For instance, `qnorm(1-0.025)` gives $z_{0.025} = 1.959964$.
  - In R, running the command `pnorm(x, mean=0, sd=1)` (or `pnorm(x)`) gives $P(N(0,1) \leq \texttt{x})$.

- Example 3. Suppose that we are interested in the annual salary of a data scientist in United States. Suppose that the salary distribution is a normal distribution with standard deviation \$14,000, and we have a random sample of 900 annual salaries of data scientists in United States. Suppose that the average annual salary in the sample is \$139,800. Give a 95% C.I. for the mean of the salary distribution. You may use 1.96 as the value of $z_{0.025}$.

Sol. $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.96$.

$$139800 \pm 1.96 \times \frac{14000}{\sqrt{900}} \approx 139800 \pm 915.$$

A 95% C.I. for the mean of the salary distribution is $[138885, 140715]$.

- If $(X_1, \cdots, X_n)$ is a random sample from a distribution with mean μ and known standard deviation σ, then by C.L.T. (Central Limit Theorem), (2) holds approximately for large n. In such case, the C.I. (confidence interval) in (3) is an approximate $(1 - a)$ confidence interval for μ since the coverage probability is approximately $(1 - a)$ for large n.

- Example 4. Suppose that we are interested in the annual salary of a data scientist in United States. Suppose that the salary distribution has standard deviation \$14,000. Suppose that we randomly choose 900 data scientists in United States and their average annual salary is \$139,800. Give an approximate 95% C.I. for the mean of the salary distribution. You may use 1.96 as the value of $z_{0.025}$.

  Sol. $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.96$.

  $$139800 \pm 1.96 \times \frac{14000}{\sqrt{900}} \approx 139800 \pm 915.$$

  Based on C.L.T., an approximate 95% C.I. for the mean of the salary distribution is $[138885, 140715]$.

- The construction of a confidence interval for μ with unknown σ involves the use of a t distribution.

- About t distributions.

  - Definition. Suppose that Z_0, Z_1, ..., Z_m are IID $N(0, 1)$ random variables. Then the distribution for

    $$\frac{Z_0}{\sqrt{(Z_1^2 + \cdots + Z_m^2)/m}}$$

    is the t distribution with m degrees of freedom, denoted by $t(m)$. The distribution of $Z_1^2 + \cdots + Z_m^2$ is the χ^2 (chi-square) distribution with m degrees of freedom, denoted by $\chi^2(m)$.

– The $t(m)$ distribution has a PDF f, where

$$f(x) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\Gamma(m/2)} \left(1 + \frac{x^2}{m}\right)^{-(m+1)/2}, \tag{4}$$

and Γ is a function on $(0, \infty)$ defined by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

for $a > 0$. The R command `gamma(a)` computes $\Gamma(a)$.

– $t(m) \to N(0,1)$ as $m \to \infty$.

- Example 5. Suppose that Z_0, Z_1, ..., Z_m are IID $N(0,1)$ random variables. Estimate the density for

$$\frac{Z_0}{\sqrt{(Z_1^2 + \cdots + Z_m^2)/m}}$$

when $m = 3$ based on 10^6 IID data using histogram and compare the estimated density with the density in (4).

R commands for carring out the above experiment:

```
m = 3
n <- 10^6
data <- rep(0, n)
for (i in 1:n){
 z <- rnorm(m+1)
 data[i] <- z[1]/sqrt(sum(z[-1]^2)/m)
 #z[1] is the first element of the vector z
 #z[-1] is the vector obtained by removing z[1] from z
}
hist(data, nclass="scott", freq=FALSE, xlim=c(-2,2))

#define f to be the PDF of t(m)
f <- function(x, df=m){
 a <- (df+1)/2
 ans <- (gamma(a)/(sqrt(df*pi)*gamma(df/2))) * (1+x^2/df)^(-a)
 return(ans)
}
curve(f, add=T, col=2)
```

Note that the above function `f` can be replaced by the function `g` defined by

```
g <- function(x){ dt(x, df=m) }
```

- Quantiles for t distributions. $t_{a,m}$ is defined such that

$$P(t(m) > t_{a,m}) = a.$$

$t_{a,m}$ is called the $(1-a)$ quantile of the $t(m)$ distribution.

  - In R, `qt(1-a, df=m)` gives $t_{a,m}$. For instance, `qt(1-0.025, df=9)` gives $t_{0.025,9} = 2.262157$.
  - $t_{a,m}$ can be found using the table "Quantiles for t distributions".

- Example 6. Find $t_{0.025,9}$ in the table "Quantiles for t distributions".

  Ans: 2.262.

- When m is large, $t_{a,m} \approx z_a$.

m	$t_{0.025,m}$	$z_{0.025}$
200	1.971896	1.959964
1000	1.962339	
10000	1.960201	

- z_a can be found in the table "Quantiles for t distributions" (look for the row with degrees of freedom $df = \infty$).

- Example 7. Find $z_{0.025}$ in the table "Quantiles for t distributions".

  Ans: 1.960.

- Estimation of μ with unknown σ. Suppose that $(X_1, \ldots, X_n)$ is a random sample from $N(\mu, \sigma^2)$ and both μ and σ are unknown. Let $\bar{X}$ and S be the sample mean and the sample standard deviation respectively. A $(1-a)$ confidence interval for μ is

$$\left[\bar{X} - t_{a/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{a/2,n-1}\frac{S}{\sqrt{n}}\right]. \tag{5}$$

- Construction of the C.I. in (5) is based on the following result.

  Fact 2 Suppose that $(X_1, \ldots, X_n)$ is a random sample from a population whose distribution is $N(\mu, \sigma^2)$. Let $\bar{X}$ be the sample mean and $S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$ be the sample standard deviation. Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1). \tag{6}$$

Replace (2) with (6) and replace the $z_{a/2}$ and σ in (3) with $t_{a/2,n-1}$ and S respectively, then we have

$$P\left(\mu \in \left[\bar{X} - t_{a/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{a/2,n-1}\frac{S}{\sqrt{n}}\right]\right) = 1 - a,$$

so $\left[\bar{X} - t_{a/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{a/2,n-1}\frac{S}{\sqrt{n}}\right]$ is a $(1-a)$ confidence interval for μ when σ is unknown.

- Example 8. Suppose that we are interested in the annual salary of a data scientist in United States. Suppose that the salary distribution is normal. Suppose that we randomly choose 49 data scientists in United States, and the average annual salary is \$139,800 and the sample standard deviation is \$14,000. Give a 95% C.I. for the mean of the salary distribution. You may use the table "Quantiles for t distributions".

  Sol. $t_{a/2,n-1} = t_{0.05/2,49-1} = t_{0.025,48} = 2.011$.

  $$139800 \pm 2.011 \times \frac{14000}{\sqrt{49}} \approx 139800 \pm 4022.$$

  A 95% C.I. for the mean of the salary distribution is $[135778, 143822]$.

- Estimation of population proportion p.
  - Population values are 0's and 1's.
  - Want to estimate p: the proportion of 1's.
  - $(X_1, \ldots, X_n)$ is a random sample (or an approximate random sample) from $Bin(1,p)$.
  - An approximate confidence interval for p with coverage probability $1-a$ is

    $$\left[\bar{X} - z_{a/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{a/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right]. \tag{7}$$

  - Construction of the C.I. in (7) is similar to that of the C.I. in (3) except that (2) is replaced by

    $$\frac{\sqrt{n}(\bar{X}-p)}{\sqrt{\bar{X}(1-\bar{X})}} \approx N(0,1).$$

- Example 9. Suppose that some union members of employees at a company are going to make a proposal of merging with another company, and they need at least 3/4 of votes for approval to make the proposal official. Suppose that 2000 union members are randomly selected and 1600 agree with the proposal. Give an approximate 95% C.I. for the proportion of union members who agree with the proposal. Based on the C.I., is it reasonable to conclude that at least 3/4 of the union members would vote for approval for the merger?

Ans. $1600/2000 = 0.8$. $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.96$.

$$0.8 \pm 1.96\sqrt{\frac{0.8 \times (1-0.8)}{2000}} = 0.8 \pm 0.018.$$

95% approximate C.I.: $[0.782, 0.818]$. It is reasonable to conclude that at least 3/4 of the union members would vote for approval for the merge since 3/4 is less than the lower bound of the 95% C.I..

- When we use $\hat{\theta}$ to estimate a parameter θ (for example, $\theta = \mu$ or p), a C.I. for θ is often of the form $[\hat{\theta} - D, \hat{\theta} + D]$. D is call the margin of error of the C.I. (信賴區間長度之一半).

- Sample size determination for estimating the population proportion. Suppose that we are given that the maximum allowable margin of error is E. Take
$$n \geq \left(\frac{z_{a/2} \cdot 0.5}{E}\right)^2,$$
then the margin of error of the C.I of p in (7) is less than or equal to E.

- Sample size determination for estimating μ when σ is known. Suppose that E is the maximum allowable margin of error. Take
$$n \geq \left(\frac{z_{a/2} \cdot \sigma}{E}\right)^2,$$
then the C.I given in (3) has marginal of error $\leq E$.

- Example 10. Suppose that a president candidate asks a survey company to give a confidence interval for the proportation of his supporters in the country. The candidate asks for 90% coverage probability and 10% maximum allowable margin of error for the confidence interval. What is the sample size required?

  Ans. $z_{a/2} = z_{0.1/2} = z_{0.05} = 1.645$.
$$\left(\frac{1.645 \times 0.5}{0.1}\right)^2 = 67.65062,$$
  so the required sample size is 68.

- Example 11. 某民調中心接受委託, 調查民眾是否贊成北北基合併成一直轄市以及首長候選人支持度. 委託人希望在95%的信心水準下, 抽樣誤差不超過正負3.4個百分點. 請問調查樣本數至少要多少?

  Ans. $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.96$.

$$\left(\frac{z_{0.025}0.5}{0.034}\right)^2 = \left(\frac{1.96 \times 0.5}{0.034}\right)^2 = 830.7958,$$

  調查樣本數至少要831.

- Example 12. Suppose that we are interested in the annual salary of a data scientist in United States. Suppose that the salary distribution is a normal distribution with standard deviation \$14,000, and we would like to obtain a 95% C.I. for the mean of the salary distribution based on a random sample of size n from the salary distribution. Suppose that the maximum allowable margin of error for the 95% C.I. is \$1,000. Find the minimum n required.

  Sol. $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.960$.

$$\left(\frac{1.960 \times 14000}{1000}\right)^2 = 752.9536,$$

  so the smallest n required is 753.