Introduction to Probability and Statistics

Cheng Mao (Georgia Tech)

Contents

This set of notes is taken and rewritten from the book *Introduction to Probability and Statistics for Engineers and Scientists* by Sheldon M. Ross. The audience is mainly non-mathematicians, so the notes are written in a quantitative but not absolutely mathematically rigorous way. It is important to know:

1. The intuition of each concept or result. (Mathematics is largely about formalizing intuition.)
2. How to do computations. (Mathematical formulation allows us to solve seemingly hard problems.)

1 Random variables

1.1 Basic notions in probability

1.1.1 Random variable

Definition. A random variable (r.v.) X is an "experiment" whose outcome is uncertain.

Example. X is the outcome of a coin flip:

$$X = \begin{cases} 1 & \text{"heads"} \quad \text{w.p. } 1/2 \\ 0 & \text{"tails"} \quad \text{w.p. } 1/2 \end{cases}$$

where "w.p." means "with probability". We can also write

$$\mathbb{P}\{X = 1\} = 1/2, \qquad \mathbb{P}\{X = 0\} = 1/2.$$

Example. Let X be the number of "heads" we get in two coin flips:

$$X = \begin{cases} 0 & \text{w.p. } 1/4 \\ 1 & \text{w.p. } 1/2 \\ 2 & \text{w.p. } 1/4 \end{cases}$$

$$\mathbb{P}\{X = 0\} = 1/4, \qquad \mathbb{P}\{X = 1\} = 1/2, \qquad \mathbb{P}\{X = 2\} = 1/4$$

Example. X is the outcome of a race among three cars (denoted by 1, 2, 3):

$$X = \begin{cases} (1,2,3) & \text{w.p. } p_1 \\ (1,3,2) & \text{w.p. } p_2 \\ (2,1,3) & \text{w.p. } p_3 \\ (2,3,1) & \text{w.p. } p_4 \\ (3,1,2) & \text{w.p. } p_5 \\ (3,2,1) & \text{w.p. } p_6 \end{cases}$$

$$\mathbb{P}\{X = (1,2,3)\} = p_1, \qquad \mathbb{P}\{X = (1,3,2)\} = p_2, \qquad \dots$$

Here we may not know $p_1, \dots, p_6$, but $p_1 + p_2 + \cdots + p_6 = 1$.

Remark. Any discrete random variable can be specified in either of these two ways.

1.1.2 Distribution

Recall that a random variable is simply an experiment. A closely related notion is a "distribution" or "law".

Definition. A (probability) distribution is the law that describes an experiment. More formally, it gives the probabilities of occurrence of different possible outcomes for the experiment.

Example. Roll a die (i.e., dice) which gives 1, 2, 3, 4, 5, or 6 at random. Let X be the number we see, and let $Y = 7 - X$.

- Are X and Y the same random variable? Of course not, and they are never equal.
- Do X and Y follow the same distribution (or law)? Yes, because either of them is equal to 1, 2, 3, 4, 5, or 6 with probability 1/6 each.

Remark. Suppose X and Y follow the same distribution.

- It is possible that $\mathbb{P}\{X = Y\} = 1$: For example, define both of them to be the number we see when rolling a die.
- It is possible that $\mathbb{P}\{X = Y\} = 0$ or $\mathbb{P}\{X \neq Y\} = 1$: The above example.
- It is possible that X and Y are independent: For example, let X and Y be the numbers we see respectively when rolling two dice.

1.1.3 Sample space

Definition. A sample space S is the set of all possible outcomes of an experiment, i.e., the set of all possible values a random variable can take.

Example. X is the outcome of a coin flip:

$$S = \{0, 1\}$$

Example. X is the number of "heads" we get in two coin flips:

$$S = \{0, 1, 2\}$$

Example. X is the outcome of a race among three cars:

$$S = \{(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)\}$$

Fact. We always have $X \in S$ by definition, so

$$\mathbb{P}\{X \in S\} = 1.$$

1.1.4 Event

Definition. There are two ways to understand the concept "event":

1. An event is a subset E of the sample space S.
2. An event is literally "what happens in the experiment", i.e., $X \in E$.

Example. Let X be the number of "heads" we get in two coin flips. The event that we have at most one "heads", i.e., $X \leq 1$, can be written as

1. $E = \{0, 1\} \subset \{0, 1, 2\}$;
2. $\{X \leq 1\} = \{X \in E\}$.

The probability of the event is

$$\mathcal{P}(E) = \mathbb{P}\{X \in E\} = \mathbb{P}\{X \leq 1\} = 1/4 + 1/2 = 3/4.$$

(The notation $\mathcal{P}(E)$ is used here for brevity, but later in the course we mainly use the more intuitive notation $\mathbb{P}\{X \in E\}$.)

Example. X is the outcome of a race among three cars. The event that car 1 wins the race can be written as

1. $E = \{(1, 2, 3), (1, 3, 2)\}$;
2. $X \in E$.

We have

$$\mathcal{P}(E) = \mathbb{P}\{X \in E\} = \mathbb{P}\{X = (1,2,3)\} + \mathbb{P}\{X = (1,3,2)\} = p_1 + p_2.$$

Fact. We always have

$$0 \leq \mathcal{P}(E) = \mathbb{P}\{X \in E\} \leq 1.$$

1.1.5 Review of set algebra

Given events E and F understood as subsets of S, we have:

- Union $E \cup F$ (either event occurs)
- Intersection $E \cap F \equiv EF$ (both events occur)
- Complement $E^c = S \setminus E$ (one and only one of the two events occurs)
- Inclusion $E \subset F$ (if E occurs, then F must occur)

Useful laws:

- DeMorgan's law $(E \cup F)^c = E^c \cap F^c$, $(E \cap F)^c = E^c \cup F^c$
- Distributive law $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$, $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$

Venn diagram is useful:

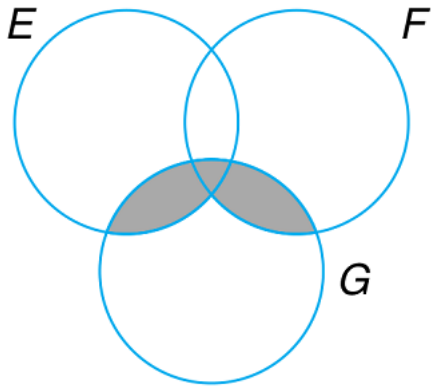


Figure 1: title

1.1.6 Probability of an event

Some simple facts:

- $\mathcal{P}(\varnothing) = 0$ where $\varnothing$ denotes the empty set
- $\mathcal{P}(S) = 1$
- $\mathcal{P}(E) \in [0,1]$ for $E \subset S$
- $\mathcal{P}(E^c) = 1 - \mathcal{P}(E)$
- If $E \subset F$, then $\mathcal{P}(E) \leq \mathcal{P}(F)$.
- If $E_1, \ldots, E_n$ are mutually exclusive, i.e., $E_i \cap E_j = \varnothing$ for $i \neq j$, then $\mathcal{P}(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n \mathcal{P}(E_i)$.
- For any $E_1, \ldots, E_n$, we have $\mathcal{P}(\bigcup_{i=1}^n E_i) \leq \sum_{i=1}^n \mathcal{P}(E_i)$. This is called the union bound.
- $\mathcal{P}(E \cup F) = \mathcal{P}(E) + \mathcal{P}(F) - \mathcal{P}(E \cap F)$

- $\mathcal{P}(E \cup F \cup G) = \mathcal{P}(E) + \mathcal{P}(F) + \mathcal{P}(G) - \mathcal{P}(E \cap F) - \mathcal{P}(E \cap G) - \mathcal{P}(F \cap G) + \mathcal{P}(E \cap F \cap G)$
- The odds of an event E is the ratio
$$\frac{\mathcal{P}(E)}{\mathcal{P}(E^c)} = \frac{\mathcal{P}(E)}{1 - \mathcal{P}(E)}.$$

1.1.7 Examples

Example. Draw a card at random from a standard 52-card deck. The random variable can be denoted by $Z = (X, Y)$, where X is the rank and Y is the suit. The sample space is

$$S = \{(x, y) : x = 1, \ldots, 13,\ y = C, D, H, S\}.$$

The events "Z is a jack, queen, or king" and "Z is a red 9, 10, or jack" can be formally described by

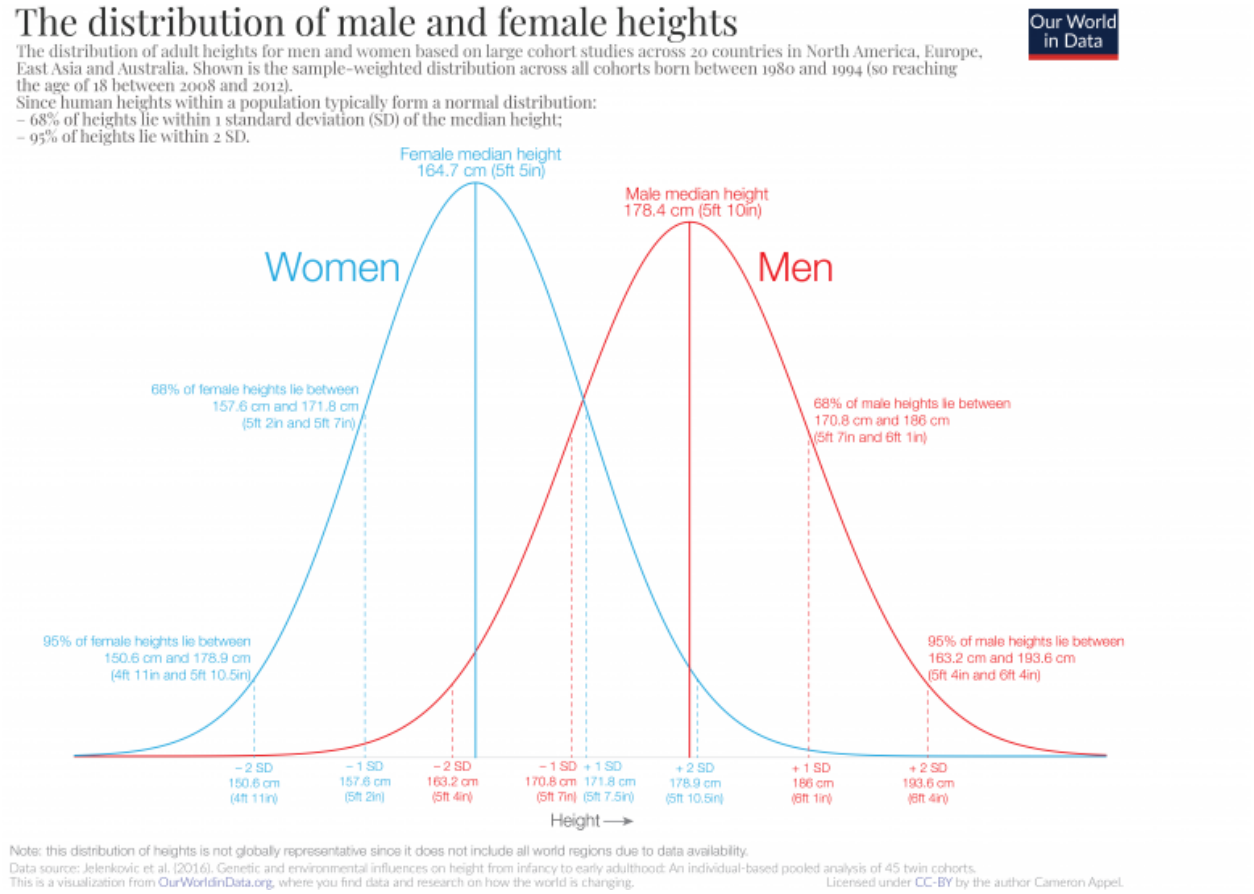
$$E = \{(x, y) : x = 11, 12, 13,\ y = C, D, H, S\}, \qquad F = \{(x, y) : x = 9, 10, 11,\ y = D, H\}$$

respectively. Then $E \cap F$ denotes the event "Z is a red jack", and $E \cup F$ denotes the event "Z is a jack, queen, king, red 9, or red 10". We have

$$\mathcal{P}(E \cap F) = \frac{2}{52} = \frac{1}{26}, \qquad \mathcal{P}(E \cup F) = \frac{12 + 4}{52} = \frac{4}{13}.$$

Example. For events $E, F \subset S$, suppose that $\mathcal{P}(E \cup F) = 0.76$ and $\mathcal{P}(E \cup F^c) = 0.87$. What is $\mathcal{P}(E)$?

Let $a = \{E \setminus F\}$, $b = \{F \setminus E\}$, $c = \{E \cap F\}$, and $d = E^c \cap F^c$. Then $\mathcal{P}(E \cup F) = a + c + b = 0.76$, $\mathcal{P}(E \cup F^c) = a + c + d = 0.87$, and $\mathcal{S} = a + b + c + d = 1$. Therefore, we can solve these equations to obtain $\mathcal{P}(E) = a + c = 0.63$.



1.2 Discrete random variables

In the previous section, we specify distributions and random variables using a case-by-case method. In general, how can we describe a distribution more formally?

Definition. The probability mass function (mass or PMF) f of a discrete random variable X taking values in S is defined by

$$f(x) = \mathbb{P}\{X = x\}$$

for $x \in S$. That is, $f(x)$ is the probability that X is equal to x.

Definition. The cumulative distribution function (distribution function or CDF) F of a real-valued random variable X is defined by

$$F(x) = \mathbb{P}\{X \leq x\}$$

for $x \in \mathbb{R}$. That is, $F(x)$ is the probability that X takes a value less than or equal to x.

We also say that f and F are the PMF and the CDF of the distribution of X, respectively.

Fact. It holds that

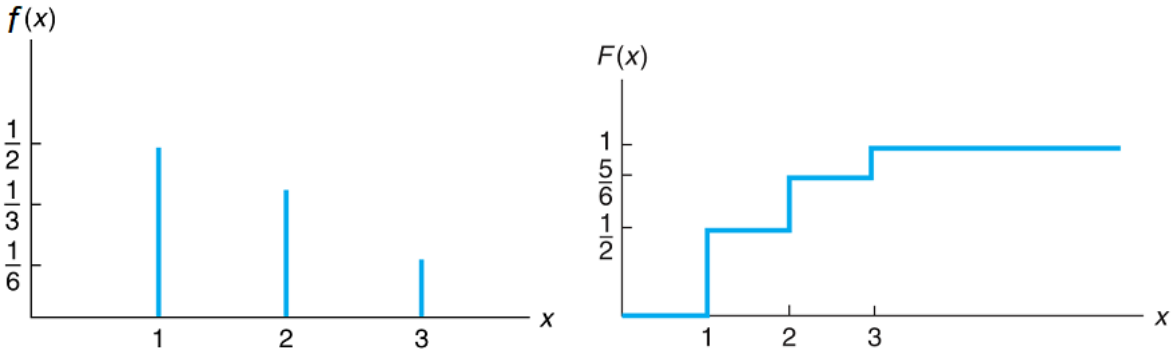
$$\mathbb{P}\{X \in S\} = \sum_{x \in S} f(x) = 1.$$

If X is integer-valued, then

$$F(x) = \sum_{y \leq x} f(y), \qquad f(x) = F(x) - F(x-1).$$

If X takes values in $\{x_1, x_2, \dots\} \subset \mathbb{R}$ where $x_{i-1} < x_i$, then

$$F(x_i) = \sum_{j=1}^{i} f(x_j), \qquad f(x_i) = F(x_i) - F(x_{i-1}).$$



Fact. We have

- $\mathbb{P}\{X \in E\} = \sum_{x \in E} f(x)$ for any event $E \subset S$;
- $\mathbb{P}\{a < X \leq b\} = F(b) - F(a)$ if X is real-valued.

Example. For the random variable X given by the above picture, we have

$$\mathbb{P}\{X \leq 2\} = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}.$$

1.3 Continuous random variables

If X is a continuous random variable, the CDF of X can still be defined in the same way.

Fact. The CDF F of a real-valued random variable X is nondecreasing and satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Example. Let X be a number chosen uniformly at random from the interval $[0, 1]$. Then the CDF F of X is given by $F(x) = 0$ if $x < 0$, $F(x) = x$ if $x \in [0, 1]$, and $F(x) = 1$ if $x > 1$.

Example. Let the CDF F of X be defined by $F(x) = 0$ if $x \le 0$ and $F(x) = 1 - e^{-x^2}$ if $x > 0$. From this definition, we can compute, for example,

$$\mathbb{P}\{1 < X \le 2\} = \mathbb{P}\{X \le 2\} - \mathbb{P}\{X \le 1\} = e^{-1} - e^{-4}.$$

Fact. For a continuous random variable X, we have

- $\mathbb{P}\{X = x\} = 0$;
- $F(x) = \mathbb{P}\{X \le x\} = \mathbb{P}\{X < x\}$;
- $\mathbb{P}\{a \le X \le b\} = \mathbb{P}\{a < X < b\}$.

What is the analogy of PMF for a continuous random variable X?

Definition. The probability density function (density or PDF) of a continuous real-valued random variable X is the function f such that

$$\mathbb{P}\{X \in E\} = \int_E f(x)\,dx \quad \text{for any } E \subset \mathbb{R}.$$

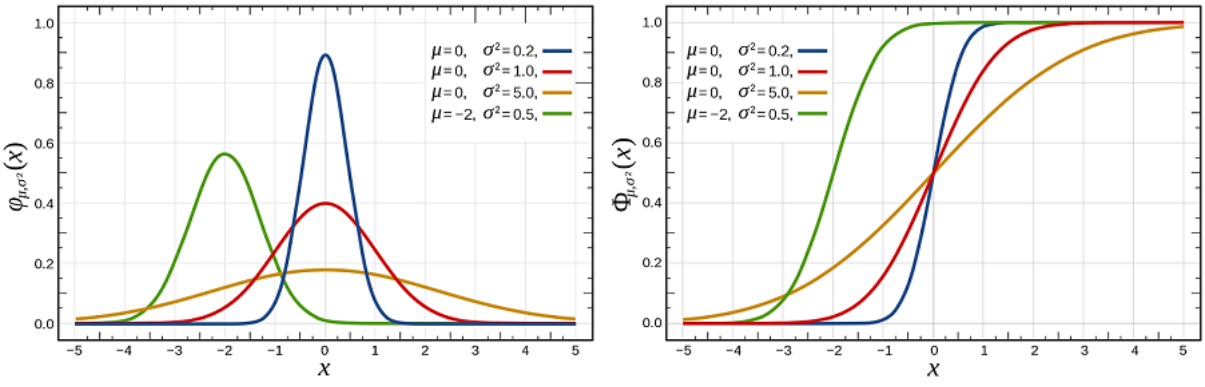
We also say that f is the PDF of the distribution of X.

Fact. It holds that

$$\mathbb{P}\{X \in \mathbb{R}\} = \int_{\mathbb{R}} f(x)\,dx = 1.$$

Moreover, we have

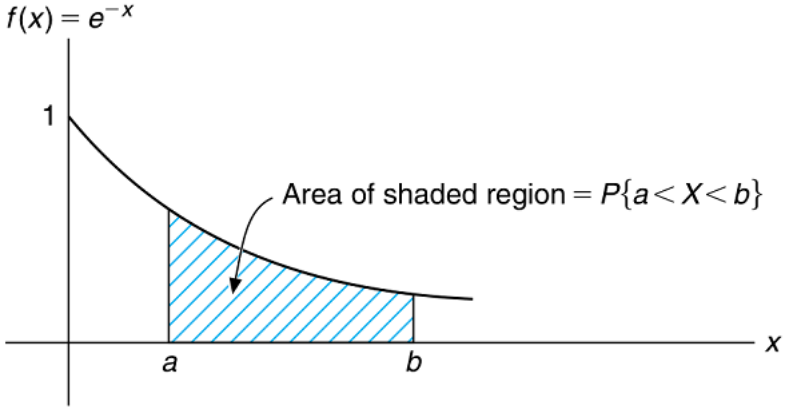
$$F(x) = \mathbb{P}\{X \le x\} = \int_{-\infty}^{x} f(y)\,dy, \qquad f(x) = F'(x) \quad \text{if } F \text{ is differentiable.}$$



Fact. We have

- $\mathbb{P}\{X = a\} = \int_a^a f(x)\,dx = 0$;
- $\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)\,dx = F(b) - F(a)$.

Intuition: $f(x)\,dx$ is the probability that X is in an infinitesimal neighborhood of x. Moreover, we have:



Example. Let X be a number chosen uniformly at random from the interval $[0, 1]$. Then the PDF f of X is given by $f(x) = 1$ if $x \in [0, 1]$ and $f(x) = 0$ otherwise.

Example. Let the CDF F of X be defined by $F(x) = 0$ if $x \leq 0$ and $F(x) = 1 - e^{-x^2}$ if $x > 0$. Then the PDF f of X is given by $f(x) = 2xe^{-x^2}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$.

Example. Let X be a random variable with PDF $f(x) = c\,x^2$ for $x \in [0, 1]$ and $f(x) = 0$ otherwise. What is c? What is $\mathbb{P}\{X \leq 0.5\}$?

We have

$$1 = \int_0^1 f(x)\,dx = \int_0^1 c\,x^2\,dx = \frac{c}{3}x^3\Big|_0^1 = \frac{c}{3},$$

so $c = 3$. Moreover,

$$\mathbb{P}\{X \leq 0.5\} = \int_0^{0.5} 3x^2\,dx = x^3\Big|_0^{0.5} = 0.125.$$

1.4 Expectation

Example. Let X be the number of "heads" we see when flipping two fair coins. What is the "expected value" of X? Intuitively, it should be

$$0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Definition. Let X be a discrete random variable taking values in S. The expectation of X is

$$\mathbb{E}[X] := \sum_{x \in S} x \cdot \mathbb{P}\{X = x\} = \sum_{x \in S} x \cdot f(x).$$

(The sum can be finite or infinite.)

Definition. Let X be a continuous random variable taking values in $\mathbb{R}$. The expectation of X is

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f(x)\,dx.$$

The expectation of X is also called the expected value or the mean of X.

Example. If X is uniformly distributed in $[0, 1]$, then $f(x) = 1$ if $x \in [0, 1]$ and $f(x) = 0$ if $x \notin [0, 1]$. Thus,

$$\mathbb{E}[X] = \int_0^1 x\,dx = 0.5.$$

Example. Let X be a random variable and E be an event. If I is the indicator random variable for the event E, i.e.,

$$I = \mathbf{1}\{X \in E\} = \begin{cases} 1 & \text{if } X \in E, \\ 0 & \text{if } X \notin E, \end{cases}$$

then

$$\mathbb{E}[I] = 1 \cdot \mathbb{P}\{X \in E\} + 0 \cdot \mathbb{P}\{X \notin E\} = \mathbb{P}\{X \in E\}.$$

1.5 Function of a random variable

Fact. If X is a random variable, then $g(X)$ is also a random variable for any function g. Its expectation is denoted by $\mathbb{E}[g(X)]$.

Example. Let X be the number of "heads" we see when flipping two fair coins. Then we have

$$\mathbb{E}[X^2] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = \frac{3}{2}.$$

Example. Let X be uniformly distributed in $[0, 1]$. Then we have

$$\mathbb{E}[X^3] = \int_0^1 x^3 \, dx = \frac{1}{4}.$$

Fact. Let X be a real-valued random variable taking values in S, and let $g : \mathbb{R} \to \mathbb{R}$ be a function.

- If X is discrete with PMF f, then

$$\mathbb{E}[g(X)] = \sum_x g(x) \cdot f(x).$$

- If X is continuous with PDF f, then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \, dx.$$

(These formulas are supposed to be very intuitive. It takes little effort to recall them once you understand what they are saying.)

Fact. The following properties are called the linearity of the expectation:

- If X is a random variable and a and b are constants, then $\mathbb{E}[b] = b$ and $\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$.
- If further g and h are real-valued functions, then $\mathbb{E}[a\,f(X) + b\,g(X)] = a\,\mathbb{E}[f(X)] + b\,\mathbb{E}[g(X)]$.

For example, if X is discrete, the first formula above holds since

$$\sum_{x \in S} (ax + b) \cdot f(x) = a \sum_{x \in S} x \cdot f(x) + b \sum_{x \in S} f(x).$$

If X is continuous, we can simply replace the sums with integrals.

1.6 Mean, variance, and moments

Definition. For a real-valued random variable X and any integer $k \geq 1$, the quantity $\mathbb{E}[X^k]$ is called the k-th moment X. In particular, the first moment $\mu := \mathbb{E}[X]$ is called the mean of X.

Definition. The quantity $\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2]$ is called the variance of X. Its square root $\sqrt{\mathrm{Var}(X)}$ is called the standard deviation of X.

Fact. It holds that

$$\mathrm{Var}(X) = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mathbb{E}[\mu^2] = \mathbb{E}[X^2] - \mu^2.$$

The following form is easier to remember:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Fact. For constants $a, b \in \mathbb{R}$, we have

$$\mathrm{Var}(aX + b) = \mathbb{E}\Big[\big(aX + b - (a\,\mathbb{E}[X] + b)\big)^2\Big] = a^2\,\mathbb{E}[(X - \mu)^2] = a^2\,\mathrm{Var}(X).$$

Example. Let X be the number of "heads" we see when flipping two fair coins. What is $\mathrm{Var}(X)$?

1. $\mathbb{E}[(X - \mu)^2] = (0 - 1)^2 \cdot \frac{1}{4} + (1 - 1)^2 \cdot \frac{1}{2} + (2 - 1)^2 \cdot \frac{1}{4} = \frac{1}{2}$;
2. $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{3}{2} - 1^2 = \frac{1}{2}$.

Example. Let X be uniformly distributed in $[0, 1]$. What is $\mathrm{Var}(X)$?

1. $\mathbb{E}[(X - \mu)^2] = \int_0^1 (x - 1/2)^2\,dx = \int_{-1/2}^{1/2} x^2\,dx = 1/12$;
2. $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \int_0^1 x^2\,dx - (1/2)^2 = 1/3 - 1/4 = 1/12$.

1.7 Moment-generating function

Definition. For a random variable X, the moment-generating function (MGF) $M(t)$ is defined by

$$M(t) = \mathbb{E}[e^{tX}].$$

If X is discrete,

$$M(t) = \sum_x e^{tx} p(x).$$

If X is continuous,

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx.$$

Fact. $M(t)$ is called the MGF because:

- $M'(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}[\frac{d}{dt}e^{tX}] = \mathbb{E}[Xe^{tX}]$, so
$$M'(0) = \mathbb{E}[X].$$
- $M''(t) = \frac{d}{dt}\mathbb{E}[Xe^{tX}] = \mathbb{E}[X\frac{d}{dt}e^{tX}] = \mathbb{E}[X^2e^{tX}]$, so
$$M''(0) = \mathbb{E}[X^2].$$
- In general, the nth derivative of $M(t)$ evaluated at $t = 0$ is equal to the nth moment of X. That is,
$$M^{(n)}(0) = \mathbb{E}[X^n], \quad n \geq 0.$$

Fact. The MGF $M(t)$ of a random variable X determines the distribution of X. Every distribution has its unique MGF.

2 Special distributions

2.1 Bernoulli distribution

We say that X is a Bernoulli random variable with parameter $p \in [0,1]$ and write $X \sim \text{Ber}(p)$ if

$$\mathbb{P}\{X = 1\} = p, \qquad \mathbb{P}\{X = 0\} = 1 - p.$$

That is, X is whether a trial is a "success" if the probability of "success" is p.

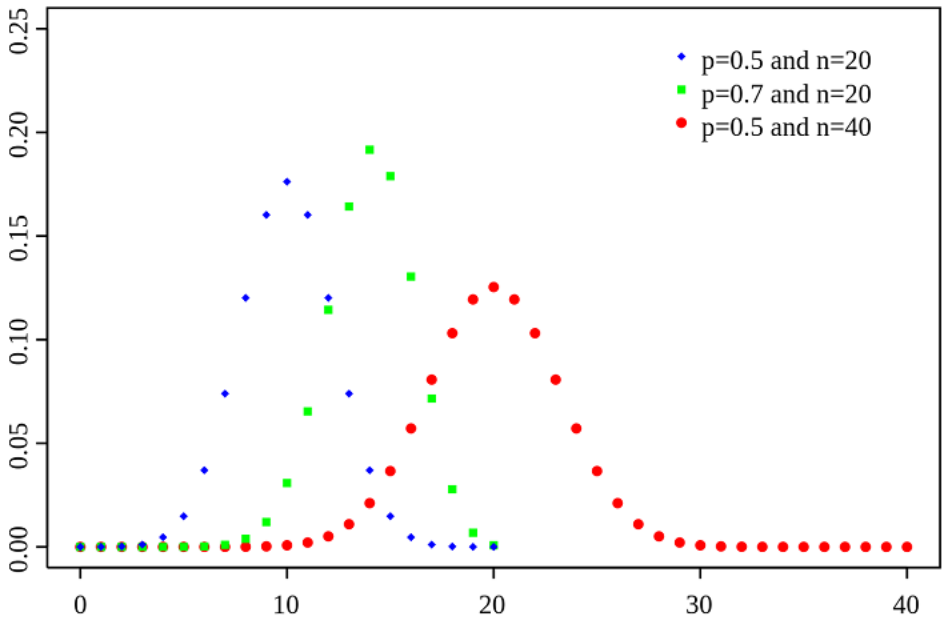
- Sample space: $\{0, 1\}$;
- Mean: $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$;
- Variance: $\text{Var}(X) = \mathbb{E}[(X - p)^2] = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p)$.

2.2 Binomial distribution

We say that X is a binomial random variable with parameters $n \geq 1$ and $p \in [0,1]$ and write $X \sim \text{Bin}(n, p)$ if X is the number of successes in n independent trials, each with success probability p. In other words, X has PMF

$$f(i) = \mathbb{P}\{X = i\} = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, 1, 2, \ldots, n.$$

(In case you have not seen this notation before, we have, for example, $\binom{9}{3} = \frac{9 \cdot 8 \cdot 7}{3 \cdot 2 \cdot 1}$.)



In R, for example, we can obtain the PMF $f(i)$ of $\text{Bin}(5, 1/6)$ at $i = 0, 1, \ldots, 5$ as follows:

```
dbinom(0:5, 5, 1/6)
```

```
## [1] 0.4018775720 0.4018775720 0.1607510288 0.0321502058 0.0032150206
## [6] 0.0001286008
```

Although we have not properly defined "independence" of random variables, it is true that, if $X_1, \ldots, X_n$ are independent $\text{Ber}(p)$ random variables, then $X = \sum_{i=1}^{n} X_i$ is a $\text{Bin}(n, p)$ random variable. Using facts about a sum of independent random variables (to be proved later), we can derive the following.

- Sample space: $\{0, 1, 2, \ldots, n\}$;

- Mean: $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$;
- Variance: $\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$.

Example. Roll five dice. The probability that we see exactly two dice showing the number 6 is

$$\binom{5}{2}(1/6)^2(5/6)^3.$$

Example. A system consists of 5 components, each of which will function with probability p independently. The system will be able to operate if at least a half of its components function. The probability that the system is able to operate is equal to

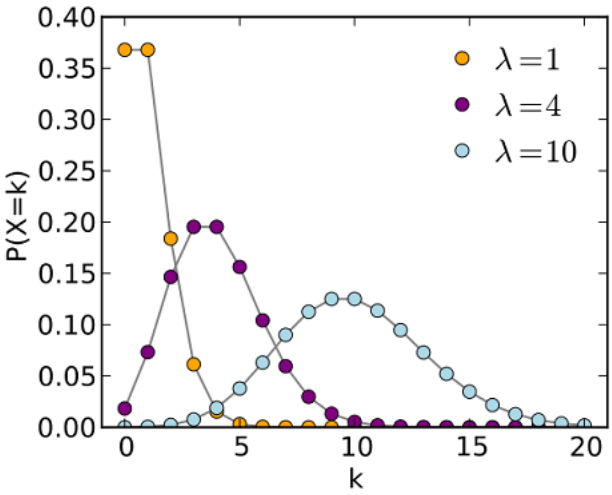
$$\binom{5}{3}p^3(1-p)^2 + \binom{5}{4}p^4(1-p) + p^5.$$

2.3 Poisson distribution

We say that X is a Poisson random variable with parameter $\lambda > 0$ and write $X \sim \mathrm{Poi}(\lambda)$ if

$$\mathbb{P}\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots,$$

where $e \approx 2.7183$ is Euler's number.



- Sample space: $\{0, 1, 2, \ldots\}$;
- MGF:

  $$M(t) = \mathbb{E}[e^{tX}] = \sum_{i=0}^{\infty} e^{ti}e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=0}^{\infty}\frac{(\lambda e^t)^i}{i!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)},$$

  and thus

  $$M'(t) = \lambda e^t e^{\lambda(e^t-1)}, \qquad M''(t) = (\lambda e^t)^2 e^{\lambda(e^t-1)} + \lambda e^t e^{\lambda(e^t-1)};$$

- Mean: $\mathbb{E}[X] = M'(0) = \lambda$;
- Variance: $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = M''(0) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Example. Suppose that the number of accidents occurring weekly on a particular stretch of a highway follows a Poisson distribution with mean equal to 3. What is the probability that there is at least one accident this week?

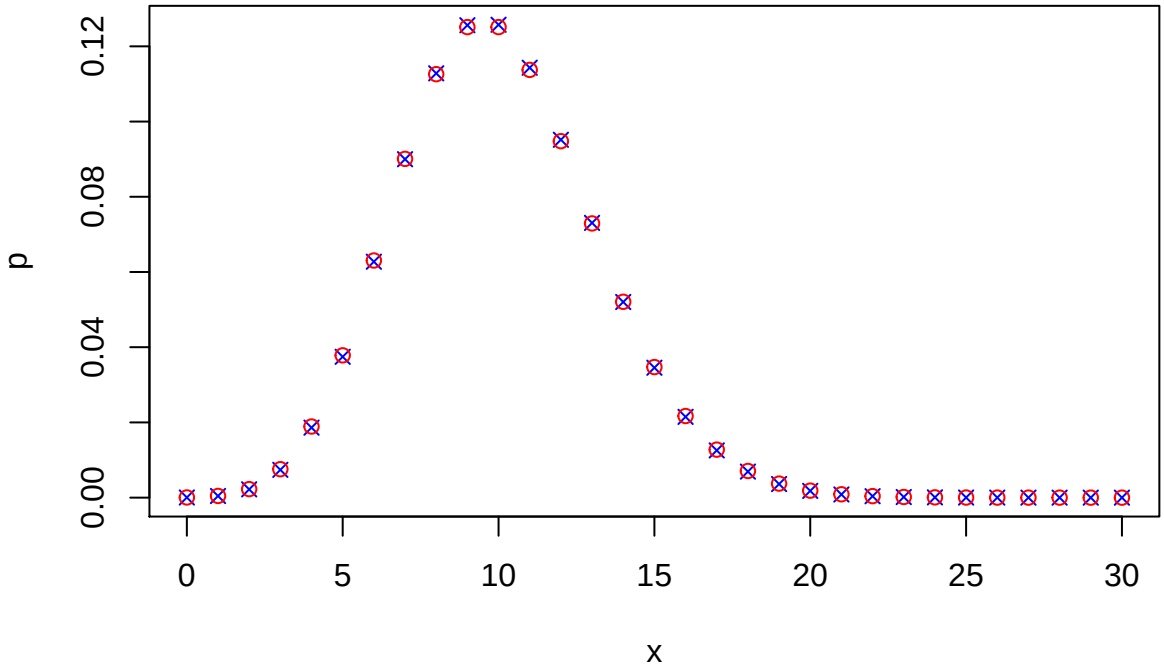
Let X be the number of accidents. As $X \sim \text{Poi}(3)$, we have

$$\mathbb{P}\{X \geq 1\} = 1 - \mathbb{P}\{X = 0\} = 1 - e^{-3}\frac{3^0}{0!} = 1 - e^{-3}.$$

Remark. For n "large" and p "small", the distributions $\text{Bin}(n, p)$ and $\text{Poi}(np)$ are "close" to each other.

Using R, we can compare the PMFs of $\text{Bin}(1000, 1/100)$ and $\text{Poi}(10)$ at values $0, 1, \ldots, 30$ (with the former denoted by blue crosses and the latter by red circles):

```
x <- c(0:30)
p = dbinom(x, 1000, 1/100)
q = dpois(x, 10)
plot(x, p, type="p", col="blue", pch=4)
lines(x, q, type="p", col="red")
```



Example. Suppose that each item produced by a certain machine is defective with probability 0.1 independently. What is the probability that there is at most 1 defective item in a sample of 10 items?

The number of defective items follows the $\text{Bin}(10, 0.1)$ distribution, so

$$\binom{10}{0}0.1^0 \cdot 0.9^{10} + \binom{10}{1}0.1^1 \cdot 0.9^9 \approx 0.7361.$$

How about the Poisson approximation, i.e., $\text{Poi}(1)$? We have

$$e^{-1}\frac{1^0}{0!} + e^{-1}\frac{1^1}{1!} \approx 0.7358.$$

2.4 Uniform distribution

Depending on the type of the sample space S, the uniform distribution, denoted by $\mathrm{Unif}(S)$, can be defined as follows.

- Discrete: A uniform random variable on a finite set S takes each value in S equally likely.
- Continuous: A uniform random variable on an interval $S = [a, b]$ has PDF $f(x) = 1/(b-a)$ for $x \in [a, b]$ and $f(x) = 0$ otherwise.
  - Sample space: $[a, b]$;
  - Mean: $\mathbb{E}[X] = (a + b)/2$;
  - Variance: $\mathrm{Var}(X) = \int_a^b (x - \frac{a+b}{2})^2 \frac{1}{b-a}\, dx = \frac{(b-a)^2}{12}$.
- 2D continuous: A uniform random variable on a region $S \subset \mathbb{R}^2$ has PDF $f(x, y) = 1/\mathrm{area}(S)$ for $(x, y) \in S$ and $f(x, y) = 0$ otherwise.

2.5 Normal distribution

We say that X is a normal (or Gaussian) random variable with mean μ and variance σ^2 and write $X \sim \mathcal{N}(\mu, \sigma^2)$ if it has PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- The sample space of X is $\mathbb{R}$.
- The mean of X is $\mathbb{E}[X] = \mu$, because

$$\begin{aligned}
\mathbb{E}[X - \mu] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2}\, dy \\
&= -\frac{\sigma}{\sqrt{2\pi}} e^{-y^2/2}\Big|_{-\infty}^{\infty} = 0.
\end{aligned}$$

- Let us check $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1$. To this end, let $s = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx$ and then

$$s^2 = \Big(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx \Big) \Big(\int_{-\infty}^{\infty} e^{-y^2/2}\, dy \Big) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy.$$

  By a change of variables $x = r \sin\theta$ and $y = r \cos\theta$ to the polar coordinates, we obtain

$$s^2 = \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}r^2} r\, dr\, d\theta = -2\pi e^{-\frac{1}{2}r^2}\Big|_0^{\infty} = 2\pi.$$

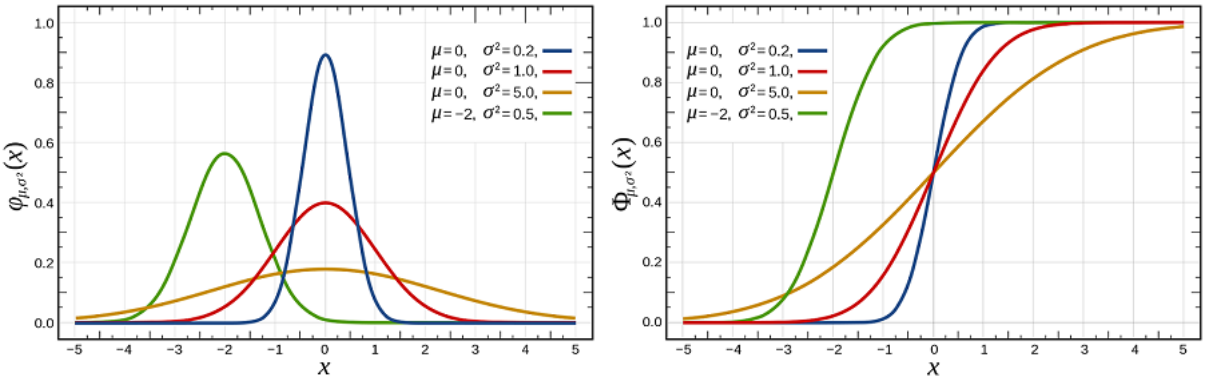
  The claim follows.

- The variance of X is $\mathrm{Var}(X) = \sigma^2$, because

$$\begin{aligned}
\mathbb{E}[(X - \mu)^2] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\, dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \Big(-y e^{-y^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \Big) \\
&= \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = \sigma^2.
\end{aligned}$$

- The MGF of X is $M(t) = e^{\mu t + \sigma^2 t^2/2}$, because

$$\begin{aligned} M(t) &= \int_{-\infty}^{\infty} \frac{e^{tx}}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2 - 2\mu x + \mu^2 - 2\sigma^2 t x}{2\sigma^2} \right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\big(x - (\mu + \sigma^2 t)\big)^2 - 2\mu\sigma^2 t - \sigma^4 t^2}{2\sigma^2} \right) dx \\ &= \exp\left(\frac{2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2} \right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\big(x - (\mu + \sigma^2 t)\big)^2}{2\sigma^2} \right) dx \\ &= \exp\left(\mu t + \sigma^2 t^2/2\right). \end{aligned}$$



2.5.1 Standard normal distribution

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and a and b are constants, then

$$a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2).$$

Hence, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

The random variable Z is said to have a standard normal distribution. The above process of translation and scaling is called standardization of the random variable X, which means making the random variable to have mean 0 and variance 1.

We use $\Phi(x)$ to denote the CDF of the standard normal distribution. There is no simple closed formula for $\Phi(x)$, but its value can be accessed in R as follows:

```
pnorm(1.5, 0, 1)
```

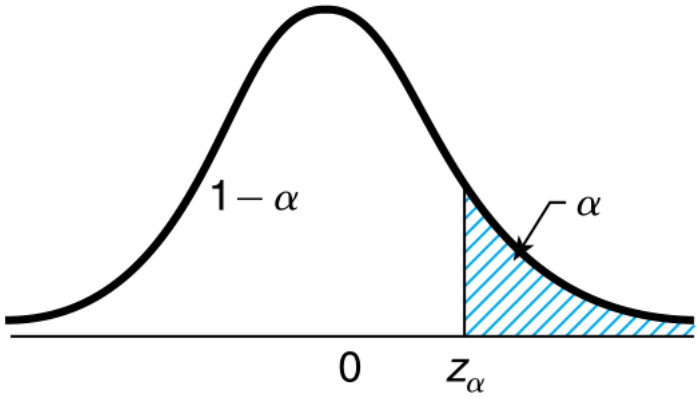
```
## [1] 0.9331928
```

By symmetry, for any $x \in \mathbb{R}$, we have

$$\Phi(-x) = \mathbb{P}\{Z \le -x\} = \mathbb{P}\{Z \ge x\} = 1 - \Phi(x).$$

Example. If $X \sim \mathcal{N}(3, 16)$, then

$$\mathbb{P}\{2 < X < 7\} = \mathbb{P}\Big\{\frac{2-3}{4} < \frac{X-3}{4} < \frac{7-3}{4}\Big\} = \Phi(1) - \Phi(-1/4) = \Phi(1) - 1 + \Phi(1/4).$$



2.5.2 Quantile

The inverse Φ^{-1} of the CDF Φ is called the quantile function of $\mathcal{N}(0,1)$. For $p \in (0,1)$, we have

$$\mathbb{P}\{Z \leq \Phi^{-1}(p)\} = p.$$

We call $\Phi^{-1}(p)$ the quantile of order p or the $(100p)$th percentile.

Moreover, for $\alpha \in (0,1)$, define $z_\alpha = \Phi^{-1}(1-\alpha)$ so that

$$\Phi(z_\alpha) = \mathbb{P}\{Z \leq z_\alpha\} = 1 - \alpha.$$

We have:

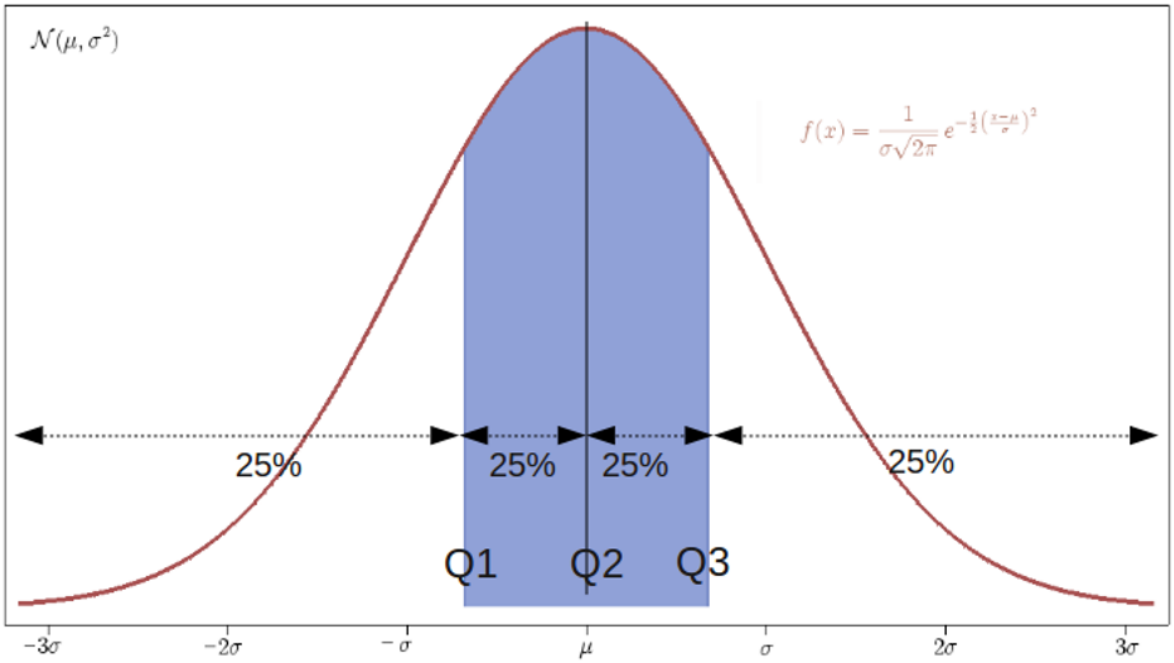
$$z_{0.05} \approx 1.645, \qquad z_{0.025} \approx 1.96, \qquad z_{0.01} \approx 2.33;$$

$$\mathbb{P}\{Z \leq 1.645\} \approx 95\%, \qquad \mathbb{P}\{Z \leq 1.96\} \approx 97.5\%, \qquad \mathbb{P}\{Z \leq 2.33\} \approx 99\%.$$

In other words, 1.645 is the 95th percentile, 1.96 is the 97.5th percentile, and 2.33 is the 99th percentile. The value of z_α can be accessed in R as follows:

```
qnorm(1-0.067, 0, 1)
```

```
## [1] 1.498513
```



There are many other special distributions, for example:

- Discrete: Rademacher distribution, geometric distribution, negative binomial distribution, and hypergeometric distribution;
- Continuous: chi-squared distribution, t-distribution, F-distribution, Beta distribution, Gamma distribution, Cauchy distribution, Laplace distribution, and Pareto distribution.

Wikipedia has a page dedicated to every commonly used distribution, so it is a great place to look up the PMF or PDF, CDF, mean, variance, MGF, and other properties of a distribution. In general, it is important to choose an appropriate distribution when fitting data.

3 Joint random variables

3.1 Definitions

Definition. Let X and Y be random variables taking values in sample spaces S and T respectively. Then the pair (X, Y) is called a joint random variable taking values in the sample space denoted by $S \times T$. A joint random variable with three or more components can be defined similarly.

Example. Recall the example of randomly selecting a card from a standard 52-card deck. Let X be the rank of the card and Y be its suit, which take values in

$$S = \{1, 2, 3, \ldots, 13\}, \qquad T = \{C, D, H, S\}$$

respectively. The joint random variable (X, Y) is the card, and $S \times T$ is the entire deck.

Definition. If discrete random variables X and Y take values in S and T respectively, the joint PMF of (X, Y) is given by

$$f(x, y) = \mathbb{P}\{X = x, Y = y\}$$

for $x \in S$ and $y \in T$.

Definition. The joint CDF of a pair of real-valued random variables (X, Y) is given by

$$F(x, y) = \mathbb{P}\{X \leq x, Y \leq y\}$$

for $x, y \in \mathbb{R}$.

Example. Let B be the number of boys and G be the number of girls in a family chosen at random from a community. The joint random variable (B, G) can be described by the following contingency table, each entry of which specifies the value of $\mathbb{P}\{B = i,\, G = j\}$:

i \ j	0	1	2	3	Row Sum $= P\{B = i\}$
0	.15	.10	.0875	.0375	.3750
1	.10	.175	.1125	0	.3875
2	.0875	.1125	0	0	.2000
3	.0375	0	0	0	.0375
Column Sum $= P\{G = j\}$	.3750	.3875	.2000	.0375	

For example, we can read from the table that 15 percent of families have no children, 20 percent have 1, 35 percent have 2, and 30 percent have 3.

Definition. If X and Y are real-valued continuous random variables, then the joint PDF of (X, Y) is the function $f(x, y)$ such that

$$\mathbb{P}\{(X, Y) \in B\} = \iint_{(x,y)\in B} f(x, y)\, dx\, dy.$$

Fact. In particular, for $B = (-\infty, a] \times (-\infty, b]$, we have

$$F(a, b) = \mathbb{P}\{X \le a, Y \le b\} = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\, dx\, dy,$$

and

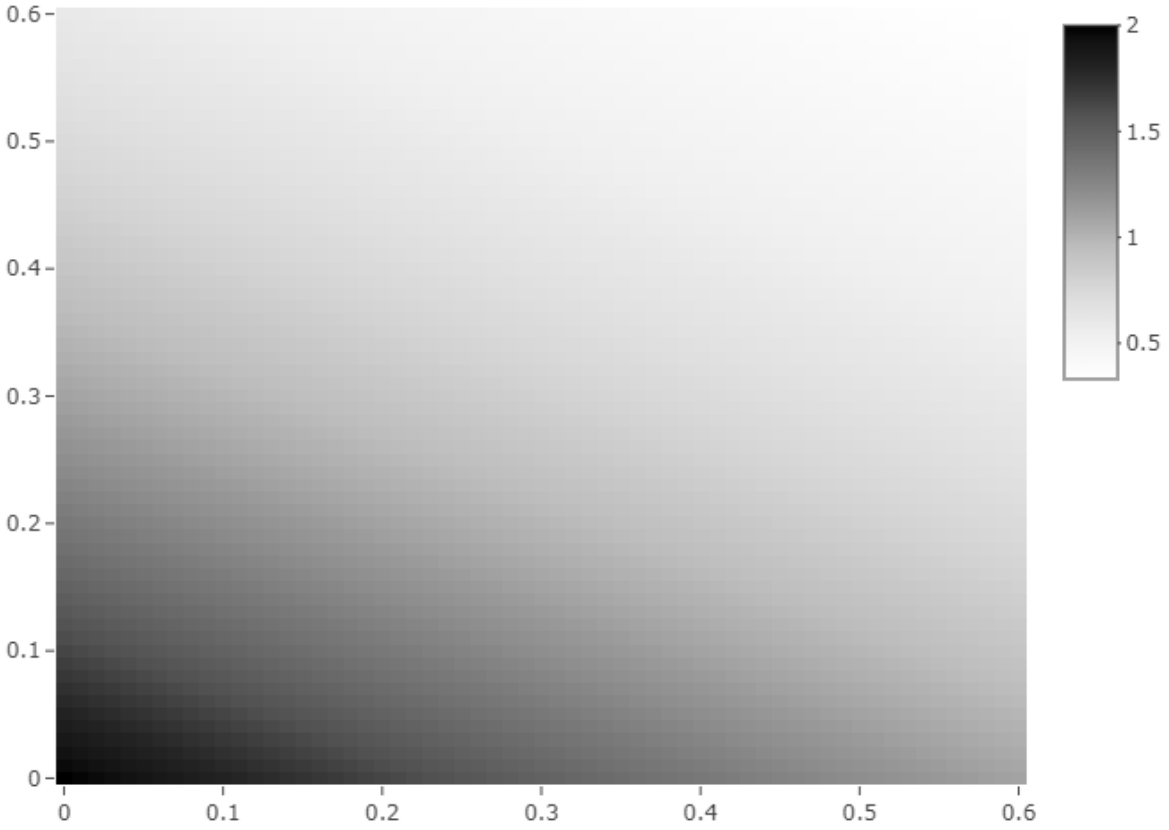
$$f(a, b) = \frac{\partial^2}{\partial a\, \partial b} F(a, b).$$

Intuition: $f(x, y)\, dx\, dy$ is the probability that (X, Y) is in an infinitesimal neighborhood of (x, y).

Example. Suppose that X and Y are continuous with joint PDF

$$f(x, y) = 2e^{-x}e^{-2y}, \quad x > 0, y > 0,$$

and $f(x, y) = 0$ otherwise.



Recall that $\mathbb{P}\{(X,Y) \in B\} = \iint_{(x,y)\in B} f(x,y)\,dx\,dy$, so we have

$$\mathbb{P}\{X > 1, Y < 1\} = \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}\,dx\,dy = e^{-1}(1-e^{-2}),$$
$$\mathbb{P}\{X < a\} = \int_0^\infty \int_0^a 2e^{-x}e^{-2y}\,dx\,dy = 1-e^{-a},$$
$$\mathbb{P}\{X < Y\} = \int_0^\infty \int_0^y 2e^{-x}e^{-2y}\,dx\,dy = 1/3.$$

3.2 Marginal distribution

Definition. When X is a component of a joint random variable (X,Y), we refer to the distribution of X as the marginal distribution of X.

The CDF F of (X,Y) is called the joint CDF of (X,Y), and the CDF F_X of X is called the marginal CDF of X. Similarly, we can define the marginal CDF F_Y.

The joint PMF f and the marginal PMFs f_X and f_Y are defined analogously.

Fact. The marginal CDF of X in the joint random variable (X,Y) is given by

$$F_X(x) = \mathbb{P}\{X \le x\} = \mathbb{P}\{X \le x, Y < \infty\} = F(x,\infty).$$

Similarly, the marginal CDF of Y is given by $F_Y(y) = F(\infty, y)$.

Fact. If X and Y are discrete random variables taking values in S and T respectively, then the marginal PMF of X is

$$f_X(x) = \mathbb{P}\{X = x\} = \sum_{y\in T} \mathbb{P}\{X = x, Y = y\} = \sum_{y\in T} f(x,y)$$

for any $x \in S$. Similarly, $f_Y(y) = \sum_{x \in S} f(x, y)$.

Such a procedure of obtaining a marginal random variable X from a joint random variable (X, Y) is called marginalization.

Example. Revisit the above table.

Fact. If X and Y are continuous real-valued random variables, then the marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

for any $x \in \mathbb{R}$. Similarly, $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$.

Example. Revisit (X, Y) with joint PDF

$$f(x, y) = 2e^{-x}e^{-2y}, \quad x > 0, y > 0,$$

and $f(x, y) = 0$ otherwise. Then the marginal PDFs of X and Y are respectively

$$f_X(x) = \int_0^{\infty} 2e^{-x}e^{-2y}\, dy = e^{-x},$$
$$f_Y(y) = \int_0^{\infty} 2e^{-x}e^{-2y}\, dx = 2e^{-2y}.$$

where $x, y > 0$.

3.3 Independence

3.3.1 Independence of events

Definition. Two events E and F are independent if

$$\mathcal{P}(E \cap F) = \mathcal{P}(E) \cdot \mathcal{P}(F).$$

Otherwise, the two events are said to be dependent.

Fact. If E and F are independent, then the following pairs of events are also independent: (1) E and F^c; (2) E^c and F; (3) E^c and F^c.

Example. Randomly select a card from a standard 52-card deck. Let E be the event that the card is an ace, and let F be the event that the card is a heart. Intuitively, these two events should be independent. Indeed, we have $\mathcal{P}(E) = 1/13$, $\mathcal{P}(F) = 1/4$, and $\mathcal{P}(E \cap F) = 1/52 = \mathcal{P}(E) \cdot \mathcal{P}(F)$.

Definition. Three events E, F, and G are mutually independent if all the following conditions hold:

$$\mathcal{P}(E \cap F \cap G) = \mathcal{P}(E) \cdot \mathcal{P}(F) \cdot \mathcal{P}(G),$$
$$\mathcal{P}(E \cap F) = \mathcal{P}(E) \cdot \mathcal{P}(F), \qquad \mathcal{P}(E \cap G) = \mathcal{P}(E) \cdot \mathcal{P}(G), \qquad \mathcal{P}(F \cap G) = \mathcal{P}(F) \cdot \mathcal{P}(G).$$

Example. Roll two dice, and denote the two numbers we see by X and Y respectively.

- Consider the events $E = \{X + Y = 7\}$, $F = \{X = 3\}$, and $G = \{Y = 4\}$. Then

  $$\mathcal{P}(E) = \mathcal{P}(F) = \mathcal{P}(G) = 1/6, \qquad \mathcal{P}(E \cap F) = \mathcal{P}(E \cap G) = \mathcal{P}(F \cap G) = 1/36.$$

  Therefore, the events E, F, and G are pairwise independent. However, we have

  $$\mathcal{P}(E \cap F \cap G) = 1/36,$$

  so the three events are not mutually independent.

- Consider the events $E = \{X + Y = 8\}$, $F = \{X = 1\}$, and $G = \{Y = 0\}$. Although

$$\mathcal{P}(E \cap F \cap G) = 0 = \mathcal{P}(E) \cdot \mathcal{P}(F) \cdot \mathcal{P}(G),$$

we have

$$\mathcal{P}(E \cap F) = 0 \neq (5/36) \cdot (1/6) = \mathcal{P}(E) \cdot \mathcal{P}(F),$$

so the events E, F, and G are not mutually independent.

3.3.2 Independence of random variables

Definition. Two random variables X and Y taking values in S and T respectively are independent if for any events $E \subset S$ and $F \subset T$,

$$\mathbb{P}\{X \in E, Y \in F\} = \mathbb{P}\{X \in E\} \cdot \mathbb{P}\{Y \in F\}.$$

Remark. What does "independence" mean in words?

1. Two events are independent if the probability that both of them happen is equal to the probability that the first happens times the probability that the second happens.
2. Two random variables are independent if any two events involving the two random variables respectively are independent.

Example. Randomly select a card from a standard 52-card deck. Let X and Y denote the rank and the suit of the card respectively. Then X and Y are independent.

It is usually difficult to prove independence using the above definition. The following two characterizations of independence are helpful sometimes.

Fact. Two random variables X and Y taking values in S and T respectively are independent if and only if

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{for all } x \in S,\, y \in T,$$

or

$$F(x, y) = F_X(x) \cdot F_Y(y) \quad \text{for all } x \in S,\, y \in T.$$

(This is true regardless of whether X and Y are discrete or continuous.)

Example. Randomly select a card from a standard 52-card deck. Let X be the rank of the card, and let Y be its suit. Then the joint PMF of (X, Y) is

$$f(x, y) = \frac{1}{52} = \frac{1}{13} \cdot \frac{1}{4} = f_X(x) \cdot f_Y(y),$$

so X and Y are independent.

Example. Revisit (X, Y) with joint PDF

$$f(x, y) = 2e^{-x}e^{-2y}, \quad x > 0, y > 0,$$

and $f(x, y) = 0$ otherwise. We have shown that

$$f_X(x) = e^{-x}, \qquad f_Y(y) = 2e^{-2y},$$

where $x, y > 0$. Hence we immediately conclude that X and Y are independent.

Fact. Let $X_1, \ldots, X_n$ be n random variables taking values in $S_1, \ldots, S_n$ respectively. Denote their joint PMF or PDF by $f(x_1, \ldots, x_n)$, and denote their respective PMFs or PDFs by $f_1(x_1), \ldots, f_n(x_n)$. Then $X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n) \quad \text{for all } x_1 \in S_1, \ldots, x_n \in S_n.$$

3.4 Conditional probability

3.4.1 Definition and examples

Example. Roll two dice. Let E be the event that the sum of the two numbers is 8, and let F be the event that the first number is 3. What is the probability of E given that F occurs?

- $F = \{(3,1),(3,2),(3,4),(3,4),(3,5),(3,6)\}$
- $E = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$
- $E \cap F = \{(3,5)\}$

So the probability is $1/6$.

The sample space is the set of all 36 possible outcomes, so

$$\frac{1}{6} = \frac{1/36}{6/36} = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)}.$$

Definition. "The probability of E given that F occurs" is called a conditional probability, and is denoted by

$$\mathcal{P}(E \,|\, F) = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)}.$$

If we are talking about random variables X and Y, then we write

$$\mathbb{P}\{X \in E \,|\, Y \in F\} = \frac{\mathbb{P}\{X \in E, Y \in F\}}{\mathbb{P}\{Y \in F\}}.$$

Fact. We have $0 \le \mathcal{P}(E \,|\, F) \le 1$ and $\mathcal{P}(F \,|\, F) = 1$. If $E_1, \ldots, E_n$ are disjoint, then $\mathcal{P}(E_1 \cup \cdots \cup E_n \,|\, F) = \mathcal{P}(E_1 \,|\, F) + \cdots + \mathcal{P}(E_n \,|\, F)$.

Example. Flip two fair coins.

1. If one coin is revealed and it turns out to be "heads", what is the probability that both are "heads"?
2. If we know that at least one coin is "heads", what is the probability that both are "heads"?

In case 1, the probability is $1/2$. In case 2, the situation seems basically the same, but it is not. Let E be the event of having two "heads", and let F be the event of having at least one "heads". Then

$$\mathcal{P}(E \,|\, F) = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

The problem is that natural language is not precise. In terms of mathematical formulas, we can let X represent the coin that is revealed and let Y represent the other coin. Then the two probabilities of interest are respectively

$$\mathbb{P}\{X = 1, Y = 1 \,|\, X = 1\}, \qquad \mathbb{P}\{X = 1, Y = 1 \,|\, X + Y \ge 1\},$$

which are obviously different.

3.4.2 Law of total probability

Let E and F be events. Since $\mathcal{P}(E) = \mathcal{P}(E \cap F) + \mathcal{P}(E \cap F^c)$, we have

$$\mathcal{P}(E) = \mathcal{P}(E \,|\, F) \cdot \mathcal{P}(F) + \mathcal{P}(E \,|\, F^c) \cdot \mathcal{P}(F^c).$$

When considering random variables X and Y, this may be written as

$$\mathbb{P}\{X \in E\} = \mathbb{P}\{X \in E \,|\, Y \in F\} \cdot \mathbb{P}\{Y \in F\} + \mathbb{P}\{X \in E \,|\, Y \in F^c\} \cdot \mathbb{P}\{Y \in F^c\}.$$

Example. In answering a multiple-choice question, a student knows the answer with probability p, and guesses randomly with probability $1-p$. If there are m choices for the question, what is the conditional probability that the students knows the answer given that the question is answered correctly?

Let E be the event that the student knows the answer, and let F be the event that the student answers the question correctly. Then

$$\mathcal{P}(F) = \mathcal{P}(F \,|\, E) \cdot \mathcal{P}(E) + \mathcal{P}(F \,|\, E^c) \cdot \mathcal{P}(E^c) = p + \frac{1-p}{m}$$

and so

$$\mathcal{P}(E \,|\, F) = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)} = \frac{p}{p + (1-p)/m} = \frac{mp}{mp - p + 1}.$$

The above identity is a special case of the law of total probability:

Fact. Let $F_1, \ldots, F_n$ form a partition of the sample space S, that is, $F_1 \cup \cdots \cup F_n = S$ and $F_1 \cap \cdots \cap F_n = \varnothing$. Let $E \subset S$ be an event. Then we have

$$\mathcal{P}(E) = \sum_{i=1}^{n} \mathcal{P}(E \cap F_i) = \sum_{i=1}^{n} \mathcal{P}(E \,|\, F_i) \cdot \mathcal{P}(F_i).$$

3.4.3 Bayes' theorem

Let E and F be events. Let X and Y be random variables. Since

$$\mathcal{P}(E \cap F) = \mathcal{P}(E \,|\, F) \cdot \mathcal{P}(F) = \mathcal{P}(F \,|\, E) \cdot \mathcal{P}(E),$$

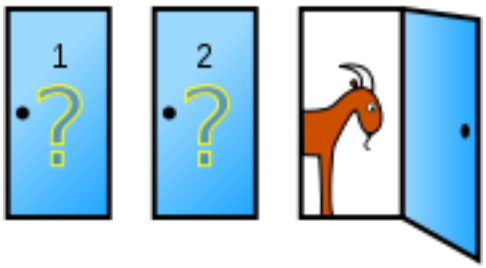
we have

$$\mathcal{P}(E \,|\, F) = \frac{\mathcal{P}(F \,|\, E) \cdot \mathcal{P}(E)}{\mathcal{P}(F)}.$$

This is known as Bayes' theorem, rule, law, or formula. When considering random variables X and Y, this may be written as

$$\mathbb{P}\{X \in E \,|\, Y \in F\} = \frac{\mathbb{P}\{Y \in F \,|\, X \in E\} \cdot \mathbb{P}\{X \in E\}}{\mathbb{P}\{Y \in F\}}.$$

Example. (Monty Hall problem) On a game show, you are given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 2, and the host, who knows what is behind each door, opens another door, say No. 3, which has a goat. The host then says to you, "Do you want to pick door No. 1 instead?" Is it to your advantage to switch your choice? What is the probability that the car is behind the door you picked at the beginning?



Is the answer 1/2? No. Without all the complication, the car is placed at random behind one of the three doors, so the probability that it is behind the door you picked has to be 1/3.

How to interpret this via conditional probability?

- Wrong: Conditional on that a goat is behind No. 3, the probability that the car is behind No. 1 is 1/2.

- Correct: Suppose that you picked No. 2. Let X be the random location of the car, and let Y be the random door that the host opened. Then

$$\mathbb{P}\{X = 2 \,|\, Y = 3\} = \frac{\mathbb{P}\{Y = 3 \,|\, X = 2\} \cdot \mathbb{P}\{X = 2\}}{\mathbb{P}\{Y = 3\}} = \frac{(1/2) \cdot (1/3)}{1/2} = \frac{1}{3}.$$

Example. A plane is missing and it is equally likely to have gone down in one of three possible regions. Let $1 - \alpha_i$ be the probability the plane can be found upon a search of the ith region when the plane is in fact in that region, $i = 1, 2, 3$. What is the conditional probability that the plane is in the ith region, $i = 1, 2, 3$, given that a search of region 1 is unsuccessful?

Let X be the region the plane is in. Let Y be the indicator that the search in region 1 is successful (i.e., $Y = 1$, successful; $Y = 0$, unsuccessful). Then

$$\mathbb{P}\{Y = 0\} = \sum_{i=1}^{3} \mathbb{P}\{Y = 0 \,|\, X = i\} \cdot \mathbb{P}\{X = i\} = \frac{\alpha_1}{3} + \frac{1}{3} + \frac{1}{3} = \frac{\alpha_1}{3} + \frac{2}{3},$$

so

$$\mathbb{P}\{X = 1 \,|\, Y = 0\} = \frac{\mathbb{P}\{Y = 0 \,|\, X = 1\} \cdot \mathbb{P}\{X = 1\}}{\mathbb{P}\{Y = 0\}} = \frac{\alpha_1/3}{\alpha_1/3 + 2/3} = \frac{\alpha_1}{\alpha_1 + 2},$$

$$\mathbb{P}\{X = 2 \,|\, Y = 0\} = \frac{\mathbb{P}\{Y = 0 \,|\, X = 2\} \cdot \mathbb{P}\{X = 2\}}{\mathbb{P}\{Y = 0\}} = \frac{1/3}{\alpha_1/3 + 2/3} = \frac{1}{\alpha_1 + 2},$$

and same for $i = 3$.

3.5 Conditional distribution

Recall that for discrete random variables X and Y, the conditional probability of $X = x$ given that $Y = y$ is

$$\mathbb{P}\{X = x \,|\, Y = y\} = \frac{\mathbb{P}\{X = x, Y = y\}}{\mathbb{P}\{Y = y\}}.$$

This motivates the following definition.

Definition. For random variables X and Y, the conditional PMF or PDF of X given that $Y = y$ is defined by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

The distribution of X conditional on $Y = y$ is given by this PMF or PDF, and the conditional CDF also follows.

Definition. For real-valued random variables X and Y, the conditional CDF of X given that $Y = y$ is

$$F_{X|Y}(x|y) = \mathbb{P}\{X \le x \,|\, Y = y\}.$$

Fact. If X and Y are continuous, then

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(x|y)\, dx.$$

Example. The joint PDF of x and y is

$$f(x, y) = \frac{12}{5} x(2 - x - y), \quad 0 < x < 1,\, 0 < y < 1,$$

and $f(x, y) = 0$ otherwise. The conditional PDF of X given that $Y = y$ where $0 < y < 1$ is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{\frac{12}{5} x(2 - x - y)}{\int_0^1 \frac{12}{5} x(2 - x - y)\, dx}.$$

The conditional probability that $0 < X < 0.5$ given that $Y = y$ where $0 < y < 1$ is

$$\mathbb{P}\{0 < X < 0.5 \,|\, Y = y\} = \int_0^{0.5} f_{X|Y}(x|y)\,dx = \int_0^{0.5} \frac{f(x,y)}{f_Y(y)}\,dx = \frac{\int_0^{0.5} \frac{12}{5}x(2-x-y)\,dx}{\int_0^1 \frac{12}{5}x(2-x-y)\,dx}.$$

3.6 Expectation of function of several random variables

Recall that for a continuous real-valued random variable X with PDF f and a function $g : \mathbb{R} \to \mathbb{R}$, the expectation $\mathbb{E}[g(X)]$ is given by

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx.$$

For a pair of random variables (X, Y) with joint PDF f and a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the expectation $\mathbb{E}[g(X, Y)]$ is analogously given by

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) \cdot f(x,y)\,dx\,dy$$

In the discrete case, if (X, Y) takes values in $S \times T$, then we have

$$\mathbb{E}[g(X,Y)] = \sum_{x\in S}\sum_{y\in T} \big(g(x,y) \cdot f(x,y)\big).$$

These formulas can be generalized to joint random variables with three or more components in the natural way.

3.6.1 Covariance

Recall that if a random variable X has mean $\mathbb{E}[X] = \mu$, then its variance is defined by $\mathbb{E}[(X-\mu)^2]$.

Definition. Suppose random variables X and Y have means μ_1 and μ_2 respectively. The covariance between X and Y is defined by

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mu_1)(Y-\mu_2)].$$

The correlation between X and Y is defined by

$$\mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}.$$

Remark. We have the following:

- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ and $\mathrm{Corr}(X,Y) = \mathrm{Corr}(Y,X)$;
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ and $\mathrm{Corr}(X,X) = 1$;
- The correlation defined above is a generalization of the correlation coefficient between two samples.

3.6.2 Useful identities for expectation, variance, and covariance

Let X, Y, $X_1, \ldots, X_n$, and $Y_1, \ldots, Y_n$ be real-valued random variables. The following formulas hold.

Expectation:

- $\mathbb{E}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \mathbb{E}[X_i]$;
- If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\cdot\mathbb{E}[Y]$;
- If X and Y are independent, and g and h are real-valued functions, then $g(X)$ and $h(Y)$ are independent, so $\mathbb{E}[g(X)\cdot h(Y)] = \mathbb{E}[g(X)]\cdot\mathbb{E}[h(Y)]$.
- If $X_1, \ldots, X_n$ are independent, then $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1]\cdots\mathbb{E}[X_n]$.

Covariance:

- $\mathrm{Cov}(\sum_{i=1}^n X_i, Y) = \sum_{i=1}^n \mathrm{Cov}(X_i, Y)$;
- $\mathrm{Cov}(\sum_{i=1}^n X_i, \sum_{j=1}^n Y_j) = \sum_{i,j=1}^n \mathrm{Cov}(X_i, Y_j)$;
- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Variance:

- If $X_1, \ldots, X_n$ are independent, then

$$\begin{aligned}\mathrm{Var}\Big(\sum_{i=1}^n X_i\Big) = \mathrm{Cov}\Big(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\Big) &= \sum_{i,j=1}^n \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^n \mathrm{Var}(X_i);\end{aligned}$$

- If $X_1, \ldots, X_n$ are independent and have the same variance, then $\mathrm{Var}(\sum_{i=1}^n X_i) = n\mathrm{Var}(X_1)$.

3.7 Sum of random variables

In this section, we discuss some results regarding a sum of independent random variables.

3.7.1 MGF

For any random variable X, let $M_X(t)$ denote its MGF. Let X, Y, and $X_1, \ldots, X_n$ be independent random variables. Then we have:

- $M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}] \cdot \mathbb{E}[e^{tY}] = M_X(t) \cdot M_Y(t)$;
- If $S := \sum_{i=1}^n X_i$, then $M_S(t) = \prod_{i=1}^n M_{X_i}(t)$.

3.7.2 Binomial

If $Y_1, \ldots, Y_n$ are independent $\mathrm{Ber}(p)$ random variables for $p \in [0, 1]$, then $X := \sum_{i=1}^n Y_i$ is a $\mathrm{Bin}(n, p)$ random variable. We have:

- $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[Y_i] = np$;
- $\mathrm{Var}(X) = \sum_{i=1}^n \mathrm{Var}(Y_i) = np(1-p)$.

For independent random variables $X_1 \sim \mathrm{Bin}(n_1, p)$ and $X_2 \sim \mathrm{Bin}(n_2, p)$, we have $X_1 + X_2 \sim \mathrm{Bin}(n_1 + n_2, p)$.

3.7.3 Poisson

For independent random variables $X_1 \sim \mathrm{Poi}(\lambda_1)$ and $X_2 \sim \mathrm{Poi}(\lambda_2)$, we have $X_1 + X_2 \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$. This is because the MGF of $X_1 + X_2$ is

$$\mathbb{E}[e^{t(X_1+X_2)}] = \mathbb{E}[e^{tX_1}] \cdot \mathbb{E}[e^{tX_2}] = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

which determines the distribution of $X_1 + X_2$.

The following fact is more advanced but interesting to know: Consider X items, where $X \sim \mathrm{Poi}(\lambda)$, and r boxes. We put each of the X items independently in the ith box with probability p_i, where $\sum_{i=1}^r p_i = 1$. Let X_i be the number of items in the ith box. Then $X_i \sim \mathrm{Poi}(\lambda p_i)$ for $i = 1, \ldots, r$, and $X_1, \ldots, X_r$ are independent.

3.7.4 Normal

For independent random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, we have $X+Y \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. More generally, if $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$, and they are independent, then

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big).$$

Example. Suppose that the yearly precipitation in an area is a normal random variable with mean 12.08 inches and standard deviation 3.1 inches. Suppose that the precipitation totals in the next two years are independent. What is the probability that the first year's precipitation exceeds the second year's by more than 3 inches?

Let the two years' precipitation totals be X and Y respectively. Then $X - Y \sim \mathcal{N}(0, 3.1^2 + 3.1^2)$, so

$$\begin{aligned}
\mathbb{P}\{X > Y + 3\} &= \mathbb{P}\{X - Y > 3\} \\
&= \mathbb{P}\Big\{\frac{X-Y}{\sqrt{3.1^2+3.1^2}} > \frac{3}{\sqrt{3.1^2+3.1^2}}\Big\} \\
&\approx \mathbb{P}\{Z > 0.6843\} \\
&= 1 - \Phi(0.6843) \approx 0.2469.
\end{aligned}$$

3.7.5 Chi-squared distribution

If $Z_1, \ldots, Z_n$ are independent standard normal random variables, then

$$X := Z_1^2 + \cdots + Z_n^2$$

is said to have a chi-squared distribution with n degrees of freedom, and we write $X \sim \chi_n^2$. Moreover,

- $\mathbb{E}[X] = \mathbb{E}[Z_1^2] + \cdots + \mathbb{E}[Z_n^2] = n$;
- $\mathrm{Var}(Z_1^2) = \mathbb{E}[(Z_1^2 - 1)^2] = 2$ and $\mathrm{Var}(X) = 2n$.

3.7.6 t-distribution

Consider independent random variables $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi_n^2$. Then the random variable

$$T := \frac{Z}{\sqrt{X/n}}$$

is said to have a t-distribution with n degrees of freedom.

Note that $X/n = \frac{1}{n}(Z_1^2 + \cdots + Z_n^2)$ is "close" to $\mathbb{E}[Z_i^2] = 1$ if n is large. Thus T is "close" to Z, a standard normal, if n is large. Moreover,

- $\mathbb{E}[T] = \mathbb{E}[Z] \cdot \mathbb{E}\big[\frac{1}{\sqrt{X/n}}\big] = 0$;
- $\mathrm{Var}(T) = \frac{n}{n-2}$.

The t-distribution is also known as Student's t-distribution, because the statistician William Sealy Gosset who studied the distribution used the pen name "Student".

4 Inequalities and limiting theorems

4.1 Markov's inequality

Proposition. (Markov's inequality) If X is a random variable that takes only nonnegative values, then for any $a > 0$,

$$\mathbb{P}\{X > a\} \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Consider the case where X is continuous with PDF f. Then

$$\mathbb{E}[X] = \int_0^\infty x f(x)\,dx \geq \int_a^\infty x f(x)\,dx \geq \int_a^\infty a f(x)\,dx = a \cdot \mathbb{P}\{X > a\}.$$

Proposition. (Chebyshev's inequality) If X is a random variable with mean μ and variance σ^2, then for any $a > 0$,

$$\mathbb{P}\{|X - \mu| > a\} \leq \frac{\sigma^2}{a^2}.$$

Proof. Apply Markov's inequality to the random variable $(X - \mu)^2$ to obtain

$$\mathbb{P}\{(X - \mu)^2 > a^2\} \leq \frac{\mathbb{E}[(X - \mu)^2]}{a^2}.$$

Example. Suppose that the number of items produced in a factory during a week is a random variable with mean 50. What can be said about the probability that this week's production will exceed 75? We have

$$\mathbb{P}\{X > 75\} \leq \frac{\mathbb{E}[X]}{75} = \frac{2}{3}.$$

If the variance of a week's production is known to be equal to 25, what can be said about the probability that this week's production will be between 40 and 60? We have

$$\mathbb{P}\{|X - 50| > 10\} \leq \frac{\sigma^2}{10^2} = \frac{1}{4}.$$

4.2 Weak law of large numbers

Theorem. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables. In other words, the random variables X_i have the same distribution and are mutually independent. Let μ and σ^2 denote the mean and the variance of X_i repectively. Then for any $\epsilon > 0$,

$$\mathbb{P}\Big\{\Big|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\Big| > \epsilon\Big\} \to 0 \quad \text{as } n \to \infty.$$

Proof. We have

$$\mathbb{E}\Big[\frac{1}{n}(X_1 + \cdots + X_n)\Big] = \frac{1}{n}\Big(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\Big) = \mu$$

and

$$\mathrm{Var}\Big(\frac{1}{n}(X_1 + \cdots + X_n)\Big) = \frac{1}{n^2}\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = \frac{\sigma^2}{n}.$$

By Chebyshev's inequality,

$$\mathbb{P}\Big\{\Big|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\Big| > \epsilon\Big\} \leq \frac{\sigma^2}{n\epsilon^2} \to 0.$$

Remark. We have $\frac{1}{n}(X_1 + \cdots + X_n) \to \mu$ "in probability" as $n \to \infty$. From a slightly different perspective, if $\sigma = 1$ and $\epsilon = \frac{10}{\sqrt{n}}$, then

$$\mathbb{P}\Big\{\Big|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\Big| \leq \frac{10}{\sqrt{n}}\Big\} > 1 - \frac{1}{100} = 0.99.$$

4.3 Central limit theorem

Theorem. Consider a sequence of i.i.d. random variables $X_1, X_2, \dots$ each with mean μ and variance σ^2. For n sufficiently large, the distribution of

$$\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu) = \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)$$

is approximately $\mathcal{N}(0,1)$, the standard normal distribution. In other words, for $Z \sim \mathcal{N}(0,1)$ and $x \in \mathbb{R}$,

$$\mathbb{P}\Big\{\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu) \le x\Big\} \approx \mathbb{P}\{Z \le x\}.$$

Remark. We have

$$\mathbb{E}\Big[\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu)\Big] = 0$$

and

$$\mathrm{Var}\Big(\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu)\Big) = 1.$$

Example. An insurance company has $n = 2500$ automobile policy holders. If the yearly claim of a policy holder is a random variable with mean $\mu = 320$ and standard deviation $\sigma = 540$, approximate the probability that the total yearly claim exceeds 830000.

If X_i is the claim of policy holder i, the total claim is $X = X_1 + \cdots + X_n$. Since $\frac{1}{540\cdot 50}(X - 2500 \cdot 320)$ is approximately $\mathcal{N}(0,1)$,

$$\begin{aligned}\mathbb{P}\{X > 830000\} &= \mathbb{P}\Big\{\frac{1}{540\cdot 50}(X - 2500\cdot 320) > \frac{1}{540\cdot 50}(830000 - 2500\cdot 320)\Big\}\\ &\approx \mathbb{P}\{Z > 10/9\} = 1 - \mathbb{P}\{Z \le 10/9\} = 1 - \Phi(10/9).\end{aligned}$$

Example. An astronomer wants to measure the distance between stars. Suppose that successive measurements are independent, and each is a random variable with mean μ being the true distance and standard deviation $\sigma = 2$ light years. How many measurements does she need to be at least 95% certain to estimate the distance within ± 0.5 light years?

Let $\bar{X}$ be the sample mean of n measurements. Since $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)$ is approximately $\mathcal{N}(0,1)$,

$$\begin{aligned}\mathbb{P}\{-0.5 < \bar{X} - \mu < 0.5\} &= \mathbb{P}\Big\{-\frac{\sqrt{n}}{4} < \frac{\sqrt{n}}{2}(\bar{X} - \mu) < \frac{\sqrt{n}}{4}\Big\}\\ &\approx \mathbb{P}\Big\{-\frac{\sqrt{n}}{4} < Z < \frac{\sqrt{n}}{4}\Big\}\\ &= 1 - \mathbb{P}\Big\{Z < -\frac{\sqrt{n}}{4}\Big\} - \mathbb{P}\Big\{Z > \frac{\sqrt{n}}{4}\Big\}\\ &= 1 - 2\cdot\mathbb{P}\Big\{Z > \frac{\sqrt{n}}{4}\Big\}\\ &= 1 - 2\big[1 - \Phi(\sqrt{n}/4)\big] = 2\,\Phi(\sqrt{n}/4) - 1.\end{aligned}$$

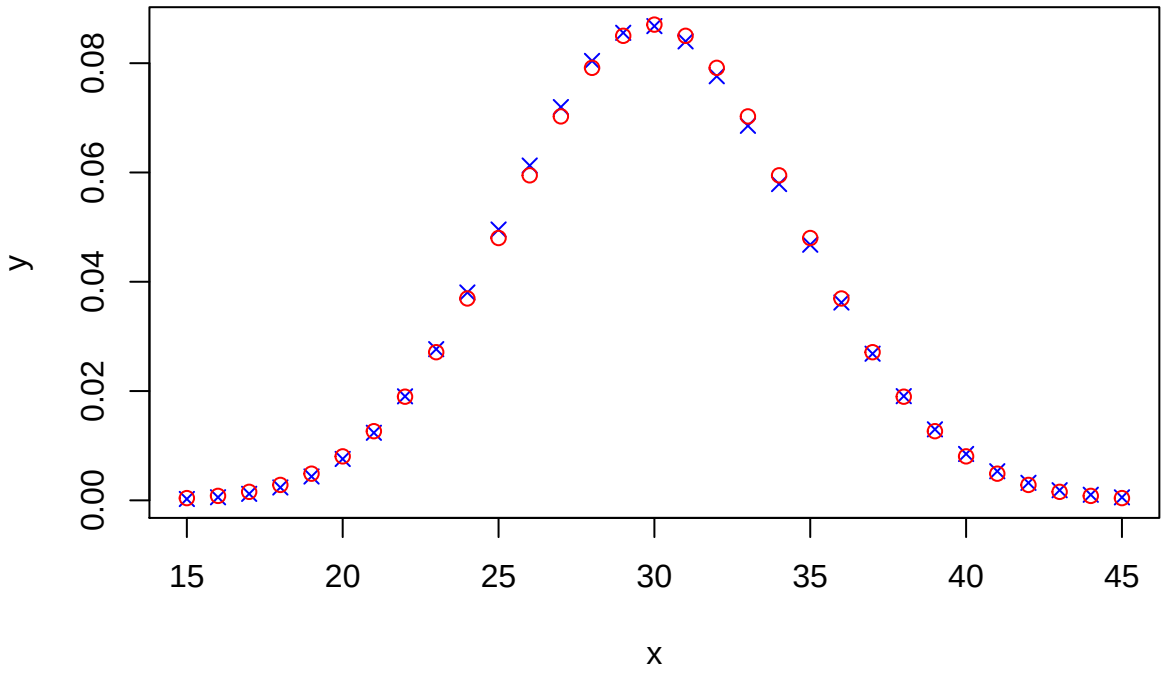
We need this quantity to be at least 0.95, that is, $\Phi(\sqrt{n}/4) \ge 0.975$. The approximation is $\sqrt{n}/4 \ge 1.96$, so the number of measurements n needs to be at least 62.

4.4 Normal approximation

By virtue of the central limit theorem, we can use the normal distribution to approximate other distributions. For example, let us consider the binomial distribution $\mathrm{Bin}(n,p)$. Recall that it can be approximated by $\mathrm{Poi}(np)$ if n is large and p is small. Instead, to obtain a normal approximation, note that $X \sim \mathrm{Bin}(n,p)$ if

$X = X_1 + \cdots + X_n$ where X_i are i.i.d. Ber(p) random variables. Thus $\frac{1}{\sigma\sqrt{n}}(X - np)$ can be approximated by $\mathcal{N}(0, 1)$, where $\sigma^2 = p(1-p)$ is the variance of Ber(p). In other words, Bin(n, p) can be approximated by $\mathcal{N}(np, np(1-p))$. Let us verify this approximation using R:

```
n = 100
p = 0.3
x = c(15:45)
y = dbinom(x, n, p)
z = dnorm(x, n*p, sqrt(n*p*(1-p)))
plot(x, y, type="p", col="blue", pch=4)
lines(x, z, type="p", col="red")
```



Example. Suppose that 45% of the population favors a certain candidate in an upcoming election. A random sample of size 200 is chosen.

(a) What is the expectation and standard deviation of the number of people X in the sample that favor the candidate?

(b) What is the probability that more than half the members of the sample favor the candidate?

Let X_i be the indicator that the ith member of the sample favors the candidate. *In such a scenario where the population is very large, we can view the random variables X_i as independent Bernoulli random variables.* Hence, approximately, $X = \sum_{i=1}^{200} X_i \sim \text{Bin}(200, 0.45)$.

(a) We have $\mathbb{E}[X] = 90$ and $\sqrt{\text{Var}(X)} = \sqrt{200 \cdot 0.45 \cdot (1 - 0.45)} = \sqrt{49.5}$.

(b) Using the *half-unit correction for continuity*, we have

$$\begin{aligned}\mathbb{P}\{X > 100\} &\approx \mathbb{P}\{X > 100.5\} \\ &= \mathbb{P}\Big\{\frac{X-90}{\sqrt{49.5}} > \frac{100.5-90}{\sqrt{49.5}}\Big\} \\ &\approx \mathbb{P}\{Z > 10.5/\sqrt{49.5}\} \\ &= 1 - \Phi(10.5/\sqrt{49.5}) \approx 0.068.\end{aligned}$$

The correction for continuity can be justified numerically using R:

```
1-pbinom(100,200,0.45)
```

```
## [1] 0.06807525
```

```
1-pnorm(10.5/sqrt(49.5))
```

```
## [1] 0.0677965
```

```
1-pnorm(10/sqrt(49.5))
```

```
## [1] 0.07760924
```

```
1-pnorm(11/sqrt(49.5))
```

```
## [1] 0.05897082
```

5 Statistical estimation

5.1 Procedure of statistical inference

The basic procedure of statistical inference is roughly as follows:

1. We are interested in some quantity in a population *(scores in a standardized test)*.
2. The quantity follows some unknown distribution $\mathcal{P}$. Possible assumption:
   - *Parametric:* Known family of distributions with unknown parameters *(normal distribution $\mathcal{N}(\mu, \sigma^2)$)*;
   - *Nonparametric:* Unknown distribution (not our focus in this course).
3. Take a sample $\{X_1, \ldots, X_n\}$ for i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}$.
4. Infer some property of the population using a statistic of the sample *(mean vs sample mean, variance vs sample variance)*.

Focusing on parametric models, we will discuss two fundamental tasks in statistics: *statistical estimation* and *hypothesis testing*.

Suppose that there is an underlying distribution $\mathcal{P}_\theta$ with unknown parameter θ (e.g., $\mathcal{P}_\theta = \text{Poi}(\lambda)$ where $\theta = \lambda$, and $\mathcal{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$). Given an i.i.d. sample $\{X_1, \ldots, X_n\}$ where $X_i \sim \mathcal{P}_\theta$ for $i = 1, \ldots, n$, we do:

- *Statistical estimation:* Estimate the parameter θ, i.e., find an estimator $\hat{\theta}$ that is close to θ. We may also estimate a function $g(\theta)$ where $g(\cdot)$ is known.
- *Hypothesis testing:* Test the *null hypothesis* H_0 against the *alternative hypothesis* H_1. For example, $H_0 : \theta = 5$ versus $H_1 : \theta \neq 5$.

In the language of statistical estimation, the parameter θ or function $g(\theta)$ to be estimated is called the *estimand*. A statistic $\hat{\theta}$ that is used to estimate θ is called an *estimator*. The value that the estimator takes is called the *estimate*.

For example, in estimating the mean θ in $\mathcal{N}(\theta, 1)$ from i.i.d. observations, θ is the estimand, $\hat{\theta} = \bar{X}$ is an estimator, and if, say $\bar{X} = 3$, then 3 is the estimate.

5.2 Basic sample statistics

5.2.1 Sample mean

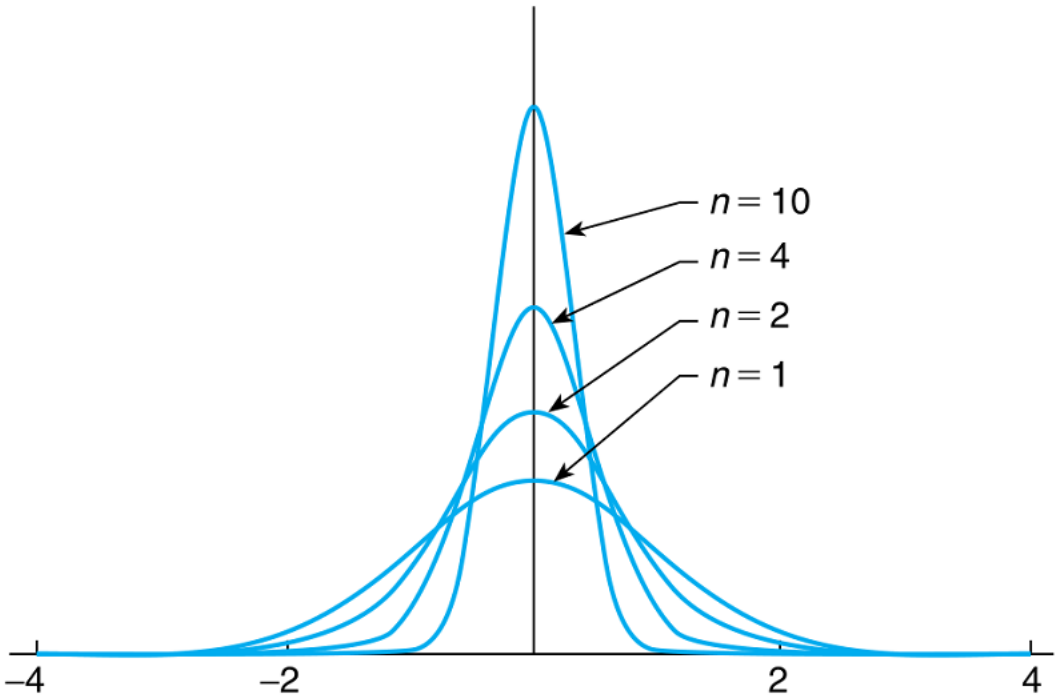
Consider i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}$, where $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$.

Definition. The expectation μ is the population mean. The average $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ is the sample mean.

We have

- $\mathbb{E}[\bar{X}] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \mu$;
- $\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2}[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)] = \sigma^2/n$.

Example. If $\mathcal{P} = \mathcal{N}(0, 1)$, then the PDF of $\bar{X}$ looks like (for varying n):



5.2.2 Sample variance

Definition. The sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Moreover, S is the sample standard deviation.

To compute the expectation of S^2, note that

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2.$$

Therefore,

$$(n-1)\mathbb{E}[S^2] = \sum_{i=1}^{n}\mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}^2] = \sum_{i=1}^{n}[\mathrm{Var}(X_i) + \mu^2] - n[\mathrm{Var}(\bar{X}) + \mu^2] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2,$$

so

$$\mathbb{E}[S^2] = \sigma^2.$$

Since the expectation of the sample variance is precisely the variance, we say that the sample variance is *unbiased*. This is the reason why we used the normalization $\frac{1}{n-1}$ instead of $\frac{1}{n}$ in the definition of S^2, because otherwise it would be *biased*.

5.2.3 Sample mean and sample variance of normal random variables

Consider i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. We have $X_1 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$ and $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. What is the distribution of S^2? How about the joint random variable $(\bar{X}, S^2)$?

Theorem. For i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, we have

- $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$;
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$;
- $\bar{X}$ and S^2 are independent.

Intuition: Recall that χ^2_n is the distribution of $\sum_{i=1}^n Z_i^2$ for independent standard normal random variables $Z_1, \dots, Z_n$. Consider the equation

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

The left-hand side is a χ^2_n random variable and the rightmost term is a χ^2_1 random variable. It is reasonable to expect that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 = (n-1)S^2/\sigma^2$$

is a χ^2_{n-1} random variable.

Corollary. For i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, the random variable

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

has t-distribution with $n-1$ degrees of freedom.

Proof. Recall that the t-distribution with $n-1$ degrees of freedom is the distribution of $\frac{Z}{\sqrt{X/(n-1)}}$ for independent random variables $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_{n-1}$. Writing the random variable in consideration as

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$$

finishes the proof.

Example. The time it takes a CPU to process a certain type of job is normally distributed with mean $\mu = 20$ seconds and standard deviation $\sigma = 3$ seconds. If a sample of $n = 15$ such jobs is observed, what is the probability that the sample variance exceeds 12?

We have

$$\mathbb{P}\{S^2 > 12\} = \mathbb{P}\Big\{\frac{14S^2}{9} > \frac{14}{9} \cdot 12\Big\} = 1 - F_{\chi^2_{14}}(56/3)$$

where $\frac{14S^2}{9} \sim \chi^2_{14}$.

```
1-pchisq(56/3,14)
```

```
## [1] 0.1780811
```

5.2.4 F-distribution

Definition. If $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ are independent, then

$$\frac{X/n}{Y/m}$$

is said to have the F-distribution with n and m degrees of freedom.

Example. If T has the t-distribution with n degrees of freedom, then T^2 has the F-distribution with 1 and n degrees of freedom.

Example. Consider two independent samples: The first sample has 10 independent normal random variables with variance 4; the second sample has 5 normal random variables having variance 2. Compute the probability that the sample variance of the second sample exceeds that of the first.

Let the sample variances of the two samples be S_1^2 and S_2^2 respectively. Then $9S_1^2/4 \sim \chi^2_9$ and $4S_2^2/2 \sim \chi^2_4$. Hence

$$\mathbb{P}\{S_2^2 > S_1^2\} = \mathbb{P}\Big\{\frac{(9S_1^2/4)/9}{(4S_2^2/2)/4} < 1/2\Big\} = F(1/2)$$

where F is the CDF of the F-distribution with 9 and 4 degrees of freedom.

5.3 Maximum likelihood estimation

5.3.1 First examples

Example. If we observe a single $X \sim \mathcal{N}(\theta, 1)$, how should we estimate θ? Intuitively, we should simply choose the estimator $\hat{\theta} = X$. Is there a principle behind this choice?

- Question: Which choice of the estimator $\hat{\theta}$ of the parameter is the "most likely" one? In other words, which normal distribution $\mathcal{N}(\hat{\theta}, 1)$ is the mostly likely to generate an observation X?
- Answer: $\hat{\theta} = X$.
- Reason: The PDF of $\mathcal{N}(\theta, 1)$ is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$. As a result, $\mathcal{N}(x, 1)$ is more likely to generate x than any other $\mathcal{N}(\theta, 1)$ for $\theta \neq x$, because

$$\frac{1}{\sqrt{2\pi}} e^{-(x-x)^2/2} = \frac{1}{\sqrt{2\pi}} > \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}.$$

  Therefore, observing X, we simply estimate the parameter by X itself.

Example. Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, why should we use the estimator $\hat{\theta} = \bar{X}$ to estimate the mean θ?

The joint distribution of $(X_1, \ldots, X_n)$ is specified by the PDF

$$\begin{aligned}
\tilde{f}(x_1, \ldots, x_n) &= f(x_1) \cdots f(x_n) \\
&= \frac{1}{\sqrt{2\pi}} e^{-(x_1-\theta)^2/2} \cdots \frac{1}{\sqrt{2\pi}} e^{-(x_n-\theta)^2/2} \\
&= \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n (x_i-\theta)^2/2}.
\end{aligned}$$

As above, we would like to choose θ so that this quantity is maximized, that is, $\sum_{i=1}^n (x_i - \theta)^2$ is minimized. The optimal choice is $\theta = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$ since the sample mean minimizes the squared error. (This can be seen by setting the derivative to zero.) Therefore, given $X_1, \ldots, X_n$, we choose the estimator $\hat{\theta} = \bar{X}$ for the parameter θ.

5.3.2 Theory

Let $\mathcal{P}_\theta$ be a distribution parametrized by θ. Denote its PDF by $f(x|\theta)$. (Here the PDF can be understood as a conditional PDF which is especially useful when studying Bayesian statistics. If this confuses you, simply understand x as the variable of the function and θ as the parameter.)

Given i.i.d. $X_1, \dots, X_n \sim \mathcal{P}_\theta$, the joint PDF is

$$\tilde{f}(x_1, \dots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta).$$

The *likelihood (function)* is equal to the joint PDF, but it is understood as a function of θ:

$$L(\theta|x_1, \dots, x_n) = \tilde{f}(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

It describes "how likely" the joint distribution parametrized by θ generates $(x_1, \dots, x_n)$. The *log-likelihood (function)* is the logarithm of the likelihood:

$$\log L(\theta|x_1, \dots, x_n) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

The *maximum likelihood estimator (MLE)* is defined to be $\theta = \hat{\theta}$ that maximizes $L(\theta|x_1, \dots, x_n)$ or $\log L(\theta|x_1, \dots, x_n)$.

Remark. The above definitions are valid for discrete random variables as well, once we use f to denote the PMF.

5.4 More examples of maximum likelihood estimation

5.4.1 Normal distribution

Consider i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ are both unknown. What are the MLEs of μ and σ?

The joint PDF is

$$\tilde{f}(x_1, \dots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(\frac{-(x_i-\mu)^2}{2\sigma^2}\Big) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big).$$

The log-likelihood is

$$\log L(\mu, \sigma|x_1, \dots, x_n) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

To maximize this over μ, we again need to minimize $\sum_{i=1}^{n}(x_i-\mu)^2$, which yields the optimal choice $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Hence the MLE of μ is

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We set $\mu = \hat{\mu}$. Then, to maximize the log-likelihood over σ, differentiate the log-likelihood with respect to σ and set the result to 0:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = 0$$

which yields $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$. Hence the MLE of σ is

$$\hat{\sigma} = \Big[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big]^{1/2}.$$

Note that the MLE is *not* equal to the sample standard deviation $S = \big[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\big]^{1/2}$.

5.4.2 Binomial distribution

Suppose n independent trials are performed, each with success probability p. Let X_i be the indicator that the ith trial succeeds. What is the MLE of p?

What is the PMF of each $X_i \sim \text{Ber}(p)$? Note that $f(1) = p$ and $f(0) = 1 - p$, so we can write

$$f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$$

where $x_i \in \{0, 1\}$. Therefore, the PMF of $(X_1, \ldots, X_n)$ is

$$\tilde{f}(x_1, \ldots, x_n|p) = \prod_{i=1}^{n} f(x_i|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

The log-likelihood is

$$\log L(p|x_1, \ldots, x_n) = \log \tilde{f}(x_1, \ldots, x_n|p) = \Big(\sum_{i=1}^{n} x_i\Big) \log p + \Big(n - \sum_{i=1}^{n} x_i\Big) \log(1-p).$$

To maximize the log-likelihood, differentiate the log-likelihood with respect to p and set the result to 0:

$$\frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\Big(n - \sum_{i=1}^{n} x_i\Big) = 0$$

which yields $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Hence the MLE of p is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

5.4.3 Poisson distribution

Consider i.i.d. $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$. What is the MLE of λ?

The Poisson PDF is $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, so the joint PDF is

$$\tilde{f}(x_1, \ldots, x_n|\lambda) = e^{-\lambda}\frac{\lambda^{x_1}}{x_1!} \cdots e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

The log-likelihood is

$$\log L(\lambda|x_1, \ldots, x_n) = -n\lambda + \Big(\sum_{i=1}^{n} x_i\Big) \log \lambda - \sum_{i=1}^{n} \log(x_i!).$$

Differentiate the log-likelihood with respect to λ and set the result to 0:

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

which yields $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Hence the MLE of λ is

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

5.4.4 Uniform distribution

Consider i.i.d. $X_1, \ldots, X_n \sim \text{Uniform}([0, \theta])$, where θ is the unknown parameter. The PDF of X_i is $f(x_i) = 1/\theta$ for $x_i \in [0, \theta]$, so the joint PDF is

$$\tilde{f}(x_1, \ldots, x_n|\theta) = 1/\theta^n$$

where $x_i \in [0, \theta]$ for each $i = 1, \ldots, n$, and $\tilde{f} = 0$ otherwise. To maximize the likelihood (which is equal to the joint PDF) over θ, we would like θ to be as small as possible. However, $x_i \le \theta$ for each $i = 1, \ldots, n$, so the smallest θ we can take is $\hat{\theta} = \max(x_1, \ldots, x_n)$. Hence the MLE of θ is

$$\hat{\theta} = \max(X_1, \ldots, X_n).$$

5.5 Interval estimation

There are two types of statistical estimation procedures, with the second being more refined than the first:

- *Point estimation:* Give an estimator $\hat{\theta}$ of the parameter θ;
- *Interval estimation:* Give an interval that contains θ with high probability.

Definition. Suppose that we have i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}_\theta$ for a real-valued parameter θ. Fix $\alpha \in (0,1)$. An interval $I \subset \mathbb{R}$ computed from the data is called a $(1-\alpha)$ confidence interval for the parameter θ if $\mathbb{P}\{\theta \in I\} \geq 1-\alpha$.

Remark. The probability makes sense because although θ may be fixed, the interval is random. Specifically, given the data $X = (X_1, \ldots, X_n)$, the confidence interval is of the form $(L(X), U(X))$ where $L(X)$ and $U(X)$ are random variables computed from the data.

For special distributions $\mathcal{P}_\theta$, we can often achieve the equality $\mathbb{P}\{\theta \in I\} = 1-\alpha$ as in the following examples.

5.5.1 Normal mean estimation with known variance

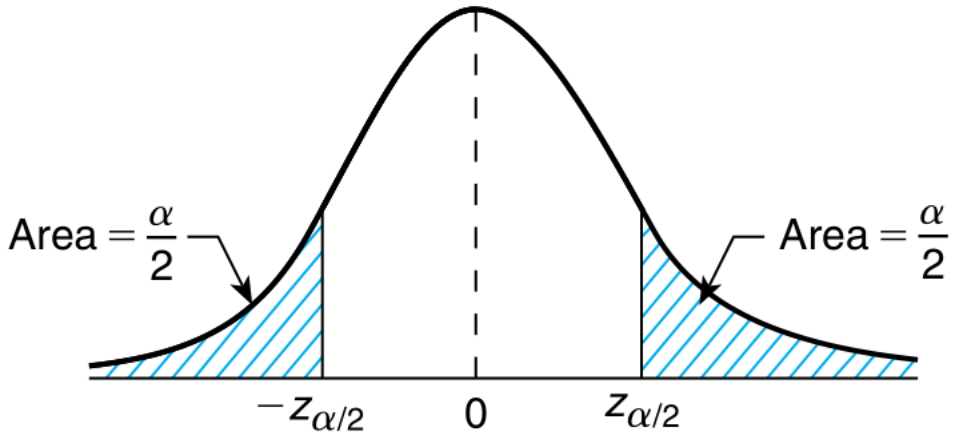
Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ is known, we can use $\bar{X}$ to estimate μ. Moreover, we have that $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim \mathcal{N}(0,1)$. Recall that $z_{\alpha/2} := \Phi^{-1}(1-\alpha/2)$ is the quantile of order $1-\alpha/2$ for $\mathcal{N}(0,1)$. Thus

$$\begin{aligned}
&\mathbb{P}\Big\{\frac{\sqrt{n}}{\sigma}|\bar{X}-\mu| \geq z_{\alpha/2}\Big\} = \alpha,\\
&\mathbb{P}\Big\{-z_{\alpha/2} < \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) < z_{\alpha/2}\Big\} = 1-\alpha,\\
&\mathbb{P}\Big\{-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big\} = 1-\alpha,\\
&\mathbb{P}\Big\{-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu-\bar{X} < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big\} = 1-\alpha,\\
&\mathbb{P}\Big\{\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big\} = 1-\alpha.
\end{aligned}$$

In other words, with probability $1-\alpha$, the population mean μ lies in the interval

$$\Big(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big),$$

which we call the $(1-\alpha)$ *two-sided confidence interval for* μ.



Similarly,

$$\Big(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\Big), \qquad \Big(-\infty, \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\Big)$$

are the $(1-\alpha)$ *one-sided confidence intervals for* μ, where the first gives a lower bound on μ and the second gives an upper bound on μ.

Example. A signal with value μ is sent from A to B, and the value received at B is $X \sim \mathcal{N}(\mu, 4)$. If the same signal is sent 9 times with independent noise, with sample mean equal to 9, what is the 95 percent two-sided confidence interval for μ? How about the 95 percent one-sided confidence interval that provides a lower bound on μ?

Recall that $z_{0.05/2} \approx 1.96$ and $z_{0.05} \approx 1.645$, so the confidence intervals are, respectively,

$$\Big(9 - 1.96 \cdot \frac{2}{3}, 9 + 1.96 \cdot \frac{2}{3}\Big), \qquad \Big(9 - 1.645 \cdot \frac{2}{3}, \infty\Big).$$

Example. Suppose that the weights of salmon grown at a hatchery are normal with standard deviation 0.3 pounds. If we want to be 95 percent certain that our estimate of the mean of a salmon's weight is correct to within ± 0.1 pounds, how large a sample is needed?

The 95 percent two-sided confidence interval is

$$\Big(\bar{X} - 1.96\frac{0.3}{\sqrt{n}}, \bar{X} + 1.96\frac{0.3}{\sqrt{n}}\Big).$$

We need $1.96\frac{0.3}{\sqrt{n}} \leq 0.1$ which gives $\sqrt{n} \geq 5.88$, so $n \geq 35$.

5.5.2 Normal mean estimation with unknown variance

Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where both μ and σ are unknown, let $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ be the sample variance. Recall that

$$T = \frac{\sqrt{n}}{S}(\bar{X} - \mu)$$

follows the t-distribution with $n-1$ degrees of freedom. Define $t_{\alpha,n-1}$ to be the quantile of order $1-\alpha$ for the t distribution with $n-1$ degrees of freedom, i.e.,

$$t_{\alpha,n-1} = F_T^{-1}(1-\alpha), \qquad \mathbb{P}\{T \leq t_{\alpha,n-1}\} = 1 - \alpha.$$

Similar to the previous case, we can derive that the $(1-\alpha)$ two-sided confidence interval for μ is

$$\Big(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\Big).$$

Moreover, the $(1-\alpha)$ one-sided confidence intervals are

$$\Big(\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\Big), \qquad \Big(-\infty, \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\Big).$$

Example. A signal with value μ is sent from A to B, and the value received at B is $X \sim \mathcal{N}(\mu, \sigma^2)$. If the same signal is sent 9 times with independent noise, with sample mean 9 and sample variance 9.5, what is the 95 two-sided percent confidence interval for μ?

It is

$$\Big(9 - t_{0.025,8}\frac{\sqrt{9.5}}{3}, 9 + t_{0.025,8}\frac{\sqrt{9.5}}{3}\Big)$$

where $t_{0.025,8} \approx 2.306$.

```
qt(0.975,8)
```

```
## [1] 2.306004
```

5.5.3 Normal variance estimation

Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown, we now consider estimating σ^2. Recall that

$$W = \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}.$$

Define $x_{\alpha,n-1}$ to be the quantile of order $1-\alpha$ for the χ^2_{n-1}, i.e.,

$$x_{\alpha,n-1} = F_W^{-1}(1-\alpha), \qquad \mathbb{P}\{W \le x_{\alpha,n-1}\} = 1-\alpha.$$

Then we have

$$\begin{aligned}
&\mathbb{P}\Big\{x_{1-\alpha/2,n-1} \le \frac{n-1}{\sigma^2} S^2 \le x_{\alpha/2,n-1}\Big\} = 1-\alpha,\\
&\mathbb{P}\Big\{\frac{x_{1-\alpha/2,n-1}}{(n-1)S^2} \le \frac{1}{\sigma^2} \le \frac{x_{\alpha/2,n-1}}{(n-1)S^2}\Big\} = 1-\alpha,\\
&\mathbb{P}\Big\{\frac{(n-1)S^2}{x_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{x_{1-\alpha/2,n-1}}\Big\} = 1-\alpha.
\end{aligned}$$

Hence the $(1-\alpha)$ two-sided confidence interval for σ^2 is

$$\Big(\frac{(n-1)S^2}{x_{\alpha/2,n-1}}, \frac{(n-1)S^2}{x_{1-\alpha/2,n-1}}\Big).$$

Similarly, the one-sided confidence intervals for σ^2 are

$$\Big(\frac{(n-1)S^2}{x_{\alpha,n-1}}, \infty\Big), \qquad \Big(0, \frac{(n-1)S^2}{x_{1-\alpha,n-1}}\Big).$$

Example. A procedure is expected to produce plates with very small deviation in their thicknesses. Suppose that 10 plates were chosen and measured. If the thicknesses of these plates have sample variance $S^2 = 1.366 \times 10^{-5}$, what is the 90 percent two-sided confidence interval for the standard deviation of the thickness of a plate?

We can compute $x_{0.05,9} \approx 16.919$ and $x_{0.95,9} \approx 3.325$, so the confidence interval for the variance is

$$\Big(\frac{9 \times 1.366 \times 10^{-5}}{16.919}, \frac{9 \times 1.366 \times 10^{-5}}{3.325}\Big).$$

The confidence interval for the standard deviation is

$$\left(\sqrt{\frac{9 \times 1.366 \times 10^{-5}}{16.919}}, \sqrt{\frac{9 \times 1.366 \times 10^{-5}}{3.325}}\right).$$

```
qchisq(1-0.05,9)
```

```
## [1] 16.91898
```

```
qchisq(1-0.95,9)
```

```
## [1] 3.325113
```

5.6 More examples of interval estimation

5.6.1 Estimation of normal difference with known variances

Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and i.i.d. $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$ where σ_1 and σ_2 are known, consider estimation of $\mu_1 - \mu_2$. We have

$$\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m),$$
$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0, 1),$$

so

$$\mathbb{P}\Big\{ -z_{\alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} < z_{\alpha/2} \Big\} = 1 - \alpha,$$
$$\mathbb{P}\Big\{ \bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \Big\} = 1 - \alpha.$$

Therefore, the $(1 - \alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\Big(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Big).$$

The $(1 - \alpha)$ one-sided confidence intervals for $\mu_1 - \mu_2$ are

$$\Big(-\infty, \bar{X} - \bar{Y} + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\Big), \qquad \Big(\bar{X} - \bar{Y} - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \infty\Big).$$

Examples. Two types of electrical cable insulation have recently been tested to determine the voltage level at which failures tend to occur. Given two samples of sizes 14 and 12 for the two types respectively, the failure voltages have sample means 50 and 65 respectively. Suppose the two samples are normally distributed with variances 40 and 100 respectively.

- Determine the 95 percent confidence interval for $\mu_1 - \mu_2$.
- Determine the value that we can assert, with 95 percent confidence, exceeds $\mu_1 - \mu_2$.

Plug $\bar{X} = 50$, $\bar{Y} = 65$, $\sigma_1^2 = 40$, $\sigma_2^2 = 100$, $n = 14$, $m = 12$, and $z_{0.025}$ into the above formula for the two-sided confidence interval.

For the second question, use the one-sided confidence interval with $z_{0.05}$ that provides the upper bound.

5.6.2 Estimation of normal difference with unknown but equal variances

Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ and i.i.d. $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$ where σ is unknown, consider estimation of $\mu_1 - \mu_2$. If σ were known, we would start from

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \sim \mathcal{N}(0, 1)$$

to derive confidence intervals as above. However, σ is unknown, so we instead consider sample variances

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

Since

$$\frac{n-1}{\sigma^2}S_1^2 \sim \chi_{n-1}^2, \qquad \frac{m-1}{\sigma^2}S_2^2 \sim \chi_{m-1}^2$$

we have

$$\frac{n-1}{\sigma^2}S_1^2 + \frac{m-1}{\sigma^2}S_2^2 \sim \chi^2_{n+m-2}.$$

It follows that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n + 1/m}} \Bigg/ \sqrt{\Big(\frac{n-1}{\sigma^2}S_1^2 + \frac{m-1}{\sigma^2}S_2^2\Big)\Big/(n+m-2)}$$

has the t-distribution with $n+m-2$ degrees of freedom. This quantity can be written as

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n + 1/m}}$$

where

$$S_p^2 := \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}.$$

As a result,

$$\mathbb{P}\Big\{ -t_{\alpha/2,n+m-2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n + 1/m}} < t_{\alpha/2,n+m-2}\Big\} = 1 - \alpha.$$

Then the $(1-\alpha)$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\Big(\bar{X} - \bar{Y} - t_{\alpha/2,n+m-2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + t_{\alpha/2,n+m-2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\Big).$$

The $(1-\alpha)$ one-sided confidence intervals for $\mu_1 - \mu_2$ are

$$\Big(-\infty, \bar{X} - \bar{Y} + t_{\alpha,n+m-2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\Big), \qquad \Big(\bar{X} - \bar{Y} - t_{\alpha,n+m-2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \infty\Big).$$

Examples. There are two different techniques employed to produce batteries. A sample of 12 batteries was produced by the first method, and a sample of 14 batteries was produced by the second method. Suppose that the sample means are 140 and 135 respectively, and the sample variances are 100 and 75 respectively. Determine the 90 percent two-sided confidence interval for the difference between the means, assuming a common variance.

Apply the above formula with $\bar{X} = 140$, $\bar{Y} = 135$, $S_1^2 = 100$, $S_2^2 = 75$, $n = 12$, $m = 14$, and $t_{0.05,24}$.

5.6.3 Binomial estimation

Given $X \sim \text{Bin}(n, p)$, we are interested in estimating p. Recall that

$$\frac{X - np}{\sqrt{np(1-p)}}$$

can be approximated by $\mathcal{N}(0,1)$. Then we have

$$\mathbb{P}\Big\{ -z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\Big\} \approx 1 - \alpha,$$

so we can define the $(1-\alpha)$ *confidence region* for p as

$$\Big\{p : -z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}\Big\}.$$

However, this is not necessarily an interval. To obtain a confidence interval, let $\hat{p} := X/n$ (which is the MLE of p) and use $\sqrt{n\hat{p}(1-\hat{p})}$ to approximate $\sqrt{np(1-p)}$. We obtain

$$\begin{aligned}
&\mathbb{P}\Big\{ -z_{\alpha/2} < \frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\alpha/2} \Big\} \approx 1-\alpha, \\
&\mathbb{P}\Big\{ -z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < X-np < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} \Big\} \approx 1-\alpha, \\
&\mathbb{P}\Big\{ -z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np-X < z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} \Big\} \approx 1-\alpha, \\
&\mathbb{P}\Big\{ X - z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} < np < X + z_{\alpha/2}\sqrt{n\hat{p}(1-\hat{p})} \Big\} \approx 1-\alpha, \\
&\mathbb{P}\Big\{ \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \Big\} \approx 1-\alpha.
\end{aligned}$$

Therefore, the approximate $(1-\alpha)$ two-sided confidence interval for p is

$$\Big(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\Big).$$

The one-sided versions can be derived similarly.

Examples. Out of a sample of 100 transistors, 80 meet a certain standard. Approximate the 95 percent confidence interval for p.

We have

$$\Big(0.8 - 1.96\sqrt{0.8\cdot 0.2/100}, 0.8 + 1.96\sqrt{0.8\cdot 0.2/100}\Big).$$

Examples. A recent poll indicated that 52% of the population was in favor of the job of the government with a margin of error of ±4%. Suppose the poll used 95% confidence interval. How many people were questioned?

The 95% confidence interval is approximately

$$\Big(\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n}\Big).$$

Plugging in $\hat{p} = 52\%$, we obtain

$$\Big(0.52 - 1.96\sqrt{0.52\cdot 0.48/n}, 0.52 + 1.96\sqrt{0.52\cdot 0.48/n}\Big).$$

Moreover, we have

$$1.96\sqrt{0.52\cdot 0.48/n} \approx 0.04,$$

so

$$n \approx 599.$$

5.7 Bayesian estimation

In this section, to ease the notation, we use $f(\cdot)$ to denote the PDF or PMF of the random variable(s) in the brackets, and we do not distinguish random variables from the specific values they take.

5.7.1 The Bayesian perspective

According to Wikipedia, *Bayesian statistics* is a theory in the field of statistics based on the *Bayesian interpretation of probability* where *probability expresses a degree of belief in an event.* The degree of belief may be based on prior knowledge about the event, such as the results of previous experiments, or on personal beliefs about the event.

Given i.i.d. $X_1, \ldots, X_n \sim \mathcal{P}_\theta$, we aim to estimate the real-valued parameter θ. Suppose that we have a prior belief that θ is itself a random variable following the distribution with PDF $f(\theta)$. This distribution of the parameter θ is called the *prior (distribution).*

Upon observing $X_1, \ldots, X_n$, how should we update our belief? Let $f(X_1, \ldots, X_n \mid \theta)$ be the joint PDF of $(X_1, \ldots, X_n)$ when the true parameter is θ. Let $f(\theta \mid X_1, \ldots, X_n)$ be the PDF of the parameter θ conditional on that we have observed $X_1, \ldots, X_n$. By the definition of conditional distribution,

$$f(\theta \mid X_1, \ldots, X_n) = \frac{f(\theta, X_1, \ldots, X_n)}{f(X_1, \ldots, X_n)} = \frac{f(\theta, X_1, \ldots, X_n)}{\int f(\theta, X_1, \ldots, X_n)\, d\theta} = \frac{f(\theta) \cdot f(X_1, \ldots, X_n \mid \theta)}{\int f(\theta) \cdot f(X_1, \ldots, X_n \mid \theta)\, d\theta}.$$

This updated distribution of the parameter θ specified by the PDF $f(\theta \mid X_1, \ldots, X_n)$ is called the *posterior (distribution)*.

Then we may estimate the true parameter by the mean of the posterior or simply the *posterior mean*:

$$\mathbb{E}[\theta \mid X_1, \ldots, X_n] = \int \theta \cdot f(\theta \mid X_1, \ldots, X_n)\, d\theta.$$

Another possible estimator of θ is the maximum a posteriori (MAP) estimator, defined as the mode $\hat{\theta}$ of the posterior $f(\theta \mid X_1, \ldots, X_n)$, i.e., the parameter $\hat{\theta}$ that maximizes $f(\theta \mid X_1, \ldots, X_n)$. Note that this is different from the MLE, which is defined as the maximizer of $f(X_1, \ldots, X_n \mid \theta)$. We will not discuss the MAP estimator in detail.

Note that the *Bayesian perspective* is different from the *frequentist perspective* which we used to derive the MLE. A rough understanding is as follows:

- *Frequentist:* The parameter θ is fixed, while the data $X_1, \ldots, X_n$ and the estimator $\hat{\theta}$ are random variables (e.g., the MLE);
- *Bayesian:* The parameter θ is a random variable, while the data $X_1, \ldots, X_n$ and the estimator $\hat{\theta}$ are fixed (e.g., the posterior mean or the MAP).

Remark. Is the MLE related to the posterior? Yes. Consider the case where $\theta \in [a, b]$ and the prior is the uniform distribution on $[a, b]$. Then the posterior has PDF

$$f(\theta \mid X_1, \ldots, X_n) = \frac{f(X_1, \ldots, X_n \mid \theta)}{\int f(X_1, \ldots, X_n \mid \theta)\, d\theta}.$$

Since the denominator does not depend on θ, maximizing $f(X_1, \ldots, X_n \mid \theta)$ is the same as maximizing $f(\theta \mid X_1, \ldots, X_n)$. In short, the MLE is equal to the MAP estimator when the prior is uniform.

5.7.2 Posterior mean

We would like to show that the posterior mean is optimal in some sense. First recall that for any random variable θ and a constant c, we have

$$\begin{aligned}
\mathbb{E}[(\theta - c)^2] &= \mathbb{E}[(\theta - \mathbb{E}[\theta] + \mathbb{E}[\theta] - c)^2] \\
&= \mathbb{E}[(\theta - \mathbb{E}[\theta])^2 + 2(\theta - \mathbb{E}[\theta])(\mathbb{E}[\theta] - c) + (\mathbb{E}[\theta] - c)^2] \\
&= \mathbb{E}[(\theta - \mathbb{E}[\theta])^2] + (\mathbb{E}[\theta] - c)^2 \qquad \text{(bias-variance trade-off)} \\
&\geq \mathbb{E}[(\theta - \mathbb{E}[\theta])^2].
\end{aligned}$$

That is, to estimate the random variable θ by a fixed number, its mean $\mathbb{E}[\theta]$ does the best job in terms of the expected squared error or the *mean squared error* (MSE). Therefore, under the posterior distribution with PDF $f(\theta \mid X_1, \ldots, X_n)$, the posterior mean is optimal in the MSE.

Example. (Binomial estimation) Given i.i.d. $X_1, \ldots, X_n \sim \mathrm{Ber}(\theta)$, what is the posterior mean of θ if the prior is the uniform distribution on $[0, 1]$? (Recall that the MLE is simply $\bar{X}$.)

The joint PMF conditional on θ is

$$f(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}.$$

Therefore, the posterior PDF is

$$\begin{aligned} f(\theta \mid X_1, \ldots, X_n) &= \frac{f(\theta) \cdot f(X_1, \ldots, X_n \mid \theta)}{\int f(\theta) \cdot f(X_1, \ldots, X_n \mid \theta)\, d\theta} \\ &= \frac{f(X_1, \ldots, X_n \mid \theta)}{\int_0^1 f(X_1, \ldots, X_n \mid \theta)\, d\theta} \\ &= \frac{\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}}{\int_0^1 \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}\, d\theta}. \end{aligned}$$

Since

$$\int_0^1 \theta^m (1-\theta)^l \, d\theta = \frac{m! \cdot l!}{(m+l+1)!},$$

we obtain

$$f(\theta \mid X_1, \ldots, X_n) = \frac{(n+1)! \cdot \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}}{(\sum_{i=1}^n X_i)! \cdot (n - \sum_{i=1}^n X_i)!} = \frac{(n+1)! \cdot \theta^x (1-\theta)^{n-x}}{x! \cdot (n-x)!},$$

where $x := \sum_{i=1}^n X_i$. Thus the posterior mean is

$$\mathbb{E}[\theta \mid X_1, \ldots, X_n] = \int_0^1 \frac{(n+1)! \cdot \theta^{x+1}(1-\theta)^{n-x}}{x! \cdot (n-x)!}\, d\theta = \frac{(n+1)! \cdot (x+1)! \cdot (n-x)!}{x! \cdot (n-x)! \cdot (n+2)!} = \frac{x+1}{n+2}.$$

We conclude that the posterior mean is

$$\hat{\theta} := \frac{1}{n+2}\Big(1 + \sum_{i=1}^n X_i\Big).$$

(Again, compare this to the MLE $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.)

Example. (Normal estimation) Consider i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \tau^2)$ where the τ is known. Suppose that we have the prior $\theta \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ are known. What is the posterior mean of θ?

The joint PDF conditional on θ is

$$f(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\tau} \exp\Big(\frac{-(X_i-\theta)^2}{2\tau^2}\Big) = \frac{1}{(2\pi)^{n/2}\tau^n} \exp\Big(-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\tau^2}\Big),$$

and the prior is

$$f(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(\frac{-(\theta-\mu)^2}{2\sigma^2}\Big).$$

Hence we can compute the posterior PDF using the formula

$$f(\theta \mid X_1, \ldots, X_n) = \frac{f(\theta) \cdot f(X_1, \ldots, X_n \mid \theta)}{\int f(\theta) \cdot f(X_1, \ldots, X_n \mid \theta)\, d\theta}.$$

The posterior is a normal distribution, with mean

$$\mathbb{E}[\theta \mid X_1, \ldots, X_n] = \int \theta \cdot f(\theta \mid X_1, \ldots, X_n)\, d\theta = \frac{\sigma^2}{n\sigma^2 + \tau^2} \sum_{i=1}^n X_i + \frac{\tau^2}{n\sigma^2 + \tau^2}\mu$$

and variance

$$\mathrm{Var}(\theta \mid X_1, \ldots, X_n) = \frac{\sigma^2\tau^2}{n\sigma^2 + \tau^2}.$$

Hence the posterior mean is

$$\hat{\theta} := \frac{n\sigma^2}{n\sigma^2 + \tau^2}\bar{X} + \frac{\tau^2}{n\sigma^2 + \tau^2}\mu.$$

Example. (Confidence interval) Suppose that a signal of value s is sent from A and the value received at B has distribution $\mathcal{N}(s, 60)$. Suppose that the signal sent from A is a priori known to have distribution $\mathcal{N}(50, 100)$. If the value received at B is equal to 40, determine the 90 percent confidence interval that will contain the actual value sent.

We have $n = 1$, $\mu = 50$, $\sigma^2 = 100$, $\bar{X} = 40$, and $\tau^2 = 60$. Thus the posterior mean is

$$\mathbb{E}[\theta \mid X_1, \dots X_n] = \frac{100}{100 + 60} \cdot 40 + \frac{60}{100 + 60} \cdot 50 = 43.75,$$

and the posterior variance is

$$\mathrm{Var}(\theta \mid X_1, \dots X_n) = \frac{100 \cdot 60}{100 + 60} = 37.5.$$

As a result,

$$\mathbb{P}\Big\{ -1.645 < \frac{s - 43.75}{\sqrt{37.5}} < 1.645 \mid X_1, \dots, X_n \Big\} \approx 0.9$$

or

$$\mathbb{P}\Big\{ 43.75 - 1.645\sqrt{37.5} < s < 43.75 + 1.645\sqrt{37.5} \mid X_1, \dots, X_n \Big\} \approx 0.9.$$

Hence the 90 percent confidence interval is

$$\Big(43.75 - 1.645\sqrt{37.5}, 43.75 + 1.645\sqrt{37.5} \Big).$$

5.7.3 Sequential estimation

An advantage of the Bayesian framework is that it allows us to do sequential estimation easily. Suppose that we are given i.i.d. random variables $X_1, X_2, X_3, \dots$ from a distribution $\mathcal{P}_\theta$ sequentially. At time n, we would like to compute the posterior of θ based on $X_1, \dots, X_n$. Recall that the posterior distribution at time n is

$$f(\theta \mid X_1, \dots, X_n) = \frac{f(\theta) \cdot f(X_1, \dots, X_n \mid \theta)}{\int f(\theta) \cdot f(X_1, \dots, X_n \mid \theta)\, d\theta} = \frac{f(\theta) \cdot \prod_{i=1}^n f(X_i \mid \theta)}{\int f(\theta) \cdot \prod_{i=1}^n f(X_i \mid \theta)\, d\theta}.$$

At time $n + 1$, the random variable X_{n+1} is given. We can now view $f(\theta \mid X_1, \dots, X_n)$ as the new prior and use the observation X_{n+1} to compute the new posterior at time $n + 1$ as follows:

$$f(\theta \mid X_1, \dots, X_n, X_{n+1}) = \frac{f(\theta \mid X_1, \dots, X_n) \cdot f(X_{n+1} \mid \theta)}{\int f(\theta \mid X_1, \dots, X_n) \cdot f(X_{n+1} \mid \theta)\, d\theta}.$$

Plugging the formula for $f(\theta \mid X_1, \dots, X_n)$ into the right-hand side, we obtain

$$f(\theta \mid X_1, \dots, X_n, X_{n+1}) = \frac{f(\theta) \cdot \prod_{i=1}^n f(X_i \mid \theta) \cdot f(X_{n+1} \mid \theta)}{\int f(\theta) \cdot \prod_{i=1}^n f(X_i \mid \theta) \cdot f(X_{n+1} \mid \theta)\, d\theta} = \frac{f(\theta) \cdot \prod_{i=1}^{n+1} f(X_i \mid \theta)}{\int f(\theta) \cdot \prod_{i=1}^{n+1} f(X_i \mid \theta)\, d\theta}.$$

Note that this is precisely the formula we would get for $f(\theta \mid X_1, \dots, X_{n+1})$ if we started from the prior $f(\theta)$ and computed the posterior using all the observations $X_1, \dots, X_{n+1}$ in a batch. In other words, we can compute the posterior sequentially given a stream of data, and this is much cheaper than computing the posterior using all the data at every time.

Example. (Normal estimation) Recall that given i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\theta, \tau^2)$ and prior $\theta \sim \mathcal{N}(\mu, \sigma^2)$, the posterior is

$$\mathcal{N}(\mu_n, \sigma_n), \quad \text{where } \mu_n := \frac{\sigma^2}{n\sigma^2 + \tau^2} \sum_{i=1}^n X_i + \frac{\tau^2}{n\sigma^2 + \tau^2}\mu \text{ and } \sigma_n^2 := \frac{\sigma^2\tau^2}{n\sigma^2 + \tau^2}.$$

Viewing this as the new prior, we compute a new posterior based on one more independent observation $X_{n+1} \sim \mathcal{N}(\theta, \tau^2)$ as follows:

$$\mathcal{N}\Big(\frac{\sigma_n^2}{\sigma_n^2 + \tau^2} X_{n+1} + \frac{\tau^2}{\sigma_n^2 + \tau^2}\mu_n, \frac{\sigma_n^2\tau^2}{\sigma_n^2 + \tau^2} \Big),$$

where

$$\begin{aligned}
&\frac{\sigma_n^2}{\sigma_n^2+\tau^2}X_{n+1}+\frac{\tau^2}{\sigma_n^2+\tau^2}\mu_n \\
&=\frac{\sigma^2\tau^2}{\sigma^2\tau^2+\tau^2(n\sigma^2+\tau^2)}X_{n+1}+\frac{\tau^2(n\sigma^2+\tau^2)}{\sigma^2\tau^2+\tau^2(n\sigma^2+\tau^2)}\Big(\frac{\sigma^2}{n\sigma^2+\tau^2}\sum_{i=1}^{n}X_i+\frac{\tau^2}{n\sigma^2+\tau^2}\mu\Big) \\
&=\frac{\sigma^2}{(n+1)\sigma^2+\tau^2}\sum_{i=1}^{n+1}X_i+\frac{\tau^2}{(n+1)\sigma^2+\tau^2}\mu=\mu_{n+1}
\end{aligned}$$

and

$$\frac{\sigma_n^2\tau^2}{\sigma_n^2+\tau^2}=\frac{\sigma^2\tau^2\tau^2}{\sigma^2\tau^2+\tau^2(n\sigma^2+\tau^2)}=\frac{\sigma^2\tau^2}{(n+1)\sigma^2+\tau^2}.$$

This distribution is indeed the posterior updated from the prior $\mathcal{N}(\mu,\sigma^2)$ using the i.i.d. observations $X_1,\dots,X_{n+1}\sim\mathcal{N}(\theta,\tau^2)$ in a batch.

6 Hypothesis testing

6.1 Basic setup of hypothesis testing

Consider i.i.d. $X_1,\dots,X_n\sim\mathcal{P}_\theta$ for an unknown parameter $\theta\in\mathbb{R}$. For example, we may have $\theta=\mu$ as the parameter and $\mathcal{P}_\mu=\mathcal{N}(\mu,\sigma^2)$ where σ is known. Given the observations, our goal is to test the *null hypothesis* H_0 against the *alternative hypothesis* H_1. For example, given i.i.d. $X_1,\dots,X_n\sim\mathcal{N}(\mu,\sigma^2)$ and a fixed constant μ_0, we test between

- $H_0:\mu=\mu_0$;
- $H_1:\mu\neq\mu_0$.

A hypothesis consisting of a single parameter that specifies the distribution is called a *simple hypothesis* (e.g., H_0 above). A hypothesis consisting of multiple parameters is called a *composite hypothesis* (e.g., H_1 above).

More precisely, testing is the task of deciding whether to *accept* or *reject* H_0, and a *test* is a decision rule that makes such a decision. A test is specified by a *critical region* or *region of rejection* $C\subset\mathbb{R}^n$: Namely, the test

- accepts H_0 if $(X_1,\dots,X_n)\notin C$;
- rejects H_0 if $(X_1,\dots,X_n)\in C$.

In the above example of normal observations, the critical region can be chosen to be

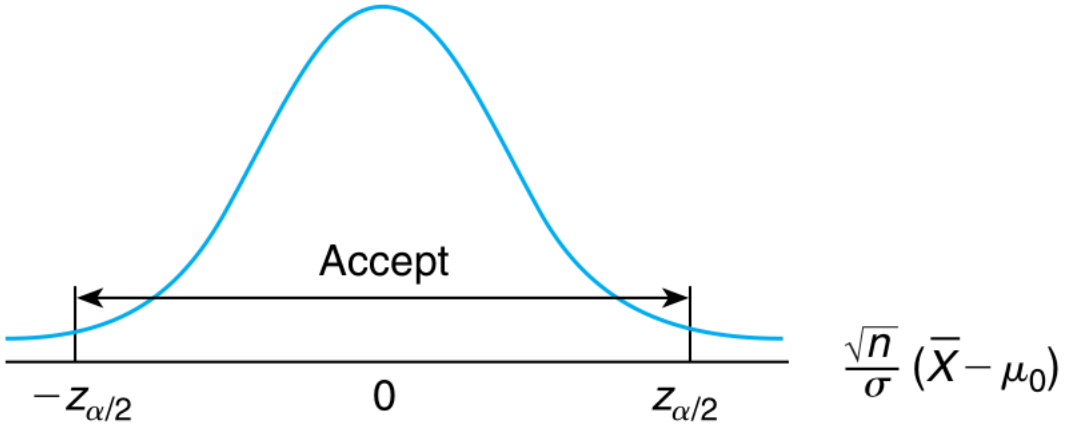
$$C=\Big\{(X_1,\dots,X_n):\Big|\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)\Big|>z_{\alpha/2}\Big\}.$$

In other words, the associated test

- accepts H_0 if $\big|\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)\big|\le z_{\alpha/2}$;
- rejects H_0 if $\big|\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)\big|>z_{\alpha/2}$.

Here $\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)$ is called the *test statistic*. Recall that a *statistic* is simply a quantity computed from the data, and a test statistic is a statistic that is used to define the test.

Two types of errors can arise in a test. *Type I error*: H_0 is true, but the test rejects it. *Type II error*: H_0 is false, but the test accepts it.



6.1.1 Significance level

We typically fix a quantity $\alpha \in (0,1)$ called the *level of significance* or *significance level* and require the test to be such that the probability of having a Type I error is no larger than α. In other words, if H_0 is true, then the probability that the test rejects H_0 is at most α. Usually α is a small constant such as 0.1, 0.05, or 0.005.

Continuing with the above example (i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and $H_0 : \mu = \mu_0$), what significance level does the test (reject H_0 if $\left|\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0)\right| > z_{\alpha/2}$) achieve? That is, if H_0 is true, what is the probability that $\left|\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0)\right| > z_{\alpha/2}$ so that we reject H_0?

If $H_0 : \mu = \mu_0$ is true, then $X_1, \ldots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$. As a result, $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \sim \mathcal{N}(0,1)$ and

$$\mathbb{P}_{\mu_0}\left\{\left|\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0)\right| > z_{\alpha/2}\right\} = \alpha,$$

where the notation $\mathbb{P}_{\mu_0}$ is used to emphasize that the probability is under the hypothesis $\mu = \mu_0$. Therefore, the test satisfies a significance level of α. We call this test the two-sided Z-test at significance level α. It is often employed for testing hypotheses about a normal mean when the variance is known.

Example. Suppose that a signal of value μ is sent from A, contaminated by random noise with distribution $\mathcal{N}(0,4)$, and then received at B. The signal is sent 5 times with independent noise and the average value received at B is $\bar{X} = 9.5$. If people at location B believe that $\mu = 8$ and would like to do a test at significance level 0.05, should they accept or reject the hypothesis?

Let the test statistic be the standardized random variable

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - 8}{2/\sqrt{5}}.$$

The test achieving the significance level 0.05 will

- accept H_0 if $\left|\frac{\bar{X}-8}{2/\sqrt{5}}\right| \le z_{0.05/2}$;
- reject H_0 if $\left|\frac{\bar{X}-8}{2/\sqrt{5}}\right| > z_{0.05/2}$.

Since $\bar{X} = 9.5$ which gives $\frac{\bar{X}-8}{2/\sqrt{5}} \approx 1.677 < 1.96 \approx z_{0.05/2}$, the hypothesis is accepted.

6.1.2 p-value

If the significance level is not specified in advance, what should we do for hypothesis testing? How do we quantify how likely a hypothesis is true? In the above example, given the value $\bar{X} = 9.5$, consider the quantity

$$\mathbb{P}\left\{|Z| > \left|\frac{9.5-8}{2/\sqrt{5}}\right|\right\} = 2\left[1 - \Phi\left(\left|\frac{9.5-8}{2/\sqrt{5}}\right|\right)\right]$$

which is the probability that $Z \sim \mathcal{N}(0,1)$ exceeds the standardized observation. We call this quantity the *p-value*. The smaller the p-value is, the more unlikely the null hypothesis is true.

For example, $\bar{X} = 9.5$ and $\bar{X} = 8.5$ give respective p-values

$$\mathbb{P}\Big\{|Z| > \Big|\frac{9.5-8}{2/\sqrt{5}}\Big|\Big\} \approx 0.094, \qquad \mathbb{P}\Big\{|Z| > \Big|\frac{8.5-8}{2/\sqrt{5}}\Big|\Big\} \approx 0.576.$$

When a significance level $\alpha = 0.1$ is required, we reject the null hypothesis $\mu_0 = 8$ if $\bar{X} = 9.5$ since the corresponding p-value is smaller than 0.1, and we accept the null hypothesis if $\bar{X} = 8.5$ since the corresponding p-value is larger than 0.1.

In general, given the value of the sample mean $\bar{X}$, the p-value is

$$\mathbb{P}\Big\{|Z| > \frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0|\Big\} = 2\Big[1-\Phi\Big(\frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0|\Big)\Big] = 2\,\Phi\Big(-\frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0|\Big),$$

where the probability is with respect to $Z \sim \mathcal{N}(0,1)$. We reject H_0 if and only if the p-value is smaller than the significance level α.

6.1.3 Power of a test

We focused on the probability of type I error above, i.e., the probability that the test rejects the null hypothesis H_0 when it is true. We consider the probability of type II error now, i.e., the probability of accepting the null hypothesis H_0 when the alternative hypothesis H_1 is true.

Consider i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$. Recall that the Z-test that rejects H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0| > z_{\alpha/2}$ achieves significance level α. Let $\mathbb{P}_\mu$ denote the probability to emphasize that the mean is μ. Define the probability of type II error as

$$\beta(\mu) = \mathbb{P}_\mu\{H_0 \text{ is accepted}\} = \mathbb{P}_\mu\Big\{\frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0| \le z_{\alpha/2}\Big\}.$$

The function $1-\beta(\mu)$ is called the *power function* or *power* of the test, i.e., the probability that the test rejects H_0 when H_1 is true. In this case, $\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \sim \mathcal{N}(0,1)$ (note that it is μ not μ_0), so

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}_\mu\Big\{-z_{\alpha/2} \le \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0) \le z_{\alpha/2}\Big\} \\
&= \mathbb{P}_\mu\Big\{-z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu) \le \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \le z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu)\Big\} \\
&= \Phi\Big(z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu)\Big) - \Phi\Big(-z_{\alpha/2} + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu)\Big).
\end{aligned}$$

Note that

- $\beta(\mu_0) = \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 2 \cdot \Phi(z_{\alpha/2}) - 1 = 1-\alpha$;
- $\lim_{\mu\to-\infty} \beta(\mu) = \Phi(\infty) - \Phi(\infty) = 0$;
- $\lim_{\mu\to\infty} \beta(\mu) = \Phi(-\infty) - \Phi(-\infty) = 0$.

Example. Suppose that a signal of value 10 is sent from A, contaminated by random noise with distribution $\mathcal{N}(0,4)$, and then received at B. The signal is sent 5 times with independent noise. If people at location B believe that the true value is 8, what is the probability of type II error of the test that rejects H_0 if $\frac{\sqrt{5}}{2}|\bar{X}-8| > 1.96$?

Since $\mu_0 = 8$ and $\mu = 10$, we have

$$\beta(10) = \Phi\Big(1.96 + \frac{\sqrt{5}}{2}(8-10)\Big) - \Phi\Big(-1.96 + \frac{\sqrt{5}}{2}(8-10)\Big) \approx 0.391.$$

```
pnorm(1.96+sqrt(5))-pnorm(-1.96+sqrt(5))
```

```
## [1] 0.3912343
```

6.2 One-sided tests

6.2.1 One-sided Z-test

In some cases, we consider the *one-sided* hypothesis testing problem with

- $H_0 : \mu \leq \mu_0$ (or $\mu = \mu_0$);
- $H_1 : \mu > \mu_0$.

Suppose that we are given i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. The one-sided Z-test

- accepts H_0 if $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \leq z_\alpha$;
- rejects H_0 if $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) > z_\alpha$.

The rationale behind the test is that it is at significance level α:

$$\mathbb{P}_{\mu_0}\Big\{\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) > z_\alpha\Big\} = \alpha,$$

as $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \sim \mathcal{N}(0, 1)$. Moreover, given the value of the sample mean $\bar{X}$, the p-value is

$$\mathbb{P}\Big\{Z > \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0)\Big\} = 1 - \Phi\Big(\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0)\Big),$$

where $Z \sim \mathcal{N}(0, 1)$ and $\bar{X}$ is a fixed value here.

Example. Suppose that a signal of value μ is sent from A, contaminated by random noise with distribution $\mathcal{N}(0, 4)$, and then received at B. The signal is sent 5 times with independent noise and the average value received at B is $\bar{X} = 9.5$. If people at location B believe that $\mu \leq 8$ and would like to do a one-sided test at significance level 0.05, should they accept or reject the hypothesis?

Similar to the two-sided case, there are two ways to approach this problem: via the test itself or via the p-value.

The test statistic is again the standardized random variable

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{9.5 - 8}{2/\sqrt{5}} \approx 1.677.$$

Since $z_{0.05} = \Phi^{-1}(1 - 0.05) \approx 1.645 < 1.677$, we have $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > z_\alpha$. As a result, the test rejects the null hypothesis H_0.

In addition, the p-value is

$$\mathbb{P}\Big\{Z > \frac{\sqrt{5}}{2}(9.5 - 8)\Big\} = 1 - \Phi\Big(\frac{\sqrt{5}}{2}(9.5 - 8)\Big) \approx 0.048 < 0.05,$$

so again the test rejects the null hypothesis H_0.

```
qnorm(1-0.05)
```

```
## [1] 1.644854
```

```
1-pnorm(sqrt(5)/2*1.5)
```

```
## [1] 0.04676626
```

6.2.2 Power of a one-sided test

As before, $\beta(\mu)$ is defined to be the probability that H_0 is accepted if the true mean is $\mu > \mu_0$ (i.e., H_1 is true):

$$\begin{aligned}\beta(\mu) &= \mathbb{P}_\mu\{H_0 \text{ is accepted}\} \\ &= \mathbb{P}_\mu\Big\{\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0) \le z_\alpha\Big\} \\ &= \mathbb{P}_\mu\Big\{\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \le z_\alpha + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu)\Big\} \\ &= \Phi\Big(z_\alpha + \frac{\sqrt{n}}{\sigma}(\mu_0-\mu)\Big).\end{aligned}$$

Note that

$$\beta(\mu_0) = \Phi(z_\alpha) = 1-\alpha, \qquad \lim_{\mu\to\infty}\beta(\mu) = \Phi(-\infty) = 0.$$

Again, the quantity $1-\beta(\mu)$ is called the power of the test.

6.2.3 The other direction

We can also consider hypothesis testing between

- $H_0 : \mu \ge \mu_0$ (or $\mu = \mu_0$);
- $H_1 : \mu < \mu_0$.

The one-sided Z-test at significance level α

- accepts H_0 if $\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0) \ge -z_\alpha$;
- rejects H_0 if $\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0) < -z_\alpha$.

The p-value is

$$\mathbb{P}\Big\{Z < \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)\Big\} = \Phi\Big(\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0)\Big).$$

Example. Suppose that cigarettes on the market have an average nicotine content of at least 1.6 mg per cigarette with standard deviation 0.8 mg. A firm claims that it has a new way to produce cigarettes with average nicotine content being less than 1.6 mg. To test this claim, a sample of 20 cigarettes from this firm were analyzed. What conclusion can be drawn, at a significance level 0.05, if the average nicotine content of the 20 cigarettes is 1.54?

We are testing between $H_0 : \mu \ge 1.6$ and $H_1 : \mu < 1.6$. How do we know which is the null and which is the alternative? This can be decided based on the following reasoning:

- We know the standard deviation of the population (cigarettes on the market) as prior knowledge, so it is more reasonable to define the null hypothesis H_0 to be the prior knowledge (i.e., $\mu \ge 1.6$).
- Rejection of H_0 is a sound, convincing statement when the significance level is low: If H_0 were true, the data is very unlikely; given the data, H_0 is very unlikely. On the other hand, accepting H_0 does not mean much—it should be understood as not rejecting H_0. Since we would only endorse the company when there is convincing evidence, we should define H_1 to be what the company claims (i.e., $\mu < 1.6$).

To test $H_0 : \mu \ge 1.6$ against $H_1 : \mu < 1.6$, we compute the test statistic

$$\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu_0) = \frac{\sqrt{20}}{0.8}(1.54-1.6) \approx -0.3354,$$

so the p-value is

$$\mathbb{P}\{Z < -0.3354\} = \Phi(-0.3354) \approx 0.369 > 0.05.$$

As a consequence, we accept H_0, that is, the evidence is not strong enough to support the company's claim.

```
pnorm(sqrt(20)/0.8*(1.54-1.6))
```

```
## [1] 0.3686578
```

Remark. In the above discussion, we assumed i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and did tests using the statistic $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$. In fact, if we have i.i.d. $X_1, \dots, X_n$ from any distribution with mean μ and variance σ^2, by the central limit theorem, the standardized statistic $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)$ is approximately $\mathcal{N}(0, 1)$ provided that n is large. Therefore, the above discussion also applies to the more general setting approximately.

6.3 t-test

Given i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ is unknown, consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Recall that if σ were known, the test statistic would be

$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0).$$

Since σ is not known in this case, we replace it by the sample standard deviation S to obtain the test statistic

$$T := \frac{\sqrt{n}}{S}(\bar{X} - \mu_0),$$

where the sample variance is defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Under H_0, recall that T follows the t-distribution with $n-1$ degrees of freedom. Let $t_{\alpha,n-1}$ denote the quantile of order $1-\alpha$ for this distribution. Then we can define the two-sided t-test at significance level α to be the test that

- accepts H_0 if $|T| \leq t_{\alpha/2,n-1}$;
- rejects H_0 if $|T| > t_{\alpha/2,n-1}$.

Given the values of the sample mean $\bar{X}$ and sample variance S, the p-value for the t-test is defined to be

$$\mathbb{P}\Big\{|T_{n-1}| > \frac{\sqrt{n}}{S}|\bar{X} - \mu_0|\Big\} = 2\Big[1 - F_{T_{n-1}}\Big(\frac{\sqrt{n}}{S}|\bar{X} - \mu_0|\Big)\Big] = 2\,F_{T_{n-1}}\Big(-\frac{\sqrt{n}}{S}|\bar{X} - \mu_0|\Big),$$

where T_{n-1} denotes a t-random variable with $n-1$ degrees of freedom and $F_{T_{n-1}}$ denotes its CDF. Similar to the Z-test, the p-value for the t-test is the probability that T_{n-1} is more extreme than the observed data.

Example. A public health official claims that the average home water use is 350 gallons a day. To verify this claim, a study of 20 randomly selected homes was instigated with the result that the average daily water uses of these 20 homes were such that $\bar{X} = 353$ and $S = 22$, do the data contradict the official's claim at a significance level $\alpha = 0.1$?

We test $H_0 : \mu = 350$ against $H_1 : \mu \neq 350$. We can compute $t_{0.05,19} \approx 1.73$, so the t-test rejects H_0 if $\frac{\sqrt{20}}{22}|\bar{X} - 350| > 1.73$. In this case, $\frac{\sqrt{20}}{22}|\bar{X} - 350| \approx 0.61$, so we accept H_0. The p-value is

$$\mathbb{P}\big\{|T_{19}| > 0.61\big\} = 2\big[1 - F_{T_{19}}(0.61)\big] \approx 0.55 > 0.1,$$

so, again, we accept H_0.

```
qt(0.95,19)
```

```
## [1] 1.729133
```

```
2*(1-pt(3*sqrt(20)/22,19))
```

```
## [1] 0.5491946
```

Given the same observations, let us now test $H_0 : \mu \le \mu_0$ (or $\mu = \mu_0$) against $H_1 : \mu > \mu_0$. The one-sided t-test at significance level α rejects H_0 if $\frac{\sqrt{n}}{S}(\bar{X} - \mu_0) > t_{\alpha,n-1}$. The p-value is

$$\mathbb{P}\Big\{T_{n-1} > \frac{\sqrt{n}}{S}(\bar{X} - \mu_0)\Big\} = 1 - F_{T_{n-1}}\Big(\frac{\sqrt{n}}{S}(\bar{X} - \mu_0)\Big).$$

The test between $H_0 : \mu \ge \mu_0$ (or $\mu = \mu_0$) and $H_1 : \mu < \mu_0$ is similar.

Example. The manufacturer of a new fiberglass tire claims that its average life will be at least 40K miles. To verify this claim, a sample of 12 tires is tested, with sample mean 37K and sample standard deviation 3K. Test the manufacturer's claim at the significance level 0.01.

We test $H_0 : \mu \ge 40$ against $H_1 : \mu < 40$. We can compute $t_{0.01,11} \approx 2.72$, so the t-test rejects H_0 if $\frac{\sqrt{12}}{3}(\bar{X} - 40) < -2.72$. In this case, $\frac{\sqrt{12}}{3}(\bar{X} - 40) \approx -3.464$, so we reject H_0. The p-value is

$$\mathbb{P}\{T_{11} < -3.464\} = F_{T_{11}}(-3.464) \approx 0.0026 < 0.01,$$

so, again, we reject H_0.

```
qt(0.99,11)
```

```
## [1] 2.718079
```

```
pt(sqrt(12)/3*(37-40),11)
```

```
## [1] 0.002647366
```

6.4 More examples of Z-test and t-test

6.4.1 Difference between normal means with known variances

Suppose that we have two independent samples: i.i.d. $X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and i.i.d. $Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$, where σ_1 and σ_2 are known. Consider testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \ne \mu_2$. The idea is that, if $\bar{X} - \bar{Y}$ is too large, then we tend to reject H_0. To make this precise, note that

$$\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m),$$

so $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0,1)$. If H_0 is true, then $\mu_1 - \mu_2 = 0$, so the test statistic is

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0,1).$$

The two-sided Z-test at significance level α rejects H_0 if

$$\frac{|\bar{X} - \bar{Y}|}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} > z_{\alpha/2}.$$

The p-value is

$$\mathbb{P}\Big\{|Z| > \frac{|\bar{X} - \bar{Y}|}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}\Big\} = 2\,\Phi\Big(-\frac{|\bar{X} - \bar{Y}|}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}\Big).$$

Example. Two new methods for producing a tire have been proposed. A tire manufacturer produces a sample of 10 tires using the first method and a sample 8 using the second. The first set is road tested at location A, and the lifetime of tires has sample mean 65 in hundred kilometers. The second set is tested

at location B, and the lifetime has sample mean 55. Suppose that the tire lifetime at the two locations is normal with standard deviation 40 and 30 respectively. If the manufacturer is interested in testing the hypothesis that there is no appreciable difference in the mean lifetime of tires produced by either method, what conclusion should be drawn at the significance level 0.05?

The test statistic is

$$\frac{65-55}{\sqrt{40^2/10+30^2/8}} \approx 0.6.$$

We can check that the hypothesis is accepted because the p-value is roughly 0.54.

```
2*pnorm(-(65-55)/sqrt(40^2/10+30^2/8))
```

```
## [1] 0.5446592
```

For testing between $H_0 : \mu_1 \leq \mu_2$ (or $\mu_1 = \mu_2$) against $H_1 : \mu_1 > \mu_2$, the one-sided Z-test rejects H_0 if

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} > z_\alpha.$$

The p-value is

$$1-\Phi\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}}\right).$$

The other side is similar and can be derived as before.

6.4.2 Difference between normal means with same but unknown variance

Consider two independent samples: i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$ and i.i.d. $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$, where σ is unknown. The sample variances are

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2, \qquad S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i-\bar{Y})^2$$

As in interval estimation, the key is to replace σ with an estimate in the statistic

$$\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{1/n+1/m}}.$$

Since

$$\frac{n-1}{\sigma^2}S_1^2 \sim \chi^2_{n-1}, \qquad \frac{m-1}{\sigma^2}S_2^2 \sim \chi^2_{m-1},$$

we have

$$\frac{n-1}{\sigma^2}S_1^2 + \frac{m-1}{\sigma^2}S_2^2 \sim \chi^2_{n+m-2}.$$

Moreover, under $H_0 : \mu_1 = \mu_2$,

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma^2/n+\sigma^2/m}} \sim \mathcal{N}(0,1).$$

It follows that

$$\frac{\bar{X}-\bar{Y}}{\sqrt{\sigma^2/n+\sigma^2/m}} \Bigg/ \sqrt{\left(\frac{n-1}{\sigma^2}S_1^2+\frac{m-1}{\sigma^2}S_2^2\right)\Big/(n+m-2)}$$

has the t-distribution with $n+m-2$ degrees of freedom, which we use as the test statistic. This statistic can be written as

$$T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{1/n+1/m}},$$

where

$$S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}.$$

Recall that the quantity S_p^2 is called the pooled estimator of the common variance σ^2.

The two-sided and one-sided t-tests can be derived in the same way as before.

Example. To study whether a type of medicine is effective for curing cold, a random group of 10 volunteers was given tablets containing the medicine. Moreover, a control group consisting of 12 other volunteers took placebo tablets. The lengths of time the cold lasted for the first group have sample mean 6.5 and sample variance 0.6 days; the lengths of time for the second group have sample mean 7.1 and sample variance 0.8. Is the medicine effective? Assuming the population variance is the same for the two groups, what conclusion can we draw at significance level 0.05?

If the medicine is effective, the time for the first group should be reduced. Therefore, we test $\mu_1 = \mu_2$ against $\mu_1 < \mu_2$. The test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{1/10 + 1/12}} \approx \frac{-0.6}{0.295} \approx -2.03,$$

where

$$S_p^2 = \frac{11S_1^2 + 9S_2^2}{20} = \frac{11 \cdot 0.6 + 9 \cdot 0.8}{20} = 0.69.$$

Since $t_{0.05,20} = F_{T_{20}}^{-1}(1 - 0.05) \approx 1.72$, the null hypothesis H_0 is rejected. Alternatively, we can see this from the p-value

$$\mathbb{P}\{T_{20} < T\} \approx F_{T_{20}}(-2.03) \approx 0.028 < 0.05.$$

```
pt(-0.6/(13.8/20*sqrt(1/10+1/12)),20)
```

```
## [1] 0.02788752
```

```
qt(0.95,20)
```

```
## [1] 1.724718
```

6.4.3 Difference between normal means with unknown and unequal variances

Consider two independent samples: i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and i.i.d. $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$, where σ_1 and σ_2 are unknown. In this case, it is difficult to do a precise analysis. Let us assume that n and m are large and do the Z-test, justified by the central limit theorem and the law of large numbers.

Pretending that the sample variances S_1 and S_2 are sufficiently close to σ_1 and σ_2 respectively, we consider the test statistic

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n + S_2^2/m}}$$

which is approximately $\mathcal{N}(0, 1)$. The rest is the same as the case of known variances.

6.5 Test statistics with other distributions

6.5.1 Testing normal variances

Suppose that we observe i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown. Consider testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$. Recall that under H_0,

$$\frac{n-1}{\sigma_0^2} S^2 \sim \chi_{n-1}^2,$$

where S^2 is the sample variance. Then we can use $\frac{n-1}{\sigma_0^2} S^2$ as the test statistic. Let $F_{\chi_{n-1}^2}$ denote the CDF of χ_{n-1}^2, and let $x_{\alpha,n-1} = F_{\chi_{n-1}^2}^{-1}(1 - \alpha)$, i.e., the quantile of order $1 - \alpha$ for χ_{n-1}^2. The two-sided χ^2-test at significance level $\alpha \in (0, 1)$ does the following:

- accept H_0 if $x_{1-\alpha/2,n-1} \leq \frac{n-1}{\sigma_0^2} S^2 \leq x_{\alpha/2,n-1}$;

- reject H_0 if $\frac{n-1}{\sigma_0^2}S^2 < x_{1-\alpha/2,n-1}$ or $\frac{n-1}{\sigma_0^2}S^2 > x_{\alpha/2,n-1}$.

In this case, the p-value is not as clear, so we do not use it.

Next, consider one-sided testing between $H_0 : \sigma^2 \le \sigma_0^2$ (or $\sigma^2 = \sigma_0^2$) and $H_1 : \sigma^2 > \sigma_0^2$. The one-sided χ^2-test at significance level α rejects H_0 if

$$\frac{n-1}{\sigma_0^2}S^2 > x_{\alpha,n-1}.$$

The p-value is

$$\mathbb{P}\Big\{X > \frac{n-1}{\sigma_0^2}S^2\Big\} = 1 - F_{\chi^2_{n-1}}\Big(\frac{n-1}{\sigma_0^2}S^2\Big),$$

where $X \sim \chi^2_{n-1}$.

The other side is analogous.

Example. A machine that controls the amount of ribbon on a tape will be judged to be effective if the variance σ^2 of the amount of ribbon on a tape is less than $0.0225\,cm^2$. If a sample of 20 tapes yields a sample variance of $S^2 = 0.025\,cm^2$, can we conclude that the machine is ineffective at a significance level 0.05?

We test $H_0 : \sigma^2 \le 0.0225$ against $H_1 : \sigma^2 > 0.0225$. Since

$$\frac{19}{0.0225}S^2 \approx 21.11 \le 30.14 \approx x_{0.05,19},$$

we accept H_0. The p-value is approximately

$$\mathbb{P}\{X > 21.11\} = 1 - F_{\chi^2_{19}}(21.11) \approx 0.33 > 0.05,$$

which gives the same conclusion.

```
19/0.0225*0.025
```

```
## [1] 21.11111
```

```
qchisq(1-0.05,19)
```

```
## [1] 30.14353
```

```
1-pchisq(19/0.0225*0.025,19)
```

```
## [1] 0.330694
```

6.5.2 Testing for binomial distributions

Suppose that we observe $X \sim \text{Bin}(n,p)$ where p is unknown. Alternatively, we may observe i.i.d. $X_1,\dots,X_n \sim \text{Ber}(p)$ and set $X = \sum_{i=1}^n X_i$. Consider testing $p \le p_0$ against $p > p_0$. A test would reject H_0 if $X > k^*$ for an appropriate threshold k^*. What k^* achieves a significance level α? Under H_0, we have

$$\mathbb{P}\{X > k\} = 1 - F_{n,p_0}(k),$$

where F_{n,p_0} denotes the CDF of $\text{Bin}(n,p)$. To have this probability bounded by α, we should define

$$k^* := \min\big\{k : 1 - F_{n,p_0}(k) \le \alpha\big\}.$$

However, it is not clear how to solve for k^* given α. The easier approach is to compute the p-value

$$\mathbb{P}\{B > X\} = 1 - F_{n,p_0}(X),$$

where $B \sim \text{Bin}(n,p_0)$. If the p-value is smaller than α, then we reject H_0.

The other one-sided case can be derived analogously. The two-sided case is more involved and we omit it.

Example. A computer chip manufacturer claims that no more than 2 percent of the chips it sends out are defective. An electronics company has purchased a large quantity of such chips. To determine if the manufacturer's claim can be taken, the company tested 300 chips and 9 of them are defective. Should the manufacturer's claim be rejected at significance level 0.05?

We test $H_0 : p \leq 0.02$ against $H_1 : p > 0.02$. The p-value is

$$1 - F_{300,0.02}(9) = 0.082 > 0.05,$$

where $B \sim \text{Bin}(300, 0.02)$, so the claim H_0 is accepted.

```
1-pbinom(9,300,0.02)
```

```
## [1] 0.08183807
```

For the same task, we may consider normal approximation and use the Z-test. If $p = p_0$, the standardized random variable

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

is approximately $\mathcal{N}(0, 1)$ by the central limit theorem. Therefore, we can use this quantity as the test statistic in the Z-test. This approximation allows us to employ the Z-test in the two-sided and both one-sided cases as before.

Example. For the above example, the p-value for the Z-test is

$$\mathbb{P}\Big\{Z \geq \frac{9.5 - 300 \cdot 0.02}{\sqrt{300 \cdot 0.02 \cdot 0.98}}\Big\} = 1 - \Phi\Big(\frac{3.5}{\sqrt{6 \cdot 098}}\Big) \approx 0.0745 > 0.05,$$

so again, we accept H_0. Note that we have used the half-unit correction for continuity.

```
1-pnorm(3.5/sqrt(6*0.98))
```

```
## [1] 0.07445734
```

6.5.3 Testing for Poisson distributions

Given i.i.d. $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$, consider testing $H_0 : \lambda \leq \lambda_0$ against $H_1 : \lambda > \lambda_0$. Then we have $X := X_1 + \cdots + X_n \sim \text{Poi}(n\lambda)$. For simplicity, we directly compute the p-value

$$\mathbb{P}\{Y > X\} = 1 - F_{n\lambda_0}(X),$$

where $Y \sim \text{Poi}(n\lambda_0)$ and $F_{n\lambda_0}$ denotes the CDF of $\text{Poi}(n\lambda_0)$.

Example. Suppose that the number of defective computer chips produced daily by a company follows a Poisson distribution. The company claims that the average number of defective chips produced daily is no greater than 25. If a sample of 5 days consists of 28, 34, 32, 38, and 22 defective chips, test the claim at significance level 0.05.

We test $H_0 : \lambda \leq 25$ against $H_1 : \lambda > 25$. The p-value is

$$1 - F_{125}(28 + 34 + 32 + 38 + 22) \approx 0.00526 < 0.05,$$

so we reject H_0.

```
1-ppois(28+34+32+38+22,125)
```

```
## [1] 0.005260669
```

Similar to binomial testing, we may also use normal approximation and the Z-test for Poisson testing.

7 Linear regression

7.1 The model

How much does a factor, such as location, food quality, service, etc., affect a restaurant's revenue? Let the factors be $x_1, \ldots, x_r$ and let the revenue be Y. Suppose that Y is determined by $x_1, \ldots, x_r$ linearly up to noise, i.e.,

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \epsilon = \beta_0 + x^\top \beta + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_r$ are unknown real-valued coefficients that we aim to estimate, and ϵ is random noise. We can also write the model in the vector form with $x = (x_1, \ldots, x_r)^\top$ and $\beta = (\beta_1, \ldots, \beta_r)^\top$ are r-dimensional vectors. Typically, we assume that the noise ϵ is a random variable with mean zero. Thus we also write

$$\mathbb{E}[Y \mid x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r = \beta_0 + x^\top \beta.$$

Remark. The above model is called the *linear regression model*, which describes the regression of the *dependent variable* Y on the set of *independent variables* $x_1, \ldots, x_r$. The quantities $\beta_0, \ldots, \beta_r$ are called the *regression coefficients*. The quantity ϵ is the random *noise*. Alternatively,

- Y can be called the regressand, endogenous variable, response variable, measured variable, criterion variable, or predicted variable;
- $x_1, \ldots, x_r$ can be called the regressors, exogenous variables, explanatory variables, covariates, input variables, predictor variables, or design points (x is known as the design vector);
- β_0 is the intercept term, and $\beta_1, \ldots, \beta_r$ are called the effects or regression parameters (β is known as the parameter vector);
- ϵ is also known as the disturbance term or simply the noise.

Let us first consider the *simple linear regression model* ($r = 1$)

$$Y = \alpha + \beta x + \epsilon.$$

Suppose that we observe n data points (x_i, Y_i) for $i = 1, \ldots, n$ following the model

$$Y_i = \alpha + \beta x_i + \epsilon_i.$$

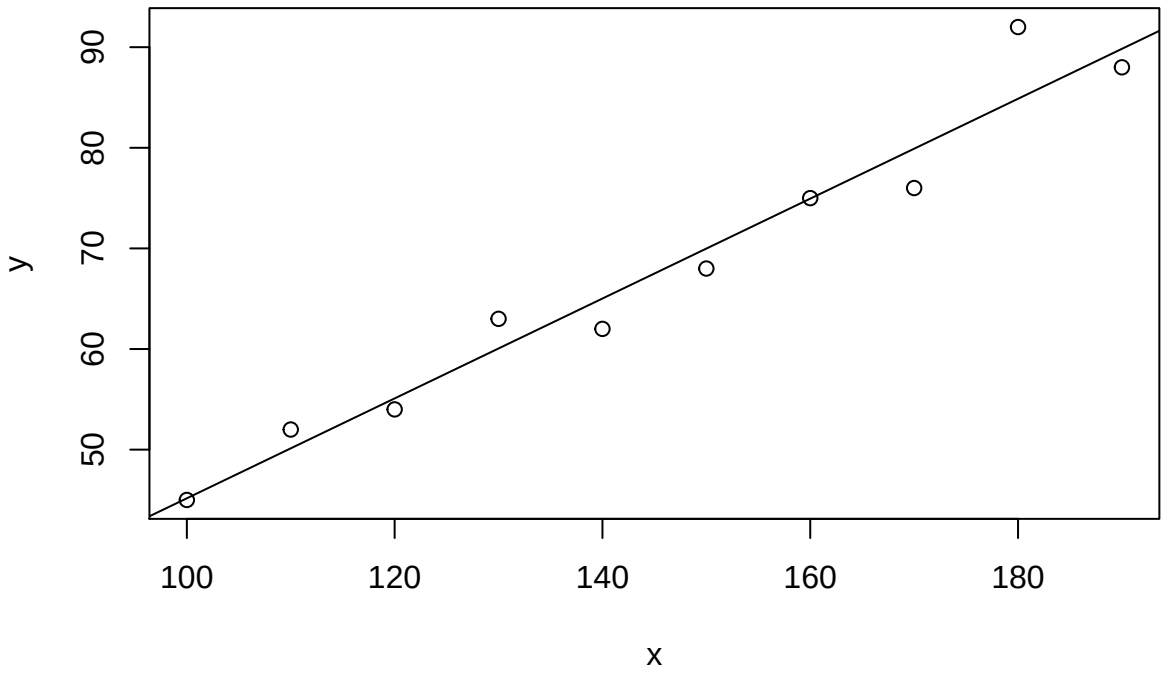
Example. Consider the observations:

$$\begin{aligned}
(x_1, Y_1) &= (100, 45), & (x_2, Y_2) &= (110, 52),\\
(x_3, Y_3) &= (120, 54), & (x_4, Y_4) &= (130, 63),\\
(x_5, Y_5) &= (140, 62), & (x_6, Y_6) &= (150, 68),\\
(x_7, Y_7) &= (160, 75), & (x_8, Y_8) &= (170, 76),\\
(x_9, Y_9) &= (180, 92), & (x_{10}, Y_{10}) &= (190, 88).
\end{aligned}$$

```
x <- seq(100,190,10)
y <- c(45,52,54,63,62,68,75,76,92,88)
fit <- lm(y ~ x)
print(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     -4.4727       0.4964
```

```
plot(x,y)
abline(fit)
```



7.2 Least squares estimators of regression coefficients

7.2.1 The estimators

Given (x_i, Y_i) for $i = 1, \ldots, n$ in a simple linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

how do we estimate the coefficients α and β? Let $\hat{\alpha}$ and $\hat{\beta}$ be estimators of α and β respectively (hypothetical at the moment). Then our prediction of Y_i would be

$$\hat{\alpha} + \hat{\beta} x_i.$$

The squared error is $(Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$, and the sum of squared errors is

$$S = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

We would like to choose $\hat{\alpha}$ and $\hat{\beta}$ so that the above sum of squares S is minimized. To this end, take the partial derivatives of S and set them to zero:

$$\frac{\partial S}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0, \qquad \frac{\partial S}{\partial \hat{\beta}} = -2 \sum_{i=1}^{n} x_i (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0.$$

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n$. Then the above two equations become

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{x}, \qquad \sum_{i=1}^n x_i Y_i = \hat{\alpha} n\bar{x} + \hat{\beta}\sum_{i=1}^n x_i^2.$$

Hence

$$\sum_{i=1}^n x_i Y_i = (\bar{Y} - \hat{\beta}\bar{x})n\bar{x} + \hat{\beta}\sum_{i=1}^n x_i^2,$$

so solving for $\hat{\beta}$ yields

$$\hat{\beta} = \frac{(\sum_{i=1}^n x_i Y_i) - n\bar{x}\bar{Y}}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2}$$

and then

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

The estimators $\hat{\alpha}$ and $\hat{\beta}$ are called the *least squares estimators*. The line $y = \hat{\alpha} + \hat{\beta}x$ is called the *estimated regression line*.

```
print(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     -4.4727       0.4964
```

```
beta = (sum(x*y)-10*mean(x)*mean(y))/(sum(x^2)-10*mean(x)^2)
alpha = mean(y)-beta*mean(x)
print(c(alpha,beta))
```

```
## [1] -4.4727273  0.4963636
```

To simplify the notation, we let

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2, \qquad S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2,$$
$$S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y},$$

and the least squares estimators are

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

Remark. If the noise terms ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ variables, then the least squares estimators are equivalent to the MLEs. This is because minimizing the sum of squares of errors is the same as maximizing the log-likelihood.

7.2.2 Distributions of the least squares estimators

We assume that the random noise terms ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ variables, so that

$$Y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2).$$

We now determine the distributions of the least squares estimators $\hat{\beta}$ and $\hat{\alpha}$ of the regression coefficients. First note that, as linear combinations of normal random variables, $\hat{\beta}$ and $\hat{\alpha}$ are normal random variables. It suffices to compute their means and variances. The mean of $\hat{\beta}$ is

$$\mathbb{E}[\hat{\beta}] = \frac{(\sum_{i=1}^n x_i \cdot \mathbb{E}[Y_i]) - n\bar{x} \cdot \mathbb{E}[\bar{Y}]}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2} = \frac{(\sum_{i=1}^n x_i(\alpha + \beta x_i)) - n\bar{x}(\alpha + \beta\bar{x})}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2} = \frac{(\sum_{i=1}^n \beta x_i^2) - n\beta\bar{x}^2}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2} = \beta,$$

so $\hat{\beta}$ is an unbiased estimator of β. Moreover,

$$\mathrm{Var}(\hat{\beta}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \mathrm{Var}(Y_i)}{\left[(\sum_{i=1}^n x_i^2) - n\bar{x}^2\right]^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{\left[(\sum_{i=1}^n x_i^2) - n\bar{x}^2\right]^2} = \frac{\sigma^2}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2}.$$

For $\hat{\alpha}$, we have

$$\mathbb{E}[\hat{\alpha}] = \mathbb{E}[\bar{Y}] - \mathbb{E}[\hat{\beta}]\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha,$$

so $\hat{\alpha}$ is an unbiased estimator of α. In addition, it can be shown that

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2) - n^2\bar{x}^2}.$$

In conclusion, we have

$$\hat{\beta} \sim \mathcal{N}\Big(\beta, \frac{\sigma^2}{S_{xx}}\Big), \qquad \hat{\alpha} \sim \mathcal{N}\Big(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}}\Big).$$

7.2.3 Residuals

The quantities $Y_i - \hat{\alpha} - \hat{\beta}x_i$, where $i = 1, \ldots, n$, are called the *residuals* (i.e., the differences between the actual responses and the predictors). The sum of squares of the residuals is

$$\mathsf{SS}_\mathsf{R} = \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{S_{xx}S_{YY} - (S_{xY})^2}{S_{xx}}.$$

It is known that

$$\frac{\mathsf{SS}_\mathsf{R}}{\sigma^2} \sim \chi^2_{n-2},$$

so

$$\mathbb{E}\Big[\frac{\mathsf{SS}_\mathsf{R}}{\sigma^2}\Big] = n - 2, \qquad \mathbb{E}\Big[\frac{\mathsf{SS}_\mathsf{R}}{n-2}\Big] = \sigma^2.$$

Consequently, $\frac{\mathsf{SS}_\mathsf{R}}{n-2}$ is an unbiased estimator of σ^2. Moreover, it is known that SS_R is independent of $\hat{\alpha}$ and $\hat{\beta}$.

7.2.4 Coefficient of determination

Consider the "variation" in Y_i, i.e., the quantity

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Two factors contribute to the variation in Y_i:

- variation in x_i, which *explains* variation in Y_i;
- variation from the noise, which is not really meaningful.

The sum of squares of residuals SS_R can be viewed as the variation coming from the noise. Therefore,

$$S_{YY} - \mathsf{SS}_\mathsf{R}$$

represents the amount of variation in Y_i that is explained by the variation in x_i. Moreover, the quantity

$$R^2 = \frac{S_{YY} - \mathsf{SS}_\mathsf{R}}{S_{YY}} = 1 - \frac{\mathsf{SS}_\mathsf{R}}{S_{YY}}$$

represents the *proportion* of the variation in Y_i that is explained by the variation in x_i. This quantity R^2 is called the *coefficient of determination*. We have $0 \le R^2 \le 1$. A value near 1 indicates a good fit, while a value near 0 indicates a poor fit.

7.2.5 Sample correlation coefficient

Recall that the sample correlation coefficient r of (x_i, Y_i) for $i = 1, \ldots, n$ is defined to be

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \cdot \left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]}}.$$

Using the identity

$$\mathsf{SS}_\mathsf{R} = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}},$$

we obtain

$$r^2 = \frac{S_{xY}^2}{S_{xx}S_{YY}} = \frac{S_{xx}S_{YY} - \mathsf{SS}_\mathsf{R}S_{xx}}{S_{xx}S_{YY}} = \frac{S_{YY} - \mathsf{SS}_\mathsf{R}}{S_{YY}} = R^2.$$

Thus $r = \pm\sqrt{R^2}$ and the coefficient of determination is the square of the sample correlation coefficient.

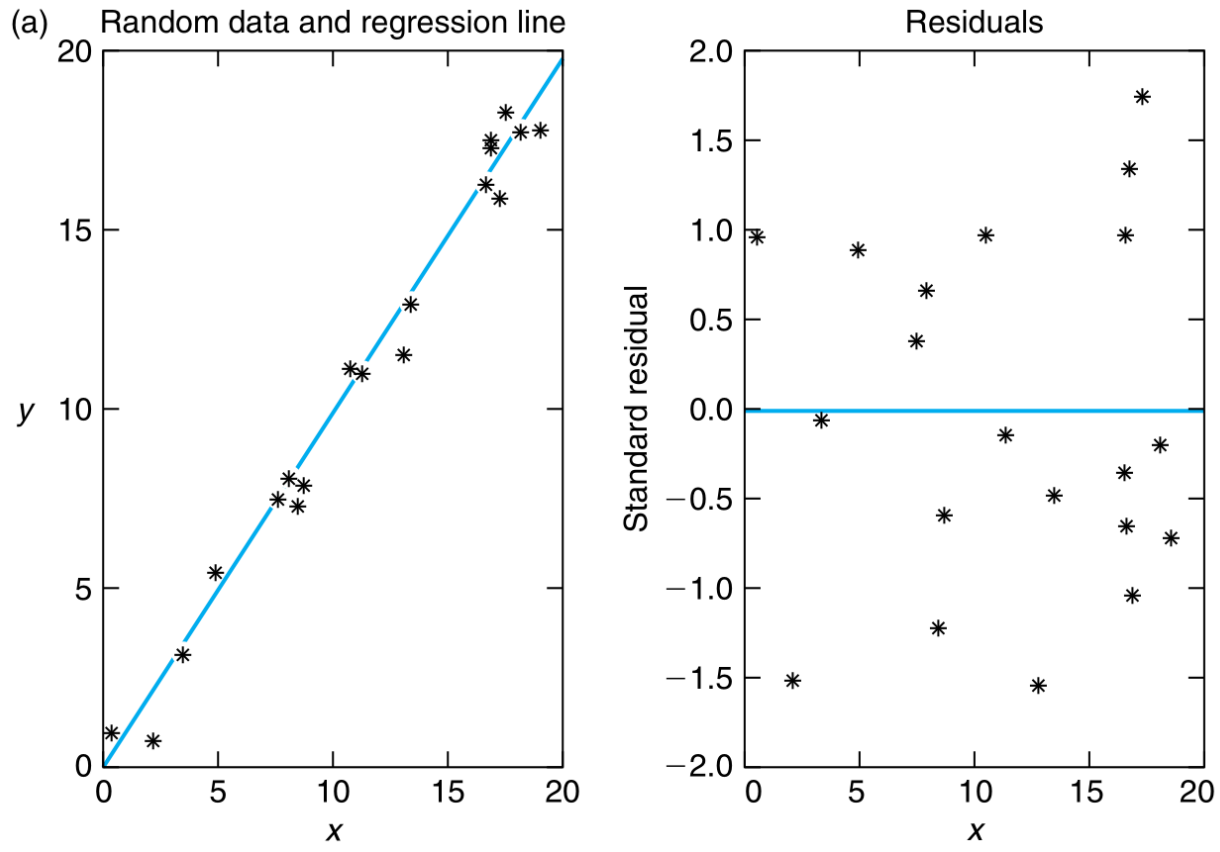
7.2.6 Validity of a linear model

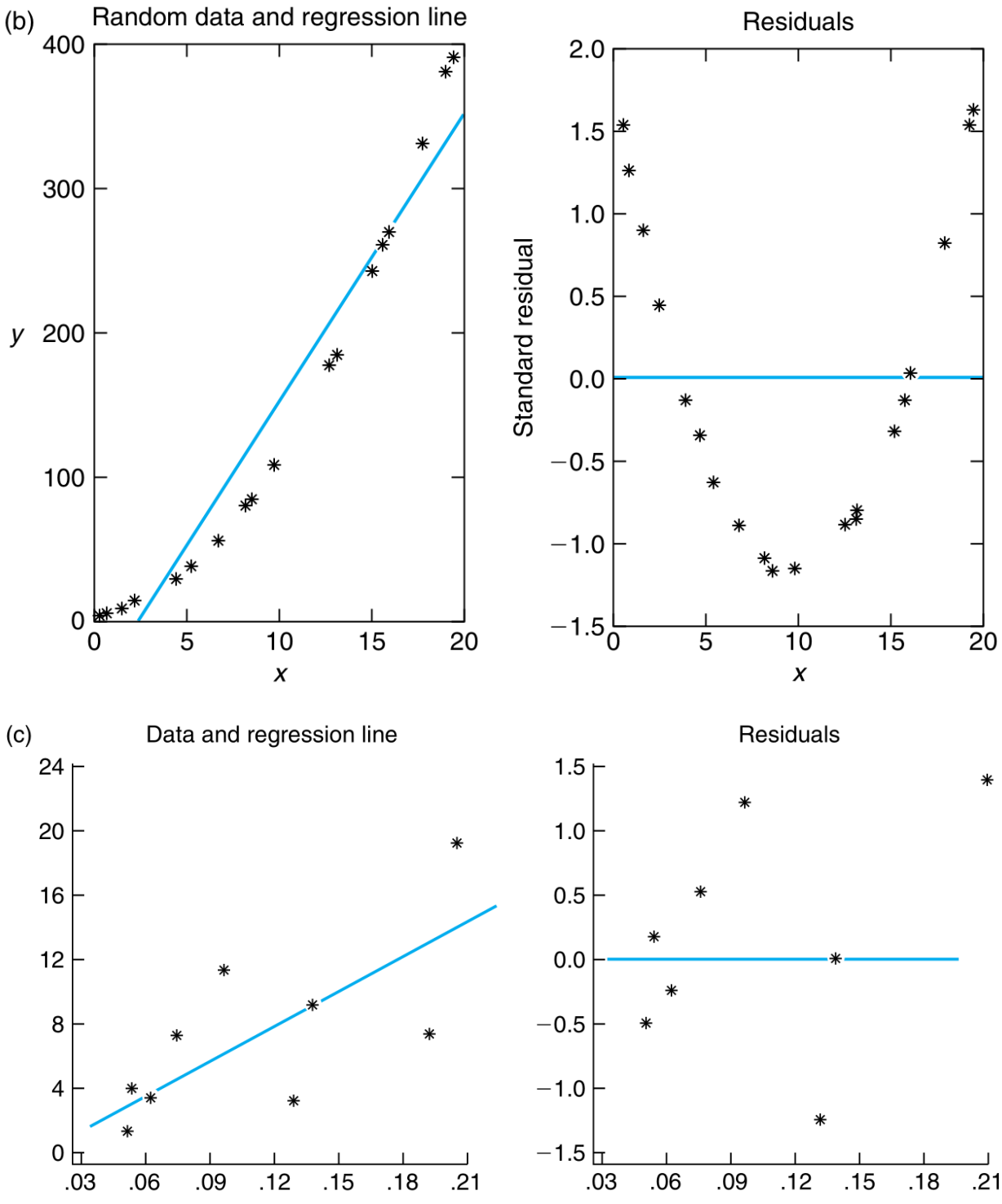
How well does a linear regression model fit the data? There are several ways to evaluate a model fit:

- Compute R^2 or r. The fit is good if R^2 is close to 1, i.e., r close to 1 or -1.
- Compute the approximately standardized residuals

  $$\frac{Y_i - \hat{\alpha} - \hat{\beta}x_i}{\sqrt{\mathsf{SS}_\mathsf{R}/(n-2)}}, \qquad i = 1, \ldots, n.$$

  If the linear model fits well, the standardized residuals are approximately standard normal.
- Draw pictures.




(b)
Random data and regression line
400
300
200
100
0
y
0
5
10
15
20
x
Residuals
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Standard residual
0
5
10
15
20
x
(c)
Data and regression line
24
20
16
12
8
4
0
.03
.06
.09
.12
.15
.18
.21
Residuals
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
.03
.06
.09
.12
.15
.18
.21

7.3 Inference in linear regression

7.3.1 Hypothesis testing for the slope

Assume the simple linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$ where $i = 1, \dots, n$. Consider testing $H_0 : \beta = 0$ against $\beta \neq 0$, i.e., whether x_i has effect on Y_i. Note that

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{xx}}} = \frac{\sqrt{S_{xx}}}{\sigma}(\hat{\beta} - \beta) \sim \mathcal{N}(0, 1),$$

Therefore, if we know σ^2, then we can do the Z-test using the test statistic $\frac{\sqrt{S_{xx}}}{\sigma}\hat{\beta}$.

For the case where σ^2 is unknown, recall that $\frac{\mathsf{SS_R}}{n-2}$ is an unbiased estimator of σ^2. As $\frac{\sqrt{S_{xx}}}{\sigma}(\hat{\beta} - \beta)$ is independent of

$$\frac{\mathsf{SS_R}}{\sigma^2} \sim \chi^2_{n-2},$$

we have that

$$\frac{\frac{\sqrt{S_{xx}}}{\sigma}(\hat{\beta} - \beta)}{\sqrt{\frac{\mathsf{SS_R}}{\sigma^2(n-2)}}} = \sqrt{\frac{(n-2)S_{xx}}{\mathsf{SS_R}}}(\hat{\beta} - \beta)$$

follows the t-distribution with $n - 2$ degrees of freedom. If H_0 is true, the above becomes

$$\sqrt{\frac{(n-2)S_{xx}}{\mathsf{SS_R}}} \cdot \hat{\beta}$$

which we use as the test statistic for the t-test.

Example. A person claims that fuel consumption of their automobile does not depend on how fast the car is driven. To test this hypothesis, the car was tested at the following speeds in miles per hour. The fuel consumption in miles per gallon attained at each of these speeds was determined with the following results.

Speed:	45	50	55	60	65	70	75
MPG:	24.2	25.0	23.3	22.0	21.5	20.6	19.8

Do these data refute the claim that fuel consumption is unaffected by the speed at which the car is being driven at the significance level 1 percent?

We compute $\bar{x} = 60$, $\bar{Y} \approx 22.343$,

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 700, \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \approx 21.757, \qquad S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = -119,$$

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} = -0.17, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \approx 32.543, \qquad \mathsf{SS_R} = \frac{S_{xx}S_{YY} - (S_{xY})^2}{S_{xx}} \approx 1.527.$$

It follows that

$$\sqrt{\frac{(n-2)S_{xx}}{\mathsf{SS_R}}} \cdot |\hat{\beta}| \approx 8.138 > 4.032 \approx |t_{0.005,5}|,$$

so we reject H_0. There is evidence that increased speeds lead to decreased fuel consumption, because $\hat{\beta}$ is negative and $Y_i \approx \hat{\alpha} + \hat{\beta}x_i$.

```
n = 7
x <- c(45,50,55,60,65,70,75)
y <- c(24.2,25.0,23.3,22.0,21.5,20.6,19.8)
x_bar = mean(x)
y_bar = mean(y)
print(c(x_bar,y_bar))
```

```
## [1] 60.00000 22.34286
```

```
s_xx = (n-1)*var(x)
s_yy = (n-1)*var(y)
s_xy = (n-1)*cov(x,y)
print(c(s_xx,s_yy,s_xy))
```

```
## [1]  700.00000   21.75714 -119.00000
```

```
beta_hat = s_xy/s_xx
alpha_hat = y_bar-beta_hat*x_bar
print(c(beta_hat,alpha_hat))
```

```
## [1] -0.17000 32.54286
```

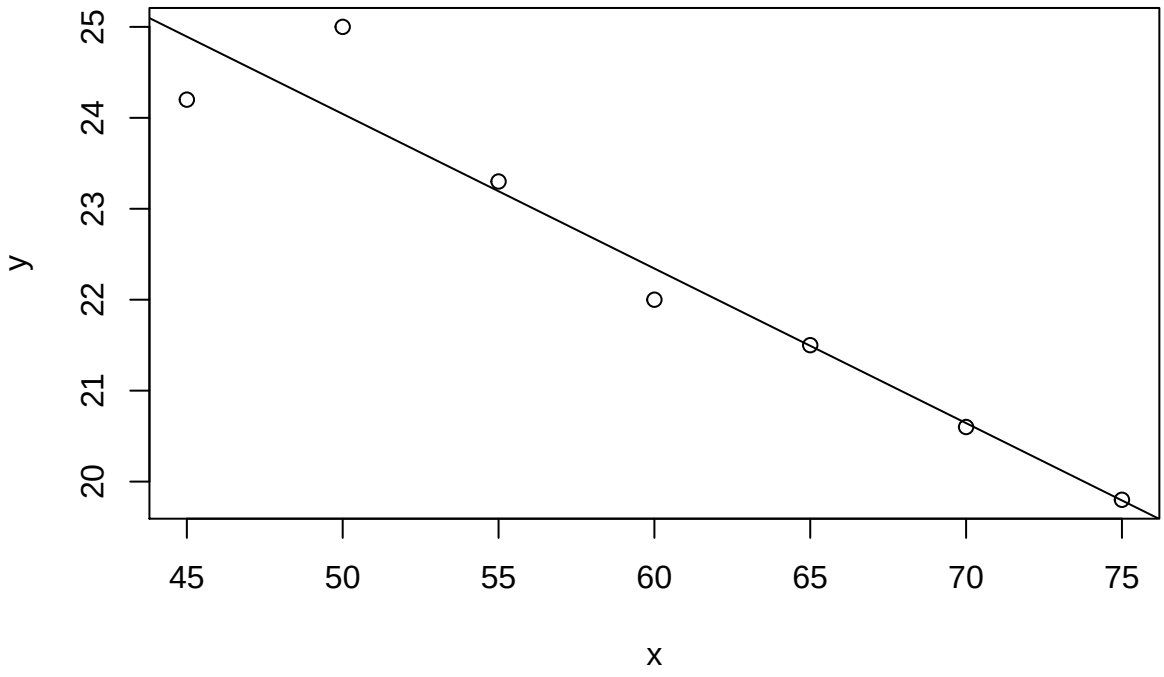
```
ss_r = (s_xx*s_yy-s_xy^2)/s_xx
ts = sqrt((n-2)*s_xx/ss_r)*beta_hat
print(c(ss_r,ts,qt(1-0.005,n-2)))
```

```
## [1]  1.527143 -8.138476  4.032143
```

```
delta = sqrt(ss_r/(n-2)/s_xx)*qt(1-0.025,n-2)
print(c(beta_hat-delta,beta_hat+delta))
```

```
## [1] -0.2236954 -0.1163046
```

```
plot(x,y)
abline(lm(y~x))
```



7.3.2 Confidence interval for the slope

Since

$$\sqrt{\frac{(n-2)S_{xx}}{\mathsf{SS}_\mathsf{R}}}(\hat{\beta}-\beta)$$

follows the t-distribution with $n-2$ degrees of freedom, we have that for any $\gamma \in (0,1)$,

$$\mathbb{P}\Big\{-t_{\gamma/2,n-2} < \sqrt{\frac{(n-2)S_{xx}}{\mathsf{SS}_\mathsf{R}}}(\hat{\beta}-\beta) < t_{\gamma/2,n-2}\Big\} = 1-\gamma,$$

$$\mathbb{P}\Big\{\hat{\beta} - \sqrt{\frac{\mathsf{SS}_\mathsf{R}}{(n-2)S_{xx}}}t_{\gamma/2,n-2} < \beta < \hat{\beta} + \sqrt{\frac{\mathsf{SS}_\mathsf{R}}{(n-2)S_{xx}}}t_{\gamma/2,n-2}\Big\} = 1-\gamma.$$

Therefore, the two-sided $(1-\gamma)$ confidence interval for β is

$$\left(\hat{\beta} - \sqrt{\frac{\mathsf{SS}_\mathsf{R}}{(n-2)S_{xx}}}t_{\gamma/2,n-2}, \hat{\beta} + \sqrt{\frac{\mathsf{SS}_\mathsf{R}}{(n-2)S_{xx}}}t_{\gamma/2,n-2}\right).$$

The one-sided versions can be derived analogously.

Example. In the above example, we can compute the two-sided 95 percent confidence interval for β, which is approximately $(-0.224, -0.116)$.

7.3.3 Inference for the intercept

Similarly, if σ is known, we can standardize $\hat{\alpha}$ to obtain the test statistic for the Z-test about α.

If σ is unknown, it can be shown that

$$\sqrt{\frac{n(n-2)S_{xx}}{\mathsf{SS}_\mathsf{R}\sum_{i=1}^n x_i^2}}(\hat{\alpha}-\alpha)$$

follows the t-distribution with $n-2$ degrees of freedom. Hence it can be used as the test statistic for the t-test. The two-sided $(1-\gamma)$ confidence interval for α is

$$\left(\hat{\alpha} - t_{\gamma/2,n-2}\sqrt{\frac{\mathsf{SS}_\mathsf{R}\sum_{i=1}^n x_i^2}{n(n-2)S_{xx}}}, \hat{\alpha} + t_{\gamma/2,n-2}\sqrt{\frac{\mathsf{SS}_\mathsf{R}\sum_{i=1}^n x_i^2}{n(n-2)S_{xx}}}\right).$$

7.3.4 Prediction of a mean response

In addition to (x_i, Y_i) for $i = 1, \ldots, n$, suppose that we are given a new variable x_0 and aim to estimate $\alpha + \beta x_0$. The natural estimator is $\hat{\alpha} + \hat{\beta}x_0$. It can be shown that

$$\frac{\hat{\alpha} + \hat{\beta}x_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\sqrt{\frac{\mathsf{SS}_\mathsf{R}}{n-2}}}$$

follows the t-distribution with $n-2$ degrees of freedom, which can be used as the test statistic for the t-test.

The two-sided $(1-\gamma)$ confidence interval for $\alpha + \beta x_0$ is

$$\left(\hat{\alpha} + \hat{\beta}x_0 - t_{\gamma/2,n-2}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\sqrt{\frac{\mathsf{SS}_\mathsf{R}}{n-2}}, \quad \hat{\alpha} + \hat{\beta}x_0 + t_{\gamma/2,n-2}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\sqrt{\frac{\mathsf{SS}_\mathsf{R}}{n-2}}\right).$$

7.3.5 Prediction of a future response

Given (x_i, Y_i) for $i = 1, \ldots, n$ and x_0, suppose that we aim to estimate $Y_0 = \alpha + \beta x_0 + \epsilon_0$ where $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$. The natural estimator is still $\hat{\alpha} + \hat{\beta} x_0$. It can be shown that

$$\frac{\hat{\alpha} + \hat{\beta} x_0 - Y_0}{\sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{\mathsf{SS_R}}{n-2}}}$$

follows the t-distribution with $n-2$ degrees of freedom, which can be used as the test statistic for the t-test. The two-sided $(1-\gamma)$ confidence interval for Y_0 is

$$\left(\hat{\alpha} + \hat{\beta} x_0 - t_{\gamma/2, n-2} \sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{\mathsf{SS_R}}{n-2}}, \quad \hat{\alpha} + \hat{\beta} x_0 + t_{\gamma/2, n-2} \sqrt{\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{\mathsf{SS_R}}{n-2}} \right).$$

7.4 Variants of a linear model

7.4.1 Transform to linearity

Consider the following situation: The amplitude of a signal at time t is

$$W(t) \approx c \cdot e^{-d \cdot t}.$$

Taking logarithms yields

$$\log W(t) \approx \log c - d \cdot t.$$

Let

$$Y = \log W(t), \qquad \alpha = \log c, \qquad \beta = -d.$$

The relation becomes

$$Y = \alpha + \beta t + \epsilon$$

for a noise term ϵ. This is the linear regression model.

Example. The following table lists the percentages of a chemical (denoted by $f(x)$) used in an experiment that is run at various temperatures in Celsius (denoted by x). Assume the relation $1 - f(x) \approx c(1-d)^x$ for constants c and d. Use the data to estimate the percentage of the chemical that would be used if the experiment were to be run at 350 degrees.

Temperature x	5	10	20	30	40	50	60	80
Percentage $f(x)$	0.061	0.113	0.192	0.259	0.339	0.401	0.461	0.551

Taking logarithms yields

$$\log(1 - f(x)) \approx \log c + x \log(1-d).$$

For

$$Y = -\log(1 - f(x)), \qquad \alpha = -\log c, \qquad \beta = -\log(1-d),$$

we obtain

$$Y = \alpha + \beta x + \epsilon.$$

Then we can do linear regression:

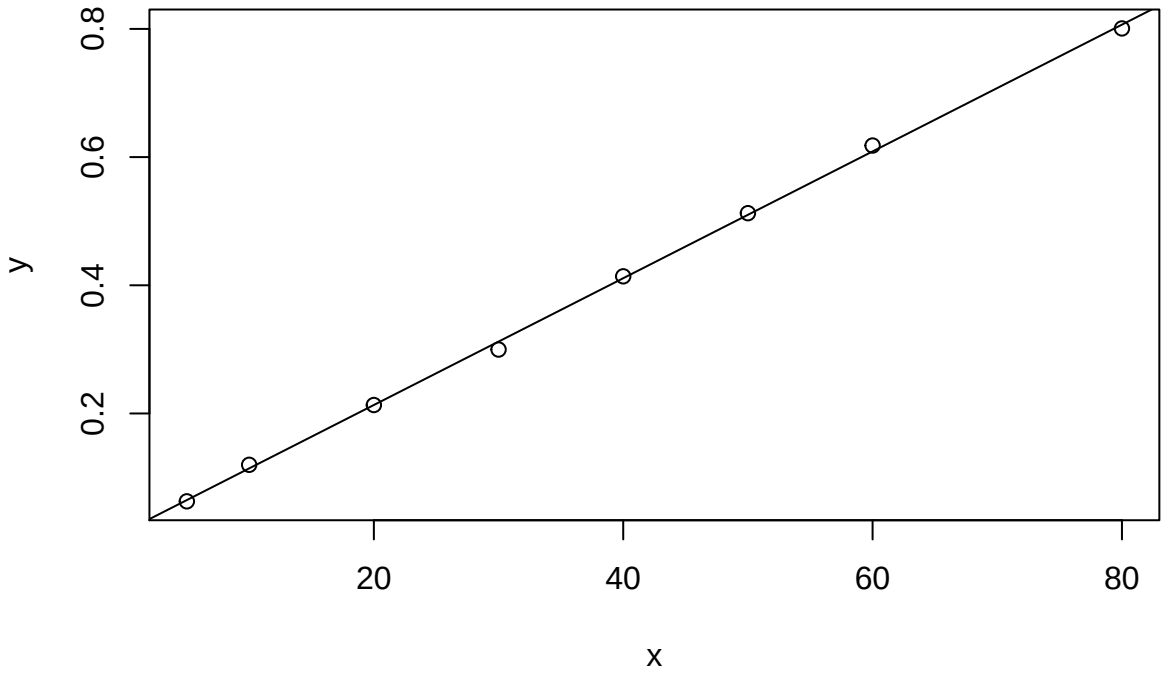
```
x <- c(5,10,20,30,40,50,60,80)
f <- c(.061,.113,.192,.259,.339,.401,.461,.551)
y = -log(1-f)
fit <- lm(y ~ x)
print(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.01545      0.00989
```

```
alpha = unname(coef(fit)[1])
beta = unname(coef(fit)[2])
c = exp(-alpha)
d = 1-exp(-beta)
x_new = 350
y_new = alpha + beta*x_new
f_new = 1-exp(-y_new)
print(c(c, d, f_new))
```

```
## [1] 0.984672526 0.009841073 0.969096383
```

```
plot(x,y)
abline(fit)
```



7.4.2 Linear regression with different variances

In the linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots n,$$

suppose that the noise terms have different variances $\mathrm{Var}(\epsilon_i) = \mathrm{Var}(Y_i) = \sigma_i^2$. If σ_i is known for each i, then it suffices to do the rescaling

$$\frac{Y_i}{\sigma_i} = \frac{\alpha}{\sigma_i} + \frac{\beta x_i}{\sigma_i} + \frac{\epsilon_i}{\sigma_i},$$

so that the noise $\tilde{\epsilon}_i = \frac{\epsilon_i}{\sigma_i}$ has variance $\mathrm{Var}(\tilde{\epsilon}_i) = 1$. Therefore, we can compute the least squares estimator

$$(\hat{\alpha}, \hat{\beta}) = \operatorname*{argmin}_{(\alpha', \beta')} \sum_{i=1}^{n} \Big(\frac{Y_i}{\sigma_i} - \frac{\alpha'}{\sigma_i} - \frac{\beta' x_i}{\sigma_i} \Big)^2 = \operatorname*{argmin}_{(\alpha', \beta')} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (Y_i - \alpha' - \beta' x_i)^2.$$

Such estimators are called weighted least squares estimators.

Given weights $w_1, \ldots, w_n$, how do we compute

$$(\hat{\alpha}, \hat{\beta}) = \operatorname*{argmin}_{(\alpha', \beta')} \sum_{i=1}^{n} w_i (Y_i - \alpha' - \beta' x_i)^2?$$

As before, let us take the partial derivatives with respect to α' and β' to obtain

$$\sum_{i=1}^{n} w_i (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0, \qquad \sum_{i=1}^{n} x_i w_i (Y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

that is,

$$\sum_{i=1}^{n} w_i Y_i = \hat{\alpha} \sum_{i=1}^{n} w_i + \hat{\beta} \sum_{i=1}^{n} w_i x_i, \qquad \sum_{i=1}^{n} w_i x_i Y_i = \hat{\alpha} \sum_{i=1}^{n} w_i x_i + \hat{\beta} \sum_{i=1}^{n} w_i x_i^2.$$

This is a system of linear equations in $\hat{\alpha}$ and $\hat{\beta}$, so we can solve for the estimators.

It is common that $\mathrm{Var}(Y_i)$ is proportional to x_i. For example, if x_i is the distance between two places and Y_i is the travel time, then it is reasonable to assume that

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

where $\mathrm{Var}(Y_i) = \mathrm{Var}(\epsilon_i) = c\, x_i$ for an unknown constant $c > 0$. From the above discussion, the system of linear equations is

$$\sum_{i=1}^{n} Y_i / x_i = \hat{\alpha} \sum_{i=1}^{n} 1/x_i + \hat{\beta} n, \qquad \sum_{i=1}^{n} Y_i = \hat{\alpha} n + \hat{\beta} \sum_{i=1}^{n} x_i.$$

Note that the constant c gets canceled out. Then we can solve for $\hat{\alpha}$ and $\hat{\beta}$.

Example. The following data represent travel times in a city, with distance x in miles and travel time Y in minutes. Assume a linear relation $Y = \alpha + \beta x + \epsilon$ and that the variance of Y is proportional to x. Find the weighted least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ in the linear regression model.

Distance	0.5	1.0	1.5	2.0	3.0	4.0	5.0	6.0	8.0	10.0
Travel time	15.0	15.1	16.5	19.9	27.7	29.7	26.7	35.9	42.0	49.4

We can compute

$$\sum_{i=1}^{n} Y_i / x_i \approx 104.22, \qquad \sum_{i=1}^{n} 1/x_i \approx 5.34, \qquad \sum_{i=1}^{n} Y_i = 277.9, \qquad \sum_{i=1}^{n} x_i = 41.$$

The system we need to solve is

$$104.22 \approx 5.34 \hat{\alpha} + 10 \hat{\beta}, \qquad 277.9 \approx 10 \hat{\alpha} + 41 \hat{\beta},$$

which yields $\hat{\alpha} \approx 12.56$ and $\hat{\beta} \approx 3.71$.

```r
x <- c(.5,1,1.5,2,3,4,5,6,8,10)
y <- c(15,15.1,16.5,19.9,27.7,29.7,26.7,35.9,42,49.4)
print(c(sum(y/x),sum(1/x),sum(y),sum(x)))
```

```
## [1] 104.221667   5.341667 277.900000  41.000000
```

```r
fit <- lm(y ~ x)
print(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      12.722        3.675
```

Note that the (unweighted) least squares estimators are slightly different.

8 Advanced linear models

8.1 Multiple linear regression

Let us consider the linear regression model where there are k factors affecting the outcome. Suppose that we observe n normally distributed data points. The model can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i = \beta^\top x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where the parameter vector $\beta \in \mathbb{R}^{k+1}$ has entries $\beta_0, \beta_1, \ldots, \beta_k$, each independent variable $x_i \in \mathbb{R}^{k+1}$ has entries $1, x_{i1}, \ldots, x_{ik}$, and the error terms ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Equivalently, we can write the model in a matrix form as

$$Y = X\beta + \epsilon,$$

where $Y, \epsilon \in \mathbb{R}^n$ are the vectors with entries Y_i and ϵ_i respectively, and $X \in \mathbb{R}^{n \times (k+1)}$ is the matrix with rows $x_i^\top$.

8.1.1 Least-squares estimator

The least squares estimators are $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ (or written as a vector $\hat{\beta} \in \mathbb{R}^{k+1}$) that minimize

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}^\top x_i)^2 = \|Y - X\hat{\beta}\|_2^2.$$

Note that there are $k+1$ unknown variables $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Taking the partial derivatives of the above quantity with respect to them, we obtain a system of $k+1$ linear equations, so we can solve for the unknown variables. However, this is too tedious. Alternatively, it is much easier to use matrix calculus to achieve the same. Namely, setting

$$\frac{d}{d\hat{\beta}} \|Y - X\hat{\beta}\|_2^2 = 2X^\top (Y - X\hat{\beta}) = 0$$

yields

$$X^\top X \hat{\beta} = X^\top Y.$$

Assuming that $n \geq k+1$ (i.e., more observations than unknown parameters, which is reasonable) and that $X^\top X \in \mathbb{R}^{(k+1)\times(k+1)}$ is invertible, we have

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Example. The following data relate the suicide rate to the population size and the divorce rate at eight different locations:

Location	Population in Thousands	Divorce Rate per 100,000	Suicide Rate per 100,000
Akron, OH	679	30.4	11.6
Anaheim, CA	1,420	34.1	16.1
Buffalo, NY	1,349	17.2	9.3
Austin, TX	296	26.8	9.1
Chicago, IL	6,975	29.1	8.4
Columbia, SC	323	18.7	7.7
Detroit, MI	4,200	32.6	11.3
Gary, IN	633	32.5	8.4

We fit a multiple linear regression model to the data of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where Y is the suicide rate, x_1 is the population, and x_2 is the divorce rate. The following R code gives $\hat{\beta}_0 \approx 3.50735$, $\hat{\beta}_1 \approx -0.00025$, and $\hat{\beta}_2 \approx 0.26095$.

```
x1 = c(679,1420,1349,296,6975,323,4200,633)
x2 = c(30.4,34.1,17.2,26.8,29.1,18.7,32.6,32.5)
y = c(11.6,16.1,9.3,9.1,8.4,7.7,11.3,8.4)
fit = lm(y ~ x1 + x2)
print(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##   3.5073534   -0.0002477    0.2609466
```

```
X = matrix(c(1,1,1,1,1,1,1,1,x1,x2), nrow=8, ncol=3)
Y = matrix(y)
print(X)
```

```
##      [,1] [,2] [,3]
## [1,]    1  679 30.4
## [2,]    1 1420 34.1
## [3,]    1 1349 17.2
## [4,]    1  296 26.8
## [5,]    1 6975 29.1
## [6,]    1  323 18.7
## [7,]    1 4200 32.6
## [8,]    1  633 32.5
```

```
beta = solve(t(X) %*% X ) %*% t(X) %*% Y
print(beta)
```

```
##               [,1]
## [1,]  3.5073533595
## [2,] -0.0002477099
## [3,]  0.2609465576
```

8.1.2 Bias, variance, and covariance

The least squares estimator is unbiased:

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^\top X)^{-1}X^\top Y] \\ &= \mathbb{E}[(X^\top X)^{-1}X^\top (X\beta + \epsilon)] \\ &= \mathbb{E}[(X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top \epsilon] \\ &= \mathbb{E}[\beta + (X^\top X)^{-1}X^\top \epsilon] \\ &= \beta + (X^\top X)^{-1}X^\top \mathbb{E}[\epsilon] = \beta.\end{aligned}$$

The covariances between entries of $\hat{\beta}$ can be computed as follows: Let $M = (X^\top X)^{-1}X^\top$ and then $\hat{\beta} = MY$. As a result

$$\hat{\beta}_{i-1} = \sum_{l=1}^{n} M_{il}Y_l$$

so

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_{i-1}, \hat{\beta}_{j-1}) &= \mathrm{Cov}\Big(\sum_{l=1}^{n} M_{il}Y_l, \sum_{r=1}^{n} M_{jr}Y_r\Big) \\ &= \sum_{l=1}^{n}\sum_{r=1}^{n} M_{il}M_{jr}\mathrm{Cov}(Y_l, Y_r) \\ &= \sum_{l=1}^{n} M_{il}M_{jl}\mathrm{Var}(Y_l) \\ &= \sigma^2 \sum_{l=1}^{n} M_{il}M_{jl} \\ &= \sigma^2 (MM^\top)_{ij}.\end{aligned}$$

This can be written in the matrix form as

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 MM^\top = \sigma^2 (X^\top X)^{-1}X^\top X(X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}.$$

In particular, $\mathrm{Var}(\hat{\beta}_i)$ are the diagonal elements of $\sigma^2(X^\top X)^{-1}$, and

$$\hat{\beta}_{i-1} \sim \mathcal{N}\Big(\beta_{i-1}, \sigma^2[(X^\top X)^{-1}]_{ii}\Big).$$

8.1.3 Residuals

The sum of squares of residuals can be similarly defined as

$$\begin{aligned}\mathsf{SS}_\mathsf{R} &= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{\beta}^\top x_i)^2 \\ &= \|Y - X\hat{\beta}\|_2^2 \\ &= (Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\ &= Y^\top Y - Y^\top X\hat{\beta} - \hat{\beta}^\top X^\top Y + \hat{\beta}^\top X^\top X\hat{\beta} \\ &= Y^\top Y - Y^\top X\hat{\beta},\end{aligned}$$

where the second-to-last equality holds because $X^\top X\beta = X^\top Y$.

It can shown that

$$\frac{\mathsf{SS}_\mathsf{R}}{\sigma^2} \sim \chi^2_{n-(k+1)},$$

so $\mathbb{E}[\frac{\mathsf{SS}_\mathsf{R}}{\sigma^2}] = n-k-1$ or

$$\mathbb{E}\Big[\frac{\mathsf{SS}_\mathsf{R}}{n-k-1}\Big] = \sigma^2.$$

Therefore, $\frac{\mathsf{SS}_\mathsf{R}}{n-k-1}$ is an unbiased estimator of σ^2. In fact, SS_R is independent of $\hat\beta = (\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_k)^\top$.

Example. In the previous example, we can compute SS_R and $\frac{\mathsf{SS}_\mathsf{R}}{n-k-1}$ as follows:

```
ssr = t(Y) %*% Y - t(Y) %*% X %*% beta
print(ssr)
```

```
##           [,1]
## [1,] 34.12123
```

```
sigma2 = ssr/(8-2-1)
print(sigma2)
```

```
##           [,1]
## [1,] 6.824247
```

The coefficient of determination is defined as

$$R^2 = 1 - \frac{\mathsf{SS}_\mathsf{R}}{\sum_{i=1}^n (Y_i - \bar Y)^2}.$$

8.1.4 Inference

Estimating a parameter: We have, for $i = 1, \ldots, k+1$,

$$\frac{\hat\beta_{i-1} - \beta_{i-1}}{\sigma\sqrt{[(X^\top X)^{-1}]_{ii}}} \sim \mathcal{N}(0,1),$$

so

$$\frac{\hat\beta_{i-1} - \beta_{i-1}}{\sigma\sqrt{[(X^\top X)^{-1}]_{ii}}\sqrt{\mathsf{SS}_\mathsf{R}/\sigma^2/(n-k-1)}} = \frac{\hat\beta_{i-1} - \beta_{i-1}}{\sqrt{[(X^\top X)^{-1}]_{ii}\cdot \mathsf{SS}_\mathsf{R}/(n-k-1)}}$$

has the t-distribution with $n-k-1$ degrees of freedom. We can use this fact to do hypothesis testing and build confidence intervals as before.

Predicting a mean response: Suppose that we are given a new covariate vector $x_0 = (1, x_{01}, \ldots, x_{0k})^\top$. Then the predicted response is

$$\sum_{i=0}^k \hat\beta_i x_{0i} = \hat\beta^\top x_0.$$

To build confidence intervals for $\sum_{i=0}^k \beta_i x_{0i}$, it suffices to use the fact that

$$\frac{\sum_{i=0}^k \hat\beta_i x_{0i} - \sum_{i=0}^k \beta_i x_{0i}}{\sqrt{x_0^\top (X^\top X)^{-1} x_0 \cdot \mathsf{SS}_\mathsf{R}/(n-k-1)}}$$

has the t-distribution with $n-k-1$ degrees of freedom.

Predicting a future response: To build confidence intervals for $Y_0 = \beta^\top x_0 + \epsilon_0$ where $\epsilon_0 \sim \mathcal{N}(0,\sigma^2)$, it suffices to use the fact that

$$\frac{\sum_{i=0}^k \hat\beta_i x_{0i} - Y_0}{\sqrt{[x_0^\top (X^\top X)^{-1} x_0 + 1] \cdot \mathsf{SS}_\mathsf{R}/(n-k-1)}}$$

has the t-distribution with $n-k-1$ degrees of freedom.

8.2 Polynomial regression

Let us consider the case where the relation between the dependent variable Y and the independent variable x is approximated by a polynomial with degree $r \geq 2$. This is a generalization of linear regression because a linear function is a polynomial of degree 1. More formally, we have

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + \epsilon,$$

where $\beta_0, \beta_1, \ldots, \beta_r$ are the regression coefficients to be estimated, and ϵ is the error term.

Suppose we observe n pairs (x_i, Y_i) for $i = 1, \ldots, n$ satisfying

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i.$$

Then the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$ are the quantities that minimize

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2 - \cdots - \hat{\beta}_r x_i^r)^2.$$

To solve for $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$, we can take the partial derivatives of the above quantity with respect to the $r+1$ variables and set them to zero. The system of $r+1$ linear equations will yield the estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$.

Example. Let $x_i = i$ for $i = 1, \ldots, 10$, and let Y_i be

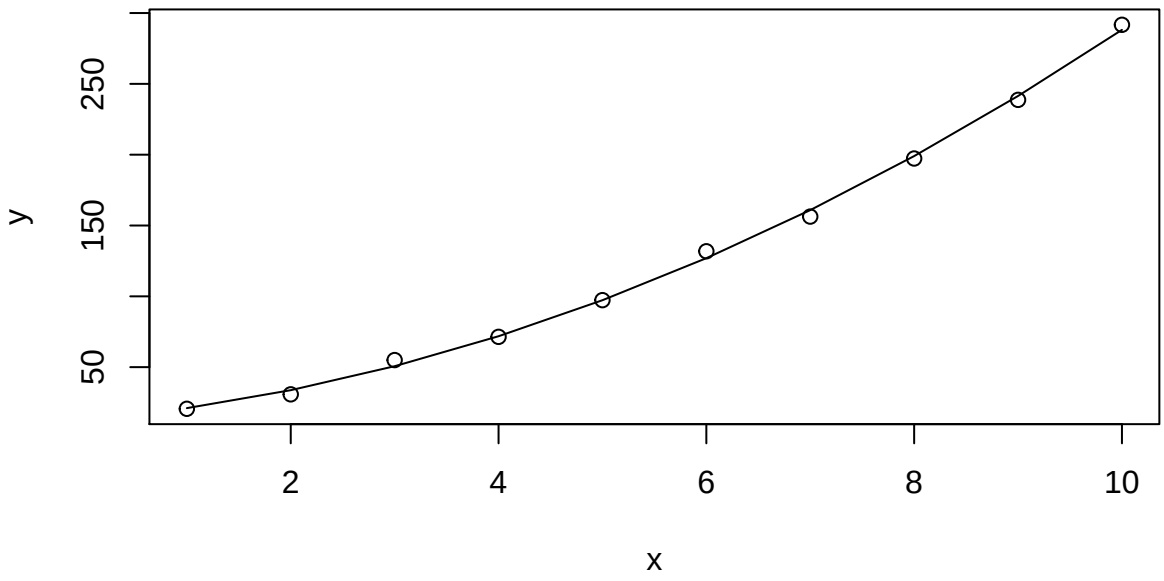
$$20.6,\ 30.8,\ 55,\ 71.4,\ 97.3,\ 131.8,\ 156.3,\ 197.3,\ 238.7,\ 291.7.$$

Let us do a quadratic fit using R:

```
x = seq(1, 10)
y = c(20.6, 30.8, 55, 71.4, 97.3, 131.8, 156.3, 197.3, 238.7, 291.7)
fit = lm(y ~ 1 + x + I(x^2))
print(fit)
```

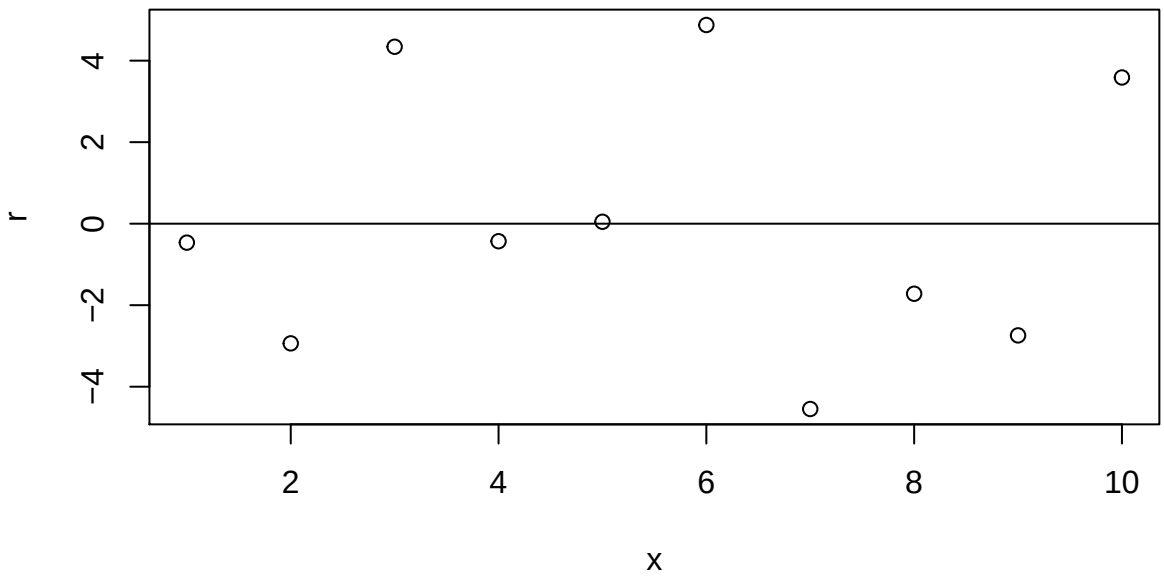
```
##
## Call:
## lm(formula = y ~ 1 + x + I(x^2))
##
## Coefficients:
## (Intercept)            x       I(x^2)
##      12.643        6.297        2.125
```

```
plot(x, y)
points(x, predict(fit), type="l")
```



Check the residuals:

```
r = y - (12.643 + 6.297*x + 2.125*x^2)
plot(x, r)
abline(0, 0)
```

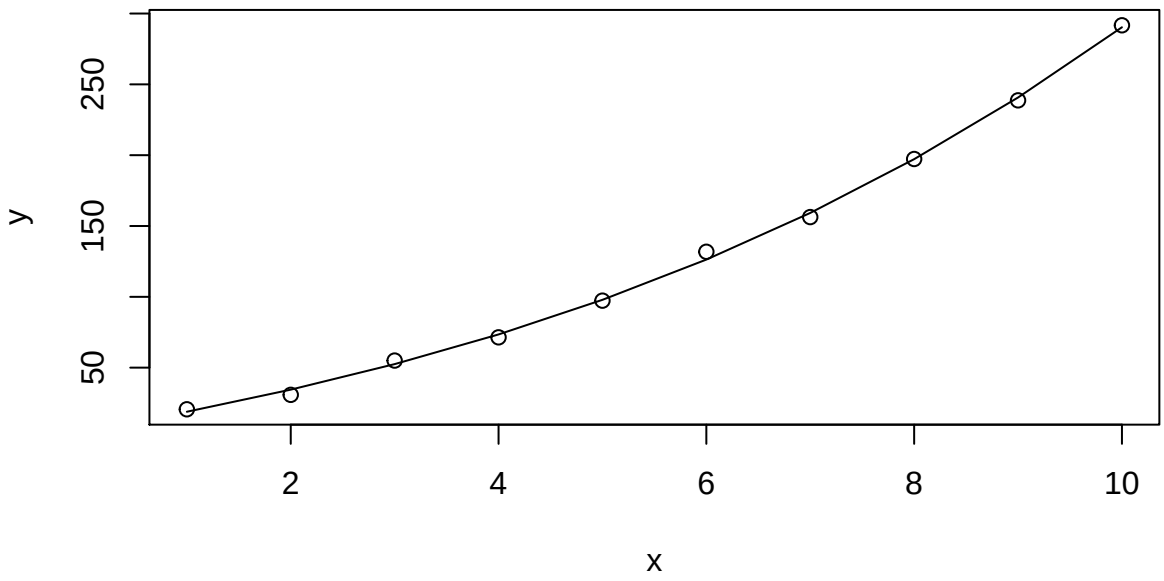


How about higher-order fit?

```
fit = lm(y ~ 1 + x + I(x^2) + I(x^3))
print(fit)
```

```
##
## Call:
## lm(formula = y ~ 1 + x + I(x^2) + I(x^3))
##
## Coefficients:
## (Intercept)            x       I(x^2)       I(x^3)
##     5.15667     12.93739      0.68526      0.08726
```

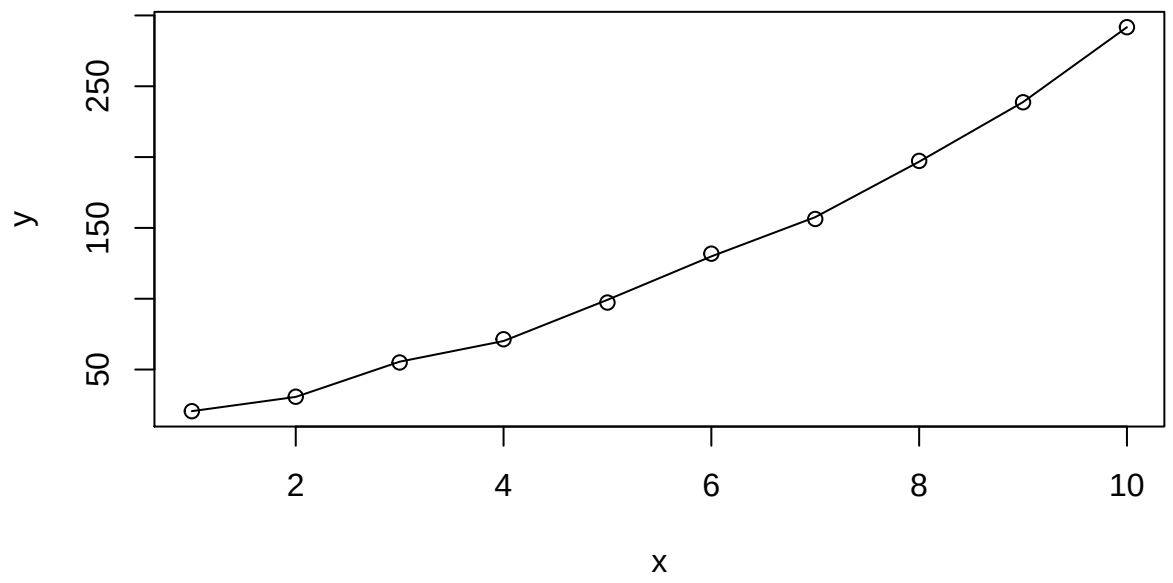
```
plot(x, y)
points(x, predict(fit), type="l")
```




```
fit = lm(y ~ 1 + x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6) + I(x^7) + I(x^8))
print(fit)
```

```
##
## Call:
## lm(formula = y ~ 1 + x + I(x^2) + I(x^3) + I(x^4) + I(x^5) +
##     I(x^6) + I(x^7) + I(x^8))
##
## Coefficients:
## (Intercept)            x       I(x^2)       I(x^3)       I(x^4)       I(x^5)
##   6.430e+02   -1.542e+03    1.457e+03   -7.043e+02    1.959e+02   -3.252e+01
##      I(x^6)       I(x^7)       I(x^8)
##   3.181e+00   -1.691e-01    3.765e-03
```

```
plot(x, y)
points(x, predict(fit), type="l")
```



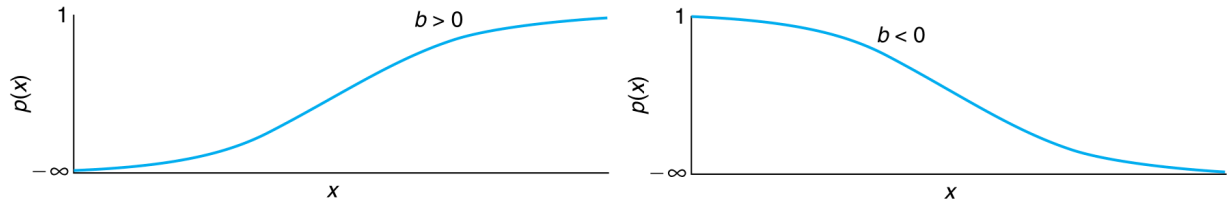
8.3 Regression with binary response

8.3.1 Logistic model

Consider independent experiments that result in a *binary* outcome: either a success or a failure. Suppose that a factor of value x in the experiment leads to a success with probability $p(x)$. The *logistic regression model* assumes

$$p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}.$$

The *logistic regression function* $p(x)$ is increasing if $b > 0$, and is decreasing if $b < 0$:



Let

$$o(x) = \frac{p(x)}{1-p(x)} = e^{a+bx}$$

be the *odds* for success. Then the log odds, called the *logit*, is a linear function:

$$\log(o(x)) = a + bx.$$

Here a and b are the parameters to be estimated.

8.3.2 Maximum likelihood estimation

Suppose that in each of n independent experiments, the independent variable x_i leads to a binary result Y_i. Here $Y_i = 1$ means a success and $Y_i = 0$ means a failure. Therefore, Y_i are independent Bernoulli random variables

$$Y_i \sim \text{Ber}\big(p(x_i)\big) = \text{Ber}\Big(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\Big).$$

Observing the data (x_i, Y_i) for $i = 1, \ldots, n$, we aim to estimate the parameters a and b.

For this, consider the PMF of each Y_i:

$$f_i(y_i \mid a, b) = \mathbb{P}\{Y_i = y_i\} = [p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i} = \Big(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\Big)^{y_i}\Big(\frac{1}{1+e^{a+bx_i}}\Big)^{1-y_i} = \frac{(e^{a+bx_i})^{y_i}}{1+e^{a+bx_i}}$$

where $y_i = 0$ or 1. Thus the joint PMF of $(Y_1, \ldots, Y_n)$ or the likelihood is

$$L(a, b \mid y_1, \ldots, y_n) = f(y_1, \ldots, y_n \mid a, b) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{(e^{a+bx_i})^{y_i}}{1+e^{a+bx_i}}$$

The log-likelihood is

$$\log L(a, b \mid y_1, \ldots, y_n) = \sum_{i=1}^{n} \log \frac{(e^{a+bx_i})^{y_i}}{1+e^{a+bx_i}} = \sum_{i=1}^{n} \big[y_i(a+bx_i) - \log(1+e^{a+bx_i})\big].$$

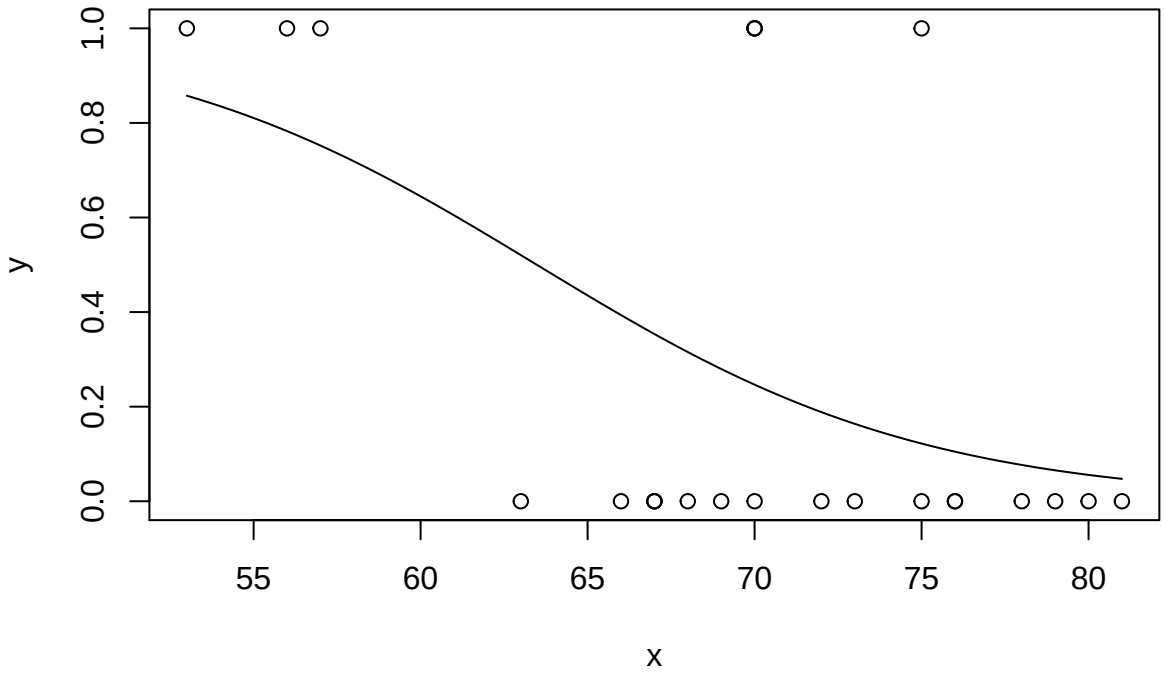
The maximum likelihood estimators $\hat{a}$ and $\hat{b}$ are the quantities a and b that maximize the above log-likelihood. There is no closed-form expression for $(\hat{a}, \hat{b})$, so we have to rely on some package to find the solution.

Example. Let us consider the relation between the temperature and whether a machine fails. The experiments are run at temperatures $53, 56, 57, 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81$. Whether the machine fails is indicated by a $0-1$ response: $1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0$. Here is how we fit the logistic regression model in R:

```
x = c(53,56,57,63,66,67,67,67,68,69,70,70,70,70,72,73,75,75,76,76,78,79,80,81)
y = c(1,1,1,0,0,0,0,0,0,0,0,1,1,1,0,0,0,1,0,0,0,0,0,0)
fit = glm(y ~ x, family = binomial)
print(fit)
```

```
##
## Call:  glm(formula = y ~ x, family = binomial)
##
## Coefficients:
## (Intercept)            x
##     10.8753      -0.1713
##
## Degrees of Freedom: 23 Total (i.e. Null);  22 Residual
## Null Deviance:       28.97
## Residual Deviance: 23.03     AIC: 27.03
```

```
plot(x,y)
a = unname(coef(fit)[1])
b = unname(coef(fit)[2])
x2 = seq(53,81,0.1)
y2 = exp(a+b*x2)/(1+exp(a+b*x2))
points(x2,y2,type="l")
```



8.3.3 Probit model

The probit model is very similar to the logistic regression model, with the difference being that the success probability is defined as

$$p(x) = \Phi(a + bx) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a+bx} e^{-y^2/2}\, dy.$$

Here Φ is the CDF of $\mathcal{N}(0, 1)$, so $p(x)$ is the probability that a standard normal random variable is less than $a + bx$.

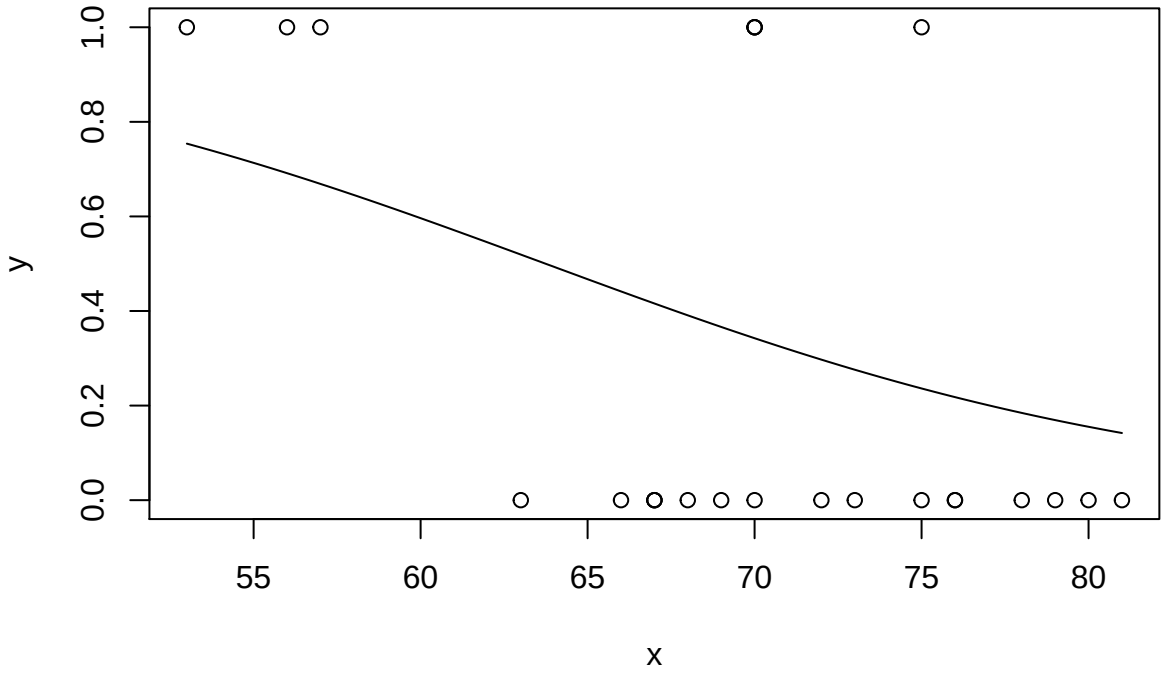
The maximum likelihood estimator can be defined similarly.

Example. We now fit a probit model to the above data:

```
fit = glm(y ~ x, family = binomial(link = "probit"))
print(fit)
```

```
##
## Call:  glm(formula = y ~ x, family = binomial(link = "probit"))
##
## Coefficients:
## (Intercept)            x
##      6.6444      -0.1042
##
## Degrees of Freedom: 23 Total (i.e. Null);  22 Residual
## Null Deviance:       28.97
## Residual Deviance: 22.98     AIC: 26.98
```

```
plot(x,y)
a = unname(coef(fit)[1])
b = unname(coef(fit)[2])
x2 = seq(53,81,0.1)
y2 = exp(a+b*x2)/(1+exp(a+b*x2))
points(x2,y2,type="l")
```



8.4 Analysis of variance

Example. A company is considering purchasing, in quantity, one of four different computer packages designed to teach employees a new programming language. To see the effectiveness of the four packages, the company can do an experiment as follows: Choose 160 employees and divide them into 4 groups of size 40. Each group uses one of the packages, and after a period of time, the effectivenss of the packages can be measured by an exam given to the employees.

The problem is, the scores of the employees in the exam will be stochastic: For example, if the average score of one group is higher than that of another group, it could be due to that one package is better than the other, but it could also be the case that this is due to random chance. How do we decide if one package is indeed better, or if the packages are interchangeable?

8.4.1 One-way analysis of variance

Throughout the discussion, the data are normally distributed with the same (but unknown) variance σ^2. Hence it suffices to specify the mean when we talk about how the data are generated.

Consider m independent samples, each of size n, consisting of independent

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \qquad i = 1, \ldots, m,\ j = 1, \ldots, n.$$

We are interested in testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ against H_1 : otherwise.

To this end, we will construct two estimators of σ^2 and compare them: The first estimator is good regardless of whether H_0 or H_1 is true, while the second is good only under H_0. As a result, comparing the two will serve as a test.

8.4.2 First estimator of variance

Since

$$(X_{ij} - \mu_i)/\sigma \sim \mathcal{N}(0, 1),$$

we have that

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \mu_i)^2/\sigma^2 \sim \chi^2_{mn}.$$

Because we do not know μ_i, the above quantity is not a statistic. But we can estimate μ_i by

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

for $i = 1, \ldots, m$. It can be shown that

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2/\sigma^2 \sim \chi^2_{nm-m}.$$

Define

$$\mathsf{SS}_\mathsf{w} = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

which is called the within samples sum of squares. Then $\mathsf{SS}_\mathsf{w}/\sigma^2 \sim \chi^2_{nm-m}$, so

$$\mathbb{E}[\mathsf{SS}_\mathsf{w}/\sigma^2] = nm - m, \qquad \mathbb{E}[\mathsf{SS}_\mathsf{w}/(nm - m)] = \sigma^2,$$

Thus $\mathsf{SS}_\mathsf{w}/(nm - m)$ is an unbiased estimator of σ^2.

8.4.3 Second estimator of variance

Let us assume $H_0 : \mu_1 = \cdots = \mu_m = \mu$. In this case, $\bar{X}_i \sim \mathcal{N}(\mu, \sigma^2/n)$, so

$$\frac{\bar{X}_i - \mu}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}}{\sigma}(\bar{X}_i - \mu) \sim \mathcal{N}(0, 1).$$

Hence

$$n\sum_{i=1}^{m}(\bar{X}_i - \mu)^2/\sigma^2 \sim \chi^2_m.$$

Furthermore, we can estimate μ by

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m}\bar{X}_i = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}.$$

It can be shown that

$$n\sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2/\sigma^2 \sim \chi^2_{m-1}.$$

Define

$$\mathsf{SS}_\mathsf{b} = n\sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2$$

which is called the between samples sum of squares. Then $\mathsf{SS}_\mathsf{b}/\sigma^2 \sim \chi^2_{m-1}$, so

$$\mathbb{E}[\mathsf{SS}_\mathsf{b}/\sigma^2] = m-1, \qquad \mathbb{E}[\mathsf{SS}_\mathsf{b}/(m-1)] = \sigma^2,$$

Thus $\mathsf{SS}_\mathsf{b}/(m-1)$ is an unbiased estimator of σ^2. In addition, it is known that this quantity tends to exceed σ^2 if H_0 is not true.

Moreover, a useful fact is that

$$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 = nm\bar{X}^2 + \mathsf{SS}_\mathsf{b} + \mathsf{SS}_\mathsf{w}.$$

8.4.4 Test statistic and F-distribution

The test statistic we use is

$$\mathrm{TS} = \frac{\mathsf{SS}_\mathsf{b}/(m-1)}{\mathsf{SS}_\mathsf{w}/(nm-m)}.$$

It is known that SS_b and SS_w are actually independent. It follows from the definition of the F-distribution that TS is an F-random variable with $m-1$ and $nm-m$ degrees of freedom. We define the quantile $f_{m-1,nm-m,\alpha}$ by

$$\mathbb{P}\{X \le f_{m-1,nm-m,\alpha}\} = 1-\alpha$$

where X denotes an F-random variable with $m-1$ and $nm-m$ degrees of freedom. Since an F-random variable takes only positive values, the F-test is similar to the chi-square test, rather than the Z-test or the t-test. In particular, we have that the p-value is

$$\mathbb{P}\{X > \mathrm{TS}\}.$$

Example. An auto rental firm is using 15 identical motors that run at a fixed speed to test 3 different brands of gasoline. Each brand of gasoline is assigned to exactly 5 of the motors. Each motor runs on 10 gallons of gasoline until it is out of fuel. The total mileages obtained by different motors are as follows:

Gas 1 :	220	251	226	246	260
Gas 2 :	244	235	232	242	225
Gas 3 :	252	272	250	238	256

Test the hypothesis that the average mileage is not affected by the type of gas used at the 5 percent level of significance.

We can compute the p-value $\mathbb{P}\{X > \mathrm{TS}\} \approx 0.115 > 0.05$ to see that H_0 is accepted.

```
x1 = c(220 , 251 , 226 , 246 , 260)
x2 = c(244 , 235 , 232 , 242 , 225)
x3 = c(252 , 272 , 250 , 238 , 256)
x_avg = (mean(x1) + mean(x2) + mean(x3))/3
n = 5
m = 3
SS_w = (n-1)*(var(x1) + var(x2) + var(x3))
SS_b = n*((mean(x1) - x_avg)^2 + (mean(x2) - x_avg)^2 + (mean(x3) - x_avg)^2)
TS = (SS_b/(m-1))/(SS_w/(n*m-m))
print(c(SS_w, SS_b, TS))
```

```
## [1] 1991.600000  863.333333    2.600924
```

```
print(1-pf(TS,m-1,n*m-m))
```

```
## [1] 0.1152489
```

```
print(qf(1-0.05,m-1,n*m-m))
```

```
## [1] 3.885294
```

8.4.5 One-way analysis of variance with unequal sample sizes

Now suppose that the samples $i = 1, \dots, m$ have sizes $n_1, \dots, n_m$ respectively, consisting of independent

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \qquad i = 1, \dots, m,\ j = 1, \dots, n_i.$$

Again, consider testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ against H_1 : otherwise.

Let

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \qquad \bar{X} = \frac{1}{m} \sum_{i=1}^{m} \bar{X}_i.$$

For similar reasons,

$$\mathsf{SS_w}/\sigma^2 \sim \chi^2_{\sum_{i=1}^m n_i - m}, \qquad \mathsf{SS_w} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

and $\mathsf{SS_w}/(\sum_{i=1}^m n_i - m)$ is an unbiased estimator of σ^2.

Moreover, if H_0 is true, then

$$\mathsf{SS_b}/\sigma^2 \sim \chi^2_{m-1}, \qquad \mathsf{SS_b} = \sum_{i=1}^{m} n_i (\bar{X}_i - \bar{X})^2,$$

and $\mathsf{SS_b}/(m-1)$ is an unbiased estimator of σ^2.

Finally, $\mathsf{SS_b}$ and $\mathsf{SS_w}$ are independent and

$$\frac{\mathsf{SS_b}/(m-1)}{\mathsf{SS_w}/(\sum_{i=1}^m n_i - m)}$$

is an F-random variable with $m-1$ and $\sum_{i=1}^m n_i - m$ degrees of freedom. We can use this quantity as the test statistic.

8.5 Two-way analysis of variance

Let us consider independent observations

$$X_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2), \qquad i = 1, \dots, m,\ j = 1, \dots n.$$

Note that the means may be all distinct, which is different from the setup in the last section. For example, X_{ij} can be the score of student j in exam i. In this section, we consider the two-factor additive model

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \qquad i = 1, \dots, m,\ j = 1, \dots n,$$

such that

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0.$$

We are interested in testing $H_0 : \alpha_1 = \cdots = \alpha_m = 0$ against H_1 : otherwise. Here H_0 means that

$$\mu_{ij} = \mu + \beta_j, \qquad j = 1, \dots n,$$

i.e., the mean is affected only by the column (e.g., which student), not the row (e.g., which exam).

The test will be constructed in a way similar to what we discussed in the last section: construct and compare two estimators of σ^2.

8.5.1 First estimator of variance

Since

$$(X_{ij} - \mu - \alpha_i - \beta_j)/\sigma \sim \mathcal{N}(0,1),$$

we have

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \mu - \alpha_i - \beta_j)^2/\sigma^2 \sim \chi^2_{nm}.$$

We need to replace μ, α_i and β_j by estimators. Define

$$\bar{X} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}X_{ij}, \qquad \bar{X}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}X_{ij}, \qquad \bar{X}_{\cdot j} = \frac{1}{m}\sum_{i=1}^{m}X_{ij}.$$

We claim:

- $\bar{X}$ is an unbiased estimator of μ;
- $\bar{X}_{i\cdot} - \bar{X}$ is an unbiased estimator of α_i;
- $\bar{X}_{\cdot j} - \bar{X}$ is an unbiased estimator of β_j.

Indeed,

$$\mathbb{E}[\bar{X}] = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mu_{ij} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\mu + \alpha_i + \beta_j) = \mu,$$

$$\mathbb{E}[\bar{X}_{i\cdot} - \bar{X}] = \frac{1}{n}\sum_{j=1}^{n}\mu_{ij} - \mu = \frac{1}{n}\sum_{j=1}^{n}(\mu + \alpha_i + \beta_j) - \mu = \alpha_i,$$

and similarly $\mathbb{E}[\bar{X}_{\cdot j} - \bar{X}] = \beta_j$.

Replacing μ, α_i and β_j in $\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \mu - \alpha_i - \beta_j)^2/\sigma^2$ by the corresponding estimators, one can show that

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2/\sigma^2 \sim \chi^2_{(n-1)(m-1)}.$$

Why is $(n-1)(m-1)$ the number of degrees of freedom? We need to estimate μ, α_i for $i = 1, \ldots, m$, and β_j for $j = 1, \ldots, n$, but $\sum_{i=1}^{m}\alpha_i = 0$ and $\sum_{j=1}^{n}\beta_j = 0$, so there are $1 + m + n - 2 = m + n - 1$ parameters. Hence there are $mn - (m + n - 1) = (m-1)(n-1)$ degrees of freedom.

We call the quantity

$$\mathsf{SS_e} = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2$$

the error sum of squares. Hence $\mathsf{SS_e}/\sigma^2 \sim \chi^2_{(n-1)(m-1)}$, and $\frac{\mathsf{SS_e}}{(n-1)(m-1)}$ is an unbiased estimator of σ^2.

8.5.2 Second estimator of variance

If H_0 is true, then

$$\mathbb{E}[\bar{X}_{i\cdot}] = \mu + \alpha_i = \mu$$

and

$$\mathrm{Var}(\bar{X}_{i\cdot}) = \sigma^2/n.$$

Hence

$$\frac{\bar{X}_{i\cdot} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \qquad n\sum_{i=1}^{m}(X_{i\cdot} - \mu)^2/\sigma^2 \sim \chi^2_m.$$

It is known that

$$n\sum_{i=1}^{m}(\bar{X}_{i\cdot}-\bar{X})^2/\sigma^2 \sim \chi^2_{m-1}.$$

We call the quantity

$$\mathsf{SS_r} = n\sum_{i=1}^{m}(\bar{X}_{i\cdot}-\bar{X})^2$$

the row sum of squares. Hence $\mathsf{SS_r}/\sigma^2 \sim \chi^2_{m-1}$, and $\frac{\mathsf{SS_r}}{m-1}$ is an unbiased estimator of σ^2.

Again, the above statements hold if H_0 is true. If H_0 is not true, then $\frac{\mathsf{SS_r}}{m-1}$ tends to be larger than σ^2.

8.5.3 Test statistics

For hypothesis testing between $H_0 : \alpha_1 = \cdots = \alpha_m = 0$ and H_1 : otherwise, the test statistic is

$$\text{TS} = \frac{\mathsf{SS_r}/(m-1)}{\mathsf{SS_e}/[(n-1)(m-1)]} = \frac{\mathsf{SS_r}(n-1)}{\mathsf{SS_e}},$$

which follows the F-distribution with $m-1$ and $(n-1)(m-1)$ degrees of freedom.

Analogously, for hypothesis testing between $H_0 : \beta_1 = \cdots = \beta_n = 0$ and H_1 : otherwise, we define the column sum of squares

$$\mathsf{SS_c} = m\sum_{j=1}^{n}(\bar{X}_{\cdot j}-\bar{X})^2.$$

The test statistic is

$$\text{TS} = \frac{\mathsf{SS_c}/(n-1)}{\mathsf{SS_e}/[(n-1)(m-1)]} = \frac{\mathsf{SS_c}(m-1)}{\mathsf{SS_e}},$$

which follows the F-distribution with $n-1$ and $(n-1)(m-1)$ degrees of freedom.

Example. The following data represent the number of different species collected at 6 stations from 1970 to 1977:

1970 :	53	35	31	37	40	43
1971 :	36	34	17	21	30	18
1972 :	47	37	17	31	45	26
1973 :	55	31	17	23	43	37
1974 :	40	32	19	26	45	37
1975 :	52	42	20	27	26	32
1976 :	39	28	21	21	36	28
1977 :	40	32	21	21	36	35

Test if the number of species depends on the location or the year, at the 5 percent level of significance.

It turns out that the p-value is very small in each case, so H_0 is rejected. That is, the number of species probably depends on both the location and the year.

```
m = 8
n = 6
x1 = c(53,35,31,37,40,43)
x2 = c(36,34,17,21,30,18)
x3 = c(47,37,17,31,45,26)
x4 = c(55,31,17,23,43,37)
x5 = c(40,32,19,26,45,37)
x6 = c(52,42,20,27,26,32)
```

```
x7 = c(39,28,21,21,36,28)
x8 = c(40,32,21,21,36,35)
X = t(matrix(c(x1,x2,x3,x4,x5,x6,x7,x8),nrow=n,ncol=m))
SS_r = n*sum((rowMeans(X)-mean(X))^2)
SS_c = m*sum((colMeans(X)-mean(X))^2)
SS_e = sum((X - matrix(rep(rowMeans(X),n),nrow=m) - t(matrix(rep(colMeans(X),m),nrow=n))
            + mean(X))^2)
TS1 = (n-1)*SS_r/SS_e
TS2 = (m-1)*SS_c/SS_e
print(1-pf(TS1,m-1,(n-1)*(m-1)))
```

```
## [1] 0.004066874
```

```
print(1-pf(TS2,n-1,(n-1)*(m-1)))
```

```
## [1] 4.929785e-10
```