

Lecture 7

5. Quantiles

Quantiles generalize the idea of median, where the number $1/2$ is replaced by **any** probability.

Definition 5.1.

Let X be a random variable with cumulative distribution function $F : \mathbb{R} \rightarrow \mathbb{R}$ and let $\alpha \in (0, 1)$. A **quantile of order α** is a number q_α satisfying the conditions

$$\begin{aligned} P(X < q_\alpha) &\leq \alpha \\ P(X > q_\alpha) &\leq 1 - \alpha, \end{aligned} \tag{5.1}$$

or, equivalently,

$$P(X < q_\alpha) \leq \alpha \leq P(X \leq q_\alpha),$$

i.e.

$$F(q_\alpha - 0) \leq \alpha \leq F(q_\alpha). \tag{5.2}$$

To interpret (5.1), a quantile is a number with the property that it **exceeds at most** $100\alpha\%$ of the data, and **is exceeded by at most** $100(1 - \alpha)\%$ of the data.

Of all quantiles, the most important are:

The **median**, the number $M = q_{1/2}$; there are at most 50% of the data to the left of the median and at most 50% to its right.

The **quartiles** are the numbers

$$Q_1 = q_{1/4}, Q_2 = M = q_{1/2}, Q_3 = q_{3/4}.$$

Remark 5.2.

1. Quantiles are useful in **statistical analysis of data**. The median roughly locates the “middle” of a set of data, while the quartiles approximately locate every 25 % of a set of data. These will be discussed again in the next chapter.
2. If X is discrete, then a quantile can take an infinite number of values, if the line $y = \alpha$ and the curve $y = F(x)$ have in common a segment line (see Figure 1).

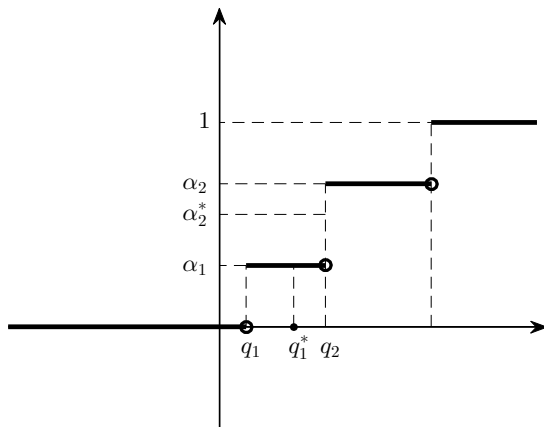


Figure 1: Quantiles for discrete variables

The case when X is continuous is more interesting and the one we will use in Statistics.

If X is continuous, then for each $\alpha \in (0, 1)$, there is a **unique** quantile q_α , given by

$$F(q_\alpha) = \alpha,$$

since F is a **continuous** function and $F(q_\alpha - 0) = \alpha = F(q_\alpha)$.

In this case, for $F : \mathbb{R} \rightarrow \mathbb{R}$ there always exists $A \subset \mathbb{R}$ such that $F : A \rightarrow [0, 1]$ is both **injective** and **surjective**, hence **invertible** (see Figure 2). Thus, in this case the unique quantile q_α is found by

$$q_\alpha = F^{-1}(\alpha). \quad (5.3)$$

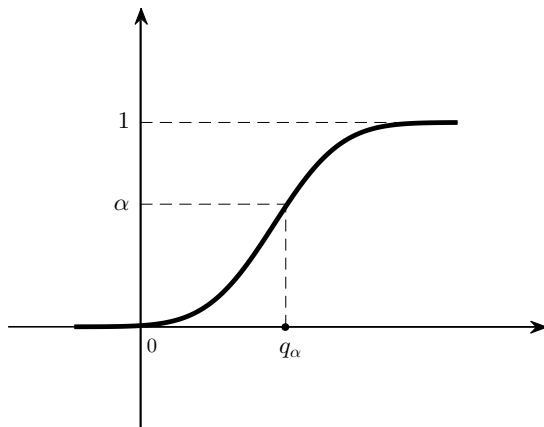


Figure 2: Quantiles for continuous variables

Now, as an interpretation, let us recall that for continuous random variables, the cdf is expressed as an integral, which means as an **area**. So we have

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

which is the area **below** the graph of the pdf f , to **the left** of q_α (see Figure 3).

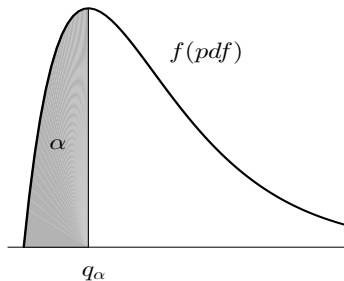


Figure 3: Quantile of order $\alpha \in (0, 1)$

6. Covariance and Correlation Coefficient

So far we have discussed numerical characteristics associated with *one* random variable. But oftentimes it is important to know if there is some kind of **relationship** between two (or more) random variables. So we need to define numerical characteristics that somehow measure that relationship.

Definition 6.1.

Let X and Y be random variables. The **covariance** of X and Y is the number

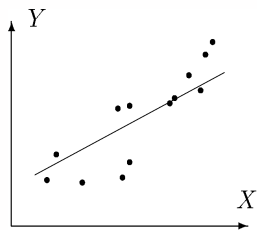
$$\text{cov}(X, Y) = E\left((X - E(X)) \cdot (Y - E(Y))\right), \quad (6.1)$$

if it exists. The **correlation coefficient** of X and Y is the number

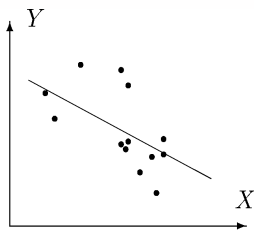
$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}, \quad (6.2)$$

if $\text{cov}(X, Y)$, $V(X)$, $V(Y)$ exist and $V(X) \neq 0$, $V(Y) \neq 0$.

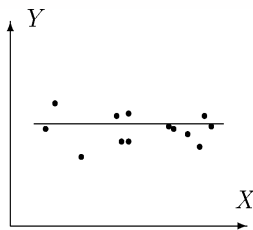
Notice the similarity between the definition of the covariance and that of the variance. The covariance measures the variation of two random variables *with respect to each other*. Just like with variance, **large values** (in absolute value) of the covariance show a **strong relationship** between X and Y , while **small absolute values** suggest a **weak relationship**. Unlike variance, covariance can also be *negative*. A negative value means that as the values of one variable increase, the values of the other decrease (see Figure 4).



(a) $\text{Cov}(X, Y) > 0$



(b) $\text{Cov}(X, Y) < 0$



(c) $\text{Cov}(X, Y) = 0$

Figure 4: Covariance

The covariance has the following properties:

Theorem 6.2.

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

- a) $\text{cov}(X, X) = V(X)$.
- b) $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$.
- c) If X and Y are independent, then $\text{cov}(X, Y) = \rho(X, Y) = 0$ (we say that X and Y are **uncorrelated**).
- d) $V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab \text{cov}(X, Y)$, for all $a, b \in \mathbb{R}$.
- e) $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$.

Proof.

a) This follows directly from definition.

b) A straightforward computation leads to

$$\begin{aligned}\text{cov}(X, Y) &= E\left((X - E(X)) \cdot (Y - E(Y))\right) \\&= E\left(XY - E(X)Y - E(Y)X + E(X)E(Y)\right) \\&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\&= E(XY) - E(X)E(Y).\end{aligned}$$

c) This follows from b), keeping in mind that X and Y are **independent**, so $E(XY) = E(X)E(Y)$.

Proof.

d)

$$\begin{aligned}
 V(aX + bY) &= E\left[\left(aX + bY - aE(X) - bE(Y)\right)^2\right] \\
 &= E\left[\left(a(X - E(X)) + b(Y - E(Y))\right)^2\right] \\
 &= E\left[a^2(X - E(X))^2 + 2ab(X - E(X))(Y - E(Y))\right. \\
 &\quad \left.+ b^2(Y - E(Y))^2\right] \\
 &= a^2V(X) + b^2V(Y) + 2ab \operatorname{cov}(X, Y).
 \end{aligned}$$

e)

$$\begin{aligned}
 \operatorname{cov}(X + Y, Z) &= E\left((X + Y - E(X) - E(Y))(Z - E(Z))\right) \\
 &= E\left((X - E(X))(Z - E(Z)) + (Y - E(Y))(Z - E(Z))\right) \\
 &= \operatorname{cov}(X, Z) + \operatorname{cov}(Y, Z).
 \end{aligned}$$



Remark 6.3.

1. Property d) of Theorem 6.2 can be generalized to **any number** of variables:

$$V\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j \operatorname{cov}(X_i, X_j).$$

2. A consequence of a) and e) of Theorem 6.2 is the following property:

$$\operatorname{cov}(aX + b, X) = aV(X), \text{ for all } a, b \in \mathbb{R}.$$

3. The converse of Theorem 6.2c) is **not true**. **Independence** is a **much stronger** condition.

Theorem 6.4.

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

- a)* $|\rho(X, Y)| \leq 1$, i.e. $-1 \leq \rho(X, Y) \leq 1$.
- b)* $|\rho(X, Y)| = 1$ if and only if there exist $a, b \in \mathbb{R}$, $a \neq 0$, such that $Y = aX + b$.

Remark 6.5.

As Theorem 6.4 states, the correlation coefficient $\rho(X, Y)$ measures the **linear “trend”** between the variables X and Y .

When $\rho = \pm 1$, there is **perfect linear correlation**, so all the points (X, Y) are on a **straight line** (see Figure 5). The closer its value is to ± 1 , the “more linear” the relationship between X and Y is.

This notion will be revisited in the next chapter.

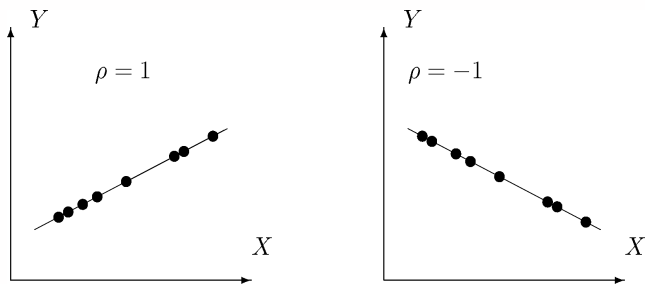


Figure 5: Perfect correlation

7. Inequalities

Inequalities can be useful in estimation theory, for **approximating** probabilities or numerical characteristics associated with a random variable.

Proposition 7.1 (Hölder's Inequality).

Let X and Y be random variables and $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$E(|XY|) \leq (E(|X|^p))^{\frac{1}{p}} \cdot (E(|Y|^q))^{\frac{1}{q}}. \quad (7.1)$$

Remark 7.2.

1. One important particular case of Hölder's inequality is for $p = q = 2$,

$$E(|XY|) \leq \sqrt{E(X^2)} \cdot \sqrt{E(Y^2)}, \quad (7.2)$$

known as **Schwarz's inequality**.

Remark (Cont).

2. A particular case of the above inequality is for $Y = 1$,

$$E(|X|) \leq \sqrt{E(X^2)}, \quad (7.3)$$

known as **Cauchy-Buniakowsky's inequality**.

Proposition 7.3 (Minkowsky's Inequality).

Let X and Y be random variables and let $p > 1$. Then

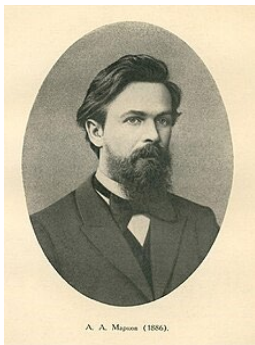
$$(E(|X + Y|^p))^{\frac{1}{p}} \leq (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}}. \quad (7.4)$$

Proposition 7.4 (Lyapunov's Inequality).

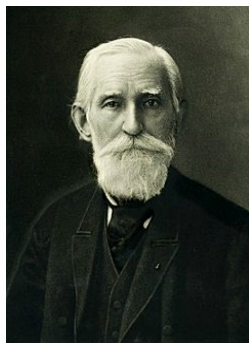
Let X be a random variable, let $0 < a < b$ and $c \in \mathbb{R}$. Then

$$(E(|X - c|^a))^{\frac{1}{a}} \leq (E(|X - c|^b))^{\frac{1}{b}}. \quad (7.5)$$

The next two inequalities are *specific* to **random variables** and are due to A. A. Markov and P. L. Chebyshev. These inequalities have many applications in **statistical analysis**.



Andrey Andreyevich Markov
(1856 - 1922)



Pafnuty Lvovich Chebyshev
(1821 - 1894)

Proposition 7.5 (Markov's Inequality).

Let X be a random variable and let $a > 0$. Then

$$P(|X| \geq a) \leq \frac{1}{a} E(|X|). \quad (7.6)$$

Proof.

Let $A = \{e \in S \mid |X(e)| \geq a\}$, with the **indicator function**

$$I_A(e) = \begin{cases} 0, & |X(e)| < a \\ 1, & |X(e)| \geq a \end{cases}.$$

Then

$$a I_A(e) = \begin{cases} 0, & |X(e)| < a \\ a, & |X(e)| \geq a \end{cases}.$$

Proof.

Now, if $|X(e)| < a$, then

$$aI_A(e) = 0 \leq |X(e)|$$

and if $|X(e)| \geq a$, then

$$aI_A(e) = a \leq |X(e)|.$$

So, either way,

$$aI_A(e) \leq |X(e)|, \quad \forall e \in S.$$

That means, as **random variables**,

$$aI_A \leq |X|,$$

which means the same thing is true for their **expected values**,

$$E(aI_A) \leq E(|X|).$$

Proof.

$$E(aI_A) \leq E(|X|).$$

The pdf of aI_A is

$$aI_A \left(\begin{array}{cc} 0 & a \\ 1 - P(|X| \geq a) & P(|X| \geq a) \end{array} \right),$$

so

$$E(aI_A) = aP(|X| \geq a).$$

Thus,

$$aP(|X| \geq a) \leq E(|X|),$$

i.e.

$$P(|X| \geq a) \leq \frac{1}{a}E(|X|).$$



Proposition 7.6 (Chebyshev's Inequality).

Let X be a random variable and let $\varepsilon > 0$. Then

$$P\left(|X - E(X)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} V(X), \quad (7.7)$$

or, equivalently,

$$P\left(|X - E(X)| < \varepsilon\right) \geq 1 - \frac{1}{\varepsilon^2} V(X), \quad (7.8)$$

Proof.

Apply Markov's inequality (7.6) to $(X - E(X))^2$ and $a = \varepsilon^2$, to get

$$P\left((X - E(X))^2 \geq \varepsilon^2\right) \leq \frac{1}{\varepsilon^2} E\left((X - E(X))^2\right),$$

Proof.

i.e.

$$P\left(|X - E(X)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} V(X),$$

and, equivalently,

$$1 - P\left(|X - E(X)| < \varepsilon\right) \leq \frac{1}{\varepsilon^2} V(X),$$

$$P\left(|X - E(X)| < \varepsilon\right) \geq 1 - \frac{1}{\varepsilon^2} V(X).$$



Example 7.7.

Suppose the number of errors in a new software, X , has expectation $E(X) = 20$. Find a bound for the probability that there are **at least** 30 errors, if the standard deviation is

a) $\sigma(X) = 2$;

b) $\sigma(X) = 5$.

Solution. According to Chebyshev's inequality, (7.7), we have

$$P(|X - 20| \geq \varepsilon) \leq \frac{(\sigma(X))^2}{\varepsilon^2}.$$

So,

$$\begin{aligned} P(X \geq 30) &= P(X - 20 \geq 10) \\ &\leq P\left((X - 20 \geq 10) \cup (X - 20 \leq -10)\right) \\ &= P(|X - 20| \geq 10) \leq \frac{(\sigma(X))^2}{100}. \end{aligned}$$

a) If $\sigma(X) = 2$, we can estimate that

$$P(X \geq 30) \leq \frac{4}{100} = 0.04.$$

b) However, for a **larger** standard deviation of $\sigma(X) = 5$, the estimation is

$$P(X \geq 30) \leq \frac{25}{100} = 0.25.$$



8. Central Limit Theorem

Central Limit Theorems are also results that can help **approximate** characteristics of random variables. First, a little bit of preparation.

Given the **special nature** of random variables, as opposed to numerical variables, there are **various types of convergence** that can be defined for sequences of such variables, having to do with probability-related notions (convergence in probability, in mean, **in distribution**, convergence almost surely, etc.)

Definition 8.1.

Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables with cumulative distribution functions $F_n = F_{X_n}$, $n \in \mathbb{N}$ and let X be a random variable with cdf $F = F_X$.

Then X_n **converges in distribution** to X , denoted by $X_n \xrightarrow{d} X$, if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x), \quad (8.1)$$

for every $x \in \mathbb{R}$, a point of continuity of F .

Remark 8.2.

Convergence in distribution is **especially important**, because the cdf of a random variable is used to compute **probabilities**.

Knowing the **limiting cdf** of a sequence of random variables makes possible the computation of probabilities (and other characteristics) in the “long run”. So such results can be helpful in estimating characteristics of random variables as n gets larger.

A statement about the limit in distribution of a sequence of random variables is called a **limit theorem**.

If the limit variable has a Normal distribution, then such a result is called a **central limit theorem**.

So, there are **many** such results, the name “Central Limit Theorem” is just *generic*.

We want to discuss a central limit theorem that applies to the following case:

Suppose X_1, X_2, \dots, X_n are **independent, identically distributed (iid)** random variables (this is a case that will be used oftenly in Statistics). Having the **same pdf**, they have the **same expectation** $\mu = E(X_i)$ and the **same standard deviation** $\sigma = \text{Std}(X_i) = \sqrt{V(X_i)}$.

We are interested in the random variable

$$S_n = X_1 + \dots + X_n.$$

This case appears in many applications and in many statistical procedures. We see right away that

$$\begin{aligned} E(S_n) &= n\mu, \\ V(S_n) &= n\sigma^2. \end{aligned}$$

How does S_n behave for large n ?

The *pure* sum S_n **diverges**. In fact, this should be anticipated because

$$V(S_n) = n\sigma^2 \rightarrow \infty,$$

so the variability of S_n grows unboundedly, as n goes to infinity.

The *average* S_n/n **converges**. Indeed, in this case, we have

$$V(S_n/n) = \frac{1}{n^2} V(S_n) = \frac{\sigma^2}{n} \rightarrow 0,$$

so the variability of S_n/n vanishes as $n \rightarrow \infty$.

An interesting case is the variable S_n/\sqrt{n} ,

$$\begin{aligned} E(S_n/\sqrt{n}) &= \sqrt{n}\mu, \\ V(S_n/\sqrt{n}) &= \sigma^2, \end{aligned}$$

which **neither diverges, nor converges**.

In fact, it behaves like some **random variable**. The following theorem (CLT) states that this variable has approximately Normal distribution for large n . In fact, the result is for its *reduced (standardized)* variable

$$\frac{S_n/\sqrt{n} - E(S_n/\sqrt{n})}{\text{Std}(S_n/\sqrt{n})} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Theorem 8.3.

Let X_1, X_2, \dots, X_n be independent, identically distributed random variables with expectation $\mu = E(X_i)$ and standard deviation $\sigma = \sigma(X_i)$ and let

$$S_n = X_1 + \dots + X_n. \quad (8.2)$$

Then, as $n \rightarrow \infty$, the reduced sum

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \in N(0, 1). \quad (8.3)$$

Remark 8.4.

1. Relation (8.3) means that

$$\begin{aligned} F_{Z_n} &\rightarrow F_{N(0,1)}, \text{ i.e.,} \\ P(Z_n \leq x) &\rightarrow P(Z \leq x), \forall x \in \mathbb{R}, \text{ as } n \rightarrow \infty. \end{aligned}$$

2. This result can be *very helpful*, since $F_{N(0,1)}(x) = \Phi(x)$ is **Laplace's function** (see equation (6.6) in Lecture 5), whose values are **known**.

3. The CLT can be used as an **approximation tool** for n “large”. In practice, it has been determined that that means $n > 30$.

Example 8.5.

A disk has free space of 330 megabytes. Is it likely to be sufficient for 300 independent images, if each image has expected size of 1 megabyte with a standard deviation of 0.5 megabytes?

Solution.

For each $i = 1, 2, \dots, n$ (i.e. for each image), let X_i denote the space it takes, in megabytes.

Then the **total** space taken by **all** 300 images will be the sum

$$S_n = X_1 + X_2 + \cdots + X_n$$

and there will be sufficient space on the disk if

$$S_n \leq 330.$$

We have $n = 300$, $\mu = 1$, $\sigma = 0.5$.

The number of images n is **large enough**, so the CLT applies to their total size S_n . Then

$$\begin{aligned}
 P(\text{sufficient space}) &= P(S_n \leq 330) \\
 &= P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{330 - n\mu}{\sigma\sqrt{n}}\right) \\
 &= P\left(Z_n \leq \frac{330 - 300 \cdot 1}{0.5 \cdot 10\sqrt{3}}\right) \\
 &= P(Z_n \leq 3.46) \\
 &\stackrel{\text{CLT}}{\approx} P(Z \leq 3.46) = \Phi(3.46) = 0.9997,
 \end{aligned}$$

a **very high** probability, hence, the available disk space is **very likely** to be sufficient.

