Lecture 6

Chapter 4. Numerical Characteristics of Random Variables

- the distribution of a random variable or a random vector, the full collection of related probabilities, contains the entire information about its behavior;
- this detailed information can be summarized in a few vital numerical characteristics describing the *average value*, *the most likely value of a random variable*, its *spread*, *variability*, etc;
- these are numbers that will provide some information about a random variable or about the relationship between random variables.

1. Expectation

Definition 1.1.

(i) If $X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ is a discrete random variable, then the **expectation (expected value, mean value)** of X is the real number

$$E(X) = \sum_{i \in I} x_i P(X = x_i) = \sum_{i \in I} x_i p_i, \tag{1.1}$$

if it exists (i.e., the series is absolutely convergent).

(ii) If X is a continuous random variable with density function $f : \mathbb{R} \to \mathbb{R}$, then its **expectation (expected value, mean value)** is the real number

$$E(X) = \int_{\mathbb{R}} x f(x) dx, \tag{1.2}$$

if it exists (i.e., the integral is absolutely convergent).

Remark 1.2.

1. The expected value can be thought of as a “long term” average value, a number that we *expect* the values of a random variable to stabilize on.

2. If $h : \mathbb{R} \to \mathbb{R}$ is a measurable function, then

$$E\big(h(X)\big) = \sum_{i \in I} h(x_i) p_i, \tag{1.3}$$

if X is discrete and

$$E\big(h(X)\big) = \int_{\mathbb{R}} h(x) f(x)\, dx, \tag{1.4}$$

if X is continuous.

It can also be interpreted as a point of equilibrium, a center of gravity.

In the discrete case, if we imagine the probabilities p_i to be weights distributed in the points x_i, then $E(X)$ would be the point that holds the whole thing in *equilibrium*. In fact, notice that the computational formula (1.1) is *actually* a weighted mean.

Consider a random variable with pdf

$$X \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}.$$

Observing this variable many times, we shall see $X = 0$ about 50% of times and $X = 1$ about 50% of times. The average value of X will then be close to 0.5, so it is reasonable to have $E(X) = 0.5$, which is what we get by (1.1).

Now, suppose that $P(X=0)=0.75$ and $P(X=1)=0.25$, i.e its pdf is now

$$X\begin{pmatrix} 0 & 1 \\ 0.75 & 0.25 \end{pmatrix}.$$

Then, in a long run, X is equal to 1 only $1/4$ of times, otherwise it equals 0. Therefore, in this case, $E(X)=0.25$.

The same interpretation would go for the continuous case, only there the "weight" would be continuously distributed, according to the density function f.

The expected value as a center of gravity is illustrated in Figure 1.

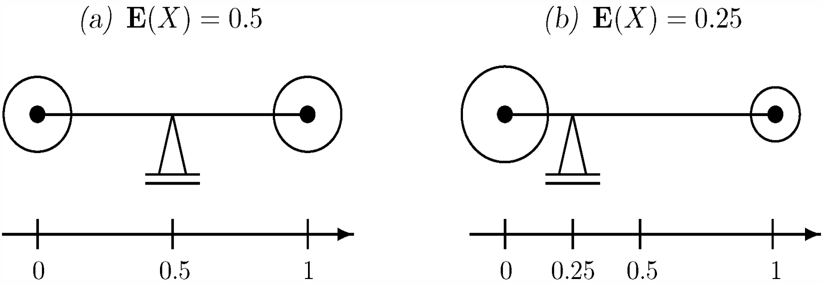

Figure 1: Expectation as a center of gravity

Example 1.3.

Let us start with a simple, intuitive example. Let X be the random variable that denotes the number shown when a die is rolled. What would be the "expected average value" of X, if the die was rolled over and over?

Solution.

Since any of the 6 numbers is equally probable to show on the die, we would expect that, in the long run, we would roll as many 1's as 6's. These would average out at

$$\frac{1+6}{2} = \frac{7}{2}.$$

Also, we would expect to roll the same number of 2's as 5's, which would also average at

$$\frac{2+5}{2} = \frac{7}{2}.$$

Finally, about the same number of 3's and 4's would be expected to show and their average is again, $\frac{7}{2}$. So, the "long term average" should be, intuitively, $\frac{7}{2}$.

On the other hand, we know that X has a Discrete Uniform $U(6)$ distribution, with pdf

$$X\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

Then, by (1.1),

$$\begin{aligned} E(X) &= \sum_{i\in I} x_i p_i \\ &= \sum_{i=1}^{6} i\cdot\frac{1}{6} = \frac{1}{6}\sum_{i=1}^{6} i \\ &= \frac{1}{6}\cdot\frac{6\cdot 7}{2} = \frac{7}{2}, \end{aligned}$$

the value we obtained intuitively.

■

Example 1.4.

Consider now a (continuous) Uniform variable $X \in U(a, b)$. That means X can take *any* value in the interval $[a, b]$, equally probable (recall Problem 3 in Seminar 2, about a spyware breaking passwords). What would be a long-run "expected average value"?

Solution.
In the long run, the variable is just as likely to take values at the beginning of the interval, as it is to take the ones towards the end of $[a, b]$. So they would average out at the value right in the middle, i.e. the midpoint of the interval,

$$\frac{a+b}{2}.$$

Indeed, since the pdf of X is

$$f(x) = \frac{1}{b-a}, \; x \in [a,b]$$

(and 0 everywhere else), by (1.2), its expected value is

$$\begin{aligned} E(X) &= \int_{\mathbb{R}} xf(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b xdx \\ &= \frac{1}{b-a} \cdot \frac{1}{2}x^2\Big|_a^b = \frac{1}{b-a} \cdot \frac{b^2-a^2}{2} = \frac{a+b}{2}. \end{aligned}$$

$\blacksquare$

Example 1.5.

The expected value of a $Bern(p), p \in (0,1)$ variable with pdf

$$X \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

is

$$E(X) \quad = \quad 0 \cdot (1-p) + 1 \cdot p \quad = \quad p. \tag{1.5}$$

Theorem 1.6.

If X and Y are either both discrete or both continuous random variables, then the following properties hold:

a) $E(aX + b) = aE(X) + b$, for all $a, b \in \mathbb{R}$.

b) $E(X + Y) = E(X) + E(Y)$.

c) If X and Y are independent, then $E(X \cdot Y) = E(X)E(Y)$.

d) If $X \leq Y$, i.e. $X(e) \leq Y(e)$, for all $e \in S$, then $E(X) \leq E(Y)$.

Proof.

We give a selected proof, only for the discrete case.

a) If X is discrete, with pdf

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I},$$

then $Y = aX + b$ is also discrete and has pdf

$$Y \begin{pmatrix} ax_i + b \\ p_i \end{pmatrix}_{i \in I}.$$

So, its expectation is

$$E(aX + b) = \sum_{i \in I} (ax_i + b) p_i = a \sum_{i \in I} x_i p_i + b \sum_{i \in I} p_i = aE(X) + b.$$

Proof.

b) For X and Y both discrete, recall that their sum has pdf

$$X + Y \begin{pmatrix} x_i + y_j \\ p_{ij} \end{pmatrix}_{(i,j) \in I \times J}, \quad p_{ij} = P(X = x_i, Y = y_j)$$

and that

$$\sum_{j \in J} p_{ij} = p_i, \ \sum_{i \in I} p_{ij} = q_j,$$

where $p_i = P(X = x_i), \ i \in I$ and $q_j = P(Y = y_j), \ j \in J$.

Proof.

Then

$$
\begin{aligned}
E(X+Y) &= \sum_{(i,j)\in I\times J} (x_i+y_j)p_{ij} = \sum_{i\in I}\sum_{j\in J}(x_i+y_j)p_{ij} \\
&= \sum_{i\in I}\sum_{j\in J} x_i p_{ij} + \sum_{j\in J}\sum_{i\in I} y_j p_{ij} \\
&= \sum_{i\in I} x_i \underbrace{\sum_{j\in J} p_{ij}}_{p_i} + \sum_{j\in J} y_j \underbrace{\sum_{i\in I} p_{ij}}_{q_j} \\
&= \sum_{i\in I} x_i p_i + \sum_{j\in J} y_j q_j \\
&= E(X) + E(Y).
\end{aligned}
$$

Proof.

c) For X and Y discrete and independent, we have

$$\begin{aligned}
E(XY) &= \sum_{i\in I}\sum_{j\in J} x_i y_j p_{ij} \overset{\text{ind}}{=} \sum_{i\in I}\sum_{j\in J} x_i y_j p_i q_j \\
&= \sum_{i\in I} x_i \underbrace{\left(\sum_{j\in J} y_j q_j\right)}_{E(Y)} p_i \\
&= E(Y)\cdot\sum_{i\in I} x_i p_i \\
&= E(X)\cdot E(Y).
\end{aligned}$$

Proof.

d) We show that if $Z \geq 0$, then $E(Z) \geq 0$.

Then by a) and b) applied to $Z = Y - X$, the property follows.

If Z is discrete, $Z \geq 0$ means its values $z_i \geq 0, \ \forall i \in I$ and then

$$E(Z) = \sum_{i \in I} z_i P(Z = z_i) \geq 0.$$

□

Remark 1.7.

1. Property b) in Theorem 1.6 can be generalized to

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i).$$

2. Property c) in Theorem 1.6 can also be generalized: If $X_1, \ldots, X_n$ are independent, then

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).$$

3. An immediate consequence of Theorem 1.6a) is the fact that

$$E\big(X - E(X)\big) = 0.$$

Example 1.8.

Let us find the expectation of a Binomial variable $X \in B(n,p), n \in \mathbb{N}$, $p \in (0,1)$.

Solution. Recall (Remark 4.8, Lecture 4) that a Binomial variable $X \in B(n,p)$ is the sum of n independent $X_i \in Bern(p)$ random variables. All variables X_i have the same expected value $E(X_i) = p$, since they have the same distribution.
Then, by the previous theorem,

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np.$$

■

Remark 1.9.

For a Normal variable $X \in N(\mu, \sigma)$, the expected value is $E(X) = \mu$.

2. Variance and Standard Deviation

Expectation shows where the average value of a random variable is located, or where the variable is expected to be, plus or minus some error. How large could this "error" be, and how much can a variable vary around its expectation? The answer to these questions can give important information about a random variable.

Knowledge of the mean value of a random variable is important, but that knowledge *alone* can be misleading. Suppose two patients in a hospital, X and Y, have their pulse (number of heartbeats per minute) checked every day. Over the course of time, they each have a mean pulse of 75, which is considered healthy. But, for patient X the pulse ranges between 70 and 80, while for patient Y, it oscillates between 40 and 110. Obviously, the second patient might have some serious health problems, which the *expected value alone* would not show.

So, next, we define some measures of variability.

Definition 2.1.

Let X be a random variable. The **variance (dispersion)** of X is the number

$$V(X) \quad = \quad E\left[\left(X - E(X)\right)^2\right], \tag{2.1}$$

if it exists. The value $\sigma(X) = \mathrm{Std}(X) = \sqrt{V(X)}$ is called the **standard deviation** of X.

Variance (and standard deviation) measure the amount of variability (spread) in the values that a random variable takes, with large values indicating a wide spread of values and small values meaning more closely knit values.

The standard deviation brings the numbers to the same "level" (e.g., measurement units), while the variance gives the squares of those numbers.

Theorem 2.2.

Let X and Y be random variables. Then the following properties hold:

a) $V(X) = E(X^2) - \big(E(X)\big)^2$.

b) $V(aX + b) = a^2 V(X)$, for all $a, b \in \mathbb{R}$.

c) If X and Y are independent, then

$$V(X + Y) = V(X) + V(Y).$$

d) If X and Y are independent, then

$$V(X \cdot Y) \quad = \quad E(X^2)E(Y^2) - \big(E(X)\big)^2\big(E(Y)\big)^2.$$

Proof.

We give a selected proof.

a) By properties of expectation in Theorem 1.6, we have

$$\begin{aligned} V(X) &= E\left[X^2 - 2E(X)X + \big(E(X)\big)^2\right] \\ &= E(X^2) - 2E(X)E(X) + \big(E(X)\big)^2 \\ &= E(X^2) - \big(E(X)\big)^2. \end{aligned}$$

b)

$$\begin{aligned} V(aX+b) &= E\left[\big(aX+b-E(aX+b)\big)^2\right] \\ &= E\left[\big(aX+b-aE(X)-b\big)^2\right] \\ &= a^2E\left[\big(X-E(X)\big)^2\right] = a^2V(X). \end{aligned}$$

Proof.

c) If X, Y are independent, then so are $X - E(X)$, $Y - E(Y)$, thus,

$$\begin{aligned} V(X+Y) &= E\left[\left(X+Y-E(X+Y)\right)^2\right] \\ &= E\left[\left((X-E(X))+(Y-E(Y))\right)^2\right] \\ &= E\left[\left(X-E(X)\right)^2\right] + 2E\left[\left(X-E(X)\right)\left(Y-E(Y)\right)\right] \\ &+ E\left[\left(Y-E(Y)\right)^2\right] \\ &\overset{\text{ind}}{=} V(X) + 2E\left(X-E(X)\right)\cdot E\left(Y-E(Y)\right) + V(Y) \\ &= V(X)+V(Y), \end{aligned}$$

since $E\left(X - E(X)\right) = 0$.

□

Remark 2.3.

1. Part a) of Theorem 2.2 provides a more practical computational formula for the variance than the definition.

Thus, if $X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ is discrete, then

$$V(X) = \sum_{i \in I} x_i^2 p_i - \Big(\sum_{i \in I} x_i p_i \Big)^2$$

and if X is continuous with density function f, then

$$V(X) = \int_{\mathbb{R}} x^2 f(x)\, dx - \Big(\int_{\mathbb{R}} x f(x)\, dx \Big)^2.$$

Remark 2.3.

2. A direct consequence of Theorem 2.2a) (since $V(X) \geq 0$) is the following inequality:

$$|E(X)| \leq \sqrt{E(X^2)},$$

which will be discussed later on in this chapter.

3. If $X = b$ is a constant random variable (i.e. it only takes that one value with probability 1), then by Theorem 2.2a), $V(X) = 0$, which is to be expected (the variable X does not vary *at all*).

Remark 2.3.

4. Part c) of Theorem 2.2 can be generalized to any number of random variables: If $X_1, \dots, X_n$ are independent, then

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i).$$

5. A consequence of parts b) and c) of Theorem 2.2 is the following property: If X and Y are independent, then

$$V(X+Y) = V(X) + V(Y) = V(X) + V(-Y) = V(X-Y).$$

Example 2.4.

Find the variance of a random variable X having
a) a Bernoulli $Bern(p)$ distribution;
b) a Binomial $B(n,p)$ distribution.

Solution.

a) We have

$$X\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \quad X^2\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

so both $E(X) = E(X^2) = p$ and thus,

$$V(X) = p - p^2 = pq.$$

b) If X is Binomial, again we use the fact that it can be written as

$$X = \sum_{i=1}^{n} X_i,$$

where $X_1, \ldots, X_n$ are independent and identically distributed with a $Bern(p)$ distribution. Then by part a), $V(X_i) = pq$, for each $i = \overline{1, n}$ and by the previous remarks,

$$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = npq.$$

■

Remark 2.5.

For a Normal variable $X \in N(\mu, \sigma)$, the variance is $V(X) = \sigma^2$ and its standard deviation is $\sigma(X) = \mathrm{Std}(X) = \sigma$. So, the parameters of a Normal variable $X \in N(\mu, \sigma)$ are its mean value and its standard deviation.

3. Median

Definition 3.1.

The **median** of a random variable X with cdf $F : \mathbb{R} \to \mathbb{R}$ is a real number M that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5. That is, M is such that

$$
\begin{aligned}
P(X > M) &\leq 1/2, \text{ i.e. } 1 - F(M) \leq 1/2, \\
P(X < M) &\leq 1/2, \text{ i.e. } F(M - 0) \leq 1/2.
\end{aligned}
$$

Comparing the mean $E(X)$ and the median M, one can tell whether the distribution of X is

- right-skewed ($M < E(X)$),
- left-skewed ($M > E(X)$), or
- symmetric ($M = E(X)$).

For *continuous* distributions, since $P(X < M) = P(X \leq M) = F(M)$ $= F(M-0)$, computing a population median reduces to solving one equation:

$$\begin{cases} P(X > M) & = & 1 - F(M) & \leq & 1/2 \\ P(X < M) & = & F(M) & \leq & 1/2 \end{cases} \Rightarrow F(M) = 1/2.$$

The Uniform distribution $U(a,b)$ has cdf $F(x) = \dfrac{x-a}{b-a},\ x \in [a,b]$. Solving the equation $F(M) = (M-a)/(b-a) = 1/2$, we find the median

$$M = \frac{a+b}{2},$$

which is also the expected value $E(X)$. That should be no surprise, knowing that the Uniform distribution is symmetric (see Figure 2(a)).
For the Exponential distribution $\text{Exp}(\lambda)$, the cdf is $F(x) = 1 - e^{-\lambda x},\ x > 0$.
Solving $F(M) = 1 - e^{-\lambda M} = 1/2$, we get

$$M = \frac{\ln 2}{\lambda} \approx \frac{0.6931}{\lambda} < \frac{1}{\lambda} = E(X),$$

since the Exponential distribution is right-skewed (see Figure 2(b)).

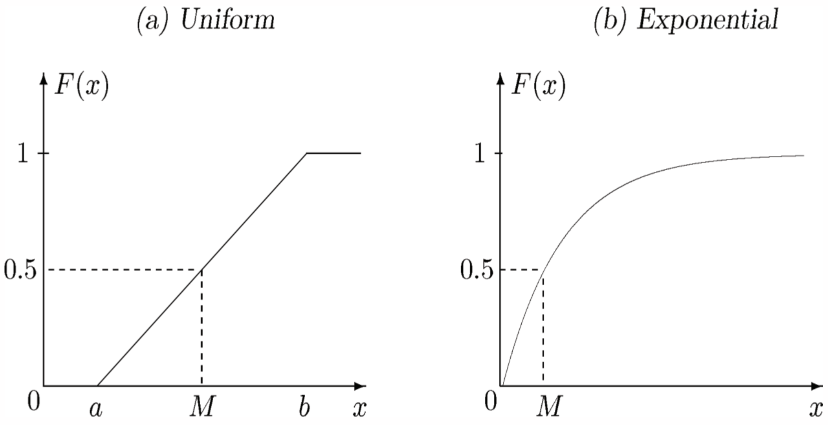

Figure 2: Median for Continuous Distributions

For *discrete* distributions, the equation $F(x) = 0.5$ has either a whole interval of roots or no roots at all (see Figure 3).

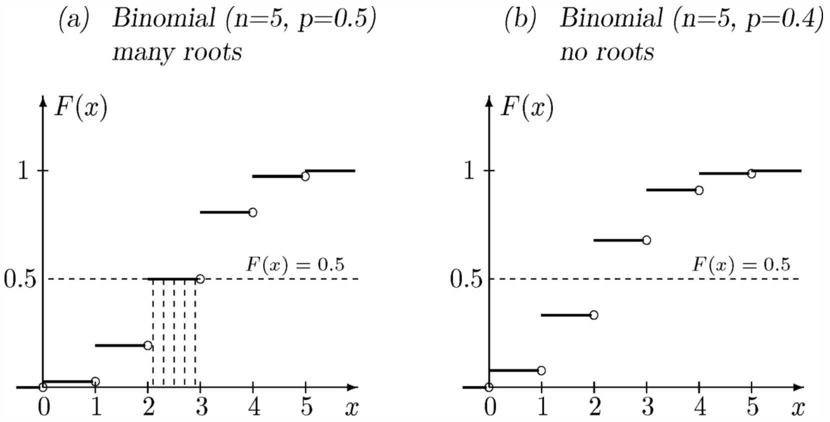

Figure 3: Median for Discrete Distributions

In the first case, the Binomial distribution $B(5, 0.5)$, with $p = 0.5$, successes and failures are equally likely. Pick, for example, $x = 2.2$ in the interval $(2, 3)$. Having fewer than 2.2 successes (i.e., at most 2) has the same chance as having more than 2.2 successes (i.e., at least 3 successes). Therefore, $X < 2.2$ with the same probability as $X > 2.2$, which makes $x = 2.2$ a central value, a median. The same is true for *any other* value in $(2, 3)$. In this case, for any number $M \in (2, 3)$, we have $F(M) = F(M - 0) = 0.5$, so any such number is a median.

In the other case, the Binomial distribution $B(5, 0.4)$ with $p = 0.4$, we have $F(1) = 0.2333$ and $F(2) = 0.5443$, so

$$\begin{array}{lll} F(x) < 0.5 & \text{for} & x < 2, \\ F(x) > 0.5 & \text{for} & x \geq 2, \end{array}$$

but there is no value of x with $F(x) = 0.5$. Then, $M = 2$ is the median. Seeing a value on either side of $M = 2$ has probability less than 0.5, which makes $M = 2$ a center value. Here, $F(M) > 0.5$ and $F(M - 0) < 0.5$.

4. Moments

The ideas of expected value and variance can be generalized.

Definition 4.1.

Let X be a random variable and let $k \in \mathbb{N}$.
The **(initial) moment of order k** of X is (if it exists) the number

$$\nu_k = E(X^k). \tag{4.1}$$

The **absolute moment of order k** of X is (if it exists) the number

$$\underline{\nu}_k = E(|X|^k). \tag{4.2}$$

The **central (centered) moment of order k** of X is (if it exists) the number

$$\mu_k = E\Big[\big(X - E(X)\big)^k\Big]. \tag{4.3}$$

Remark 4.2.

1. If X is a discrete random variable with pdf $\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, then for every $k \in \mathbb{N}$,

$$\nu_k = \sum_{i\in I} x_i^k p_i, \quad \underline{\nu}_k = \sum_{i\in I} |x_i|^k p_i, \quad \mu_k = \sum_{i\in I} \big(x_i - E(X)\big)^k p_i.$$

If X is a continuous random variable with pdf f, then for every $k \in \mathbb{N}$,

$$\nu_k = \int_{\mathbb{R}} x^k f(x)\,dx, \quad \underline{\nu}_k = \int_{\mathbb{R}} |x|^k f(x)\,dx, \quad \mu_k = \int_{\mathbb{R}} \Big(x - E(X)\Big)^k f(x)\,dx.$$

Remark 4.2.

2. The expectation of a random variable X is the moment of order 1,

$$E(X) = \nu_1.$$

The variance of a random variable X is the central moment of order 2,

$$V(X) = \mu_2 = \nu_2 - \nu_1^2.$$

For any random variable X, the central moment of order 1 is 0,

$$\mu_1 = E\big(X - E(X)\big) = E(X) - E(X) = 0.$$

3. An important property of the moments of a random variable X, which we just state, without proof, is the following: If $\underline{\nu}_n = E(|X|^n)$ exists for some $n \in \mathbb{N}$, then $\nu_k,\ \underline{\nu}_k$ and μ_k also exist, for all $k = \overline{1, n}$.