# Empirical Processes, and the Kolmogorov–Smirnov Statistic<br>Math 6070, Spring 2013

Davar Khoshnevisan
University of Utah

March 1, 2013

## Contents

## 1 Some Basic Theory

We wish to address the following fundamental problem: Let $X_1, X_2, \ldots$ be an i.i.d. sample from a distribution function $F$. Then, what does $F$ "look like"? There are many way to interpret the last question, but no matter what we mean by "look like," we need to start by a statistical estimate for the unknown "parameter" $F$.

Consider the *empirical distribution function*

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbf{I}\{X_i \leqslant x\}.$$

[Plot it!] We begin with some elementary facts about $\hat{F}_n$.

## 1.1 Consistency and Unbiasedness at a Point

Fix $x \in \mathbf{R}$. Then, $n\hat{F}_n(x) \sim \text{Binomial}(n\,, F(x))$. Consequently,

$$\mathrm{E}\left[\hat{F}_n(x)\right] = \frac{1}{n}\mathrm{E}\left[n\hat{F}_n(x)\right] = F(x).$$

That is, $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$ for each fixed $x$. Also,

$$\mathrm{Var}\left(\hat{F}_n(x)\right) = \frac{1}{n^2}\mathrm{Var}\left(n\hat{F}_n(x)\right) = \frac{F(x)\big[1-F(x)\big]}{n}.$$

Consequently, by the Chebyshev inequality, $\hat{F}_n(x) \overset{\mathrm{P}}{\to} F(x)$. Therefore, $\hat{F}_n(x)$ is a consistent, unbiased estimator of $F(x)$ for each fixed $x \in \mathbf{R}$. Based on this, we can construct nice confidence intervals for $F(x)$ for a fixed $x$. But what if we wanted to have a confidence set for $(F(x_1)\,, F(x_2))$, or wished to know something about the entire function $F$?

## 1.2 The Kolmogorov–Smirnov Statistic

Define the *Kolmogorov–Smirnov statistic* $D_n$ by

$$D_n := \max_{-\infty<x<\infty}\left|\hat{F}_n(x) - F(x)\right|. \tag{1}$$

We intend to prove the following:

**Theorem 1 (Glivenko–Cantelli)** *As $n \to \infty$, $D_n \overset{\mathrm{P}}{\to} 0$.*

In particular, we can fix $\varepsilon > 0$ small and deduce that if $n$ is large then with high probability, $D_n \leqslant \varepsilon$ (say!). If so, then we could plot

$$\mathscr{C}_n(\varepsilon) := \left\{(x\,, y) : \ \left|\hat{F}_n(x) - y\right| \leqslant \varepsilon\right\}.$$

This gives us a good idea of the shape of $F$ because $|D_n| \leqslant \varepsilon$ if and only if the graph of $F$ lies in $\mathscr{C}_n(\varepsilon)$. [Recall that the graph of $F$ is the collection of all pairs $(x\,, F(x))$.]

Before we prove Theorem 1 we take another look at $D_n$ and its so-called "distribution-free" property.

**Theorem 2 (The Distribution-Free Property)** *The distribution of $D_n$ is the same for all continuous underlying distribution functions $F$.*

**Proof.** I first prove this in the slightly simpler case that $F$ is strictly increasing. In this case $F^{-1}$ [the inverse] exists and is strictly increasing also. Therefore,

$$\begin{aligned} D_n &= \max_{-\infty<x<\infty} \left| \hat{F}_n(x) - F(x) \right| \\ &= \max_{0\leqslant y\leqslant 1} \left| \hat{F}_n\left(F^{-1}(y)\right) - F\left(F^{-1}(y)\right) \right| \\ &= \max_{0\leqslant y\leqslant 1} \left| \hat{F}_n\left(F^{-1}(y)\right) - y \right|. \end{aligned}$$

Now,

$$\hat{F}_n\left(F^{-1}(y)\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{I}\left\{X_i \leqslant F^{-1}(y)\right\} = \frac{1}{n}\sum_{i=1}^n \mathbf{I}\left\{F(X_i) \leqslant y\right\},$$

and this is the empirical distribution function for the i.i.d. random sample $F(X_1), \ldots, F(X_n)$.[1] The latter is a sample from the Unif$(0\,,1)$ distribution, and thus we find that the distribution of $D_n$ is the same as the Kolmogorov–Smirnov statistic for a Unif$(0\,,1)$ sample. This proves the result in the case that $F^{-1}$ exists. In the general case, $F^{-1}$ does not necessarily exists. However, consider $F^{+}(x) := \min\{y :\ F(x) > x\}$. Then, this has the property that $F^{+}(x) \leqslant z$ if and only if $x \leqslant F(z)$. Plug $F^{+}$ in place of $F^{-1}$ everywhere in the previous proof. □

## 1.3 Order Statistics

Let $X_1, \ldots, X_n$ be i.i.d. random variables, all with the same distribution function $F$. The *order statistics* are the resulting ordered random variables

$$X_{1:n} \leqslant \ldots \leqslant X_{n:n}.$$

[1] In other words, if $X \sim F$ and $F$ has nice inverse, then $F(X) \sim$ Unif$(0\,,1)$. Here is the reason: $\mathrm{P}\{F(X) \leqslant a\} = \mathrm{P}\{X \leqslant F^{-1}(a)\} = F(F^{-1}(a)) = a$, when $0 \leqslant a \leqslant 1$. Otherwise, this probability is zero [when $a < 0$] or one [when $a > 1$].

Thus, $X_{1:n} := \min_{1\leqslant j\leqslant n} X_j$, $X_{2:n} := \min(\{X_i\}_{i=1}^n \setminus \{X_{1:n}\})$, …, $X_{n:n} := \max_{1\leqslant j\leqslant n} X_j$.

If $F$ is continuous, then the order statistics attain unique values with probability one. That is, the following occurs with probability one:

$$X_{1:n} < \cdots < X_{n:n}. \tag{2}$$

We prove this in the absolutely-continuous case. Consider the event $E_{i,j} := \{X_i = X_j\}$. Then, the continuity of $F$ implies that whenever $i \neq j$,

$$\mathrm{P}(E_{i,j}) = \iint_{\{x=y\}} f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y = 0,$$

where $f$ is the common density function of $X_i$ and $X_j$. Thus, by Boole's inequality,[2]

$$\begin{aligned}\mathrm{P}\left\{X_i = X_j \text{ for some } 1\leqslant i \neq j \leqslant n\right\} &= \mathrm{P}\left(\bigcup_{1\leqslant i\neq j\leqslant n} E_{i,j}\right)\\ &\leqslant \sum_{1\leqslant i\neq j\leqslant n} \mathrm{P}(E_{i,j})\\ &= 0.\end{aligned}$$

This implies (2).

Now suppose $F$ is continuous, so that (2) is in force. Then, $D_n$ is distribution-free. So we can assume without loss of generality that $X_1, \ldots, X_n \sim$ Unif$(0\,,1)$. Also, in this case, $D_n = \max_{0\leqslant x\leqslant 1} |\hat{F}_n(x) - x|$. Now plot the function $|\hat{F}_n(x) - x|$ to find that the maximum [in $D_n$] can occur only at the order statistics. [This is a property of convex functions.] That is, $D_n = \max_{1\leqslant i\leqslant n} |\hat{F}_n(X_{i:n}) - X_{i:n}|$. But another graphical inspection shows that $\hat{F}_n(X_{i:n}) = i/n$. Therefore,

$$D_n = \max_{1\leqslant i\leqslant n} \left| X_{i:n} - (i/n) \right|. \tag{3}$$

Next, we study the distribution of each $X_{i:n}$ in the Unif$(0\,,1)$ case.

**Theorem 3** *Suppose* $X_1, \ldots, X_n \sim$ Unif$(0\,,1)$ *are i.i.d. Then, for all* $1 \leqslant k \leqslant n$*, the kth order statistic* $X_{k:n}$ *has an absolutely continuous distribution with density*

$$f_{X_{k:n}}(t) = k\binom{n}{k} t^{k-1}(1-t)^{n-k}, \qquad 0 \leqslant t \leqslant 1.$$

[2] Recall that *Boole's inequality* states that $\mathrm{P}(A_1 \cup \cdots \cup A_k) \leqslant \mathrm{P}(A_1) + \cdots + \mathrm{P}(A_k)$.

**Proof.** First let us estimate carefully the probability

$$F_{X_{k:n}}(a+\varepsilon) - F_{X_{k:n}}(a) = \mathrm{P}\left\{a < X_{k:n} \leqslant a+\varepsilon\right\},$$

where $a \in (0\,,1)$ and $\varepsilon$ is a small constant that is small enough to ensure that $a+\varepsilon < 1$. For every subinterval $J$ of $(0\,,1)$, let $N_J$ denote the total number of data points that fall in $J$; that is,

$$N_J := \sum_{j=1}^{n} \mathbf{I}\left\{X_j \in J\right\}.$$

Since the $X_j$ have a uniform distribution, it follows that $N_J \sim \mathrm{Bin}(n\,,|J|)$. In particular, we may apply a Taylor expansion or two in order to see that the follows are valid as $\varepsilon \downarrow 0$:

$$\begin{aligned}
&\mathrm{P}\left\{N_{(a,a+\varepsilon]} = 0\right\} = (1-\varepsilon)^n \approx 1 - n\varepsilon,\\
&\mathrm{P}\left\{N_{(a,a+\varepsilon]} = 1\right\} = n\varepsilon(1-\varepsilon)^{n-1} \approx n\varepsilon\left(1-(n-1)\varepsilon\right) = n\varepsilon.
\end{aligned}$$

This means that:

1. It is unlikely to have any data in $(a+a+\varepsilon]$ when $\varepsilon \approx 0$; in fact the probability of such an event is $\mathrm{P}\{N_{(a,a+\varepsilon]} \neq 0\} \approx n\varepsilon$; and yet

2. For small $\varepsilon > 0$, if we happen to see the unlikely event that $N_{(a,a+\varepsilon]} \geqslant 1$ then chances are very high that we saw only one data point in $(a\,,a+\varepsilon]$; i.e.,

$$\mathrm{P}\left(N_{(a,a+\varepsilon]} = 1 \;\middle|\; N_{(a,a+\varepsilon]} \neq 0\right) = \frac{\mathrm{P}\{N_{(a,a+\varepsilon]} = 1\}}{\mathrm{P}\{N_{(a,a+\varepsilon]} \neq 0\}} \approx 1.$$

Therefore, for small $\varepsilon$, it follows that

$$\mathrm{P}\left\{a < X_{k:n} \leqslant a+\varepsilon\right\} \approx \mathrm{P}\left\{N_{(0,a)} = k-1\,, N_{(a,a+\varepsilon]} = 1\,, N_{(a+\varepsilon,1)} = n-k\right\}.$$

But the random vector $(N_{(0,a)}\,, N_{(a,a+\varepsilon]}\,, N_{(a+\varepsilon,1)})$ has a multinomial distribution. Therefore, if $i, j, l$ are integers between 1 and $n$ such that $i+j+l = n$, then

$$\mathrm{P}\left\{N_{(0,a)} = i\,, N_{(a,a+\varepsilon]} = j\,, N_{(a+\varepsilon,1)} = l\right\} = \frac{n!}{i!\cdot j!\cdot l!} a^i \varepsilon^j (1-a-\varepsilon)^l.$$

Plug in $i := k-1$, $j := 1$ and $l := n-k$ to see that

$$\begin{aligned}
\mathrm{P}\left\{a < X_{k:n} \leqslant a+\varepsilon\right\} &\approx \frac{n!}{(k-1)!(n-k)!} a^k \varepsilon (1-a-\varepsilon)^{n-k} \\
&= k\binom{n}{k} a^k \varepsilon (1-a-\varepsilon)^{n-k} \\
&\approx \varepsilon k \binom{n}{k} a^k (1-a)^{n-k}.
\end{aligned}$$

Divide by $\varepsilon$ and let $\varepsilon \to 0$ to see that

$$\lim_{\varepsilon \downarrow 0} \frac{F_{X_{k:n}}(a+\varepsilon) - F_{X_{k:n}}(a)}{\varepsilon} = \lim_{\varepsilon \downarrow 0} \frac{\mathrm{P}\left\{a < X_{k:n} \leqslant a+\varepsilon\right\}}{\varepsilon} = k\binom{n}{k} a^k (1-a)^{n-k}.$$

This does the job. □

Let us compute the moments too.

**Theorem 4** *Suppose $X_1, \ldots, X_n \sim \mathrm{Unif}(0\,,1)$ are i.i.d. Then, for all $1 \leqslant k \leqslant n$ and $p \geqslant 1$,*

$$\mathrm{E}\left[X_{k:n}^p\right] = \frac{n!}{(k-1)!} \cdot \frac{\Gamma(k+p)}{\Gamma(n+p+1)}.$$

**Proof.** We use the density to find that $\mathrm{E}[X_{k:n}^p] = k\binom{n}{k}\int_0^1 t^{k+p-1}(1-t)^{n-k}\,\mathrm{d}t$. Now recall *Beta functions*:

$$\mathrm{B}(r\,,s) := \int_0^1 t^{r-1}(1-t)^{s-1}\,\mathrm{d}t,$$

for all $r, s > 0$. These functions can be written as follows: $\mathrm{B}(r\,,s) = \Gamma(r)\Gamma(s)/\Gamma(r+s)$. Therefore,

$$\mathrm{E}\left[X_{k:n}^p\right] = k\binom{n}{k}\mathrm{B}(k+p\,,n-k+1) = k\binom{n}{k}\frac{\Gamma(k+p)\Gamma(n-k+1)}{\Gamma(n+p+1)}.$$

Cancel terms, and recall that $\Gamma(k+1) = k!$ to finish. □

### 1.4 Proof of the Glivenko–Cantelli Theorem

I will prove the Glivenko–Cantelli theorem in the slightly less general setting where $F$ is continuous. In this case, $D_n$ has the same distribution as in the case that the $X$'s are Unif$(0\,,1)$. Now, according to Theorem 4,

$$\begin{aligned}
\mathrm{E}\left[X_{k:n}\right] &= \frac{k}{n+1},\\
\mathrm{E}\left[X_{k:n}^2\right] &= \frac{k}{n+1}\cdot\frac{k+1}{n+2},\\
\mathrm{E}\left[X_{k:n}^3\right] &= \frac{k}{n+1}\cdot\frac{k+1}{n+2}\cdot\frac{k+2}{n+3},\\
\mathrm{E}\left[X_{k:n}^4\right] &= \frac{k}{n+1}\cdot\frac{k+1}{n+2}\cdot\frac{k+2}{n+3}\cdot\frac{k+3}{n+4}.
\end{aligned}$$

Let $\mu := \mathrm{E}X_{k:n}$, and apply the binomial theorem to find that

$$\begin{aligned}
\mathrm{E}\left[(X_{k:n}-\mu)^4\right] &= \mathrm{E}\left[X_{k:n}^4\right] - 4\mathrm{E}\left[X_{k:n}^3\right]\mu + 6\mathrm{E}\left[X_{k:n}^2\right]\mu^2 - 4\mu^4 + \mu^4\\
&= \mathrm{E}\left[X_{k:n}^4\right] - 4\mathrm{E}\left[X_{k:n}^3\right]\mu + 6\mathrm{E}\left[X_{k:n}^2\right]\mu^2 - 3\mu^4.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathrm{E}\left[X_{k:n}^2\right] = \mu\cdot\frac{k+1}{n+2} &= \mu^2 + \mu\left(\frac{k+1}{n+2} - \frac{k}{n+1}\right)\\
&= \mu^2 + \mu\cdot\frac{n-k+1}{(n+1)(n+2)}.
\end{aligned}$$

Therefore, because $n-k+1 \leqslant n+1$,

$$\mathrm{Var}X_{k:n} = \mu\cdot\frac{n-k+1}{(n+1)(n+2)} \leqslant \frac{\mu}{n+2} \leqslant \frac{n}{(n+1)(n+2)} \approx \frac{1}{n}, \tag{4}$$

when $n$ is large. This shows that $X_{k:n} \approx \mathrm{E}X_{k:n}$ with high probability, thanks to the Chebyshev inequality. We need a slightly better estimate for our purposes. With that in mind, let us note that

$$\begin{aligned}
\mathrm{E}\left[(X_{k:n}-\mu)^4\right] &= \mathrm{E}\left[X_{k:n}^4\right] - 4\mathrm{E}\left[X_{k:n}^3\right]\mu + 6\mathrm{E}\left[X_{k:n}^2\right]\mu^2 - 4\mu^4 + \mu^4\\
&= \mathrm{E}\left[X_{k:n}^4\right] - 4\mathrm{E}\left[X_{k:n}^3\right]\mu + 6\mathrm{E}\left[X_{k:n}^2\right]\mu^2 - 3\mu^4\\
&= \left(\frac{k}{n+1}\cdot\frac{k+1}{n+2}\cdot\frac{k+2}{n+3}\cdot\frac{k+3}{n+4}\right) - 4\left(\frac{k}{n+1}\cdot\frac{k+1}{n+2}\cdot\frac{k+2}{n+3}\cdot\frac{k}{n+1}\right)\\
&\quad + 6\left(\frac{k}{n+1}\cdot\frac{k+1}{n+2}\cdot\frac{k}{n+1}\cdot\frac{k}{n+1}\right) - 3\left(\frac{k}{n+1}\cdot\frac{k}{n+1}\cdot\frac{k}{n+1}\cdot\frac{k}{n+1}\right).
\end{aligned}$$

We factor a $k/(n+1)$ from the entire expression and use the fact that $k/(n+1) \leqslant n/(n+1) \leqslant 1$ to see that

$$\begin{aligned}
\mathrm{E}\left[(X_{k:n}-\mu)^4\right] &\leqslant \left(\frac{k+1}{n+2}\cdot\frac{k+2}{n+3}\cdot\frac{k+3}{n+4}\right) - 4\left(\frac{k+1}{n+2}\cdot\frac{k+2}{n+3}\cdot\frac{k}{n+1}\right)\\
&\qquad + 6\left(\frac{k+1}{n+2}\cdot\frac{k}{n+1}\cdot\frac{k}{n+1}\right) - 3\left(\frac{k}{n+1}\cdot\frac{k}{n+1}\cdot\frac{k}{n+1}\right)\\
&\leqslant \frac{\text{constant}}{n^2},
\end{aligned}$$

after some painful (but otherwise direct) computations. The "constant" here does not depend on $k$ nor on $n$. [This is because $k \leqslant n$.] Consequently, by the Boole and Chebyshev inequalities,

$$\begin{aligned}
\mathrm{P}\left\{\max_{1\leqslant k\leqslant n}|X_{k:n}-\mathrm{E}X_{k:n}| \geqslant \varepsilon\right\} &= \mathrm{P}\left(\bigcup_{k=1}^{n}\left\{|X_{k:n}-\mathrm{E}X_{k:n}| \geqslant \varepsilon\right\}\right)\\
&\leqslant \sum_{k=1}^{n}\mathrm{P}\left\{|X_{k:n}-\mathrm{E}X_{k:n}| \geqslant \varepsilon\right\}\\
&\leqslant \frac{\text{constant}}{n\varepsilon^4}.
\end{aligned}$$

Therefore, $\max_{1\leqslant k\leqslant n}|X_{k:n}-\mathrm{E}X_{k:n}| \xrightarrow{\mathrm{P}} 0$. But

$$\left|\mathrm{E}X_{k:n}-\frac{k}{n}\right| = k\left|\frac{1}{n+1}-\frac{1}{n}\right| = \frac{k}{n(n+1)} \leqslant \frac{1}{n}.$$

Therefore, $\max_{1\leqslant k\leqslant n}|(k/n)-\mathrm{E}X_{k:n}| \to 0$, whence Theorem 1 follows. □

## 2 Confidence Intervals and Tests at a Point

We wish to describe asymptotic $(1-\alpha)$-level confidence intervals for $F(x)$ for a given $x \in \mathbf{R}$. Recall that $n\hat{F}_n(x) \sim \text{Binomial}(n\,, F(x))$. Therefore, by the central limit theorem,

$$\frac{n\left[\hat{F}_n(x)-F(x)\right]}{\left(nF(x)\left[1-F(x)\right]\right)^{1/2}} \xrightarrow{d} \mathrm{N}(0\,,1).$$

Also, $\hat{F}_n(x) \xrightarrow{\mathrm{P}} F(x)$. Thus, Slutsky's theorem implies that as $n\to\infty$,

$$n^{1/2}\,\frac{\hat{F}_n(x)-F(x)}{\left(\hat{F}_n(x)\left[1-\hat{F}_n(x)\right]\right)^{1/2}} \xrightarrow{d} \mathrm{N}(0\,,1).$$

Thus, an asymptotic $(1-\alpha)$-level confidence interval for $F(x)$ is

$$\mathscr{C}_n(\alpha) := \left[ \hat{F}_n(x) - z_{\alpha/2}\sqrt{\frac{\hat{F}_n(x)\big[1-\hat{F}_n(x)\big]}{n}}\, , \hat{F}_n(x) + z_{\alpha/2}\sqrt{\frac{\hat{F}_n(x)\big[1-\hat{F}_n(x)\big]}{n}} \right].$$

Likewise, suppose we were to test

$$H_0:\ F(x) = F_0(x) \quad \text{versus} \quad H_1:\ F(x) \neq F_0(x),$$

where $F_0$ is a known, fixed distribution function. Then, an asymptotic $(1-\alpha)$-level test can be based on rejecting $H_0$ if and only if

$$\frac{\left|\hat{F}_n(x) - F_0(x)\right|}{\sqrt{F_0(x)\big[1-F_0(x)\big]}} > \frac{z_{\alpha/2}}{\sqrt{n}}.$$

**Something to think about:** How would you find a sensible, asymptotic level $(1-\alpha)$ confidence interval for $\mathrm{P}\{a < X_1 \leqslant b\}$, where $a < b$ are known and fixed? Also, how would you test $H_0 : \mathrm{P}\{a < X_1 \leqslant b\} = F_0(b) - F_0(a)$ versus $H_1 :\ \mathrm{P}\{a < X_1 \leqslant b\} \neq F_0(b) - F_0(a)$ for a known $F_0$?

# 3 Empirical-Process Theory

Now consider the more realistic problem of finding simultaneous confidence intervals for $F(x)$, simultaneously over all $x \in \mathbf{R}$. Or suppose we wish to test $H_0 :\ F = F_0$ versus $H_1 :\ F \neq F_0$, where $F_0$ is known. [These are essentially the same problem.]

Suppose we knew the exact distribution of $D_n(F_0) := \max_x |\hat{F}_n(x) - F_0(x)|$. Then, we can find $\delta_\alpha(n)$ such that

$$\mathrm{P}_F\{D_n(F) \leqslant \delta_\alpha(n)\} \geqslant 1 - \alpha, \tag{5}$$

for all $F$. Now consider the confidence interval

$$\mathbf{C}_n(\alpha) := \{F :\ D_n(F) \leqslant \delta_\alpha(n)\}.$$

This has level $1-\alpha$. Note that $\delta_\alpha(n)$ does not depend on $F$, because of the distribution-free nature of $D_n$! Moreover, $\delta_\alpha(n)$ can be computed by simulation: Without loss of generality, assume that $F$ is the Unif$(0\,,1)$ distribution function. In this case, $D_n = \max_{1\leqslant j\leqslant n} |X_{j:n} - (j/n)|$, whose distribution can be simulated by Monte-Carlo. [This will be the next Project.]

However, for theoretical purposes, it may help to be able to construct an asymptotic level-$(1-\alpha)$ confidence interval. This is helpful also when $n$ is large. [See equation (7) on page 14 below, for an example.] To do so, we need to know how fast $D_n$ converges to zero.

In order to understand this question, we assume that $X_1, \ldots, X_n$ are distributed as Unif$(0\,,1)$, so that $F(x) = x$ forall $0 \leqslant x \leqslant 1$. And first consider the random vector

$$\begin{aligned}\sqrt{n}\begin{bmatrix}\hat{F}_n(x_1) - F(x_1)\\ \vdots \\ \hat{F}_n(x_k) - F(x_k)\end{bmatrix} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix}\mathbf{I}\{X_1 \leqslant x_1\} - F(x_1)\\ \vdots \\ \mathbf{I}\{X_1 \leqslant x_k\} - F(x_k)\end{bmatrix}\\ &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{Z}_i,\end{aligned}$$

where $x_1, \ldots, x_k$ are fixed. Evidently, $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots$ are i.i.d. $k$-dimensional random vectors with $\mathrm{E}\boldsymbol{Z}_1 = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{Z}_1) = \boldsymbol{Q}$, where

$$\boldsymbol{Q}_{i,i} = F(x_i)\{1 - F(x_i)\}, \qquad 1 \leqslant i \leqslant k,$$

and

$$\begin{aligned}\boldsymbol{Q}_{i,j} &= F(\min(x_i\,, x_j)) - F(x_i)F(x_j)\\ &= F(\min(x_i\,, x_j))\{1 - F(\max(x_i\,, x_j))\} \qquad 1 \leqslant i \neq j \leqslant k.\end{aligned}$$

By the multidimensional central limit theorem, $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{Z}_i$ converges in distribution to $N_k(\mathbf{0}\,, \boldsymbol{Q})$. In particular, under $F$,

$$\sqrt{n}\,\max_{1\leqslant i\leqslant k}\left|\hat{F}_n(x_i) - F(x_i)\right| \xrightarrow{d} \max_{1\leqslant i\leqslant k}|W_i|, \tag{6}$$

where $\boldsymbol{W} = (W_1\,, \ldots, W_k)' \sim N_k(\mathbf{0}\,, \boldsymbol{Q})$. Now choose $x_1, \ldots, x_k$ to be a very fine partition of $[0\,,1]$ to "see" that the left-hand side is very close to $\sqrt{n}D_n$. Therefore, one may surmise that $\sqrt{n}D_n$ converges in distribution. In order to guess the asymptotic limit, we need to understand the right-hand side better. This leads us to "Gaussian processes," particularly to "Brownian motion," and "Brownian bridge."

## 3.1 Gaussian Processes

Let $A$ be a set. A *Gaussian process* $G$, indexed by $A$, is a collection $\{G(x)\}_{x\in A}$ of Gaussian random variables such that for all $x_1, \ldots, x_k \in A$, the random vector $(G(x_1)\,, \ldots, G(x_k))'$ is multivariate normal. The function $m(x) := \mathrm{E}[G(x)]$ is the *mean function* and $C(x\,,y) := \mathrm{Cov}(G(x)\,, G(y))$ is the *covariance function*. A *centered* Gaussian process is one whose mean function is identically zero.

Note that for all $x, y \in A$,

$$\mathrm{E}\left[|G(x) - G(y)|^2\right] = C(x\,,x) + C(y\,,y) - 2C(x\,,y).$$

We will need the following "hard facts":

**Theorem 5** *The distribution of $G$ is determined by the functions $m$ and $C$.*

**Theorem 6** *Suppose $G$ is centered Gaussian and $A \subset \mathbf{R}$. Suppose also that there exist constants $K, \eta > 0$ such that for all $x, y \in A$*

$$\mathrm{E}\left[|G(x) - G(y)|^2\right] \leqslant K|x - y|^\eta.$$

*Then, with probability one, the random function $G$ is continuous.*

**Example 7 (Brownian Motion)** The *Brownian motion $B$* (also known as the *Wiener process* as well as the *Bachelier–Wiener process*) is a centered Gaussian process indexed by $[0\,,\infty)$, whose covariance function is given by

$$C(s\,,t) := \min(s\,,t), \qquad s, t \geqslant 0.$$

Note that whenever $0 \leqslant s \leqslant t$,

$$\begin{aligned}\mathrm{E}\left(|B(t) - B(s)|^2\right) &= s + t - 2\min(s\,,t)\\ &= |s - t|.\end{aligned}$$

Therefore, $B$ is a random continuous function.

**Example 8 (The Brownian Bridge)** The *Brownian bridge* is a centered Gaussian process $B^\circ$ indexed by $[0\,,1]$, whose covariance function is

$$C(s\,,t) = \min(s\,,t) - st.$$

Suppose $0 \leqslant s \leqslant t \leqslant 1$. Then,

$$\begin{aligned}\mathrm{E}\left(|B^\circ(t) - B^\circ(s)|^2\right) &= s - s^2 + t - t^2 - 2\left[\min(s\,,t) - st\right]\\ &= |t - s| - |t - s|^2\\ &\leqslant |t - s|.\end{aligned}$$

Therefore, $B^\circ$ is a continuous random function. In fact, the Brownian bridge is related to the Brownian motion in a nice way. Let $B$ denote Brownian motion, and define

$$b(t) = B(t) - tB(1), \qquad 0 \leqslant t \leqslant 1.$$

Linear combinations of multivariate normals are themselves multivariate normals. Therefore, $b$ is a centered Gaussian process indexed by $[0\,,1]$. Its covariance is computed as follows: If $0 \leqslant s, t \leqslant 1$, then

$$\begin{aligned}\text{Cov}(b(s)\,,b(t)) &= \text{E}[b(s)b(t)]\\ &= \text{E}\left[B(t)B(s)\right] - t\text{E}[B(t)B(1)] - s\text{E}[B(s)B(1)] + st\text{Var}B(1)\\ &= \min(s\,,t) - t\min(t\,,1) - s\min(s\,,1) + st\\ &= \min(s\,,t) - st.\end{aligned}$$

That is, $b$ is a Brownian bridge. Because $B$ is continuous, so is $b$, and therefore this gives an alternative proof that Brownian bridge is continuous.

**Example 9 (The OU Process)** This example is not needed in this course. However, it plays an important role in the classical theory of Gaussian processes and in particular Brownian motion; I would be remiss if I said nothing about this example.

Let $B$ be a Brownian motion, and define

$$U(t) := \text{e}^{-t/2} B\left(\text{e}^t\right) \qquad (t \geqslant 0).$$

The stochastic process $U$ is clearly a mean-zero Gaussian process. It covariance function can be computed as follows: If $t, s \geqslant 0$, then

$$\begin{aligned}\text{Cov}(U(s)\,,U(t)) &= \frac{1}{\text{e}^{(s+t)/2}}\text{Cov}\left(B\left(\text{e}^s\right), B\left(\text{e}^t\right)\right)\\ &= \frac{1}{\text{e}^{(s+t)/2}}\min\left(\text{e}^s\,,\text{e}^t\right)\\ &= \text{e}^{-|t-s|/2}.\end{aligned}$$

In order to see the last line, consider the two cases $s \geqslant t \geqslant 0$ and $0 \leqslant s < t$ separately.

The process $U$ is "stationary" in the sense that the joint distributions of $(U(t_1)\,,\ldots,U(t_k))$ and $(U(\tau+t_1)\,,\ldots,U(\tau+t_k))$ are the same for every $t_1,\ldots,t_k \geqslant 0$ and for all "shifts" $\tau \geqslant 0$. This property can be seen from inspecting the covariance function.

### 3.2 A Return to Empirical Processes

Once again, let $X_1, X_2, \ldots, X_n$ be i.i.d. Unif$(0\,,1)$ random variables. Now consider the random function,

$$E_n(x) := \sqrt{n}\left[\hat{F}_n(x) - x\right], \qquad 0 \leqslant x \leqslant 1.$$

Note that $\max_{0\leqslant x\leqslant 1} |E_n(x)| = D_n(F)$ is our friend, the Kolmogorov–Smirnov statistic. If we reconsider our derivation of (6) we find a proof of the following fact: For all $0 \leqslant x_1, \ldots, x_k \leqslant 1$,

$$\sqrt{n}\begin{bmatrix} E_n(x_1) \\ \vdots \\ E_n(x_k) \end{bmatrix} \overset{d}{\to} \begin{bmatrix} B^\circ(x_1) \\ \vdots \\ B^\circ(x_k) \end{bmatrix}.$$

In particular,

$$\sqrt{n}\max_{1\leqslant j\leqslant k} |E_n(x_j)| \overset{d}{\to} \max_{1\leqslant j\leqslant k} |B^\circ(x_j)|.$$

It turns out that a little more is true. Namely, that

$$\sqrt{n}\ D_n := \max_{0\leqslant x\leqslant 1} |E_n(x)| \overset{d}{\to} \max_{0\leqslant x\leqslant 1} |B^\circ(x)|\,.$$

The advantage here is that the distribution of the random variable on the right-hand side is known. [A lot is known about Brownian motion, and $B^\circ$ is related to the latter Gaussian process because we can think of $B^\circ$ as $B^\circ(t) = B(t) - tB(1)$.] As an example, we have

$$\mathrm{P}\left\{\max_{0\leqslant t\leqslant 1} |B^\circ(t)| > x\right\} = 2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2x^2}.$$

Thus, if $n$ is large then

$$\mathrm{P}\left\{D_n > \frac{x}{\sqrt{n}}\right\} \approx 2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2x^2} = 2\mathrm{e}^{-2x^2} - 2\mathrm{e}^{-8x^2} \pm \cdots.$$

The series converges absolutely, and rapidly. In fact, the first term in the infinite sum typically gives a very good approximation, for instance if $x \geqslant 1$. [For $x = 1$, the first term is $2\mathrm{e}^{-2} \approx 0.271$, whereas the second term is $2\mathrm{e}^{-8} \approx 0.00067$.]

For instance, in order to find an asymptotic level-$(1-\alpha)$ confidence interval for $F(t)$, simultaneously for all $t$, we need to find a number $x$ such that

$\mathrm{P}\{D_n > x/\sqrt{n}\} \approx \alpha$. [Earlier, this $x/\sqrt{n}$ was called $\delta_\alpha(n)$. See equation (5).] When $\alpha \leqslant 0.2$, a very good approximation then is given by $2\mathrm{e}^{-2x^2} \approx \alpha$. Solve for $x$ to find that

$$\delta_\alpha(n) \approx \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\alpha}\right)}, \qquad \text{for large } n \text{ and } 0 < \alpha \leqslant 0.2. \tag{7}$$

# 4 Tests of Normality

We can use the $\chi^2$-test of Pearson to test whether a certain data has the $N(\mu_0\,, \sigma_0^2)$ distribution, where $\mu_0$ and $\sigma_0$ are known. Now we wish to address the same problem, but in the more interesting case that $\mu_0$ and $\sigma_0$ are *unknown*. [For instance, you may wish to know whether or not you are allowed to use the usual homoskedasticity assumption in the usual measurement-error model of linear models.]

Here we discuss briefly some solutions to this important problem. A good way to go about it this: First try the quick-and-dirty solution (qq-plots). If this rejects normality, then your data is probably not normal. Else, try one or two more of the more sophisticated methods below. Used in combination, they provide a good picture of whether or not your data is normally distributed. Of course, you should *always* plot a histogram of your data, as well.

## 4.1 QQ-Plots: A Visual Test

Recall that $\Phi$ denotes the standard-normal distribution function. That is, for all $t \in (-\infty\,, \infty)$,

$$\Phi(t) := \int_{-\infty}^{t} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\, \mathrm{d}u.$$

**Lemma 10** *$X$ is normally distributed if and only if $\Phi^{-1}(F_X(t))$ is a linear function of $t$.*

**Proof:** Suppose $X \sim N(\mu_0\,, \sigma_0)$. Then for all $t \in (-\infty\,, \infty)$,

$$\begin{aligned} F_X(t) &= \mathrm{P}\{X \leqslant t\} \\ &= \mathrm{P}\left\{\frac{X - \mu_0}{\sigma_0} \leqslant \frac{t - \mu_0}{\sigma_0}\right\} \\ &= \Phi\left(\frac{t - \mu_0}{\sigma_0}\right). \end{aligned}$$

Equivalently,

$$\Phi^{-1}(F_X(t)) = \frac{t-\mu_0}{\sigma_0}.$$

We have shown that if $X$ is normal, then $\Phi^{-1}(F_X)$ is a linear function. The converse is also true, and holds for similar reasons. □

Next we recall the Kolmogorov–Smirnov statistic $D_n$. Suppose $F$ is continuous, and $\alpha \in (0\,,1)$ is fixed. Then we can find $c$ such that

$$\mathrm{P}_F\{D_n \leqslant c\} = 1-\alpha.$$

Therefore,

$$\mathrm{P}_F\left\{\hat{F}_n(t) - c \leqslant F(t) \leqslant \hat{F}_n(t) + c \quad \text{for all } t\right\} = 1-\alpha.$$

Because $\Phi$ is strictly increasing, so is $\Phi^{-1}$. Therefore,

$$\mathrm{P}_F\left\{\Phi^{-1}\left(\hat{F}_n(t) - c\right) \leqslant \Phi^{-1}(F(t)) \leqslant \Phi^{-1}\left(\hat{F}_n(t) + c\right) \quad \text{for all } t\right\} = 1-\alpha.$$

Now suppose we have plotted the curves of $\Phi^{-1}(\hat{F}_n(t) \pm c)$ and found that we cannot draw a straight line between the two curves (Lemma 10). Thus, we can reject "$H_0$ : normal" at level $1-\alpha$; else, we do not reject. [This is a *visual test.*] Note that it is not hard to plot the two curves because $\hat{F}_n$ is constant between $X_{i:n}$ and $X_{i+1:n}$. Therefore, we need only plot the $2n$ points: $(X_{i:n}\,, \Phi^{-1}((i/n) \pm c))$ for $i = 1, \ldots, n$.

An even simpler visual method is to just plot the curve $\Phi^{-1}(\hat{F}_n)$, and see if it looks like a straight line. We need to only plot things at the order statistics. Thus, we plot the $n$ points: $(X_{i:n}\,, \Phi^{-1}((i/n)))$ for $1 \leqslant i \leqslant n$. The resulting plot is called a *qq-plot*.

Figures 1 and 2 contain four distinct examples. I have used qq-plot in the prorgam environment "`R`." The image on the left-hand side of Figure 1 shows a simulation of $10,000$ standard normal random variables (in `R`, you type `x=rnorm(10000,0,1)`), and its qq-plot is drawn on typing `qqnorm(x)`. In a very strong sense, **this figure is the benchmark**.

The image on the right-hand side of Figure 1 shows a simulation of $10,000$ standard Cauchy random variables. That is, the density function is $f(x) = (1/\pi)(1+x^2)^{-1}$. This is done by typing `y=rcauchy(10000,0,1)`, and the resulting qq-plot is produced upon typing `qqnorm(y)`. We know that the Cauchy has much fatter tails than normals. For instance,

$$\mathrm{P}\{\text{Cauchy} > a\} = \frac{1}{\pi}\int_a^\infty \frac{\mathrm{d}x}{1+x^2} \sim \frac{1}{\pi a} \quad (\text{as } a \to \infty),$$

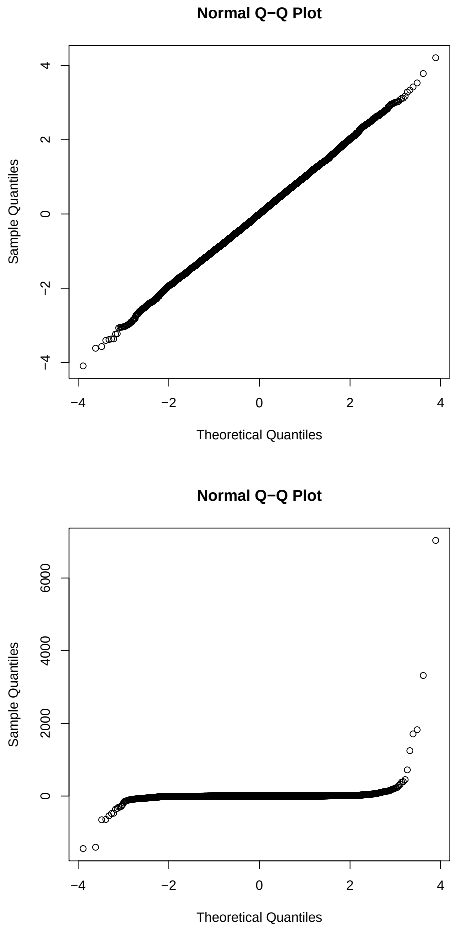

**Figure 1:** Left is N$(0\,,1)$ data. Right is Cauchy data

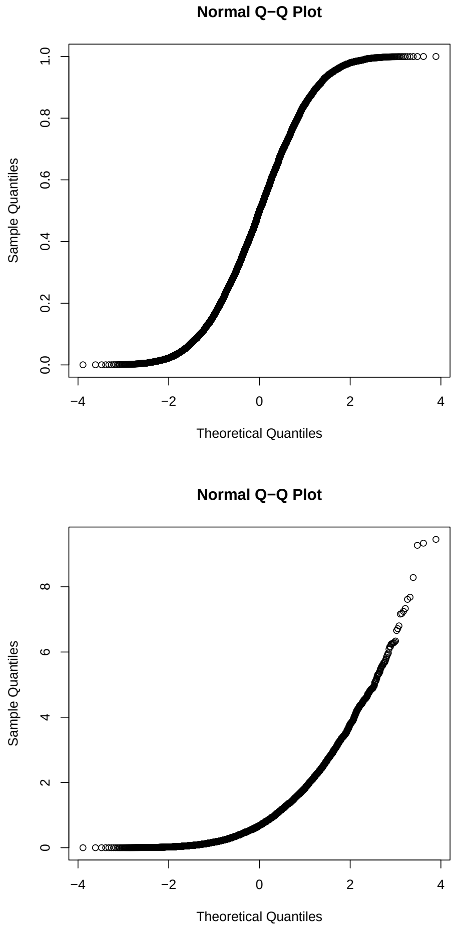

**Figure 2:** Left is the qq-plot of Unif$(0\,,1)$. Right is the qq-plot of a Gamma$(1\,,1)$.

whereas $\mathrm{P}\{\mathrm{N}(0\,,1) > a\}$ decays faster than exponentially. Therefore,

$$\mathrm{P}\{\mathrm{N}(0\,,1) > a\} \ll \mathrm{P}\{\mathrm{Cauchy} > a\} \quad \text{when } a \text{ is large.}$$

This heavy-tailedness can be read off in the right-figure in Figure 1: The Cauchy qq-plot grows faster than linearly on the right-hand side. *And this means that the standard Cauchy distribution has fatter right-tails.* Similar remarks apply to the left-tails.

The left-image in Figure 2 shows the result of the qq-plot of a simulation of $10,000$ iid Unif$(0\,,1)$ random variables. [To generate these uniform random variables you type, `rUnif(10000,0,1)`.]

Now Unif$(0\,,1)$ random variables have much smaller tails than normals because uniforms are in fact *bounded*. This fact manifests itself in the left-image of Figure 2. For instance, we can see that the right-tail of the qq-plot for Unif$(0\,,1)$ grows less rapidly than linearly. And this shows that the right-tail of a uniform is much smaller than that of a normal. Similar remarks apply to the left-tails.

A comparison of the three figures mentioned so far should give you a feeling for how sensitive qq-plots are to the effects of tails. [All three are from distributions that are symmetric about their median.] Finally, the right-most image in Figure 2 shows an example of $10,000$ Gamma random variables with $\alpha = \beta = 1$. You generate them in `R` by typing `x=rgamma(10000,1,1)`. Gamma distributions are inherently *asymmetric*. You can see this immediately in the qq-plot for Gammas; see the right-image in Figure 2. Because Gamma random variables are non-negative, the left-tail is much smaller than that of a normal. Hence, the left-tail of the qq-plot grows more slowly than linearly. The right-tail however is fatter. [This is always the case. However, for the sake of simplicity consider the special case where Gamma=Exponential.] This translates to the faster-than-linear growth of the right-tail of the corresponding qq-plot (right-image in Figure 2).

I have shown you Figures 1 and 2 in order to high-light the basic features of qq-plots in ideal settings. By "ideal" I mean "simulated data," of course.

Real data does not generally lead to such sleek plots. Nevertheless one learns a lot from simulated data, mainly because simulated data helps identify key issues without forcing us to have to deal with imperfections and other flaws.

But it is important to keep in mind that it is real data that we are ultimately after. And so the histogram and qq-plot of a certain real data set are depicted in Figure 3. Have a careful look and ask yourself a number of questions: Is the data normally distributed? Can you see how the shape of the histogram manifests itself in the shape and gaps of the qq-plot? Do the tails look like those of a normal distribution? To what extent is the ”gap” in the histogram “real”? By this I mean to ask what do you think might happen if we change the bin-size of the histogram in Figure 3?

### 4.2 A Non-Parametric Test

Another, more quantitative, test is based on a variant of the Kolmogorov–Smirnov statistic; see (1) for the latter. We are interested in “$H_0:$ normal,” and under the stated null hypothesis the Kolmogorov–Smirnov statistic becomes

$$D_n := \max_{-\infty<x<\infty} \left| \hat{F}_n(x) - \Phi\left(\frac{x-\mu}{\sigma}\right) \right|.$$

But $\mu$ and $\sigma$ are unknown. So we do the obvious thing and estimate them to obtain the modified KS-statistic,

$$D_n^* := \max_{-\infty<x<\infty} \left| \hat{F}_n(x) - \Phi\left(\frac{x-\bar{X}_n}{s_n}\right) \right|.$$

Here, $\bar{X}_n$ and $s_n$ are the usual estimates for the mean and SD:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad s_n^2 := \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2.$$

Note that if we replace $X_i$ by $(X_i-\mu)/\sigma$, then we do not change the value of $D_n^*$. Therefore, $D_n^*$ is distribution-free among all $N(\mu\,,\sigma^2)$ distributions. Hence, we can compute its distribution by simulation, using iid $\mathrm{N}(0\,,1)$’s.[3] When we wish to simulate $D_n^*$, it may help to recognize that

$$D_n^* := \max_{1\leqslant i\leqslant n} \max(A_i\,, B_i),$$

where

$$A_i := \left| \frac{i}{n} - \Phi\left(\frac{X_{i:n}-\bar{X}_n}{s_n}\right) \right|, \qquad \text{and} \qquad B_i := \left| \frac{i-1}{n} - \Phi\left(\frac{X_{i:n}-\bar{X}_n}{s_n}\right) \right|.$$

[3] For further literature on $D_n^*$ see J. Durbin (1973), *Distribution theory for Tests Based on the Sample Distribution Function*, Regional Conf. Series in Appl. Math. **9**, Society for Applied and Industrial Mathematics, Philadelphia, Pennsylvania.

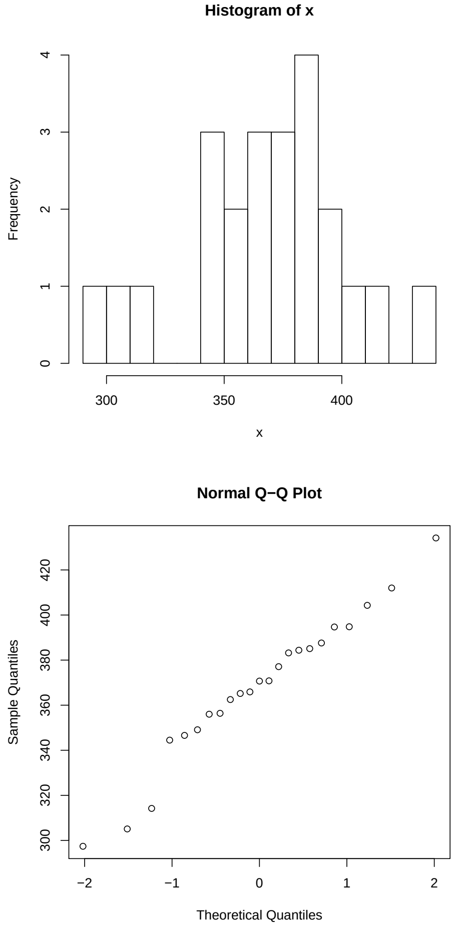

**Figure 3:** Histogram and qq-plot of data

### 4.3 A Test of Skewness and/or Kurtosis

Suppose $\mathrm{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2 > 0$, and $X$ has finite absolute fourth-moment. Define

$$\gamma_1 := \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\sigma^3} \qquad \text{and} \qquad \gamma_2 := \frac{\mathrm{E}\left[(X-\mu)^4\right]}{\sigma^4}.$$

If $X \sim N(\mu\,,\sigma^2)$ then $\gamma_1 = 0$ (zero *skewness*), and $\gamma_2 = 3$ (zero *kurtosis* $:= \gamma_2 - 3$). Therefore, normal data must have $\gamma_1 = 0$ and $\gamma_2 = 3$. These two parameters do not determine the distribution, but are particularly sensitive to the general "normal" shape of $F_X$. Their estimates are respectively,

$$\hat\gamma_1 := \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar X_n)^3}{s_n^3} \qquad \text{and} \qquad \hat\gamma_2 := \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar X_n)^4}{s_n^4}.$$

Note that both are consistent. Indeed, by the law of large numbers,

$$\hat\gamma_1 \overset{\mathrm{P}}{\to} \gamma_1 \qquad \text{and} \qquad \hat\gamma_2 \overset{\mathrm{P}}{\to} \gamma_2 \qquad \text{as } n \to \infty.$$

Also note that if we replace $X_i$ by $(X_i - \mu)/\sigma$, then $\hat\gamma_1$ and $\hat\gamma_2$ do not change. Therefore, we can simulate $\mathrm{N}(0\,,1)$'s to simulate the distribution of $\hat\gamma_1$ and $\hat\gamma_2$ respectively. In this way we can perform the test $H_0 :\ X \sim \mathrm{N}(0\,,1)$ by checking to see if $H_0 :\ \gamma_2 = 3$.

### 4.4 A Test Based on Correlation

There is another idea for testing "$H_0$ : Normal" that you should know about. Recall that the data is deemed to be normally distributed if and only if the qq-plot (large sample) looks linear. We can view the qq-plot as a scatterplot of two-dimensional data. Therefore, we can test to see if the correlation is high. A commonly-used statistic is the "$R^2$-statistic,"

$$R^2 := \frac{\left[\sum_{i=1}^n (X_{i:n} - \bar X_n)\Phi^{-1}\left(\frac{i-\frac12}{n}\right)\right]^2}{\sum_{i=1}^n (X_i - \bar X_n)^2 \cdot \sum_{j=1}^n \left[\Phi^{-1}\left(\frac{j-\frac12}{n}\right)\right]^2}.$$

[Why is this the square of a sample-correlation? If you want, you can pretend that $(i-\frac12)/n$ is $i/n$ here.] Under $H_0$ this $R^2$ statistic should be very close to one. Moreover, $0 \leqslant R^2 \leqslant 1$, thanks to the Cauchy–Schwarz inequality.

Note that if we replace $X_i$ by $(X_i - \mu)/\sigma$—where $\mu$ and $\sigma$ are the true mean and SD—we do not alter the value of $R^2$. Therefore, $R^2$ is distribution-free among all normals, and as such its distribution can be simulated. The form of the test is as follows: First, use simulation to create a table of probabilities for the distribution of $R^2$; then, find $c$ such that $\mathrm{P}\{R^2 \geqslant c\} = 1 - \alpha$. Finally, look at the sample-$R^2$. If it is less than $c$ then reject $H_0$ at $(1 - \alpha) \times 100\%$ level.

There are other, related, tests. A notable one is the one based on the so-called "Shapiro-Wilks" $W^2$ statistic. Here are some references to this and more in the literature:

- Shapiro and Wilks (1965). *Biometrika* **52**, 591–611
- Shapiro, Wilks, and Chen (1968). *J. Amer. Stat. Assoc.* **63**, 1343–1372
- Shapiro and Francia (1972). *J. Amer. Stat. Assoc.* **67**, 215–216
- Venter and de Wet (1972). *S. African Stat. J.* **6**, 135–149