

Interval Estimation

It is a common requirement to efficiently estimate population parameters based on simple random sample data. In the R tutorials of this section, we demonstrate how to compute the estimates.

The steps are to be illustrated with a built-in data frame named `survey`. It is the outcome of a Statistics student survey in an Australian university.

The data set belongs to the MASS package, which has to be pre-loaded into the R workspace prior to use.

```
> library(MASS)           # load the MASS package
```

```
> head(survey)
  Sex  Wr.Hnd NW.Hnd ...
1 Female 18.5   18.0 ...
2  Male 19.5   20.5 ...
3  Male 18.0   13.3 ...
.....
```

For further details of the survey data set, please consult the R documentation.

```
> help(survey)
```

Point Estimate of Population Mean

For any particular random sample, we can always compute its sample mean. Although most often it is not the actual population mean, it does serve as a good point estimate.

For example, in the data set `survey`, the survey is performed on a sample of the student population. We can compute the sample mean and use it as an estimate of the corresponding population parameter.

Problem

Find a point estimate of mean university student height with the sample data from `survey`.

Solution

For convenience, we begin with saving the survey data of student heights in a variable `height.survey`.

```
> library(MASS)                # load the MASS package  
> height.survey = survey$Height
```

It turns out not all students have answered the question, and we must filter out the missing values. Hence we apply the `mean` function with the `"na.rm"` argument as `TRUE`.

```
> mean(height.survey)
```

```
[1] NA
```

```
> mean(height.survey, na.rm=TRUE) # skip missing values
```

```
[1] 172.38
```

Answer

A point estimate of the mean student height is 172.38 centimeters.

Interval Estimate of Population Mean with Known Variance

After we found a point estimate of the population mean, we would need a way to quantify its accuracy. Here, we discuss the case where the population variance σ^2 is assumed known.

Let us denote the $100(1 - \alpha/2)$ percentile of the standard normal distribution as $z_{\alpha/2}$. For random sample of sufficiently large size, the end points of the interval estimate at $(1 - \alpha)$ confidence level is given as follows:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Problem

Assume the population standard deviation σ of the student height in survey is 9.48. Find the margin of error and interval estimate at 95% confidence level.

Solution

We first filter out missing values in survey\$Height with the na.omit function, and save it in height.response.

```
> library(MASS) # load the MASS package
> height.response = na.omit(survey$Height)
```

Then we compute the standard error of the mean.

```
> n = length(height.response)
> sigma = 9.48 # population standard deviation
> sem = sigma/sqrt(n); sem # standard error of the mean
[1] 0.65575
```

Since there are two tails of the normal distribution, the 95% confidence level would imply the 97.5th percentile of the normal distribution at the upper tail. Therefore, $z_{\alpha/2}$ is given by `qnorm(.975)`. We multiply it with the standard error of the mean `sem` and get the margin of error.

```
> E = qnorm(.975)*sem; E          # margin of error
[1] 1.2852
```

We then add it up with the sample mean, and find the confidence interval as told.

```
> xbar = mean(height.response)    # sample mean
> xbar + c(-E, E)
[1] 171.10 173.67
```

Answer

Assuming the population standard deviation σ being 9.48, the margin of error for the student height survey at 95% confidence level is 1.2852 centimeters. The confidence interval is between 171.10 and 173.67 centimeters.

Alternative Solution

Instead of using the textbook formula, we can apply the `z.test` function in the `TeachingDemos` package. It is not a core R package, and must be installed and loaded into the workspace beforehand.

```
> library(TeachingDemos)          # load TeachingDemos package
> z.test(height.response, sd=sigma)
```

One Sample z-test

```
data: height.response
z = 262.88, n = 209.000, Std. Dev. = 9.480,
Std. Dev. of the sample mean = 0.656, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 171.10 173.67
sample estimates:
mean of height.response
      172.38
```

Interval Estimate of Population Mean with Unknown Variance

After we found a point estimate of the population mean, we would need a way to quantify its accuracy.

Here, we discuss the case where the population variance is not assumed.

Let us denote the $100(1 - \alpha/2)$ percentile of the Student t distribution with $n - 1$ degrees of freedom as $t_{\alpha/2}$. For random samples of sufficiently large size, and with standard deviation s , the end points of the interval estimate at $(1 - \alpha)$ confidence level is given as follows:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Problem

Without assuming the population standard deviation of the student height in survey, find the margin of error and interval estimate at 95% confidence level.

Solution

We first filter out missing values in `survey$Height` with the `na.omit` function, and save it in `height.response`.

```
> library(MASS)                # load the MASS package
> height.response = na.omit(survey$Height)
```

Then we compute the sample standard deviation.

```
> n = length(height.response)
> s = sd(height.response)      # sample standard deviation
> SE = s/sqrt(n); SE          # standard error estimate
[1] 0.68117
```

Since there are two tails of the Student t distribution, the 95% confidence level would imply the 97.5th percentile of the Student t distribution at the upper tail. Therefore, $t_{\alpha/2}$ is given by `qt(.975, df=n-1)`. We multiply it with the standard error estimate `SE` and get the margin of error.

```
> E = qt(.975, df=n-1)*SE; E      # margin of error
[1] 1.3429
```

We then add it up with the sample mean, and find the confidence interval.

```
> xbar = mean(height.response)    # sample mean
> xbar + c(-E, E)
[1] 171.04 173.72
```

Answer

Without assumption on the population standard deviation, the margin of error for the student height survey at 95% confidence level is 1.3429 centimeters. The confidence interval is between 171.04 and 173.72 centimeters.

Alternative Solution

Instead of using the textbook formula, we can apply the `t.test` function in the built-in stats package.

```
> t.test(height.response)
```

One Sample t-test

```
data: height.response
t = 253.07, df = 208, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 171.04 173.72
sample estimates:
mean of x
 172.38
```

Sampling Size of Population Mean

The quality of a sample survey can be improved by increasing the sample size.

The formula below provide the sample size needed under the requirement of population mean interval estimate at $(1 - \alpha)$ confidence level, margin of error E , and population variance σ^2 .

Here, $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

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

Problem

Assume the population standard deviation σ of the student height in survey is 9.48. Find the sample size needed to achieve a 1.2 centimeters margin of error at 95% confidence level.

Solution

Since there are two tails of the normal distribution, the 95% confidence level would imply the 97.5th percentile of the normal distribution at the upper tail. Therefore, $z_{\alpha/2}$ is given by `qnorm(.975)`.

```
> zstar = qnorm(.975)
> sigma = 9.48
> E = 1.2
> zstar^2 * sigma^2 / E^2
[1] 239.75
```

Answer

Based on the assumption of population standard deviation being 9.48, it needs a sample size of 240 to achieve a 1.2 centimeters margin of error at 95% confidence level.

Point Estimate of Population Proportion

Multiple choice questionnaires in a survey are often used to determine the the proportion of a population with certain characteristic.

For example, we can estimate the proportion of female students in the university based on the result in the sample data set survey.

Problem

Find a point estimate of the female student proportion from survey.

Solution

We first filter out missing values in survey\$Sex with the na.omit function, and save it in gender.response.

```
> library(MASS)                # load the MASS package
> gender.response = na.omit(survey$Sex)
> n = length(gender.response)    # valid responses count
```

To find out the number of female students, we compare gender.response with the factor 'Female', and compute the sum.

Dividing it by n gives the female student proportion in the sample survey.

```
> k = sum(gender.response == "Female")
> pbar = k/n; pbar
[1] 0.5
```

Answer

The point estimate of the female student proportion in survey is 50%.

Interval Estimate of Population Proportion

After we found a point sample estimate of the population proportion, we would need to estimate its confidence interval.

Let us denote the $100(1 - \alpha/2)$ percentile of the standard normal distribution as $z_{\alpha/2}$. If the samples size n and population proportion p satisfy the condition that $np \geq 5$ and $n(1 - p) \geq 5$, then the end points of the interval estimate at $(1 - \alpha)$ confidence level is defined in terms of the sample proportion as follows.

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

Problem

Compute the margin of error and estimate interval for the female students proportion in survey at 95% confidence level.

Solution

We first determine the proportion point estimate. Further details can be found in the previous tutorial.

```
> library(MASS) # load the MASS package
> gender.response = na.omit(survey$Sex)
> n = length(gender.response) # valid responses count
> k = sum(gender.response == "Female")
> pbar = k/n; pbar
[1] 0.5
```

Then we estimate the standard error.

```
> SE = sqrt(pbar*(1-pbar)/n); SE # standard error
[1] 0.032547
```

Since there are two tails of the normal distribution, the 95% confidence level would imply the 97.5th percentile of the normal distribution at the upper tail. Therefore, $z_{\alpha/2}$ is given by `qnorm(.975)`. Hence we multiply it with the standard error estimate SE and compute the margin of error.

```
> E = qnorm(.975)*SE; E           # margin of error
[1] 0.063791
```

Combining it with the sample proportion, we obtain the confidence interval.

```
> pbar + c(-E, E)
[1] 0.43621 0.56379
```

Answer

At 95% confidence level, between 43.6% and 56.3% of the university students are female, and the margin of error is 6.4%.

Alternative Solution

Instead of using the textbook formula, we can apply the prop.test function in the built-in stats package.

```
> prop.test(k, n)
```

```
1-sample proportions test without continuity
correction
```

```
data: k out of n, null probability 0.5
```

```
X-squared = 0, df = 1, p-value = 1
```

```
alternative hypothesis: true p is not equal to 0.5
```

```
95 percent confidence interval:
```

```
0.43672 0.56328
```

```
sample estimates:
```

```
p
0.5
```

Sampling Size of Population Proportion

The quality of a sample survey can be improved by increasing the sample size.

The formula below provide the sample size needed under the requirement of population proportion interval estimate at $(1 - \alpha)$ confidence level, margin of error E , and planned proportion estimate p . Here, $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

$$n = \frac{(z_{\alpha/2})^2 p (1 - p)}{E^2}$$

Problem

Using a 50% planned proportion estimate, find the sample size needed to achieve 5% margin of error for the female student survey at 95% confidence level.

Solution

Since there are two tails of the normal distribution, the 95% confidence level would imply the 97.5th percentile of the normal distribution at the upper tail. Therefore, $z_{\alpha/2}$ is given by `qnorm(.975)`.

```
> zstar = qnorm(.975)
> p = 0.5
> E = 0.05
> zstar^2 * p * (1-p) / E^2
[1] 384.15
```

Answer

With a planned proportion estimate of 50% at 95% confidence level, it needs a sample size of 385 to achieve a 5% margin of error for the survey of female student proportion.

Some Extra Insight:

Confidence interval isn't always right

The fact that not all confidence intervals contain the true value of the parameter is often illustrated by plotting a number of random confidence intervals at once and observing.

```
> m = 50; n=20; p = .5;                # toss 20 coins 50 times
> phat = rbinom(m,n,p)/n                # divide by n for proportions
> SE = sqrt(phat*(1-phat)/n)            # compute SE
> alpha = 0.10;zstar = qnorm(1-alpha/2)
> matplot(rbind(phat - zstar*SE, phat + zstar*SE), rbind(1:m,1:m),type="l",lty=1)
> abline(v=p)                           # draw line for p=0.5
```

Comparing p-values from t and z

One may be tempted to think that the confidence interval based on the t statistic would always be larger than that based on the z statistic as always $t^* > z^*$. However, the standard error SE for the t also depends on s which is variable and can sometimes be small enough to offset the difference.

To see why t^* is always larger than z^* , we can compare side-by-side boxplots of two random sets of data with these distributions.

```
> x=rnorm(100);y=rt(100,9)
> boxplot(x,y)
> qqnorm(x);qqline(x)
> qqnorm(y);qqline(y)
```

which gives (notice the symmetry of both, but the larger variance of the t distribution).

And for completeness, this creates a graph with several theoretical densities.

```
> xvals=seq(-4,4,.01)
> plot(xvals,dnorm(xvals),type="l")
> for(i in c(2,5,10,20,50)) points(xvals,dt(xvals,df=i),type="l",lty=i)
```

The R Stats Package

ansari.test	Performs the Ansari-Bradley two-sample test for a difference in scale parameters.
bartlett.test	Performs Bartlett's test of the null that the variances in each of the groups (samples) are the same.
binom.test	Performs an exact test of a simple null hypothesis about the probability of success in a Bernoulli experiment.
Box.test	Compute the Box–Pierce or Ljung–Box test statistic for examining the null hypothesis of independence in a given time series. These are sometimes known as ‘portmanteau’ tests.
chisq.test	chisq.test performs chi-squared contingency table tests and goodness-of-fit tests.
cor.test	Test for association between paired samples, using one of Pearson's product moment correlation coefficient, Kendall's tau or Spearman's rho.
fisher.test	Performs Fisher's exact test for testing the null of independence of rows and columns in a contingency table with fixed marginals.
fligner.test	Performs a Fligner-Killeen (median) test of the null that the variances in each of the groups (samples) are the same.
friedman.test	Performs a Friedman rank sum test with unreplicated blocked data.
kruskal.test	Performs a Kruskal-Wallis rank sum test.
ks.test	Performs one or two sample Kolmogorov-Smirnov tests.
mantelhaen.test	Performs a Cochran-Mantel-Haenszel chi-squared test of the null that two nominal variables are conditionally independent in each stratum, assuming that there is no three-way interaction.
mauchly.test	Tests whether a Wishart-distributed covariance matrix (or transformation thereof) is proportional to a given matrix.
mcnemar.test	Performs McNemar's chi-squared test for symmetry of rows and columns in a two-dimensional contingency table.
mood.test	Performs Mood's two-sample test for a difference in scale parameters.
oneway.test	Test whether two or more samples from normal distributions have the same means. The variances are not necessarily assumed to be equal.
poisson.test	Performs an exact test of a simple null hypothesis about the rate parameter in Poisson distribution, or for the ratio between two rate parameters.

PP.test	Computes the Phillips-Perron test for the null hypothesis that x has a unit root against a stationary alternative.
prop.test	can be used for testing the null that the proportions (probabilities of success) in several groups are the same, or that they equal certain given values.
quade.test	Performs a Quade test with unreplicated blocked data.
shapiro.test	Performs the Shapiro-Wilk test of normality.
t.test	Student's t-Test
var.test	Performs an F test to compare the variances of two samples from normal populations.
wilcox.test	Performs one- and two-sample Wilcoxon tests on vectors of data; the latter is also known as 'Mann-Whitney' test.