Statistical Inference : Confidence Intervals

Population: the form of the distribution is assumed known, but the parameter(s) which determines the distribution is unknown

Sample: Draw a set of random sample from the population (i.i.d)

Point estimation (MME, MLE)
Confidence Intervals:

- Confidence intervals for a population mean
- Confidence intervals for difference between two means
- Confidence intervals for variances
- Confidence intervals for proportions
- Sample Size

Confidence intervals for the mean of a single population

CI for μ

1. A set of random sample (i.i.d) from a normally distributed population.
   (i) when the variance $\sigma^2$is **known**.
   (ii) when the variance σ^2 is **unknown**.
2. Sample is **NOT** from a normal distribution.
   (a) When n is large (CLT $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$)
   (b) When n is less than 30 and underlying distribution is less normal—Non-parameter methods

CIs for μ when the variance σ^2 is known

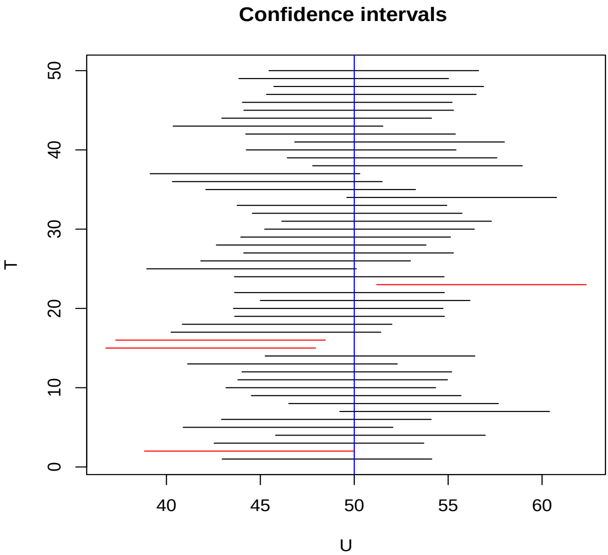
Assume the population $X \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. We draw a set of random sample of size n, let $\overline{X}$ be the sample average, and we can work out the probability that the random interval

$$[\ \overline{X} - z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})\ , \overline{X} + z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})\ ]$$

contains the unknown mean μ is $1 - \alpha$, i.e.,

$$\mathbf{P}(L \leq \mu \leq U) = 1 - \alpha$$

$[L, U]$ is a set of random intervals that contains μ with probability $1 - \alpha$;
If we replicate the sampling process 100 times, we have 100 different confidence intervals. It should be true that about 95% of them would contain the population mean μ.



Confidence intervals
T
U
0
10
20
30
40
50
40
45
50
55
60

Once the sample is observed and the sample average is computed to equal to $\overline{x}$, we call the interval

$$[\ \overline{x} \pm z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})\ ]$$

a $100(1-\alpha)\%$ **confidence intervals** for the unknown mean μ. We are $100(1-\alpha)\%$ confidence that $[\overline{x} \pm z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})]$ will contain μ

sample size n
confidence coefficient $1-\alpha$

Example

Assume the population $X \sim \mathcal{N}(\mu, 81)$, we draw a set of random sample of size $n = 10$, and have

$$60.50\ 66.18\ 48.10\ 41.21\ 53.66\ 36.49\ 54.80\ 56.04\ 43.48\ 42.41$$

Find a 95% confidence interval for μ.

Example

Assume the population $X \sim \mathcal{N}(\mu, 81)$, we draw a set of random sample of size $n = 10$, and have

$$60.50\ 66.18\ 48.10\ 41.21\ 53.66\ 36.49\ 54.80\ 56.04\ 43.48\ 42.41$$

Find a 95% confidence interval for μ.

$[\overline{x} \pm 1.96(\frac{\sigma}{\sqrt{n}})] = [44.71, 55.87]$ is a 95% confidence interval for μ.

For a particular sample and $\overline{x}$ was computed, the interval either does or does not contain the mean μ. We can't say there is 95% chance that the μ will fall between 44.71 and 55.87. We can only say that we have 95% confidence that the population mean will fall between [44.71, 55.87].
(we provide information about the uncertainty of the estimate)

CI for μ when σ^2 is also unknown.

Recall: T-**distribution** According to the definition of a T random variable: $Z \sim \mathcal{N}(0,1)$ and $V = \chi^2(r)$, Z, V are independent

$$T = \frac{Z}{\sqrt{V/r}}$$

has a t-distribution with r degrees of freedom.

Recall: Normal and χ^2 distributions

Given $X_1, \cdots, X_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution where μ and σ^2 are unknown parameters, let $\overline{X}$ be the sample average, and $S^2 = \sum (X_i - \overline{X})^2/(n-1)$ the sample variance. Define $W = (n-1)S^2/\sigma^2$ (ie sum of squares divided by σ^2, then W is a chi-square distribution with $r = n-1$ degrees of freedom. That is

$$W = \frac{(n-1)S^2}{\sigma^2} \ \sim \ \chi^2(n-1)$$

$E(W) = 2(r/2) = r$, $Var(W) = 4(r/2) = 2r$. Thus a random variable $W \sim \chi^2(v)$ have mean v and variance $2v$ and the mgf of W is $M_w(t) = \left(\frac{1}{1-2t}\right)^{\frac{r}{2}}, t < 1/2$

CI for μ when σ^2 is also unknown.

We have

$$T = \frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \Big/ (n-1)}} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

have a t distribution with $r = n - 1$ degrees of freedom (recall many properties of t-distribution?)

Random Intervals

$$[\ \overline{X} - t_{\alpha/2}(n-1)(\frac{S}{\sqrt{n}}), \overline{X} + t_{\alpha/2}(n-1)(\frac{S}{\sqrt{n}})\ ]$$

CI for μ when σ^2 is also unknown.

We have

$$T = \frac{\sqrt{n}(\overline{X}-\mu)/\sigma}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

have a t distribution with $r = n-1$ degrees of freedom (recall many properties of t-distribution?)

Random Intervals

$$[\ \overline{X} - t_{\alpha/2}(n-1)(\frac{S}{\sqrt{n}}), \overline{X} + t_{\alpha/2}(n-1)(\frac{S}{\sqrt{n}})\ ]$$

Once a random sample is observed, we compute $\overline{x}$ and s^2 and

$$[\overline{x} \pm t_{\alpha/2}(n-1)(\frac{s}{\sqrt{n}})]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

CIs for difference of two means

Two independent normal distributions

1. When both variances are known.
2. If the variances are unknown and the sample sizes are large
3. If the variances are unknown
   (a) assume common unknown equal variance
   (b) unequal variance
      (i) sample sizes are large
      (ii) sample sizes are small

Paired data, Match data, dependent data

Both variances are known

Two independent random samples of sizes n and m from the two normal distributions

$X_1, \cdots, X_n \sim \mathcal{N}(\mu_x, \sigma_X^2)$, and $Y_1, \cdots, Y_m \sim \mathcal{N}(\mu_y, \sigma_Y^2)$.

Then we have $\overline{X} \sim \mathcal{N}(\mu_X, \sigma_X^2/n)$, and $\overline{Y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2/m)$.

Let $W = \overline{X} - \overline{Y}$, then

$$W \sim \mathcal{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$$

Both variances are known

Once the samples are drawn

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

is a 100(1-α)% CI for $\mu_X - \mu_Y$

Sample sizes are large and variances are unknown

We replace variances with the sample variances s_X^2, s_Y^2 where they are the values of the respective unbiased estimates of the variances.

That is

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

is an approximate 100(1-α)% CI for $\mu_X - \mu_Y$

Sample sizes are small and variances are unknown

a. Assumed common variance

Estimate for the common variance: equal variance
$\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Denote

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

which is an unbiased estimator for the common variance σ^2.

Estimate for the common variance

Since the random samples are from two independent normal distribution with common variance , we have

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1), \quad \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(m-1)$$

and they are independent. Thus

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n+m-2)$$

and $\mathbf{E}(U) = n+m-2$, thus we have $\mathbf{E}(S_p^2) = \sigma^2$

(a). Common variance assumption

we have

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_Y)}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}}$$

but we don't know σ^2 so we have

$$T = \frac{Z}{\sqrt{U/r}} = \frac{[\overline{X} - \overline{Y} - (\mu_x - \mu_Y)]/\sqrt{\sigma^2(1/n + 1/m)}}{\sqrt{[\frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2}]/(n+m-2)}}$$

$$= \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_Y)}{\sqrt{[\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}](\frac{1}{n} + \frac{1}{m})}}$$

has a t-distribution with $r = n + m - 2$ degrees of freedom.

A 00$(1-\alpha)$% CI for $\mu_X - \mu_Y$ is

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(n+m-2)\sqrt{s_p^2(\frac{1}{n} + \frac{1}{m})}$$

(b) not equal variances

$$W = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}}$$

1. If n and m are large enough and the underlying distributions are close to normal -> use normal distribution to construct a CI
2. * If n and m are **small** -> approximating Student's t distribution has r degrees of freedom (Welch t) where

$$\frac{1}{r} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1} \quad \text{and} \quad c = \frac{s_X^2/n}{s_X^2/n + s_Y^2/m}$$

$$r = \frac{(s_X^2/n + s_Y^2/m)^2}{\frac{1}{n-1}(s_X^2/n)^2 + \frac{1}{m-1}(s_Y^2/m)^2}$$

If r is not an integer, then we use the greatest integer in r, i.e., $\lfloor r \rfloor$ the "floor" is the number of degrees of freedom

We have

$$\overline{x} - \overline{y} \pm t_{\alpha/2}(r)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

is a 100(1-α)% CI for $\mu_X - \mu_Y$.

Paired (Match) samples

X_i and Y_i are measurements taken from the same subject. X_i and Y_i are dependent random variables.

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be n pairs of dependent measurements.

Let $D_i = X_i - Y_i$, $i = 1, \cdots, n$. Suppose D_i can be thought of as a random sample from $N(\mu_D, \sigma_D^2)$, where μ_D and σ_D^2 are the mean and standard deviation of each difference.

To form a CI for $\mu_X - \mu_Y$, use

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

where $\overline{D}$ and S_D are the sample mean and sample standard deviation of the n differences. T is a t statistic with $n-1$ degrees of freedom.

Thus the CI for $\mu_D = \mu_X - \mu_Y$ is

$$\overline{d} \ \pm \ t_{\alpha/2}(n-1)\frac{s_D}{\sqrt{n}}$$

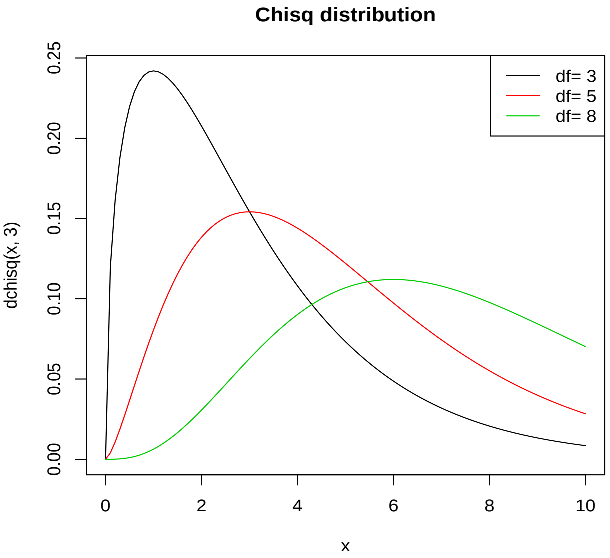
where $\overline{d}$ and s_D are the observed mean and standard deviation of the sample. (this is the same as the CI for a single mean).

Confidence intervals for variances

Recall : Chi-Square distribution
Given $X_1, \cdots, X_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution where μ and σ^2 are unknown parameters and let $S^2 = \sum(X_i - \overline{X})^2/(n-1)$.

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

chi-squared R.V.s



Example: chi-squared distribution

Let S^2 be the sample variance of a random sample of size 6 is drawn from a $\mathcal{N}(\mu, 12)$ distribution. Find
$\mathbf{P}(2.76 < S^2 < 22.2)$

Example: chi-squared distribution

Let S^2 be the sample variance of a random sample of size 6 is drawn from a $\mathcal{N}(\mu, 12)$ distribution. Find $\mathbf{P}(2.76 < S^2 < 22.2)$

Let $W = \frac{(n-1)S^2}{\sigma^2}$, then $W \sim \chi^2(5)$.

So $\mathbf{P}(2.76 < S^2 < 22.2) = \mathbf{P}(\frac{5}{12}(2.76) < \frac{(n-1)S^2}{\sigma^2} < \frac{5}{12}(22.2)) = \mathbf{P}(1.15 < W < 9.25)$ and $\mathbf{P}(2.76 < S^2 < 22.2) = \mathbf{P}(W < 9.25) - \mathbf{P}(W < 1.15) = pchisq(9.25, 5) - pchisq(1.15, 5) = 0.85$.

CI for variance σ^2

Let $X_1, \cdots, X_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, find the $100(1-\alpha)\%$ CI for σ^2

CI for variance σ^2

Let $X_1, \cdots, X_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, find the $100(1-\alpha)\%$ CI for σ^2

Select constants a and b from $\chi^2(n-1)$ such that

$$\mathbf{P}(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b) = 1 - \alpha$$

we select $a = \chi^2_{1-\alpha/2}(n-1)$ and $b = \chi^2_{\alpha/2}(n-1)$, and we have

$$\begin{aligned} 1 - \alpha &= \mathbf{P}(\frac{a}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{b}{(n-1)S^2}) \\ &= \mathbf{P}(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}) \end{aligned}$$

The probability that the random interval $[(n-1)S^2/b, (n-1)S^2/a]$ contains the unknown σ^2 is $1 - \alpha$.

Example: CI for variance

$X_1, \cdots, X_{13} \sim \mathcal{N}(\mu, \sigma^2)$, we have $\overline{x} = 18.97$ and $\sum_{i=1}^{13}(x_i - \overline{x})^2 = 128.41$ find the 90% CIs for σ^2.

From chi-squared table we have $\chi^2_{0.95}(12) = 5.226$ and $\chi^2_{0.05}(12) = 21.03$ (5 quantile and 95 quantile from a chi-squared distribution with 12 degrees of freedom respectively).

A 90% CIs for σ^2 is

$$[\ \frac{128.4}{21.03}, \frac{128.4}{5.226}\ ] = [6.11, 24.57]$$

Example

Given $X_1, \cdots, X_n$ is a random sample from a Exponential(λ) distribution (mean=$1/\lambda$).

1. Let $W = 2\lambda \sum_{i=1}^{n} X_i$, show $W \sim \chi^2(2n)$ (hint: use Moment generating function)
2. Find a 90% CIs for λ.

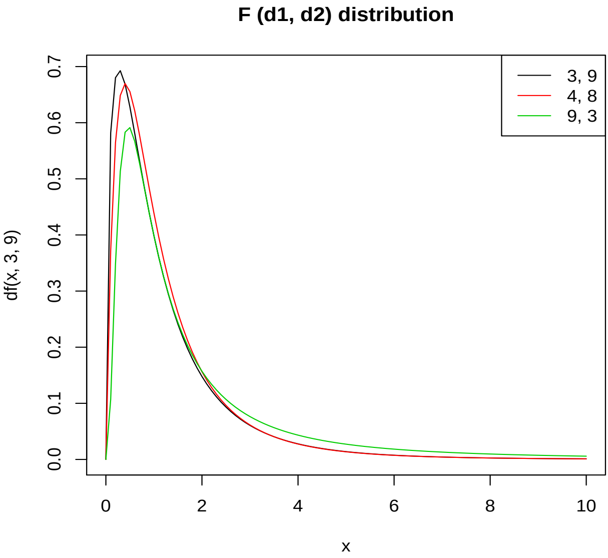
CI for the ratio of variances

Recall: F distribution

$W_1 \sim \chi^2(v_1)$, $W_2 \sim \chi^2(v_2)$ and W_1, W_2 are independent random variables. Then a random variable F which can be expressed as

$$F = \frac{W_1/v_1}{W_2/v_2}$$

is said to be distributed as a F distribution with degrees of freedom v_1 and v_2, denoted by $F(v_1, v_2)$ or F_{v_1,v_2}


F (d1, d2) distribution
3, 9
4, 8
9, 3
df(x, 3, 9)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10

F-distribution

Reciprocal of an F

Let the r.v. $F \sim F(v_1, v_2)$ and let $Y = 1/F$. Then Y has a pdf.

$$
\begin{aligned}
f(y)^* &= g(F)\left|\frac{dF}{dy}\right| \\
&= \frac{v_1^{v_1/2} y^{1-(v_1/2)} v_2^{v_2/2} y^{(v_1+v_2)/2}}{B(\frac{v_1}{2}, \frac{v_2}{2})(v_2 y + v_1)^{(v_1+v_2)/2}} \frac{1}{y^2} \\
&= \frac{v_1^{v_1/2} v_2^{v_2/2} y^{(v_2/2)-1}}{B(\frac{v_2}{2}, \frac{v_1}{2})(v_1 + v_2 y)^{(v_1+v_2)/2}} \quad y \in [0, \infty)
\end{aligned}
$$

That is if $F \sim F(v_1, v_2)$ and $Y = 1/F$, then $Y \sim F(v_2, v_1)$

CI for σ_X^2/σ_Y^2 from two ind. Normal

Given S_X^2, S_Y^2 are unbiased estimates of σ_X^2, σ_Y^2 derived from samples of size n and m , respectively, from two independent normal populations. Find a $100(1-\alpha)\%$ CI for σ_X^2/σ_Y^2.

CI for σ_X^2/σ_Y^2 from two ind. Normal

Given S_X^2, S_Y^2 are unbiased estimates of σ_X^2, σ_Y^2 derived from samples of size n and m , respectively, from two independent normal populations. Find a $100(1-\alpha)\%$ CI for σ_X^2/σ_Y^2.
$(n-1)S_X^2/\sigma_X^2 \sim \chi^2(n-1)$, $(m-1)S_Y^2/\sigma_Y^2 \sim \chi^2(m-1)$

$$\frac{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)}{\frac{(n-1)S_1^2}{\sigma_1^2}/(n-1)} = \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2}$$

follow a F distribution with degrees of freedom $(m-1)$ and $(n-1)$ i.e.,

$$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim F(m-1, n-1)$$

$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim F(m-1, n-1)$

So we select $c = F_{1-\alpha/2}(m-1, n-1)$ and $d = F_{\alpha/2}(m-1, n-1)$, and

$$\mathbf{P}(c \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq d) = 1 - \alpha$$

That is

$$\mathbf{P}\big(c \ \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq d \ \frac{S_X^2}{S_Y^2}\big) = 1 - \alpha$$

$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim F(m-1, n-1)$

So we select $c = F_{1-\alpha/2}(m-1, n-1)$ and $d = F_{\alpha/2}(m-1, n-1)$, and

$$\mathbf{P}\left(c \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq d\right) = 1-\alpha$$

That is

$$\mathbf{P}\left(c\ \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq d\ \frac{S_X^2}{S_Y^2}\right) = 1-\alpha$$

Often from table we have
$c = F_{1-\alpha/2}(m-1, n-1) = 1/F_{\alpha/2}(n-1, m-1)$ and $d = F_{\alpha/2}(m-1, n-1)$, let s_x^2 and s_y^2 be the realization of S_X^2 and S_Y^2, then a $100(1-\alpha)\%$ CIs for σ_X^2/σ_Y^2 is

$$\left[\ \frac{1}{F_{\alpha/2}(n-1, m-1)}\frac{s_X^2}{s_Y^2},\ F_{\alpha/2}(m-1, n-1)\frac{s_X^2}{s_Y^2}\ \right]$$

Example

From two ind Normal with unknown means and variances, we have $(12)s_X^2 = 128.4$ from a random sample of size 13 and $(8)s_Y^2 = 36.72$ from a random sample of size 9. Find a 98% CIs for σ_X^2/σ_Y^2.

Example

From two ind Normal with unknown means and variances, we have $(12)s_X^2 = 128.4$ from a random sample of size 13 and $(8)s_Y^2 = 36.72$ from a random sample of size 9. Find a 98% CIs for σ_X^2/σ_Y^2.

$$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim F(8, 12)$$

From F-table we have $F_{0.01}(12, 8) = 5.67$ and $F_{0.01}(8, 12) = 4.50$, so a 98% CIs for σ_X^2/σ_Y^2 is

$$[\ (\frac{1}{5.67})\frac{128.4/12}{36.72/8}, (4.50)\frac{128.4/12}{36.72/8}\ ]$$

Confidence intervals for proportions (p)

Estimate proportions. Construct a CI for p in the $\text{Bin}(n, p)$ distribution.

Assume that sampling is from a binomial population and hence that the problem is to estimate p in the $\text{Bin}(n, p)$ distribution where p is unknown.

recall:

Given Y is distributed as $\text{Bin}(n, p)$, an unbiased estimate of p is $\hat{p} = \frac{Y}{n}$.

$$E(\hat{p}) = E(\frac{Y}{n}) = p$$

and

$$Var(\hat{p}) = \frac{1}{n^2}\text{Var}(Y) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Confidence intervals for proportions (p)

For large n,

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

can be approximated by the standard normal $\mathcal{N}(0, 1)$.

Thus an approximate $100(1-\alpha)\%$ CI for p is obtained by considering

$$\mathbf{P}\left(-z_{\alpha/2} < \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) = 1 - \alpha$$

Confidence intervals for proportions (p)

For large n,

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{(Y/n) - p}{\sqrt{p(1-p)/n}}$$

can be approximated by the standard normal $\mathcal{N}(0, 1)$.

Thus an approximate $100(1-\alpha)\%$ CI for p is obtained by considering

$$\mathbf{P}(-z_{\alpha/2} < \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}) = 1 - \alpha$$

Replace the variance of $\hat{p} = Y/n$ by its estimate $\hat{p}(1-\hat{p})/n$, giving a simple expression for the CI for p is

$$[\hat{p} \ \pm \ z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] = [\frac{Y}{n} \ \pm \ z_{\alpha/2}\sqrt{\frac{(Y/n)(1-Y/n)}{n}}]$$

Example

Assume $Y \sim \text{Bin}(n, p)$, we have $n = 36$ and $y/n = 0.222$, find an approximate 90% CIs for p

Poll $n = 351$ and $y = 185$ say yes, find 95% CI for p?

Example

Assume $Y \sim \text{Bin}(n, p)$, we have $n = 36$ and $y/n = 0.222$, find an approximate 90% CIs for p

$$[\ 0.222 \pm 1.645\sqrt{\frac{(0.222)(1-0.222)}{36}}\ ]$$

Poll $n = 351$ and $y = 185$ say yes, find 95% CI for p?

Example

Assume $Y \sim \text{Bin}(n, p)$, we have $n = 36$ and $y/n = 0.222$, find an approximate 90% CIs for p

$$[\ 0.222 \pm 1.645\sqrt{\frac{(0.222)(1-0.222)}{36}}\ ]$$

Example

Poll $n = 100$ and $y = 51$ say yes, find 95% CI for p .41, 0.61

Poll $n = 351$ and $y = 185$ say yes, find 95% CI for p?

CI for difference of two proportions

民調的解讀

「....施政滿意度4成4。本次調查是以台灣地區住宅電話簿為抽樣清冊， 並以電話的後四碼進行隨機抽樣。共成功訪問1056位台灣地區20歲以上民眾。在95%的信心水準下，抽樣誤差為正負3.0百分點。

1. 這項民調的母體是什麼？樣本數為多少？
2. 受訪民眾中對施政滿意約有多少人？
3. 算出這次調查的信賴區間？

民調的解讀

1. 在本次調查中，母體是台灣地區20歲以上的民眾，樣本則是成功訪問的1056人，「滿意度4成4」表示在1056位受訪者中，約有44%的人表示滿意(即約有456人回答滿意
2. 區間$[0.44-0.03, 0.44+0.03]=[0.41, 0.47]$，稱為信賴區間 (信賴區間：[估計值-最大誤差，估計值+最大誤差])
   假設母體真正的滿意比例是p,這次的調查推估p的值可能會落在0.41到0.47的範圍內。
3. 95%的信心水準: p是不可知的，而抽樣都會有誤差，並不能保證真正的比例p一定會在我們所推估的區間內。「如果我們抽樣很多次，每次都會得到一個信賴區間，那麼這麼多的信賴區間中，約有95%的區間會涵蓋真正的p值。
4. 而我們有95%的信心說，真正的滿意度會落在我們所得出的區間中。

Sample Size for proportion

某報對於台北市市長施政滿意程度進行民調，民調結果如下： 「滿意度為六成三，本次民調共成功訪問n位台北市20歲以上的成年民眾，在95%的信心水準下，抽樣誤差為正負3.2百分點。」 求n?

Sample Size for proportion

某報對於台北市市長施政滿意程度進行民調，民調結果如下：「滿意度為六成三，本次民調共成功訪問n位台北市20歲以上的成年民眾，在95%的信心水準下，抽樣誤差為正負3.2百分點。」 求n?

$z_{0.025} = 1.96$ and $(1.96)\sqrt{\frac{(0.63)(1-0.63)}{n}} = 0.032$
we have $n = (0.63)(0.37)(1.96)^2/0.032^2 = 864$

Sample Size for proportion

某報對於台北市市長施政滿意程度進行民調，民調結果如下：「滿意度為六成三，本次民調共成功訪問n位台北市20歲以上的成年民眾，在95%的信心水準下，抽樣誤差為正負3.2百分點。」 求n?

$z_{0.025} = 1.96$ and $(1.96)\sqrt{\frac{(0.63)(1-0.63)}{n}} = 0.032$
we have $n = (0.63)(0.37)(1.96)^2/0.032^2 = 864$

The maximum error of the estimate for 98% confidence coefficient is 0.01 for $\hat{p} = 0.08$, find the n

Sample Size for proportion

某報對於台北市市長施政滿意程度進行民調，民調結果如下：「滿意度為六成三，本次民調共成功訪問n位台北市20歲以上的成年民眾，在95%的信心水準下，抽樣誤差為正負3.2百分點。」求n?

$z_{0.025} = 1.96$ and $(1.96)\sqrt{\frac{(0.63)(1-0.63)}{n}} = 0.032$
we have $n = (0.63)(0.37)(1.96)^2/0.032^2 = 864$

The maximum error of the estimate for 98% confidence coefficient is 0.01 for $\hat{p} = 0.08$, find the n

$z_{0.01} = 2.326$ and $(2.326)\sqrt{\frac{(0.08)(0.92)}{n}} = 0.01$
we have $n = (0.08)(0.92)(2.326)^2/0.01^2 = 3982$

unknown $\hat{p}$

For estimating p, we have $p^*(1-p^*) \leq 1/4$. Hence we need

$$n = 2.326^2/(4 * 0.01^2) = 13530$$

for the maximum error of the estimate for 98% confidence coefficient is 0.01.

95% confidence coefficient for $\epsilon = 0.01$, we have $n = 9604$

95% confidence coefficient for $\epsilon = 0.03$, we have $n = 1067$

Sample Size and CIs for given $\hat{p}$

The 95% CI for the proportion of people of supporting A when there is 51% people support A in polls of 100, 400 or 10,000 sample.

[0.41, 0.61], [0.46, 0.56], [0.50, 0.52]

$[0.51 \pm 0.1], [0.51 \pm 0.05], [0.51 \pm 0.01]$

Sample Size for mean

$100(1-\alpha)\%$ CI for μ is $[\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})]$. Denote such interval as $\overline{x} \pm \epsilon$. we sometime call $\epsilon = z_{\alpha/2}(\sigma/\sqrt{n})$ the maximum error of the estimate

$$n = \frac{(z_{\alpha/2})^2(\sigma)^2}{\epsilon^2}$$

where it is assumed that σ^2 is known.

Example

we want the 95% CIs for μ to be $\overline{x} \pm 1$ for a normal population with standard deviation $\sigma = 15$, find the sample size.

Example

we want the 95% CIs for μ to be $\overline{x} \pm 1$ for a normal population with standard deviation $\sigma = 15$, find the sample size.

$$1.96\left(\frac{15}{\sqrt{n}}\right) = 1$$

we have $n \approx 864.35 = 865$.

Example

we want the 95% CIs for μ to be $\overline{x} \pm 1$ for a normal population with standard deviation $\sigma = 15$, find the sample size.

$$1.96\left(\frac{15}{\sqrt{n}}\right) = 1$$

we have $n \approx 864.35 = 865$.

The 80% CIs for μ is $\overline{x} \pm 2$, then we have

$$1.282\left(\frac{15}{\sqrt{n}}\right) = 2$$

where $z_{0.1} = 1.282$ and thus $n = 93$

Pivotal quantity I

A very useful method for finding confidence intervals uses a pivotal quantity.

What is a **pivotal quantity**? A pivotal quantity is a function of data and the unknown parameter, say $g(\mathbf{X}, \theta)$, and the distribution of $g(\mathbf{X}, \theta)$ does not depend on the unknown parameter.

Example

Given $X_1, \cdots, X_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

Pivotal quantity II

1. When σ is known, $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is a pivotal quantity.
   $Z \sim \mathcal{N}(0, 1)$
2. When σ is unknown, $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ is a pivotal quantity where S is the sample deviation. $T \sim t(n-1)$
3. $W = (n-1)S^2/\sigma^2$ is a pivotal quantity. $W \sim \chi^2(n-1)$

$$Y \sim \mathrm{Bin}(n, p),\ \frac{(Y/n) - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1)$$