1. Empirical distribution function $\hat{F}$
2. Binomial process
3. Deviation between $\hat{F}$ and $F$

Material from Khmaladze E.V. book
Statistical methods with application to demography and life insurance

Estate V. Khmaladze
In demography, statistical problems begin where they always begin: we do not know the true distribution function $F$ of the duration of life; but we have data, based on which we have to make inference about this unknown $F$, test different hypotheses which we can formulate about $F$, or estimate $F$ and/or some of its parameters.

Suppose we have $n$ independent, identically distributed durations of life $T_1, T_2, \ldots, T_n$, that is, $n$ independent random variables, each with the same distribution function $F$. Let us consider a function of a random variable $T$ and of a point $x$, defined as

$$I_{\{T < x\}} = \begin{cases} 1, & \text{if } T < x; \\ 0, & \text{if } T \geq x. \end{cases}$$

More generally, let $A$ be a subset of $[0, \infty)$, say, an interval $[x_1, x_2]$, and let $I_{\{T \in A\}}$ be a function of $T$ and $A$ defined as

$$I_{\{T \in A\}} = \begin{cases} 1, & \text{if } T \in A; \\ 0, & \text{if } T \not\in A. \end{cases} \quad (3.1)$$

The function (3.1) is called an indicator function of a set, or event, $A$. Indicators are very useful and will frequently be used below. In particular, consider

$$z_n(x; T_1, T_2, \ldots, T_n) = \sum_{i=1}^{n} I_{\{T_i < x\}}. \quad (3.2)$$
The empirical distribution function of duration of life

For a given $x$ this function $z_n(x; T_1, T_2, \ldots, T_n)$ is equal to the number among $n$ individuals who have durations of life less than $x$. However, since $T_1, T_2, \ldots, T_n$ are random, $z_n(x; T_1, T_2, \ldots, T_n)$ is also a random variable for each $x$, that is, a random function in $x$ or a random process. For any given values of $T_1, T_2, \ldots, T_n$ the trajectory of $z_n(x; T_1, T_2, \ldots, T_n)$ is a piece-wise constant, non-decreasing function of $x$. It equals 0 at $x = 0$ and equals $n$ for $x > \max T_i$. It has jumps at $x = T_1, T_2, \ldots, T_n$, each of height 1.

\textbf{Exercise.} Suppose we are given $n = 3$ life times (in years), $T_1 = 75.5$, $T_2 = 48.6$, $T_3 = 69.1$.

Draw the graph of $z_n(x; T_1, T_2, T_3)$ in $x$. Is $z_n(x; T_1, T_2, T_3)$ continuous in $x$ from the left? And from the right? If it were that $T_1 = 48.6$, $T_2 = 69.1$, $T_3 = 75.5$, would $z_n(x; T_1, T_2, T_3)$, as a function of $x$, change? $\triangle$

From now on we will drop $T_1, T_2, \ldots, T_n$ from the notation for $z_n$ and write simply $z_n(x)$.

An empirical distribution function of random variables $T_1, T_2, \ldots, T_n$ is defined as

$$
\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(T_i < x)}.
$$

From the properties of $z_n(x)$ we can easily see that $\hat{F}_n(x)$ is a piece-wise constant, non-decreasing function of $x$. It equals 0 at $x = 0$ and equals 1 for $x > \max T_i$. It has jumps at $x = T_1, T_2, \ldots, T_n$ of height $1/n$ each.

Let us clarify the question of jump-points. The empirical distribution function has its first jump, from 0 to $1/n$, at the point $T_{(1)} = \min T_i$; its second jump, from $1/n$ to $2/n$, occurs at the point $T_{(2)}$, which is the second smallest duration of life; and so on until the last jump, from $(n-1)/n$ to 1, at the point $T_{(n)} = \max T_i$. The random times $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ are called order statistics (based on $T_1, T_2, \ldots, T_n$). Although we assumed that random variables $T_1, T_2, \ldots, T_n$ were independent, the order statistics can by no means be regarded as independent. For example, the distribution of $T_{(i)}$ very much depends on the values of $T_{(i-1)}$ and $T_{(i+1)}$: it can only take values between $T_{(i-1)}$ and
\( T_{(i+1)} \), so even the range depends on the surrounding order statistics. Any permutation of \( T_1, T_2, \ldots, T_n \) does not change the order statistics. Therefore, \( \hat{F}_n(x) \) depends not directly on \( T_1, T_2, \ldots, T_n \), but only on their order statistics \( T_{(1)}, T_{(2)}, \ldots, T_{(n)} \):

\[
\frac{1}{n} \sum_{i=1}^{n} I_{(T_i < x)} = \frac{1}{n} \sum_{i=1}^{n} I_{(T_{(i)} < x)},
\]

(3.4)

Although these two sums are equal, we will see later on that they suggest quite different methods for the study of the empirical distribution function: the left side represents it as a sum of independent random functions, as each \( I_{(T_i < x)} \) is a random function in \( x \); while the right side represents it as a point process with stopping times \( T_{(1)}, T_{(2)}, \ldots, T_{(n)} \) (cf. Lecture 8 below).

The empirical distribution function is one of the central objects in non-parametric statistics, if not in all of mathematical statistics. There is a huge body of publications on the subject and the interested reader may gain wider understanding of the theory of the empirical distribution function by reading, e.g., the monograph Shorack and Wellner (2009). Below we will mention some classical facts concerning the asymptotic behavior of \( \hat{F}_n \), and then we shall use them.

Let us begin with the following statement:

"If \( T_1, T_2, \ldots, T_n \) are independent and identically distributed random variables, then for every fixed \( x \) the random variable \( z_n(x) \) has the binomial distribution:

\[
P\{z_n(x) = k\} = \binom{n}{k} b(k; n, F(x)) = \binom{n}{k} F^k(x) [1 - F(x)]^{n-k}, \quad k = 0, 1, \ldots, n.
\]

(3.5)"

In fact, at fixed \( x \) each indicator \( I_{(T_i < x)} \) is a Bernoulli random variable with two values, 0 and 1, and

\[
P\{I_{(T_i < x)} = 1\} = P\{T_i < x\} = F(x).
\]

Since random variables \( T_1, T_2, \ldots, T_n \) are independent, then these Bernoulli random variables are also independent. Therefore \( z_n(x) \) is the sum of \( n \) independent Bernoulli random variables with the probability of "success" \( F(x) \), and therefore is indeed a binomial random variable with the distribution (3.5).
The empirical distribution function of duration of life

Once we know the distribution of the random variables \( z_n(x) \) we can know the distribution of \( \hat{F}_n(x) \):

\[
P\left\{ \hat{F}_n(x) = \frac{k}{n} \right\} = b(k; n, F(x)), \quad k = 0, 1, \ldots, n.
\]

As a result

\[
\mathbb{E}\hat{F}_n(x) = F(x) \quad \text{and} \quad \text{Var}\hat{F}_n(x) = \frac{1}{n} F(x) \left[ 1 - F(x) \right]. \quad (3.6)
\]

\[\Diamond \] **Exercise.** Find \( \mathbb{E}[I_{T_i < x}] \) and \( \text{Var}[I_{T_i < x}] \), and then prove (3.6). \[\triangle \]

The first equality in (3.6) shows that if random variables \( T_1, T_2, \ldots, T_n \) are identically distributed with distribution function \( F(x) \), then the empirical distribution function \( \hat{F}_n(x) \) is an unbiased estimator of this common \( F(x) \). The second equality shows that if \( T_1, T_2, \ldots, T_n \) are also independent, then the variance \( \hat{F}_n(x) \) converges to 0 at the rate of \( \frac{1}{n} \), i.e. \( \hat{F}_n(x) \) is a consistent estimator in the mean-square sense. We will now prove that \( \hat{F}_n(x) \) is consistent in a much stronger sense.

First of all let us look at the following important inequality.

The exponential inequality for binomial distribution, see for instance Shorack and Wellner (2009), p. 440, states that

*If random variables \( \xi_n \) have the binomial distribution*

\[
P\left\{ \frac{\xi_n}{n} - p > \varepsilon \right\} \leq e^{-na(p, \varepsilon)}, \quad (3.7)
\]

where \( a(p, \varepsilon) = p \int_0^{e/\varepsilon} \ln(1 + y) \, dy \).

Consequently, if \( T_1, T_2, \ldots, T_n \) are independent and identically distributed, then

\[
P\left\{ |\hat{F}_n(x) - F(x)| > \varepsilon \right\} \leq e^{-na(F(x), \varepsilon)} + e^{-ne(F^*(x), \varepsilon)}. \quad (3.8)
\]

Proofs of the inequalities (3.7) and (3.8) will be given below; but for now, let us look at how (3.8) can be applied.
For any fixed $x > 0$ and for arbitrarily small fixed $\varepsilon > 0$, values of $a(F(x), \varepsilon)$ and $a(F^*(x), \varepsilon)$ are fixed numbers. Consequently, (3.8) shows that the probability that the empirical distribution function $\hat{F}_n(x)$ deviates from its mean $F(x)$ by more than $\varepsilon$ decreases exponentially with increasing number of observations $n$. From here, in turn, it follows that

for each fixed $x > 0$

\[ \hat{F}_n(x) \to F(x), \quad n \to \infty, \tag{3.9} \]

with probability 1.

This is a much stronger statement than the convergence of the variance $\hat{F}_n(x)$ to 0.

Proof of (3.9). Let

\[ A_n = \{ |\hat{F}_n(x) - F(x)| > \varepsilon \}. \]

Then, according to (3.8),

\[ \sum_{n=1}^{\infty} P(A_n) \leq \sum_{n=1}^{\infty} \left[ e^{-na(F(x),\varepsilon)} + e^{-na(F^*(x),\varepsilon)} \right] < \infty, \]

which, using the Borel–Cantelli Lemma, implies that the probability of $A_n$ occurring infinitely often is 0 (see, e.g., Shiryaev (1980)). In other words, the events $A_n, n = 1, 2, \ldots$, will stop occurring and $\hat{F}_n(x)$ will stay within an $\varepsilon$ neighborhood of $F(x)$.

Now we come back to the exponential inequality (3.8).

Note first the inequality

\[ P\{ f(X) > c \} \leq \frac{Ef(X)}{c}, \]

true for a positive increasing function $f(X)$ of a random variable $X$, which is called the Markov inequality, see, e.g., Shiryaev (1980). The Markov inequality is a generalization of the Chebyshev inequality, and can be proved in exactly the same way. Now note that

\[ P\left\{ \frac{\xi_n}{n} - p \geq \varepsilon \right\} = P\{\xi_n - np \geq n\varepsilon\} = P\{e^{\lambda(\xi_n - np)} \geq e^{n\lambda\varepsilon}\} \]
The empirical distribution function of duration of life for any positive value of the parameter $\lambda$. We will select a specific value later. Let us use the Markov inequality

$$P\{e^{\lambda(\xi_n-np)} \geq e^{n\lambda\varepsilon}\} \leq \frac{Ee^{\lambda(\xi_n-np)}}{e^{n\lambda\varepsilon}}.$$  

For binomial $\xi$ it is, however, easy to calculate the expected value $Ee^{\lambda(\xi_n-np)}$ on the right (for this it is sufficient to use the binomial formula):

$$Ee^{\lambda\xi} = (1 + p(e^\lambda - 1))^n.$$  

Using the inequality $(1 + x)^n < e^{nx}$ for this last expression we finally obtain

$$P\left\{\frac{\xi_n}{n} - p \geq \varepsilon\right\} \leq e^{np(e^\lambda-1)}e^{-n\lambda(p-p+n\lambda \varepsilon)} = e^{np(e^\lambda-1-\lambda)-n\lambda \varepsilon}.$$  

Not for all values of $\lambda$ does the right hand side produce a useful inequality. Hence, it is important to choose $\lambda$ which minimizes the exponent. Differentiating with respect to $\lambda$ we find:

$$p(e^\lambda - 1) = \varepsilon, \quad \lambda = \ln \left(1 + \frac{\varepsilon}{p}\right).$$  

For this value of $\lambda$ the exponent becomes

$$np(e^{\lambda} - 1 - \lambda) - n\lambda \varepsilon = n\varepsilon - n(p + \varepsilon)\ln \left(1 + \frac{\varepsilon}{p}\right)$$  

$$= np \left[\frac{\varepsilon}{p} - \left(1 + \frac{\varepsilon}{p}\right)\ln \left(1 + \frac{\varepsilon}{p}\right)\right]$$  

$$= -np \left(\frac{\varepsilon}{p}\right),$$

because the expression in square brackets is equal to the integral

$$-\int_0^{\varepsilon/p} \ln(1+x) \, dx.$$  

Therefore, the inequality (3.7) is proved.

Now let us show that (3.8) follows from (3.7). If random variables $T_1, T_2, \ldots, T_n$ are independent and identically distributed, then $z_n(x) =$
\[ n \hat{F}_n(x) \text{ and } n - z_n(x) = nF^*_n(x) \text{ are both binomial random variables;} \]
and since
\[
P \left\{ \left| \hat{F}_n(x) - F(x) \right| > \varepsilon \right\} = P \left\{ \hat{F}_n(x) - F(x) > \varepsilon \right\} + P \left\{ \hat{F}_n^*(x) - F^*(x) > \varepsilon \right\}, \tag{3.10}
\]
we can apply inequality (3.7) to both summands on the right, and this leads to (3.8).

One can prove a still stronger statement on convergence of the empirical distribution function \( \hat{F}_n(x) \) to \( F(x) \) by using the so-called Glivenko–Cantelli theorem (Glivenko 1933), which is the theorem for which we were actually preparing. According to this theorem,

if \( T_1, T_2, \ldots, T_n \) are independent and identically distributed with distribution function \( F(x) \), then
\[
\sup_x \left| \hat{F}_n(x) - F(x) \right| \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty, \tag{3.11}
\]
with probability 1.

Therefore, \( \hat{F}_n(x) \) converges to \( F(x) \) not only at each particular \( x \), but also converges uniformly in \( x \).

\[ \diamond \text{ Exercise.} \] It is well known that convergence of a sequence of functions \( \varphi_n(x) \) to a limiting function \( \varphi(x) \) at every \( x \) does not imply uniform convergence of \( \varphi_n \) to \( \varphi \).

a) Construct a corresponding example with \( \varphi \) continuous on \([0, 1]\).

b) Nevertheless, in the case of distribution functions, point-wise convergence (3.9) implies uniform convergence (3.11), because distribution functions \( F \) have two specific properties – they are non-decreasing and bounded. Make sure that you see this without resorting to the proof below. \( \triangle \)

Let us prove (3.11). In this proof it is not necessary for \( F(x) \) to be continuous – it can be any distribution function. For a given \( F(x) \) let us consider its graph
\[
\Gamma_F = \{(u,x) : F(x) \leq u \leq F(x+)\}. \tag{3.12}
\]

If \( F(x) \) is continuous in \( x \), then at this \( x \) there is only one value of \( F(x) \),
\( u = F(x) \); if \( F(x) \) has a jump at \( x \), then values of \( u \) fill in the whole interval from \( F(x) = F(x^-) \) to \( F(x^+) \).

Let us take an arbitrarily large integer \( N \) and consider it as fixed. Choose \( x_i \) as

\[
x_i = \min \left\{ x : F(x) \leq \frac{i}{N} \leq F(x^+) \right\}, \quad i = 1, \ldots, N - 1,
\]

and suppose \( x_0 = -\infty \) and \( x_N = \infty \). With this choice of points, if \( x_i < x_{i+1} \), then

\[
0 \leq F(x_{i+1}) - F(x_i^+) \leq \frac{1}{N},
\]

while if \( x_i = x_{i+1} \) then \( F(x_{i+1}) - F(x_i) = 0 \). Note also that (3.9) is true, and can be proved in exactly the same way, also for the limits from the right, \( \hat{F}_n(x^+) \to F(x^+) \), which is useful if \( F \) happens to be discontinuous at \( x \). Since the number of distinct \( x_i \) is not exceeding finite \( N \), from (3.8) follows, that

\[
\max_{0 \leq i \leq N} |\hat{F}_n(x_i) - F(x_i)| \to 0 \quad \text{and} \quad \max_{0 \leq i \leq N} |\hat{F}_n(x_i^+) - F(x_i^+)| \to 0,
\]
as \( n \to \infty \), with probability 1.
Now for any \( x \) such that that \( x_i < x < x_{i+1} \), we have

\[
\hat{F}_n(x) - F(x) \leq \hat{F}_n(x_{i+1}) - F(x_i) \leq \hat{F}_n(x_{i+1}) - F(x_{i+1}) + \frac{1}{N}
\]

Similarly,

\[
\hat{F}_n(x) - F(x) \geq \hat{F}_n(x_i) - F(x_i) \geq \hat{F}_n(x_i) - F(x_i) - \frac{1}{N}.
\]

Therefore, for all such \( x \)

\[
|\hat{F}_n(x) - F(x)| \leq \max_{0 \leq i \leq N} \left( \max \left| \hat{F}_n(x_i) - F(x_i) \right|, \left| \hat{F}_n(x_i) - F(x_i) \right| \right) + \frac{1}{N}.
\]

If \( x = x_i \) for some \( i \), the above inequality is still true. Since its right hand side is independent of \( x \), we will obtain

\[
0 \leq \sup_x |\hat{F}_n(x) - F(x)| \leq \\
\max_{0 \leq i \leq N} \left( \max \left| \hat{F}_n(x_i) - F(x_i) \right|, \left| \hat{F}_n(x_i) - F(x_i) \right| \right) + \frac{1}{N}.
\]

The right hand side here converges to \( \frac{1}{N} \) with probability 1. Consequently,

\[
0 \leq \limsup_{n \to \infty} |\hat{F}_n(x) - F(x)| \leq \frac{1}{N}.
\]

Since \( N \) can be arbitrarily large, this limit is equal to 0, which proves (3.11).