

Chapter 2

2. Probability and sampling distributions in R

Objectives

After completing this unit, students will be able to:

- Use R as a set of statistical tables
- Examine the distribution of a set of data
- carryout simulation for estimation and hypothesis test,
- Perform one and two sample tests in R

Probability functions in R

The following table describes functions related to probability distributions.

Function	Description
dnorm(x)	normal density function (by default m=0 sd=1) # plot standard normal curve x <- seq(-3, 3, 0.2) y <- dnorm(x) plot(x, y, type='l', xlab="Normal Deviate", ylab="Density")
pnorm(q)	cumulative normal probability for q (area under the normal curve to the left of q) pnorm(-0.35) # the area under the standard normal curve to the left of -0.35 pnorm(0)-pnorm(-1.45) # $p(-1.45 < z < 0)$ pnorm(87.2, mean=80, sd=4.8) # $p(x < 87.2)$ $x \sim N(\mu=80, sd=4.8)$
qnorm(p)	normal quantile. value at the p percentile of normal distribution qnorm(.9) # is 1.28 90th percentile qnorm(0.025, lower.tail=FALSE) # is 1.959964. i.e, $Z_{0.025}$ (the value of Z under the standard normal curve the area to its right is 0.025)
rnorm(n, m=0,sd=1)	n random normal deviates with mean m and standard deviation sd. #50 random normal variates with mean=50, sd=10 x <- rnorm(50, m=50, sd=10)

dbinom(x, size, prob) pbinom(q, size, prob) qbinom(p, size, prob) rbinom(n, size, prob)	binomial distribution where <i>size</i> is the sample size and <i>prob</i> is the probability of a heads (<i>p</i>) # prob of 0 to 5 heads of fair coin out of 10 flips (i.e, $p(x \leq 5)$) <code>sum(dbinom(0:5, 10, 0.5))</code> # prob of 5 or less heads of fair coin out of 10 flips <code>pbinom(5, 10, .5)</code>
dpois(x, lamda) ppois(q, lamda) qpois(p, lamda) rpois(n, lamda)	poisson distribution with $m=std=lamda$ #probability of 0,1, or 2 events with $lamda=4$ <code>dpois(0:2, 4)</code> # probability of at least 3 events with $lamda=4$ (i.e, $p(x \geq 3)$) <code>1- ppois(2,4)</code>
dunif(x, min=0, max=1) punif(q, min=0, max=1) qunif(p, min=0, max=1) runif(n, min=0, max=1)	uniform distribution, follows the same pattern as the normal distribution above. #10 uniform random variates <code>x <- runif(10)</code>

2.1. R as a set of statistical tables

One convenient use of R is to provide a comprehensive set of statistical tables. Functions are provided to evaluate the cumulative distribution function $P(X \leq x)$, the probability density function and the quantile function (given q , the smallest x such that $P(X \leq x) > q$), and to simulate from the distribution.

Here are some common statistical distribution with their R name and additional arguments

Distribution	R name	additional arguments
Binomial	Binom	<i>size, prob</i>
chi-squared	Chisq	<i>df, ncp</i>
Exponential	Exp	<i>Rate</i>
F	F	<i>df1, df2, ncp</i>
Logistic	Logis	<i>location, scale</i>
Normal	Norm	<i>mean, sd</i>
Poisson	Pois	<i>Lambda</i>
Student's t	T	<i>df, ncp</i>
Uniform	Unif	<i>min, max</i>

The functions available for each distribution follow this format:

Name	description
d name()	density or probability function
p name()	cumulative density function
q name()	quantile function
r name()	random deviates

quantiles

The quantile function is the inverse of the cumulative distribution function. The **p**-quantile is the value with the property that there is probability **p** of getting a value less than or equal to it. The median is by definition the 50% quantile.

Tables of statistical distributions are almost all given in terms of *quantiles*.

Z- table

To find the value of $Z_{0.025}$

```
qnorm(0.025, lower.tail = FALSE)
qnorm(1-0.025) # alternatively,
```

t- table: values of t exceeded with given probability

To find the value of $t_{0.025}(10)$

```
qt(0.025, 10, lower.tail=FALSE)
```

F - Table

To find the value of $F_{0.025}(2, 7)$: upper 2.5% point for an F(2, 7) distribution

```
qf(0.025, 2, 7, lower.tail = FALSE)
```

2.2. Examining the distribution of a set of data

Given a (univariate) set of data we can examine its distribution in a large number of ways. The simplest is to examine the numbers. Two slightly different summaries are given by summary and fivenum and a display of the numbers by stem (a “stem and leaf” plot).

```
> attach(faithful)
> summary(eruptions)
Min. 1st Qu. Median Mean 3rd Qu. Max.
```

```
1.600 2.163 4.000 3.488 4.454 5.100
> fivenum(eruptions)
[1] 1.6000 2.1585 4.0000 4.4585 5.1000
> stem(eruptions)
```

A stem-and-leaf plot is like a histogram, and R has a function `hist` to plot histograms.

```
> hist(eruptions)
## make the bins smaller, make a plot of density
> hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE)
> lines(density(eruptions, bw=0.1))
> rug(eruptions) # show the actual data points
```

2.4. One and Two-sample t-tests

The R function `t.test()` can be used to perform both one and two sample t-tests on vectors of data. The function contains a variety of options and can be called as follows:

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Here **x** is a numeric vector of data values and **y** is an optional numeric vector of data values. If **y** is excluded, the function performs a one-sample t-test on the data contained in **x**, if it is included it performs a two-sample t-tests using both **x** and **y**.

The option **mu** provides a number indicating the true value of the mean (or difference in means if you are performing a two sample test) under the null hypothesis. The option **alternative** is a character string specifying the alternative hypothesis, and must be one of the following: **"two.sided"** (which is the default), **"greater"** or **"less"** depending on whether the alternative hypothesis is that the mean is different than, greater than or less than **mu**, respectively. For example the following call:

```
t.test(x, alternative = "less", mu = 10)
```

It performs a one sample t-test on the data contained in **x** where the null hypothesis is that $\mu = 10$ and the alternative is that $\mu < 10$.

The option **paired** indicates whether or not you want a paired t-test (TRUE = yes and FALSE = no). If you leave this option out it defaults to FALSE.

The option **var.equal** is a logical variable indicating whether or not to assume the two variances as being equal when performing a two-sample t-test. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used. If you leave this option out it defaults to FALSE.

Finally, the option **conf.level** determines the confidence level of the reported confidence interval for μ in the one-sample case and $\mu_1 - \mu_2$ in the two-sample case.

i. One-sample t-tests

Example: An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were:

0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418

Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g?

Solution

Let μ be the mean level of Salmonella in all batches of ice cream. Here the hypothesis of interest can be expressed as:

$$H_0: \mu = 0.3$$

$$H_1: \mu > 0.3$$

Hence, we will need to include the options **alternative="greater"**, **mu=0.3**. Below is the relevant R-code:

```
x <- c(0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418)
```

```
t.test(x, alternative="greater", mu=0.3)
```

```
One Sample t-test
```

```
data: x
```

```
t = 2.2051, df = 8, p-value = 0.02927
```

```
alternative hypothesis: true mean is greater than 0.3
```

```
95 percent confidence interval:
```

```
0.3245133      Inf
```

```
sample estimates:
```

```
mean of x
```

```
0.4564444
```

From the output we see that the $p\text{-value} = 0.029 < 0.05$. The decision is reject the null hypothesis. Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3 MPN/g.

ii. Two-sample t-tests

Example: Six subjects were given a drug (treatment group) and an additional 6 subjects a placebo (control group). Their reaction time to a stimulus was measured (in ms). We want to perform a two-sample t-test for comparing the means of the treatment and control groups.

Let μ_1 be the mean of the population taking medicine and μ_2 the mean of the untreated population. Here the hypothesis of interest can be expressed as:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Here we will need to include the data for the treatment group in x and the data for the control group in y . We will also need to include the options **alternative="less"**, **mu=0**.

Finally, we need to decide whether or not the standard deviations are the same in both groups.

Below is the relevant R-code when assuming **equal standard deviation**:

```
Control <- c(91, 87, 99, 77, 88, 91)
Treat <- c(101, 110, 103, 93, 99, 104)
t.test(Control, Treat, alternative="less", var.equal=TRUE)

Two Sample t-test

data: Control and Treat
t = -3.4456, df = 10, p-value = 0.003136
alternative hypothesis: true difference in means is less than 0
```

Below is the relevant R-code when **not assuming equal variances**:

```
t.test(Control, Treat, alter = "less")

Welch Two Sample t-test

data: Control and Treat
t = -3.4456, df = 9.4797, p-value = 0.003391
alternative hypothesis: true difference in means is less than 0
```

Here the pooled t-test and the Welsh t-test give roughly the same results (p-value = 0.00313 and 0.00339, respectively).

Comparison of variances

Even though it is possible in R to perform the two sample t-test without the assumption that the variances are the same, you may still be interested in testing that assumption, and R provides the **var.test** function for that purpose, implementing an F test on the ratio of the group variances.

```
var.test(Treat, Control)
```

```
      F test to compare two variances

data:  Treat and Control
F = 0.62038, num df = 5, denom df = 5, p-value = 0.6131
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08680986 4.43344648
sample estimates:
ratio of variances
      0.6203764
```

It is seen that the test is not significant, so there is no evidence against the assumption of the variances being identical. i.e, there is evidence that the assumption of equal variances

iii. Paired t-tests

There are many experimental settings where each subject in the study is in both the treatment and control group. For example, in a matched pairs design, subjects are matched in pairs and different treatments are given to each subject in the pair. The outcomes are thereafter compared pair-wise. Alternatively, one can measure each subject twice, before and after a treatment. In either of these situations we can't use two-sample t-tests since the independence assumption is not valid. Instead we need to use a paired t-test. This can be done using the option **paired=TRUE**.

Example: A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the

same cars using the other kind of gasoline. We use a paired t-test to determine whether cars get significantly better mileage with premium gas.

Below is the relevant R-code:

```
reg <- c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
prem <- c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
t.test(prem, reg, alternative="greater", paired=TRUE)
```

Paired t-test

```
data:  prem and reg
t = 4.4721, df = 9, p-value = 0.0007749
alternative hypothesis: true difference in means is greater than 0
```

The results show that the t-statistic is equal to 4.47 and the p-value is 0.00075. Since the p-value $< \alpha$ -value = 0.05, we reject the null hypothesis. There is strong evidence of a mean increase in gas mileage between regular and premium gasoline.

Summary of one and two sample t test

t-tests

The `t.test()` function produces a variety of t-tests. Unlike most statistical packages, the default assumes unequal variance and applies the Welch df modification.

Independent 2-sample t-test

```
t.test(y1,y2) # where y1 and y2 are numeric
```

Independent 2-group t-test

```
t.test(y~x) # where y is numeric and x is a binary factor
```

Paired t-test

```
t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric
```

Paired/dependent 2-group t-test

```
t.test(y~x, paired = TRUE) # where y is numeric and x is a binary factor
```

One sample t-test

```
t.test(y,mu=3) # Ho: mu=3
```

You can use the **var.equal = TRUE** option to specify equal variances and a pooled variance estimate. You can use the **alternative="less"** or **alternative="greater"** option to specify a one tailed test.

Self-test exercises

1. The results of two independent samples from two populations are listed below:

y1: 17, 19, 10, 29, 27, 21, 17, 17, 14, 20

y2: 26, 24, 26, 29, 15, 29, 31, 25, 18, 26

- a) Enter the data into R.
 - b) Show the Boxplot of the two samples side by side.
 - c) Test equality of two variances, assuming normality of the two samples. Use $\alpha = 0.05$.
 - d) Test equality of two means, provided that the two samples are from normal distribution. Assume the population variances are equal. Use $\alpha = 0.05$.
 - e) Find a 99% confidence interval for the difference in means. Assume the population variances are not equal
2. A dietitian wishes to see if a person's cholesterol level will change if the diet is supplemented by a certain mineral. Six subjects were pretested, and then they took the mineral supplement for a 6-week period. The results are shown in the table. (Cholesterol level is measured in milligrams per deciliter.) Can it be concluded that the cholesterol level has been changed at a $\alpha = 0.10$? Assume the variable is approximately normally distributed.

Subject	1	2	3	4	5	6
Before (x_1)	210	235	208	190	172	244
After (x_2)	190	170	210	188	173	228