Quick introduction to probability and statistics

Nemesh N. T.

1 Probability theory

Probability theory study randomness in rigorous mathematical fashion. There following notions are the key concepts in probability theory:

- Probability space Ω — describes possible outcomes we can expect while studying a random phenomena;
- Probability function — describes how probable a specific random outcome;
- Random variable — a quantity whose value is random;
- Distribution function — a function that describes how probable a specific value is for a given random variable
- Expected value — an average value of a random variable;
- Variance — a quantity describing volatility of a random variable.

1.1 Probability space

We shall discuss all these notions, give definitions and list their properties.

Definition 1.1.1 *Probablity space is a pair* $(\Omega, \mathbb{P})$*, where*

- Ω *is a set of all elementary events*
- $\mathbb{P}$ *is a probablity function* $\mathbb{P} : \Omega \to [0, 1]$

The function $\mathbb{P}$ *must satisfy equations*

- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ *for any* $A, B \subset \Omega$ *such that* $A \cap B = \varnothing$

Remark 1.1.2 *In general*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Any single element $\omega \in \Omega$ is called an elementary event.
Any subset of $A \subset \Omega$ is called an event.

Example 1.1.3 *Tossing a cube once.* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *Elementary events* 1, 2, …, 6. *Event — any subset of* Ω, *e.g.* $\{1, 4, 5\}$.

Example 1.1.4 *Tossing a coin once. Then* $\Omega = \{H, T\}$, $\mathbb{P}(H) = 1/2$, $\mathbb{P}(T) = 1/2$

Example 1.1.5 *Tossing a coin twice. Then* $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$, $\mathbb{P}((H,H)) = \mathbb{P}((H,T)) = \mathbb{P}((T,H)) = \mathbb{P}((T,T)) = 1/4$.

Example 1.1.6 *Throwing a dot on a segment. Let* $\Omega = [0,1]$, $\mathbb{P}(A) = length(A)$. *If* $A = [0.2, 0.4]$, *then* $P(A) = 0.2$. *If* $B = [0.7, 0.8]$, *then* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) = 0.3$

Definition 1.1.7 *Two events* $A, B \subset \Omega$ *are called inpedendent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

This mathematical definition of independent events is consistent with our usual understanding of independent events.

Example 1.1.8 *Tossing a coin twice. Then* $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$, $\mathbb{P}((H,H)) = \mathbb{P}((H,T)) = \mathbb{P}((T,H)) = \mathbb{P}((T,T)) = 1/4$. *Let A denote the event that first toss gave tails, and B denote the event that second toss gave heads. Then*

$$A = \{(T,H),(T,T)\}, \quad B = \{(T,H),(H,H)\}, \quad A \cap B = \{(T,H)\},$$

$$\mathbb{P}(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad \mathbb{P}(A \cap B) = \frac{1}{4}$$

As we see A and B are independent in usual sense and $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

Definition 1.1.9 *Events* $A_1, \ldots, A_n$ *are called independent if*

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_m}) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_m})$$

for any $\{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$

Definition 1.1.10 *Let A and B be two events and* $P(B) \neq 0$, *then the conditional probability of A given B is defined by*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Example 1.1.11 *Conditional probability measures how probable the event A is* <u>*given*</u> *that event B already happened. For example, consider experiment where we toss a coin twice. Then* $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$. *Let A be the event that after two tosses we got two different outcomes, and let B be the event that there was at first toss we got head. In this case*

$$A = \{(H,T),(T,H)\} \qquad B = \{(H,T),(H,H)\}$$

Then

$$\mathbb{P}(A) = \mathbb{P}(\{(H,T),(T,H)\}) = \frac{1}{2} \qquad \mathbb{P}(B) = \mathbb{P}(\{(H,T),(H,H)\}) = \frac{1}{2}$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(H,T)\}) = \frac{1}{4} \qquad \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

Remark 1.1.12 *If A and B are independent events and* $\mathbb{P}(B) \neq 0$, *then*

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

Indeed,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

1.2 Random variables

Definition 1.2.1 *Random varialbe is a function $X : \Omega \to \mathbb{R}$ defined on a probability space.*

Example 1.2.2 *Let A be any event in a probability space Ω. Then the funciton*

$$1_A : \Omega \to \mathbb{R} : \omega \to \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

Example 1.2.3 *Tossing a cube once. $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = \mathbb{P}(4) = \mathbb{P}(5) = \mathbb{P}(6) = 1/6$. Then*

$$X : \Omega \to \mathbb{R} : \omega \to \omega + 10$$

is a random variable.

Example 1.2.4 *Suppose we are tossing a coin. With probability p we get tails and $X = 1$, otherwise $X = 0$. This random variable is called Bernoulli random variable. We write this fact as $X \sim Ber(p)$*

Example 1.2.5 *Suppose we are tossing a coin n times. If X is the numbers of tails after n tosses, with say that X is binomial random variable. We write this fact as $X \sim Bin(n, p)$. Clearly X can be represented as sum of n Bernoulli random variables $X_1, \dots, X_n$, i.e. $X = X_1 + \dots + X_n$. Indeed, just pick $X_i = 1$ if we got tails at i-th toss and $X_i = 0$ if we got head at i-th toss.*

Example 1.2.6 *Let Ω be a unit square (i.e. $\Omega = [0, 1] \times [0, 1]$), let $\mathbb{P}(A) = area(A)$. Then*

$$X : \Omega \to \mathbb{R} : (x, y) \to \sqrt{x^2 + y^2}$$

is a random variable.

Definition 1.2.7 *If $X : \Omega \to \mathbb{R}$ is a random variable and A is some subset of real numbers then*

$$\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$$

In particular, for a real number $a \in \mathbb{R}$ we have

$$\{X = a\} = \{\omega \in \Omega : X(\omega) = a\}$$

Definition 1.2.8 *Two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are called independent if events*

$$\{X = k\}, \quad \{Y = l\}$$

are independent for any $k, l \in \mathbb{R}$.

This definition of independence is consistent with usual understanding of independent random quantities. Checking by definition that two random variables are independent is tedios. Usually it is clear from the context of the problem being solved that two random variables are independent.

Definition 1.2.9 *Random variables $X_1, \dots, X_n$ are called independent if events*

$$\{X_{i_1} = k_1\}, \dots, \{X_{i_m} = k_m\}$$

are independent for any $\{i_1, \dots, i_m\} \subset \{1, \dots, n\}$ and any $k_1, \dots, k_m \in \mathbb{R}$.

1.3 Distributions of random variables

Definition 1.3.1 *Cumulative density function is a function defined by*

$$F_X : \mathbb{R} \to [0,1] : t \mapsto \mathbb{P}(X \leq t)$$

Example 1.3.2 *Tossing a coin twice. Again* $\Omega = \{(i,j), i,j \in \{1,\dots,6\}\}$, $\mathbb{P}((i,j)) = \frac{1}{36}$ *for all* $(i,j) \in \Omega$. *Again consider random variable*

$$X : \Omega \to \mathbb{R} : (i,j) \mapsto i + j$$

Now we shall compute F_X *for all* $t \in \mathbb{R}$.

$$F_X(t) = \begin{cases} 0 & \text{if } t < 2 \\ \mathbb{P}(X=2) & \text{if } 2 \leq t < 3 \\ \mathbb{P}(X=2) + \mathbb{P}(X=3) & \text{if } 3 \leq t < 4 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=4) & \text{if } 4 \leq t < 5 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=5) & \text{if } 5 \leq t < 6 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=6) & \text{if } 6 \leq t < 7 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=7) & \text{if } 7 \leq t < 8 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=8) & \text{if } 8 \leq t < 9 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=9) & \text{if } 9 \leq t < 10 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=10) & \text{if } 10 \leq t < 11 \\ \mathbb{P}(X=2) + \ldots + \mathbb{P}(X=11) & \text{if } 11 \leq t < 12 \\ 1 & \text{if } 12 \leq t \end{cases} = \begin{cases} 0 & \text{if } t < 2 \\ \frac{1}{36} & \text{if } 2 \leq t < 3 \\ \frac{3}{36} & \text{if } 3 \leq t < 4 \\ \frac{6}{36} & \text{if } 4 \leq t < 5 \\ \frac{10}{36} & \text{if } 5 \leq t < 6 \\ \frac{15}{36} & \text{if } 6 \leq t < 7 \\ \frac{21}{36} & \text{if } 7 \leq t < 8 \\ \frac{26}{36} & \text{if } 8 \leq t < 9 \\ \frac{30}{36} & \text{if } 9 \leq t < 10 \\ \frac{33}{36} & \text{if } 10 \leq t < 11 \\ \frac{35}{36} & \text{if } 11 \leq t < 12 \\ 1 & \text{if } 12 \leq t \end{cases}$$

Example 1.3.3 *Throwing dot on a line. Let* $\Omega = [0,1]$, $\mathbb{P}(A) = \text{length}(A)$ *and*

$$X : \Omega \to \mathbb{R} : \omega \mapsto \omega$$

We shall compute F_X *for all* $t \in \mathbb{R}$

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\}) = \mathbb{P}(\{\omega \in [0,1] : \omega \leq t\})$$

$$= \begin{cases} \text{length}(\varnothing) & t < 0 \\ \text{length}([0,t]) & 0 \leq t \leq 1 \\ \text{length}([0,1]) & 1 < t \end{cases} = \begin{cases} 0 & t < 0 \\ t & 0 \leq t \leq 1 \\ 1 & 1 < t \end{cases}$$

From these examples we see that F_X is always an increasing function. If X attains only finitely many values, then F_X has discontinuities, and X is called a discrete random variable. If F_X is continuous, then X is called a continuous random variable.

Definition 1.3.4 *Probability density function is a function defined by*

- $f_X : \mathbb{R} \to [0,1] : t \mapsto \mathbb{P}(X = t)$ *if* X *is discrete;*
- $f_X : \mathbb{R} \to \mathbb{R} : t \mapsto F_X'(t)$ *if* X *is continuous.*

Theorem 1.3.5 *If X is a continuous random variable, then*

$$F_X(t) = \int_{-\infty}^{t} f_X(s)ds$$

Example 1.3.6 *Tossing a cube twice. Define $\Omega = \{(i,j), i,j \in \{1,\dots,6\}\}$, $\mathbb{P}((i,j)) = \frac{1}{36}$ for all $(i,j) \in \Omega$. Consider a random variable*

$$X : \Omega \to \mathbb{R} : (i,j) \mapsto i + j$$

Now we shall compute f_X for reasonable values of X. That is for $t \in \{2,3,\dots,12\}$

$$f_X(2) = \mathbb{P}(X = 2) = \mathbb{P}(\{(1,1)\}) = \frac{1}{36}$$

$$f_X(3) = \mathbb{P}(X = 3) = \mathbb{P}(\{(1,2),(2,1)\}) = \frac{2}{36}$$

$$f_X(4) = \mathbb{P}(X = 4) = \mathbb{P}(\{(1,3),(3,1),(2,2)\}) = \frac{3}{36}$$

$$\dots$$

After careful gazing we get

$$f_X(k) = \begin{cases} \frac{6-|k-7|}{36} & \textit{if } k \in \{2,\dots,12\} \\ 0 & \textit{otherwise} \end{cases}$$

Example 1.3.7 *Throwing dot on a line. Let $\Omega = [0,1]$, $\mathbb{P}(A) = length(A)$ and*

$$X : \Omega \to \mathbb{R} : \omega \mapsto \omega$$

We shall compute f_X for all $t \in \mathbb{R}$

$$f_X(t) = F_X'(t) = \begin{cases} 0' & t < 0 \\ t' & 0 \le t \le 1 \\ 1' & 1 < t \end{cases} = \begin{cases} 0 & t < 0 \\ 1 & 0 \le t \le 1 \\ 0 & 1 < t \end{cases}$$

Example 1.3.8 ***(Uniform distribution)*** *We say that a random variable is uniformly distributed on a segment $[a,b]$ if it has probability density function of the form*

$$f_X(t) = \begin{cases} \frac{1}{b-a} & \textit{if } t \in [a,b] \\ 0 & \textit{if } t \notin [a,b] \end{cases}$$

We express this fact as $X \sim Unif(a,b)$

The following example is central to the whole theory

Definition 1.3.9 ***(Normal distribution)*** *A random variable X is called normal if it has probability density function of the form*

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

In this case we shall write $X \sim Norm(\mu, \sigma^2)$.

Remark 1.3.10 *Consider function*

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

then for any X with normal distribution with parameters μ, σ we have

$$F_X(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

In other words Φ is the cumulative density function of the normal random variable with parameters $\mu = 0$, $\sigma^2 = 1$. Normal random variables with such parameters are called standard normal random variables.

Later we shall discuss the meaning of parameters μ and σ.

Example 1.3.11 ***(Binomial distribution)*** *Let X be a binomial random variable with parameters n and p. In other words X is a number of tails after n tosses of coin such that at each toss tails have probability p. Clearly X attains values $\{0, \ldots, n\}$. We shall compute $f_X(k)$ for $k \in \{0, \ldots, n\}$. By definition $f_X(k) = \mathbb{P}(X = k)$. An event $\{X = k\}$ consist of some number elemtary events $\{\omega_1, \ldots \omega_N\}$. The number N of these elementary events equals the number of ways to pick k tosses out of n tosses that will end up with tails. This is a standard fact from combinatorics, that the latter number is $\binom{n}{k}$. Each elementary event ω_i corresponds to the series of tosses with exactly k tails and $n-k$ heads, so $\mathbb{P}(\{\omega_i\}) = p^k(1-p)^{n-k}$. Thus*

$$\begin{aligned} f_X(k) &= \mathbb{P}(X = k) \\ &= \mathbb{P}(\{\omega_1, \ldots, \omega_N\}) \\ &= \mathbb{P}(\{\omega_1\}) + \ldots + \mathbb{P}(\{\omega_N\}) \\ &= p^k(1-p)^{n-k} + \ldots + p^k(1-p)^{n-k} \\ &= \binom{n}{k} p^k(1-p)^{n-k} \end{aligned}$$

Definition 1.3.12 ***(Poisson distribution)*** *Let X be a discrete random variable with probability density function defined by*

$$f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

then we say that X is Poisson random variable with parameter λ. We denote this fact as $X \sim Pois(\lambda)$.

1.4 Quantiles of random variables

Definition 1.4.1 *An α-quantile of a random variable X is the smallest number q such that*

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha$$

In other words this is the smallest number q such that $F_X(q) \geq \alpha$. Notation: $q_\alpha(X)$.

Put differently $q_\alpha(X)$ is the smallest value q such that at least 100α percent of the values of X are less than q.

Example 1.4.2 *Tossing a coin twice. Again* $\Omega = \{(i,j), i,j \in \{1,\ldots,6\}\}$*,* $\mathbb{P}((i,j)) = \frac{1}{36}$ *for all* $(i,j) \in \Omega$*. Again consider random variable*

$$X : \Omega \to \mathbb{R} : (i,j) \mapsto i + j$$

Then $q_{0.25}(X) = 4$ *because 25 percent of the values are less than 4 and 4 is the smallest possible constant here.*

Notation:

- $q_{0.25}(X)$ — first quartile of X
- $q_{0.50}(X)$ — second quartile of X or median of X
- $q_{0.75}(X)$ — third quartile of X

Question: Find the median value of the total score after tossing two cubes. Find the 0.99 quantile.

Remark 1.4.3 *If* X *is a continuous random variable, then* $q_\alpha(X)$ *is a (necessarily unique) root of the equation* $F_X(q) = \alpha$*.*

Example 1.4.4 *Throwing dot on a line. Let* $\Omega = [0,1]$*,* $\mathbb{P}(A) =$ *length*(A) *and*

$$X : \Omega \to \mathbb{R} : \omega \mapsto \omega$$

As we already know

$$F_X(t) = \begin{cases} 0 & t < 0 \\ t & 0 \le t \le 1 \\ 1 & 1 < t \end{cases}$$

Since X *is a continuous random variable* $q_{0.25}(X)$*, is a root of the equation* $F_X(q) = 0.25$*. Clearly, q=0.25, so* $q_{0.25}(X) = 0.25$*.*

1.5 Expected value of a random variable

Definition 1.5.1 *Expected value of a random variable* X *is a number*

- $\mathbb{E}[X] = \sum_{k \in \mathcal{X}} k\mathbb{P}(X = k)$ *if* X *is discrete (*$\mathcal{X}$ *is the set of values attained by* X*);*
- $\mathbb{E}[X] = \int_{-\infty}^{+\infty} t f_X(t) dt$ *if* X *is continuous;*

Another notation for expected value is m_X*.*

Simply speaking expected value of a random variable is an average value of that variable.

Example 1.5.2 *Tossing a coin twice. Again* $\Omega = \{(i,j), i,j \in \{1,\ldots,6\}\}$*,* $\mathbb{P}((i,j)) = \frac{1}{36}$ *for all* $(i,j) \in \Omega$*. Again consider random variable*

$$X : \Omega \to \mathbb{R} : (i,j) \mapsto i + j$$

Then $\mathcal{X} = \{2, 3, \ldots, 12\}$*. Therefore*

$$\mathbb{E}[X] = 2 \cdot \mathbb{P}(X = 2) + 3\mathbb{P}(X = 3) + \ldots + 12\mathbb{P}(X = 12) = 7$$

Example 1.5.3 *Let X be a Bernoulli random variable with parameter p. Then $\mathcal{X} = \{0, 1\}$, so*

$$\mathbb{E}[X] = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p$$

Example 1.5.4 *Let X be a normal random variable with parameters μ and σ. Then*

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} t f_X(t) dt = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \ldots = \mu$$

Thus parameter μ in the normal distribution is the average value of the random variable.

Theorem 1.5.5 *Suppose we are given two random variables X and Y, then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Even more if a and b — real numbers, then

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[y]$$

Theorem 1.5.6 *If X and Y are two independent random variables then*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Example 1.5.7 *Let X be a binomial random variable with parameters n and p. Then X can be represented as sum of n Bernoulli random variables $X_1, \ldots, X_n$, i.e. $X = X_1 + \ldots + X_n$. From previous example we know that $\mathbb{E}[X_1] = \ldots = \mathbb{E}[X_n] = p$. Therefore*

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = p + \ldots + p = np$$

Theorem 1.5.8 ***(Law of the unconscious statistician, a.k.a. LOTUS)*** *Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ be any funciton, then*

- $\mathbb{E}[g(X)] = \sum_{k \in \mathcal{X}} g(k)\mathbb{P}(X = k)$ *if X is discrete ($\mathcal{X}$ — values attained by X);*
- $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(t) f_X(t) dt$ *if X is continuous.*

Example 1.5.9 *Let X be a bernoulli random variable with parameter p. Then $\mathcal{X} = 0, 1$ and by the LOTUS*

$$\mathbb{E}[X^2] = 0^2 \cdot \mathbb{P}(X = 0) + 1^2 \cdot \mathbb{P}(X = 1) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p$$

1.6 Variance and standard deviation of a random variable

Definition 1.6.1 *Variance of a random variable X is the expected value of the random variable $(X - \mathbb{E}[X])^2$. In other words*

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[x])^2]$$

Remark 1.6.2 *One can show that*

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

If X is a continuous random variable, then

$$\mathbb{V}[X] = \int_{-\infty}^{+\infty} (t - \mathbb{E}[X])^2 f_X(t) dt = \int_{-\infty}^{+\infty} t^2 f_X(t) dt - \left(\int_{-\infty}^{+\infty} t f_X(t) dt\right)^2$$

Variance shows how volatile values of X are. Variance shows how much they are different from the expexted value. In practice it is more convinient to work with another quantity called standard deviation. This characteristic has the advantage that it is measured in the same units as the expected value.

Definition 1.6.3 *The standard deviation of a random variable X is defined as*

$$s_X = \sqrt{\mathbb{V}[X]}$$

Example 1.6.4 *Let X be a Bernoulli random variable with parameter p. As we showed earlier $\mathbb{E}[X] = p$ and $\mathbb{E}[X^2] = p$, so*

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

$$s_X = \sqrt{p(1-p)}$$

Example 1.6.5 *Let X be a normal random variable with parameters μ and σ. As we already know $\mathbb{E}[X] = \mu$, so*

$$\mathbb{V}[X] = \int_{-\infty}^{+\infty} (t - \mathbb{E}[X])^2 f_X(t) dt = \int_{-\infty}^{+\infty} (t-\mu)^2 \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \ldots = \sigma^2$$

$$s_X = \sqrt{\mathbb{V}[X]} = \sigma$$

Thus parameter σ in the normal distribution is the standard deviation of the random variable.

Theorem 1.6.6 *Suppose we are given two* <u>independent</u> *random variables X and Y, then*

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

Even more if a and b are real numbers, then

$$\mathbb{V}[aX+bY] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y]$$

Example 1.6.7 *Let X be a binomial random variable, then X can be represented as a sum of n Bernoulli variables $X = X_1 + \ldots + X_n$, where $X_1, \ldots, X_n \sim Ber(p)$. As we showed earlier $\mathbb{V}[X_1] = \ldots = \mathbb{V}[X_n] = p(1-p)$, so from independece of $X_1, \ldots, X_n$ and previous remark we get*

$$\mathbb{V}[X] = \mathbb{V}[X_1 + \ldots + X_n] = \mathbb{V}[X_1] + \ldots + \mathbb{V}[X_n] = p(1-p) + \ldots + p(1-p) = np(1-p)$$

$$s_X = \sqrt{\mathbb{V}[X]} = \sqrt{np(1-p)}$$

1.7 De Moivre–Laplace theorem

The following thoerem states informally that for big n binomial random variables behave like normal varaibles

Theorem 1.7.1 *Let X be a binomial random variable with parameters n and p where $0 < p < 1$. Then $f_X(k) \approx f_Y(k)$ for the normal random variable Y with parameters $\mu = np$ and $\sigma^2 = np(1-p)$. More explicitly*

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \underset{n\to\infty}{\rightarrow} \frac{1}{\sqrt{2\pi \cdot np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}$$

This theorem has another form (so called integral form).

Theorem 1.7.2 *Let X be a binomial random variable with parameters n and p where $0 < p < 1$. Then $F_X(t) \approx F_Y(t)$ for the normal random variable Y with parameters $\mu = np$ and $\sigma = np(1-p)$. More explicitly*

$$F_X(k) = \mathbb{P}(X \leq t) \underset{n\to\infty}{\to} \int_{-\infty}^{k} \frac{1}{\sqrt{2\pi \cdot np(1-p)}} e^{-\frac{(s-np)^2}{2np(1-p)}} ds = \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$

These theorems assumes that n has to be big enough. How big n shold be in practice?

Remark 1.7.3 *Let X be a binomial random variable with parameters n and p where $pn > 10$ and $n(1-p) > 10$. Then*

$$F_X(k) \approx \Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

1.8 Law of large numbers

The following theorem states informally that average of a big number of equally distributed random variables approaches their expected value. This is one of the foundational theorems of probability theory which gives a firm basis for applications in real world problems.

Definition 1.8.1 *We say that a sequence of random variables $X_1, \ldots, X_n$ is i.i.d. if its random variables are independent and identically distributed. In other words they are independent and have the same cumulative density function.*

Clearly, if $X_1, \ldots, X_n$ are i.i.d., then these random variables have equal characterstics like quantiles, mean, standard deviation and many others.

Theorem 1.8.2 ***(Weak law of large numbers)**. Let $X_1, \ldots, X_n$ be i.i.d. with expected value μ, then for all $\epsilon > 0$ we have*

$$\mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| < \epsilon\right) \underset{n\to\infty}{\to} 1$$

Theorem 1.8.3 ***(Strong law of large numbers)**. Let $X_1, \ldots, X_n$ be i.i.d. with expected value μ, then*

$$\mathbb{P}\left(\frac{X_1 + \ldots + X_n}{n} = \mu\right) \underset{n\to\infty}{\to} 1$$

1.9 Central limit theorem

Cental limit theorem is no doubt the most important theorem of probability theory. Most non-trivial results are based on this fact. It essentially says that the average of i.i.d. random variables behave like a normal random variable. Compare this with the law of large numbers.

Before stating the theorem we shall give a short remark on characteristics of the average of i.i.d. random variables.

Remark 1.9.1 *Let $X_1, \dots, X_n$ be i.i.d. random variables. Let $Y = \frac{1}{n}(X_1 + \ldots + X_n)$ be their average. Since $X_1, \dots, X_n$ identically distributed, then they have the same expected value and varianece*

$$\mathbb{E}[X_1] = \ldots = \mathbb{E}[X_n] = \mu, \quad \mathbb{V}[X_1] = \ldots = \mathbb{V}[X_n] = \sigma^2$$

Then

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \frac{1}{n}\mathbb{E}[X_1 + \ldots + X_n] = \frac{1}{n}(\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]) = \frac{1}{n} \cdot n\mu = \mu$$

$$\mathbb{V}[Y] = \mathbb{V}\left[\frac{1}{n}(X_1 + \ldots + X_n)\right] = \left(\frac{1}{n}\right)^2 \mathbb{V}[X_1 + \ldots + X_n] = \frac{1}{n^2}(\mathbb{V}[X_1] + \ldots + \mathbb{V}[X_n]) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Therefore

$$m_Y = \mu, \quad s_Y = \frac{\sigma}{\sqrt{n}}$$

Theorem 1.9.2 *Let $X_1, \dots, X_n$ be i.i.d. with expected value μ and variance σ^2. Let $Y = \frac{1}{n}(X_1 + \ldots + X_n)$ be their average. Then the random variable $Z_Y = \sqrt{n}\frac{Y-\mu}{\sigma}$ has cumulative density function approximately equal to the cumulative density funciton of the standard normal distribution:*

$$F_{Z_Y}(t) \underset{n\to\infty}{\to} \Phi(t)$$

More explicitly

$$\mathbb{P}\left(\frac{\frac{1}{n}(X_1 + \ldots + X_n) - \mu}{\frac{1}{\sqrt{n}}\sigma} < t\right) \underset{n\to\infty}{\to} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{s^2}{2}} ds$$

Remark 1.9.3 *One can note that De Moivre-Laplace theorem is nothing more than a central limit theorem applied to i.i.d. Bernoulli random variables.*

2 Statistics

In probability theory we study characteristics and behaviour of random variables assuming that we know the probability space or at least the cumulative density functions. In real world it is not possible to get the exact description of a probability space for our problem or a precise formula for the densities of random variables.

The best we can do is to make a good guess about random variable distributions based on some numbers of observations. For example, tossing a coin 10000 times and observing heads in 5053 time we can be quite confident that this coin is described by a Bernoulli random variable with parameter $p = 1/2$. To be absolutely sure about distribution of the coin we had to make infinitely many tossing, which is impossible in practice. Therefore we need to study random variables given only finitely many observations.

The goal of statistics is to give us tools to

1. approximately recover distributions of random variables given finite number of observations;
2. approximately compute random variables characteristics given finite number of observations;
3. for a given level of confidence answer questions regarding random variable behaviour;

4. predict values of dependent random variables given finite number of observations of independent random variables.

These four big problems have their names: distribution estimation, point estimation, hypothesis testing and regression respectively.

2.1 Samples, observations, sample statistics

Definition 2.1.1 *A sample $\mathscr{X}$ of size n is a sequence of random variables $X_1, \ldots, X_n$ defined on some probability space Ω.*

Definition 2.1.2 *Let X be a random variable. A sample $\mathscr{X}$ from X of size n is a sequence $X_1, \ldots, X_n$ of i.i.d. random variables with the same distribution as X. We shall denote this fact as $\mathscr{X} \sim X$.*

Example 2.1.3 *Assume we have a fair coin and toss it n times. Outcomes of the tossed coin are described by a random variable X. Let X_i denote the random variable describing coin side on the i-th toss. Then $\mathscr{X} = (X_1, \ldots, X_n)$ is a sample of the size n of the random variable X, where X is a random variable representing outcome of the flipped coin.*

Definition 2.1.4 *Let $\mathscr{X}$ be a sample of the random variable $X : \Omega \to \mathbb{R}$ defined on a probability space Ω. For any fixed elementary event $\omega \in \Omega$ the sequence of numbers $x = (X_1(\omega), \ldots, X_n(\omega))$ is called on observation of X of size n.*

Example 2.1.5 *Assume we have a fair coin and we toss it 5 times. Outcomes of the tossed coin are described by a random variable X. Suppose we got the following outcomes $x = (H, H, T, T, H)$. Then x is called the observations of the random variable X.*

Definition 2.1.6 *Let $\mathscr{X}$ be a sample of size n. A statistic T is a random variable which is a function of sample $\mathscr{X}$.*

Example 2.1.7 *Let $\mathscr{X}$ be a sample of size n of the random variable X. Then we define the following statistics*

- *sample mean*
$$m(\mathscr{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- *sample k-th moment*
$$m_k(\mathscr{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$
- *sample variance*
$$s_b^2(\mathscr{X}) = \frac{1}{n}\sum_{i=1}^{n} (X_i - m(\mathscr{X}))^2$$
- *unbiased sample variance*
$$s^2(\mathscr{X}) = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - m(\mathscr{X}))^2$$

Remark 2.1.8 *Clearly,*

$$m(\mathscr{X}) = m_1(\mathscr{X}) \qquad s_0^2(\mathscr{X}) = \frac{n}{n-1} s_b^2(\mathscr{X})$$

One can show that

$$s_b^2(\mathscr{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = m_2(\mathscr{X}) - m(\mathscr{X})^2$$

Remark 2.1.9 *Let $\mathscr{X}$ be a sample from the random variable X, then*

$$\mathbb{E}[m_k(\mathscr{X})] = \mathbb{E}[X^k]$$

Indeed,

$$\mathbb{E}[m_k(\mathscr{X})] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^k\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i^k\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^k] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X^k] = \frac{1}{n}\cdot n\mathbb{E}[X^k] = \mathbb{E}[X^k]$$

Remark 2.1.10 *Let $\mathscr{X}$ be a sample from the random variable X, then*

$$\mathbb{E}[m(\mathscr{X})] = m_X, \qquad \mathbb{E}[s_b^2(\mathscr{X})] = \frac{n-1}{n} s_X^2, \qquad \mathbb{E}[s^2(\mathscr{X})] = s_X^2$$

Indeed,

$$\mathbb{E}[m(\mathscr{X})] = \mathbb{E}[m_1(\mathscr{X})] = \mathbb{E}[X^1] = m_X$$

Now note that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \sum_{i\neq j}\mathbb{E}[X_i X_j]\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}[X^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \frac{1}{n^2}\left(n\mathbb{E}[X^2] + \sum_{i\neq j}\mathbb{E}[X]\mathbb{E}[X]\right) \\
&= \frac{1}{n^2}\left(n\mathbb{E}[X^2] + (n^2-n)\mathbb{E}[X]\mathbb{E}[X]\right) \\
&= \frac{1}{n}\left(\mathbb{E}[X^2] + (n-1)\mathbb{E}[X]^2\right)
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}[s_b^2(\mathscr{X})] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \mathbb{E}[m_2(\mathscr{X})] - \frac{1}{n}\left(\mathbb{E}[X^2] + (n-1)\mathbb{E}[X]^2\right) \\
&= \mathbb{E}[X^2] - \frac{1}{n}\left(\mathbb{E}[X^2] + (n-1)\mathbb{E}[X]^2\right) \\
&= \frac{n-1}{n}(\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\
&= \frac{n-1}{n}\mathbb{V}[X] \\
&= \frac{n-1}{n}s_X^2
\end{aligned}$$

and

$$\mathbb{E}[s_0^2(\mathscr{X})] = \mathbb{E}[\frac{n}{n-1}s_b^2(\mathscr{X})] = \frac{n}{n-1}\mathbb{E}[s_b^2(\mathscr{X})] = \frac{n}{n-1}\frac{n-1}{n}s_X^2 = s_X^2$$

2.2 Distribution estimates

As we have seen earlier the most single important characteristic of a random variable is its distribution function. Given a set of observations of a random variable one can construct an approximation of this distribution function. The construction is pretty straightforward.

Definition 2.2.1 *Let $\mathscr{X}$ be a sample from a random variable X. Then the parametric random variable*

$$F_{\mathscr{X}}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i<t\}}$$

is called the empirical cumulative distribution function.

For every t this statistic gives a good approximation for $F_X(t)$.

Theorem 2.2.2 *Let $\mathscr{X}$ be a sample from a random variable X. Then for all $t \in \mathbb{R}$ and $\epsilon > 0$*

$$\mathbb{P}(|F_{\mathscr{X}}(t) - F_X(t)| > \epsilon) \underset{n\to\infty}{\to} 0$$

Even stronger result is true

Theorem 2.2.3 *Let $\mathscr{X}$ be a sample from a random variable X. Then for all $\epsilon > 0$*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |F_{\mathscr{X}}(t) - F_X(t)| > \epsilon\right) \underset{n\to\infty}{\to} 0$$

Example 2.2.4 *Let $x = (1, 2, 4, 2, 4, 1, 3, 1)$ be an observation of a random variable X. Find a function that approximates cumulative distribution function of X. The observation x corresponds to some elementary event ω, that is $(X_1(\omega), \ldots, X_8(\omega)) = (1, 2, 4, 2, 4, 1, 3, 1)$. The desired approximation will be*

$$\begin{aligned}
F_{\mathscr{X}}(t)(\omega) &= \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i(\omega)<t\}} \\
&= \frac{1}{8}\sum_{i=1}^{8} 1_{\{X_i(\omega)<t\}} \\
&= \frac{1}{8}\left(1_{1<t} + 1_{2<t} + 1_{4<t} + 1_{2<t} + 1_{4<t} + 1_{1<t} + 1_{3<t} + 1_{1<t}\right) \\
&= \frac{1}{8}\left(3 \cdot 1_{1<t} + 2 \cdot 1_{2<t} + 1_{3<t} + 2 \cdot 1_{4<t}\right)
\end{aligned}$$

Therefore

$$F_{\mathscr{X}}(t)(\omega) = \begin{cases} 0 & < t \leq 1 \\ \frac{3}{8} & 1 < t \leq 2 \\ \frac{5}{8} & 2 < t \leq 3 \\ \frac{6}{8} & 3 < t \leq 4 \\ 1 & 4 < t \end{cases}$$

2.3 Point estimates

Suppose we study a random variable X. We have n observations $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and we want to know the distribution of X. In practice we usually have a good guess of what type of distribution X should have. It might be normal distribution or a binomial distribution or Poisson distribution. All these classes of distributions are parametric meaning that you need to specify some parameters to pick a concrete distribution of the class. For example you need to know exact values of μ and σ^2 to speak about normal distribution $Norm(\mu, \sigma^2)$. Now given observations $x \in \mathbb{R}^n$ and a class of distributions X belongs to we can hope to find parameters of the specific distribution of X. This is the primary goal of point estimations theory.

Definition 2.3.1 *Let $\mathcal{D}$ be a family of distributions. We say that $\mathcal{D}$ is a parametric family if there is a set of parameters $\Theta \subset \mathbb{R}^k$ such that any distribution $D \in \mathcal{D}$ is uniquely determined by some group of parameters $\theta \in \Theta$. We write this fact as $\mathcal{D} = \{D_\theta : \theta = (\theta_1, \ldots, \theta_k) \in \Theta\}$.*

Example 2.3.2 *Let $\Theta = \mathbb{R}_+$. Then the class of Poisson distributions $\mathcal{P}$ is a parametric family because*

$$\mathcal{P} = \{Pois(\lambda) : \lambda \in \Theta\}$$

Example 2.3.3 *Let $\Theta = \mathbb{R} \times \mathbb{R}_+$. Then the class of normal distributions $\mathcal{N}$ is a parametric family because*

$$\mathcal{N} = \{Norm(\mu, \sigma^2) : (\mu, \sigma^2) \in \Theta\}$$

Example 2.3.4 *Let $\Theta = \mathbb{N} \times [0, 1]$. Then the class of binomial distributions $\mathcal{B}$ is a parametric family*

$$\mathcal{B} = \{Bin(n, p) : (n, p) \in \Theta\}$$

Definition 2.3.5 *Let X be a random variable with distribution from a parametric family $\mathcal{D} = \{D_\theta : \theta = (\theta_1, \dots, \theta_k) \in \Theta\}$. Let $\mathscr{X}$ be a sample of X. A statistic T is called a point estimation of $\theta_i \in \mathbb{R}$ if*

$$\mathbb{E}[T(\mathscr{X})] = \theta_i$$

Example 2.3.6 *Let $X \sim Norm(\mu, \sigma^2)$ be a normal random variable. Let $\mathscr{X}$ be a sample from X. Then $m(\mathscr{X})$ is a point estimation of μ and $s_0(\mathscr{X})$ is a point estimation of σ^2. Indeed, since X is normal, then $m_X = \mu$ and $s_X^2 = \sigma^2$. Now using remark 2.1.10 we get*

$$\mathbb{E}[m(\mathscr{X})] = m_X = \mu, \qquad \mathbb{E}[s_0^2(\mathscr{X})] = s_X = \sigma^2$$

Example 2.3.7 *Let $X \sim Unif(a, b)$ be a random variable uniformly distributed on $[a, b]$. Let $\mathscr{X}$ be a sample from X. Consider statistics $L(\mathscr{X}) = \min(X_1, \dots, X_n)$ and $U(\mathscr{X}) = \max(X_1, \dots, X_n)$ then one can show that*

$$\mathbb{E}[L(\mathscr{X})] = a + \frac{1}{n+1}(b-a), \qquad \mathbb{E}[U(\mathscr{X})] = a + \frac{n}{n+1}(b-a)$$

Therefore

$$a = \frac{n\mathbb{E}[L(\mathscr{X})] - \mathbb{E}[U(\mathscr{X})]}{n-1} = \mathbb{E}\left[\frac{nL(\mathscr{X}) - U(\mathscr{X})}{n-1}\right],$$
$$b = \frac{n\mathbb{E}[U(\mathscr{X})] - \mathbb{E}[L(\mathscr{X})]}{n-1} = \mathbb{E}\left[\frac{nU(\mathscr{X}) - L(\mathscr{X})}{n-1}\right]$$

These equalities show that statistics

$$A(\mathscr{X}) = \frac{nL(\mathscr{X}) - U(\mathscr{X})}{n-1}, \qquad B(\mathscr{X}) = \frac{nU(\mathscr{X}) - L(\mathscr{X})}{n-1}$$

are point estimates for a and b.

Example 2.3.8 *Let $x = (1, 2, 3, 1, 3, 1, 4)$ be observations of the random variable X with uniform distribution on some segment $[a, b]$. Our observation corresponds to some elementary event ω, so $(X_1(\omega), \dots, X_n(\omega)) = (1, 2, 3, 1, 3, 1, 4)$ In our case $n = 7$ and*

$$L(\mathscr{X})(\omega) = \min(1, 2, 3, 1, 3, 1, 4) = 1, \qquad U(\mathscr{X})(\omega) = \max(1, 2, 3, 1, 3, 1, 4) = 4$$

$$A(\mathscr{X})(\omega) = \frac{nL(\mathscr{X})(\omega) - U(\mathscr{X})(\omega)}{n-1} = \frac{1}{2}, \qquad B(\mathscr{X})(\omega) = \frac{nU(\mathscr{X})(\omega) - L(\mathscr{X})(\omega)}{n-1} = \frac{9}{2}$$

Therefore $a \approx 0.5$, $b \approx 4.5$.

2.4 Hypothesis testing

Given observation of a random variable X we can make several guesses about random variable distribution. These guesses are called hypotheses. Our goal is to construct a function that chooses one of the hypothesis given observation of X. Such functions are called a criteria. Since we can inspect only finitely many observations there is always a possibility that our criterion chooses a wrong hypothesis. We want this to happen as rarely as possible.

Definition 2.4.1 *Let $\mathscr{X}$ be a sample. A hypothesis H is any proposition regarding $\mathscr{X}$.*

Example 2.4.2 *Let $\mathscr{X}$ be a sample. The following statements are hypotheses:*

- *all random variables $X_1, \dots, X_n$ in $\mathscr{X}$ are independent;*
- *all random variables $X_1, \dots, X_n$ in $\mathscr{X}$ are i.i.d;*
- *$\mathscr{X}$ is a sample from Bernoulli random variable with $p = 1/2$;*
- *$\mathscr{X}$ is a sample from normal random variable with $\mu \in [0.25, 0.75]$ and $\sigma^2 \in [1, 2]$;*

Definition 2.4.3 *A hypothesis H is called simple if it has the form: $\mathscr{X}$ is a sample from distribution D. Otherwise H is composite.*

Example 2.4.4 *Here are few examples of simple and composite hypotheses:*

- *$\mathscr{X} \sim Norm(1, 2^2)$ — simple hypothesis;*
- *$\mathscr{X} \sim Unif(1, 5)$ — simple hypothesis;*
- *$\mathscr{X} \sim Norm(\mu, 2^2)$ where $\mu \in (-1, 1)$ — composite hypothesis;*
- *all random variables in $\mathscr{X} = (X_1, \dots, X_n)$ are independent — composite hypothesis;*
- *all random variables in $\mathscr{X} = (X_1, \dots, X_n)$ are i.i.d. — composite hypothesis.*

Now we shall formalize the notion of criterion for hypotheses testing.

Definition 2.4.5 *Let $H_1, \dots H_k$ be a set of hypotheses. Then any function of the form*

$$\delta : \mathbb{R}^n \to \{H_1, \dots, H_k\}$$

is called a criterion.

Remark 2.4.6 *In practice we usually consider criteria with two hypotheses $\{H_1, H_2\}$. The hypothesis H_1 is called the null hypothesis, and H_2 is called the alternative.*

It is rarely possible to develop a criterion that does not make mistakes. In order to quantify mistakes that a criterion can make we give the following definition.

Definition 2.4.7 *Let $\delta : \mathbb{R}^n \to \{H_1, \dots, H_k\}$ be a criterion. We say that δ made an error of the i-th kind on the observation $x = (x_1, \dots, x_n)$ if the hypothesis H_i is true but $\delta(x_1, \dots, x_n) \neq H_i$. The probability of the error of the i-th kind is defined by*

$$\alpha_i(\delta) = \mathbb{P}(\delta(X_1, \dots, X_n) \neq H_i | H_i)$$

Example 2.4.8 *Let $\mathscr{X}$ be a sample of size $2n$ from Bernoulli random variable. Consider two hypotheses:*

$$H_1 = \{\mathscr{X} \sim Ber(0.5)\} \qquad H_2 = \{\mathscr{X} \sim Ber(p), p > 0.5\}$$

For example, for these two hypotheses we can define a criterion

$$\delta(x_1, \dots, x_{2n}) = \begin{cases} H_1 & \text{if } \overline{x} = 0.5 \\ H_2 & \text{otherwise} \end{cases}$$

where $\overline{x} = \frac{1}{2n}\sum_{i=1}^{2n} x_i$. Intuitively this criterion we rarely choose the null hypothesis and often make the error of the first kind. We shall compute exact values for the errors of the first and second kind. For the beginning note that $\mathbb{P}(\overline{x} \neq 0.5) = 1$, so

$$\alpha_1(\delta) = \mathbb{P}(\delta(x) \neq H_1|H_1) = \frac{\mathbb{P}(\delta(x) \neq H_1 \cap H_1)}{\mathbb{P}(H_1)} = \frac{\mathbb{P}(\overline{x} \neq 0.5 \cap H_1)}{\mathbb{P}(H_1)} = \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_1)} = 1$$

$$\alpha_2(\delta) = \mathbb{P}(\delta(x) \neq H_2|H_2) = \frac{\mathbb{P}(\delta(x) \neq H_2 \cap H_2)}{\mathbb{P}(H_2)} = \frac{\mathbb{P}(\overline{x} = 0.5 \cap H_2)}{\mathbb{P}(H_2)} = \frac{0}{\mathbb{P}(H_2)} = 0$$

Our criterion almost always makes the error of the first kind and never makes the error of the second kind.

Clearly the problem with this criterion is that it requires exact equality for the average rate of heads in Bernoulli trials. The better criterion would check that $\overline{x}$ falls into some neighbourhood of 0.5. In this case the probability of the error of the first kind would fall dramatically but there would be a room for the errors of the second kind.

Remark 2.4.9 *In practice we usually consider criteria δ with two hypotheses. In this case we denote $\alpha_1(\delta)$ as $\alpha(\delta)$ and $\alpha_2(\delta)$ as $\beta(\delta)$. If the criterion in question is clear from the context, we denote probabilities of the errors as α and β. Even more, the quantity $1 - \beta$ is called the power of the criterion δ.*

Errors of the first and second kind are somewhat opposite to each other. If you minimize α, the β goes bigger and if you try to make β smaller you get larger α.

Example 2.4.10 *Let $\mathscr{X}$ be a sample of size 1 from a normal random variable X. We have two simple hypotheses regarding distribution of X:*

$$H_1 = \{\mathscr{X} \sim Norm(0, 1)\} \qquad H_2 = \{\mathscr{X} \sim Norm(1, 1)\}$$

Consider criterion

$$\delta(x_1) = \begin{cases} H_1 & \textit{if } x_1 \leq b \\ H_2 & \textit{if } x_1 > b \end{cases}$$

Note, that this criterion depends on parameter b. From definition we have

$$\alpha = \mathbb{P}(\delta(x_1) \neq H_1|H_1) = \mathbb{P}(x_1 > b|H_1) \qquad \beta = \mathbb{P}(\delta(x_1) \neq H_2|H_2) = \mathbb{P}(x_1 \leq b|H_2)$$

As b grows bigger we get smaller values of α and larger values β and vice versa.

2.5 Statistical tests

From now on we shall build criteria for some specific but practically important case. Suppose we have a sample $\mathscr{X}$ from a random variable X. We consider only two hypotheses H_1 and $H_2 = \{H_1 \text{ is not true }\}$. For these to hypotheses we shall build criteria of the form

$$\delta(x) = \begin{cases} H_1 & \text{if} \quad |\rho(x)| \leq C \\ H_2 & \text{if} \quad |\rho(x)| > C \end{cases}$$

Our goal is to invent a function ρ such that the criterion δ 'mostly' gives correct answers. By 'mostly' we do not expect to make more than specified percent (say α) of errors of the first kind. To tweak the error rate of our criterion we need to choose C such that $\alpha(\delta) \leq \alpha$.

Definition 2.5.1 *Let $\mathscr{X}$ be a sample of size n from X. Consider two hypotheses H_1 and $H_2 = \{H_1$ is not true $\}$. Let $\rho(\mathscr{X})$ be a statistic (called test statistic) such that*

- *if H_1 is true, then cumulative density function of $\rho(\mathscr{X})$ pointwise converges to cumulative density function of some continuous random variable η*

$$F_{\rho(\mathscr{X})}(t) \underset{n\to\infty}{\to} F_\eta(t) \quad \textit{for all} \quad t \in \mathbb{R}$$

- *if H_1 is not true, then*

$$\mathbb{P}(|\rho(\mathscr{X})| > \epsilon) \underset{n\to\infty}{\to} 1 \quad \textit{for all} \quad \epsilon > 0.$$

Then a statistical test is a criterion of the form

$$\delta(x) = \begin{cases} H_1 & \textit{if} \quad |\rho(x)| \le C \\ H_2 & \textit{if} \quad |\rho(x)| > C \end{cases}$$

where $C > 0$.

Remark 2.5.2 *One needs to clarify requirements to test statistic in the previous definition. The first requirement says that if H_1 holds true, then $\rho(\mathscr{X})$ attains the same values as η. If H_2 holds true, then for n big enough $\rho(\mathscr{X})$ will attain big values.*

In the very definition of the criterion in statistical tests we assume that test statistic has distribution close to some distribution which does not depend on sample being studied. Now we shall discuss typical distributions encountered in statistical tests.

Definition 2.5.3 *Let $X_1, \ldots, X_n \sim Norm(0,1)$. Consider random variable $X = X_1^2 + \ldots + X_n^2$. Its distribution is called the χ^2 distribution with n degrees of freedom. Notation: $X \sim Chi(n)$.*

Definition 2.5.4 *Let $X_1 \sim Norm(0,1)$ and $X_2 \sim \chi^2_n$. Consider random variable $X = \frac{X_1}{\sqrt{X_2/n}}$. Its distribution is called the student distribution with n degrees of freedom. Notation: $X \sim St(n)$.*

Definition 2.5.5 *Let $X_1 \sim Chi(k)$ and $X_2 \sim Chi(n)$. Consider random variable $X = \frac{X_1/k}{X_2/n}$. Its distribution is called the Fisher distribution with parameters k and n. Notation $X \sim F(k,n)$.*

Definition 2.5.6 *We say that a random variable X has Kolmogorov's distribution if its cumulative distribution function is*

$$F_X(t) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2t^2}.$$

Notation $X \sim Kolm$

Now we shall list a few constructions that lead to these distributions

Remark 2.5.7 *Let $\mathscr{X} = (X_1, \ldots, X_n) \sim Norm(\mu, \sigma^2)$. Then*

- $\sqrt{n}\frac{m(\mathscr{X})-\mu}{\sigma} \sim Norm(0,1)$;

- $\frac{(n-1)s^2(\mathscr{X})}{\sigma^2} \sim Chi(n-1)$;
- $\sqrt{n}\frac{m(\mathscr{X})-\mu}{s_0(\mathscr{X})} \sim St(n-1)$.

Now we shall discuss statistical tests used in practice.

Remark 2.5.8 *Suppose we are given a statistical test. It must be used as follows:*

- *Choose level of confidence α (i.e. probability of errors of the first kind);*
- *Find C such that $\mathbb{P}(|\eta| \geq C) = \alpha$;*
- *Given observations x compute $\rho(x)$;*
- *If $|\rho(x)| > C$ we reject H_1, otherwise reject H_2.*

Example 2.5.9 ***(Kolmogorov's test)*** *This criterion tests if a given sample $\mathscr{X}$ of size n was sampled from a continuous random variable X with cumulative distribution function F_X.*

- *hypotheses:*
$$H_1 = \{\mathscr{X} \sim X\}, \qquad H_2 = \{H_1 \text{ is not true}\}$$
- *test statistic:*
$$\rho(\mathscr{X}) = \sqrt{n}\sup_{t\in\mathbb{R}} |F_{\mathscr{X}}(t) - F_X(t)|$$
- *distribution of test statistic if H_1 is true: $\eta \sim Kolm$*

Example 2.5.10 ***(z-test)*** *This criterion checks if a given sample $\mathscr{X}$ of size n sampled from a normal random variable $X \sim Norm(\mu, \sigma^2)$ (with unknown a and known σ) has mean equal to μ_0.*

- *hypotheses:*
$$H_1 = \{\mu = \mu_0\}, \qquad H_2 = \{H_1 \text{ is not true }\}$$
- *test statistic:*
$$\rho(\mathscr{X}) = \sqrt{n}\frac{m(\mathscr{X}) - \mu_0}{\sigma}$$
- *distribution of test statistic if H_1 is true: $\eta \sim St(n-1)$*

Example 2.5.11 ***(t-test)*** *This criterion checks if a given sample $\mathscr{X}$ of size n sampled from a normal random variable $X \sim Norm(\mu, \sigma^2)$ (with unknown a and unknown σ) has mean equal to μ_0.*

- *hypotheses:*
$$H_1 = \{\mu = \mu_0\}, \qquad H_2 = \{H_1 \text{ is not true }\}$$
- *test statistic:*
$$\rho(\mathscr{X}) = \sqrt{n}\frac{m(\mathscr{X}) - \mu_0}{s_0(\mathscr{X})}$$
- *distribution of test statistic if H_1 is true: $\eta \sim Norm(0,1)$*

Example 2.5.12 ***(Pirson's test)*** *This criterion tests if a given sample $\mathscr{X}$ of size n was sampled from a continuous random variable X satisfying certain restrictions on its distribution. Let $A_1, \dots, A_k$ be a sequence of disjoint segments whose union contains all possible values of X. Let $p_1, \dots, p_k$ be the expected probabilities that X fall into segments $A_1, \dots, A_k$ respectively. Clearly $p_1, \dots, p_k$ must sum up to 1.*

- *hypotheses:*
$$H_1 = \{\mathbb{P}(X_1 \in A_i) = p_i \textit{ for all } i \in \{1, \dots, k\}\}, \qquad H_2 = \{H_1 \textit{ is not true}\}$$
- *test statistic:*
$$\rho(\mathscr{X}) = \sum_{i=1}^{k} \frac{(\nu_i(\mathscr{X}) - np_i)^2}{np_i}$$
where
$$\nu_i(\mathscr{X}) = \sum_{j=1}^{n} 1_{X_j \in A_i} \quad \textit{number of } X_i \textit{ that fall into } A_i$$
- *distribution of test statistic if H_1 is true: $\eta \sim Chi(k-1)$*

In practice we do not explicitly specify probabilities $p_1, \dots, p_k$, but compute them from distribution of some random variable X using formulae

$$p_i = \mathbb{P}(X \in A_i)$$

This approach gives rise to the false belief that Pirson's test check that a sample $\mathscr{X}$ was sampled from random variable X. This would be true if the number of segments $A_1, \dots, A_k$ grew to infinity while their sizes would uniformly approach zero.

Example 2.5.13 ***(Two-sample Kolmogorov's test)*** *This criterion tests if a given a sample $\mathscr{X}$ of size n from random variable X and a sample $\mathscr{Y}$ of size m from random variable Y have the same distribution.*

- *hypotheses:*
$$H_1 = \{F_X = F_Y\}, \qquad H_2 = \{H_1 \textit{ is not true }\}$$
- *test statistic:*
$$\rho(\mathscr{X}, \mathscr{Y}) = \sqrt{\frac{mn}{m+n}} \sup_{t \in \mathbb{R}} |F_{\mathscr{X}}(t) - F_{\mathscr{Y}}(t)|$$
- *distribution of test statistic if H_1 is true: $\eta \sim Kolm$*

Example 2.5.14 ***(Two-sample Fisher's test)*** *This criterion tests if a given sample $\mathscr{X}$ of size n from normal random variable $X \sim Norm(\mu_X, \sigma_X^2)$ and a sample $\mathscr{Y}$ of size m from random variable $Y \sim Norm(\mu_Y, \sigma_Y^2)$ have the same standard deviation.*

- *hypotheses:*
$$H_1 = \{\sigma_X = \sigma_Y\}, \qquad H_2 = \{H_1 \textit{ is not true}\}$$
- *test statistic:*
$$\rho(\mathscr{X}, \mathscr{Y}) = \frac{s^2(\mathscr{X})}{s^2(\mathscr{Y})}$$

- *distribution of test statistic if H_1 is true: $\eta \sim F(n-1, m-1)$*

Example 2.5.15 ***(Two-sample Student's test)*** *This criterion tests if a given sample $\mathscr{X}$ of size n from normal random variable $X \sim Norm(\mu_X, \sigma^2)$ and a sample $\mathscr{Y}$ of size m from random variable $Y \sim Norm(\mu_Y, \sigma^2)$ have the same standard deviation.*

- *hypotheses:*

$$H_1 = \{\mu_X = \mu_Y\}, \qquad H_2 = \{H_1 \textit{ is not true } \}$$

- *test statistic:*

$$\rho(\mathscr{X}, \mathscr{Y}) = \sqrt{\frac{mn(n+m-2)}{m+n}} \frac{m(\mathscr{X}) - m(\mathscr{Y})}{\sqrt{(n-1)s^2(\mathscr{X}) + (m-1)s^2(\mathscr{Y})}}$$

- *distribution of test statistic if H_1 is true: $\eta \sim St(n-1, m-1)$*

Example 2.5.16 ***(Pirson's test)*** *This criterion tests if a given sample $\mathscr{X}$ of size n of a random variable X and a sample $\mathscr{Y}$ of size n from a random variable Y has the property that X and Y are independent. Let $A_1, \dots, A_k$ be a sequence of disjoint segments whose union contains all possible values of X. Analogously, let $B_1, \dots, B_l$ be a sequence of disjoint segments whose union contains all possible values of Y. Let $p_{i,j}$ be the expected probability that $X \in A_i$ and $Y \in B_j$. Clearly all $p_{i,j}$ must sum up to 1.*

- *hypotheses:*

$$H_1 = \{X \textit{ and } Y \textit{ are independent}\}, \qquad H_2 = \{H_1 \textit{ is not true } \}$$

- *test statistic:*

$$\rho(\mathscr{X}, \mathscr{Y}) = n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(\nu_{i,j} - \frac{1}{n}\nu_{i,*}\nu_{*,j}\right)^2}{\nu_{i,*}\nu_{*,j}}$$

where

$$\nu_{i,j} = \sum_{s=1}^{n} 1_{X_s \in A_i} 1_{Y_s \in B_j} \quad \textit{number of pairs } (X_s, Y_s) \textit{ that fall into } (A_i, B_j)$$

$$\nu_{i,*} = \sum_{j=1}^{l} \nu_{i,j} \quad \textit{number of } X_s \textit{ that fall into } A_i$$

$$\nu_{*,j} = \sum_{i=1}^{k} \nu_{i,j} \quad \textit{number of } Y_s \textit{ that fall into } B_j$$

- *distribution of test statistic if H_1 is true: $\eta \sim Chi((k-1) \cdot (l-1))$*

Example 2.5.17 ***(Bartlett test)*** *This criterion tests if samples from different normal variables have the same variance. Let $\mathscr{X}_1, \dots, \mathscr{X}_k$ be samples of sizes $n_1, \dots, n_k$ respectively. Suppose $\mathscr{X}_i \sim Norm(\mu_i, \sigma_i^2)$ and $n_i > 3$.*

- *hypotheses:*

$$H_1 = \{\sigma_i = \sigma_j \textit{ for alli}, j \in \{1, \dots, k\}\}, \qquad H_2 = \{H_1 \textit{ is not true}\}$$

- *test statistic:*

$$\rho(\mathscr{X}_1,\ldots,\mathscr{X}_k) = \frac{(N-k)\ln(s_p^2(\mathscr{X}_1,\ldots,\mathscr{X}_k)) - \sum_{i=1}^k (n_i-1)\ln(s^2(\mathscr{X}_i))}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^k \frac{1}{n_i-1} - \frac{1}{N-k}\right)}$$

  where

$$N = \sum_{i=1}^k n_i, \qquad s_p^2(\mathscr{X}_1,\ldots,\mathscr{X}_k) = \frac{1}{N-k}\sum_{i=1}^k (n_i-1)s^2(\mathscr{X}_i)$$

- *distribution of test statistic if H_1 is true: $\eta \sim Chi(k-1)$*

Example 2.5.18 ***(ANOVA test)*** *This criterion tests if samples from different normal variables have the same mean. Let $\mathscr{X}_1,\ldots,\mathscr{X}_k$ be samples of sizes $n_1,\ldots,n_k$ respectively. Suppose $\mathscr{X}_i \sim Norm(\mu_i,\sigma^2)$. Note: all samples must be sampled from normal variables with equal variances.*

- *hypotheses:*

$$H_1 = \{\mu_i = \mu_j \text{ for alli}, j \in \{1,\ldots,k\}\}, \qquad H_2 = \{H_1 \text{ is not true}\}$$

- *test statistic:*

$$\rho(\mathscr{X}_1,\ldots,\mathscr{X}_k) = \frac{s_i^2(\mathscr{X}_1,\ldots,\mathscr{X}_k)}{s_p^2(\mathscr{X}_1,\ldots,\mathscr{X}_k)}$$

  where

$$N = \sum_{i=1}^k n_i,$$

$$M(\mathscr{X}_1,\ldots,\mathscr{X}_k) = \frac{1}{N}\sum_{i=1}^k n_i m(\mathscr{X}_i) \quad \textit{overall mean}$$

$$s_i^2(\mathscr{X}_1,\ldots,\mathscr{X}_k) = \frac{1}{k-1}\sum_{i=1}^k n_i(m(\mathscr{X}_i) - M(\mathscr{X}_1,\ldots,\mathscr{X}_k))^2 \quad \textit{inter sample variance}$$

$$s_p^2(\mathscr{X}_1,\ldots,\mathscr{X}_k) = \frac{1}{N-k}\sum_{i=1}^k (n_i-1)s^2(\mathscr{X}_i) \quad \textit{pooled variance}$$

- *distribution of test statistic if H_1 is true: $\eta \sim F(k-1, n-k)$*