Probability Distributions

Giovanni Saraceno

Contents

Probability Distributions

The concepts of probability and randomness are central to statistics. In particular, to understand statistical methods, it is important to view data as samples derived from distributions.

This section outlines the basic ideas of probability and functions in `R` for random sampling and handling theoretical distributions.

Random Sampling

Early examples in probability theory primarily dealt with gambling and games where the core concept was random sampling, such as shuffling a deck of cards or drawing numbered balls. In `R`, we can simulate such situations using the `sample` function. For example, to draw 5 numbers (randomly!) from the set `1:90`, we can use:

```
sample(1:90, 5)
```

```
## [1] 56 75 13 67 27
```

The first argument, `x`, is a vector indicating the set to sample from, while `size` specifies the sample size. By default, the sample function samples **without replacement** (i.e., the sample cannot contain duplicate values), so the sample size cannot exceed the length of the set. To sample **with replacement**, use the option `replace = TRUE`. For example, to simulate 10 coin tosses:

```
sample(c("H", "T"), 10, replace = TRUE)
```

```
##  [1] "H" "T" "H" "T" "H" "T" "T" "T" "T" "T"
```

In a fair coin toss, the events "Heads" and "Tails" are equally likely (i.e., each has a probability of $\frac{1}{2}$). In R, we can also consider cases where the events are not equally probable using the `prob` option:

```
sample(c("H", "T"), 10, replace = TRUE, prob = c(0.9, 0.1))
```

```
##  [1] "H" "H" "H" "H" "T" "H" "H" "T" "H" "H"
```

Note: The sum of the values in the prob vector must equal 1.

Combinatorics

Consider the example of sampling 5 numbers without replacement. The probability of a specific number being drawn first is $\frac{1}{90}$, for the second $\frac{1}{89}$, and so on. Thus, the probability of a specific sample is:

```
1 / prod(90:86)
```

```
## [1] 1.896126e-10
```

This is the probability of drawing specific numbers. If this scenario corresponds to a lottery, we are interested in the probability of guessing a specific set of 5 numbers. In this case, we must account for all possible orders of the 5 numbers, which is 5! or $5 \times 4 \times 3 \times 2 \times 1$. The probability of winning the lottery is:

```
factorial(5) / prod(90:86)
```

```
## [1] 2.275351e-08
```

Alternatively, we can calculate the total number of ways to choose 5 elements from 90 using the binomial coefficient:

$$\binom{90}{5} = \frac{90!}{5!85!} = 43949268$$

In R, we use the choose function:

```
1 / choose(90, 5)
```

```
## [1] 2.275351e-08
```

Distributions in R

Consider independent replications of a given experiment. From a probabilistic perspective, we are often less interested in individual outcomes (success or failure) and more focused on the total number of successes. This result is random and thus described by a random variable.

Random variables are categorized as discrete or continuous.

A discrete random variable X takes values in a discrete set and is characterized by its probability mass function $f(x) = P(X = x)$ or its cumulative distribution function $F(x) = P(X \leq x)$.

A continuous random variable can take values in 'R' and is characterized by its density function $f(x)$ and distribution function (or cumulative distribution function):

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

`R` includes implementations of major probability distributions, both discrete and continuous, as these are central to statistical modeling and hypothesis testing (discussed later), replacing traditional statistical tables. Examples include:

Distribution	**R Name**
Binomial	`binomial`
Chi-squared	`chisq`
Exponential	`exp`
Geometric	`geom`
Poisson	`pois`
Normal	`norm`
t-Student	`t`

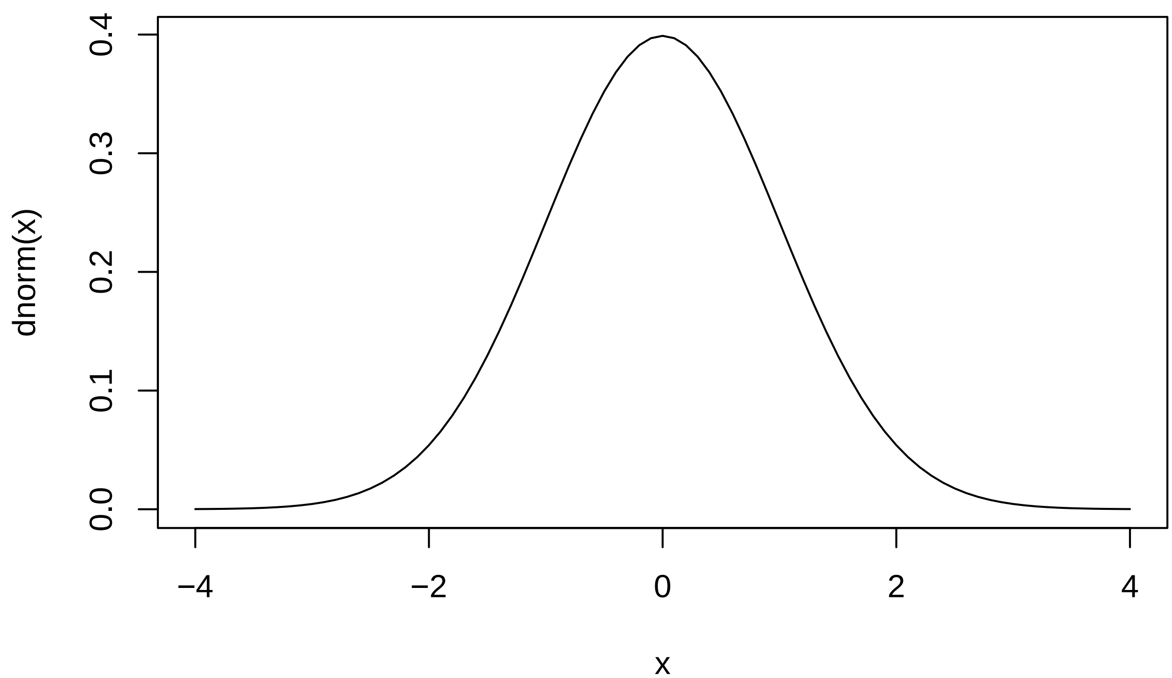
Each distribution allows for four fundamental operations:Probability density/mass, Cumulative distribution, Quantiles, Random number generation. By adding a prefix to the distribution name in `R`, we can compute these quantities:

prefix	function
d	Density‘
p	Cumulative distribution
q	Quantile
r	Random Generation

Density

The density function is rarely used directly but is helpful for plotting:

```
x <- seq(-4, 4, 0.1)
plot(x, dnorm(x), type = "l")
```



or using `gggplo2`

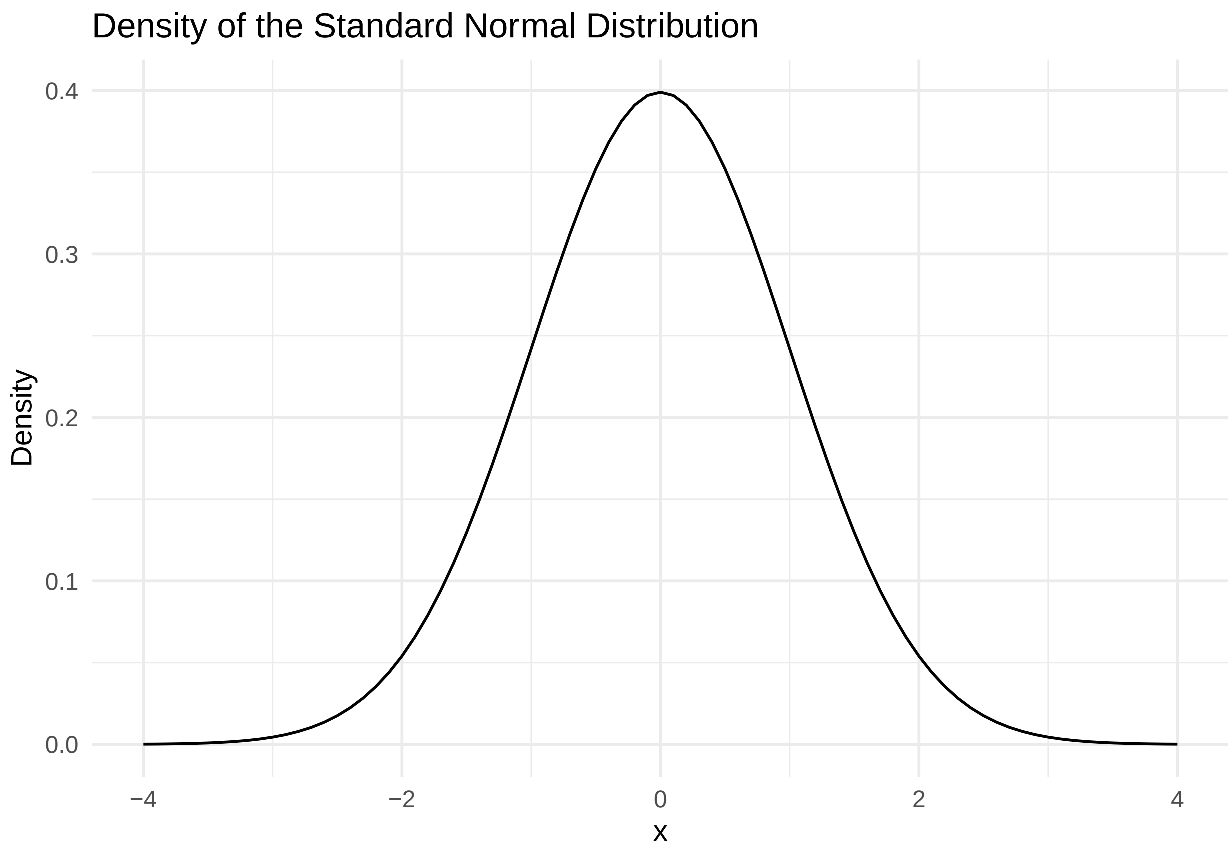
```
library(ggplot2)
```

```
## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.3.3
```

```
x <- seq(-4, 4, 0.1)
df <- data.frame(x = x, density = dnorm(x))

ggplot(df, aes(x = x, y = density)) +
  geom_line() +
```

```r
  labs(title = "Density of the Standard Normal Distribution",
       x = "x", y = "Density") +
  theme_minimal()
```

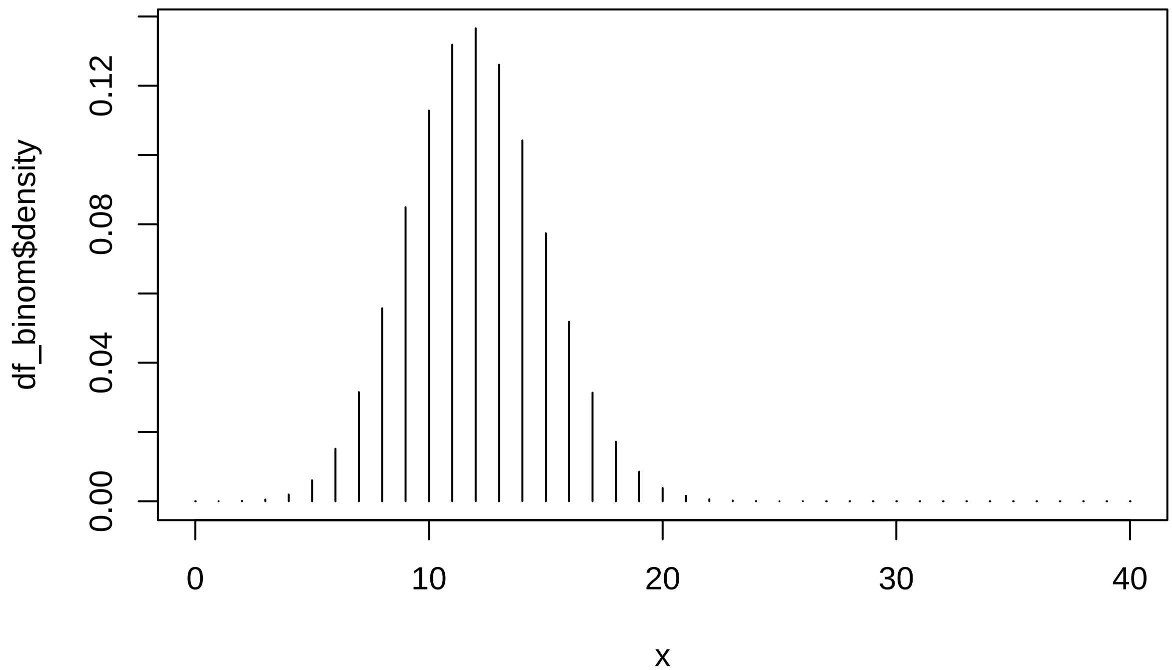


Alternatively, we can use:

```r
curve(dnorm(x), from = -4, to = 4)
```

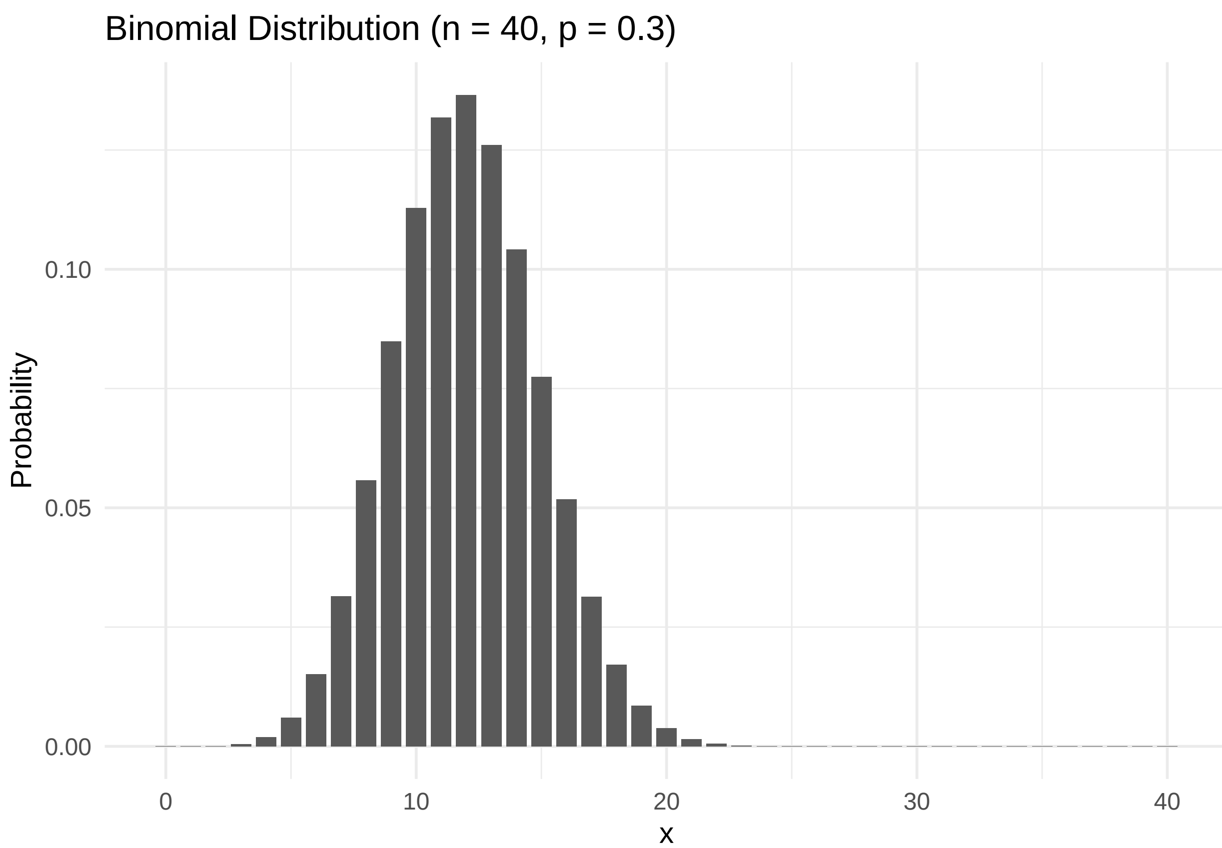
For discrete distributions, a "stick diagram" is often used. For example, the binomial distribution with $n = 40, p = 0.3$:

```r
x <- 0:40
df_binom <- data.frame(x = x, density = dbinom(x, size = 40, prob = 0.3))
plot(x, df_binom$density, type = "h")
```



or

```
ggplot(df_binom, aes(x = x, y = density)) +
  geom_col(width = 0.8) +
  labs(title = "Binomial Distribution (n = 40, p = 0.3)",
       x = "x", y = "Probability") +
  theme_minimal()
```



Cumulative Distribution

These functions are used to calculate probabilities. For example, suppose a biochemical measure in healthy individuals is normally distributed with a mean of 132 and a standard deviation of 13. For a patient with a value of 160:

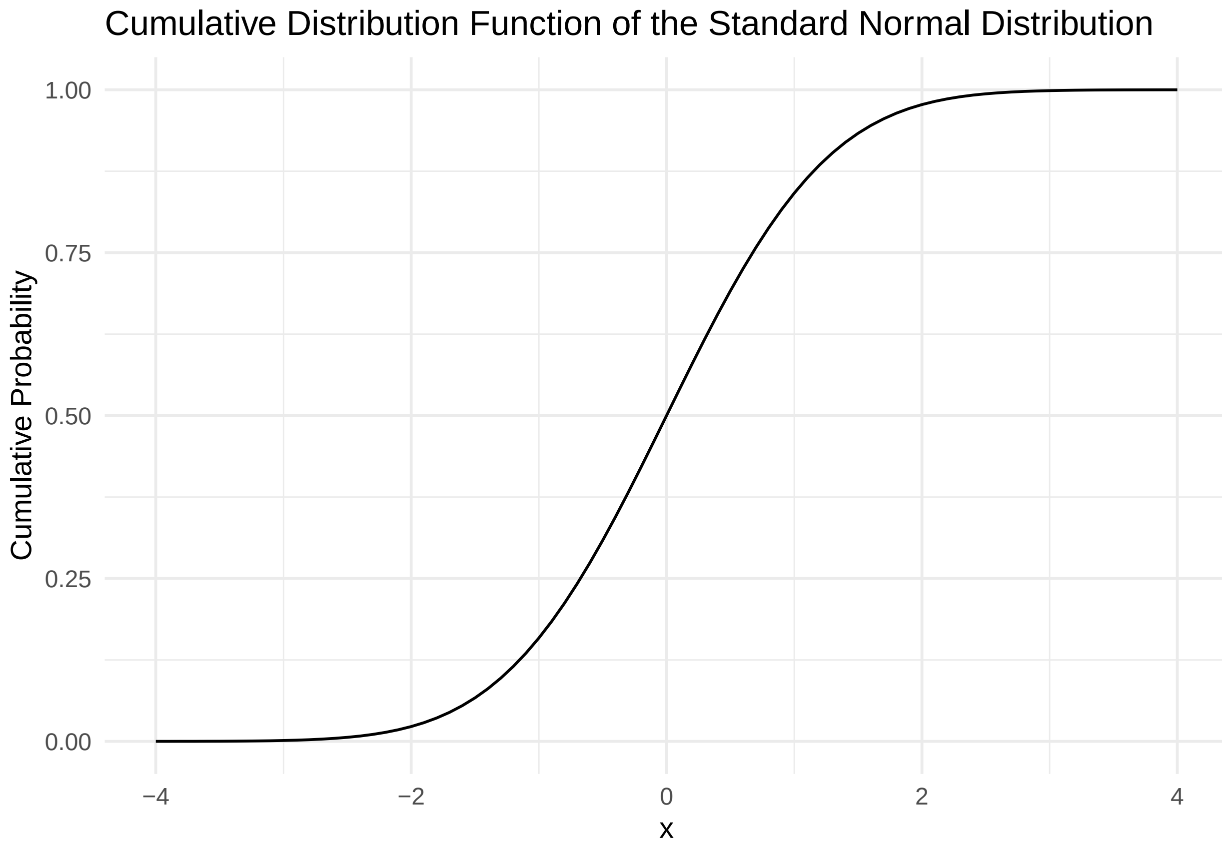
```
1 - pnorm(160, mean = 132, sd = 13)
```

```
## [1] 0.01562612
```

This indicates that approximately 1.5% of the general population exceeds this value. We can visualize the cumulative distribution function of a normal distribution:

```
x <- seq(-4, 4, 0.1)
df_cdf <- data.frame(x = x, cdf = pnorm(x))

ggplot(df_cdf, aes(x = x, y = cdf)) +
  geom_line() +
  labs(title = "Cumulative Distribution Function of the Standard Normal Distribution",
       x = "x", y = "Cumulative Probability") +
  theme_minimal()
```



Quantiles

Quantile functions are the inverse of cumulative distribution functions. For instance, to find z such that $\Phi(z) = 0.9$:

```
qnorm(0.9)
```

```
## [1] 1.281552
```

Random Numbers

We can generate random numbers as realizations of random variables:

```
rnorm(10, mean = 3, sd = 2)
```

```
##  [1] 3.2375325 3.2586252 0.4884656 3.2410743 4.4755755 2.3242337 3.8388310
##  [8] 2.1891753 5.6830545 0.3474844
```

```
rbinom(15, size = 50, prob = 0.2)
```

```
##  [1]  9 11 13  7 11  9  8  8 10 22 14 10  6  7  9
```

Random data is useful for studying mathematical approximations via simulation study.

Exercises

Exercise 1 Calculate the probability of each of the following events:

1. $(X > 3)$, where $X \sim N(0, 1)$:

```
## [1] 0.001349898
```

2. $(X > 42)$, where $X \sim N(35, 36)$:

```
## [1] 0.1216725
```

3. $(X = 10)$, where $X \sim Bin(10, 0.8)$:

```
## [1] 0.1073742
```

4. $(X < 0.9)$, where $X \sim N(0, 1)$:

```
## [1] 0.8159399
```

5. $(X > 6.5)$, where $X \sim \chi^2_2$:

```
## [1] 0.03877421
```

Exercise 2 It is known that 5% of the normal distribution lies outside the interval $(-2s, 2s)$, centered at the mean. What are the corresponding intervals for 1%, 0.5%, and 0.1%? What is the position of the quantiles expressed in terms of the standard deviation s?

Exercise 3 Consider a disease where the probability of post-operative complications is 20%. A surgeon suggests a new procedure and tests it on 10 patients, none of whom have complications. What is the probability of operating on 10 patients successfully (without complications) using the traditional procedure?

Exercise 4 The toss of a coin can be simulated in R using rbinom instead of sample. How exactly can this be done?
Hint: Simulate 10 tosses of a fair coin using rbinom.