

Introduction to R*

Section 5: Environments, Running R , Libraries and some probability distributions

Wim R.M. Cardoen

Last updated: 07/25/2023 @ 12:46:34

Contents

| | | |
|----------|--------------------------------------|----------|
| 1 | R Environments | 2 |
| 2 | Running R | 2 |
| 3 | R Packages | 2 |
| 3.1 | Installation of R packages | 2 |
| 3.2 | Using R packages | 2 |
| 4 | Probability distributions | 2 |
| 4.1 | Discrete distributions | 2 |
| 4.1.1 | Examples | 3 |
| 4.2 | Continuous distributions | 4 |
| 4.2.1 | Examples | 5 |
| 4.3 | Exercises | 6 |
| | Bibliography | 7 |

*© - Wim R.M. Cardoen, 2022 - The content can neither be copied nor distributed without the **explicit** permission of the author.

1 R Environments

under construction

2 Running R

under construction

3 R Packages

under construction

3.1 Installation of R packages

3.2 Using R packages

4 Probability distributions

R comes with the most important probability distributions installed. For the theoretical underpinnings, see e.g. (Casella & Berger, 2002).

Probability distributions can be grosso modo classified into:

- discrete distributions
- continuous distributions

4.1 Discrete distributions

Let $\mathbb{P}(X = k; \{\xi\})$ be a discrete probability mass function when the random variable $X = k$ and which depends on the parameter set $\{\xi\}$.

Let **keyword** be the (variable) name of the corresponding distribution. Then,

- **dkeyword**(k, \dots) : calculates the probability $\mathbb{P}(X = k)$
- **pkeyword**(k, \dots) : calculates the cumulative probability function (CDF) at k :
$$F(X = k; \{\xi\}) := \sum_{j=0}^k \mathbb{P}(X = j)$$
- **qkeyword**(p, \dots): calculates the value of k where $p = F(k; \{\xi\})$ or $k = \lceil F^{-1}(p; \{\xi\}) \rceil$
- **rkeyword**(n, \dots): generates a vector of n random values sampled from the distribution **keyword**.

Some common discrete probability distributions (implemented in R) are displayed in Table 1.

| keyword | Name | $\mathbb{P}(X = k; \{\xi\})$ | Parameter set ($\{\xi\}$) |
|---------|-------------------|--|---|
| binom | Binomial | $\binom{n}{k} p^k (1-p)^{n-k}$ | $0 \leq p \leq 1$ |
| nbinom | Negative Binomial | $\binom{k+r-1}{k} (1-p)^k p^r$ | $0 \leq p \leq 1; r > 0$ |
| hyper | Hypergeometric | $\frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$ | $N \in \{0, 1, 2, \dots\}; K, n \in \{0, 1, \dots, N\}$ |
| pois | Poisson | $\frac{\lambda^k e^{-\lambda}}{k!}$ | $0 < \lambda < \infty$ |

Table 1: A few common discrete probability distributions.

4.1.1 Examples

- Let's consider the following distribution: `binom(p = 0.3, n = 5)`.

```
n <- 5
p <- 0.3
# Alternative code for the binom distribution
mybinom <- function(n,p){
  v <- vector(mode="double", length=(n+1))
  for(k in 0:n){
    v[k+1] <- choose(n,k)*p^k*(1-p)^(n-k)
  }
  return(v)
}
pvec <- mybinom(n,p)
```

- Value of the PMF at $k = \{0, 1, \dots, 5\}$:

```
for(k in 0:n){
  cat(sprintf(" P(X=%d):%8.6f and should be %8.6f\n",
    k, dbinom(k, size=n, p), pvec[k+1]))
}
```

```
P(X=0):0.168070 and should be 0.168070
P(X=1):0.360150 and should be 0.360150
P(X=2):0.308700 and should be 0.308700
P(X=3):0.132300 and should be 0.132300
P(X=4):0.028350 and should be 0.028350
P(X=5):0.002430 and should be 0.002430
```

- Value of the CDF at $k = \{0, 1, \dots, 5\}$:

```
for(k in 0:n){
  cat(sprintf(" F(X=%d):%8.6f and should be %8.6f\n",
    k, pbinom(k, size=n, p), sum(pvec[1:(k+1)]) ))
}
```

```
F(X=0):0.168070 and should be 0.168070
```

```

F(X=1):0.528220 and should be 0.528220
F(X=2):0.836920 and should be 0.836920
F(X=3):0.969220 and should be 0.969220
F(X=4):0.997570 and should be 0.997570
F(X=5):1.000000 and should be 1.000000

```

– The quantile function:

```

pvec <- c(0.0, 0.25, 0.50, 0.75, 1.00)
for(item in pvec){
  cat(sprintf(" P:%4.2f => k=%d\n",
    item, qbinom(item,size=n, prob=p)))
}

```

```

P:0.00 => k=0
P:0.25 => k=1
P:0.50 => k=1
P:0.75 => k=2
P:1.00 => k=5

```

– Sampling random numbers from the distribution:

```

tot <- 15
vec <- rbinom(tot,size=n, prob=p)
print(vec)

```

```
[1] 2 1 2 1 1 1 3 4 1 1 2 0 0 1 2
```

4.2 Continuous distributions

Let $f(x; \{\xi\})$ be a continuous probability density function (pdf), which depends on the variable x and the parameter set $\{\xi\}$.

Let **keyword** be the (variable) name of the corresponding distribution. Then,

- **dkeyword**(x, \dots) : calculates the value of the pdf at x , i.e.
 $f(x; \{\xi\})$
- **pkeyword**(x, \dots) : calculates the cumulative probability function (cdf) at x :
$$F(x; \{\xi\}) := \int_{-\infty}^x f(t; \{\xi\}) dt$$
- **qkeyword**(p, \dots): calculates the value of x where $p = F(x; \{\xi\})$ or
 $x = F^{-1}(p; \{\xi\})$
- **rkeyword**(n, \dots): generates a vector of n random values sampled from the distribution **keyword**.

Some common continuous probability distributions (implemented in **R**) are displayed in Table 2.

| keyword | Name | $f(x; \{\xi\})$ | $\text{Dom}(x)$ | Parameter set ($\{\xi\}$) |
|---------------|-------------|--|------------------------|---|
| unif | Uniform | $\frac{1}{(b-a)}$ | $a \leq x \leq b$ | a, b |
| norm | Normal | $\frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ | $-\infty < x < \infty$ | $-\infty < \mu < \infty, \sigma > 0$ |
| cauchy | Cauchy | $\frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{x-\theta}{\sigma}\right)^2}$ | $-\infty < x < \infty$ | $-\infty < \theta < \infty, \sigma > 0$ |
| t | Student's t | $\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\nu\pi} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}$ | $-\infty < x < \infty$ | $\nu = 1, 2, \dots$ |
| chisq | Chi-squared | $\frac{1}{\Gamma(\nu/2)2^{(\nu/2)}} x^{(\nu/2)-1} e^{-\frac{x}{2}}$ | $0 \leq x < \infty$ | $\nu = 1, 2, \dots$ |
| f | F | $\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{(\nu_1-2)/2}}{\left(1 + \left(\frac{\nu_1}{\nu_2}\right)x\right)^{(\nu_1+\nu_2)/2}}$ | $0 \leq x < \infty$ | $\nu_1, \nu_2 = 1, 2, \dots$ |
| exp | Exponential | $\lambda e^{-\lambda x}$ | $0 \leq x < \infty$ | $\lambda > 0$ |

Table 2: A few common continuous probability distributions.

where $\Gamma(x)$ stands for the gamma function which has the following mathematical form:

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$$

4.2.1 Examples

- Let's consider the following distribution: $N(\mu = 5.0, \sigma^2 = 4.0)$.
Therefore, **distro:norm**

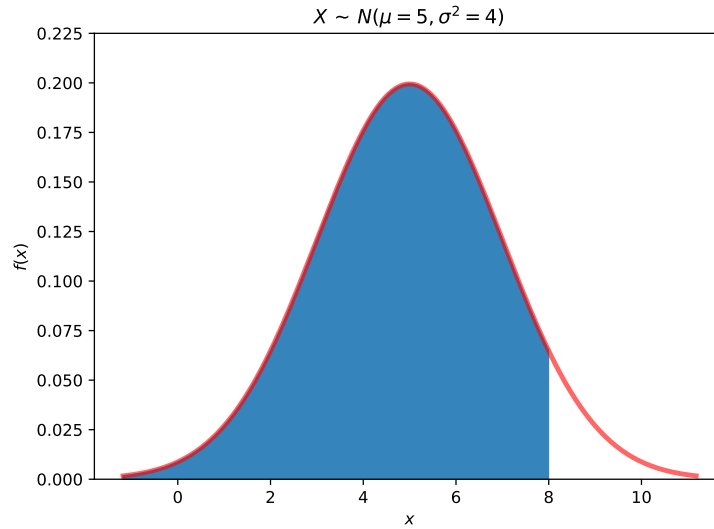


Figure 1: Plot of the normal distribution (red). The area under the curve (blue) represents the cumulative probability at $x = 8.0$.

```
x <- 8.0
mu <- 5.0
sigma <- 2.0
```

- Value of the PDF at x :

```
cat(sprintf("The density at %f is %12.10f\n", x, dnorm(x,mean=mu, sd=sigma)))
```

```
The density at 8.000000 is 0.0647587978
```

- Value of the CDF at x :

```
prob <- pnorm(x,mean=mu,sd=sigma)
cat(sprintf("The Cumulative Probability at %f is %12.10f", x, prob))
```

```
The Cumulative Probability at 8.000000 is 0.9331927987
```

- The quantile function:

```
cat(sprintf("The point where the Cumulative Probability is %12.10f: %8.4f",
            prob, qnorm(prob, mean=mu, sd=sigma)))
```

```
The point where the Cumulative Probability is 0.9331927987: 8.0000
```

- Sampling random numbers from the distribution:

```
vec <- rnorm(n=10, mean=mu, sd=sigma)
print(vec)
```

```
[1] 3.508793 7.898552 5.345508 4.602226 2.957217 6.018670 4.606252 2.411251
[9] 5.699139 6.586581
```

4.3 Exercises

1. Generate vectors with 10^4 , 10^5 , 10^6 , 10^7 random numbers from the $\chi^2(\nu = 5)$ distribution. Calculate the mean, as well as the variance for each of those vectors. (`mean()`, `var()`)
Note: If $X \sim \chi^2(\nu) \Rightarrow \mathbb{E}[X] = \nu$ and $\mathbb{V}[X] = 2\nu$
2. A brewer from the far-away lands of Hatu wants to follow the land's alcohol ordinance (i.e. a maximum of 5% ethanol per volume). In order to comply with the law he sent a batch of independent samples to a certified lab. The lab results are to be found in the file `data/beer.csv`.

His plan is to perform a simple one-sided hypothesis test:

$$H_0 : \mu_0 = 5.0$$

$$H_1 : \mu \neq \mu_0$$

He assumes that the alcohol % per volume is normally distributed (over the different samples/batches of beer) i.e. $N(\mu, \sigma^2)$ where μ is supposed to be 5.0 but where σ^2 is unknown.

Let $c_{1-\alpha}$ be the (critical) point that separates the acceptance region (\mathcal{A}) with $\mathbb{P}(\mathcal{A}) = 1 - \alpha$ from

the rejection region (\mathcal{R}) with $\mathbb{P}(\mathcal{R}) = \alpha$. Therefore,

$$\begin{aligned}\alpha &= \mathbb{P}(\bar{X} > c_{1-\alpha} | \mu = \mu_0) \\ &= \mathbb{P}\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > \frac{c_{1-\alpha} - \mu_0}{s/\sqrt{n}}\right) \\ &= \mathbb{P}\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > t_{1-\alpha}(\nu = n - 1)\right)\end{aligned}$$

where t stands for **Student's t** distribution, s^2 is the sample variance, n the number of measurements.

- Read the lab results from the file `data/beer.csv`.
- Calculate the numerical value (τ) of the test-statistic T , given by:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t(\nu = n - 1)$$

- Determine $t_{0.95}(\nu = n - 1)$, i.e. the critical point such that $\mathbb{P}(\mathcal{R}) = 0.05$.
 - Decide whether the brewer should reject H_0 (i.e. reject if $\tau > t_{0.95}(\nu = n - 1)$).
 - What is the probability of the area under the curve for $t \in [\tau, +\infty)$?
3. Check this [link](#) if you are interested in the origin of **Student's t** distribution.

Bibliography

Casella G. & Berger R.L. (2002). Statistical Inference. Duxbury Advanced Series in Statistics and Decision Sciences. Thomson Learning.