CSMMA16

Introduction to R: Probability and Statistics

Aim

The aim of this session is to continue your introduction to R, specifically considering aspects of probability and statistics. Refer to the handout on this topic.

Generating random data

You can use R to generate random data from all the common distributions. To generate normally distributed data use the`rnorm` function, for uniformly distributed data use `runif` and for binomial data `rbinom`. To generate a random sample $\boldsymbol{x}$ of n iid normally distributed random variables with mean μ and standard deviation σ, the R command is

```
> x <- rnorm(n,mu,sigma)
```

Randomly generate 1000 numbers from a normal distribution with a mean of 10 and a standard deviation of 0.5 and save it in the variable x. Plot the data using

```
> plot(x)
```

Is this what you would expect? Calculate the mean and standard deviation in R. Are they the same as the theoretical mean and standard deviation? Explain why. Next, plot a histogram of x. The data should be approximately normally distributed.

Write a function to generate m datasets, each of sample size nn, from a binomial distribution with $p = 0.01$ and trial size 10 and to calculate the mean of each dataset. Do not write your own loop in R; instead use the `apply` function. Use this function in the `apply` function to generate 10000 means each from studies with sizes $nn =$
(a) 10 and (b) 100, respectively.
Plot histograms of the sample means for each sample size so that the two histograms are in the same figure. Use the Central Limit Theorem to explain your findings.

Effect of Variance

This exercise aims to demonstrate the effect of variance in the data on our ability to detect significance.

Generate a random sample of size 25 from a normal distribution with mean of 1.1 and a standard deviation of 0.8. Use the R command `t.test` to test the hypothesis that the data comes from a normal distribution with a mean of 1, even though you know the theoretical mean is 1.1. Record your findings. Repeat the exercise but replace the standard deviation by 0.1. Compare the two cases and comment on the effect of the variance in the data on hypothesis testing.

Understanding a confidence interval

Write a function that uses the R command `t.test` to calculate and return a 95% confidence interval for the mean μ based on a randomly generated sample of size 25 from a normal distribution with mean of $\mu = 10$ and a standard deviation of $\sigma = 1$. Use this function to generate 100 confidence intervals and to count how often the intervals contain the true mean μ. Comment on your results.

Paired t-test vs two-sample t-test

This exercise aims to illustrate the importance of selecting the correct t-test when comparing two sets of data.

Generate a random sample $\boldsymbol{x}$ of size 20 from a normal distribution with mean of 100 and standard deviation of 4. Then generate a variable $y = 2 + x + \epsilon$, where ϵ represents noise on the data which is normally distributed with a mean of zero and a standard deviation of 1. You have just generated 20 pairs of data, with each element in $\boldsymbol{y}$ associated with the corresponding element in $\boldsymbol{x}$ by adding a number 2 with some uncertainty due to noise.

Plot $\boldsymbol{x}$ and $\boldsymbol{y}$ superimposed with each other in R and draw lines showing their means. In order to get all points on the graph you may need to use `min` and `max` functions to define the plot area.

Use the R command to test if the means of X and Y are significantly different at 5% level of significance. Now use `t.test` with `paired=TRUE` for the same hypothesis test. Discuss your results.