

Chapter 2

Probability and Distributions

Suam habet fortuna rationem (Chance has its reason).
—Petronius

David Hume, the renowned philosopher in his *Treatise of Human Nature*, describes probability as the amount of evidence that accompanies uncertainty, a reasoning from conjecture. Probability theory is the branch of mathematics dealing with the study of uncertainty. It provides a means of quantifying uncertainty.

Since machine learning relies on the probabilistic nature of data, probability theory has an important role to play in the design of learning algorithms.

While the evidence associated with uncertainty may allow us to reason about uncertain statements, information theory allows to measure the uncertainty or randomness of a probability distribution by a mathematical parameter known as *entropy*.

If an event is unpredictable because of a random variable, it is called a *stochastic* (nondeterministic) event. In a *deterministic* event, there is no presence of randomness and the outcomes can be precisely determined. Randomness and uncertainty are intrinsic to machine learning and may arise from different sources.

2.1 Sources of Uncertainty

The possible sources of uncertainty in machine learning are:

- Randomness inherent in the data, resulting in different learning models because models train differently with different data sets.
- A deterministic data set, which may appear stochastic if its observations are incomplete/missing, will result in an uncertain model.

- Randomness present in the learning model itself, i.e., initializing a set of weights specific to a model could induce randomness in the model. The *k-means* algorithm used in clustering is particularly vulnerable in this regard.

2.2 Random Experiment

A **random experiment** is one where we cannot predict the outcome with certainty before the experiment is conducted. Frequency-based probability (*frequentist*) theory assumes that the experiment can be indefinitely repeated, while the conditions of the experiment remain unchanged. The repetitions may be in time domain (repeatedly tossing a single coin) or in the space domain (tossing a bunch of similar coins all at once). Since we are concerned with the long-term behavior of the outcomes of the experiment, the “repeatability” aspect of a random experiment assumes importance. Therefore, a random experiment should clearly state, what constitutes an outcome.

The **sample space** of a random experiment constitutes the set S , which includes all the possible outcomes of a random experiment.

The **parameter** of a model refers to a non-random quantity which does not change once its value is chosen.

A **random variable** is the function defining the outcome of a random experiment. It is a description of the states that it can possibly take. Depending on the possible values of the outcome, a random variable may be defined to be **discrete** or **continuous**. A discrete random variable has a finite or countably infinite¹ number of states. A continuous random variable is associated with an infinite number of possible values and is defined over an interval.

2.3 Probability

The probability P of a success is defined by the frequency of that event based over past observations. Therefore, the probability of the number of successes k over n observations is dependent on the rate at which the events take place and is known as **Frequentist probability**.

In contrast, **Bayesian probability** is probability of an event, based on prior knowledge of conditions that might be related to the event. It is presented in the form of an inference, which is the **prior probability** distribution and a **posterior probability** distribution.

Suppose we have a parameter ρ which is defined by a normal distribution and our objective is to estimate the parameter ρ . In this context, Bayesian probability can be defined by the following:

¹A countably infinite set may take forever to count, but we can get to any particular element in the set in a finite amount of time.

- **Likelihood**—is the conditional density of the data given ρ , i.e., $\mathcal{L}(\text{data}, \rho)$. Suppose if ρ also has some inherent noise defined by a normal distribution with zero mean and variance of 1, i.e., $\mathcal{N}(\mu = 0, \sigma = 1)$, the likelihood of the data (denoted as Y), given ρ , can be written as $\mathcal{L}(Y, \rho) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y-\rho)^2}{2}}$.
- **Prior**—is the distribution of ρ in the absence of any data. If it follows a normal distribution, it will have a mean μ and a variance σ and can be defined as $\mathcal{N}(\mu, \sigma)$.
- **Posterior**—is the conditional distribution of ρ , given the data, which is proportional to the likelihood and the prior.

To summarize the above, we have a distribution of ρ defined as $Y = \rho + \mathcal{N}(0, 1)$. The likelihood is $p(Y | \rho) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y-\rho)^2}{2}}$ and the prior distribution is $p(\rho) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\rho-\mu)^2}{2\sigma^2}}$ (i.e., normally distributed).

The posterior probability distribution is therefore proportional to the product of the prior and the likelihood and is defined as $p(\rho | Y) \propto \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y-\rho)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\rho-\mu)^2}{2\sigma^2}}$.

2.3.1 Marginal Probability

The probability of a subset of events irrespective of any other event over a set of events is the marginal probability of the event subset. For instance, the probability of a dice coming up with a particular number is the marginal probability of the number. Marginal probability can be represented by the following equations:

$$P(X = x) = \sum_y P(X = x, Y = y) \text{—for discrete variables} \quad (2.3.1)$$

$$p(x) = \int p(x, y) dy \text{—for continuous variables}$$

y represents all the possible values of the random variable Y . We are holding X constant ($X = x$) while iterating over all the possible Y values and summing up the joint probabilities.

```
# CALCULATE MARGINAL PROBABILITY
# Create a sample outcome of rolling a six-sided die twice
S <- rolldie(times = 2, nsides = 6, makespace = TRUE)
# Find the marginal probability (Px = X1)
marginal(S, vars = c("X1"))
```

```

X1      probs
1  1 0.1666667
2  2 0.1666667
3  3 0.1666667
4  4 0.1666667
5  5 0.1666667
6  6 0.1666667
```

2.3.2 Conditional Probability

The probability of the event X occurring when the secondary event Y has happened is the conditional probability of X given Y . Mathematically, the conditional probability that $X = x$ given $Y = y$ can be represented as $P(X = x | Y = y)$, which can be written as

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \quad (2.3.2)$$

```
# CALCULATE CONDITIONAL PROBABILITY
df <- data.frame('condition' = c('disease', 'no disease'),
                 'treatment' = c(5, 195),
                 'no treatment' = c(100, 500),
                 check.names=F)
```

Condition	Treatment	No treatment
Disease	5	100
No disease	195	500

```
# Find the conditional probability of a disease given a treatment
# P(disease | treatment) = p(disease & treatment)/p(treatment)
5/200
```

```
[1] 0.025
```

2.3.3 The Chain Rule

We can rearrange the formula of conditional probability to get a chain of conditional probabilities represented as $P(X, Y) = P(X | Y)P(Y)$.

We can extend the above for n variables

$$P(X_1, X_2, \dots, X_n) = P(X_1 | X_2, \dots, X_n) \times P(X_2 | X_3, \dots, X_n) \\ \times P(X_{n-1} | X_n) \times P(X_n). \quad (2.3.3)$$

Equation 2.3.3 is known as the chain rule of conditional probabilities.

2.4 Bayes' Rule

It follows from Eq. 2.3.3 that $P(X, Y) = P(X | Y)P(Y)$. By symmetry, the same equation can also be written as $P(X, Y) = P(Y | X)P(X)$. The Bayes' rule can now be defined as

$$P(X | Y) = \frac{P(Y | X)P(X)}{P(Y)} \quad (2.4.1)$$

Bayes' rule can be thought of as updating our belief about a hypothesis X in the light of the evidence Y . The posterior belief $P(X | Y)$ is calculated by multiplying our prior belief $P(X)$ by the likelihood $P(Y | X)$. Given prior state knowledge, the Bayes' rules tell us how to update the belief based on the observations.

Fundamentally, the difference between frequentists and Bayesians revolves around the definition of probability. For frequentists, probabilities are related to the frequencies of the events. For Bayesians, probabilities are related to our own belief about an event and the likelihood of the data, given the event. Let us explore this with an example.

Frequentist Approach: A coin when flipped ends up “head” with probability P (the value of P is unknown) and “tail” with probability $(1 - P)$. Trying to estimate P , we flip the coin 14 times. It ends up “head” 10 times. We need to know if the next two tosses will get two “heads” in a row. We would say that the best (maximum likelihood) estimate for p is $\frac{10}{14} \sim 0.714$. The probability of two heads is $0.714^2 \sim 0.51$; we therefore predict two “heads”.

Bayesian Approach: In this approach, p is not a value but a distribution. The Bayesian approach would treat p as a random variable with its own distribution of possible values. What is the probability of a given value of p , given the data

$$P(p | data) = \frac{P(data | p)P(p)}{P(data)}$$

Let us assume that we have tossed the coin three times, each consisting of 14 tosses and we come up with the following results:

$$\begin{aligned} P(H = 7) &= 0.5 && -7 \text{ heads} \\ P(H = 8) &= 0.57 && -8 \text{ heads} \\ P(H = 9) &= 0.64 && -9 \text{ heads} \end{aligned}$$

The above is the **prior** distribution, which represents the knowledge we have, about how the data is generated prior to observing them.

The **likelihood** function $P(data|p)$ is the likelihood for the *data*, as it tells us how likely the data is, given the model specified by any value of p . The data consists of 10 “heads” and 4 “tails”. Therefore,

$$P(data|p) = p^{10} \cdot (1 - p)^4$$

or,

$$\begin{aligned}P(data|p = 0.5) &= 0.5^7 \cdot (1 - 0.5)^7 = 6.10e^{-05} \\P(data|p = 0.57) &= 0.57^8 \cdot (1 - 0.57)^6 = 7.04e^{-05} \\P(data|p = 0.64) &= 0.64^9 \cdot (1 - 0.64)^5 = 0.00011\end{aligned}$$

Therefore, $P(data)$ (also known as *evidence*) can be calculated as

$$\begin{aligned}P(data) &= \sum_p P(data|p) \cdot P(p) \\P(data) &= 6.1e^{-05} * 0.5 + 7.04e^{-05} * 0.57 + 0.00011 * 0.64 = 0.00014\end{aligned}$$

The posterior probabilities are

$$\begin{aligned}P(p = 0.5|data) &= \frac{P(data|p = 0.5) \cdot P(H = 7)}{P(data)} = 0.217 \\P(p = 0.57|data) &= \frac{P(data|p = 0.57) \cdot P(H = 8)}{P(data)} = 0.287 \\P(p = 0.64|data) &= \frac{P(data|p = 0.64) \cdot P(H = 9)}{P(data)} = 0.496\end{aligned}$$

Considering the conditional independence between two tosses, $P(HH|p) = [P(H|p)]^2$, we can predict the probability of two heads as follows:

- if $p = 0.5$, $P(HH|p) = 0.217^2 = 0.047$
- if $p = 0.57$, $P(HH|p) = 0.287^2 = 0.082$
- if $p = 0.64$, $P(HH|p) = 0.496^2 = 0.246$

Since all the values are less than 0.5, we predict two “tails”.

Summing it up in an application,

```
likelihood = function(p,heads){
  p^heads * (1 - p)^(14-heads)
}
likelihood = c(likelihood(0.5, 7), likelihood(0.57, 8), likelihood(0.64, 9))
prior = c(0.5, 0.57, 0.64)
evidence = sum(prior * likelihood)
posterior_prob = prior * likelihood / evidence
```

Likelihood values are 6.103516e-05 7.043842e-05 0.0001089262

Evidence = 0.0001403802

Posterior Probabilities are 0.2173923 0.2860082 0.4965995

2.5 Probability Distribution

A **probability distribution** (or probability measure) is a description of the outcome of a random experiment, i.e., it is a description of a random phenomenon in terms of the probabilities of its possible states. If the random variable takes any value between two specified limits, it is a continuous variable, or else a discrete variable and the probability distribution would depend on whether the variables are discrete or continuous.

2.5.1 Discrete Probability Distribution

A distribution of discrete variables is described using a **probability mass function** (PMF), denoted as P . It maps the state of a random variable to the probability of the random variable of assuming that state. If the probability that $X = x$ is 1, it indicates that $X = x$ is certain and a probability of 0 indicates that $x = x$ is impossible.

The PMF, P of a random variable X , needs to satisfy the **Kolmogorov axioms**:

- The domain of P must be the set of all possible states of X ;
- $0 \leq P(x) \leq 1$ for all $x \in X$; and
- $\sum_{x \in X} P(x) = 1$.

2.5.2 Continuous Probability Distribution

For continuous variables, we define the probability which the variable can take as **probability density function** (PDF), and the probability density function p needs to satisfy the **Kolmogorov axioms**:

- The domain of p must be the set of all possible states of X ;
- $p(x) \geq 0$, for all $x \in X$; and
- $\int p(X)dx = 1$.

A probability density function $p(x)$ gives the probability of existing inside an infinitesimal region δx , and is given by $p(x)\delta x$. The fact that $p(x) = p(X = x) = 0$ may seem paradoxical, but it can be thought of as an interval which has a positive length composed of points having zero length.

2.5.3 Cumulative Probability Distribution

The **cumulative distribution function** (CDF) is the probability that the random variable takes a value less than or equal to x , i.e., it is the probability of having a

value less than x

$$\begin{aligned} f(x) &= \int_{-\infty}^x f(t)dt \text{ for a continuous distribution} \\ F(x) &= \sum_{t \leq x} f(t) \text{ for a discrete distribution} \end{aligned} \quad (2.5.1)$$

2.5.4 Joint Probability Distribution

Joint probability is the distribution of many variables occurring at the same time. $P(X = x, Y = y)$ denotes the joint probability that $X = x$ and $Y = y$ occurs simultaneously.

2.6 Measures of Central Tendency

The **expected value** of a function $f(x)$ having a probability distribution $P(X)$ is the **mean** value of $f(x)$. The expectation or the expected value is computed as

$$\begin{aligned} E[f(x)] &= \sum_{x \in S} P(x)f(x) \text{ —for discrete variables} \\ E[f(x)] &= \int_S p(x)f(x)dx \text{ —for continuous variables} \end{aligned} \quad (2.6.1)$$

Median is any x drawn from a probability distribution, having a value m such that half of the population has values of x less than m and the other half, from the population has values of x greater than m and satisfies the following equation:

$$\begin{aligned} \int_{-\infty}^m p(x)dx &= \frac{1}{2} && \text{for continuous distributions} \\ P(x \leq m) &\geq \frac{1}{2} \text{ and } P(x \geq m) \geq \frac{1}{2} && \text{for discrete distributions} \end{aligned} \quad (2.6.2)$$

Mode of a discrete probability distribution is a value x which maximizes the probability mass function and that of a continuous probability distribution and, is a value x which maximizes the probability density function. It is the value that is most likely to be sampled from a distribution.

The above three measures are influenced by the **shape** of the distribution as well as by the presence of **outliers**.

2.7 Dispersion

Percentile is a measure indicating the value below which a given percentage of observations in a group of observations fall. The k th percentile of a set of observations is defined such that $k\%$ of the observations will lie below and $(100 - k)\%$ of the observations will lie above the defined percentile.

- The 25th percentile is the *lower quartile*;
- The 50th percentile is the *median*; and
- The 75th percentile is the *upper quartile*.

Sometimes, **quantiles** are more commonly used, which are the values taken from regular intervals of the quantile function of a random variable. The quantile q of a probability distribution is the inverse of its cumulative distribution function f . The quantile of order n for a distribution f can be written as $f(x) = P(X \leq x) \geq n$.

The **variance** and **standard deviation** of a vector $\mathbf{X} = [x_1, x_2, \dots, x_n]$ are both measures of the spread of the distribution about the mean. It is a measure of how much the values of $f(\mathbf{X})$ vary as we sample different values of \mathbf{X} . The square root of the variance is the standard deviation:

$$Var(f(\mathbf{X})) = E([f(\mathbf{X}) - E[f(\mathbf{X})])^2 \quad (2.7.1)$$

The variance of \mathbf{X} is written as

$$\begin{aligned} Var(\mathbf{X}) &= \sigma_{\mathbf{X}}^2 \\ &= \frac{1}{n-1} \mathbf{X} \mathbf{X}^\top \end{aligned} \quad (2.7.2)$$

In Eq. 2.7.2, $\mathbf{X} \mathbf{X}^\top$ is a dot product, which gives us the length of the vector \mathbf{X} . If the length is large, it intuitively tells us that the variance is high and *vice versa*.

2.8 Covariance and Correlation

Covariance gives an indication as to how much two vectors are linearly related (i.e., how much, each one of the vectors are in each other), as well as the scale of these variables. Let us consider two vectors $\mathbf{X} = [x_1, x_2, \dots, x_n]$ and $\mathbf{Y} = [y_1, y_2, \dots, y_n]$

$$Cov(f(\mathbf{X}), f(\mathbf{Y})) = E[(f(\mathbf{X}) - E[f(\mathbf{X})])(f(\mathbf{Y}) - E[f(\mathbf{Y})])] \quad (2.8.1)$$

The covariance of \mathbf{X} and \mathbf{Y} can be written as

$$\begin{aligned}
Cov(\mathbf{XY}) &= \sigma_{\mathbf{XY}}^2 \\
&= \frac{1}{n-1} \mathbf{XY}^\top
\end{aligned} \tag{2.8.2}$$

Equation 2.8.2 is again a dot product of \mathbf{A} and \mathbf{B} . If this value is 1, the two vectors are the same and, if it is 0, the two vectors are orthogonal (perpendicular) to each other, i.e., the two vectors have nothing to do with each other and are statistically independent. So covariance defines the measure of statistical dependence between the two vectors.

$$Cov(\mathbf{XY}) = \begin{pmatrix} \sigma_{x_1x_1}^2 & \sigma_{x_1y_1}^2 & \cdots & \sigma_{x_1x_n}^2 & \sigma_{x_1y_n}^2 \\ \sigma_{x_1y_1}^2 & \sigma_{y_1y_1}^2 & \cdots & \sigma_{x_ny_1}^2 & \sigma_{y_1y_n}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{x_1x_n}^2 & \sigma_{x_ny_1}^2 & \cdots & \sigma_{x_nx_n}^2 & \sigma_{x_ny_n}^2 \\ \sigma_{x_1y_n}^2 & \sigma_{y_1y_n}^2 & \cdots & \sigma_{x_ny_n}^2 & \sigma_{y_ny_n}^2 \end{pmatrix} \tag{2.8.3}$$

The diagonals of the matrix in Eq. 2.8.3 are the variance terms and the covariances of all the \mathbf{X} and \mathbf{Y} pairs are the off-diagonal terms (which, incidentally, are symmetric and is a *Hermitian* matrix).

Now, if the covariance terms are small in value, it implies that the variables are not dependent on each other but, if some (or all) of the covariance terms are high in value, it means that they are *redundant* variables, because they have a lot of dependency on each other.

The fact that redundant variables “may not” contribute toward determining the target variable is important in most machine learning algorithms, as they can become confounders in a model. So, ideally, the extreme requirement is that the non-diagonal terms should be either zero or close to zero.

Now, if the off-diagonal terms are either 0 or close to 0 and, if we reorient the diagonal terms from the largest value (i.e., $(i = 1, j = 1)$ to the smallest value $(i = n, j = n)$), we are in a position to identify the most important predictor variables, which contribute to determining the target variable. In other words, we can answer the question—**are all the predictor variables important OR redundant?**

At this stage, you may want to refer to Sect. 1.11 on *SVD*.

Correlation measures how much the two variables (vectors) are related to each other:

$$Cor(f(\mathbf{X}), f(\mathbf{Y})) = \frac{\sigma_{\mathbf{XY}}^2}{\sqrt{\sigma_{\mathbf{X}}^2 \sigma_{\mathbf{Y}}^2}} \tag{2.8.4}$$

2.9 Shape of a Distribution

The shape of a distribution is quantitatively measured by **Skewness** and **Kurtosis**. A distribution is said to be right-skewed (or positively skewed) if the right tail is stretched from the center. A left-skewed (or negatively skewed) distribution is stretched to the left side. A symmetric distribution is balanced about its center.

Some distributions have a flat shape with thin tails and are called *platykurtic*. Distributions with a steep peak and with heavy tails are called *leptokurtic*. Skewness is a measure of the absence of symmetry, while Kurtosis is the degree to which the distribution is peaked.

$$skew(x) = E\left[\left(\frac{x - \mu}{\sigma}\right)^3\right] \quad (2.9.1)$$

$$kurt(x) = E\left[\left(\frac{x - \mu}{\sigma}\right)^4\right], \quad (2.9.2)$$

where

$$\sigma^2 = var(x)$$

$$\mu = E(x)$$

2.10 Chebyshev's Inequality

Chebyshev's inequality states that **at least** $(1 - \frac{1}{k^2})$ of the data of a distribution will lie within k standard deviations away from the mean (k being a real number greater than 0). It provides a way to know what fraction of the data will fall within k standard deviations from the mean for any distribution. This is illustrated in Fig. 2.1.

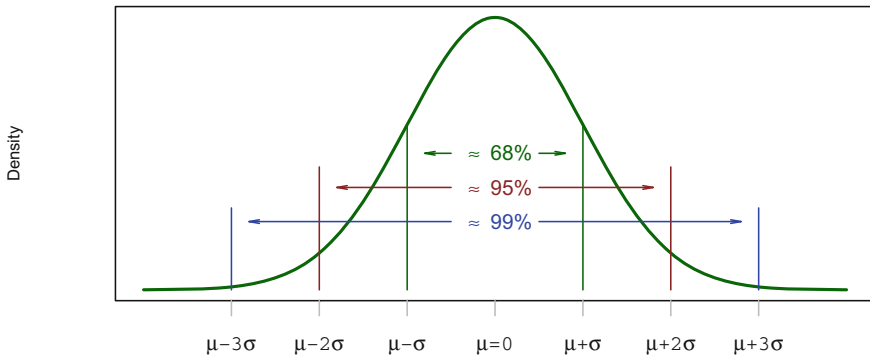


Fig. 2.1 Chebyshev's theorem, as applicable to a Gaussian distribution

2.11 Common Probability Distributions

We walk through two commonly used distributions within the discrete and continuous distribution spaces.

2.11.1 Discrete Distributions

Bernoulli Distribution

The Bernoulli distribution has two possible outcomes and is controlled by a parameter ϕ , which is the probability of the outcome registering a success (i.e., the probability is 1). The Bernoulli distribution can be defined as having the following properties:

$$P(x = 1) = \phi \quad (2.11.1)$$

$$P(x = 0) = 1 - \phi \quad (2.11.2)$$

$$P(x = x) = \phi^x (1 - \phi)^{1-x} \quad (2.11.3)$$

$$E_x[x] = \phi \quad (2.11.4)$$

$$Var_x(x) = \phi(1 - \phi) \quad (2.11.5)$$

$$f(x) = P(x = x) = \binom{n}{x} \cdot p^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} \quad (2.11.6)$$

x is the number of successes,
 n is the number of experiments, and
 p is the probability of success.

The probability of obtaining seven tails by flipping a coin eight times can be obtained as

```
# BERNOULLI DISTRIBUTION
# Probability of having seven tails by flipping a fair coin eight times
dbinom(x=7, size=8, prob=0.5)
```

```
[1] 0.03125
```

Approximately, 3% are the chances of obtaining seven tails by flipping a coin eight times.

Multinomial and Multinoulli Distributions

The Bernoulli distribution has only two outcomes from one trial. The binomial distribution is an extension of the bernoulli distribution having two outcomes for each of the multiple trials. The multinoulli distribution generalizes the bernoulli

distribution and has more than two outcomes from a single trial. The multinomial distribution extends this even further by having multiple outcomes from multiple trials.

For continuous variables, there exist uncountable number of states, which are governed by a small number of parameters. These parameters which define a continuous distribution, necessarily, impose strict limits on such distributions.

Let us suppose we have a random vector (x_1, x_2, x_3) having a multinomial distribution. The probabilities associated with the occurrence of each variable x_1, x_2, x_3 are $p_1 = 0.1$, $p_2 = 0.4$, and $p_3 = 0.5$. The number of times each variable occurs is $x_1 = 5, x_2 = 5, x_3 = 10$, i.e., 20 occurrences. The multinomial probability for these multiple outcomes can be found as follows:

```
# MULTINOMIAL DISTRIBUTION
dmultinom(x = c(5,5,10), size = 20, prob = c(0.1, 0.4, 0.5))
```

```
[1] 0.004655851
```

2.11.2 Continuous Distributions

Gaussian Distribution

If the number of events is large, the Gaussian distribution may be used to describe the events. The most commonly used Gaussian distribution is the **normal distribution**, where the density function is represented by the equation

$$f(x) = \mathcal{N}(x; \mu, \sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (2.11.7)$$

The parameter μ , which defines the central peak of the X -coordinate, is also the mean of the distribution, i.e., $E[X] = \mu$ and $\mu \in \mathbb{R}$. The distribution has a standard deviation represented by the symbol σ , where $\sigma \in (0, \infty)$. The variance is σ^2 .

The **cumulative density function** for the Gaussian distribution is

$$F(x) = \int_{-\infty}^x \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (2.11.8)$$

Figure 2.2 depicts the PDF and Fig. 2.3 depicts the CDF of the **standard normal distribution**, which has $\mu = 0$ and $\sigma = 1$.

If we do not have any prior knowledge about the form of a distribution, the normal distribution may be assumed as the default choice for two major reasons:

- The **Central Limit Theorem** proves that given sufficiently large samples (sample sizes greater than 30) from a population, all of the samples are normally distributed. The mean of the sample will be closer to the mean of the population, as the sample size increases, irrespective of whether the sample is normally distributed

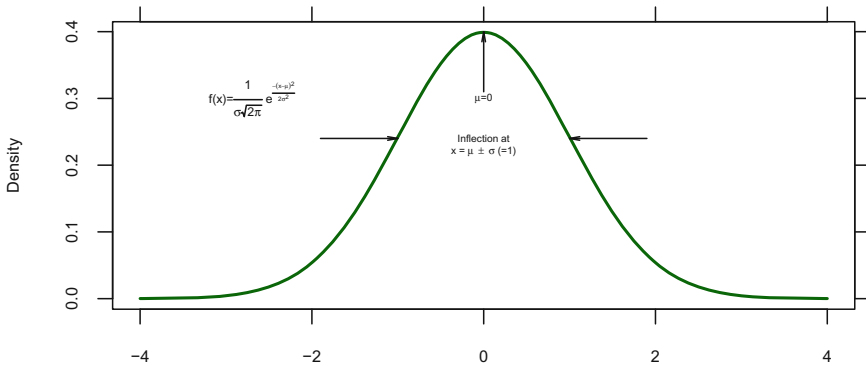


Fig. 2.2 PDF of Standard Normal

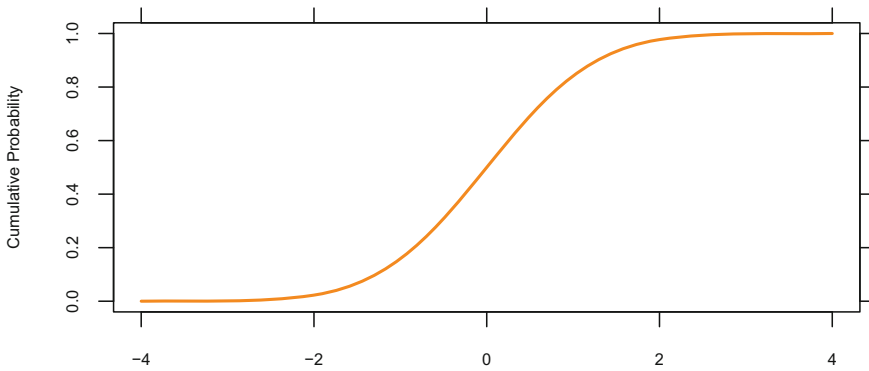


Fig. 2.3 CDF of Standard Normal

or not. The variance of the samples will approximate closely to the variance of the population divided by the sample size. The important take-away is that in practice, many distributions can be modeled successfully as approximations to the Gaussian distribution.

- **Entropy** is a mathematical formula used to measure information gain. The normal distribution has the maximum entropy among all continuous distributions having a fixed mean and variance. What this implies is that the normal distribution incorporates the minimum amount of prior knowledge into a model.

Logistic Distribution

The logistic distribution is used in a certain type of regression known as *logistic regression*. This distribution is symmetrical, unimodal (having one peak), and is similar in shape to the Gaussian distribution. The logistic distribution tends to have slightly fatter tails.

The CDF of the logistic distribution is

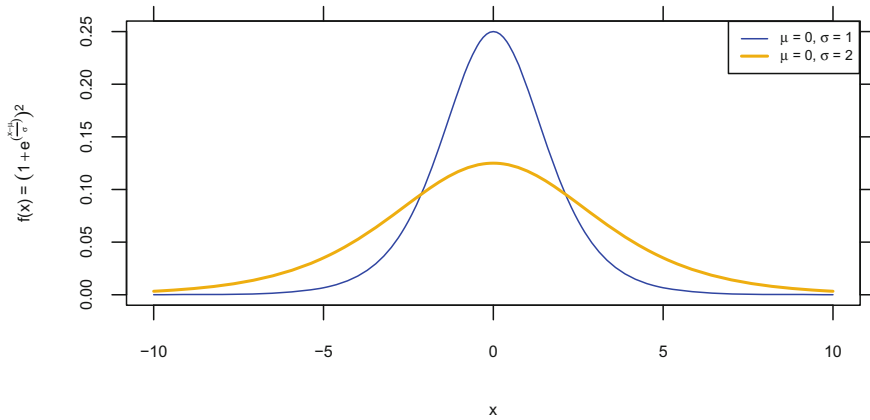


Fig. 2.4 PDF—Logistic Distribution

$$F(x) = \frac{e^x}{(1 + e^x)} \quad (2.11.9)$$

Differentiating the above equation wrt x , we get the PDF

$$f(x) = \frac{e^x}{(1 + e^x)^2} \quad (2.11.10)$$

The quantile of the distribution is

$$F^{-1}(x) = \ln\left(\frac{p}{1-p}\right) \quad (2.11.11)$$

Figures 2.4 and 2.5 show the plots of the logistic distribution.

An interesting aspect of this distribution is that the ratio, $\frac{p}{1-p}$, is called the **odds** in favor of an event happening with probability p . The natural logarithm of the *odds ratio*, $\ln\left(\frac{p}{1-p}\right)$ is the logarithmic odds and is called the **logit**.

2.11.3 Summary of Probability Distributions

R has a range of probability distributions and for each of them, four functions are available—the pdf (which has a prefix **d**); the cdf (prefix **p**); quantiles of the distribution (**q**); and the random number generator (**r**). Each letter can be prefixed to the function names in Table 2.1.

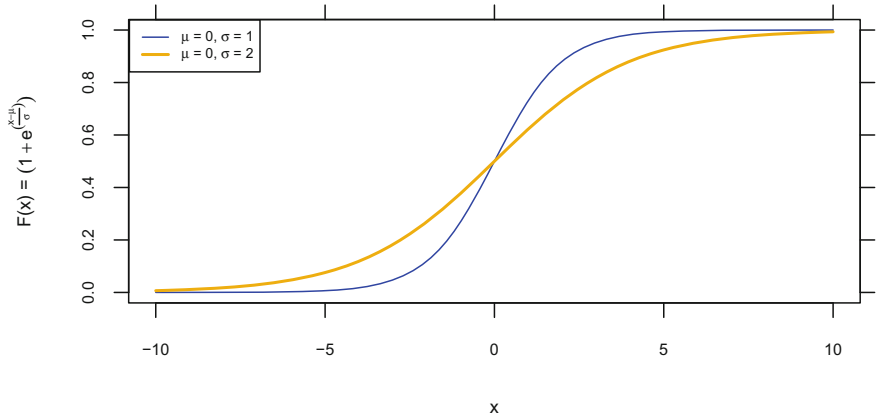


Fig. 2.5 CDF—Logistic Distribution

Table 2.1 Probability distributions and their respective parameters

R function	Distribution	Parameters
beta	beta	shape1, shape2
binom	binomial	sample size, probability
cauchy	Cauchy	location, scale
exp	exponential	rate
chisq	chi-squared	degrees of freedom
F	Fishers F	df1, df2
gamma	gamma	shape
geom	geometric	probability
lnorm	lognormal	mean, standard deviation
logis	logistic	location, scale
norm	normal	mean, standard deviation
pois	Poisson	mean
signrank	Wilcoxon signed rank statistic	sample size
t	Students t	degrees of freedom
Weibull	Weibull	shape

2.12 Tests for Fit

Tests for fit measure how well do the observed data correspond to the fitted (assumed) model. They test the following hypotheses:

H_0 : The data fits the model

H_A : The data does not fit the model

There are many different tests for fit, namely Kolmogorov–Smirnov test, Anderson–Darling test, etc.; however, we shall discuss only the chi-square test.

2.12.1 Chi-Square Distribution

A standard normal deviate is a normal distribution with zero mean and unit standard deviation (i.e., $\mathcal{N}(0, 1)$). The chi-square distribution is a distribution of the sum of squared standard normal deviates. The degrees of freedom of the chi-square distribution is the number of standard normal deviates. The distribution of a squared single normal deviate is a chi-square distribution having a single degree of freedom (χ_1^2).

If we consider Z_1, \dots, Z_k as independent, standard normal variables, the sum of their squares is $\sum_i^k Z_i^2 \sim \chi_k^2$.

If $Z_1 \sim \mathcal{N}(0, 1)$, how do we find the pdf of $(Z_1)^2$? We have seen earlier that the pdf of x is

$$f(x) = \mathcal{N}(x; \mu, \sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

If $g(x)$ is the pdf of $(Z_1)^2$ the pdf for the chi-square distribution can be derived as

$$\begin{aligned} g(x) &= \frac{d}{dx} p(x^2 \leq x) \\ &= \frac{d}{dx} p(-\sqrt{x} \leq x^2 \leq \sqrt{x}) \\ &= \frac{d}{dx} \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-u^2/2} du \\ &= \frac{2}{2\pi} \frac{d}{dx} \int_0^{\sqrt{x}} e^{-u^2/2} du \\ &= \frac{2}{2\pi} e^{-\sqrt{x^2}/2} \frac{d}{dx} \sqrt{x} \\ &= \frac{2}{2\pi} e^{-\frac{x}{2}} \frac{1}{2\sqrt{x}} \\ &= \frac{e^{-x/2}}{2\pi\sqrt{x}} \\ &= \frac{1}{2\pi} x^{(\frac{1}{2}-1)} e^{-x/2} \end{aligned} \tag{2.12.1}$$

The last term in Eq. (2.12.1) resembles another function known as the *Gamma function*, which defines the pdf of the chi-square distribution.

The degrees of freedom is the mean of the chi-square distribution. As the degrees of freedom increase, the chi-square distribution tends toward the normal distribution.

The chi-square test is used to test “goodness-of-fit” of data to a model. There are different types of “chi-square” test, as well as other tests that use the chi-square distribution. All of them estimate the probability of observing the results under the given hypotheses. If that probability is low, then one can confidently reject the hypothesis. Many test statistics are approximately distributed as chi-square.

A chi-square distribution having different degrees of freedom is shown in Fig. 2.6.

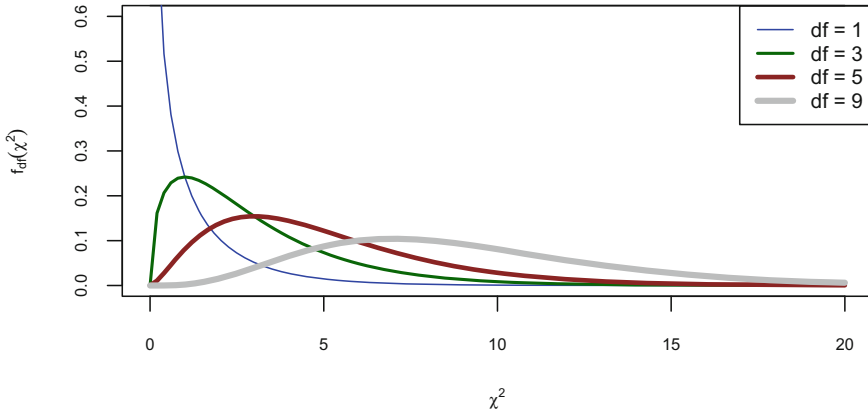


Fig. 2.6 Chi-square distributions with different degrees of freedom

2.12.2 Chi-Square Test

Of the many chi-squares tests, two important ones are as follows:

- Tests of deviations of differences between theoretically expected and observed frequencies (*one-way tables*). This is also known as the **goodness-of-fit test** and determines if a sample data matches a population. It measures the “goodness-of-fit” between the observed and expected data.
- Test of relationship between categorical variables (*contingency tables*). This test compares two categorical variables in a contingency table to see if they are related and is also called **test for independence**. In a more general sense, it is a test to see whether distributions of categorical variables differ from each another and determines whether there is a significant association between them.

Chi-square tests use the **chi-square test statistic** and the equation for this statistic is

$$\chi_c^2 = \sum \frac{(\text{observed Value} - \text{expected Value})^2}{\text{expected Value}} \quad (2.12.2)$$

Let us consider the following data on weekly customer visits in Table 2.2.

We need to find out if the observed and expected values come from the same distribution, given a significance level $\alpha = 0.05$.

First, let us define our hypotheses space:

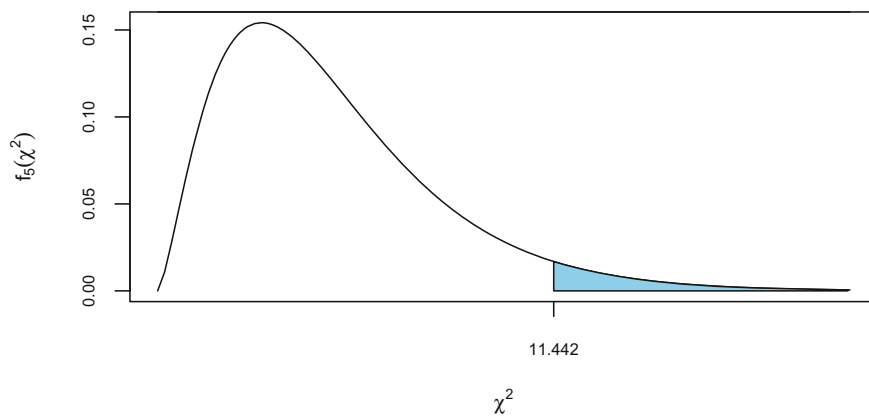
Null Hypothesis H_0 : There is no difference between the observed and expected customer visits.

Alternate Hypothesis H_A : There is a difference between the observed and expected customer visits.

The hypothesis is tested using Pearson’s χ^2 -test.

Table 2.2 Observed and expected customer visits data

Day	Observed no of visits	Expected probability (%)
Mon	30	10
Tue	14	10
Wed	34	15
Thur	45	20
Fri	57	30
Sat	20	15

**Fig. 2.7** Shaded area depicting the p-value of a chi-square distribution

```
# GOODNESS of FIT
chisq.test(x = c(30, 14, 34, 45, 57, 20),
           p = c(0.1, 0.1, 0.15, 0.2, 0.3, 0.15))
```

Chi-squared test for given probabilities

data: c(30, 14, 34, 45, 57, 20)

X-squared = 11.442, df = 5, p value = 0.04329

p value = 0.04329 and $\chi^2 = 11.442$

p value is the probability of obtaining a result equal to or more extreme (toward the direction of H_A) than what was actually observed, when the null hypothesis is true. This probability is the area under the curve at the point and beyond, given the null hypothesis.

The chi-square distribution showing the p value corresponding to the χ^2 value is shown in Fig. 2.7.

A small p value, in this case, implies that it is unlikely to find any difference between the observed and expected values due to chance, in the absence of any substantial difference between the observed and expected customer visits.

Table 2.3 Survey data

	Exercise: Frequently	Exercise: None	Exercise: Some
Heavy	7	1	3
Never	87	18	84
Occasional	12	3	4
Regular	9	1	7

Since 0.04329 is smaller the defined significance level, we reject H_0 . We therefore conclude that there exists a significant difference between the observed and expected customer visits, i.e., it is not a good fit.

For the purpose of understanding the test for independence, we use the “survey” data table in the MASS library. In Table 2.3, the Smoker’s column records students’ smoking habit as either “Heavy”, “Regular”, “Occasional”, and “Never”, while the remaining columns record their exercise level as “Freq” (Frequently), “None”, and “Some”.

We would like to know if the smoking habit of students is independent of their exercise levels at a significance level, $\alpha = 0.05$ (5%).

The hypothesis space is defined as follows:

H_0 : Students smoking habit and exercise level are independent.

H_a : Students smoking habit and exercise level are not independent.

```
# TEST for INDEPENDENCE
chisq.test(Survey.data, simulate.p.value = TRUE)
```

Pearson’s chi-squared test with simulated p value (based on 2000 replicates)

```
data: Survey.data
X-squared = 5.4885, df = NA, p value = 0.4868
```

As the p value is greater than 0.05 (the level of significance), we accept the null hypothesis and we can conclude that the smoking habit of the students is independent of the exercise level they indulge.

2.13 Ratio Distributions

A ratio distribution is the ratio of two random variables having two different known distributions.

2.13.1 Student's *t*-Distribution

The *t*-distribution is the ratio of a normal random variable and an independent chi-distributed random variable (i.e., the square root of a chi-squared distribution).

If **U** and **V** are two random variables, where **U** is a standard normal distribution and **V** is a chi-squared distribution with *m* degrees of freedom, then the student's *t*-distribution having *m* degrees of freedom can be written as

$$t = \frac{U}{\sqrt{\frac{V}{m}}} \sim t_m \quad (2.13.1)$$

The *t*-distribution is used to estimate population parameters when the sample size is small and/or when the population variance σ^2 is unknown. *t*-distributions describe the samples drawn from a full population (where the full population is described by a normal distribution). The *t*-distribution for different sample sizes are different—the larger the sample, the more it resembles a normal distribution.

The student *t*-distribution with different degrees of freedom is shown in Fig. 2.8.

t-Statistic

The *t*-statistic in a *t* test is a test statistic and is used to compare the actual sample mean and the population mean. A significant difference indicates that the hypothesized value for μ should be rejected. The *t* test uses the *t*-statistic, *t*-distribution, and degrees of freedom to find the *p* value. The *p* value determines whether the population means differ.

The most common hypothesis test involves testing the null hypothesis:

$H_0 : \hat{\mu} - \mu = 0$, i.e., there is no difference between sample and population mean.

$H_a : \hat{\mu} - \mu \neq 0$, i.e., there exists a difference between the sample and population means.

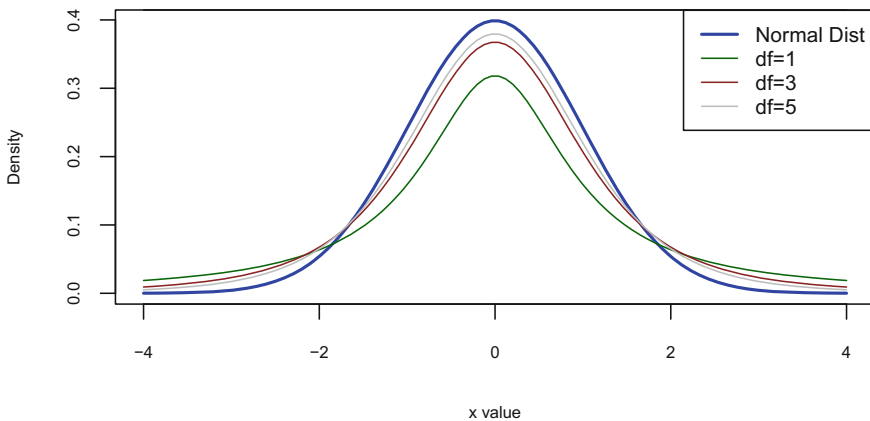


Fig. 2.8 Plot of *t*-distributions, with different degrees of freedom

The hypothesis test attempts to decide the following:

- Is the difference between $\hat{\mu}$ and μ simply due to chance (sampling error)?
- Is the discrepancy between $\hat{\mu}$ and μ more than as would be expected by chance, i.e., it tests whether the sample mean is significantly different from the population mean.

The critical step for the above hypothesis test is to calculate exactly how much difference between $\hat{\mu}$ and μ is reasonable to expect. And, this depends on the **standard error** $SE(\hat{\mu})$, which tells us the average amount by which the estimate $\hat{\mu}$ differs from the actual value of μ , and is written as follows:

$$SE(\hat{\mu}) = s/\sqrt{n} \tag{2.13.2}$$

s is the sample standard deviation and n is the sample size
Equation 2.13.2 also tells us that the deviation, $SE(\hat{\mu})$, shrinks with larger values of n .

The t-statistic is thus defined as

$$t = \frac{(\hat{\mu} - \mu)}{SE(\mu)} \tag{2.13.3}$$

The above equation measures the number of standard deviations, $\hat{\mu}$ is from μ . The t-statistic thus defines how much difference between $\hat{\mu}$ and μ is reasonable to expect—if the ratio is large, the difference would be significantly greater than what could be attributed to chance and accordingly we reject H_0 . But, if the ratio is small, the difference is not significant and we accept H_0 .

For the purpose of our analysis, we use the mtcars data set (from the *stats* library). This data set was extracted from the 1974 Motor Trend US magazine, and comprises 11 different aspects of automobile design and performance for 32 automobile models (1973 to 1974). Table 2.4 shows a brief description of the variables in the data set:

Table 2.4 mtcars data

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

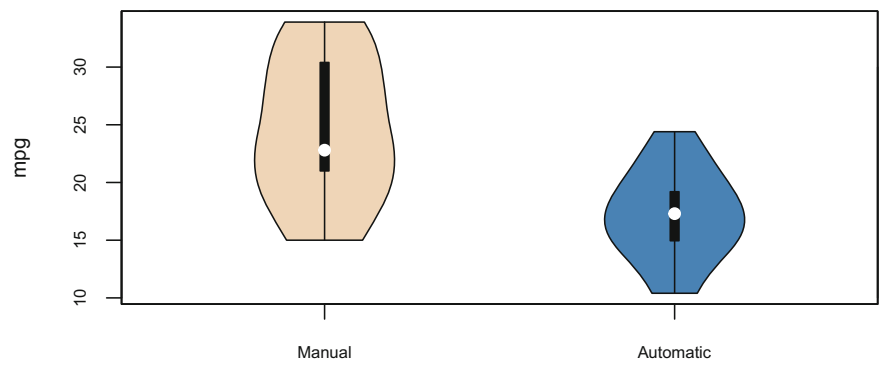


Fig. 2.9 Violin Plot of Transmission vs MPG

Table 2.5 T Test Results

t.statistic	df	p.value	automatic.mean	manual.mean
−3.767	18.332	0.001	17.147	24.392

It is worthwhile to visualize in Fig. 2.9 how MPG (miles per gallon) scores, for “automatic” and “manual” transmission, respectively.

It appears from the plot that automatic cars have lower miles per gallon than manual cars. To find out, which of the two, automatic or manual transmission, is better for MPG, we proceed as follows:

H_0 : Cars with automatic transmission use more fuel than cars with manual transmission.

H_a : Cars with automatic transmission **do not** use more fuel than cars with a manual transmission.

The two-sample t test is used to compare the means of the two samples (if they have different means) and the results are shown in Table 2.5.

```
# TESTING SIGNIFICANCE for TWO MEANS
test <- t.test(mpg ~ am, data = mtcars)
```

The p value (i.e., the probability of the difference in means, between the two groups) is very low (much lower than 0.05), and therefore we reject H_0 , i.e., cars with automatic transmission **do not** use more fuel than cars with a manual transmission. The t test shows that the apparent pattern in Fig. 2.9 happened by random chance, i.e., the samples that were picked were a group of automatic cars with high fuel efficiency and a group of manual cars with low fuel efficiency.

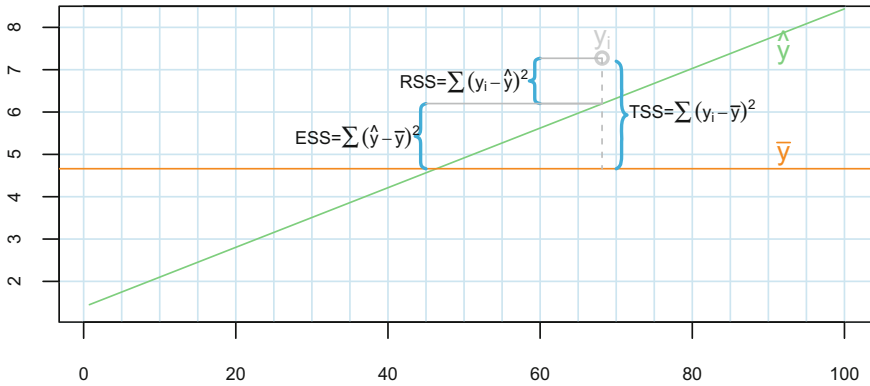


Fig. 2.10 Explained and unexplained variances of a regression model

2.13.2 *F-Distribution*

The F-distribution arises while dealing with ratios of variances. The F-distribution can be used among others for testing the equality of two population variances and testing the validity of a multiple regression model. The F-distribution has two important properties:

- It is defined only for positive values.
- It is positively skewed and thus not symmetric about its mean.

F-Statistic

An F-Statistic is a value obtained during a regression analysis to find out if the means between two populations are significantly different. It is similar to a t-statistic; a t-statistic tells us if a single variable is statistically significant and an F-statistic tells us if a group of variables are jointly significant.

The F-Statistic is defined as the quotient of the following ratio:

$$F = \frac{\text{Effect (Explained Variance)}}{\text{Error (Unexplained Variance)}}$$

The explained and unexplained variance of a regression model are shown in Fig. 2.10.

Total Sum of Squares (TSS) measures the variation of y_i values around their mean \bar{y} . Residual Sum of Squares (RSS) is the variation attributable to factors other than the relationship between x and y . Explained Sum of Squares (ESS) is the explained variation attributable to the relationship between x and y . The total variation is made up of two parts and is represented as

$$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2 \quad (2.13.4)$$

As shown in Fig. 2.10, the explained/unexplained variance can be written as

$$TSS - RSS = \text{Explained Variance} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (2.13.5)$$

$$RSS = \text{Unexplained Variance} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.13.6)$$

The F-Test

In the process of estimating the parameters $\beta_0, \beta_1, \dots, \beta_n$, we are interested to find if at least one or a subset of the predictors X_1, X_2, \dots, X_n are useful in predicting the dependent variable, Y . The steps involved are as follows:

1. Hypotheses statement:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_n = 0$$

$$H_a : \text{For at least one of } j, \beta_j \neq 0$$

2. The F-test-statistic is computed assuming H_0 is true:

$$\begin{aligned} F &= \frac{\text{Explained Variance}}{\text{Unexplained Variance}} \\ &= \frac{(TSS - RSS)/p}{RSS/(m - p - 1)} \end{aligned} \quad (2.13.7)$$

3. Determine the p value of the F-statistic.
4. Reject the null hypothesis if the p value for the F-statistic is below the level of significance α (usually, 0.05).

The above implies that when there is no relationship between the response and the predictors (i.e., parameters β_1, \dots, β_n are all zero), F-statistic is very close to or

Table 2.6 Data on blood pressure, age, and weight

BP	Age	Wt
132	52	173
143	59	184
153	67	194
162	73	211
154	64	196
168	74	220
137	54	188
149	61	188
159	65	207
128	46	167
166	72	217

Table 2.7 Summary of linear regression model

	Estimate	Std. error	t value	Pr(> t)
(Intercept)	30.99	11.94	2.595	0.03186
Age	0.8614	0.2482	3.47	0.00844
Wt	0.3349	0.1307	2.563	0.03351

equal to 1. On the other hand, if there exists some relationship between the response and the predictors (i.e., parameters β_1, \dots, β_n are **not** all zero), we would expect the F-statistic to be greater than 1.

We use a multiple regression model on the data set shown in Table 2.6 to predict BP from the predictors “Age” and “Wt”. The summary of the regression model is shown in Table 2.7.

The regression equation is **BP** = 31 + 0.86**Age** + 0.33**Wt**

The model F-statistic = 168.8 corresponding to a very small p value. Thus, the probability of accepting the null hypothesis is extremely low, and therefore we can assume there is a statistically significant relationship between the variables.

This concludes our discussion on probability and distributions. We are now ready to discuss the concepts of machine learning.



<http://www.springer.com/978-981-10-6807-2>

Machine Learning with R

Ghatak, A.

2017, XIX, 210 p. 56 illus., Hardcover

ISBN: 978-981-10-6807-2