Fitting time series to a Rayleigh distribution

(v.1 January 24, 2023)

Fernando Pérez Fontán, Vicente Pastoriza Santos and Fernando Machado Domínguez
University of Vigo. Spain

fpfontan@uvigo.es, vpastoriza@uvigo.es, fmachado@uvigo.es

`https://github.com/RadioPropagationChannel`

Contents:

1. Mobile communications propagation basics

We would like to briefly discuss some of the characteristics of the mobile/wireless propagation channel [1]. We will concentrate here on systems with base stations producing fairly large coverage areas such as those in so called **macrocells**. This is the classical case, however, cell sizes tend to be reduced in exchange for capacity and reduced interference while showing specific propagation characteristics. The modeling techniques involved in microcell land mobile systems (see elsewhere on this site) have also many similarities with those used in other **point-to-area systems** such as sound and TV broadcasting or fixed Internet access.

The similarities between fixed and mobile wireless channels over the frequency bands of interest (mostly UHF bands) not only include the mechanisms giving rise to path loss. They are also subjected to shadowing and multipath effects, even though these are normally much milder in the fixed systems case, especially because we have more choices as to the positioning of the terminal and use directive antennas.

Depending on the BS(Base Station)/AP (Access Point) height, cells of larger or smaller size can be created. The classical cellular environment with tall masts above rooftops gives rise, as said, to so-called **macrocells**. As the BS antenna height becomes smaller and goes below the surrounding rooftops, so-called **microcells** are generated. BSs within buildings give rise to **picocells**. Further, when satellites are used, which means much higher "BS antenna heights", **megacells** are originated.

In suburban areas, man-made structures [2] such as buildings or small houses, with sizes ranging from a few to tens of meters, significantly influence the wireless propagation channel. In urban areas, the size of structures can even be larger. Likewise, in rural and suburban environments, features such as isolated trees or groups of trees, etc. may reach similar dimensions. These features have the same or greater sizes than the transmitted wavelength (**metric**, **decimetric**, **centimetric waves** and now, **millimeter waves**) and may both block (diffraction) and scatter the radio signal causing specular and/or diffuse reflections. These contributions

may reach the mobile station, MS, (or, in more recent terminology, the user equipment, UE) by way of multiple paths, in addition to that of the direct signal. In many cases, these echoes make it possible that a sufficient amount of energy reaches the user terminal, thus making the communication link feasible. This is especially so when the direct signal is blocked. Hence, in addition to the expected distance dependent power decay, two main effects are present in mobile propagation: **shadowing/blockage** and **multipath**.

We can identify three different rates of change in the received signal as a function of the distance (spatial variations) between BS/AP and MS/UE, namely, **very slow variations** due to range, **slow** or **long-term variations** due to shadowing and **fast** or **short-term variations** due to multipath. In this chapter we address so-called narrowband effects which describe the channel's behavior around a single one frequency, e.g. the carrier. We treat wideband and spatial effects somewhere else on this site).

Coming back to the signal amplitude variations, time variations, when the terminal remains stationary also take place. These are of significance in fixed links where spatial variations are still present but we cannot observe them. In this fascicle, we concentrate here on position dependent variations.

While in conventional macrocells BS heights are in the order of 30 m or so, and are normally set on elevated sites with no or few blocking/scattering elements in their surroundings, MS antenna heights are usually smaller than those of local, natural and man-made features. Typical values range from 1.5 or so for handheld terminals to 3 m for vehicular terminals. For other large cell radio communication systems e.g. for TV broadcasting or fixed wireless access operating in the same frequency bands, the propagation channel will present a milder behavior given that, in these cases, the receive antennas are usually directive and are normally sited well above the ground and clear of near obstacles. Both the shadowing effect on the direct signal and the amount of multipath can be considerably reduced by carefully choosing the antenna location.

Other operating scenarios where both ends of the link are surrounded by obstacles are indoor communications where walls, the ceiling and floor or the various pieces of furniture will clearly determine the propagation conditions.

Two representative and extreme scenarios may be considered:

(a) the case where a strong direct signal is available together with a number of weaker multipath echoes, i.e., **line-of-sight** (LOS) conditions; and
(b) the case where a number of weak multipath echoes is received and no direct signal is available, **non-line-of-sight** (NLOS) conditions.

Case (a) occurs in open areas or in very specific spots in city centers, in places such as crossroads or large squares with a good visibility of BS. Sometimes, there might not be a direct LOS signal but a strong specular reflection off a smooth surface such as that of a large building could give rise to similar conditions. This situation may be modeled by a Rice distribution for the variations of the received RF signal envelope: **Rice case**. Under these conditions, the received signal will be strong and with moderate fluctuations (Figure 1.left). Note the reference used in this plot. Signal levels are referred to what we would have in LOS conditions and where no multipath were present (0 dB level). We present the Rice distribution somewhere else on this site.

Case (b) will typically be found in dense built-up urban environments. This is a worst-case scenario since the direct signal is completely blocked and the overall received signal is only due to multipath, thus being weaker and subjected to marked variations (Figure 1.right). Note that the received signal oscillates around levels well below that of the direct, LOS signal (0 dB in the plot). This kind of situation may also occur in rural environments where the signal is obstructed by dense masses of trees: wooded areas or tree alleys. The received signal's amplitude variations in this situation are normally modeled with a Rayleigh distribution: **Rayleigh case**. The Rayleigh distribution is presented in a section below.

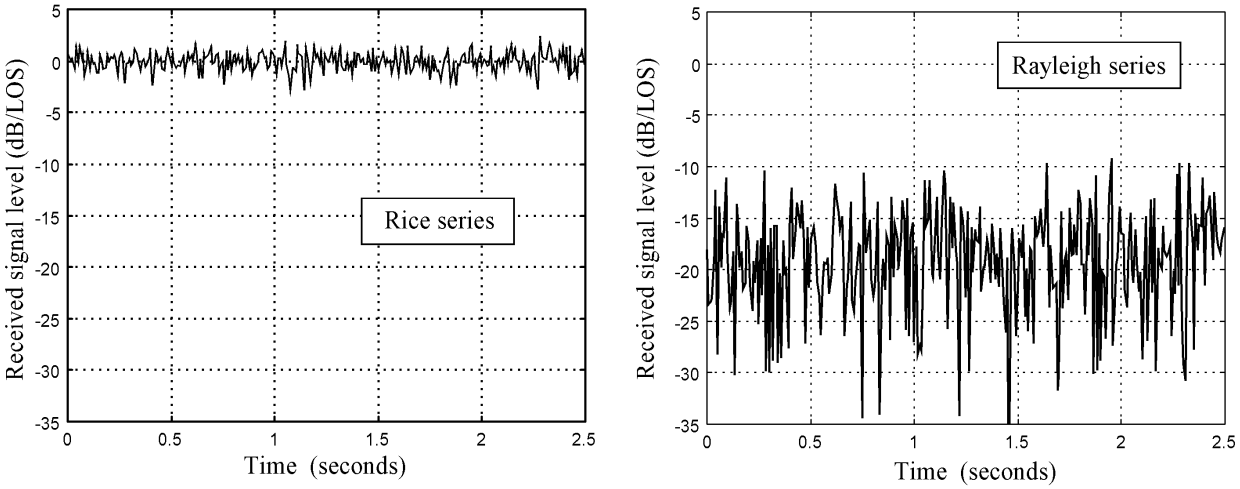


Figure 1 Rice and Rayleigh distributed time series. Frequency 900 MHz, mobile speed 10 m/s [3]

The received field strength, E, or the received voltage, v, may be represented in the time domain, $r(t)$, or in the traveled distance domain, $r(x)$. Figure 2 shows a typical mobile communications scenario with the MS driving away from BS along a radial route so that the link profile is the same as that of the terrain profile. The figure also shows a sketch of the received signal as a function of the distance from BS. In addition to the obvious distance dependent decay, the first thing to be noted is that the signal is subjected to strong oscillations as MS travels away from BS. On Figure 2 we illustrate the received signal variations as a function of distance and showing (dashed line). Note that, superposed, we will find the fast Rayleigh variations (solid line). This last component oscillates much faster than illustrated. Refer better Figure 1 to for a more realistic illustration.

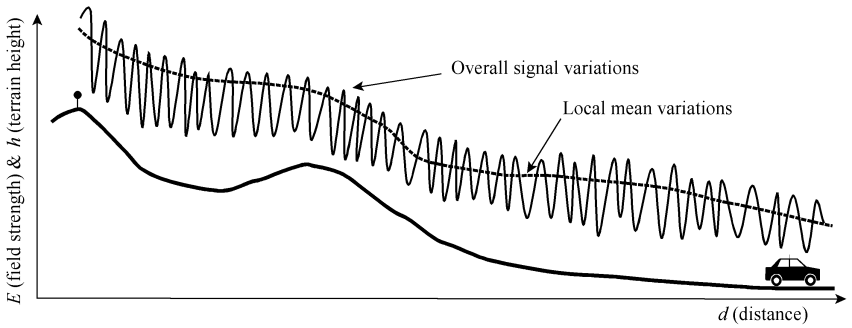


Figure 2 Variations in the received signal with the movement of the mobile [1].

For carrying out propagation channel measurements, the mobile speed, V, should preferably remain constant. Of course, there are ways around this. In such cases, the traversed distance needs to be recorded too. In our simulations, we will assume a constant MS speed. For a constant terminal speed, V, it is straightforward to make the conversion between the signal represented in the time domain, $r(t)$, and the signal represented in the (traveled) distance domain, $r(x)$ $(t = x/V)$.

Variable x may either be expressed in meters or in wavelengths. Based on such signal recordings plotted in the distance domain, it is possible to separate and study individually the fast and slow variations due, respectively, to multipath and blockage/shadowing, as illustrated in Figure 3. Note how the solid line in Figure 3.left corresponds to the slow or local mean variations. Figure 3.right shows the fast variations once separated from the overall shadowing-plus-multipath variations.

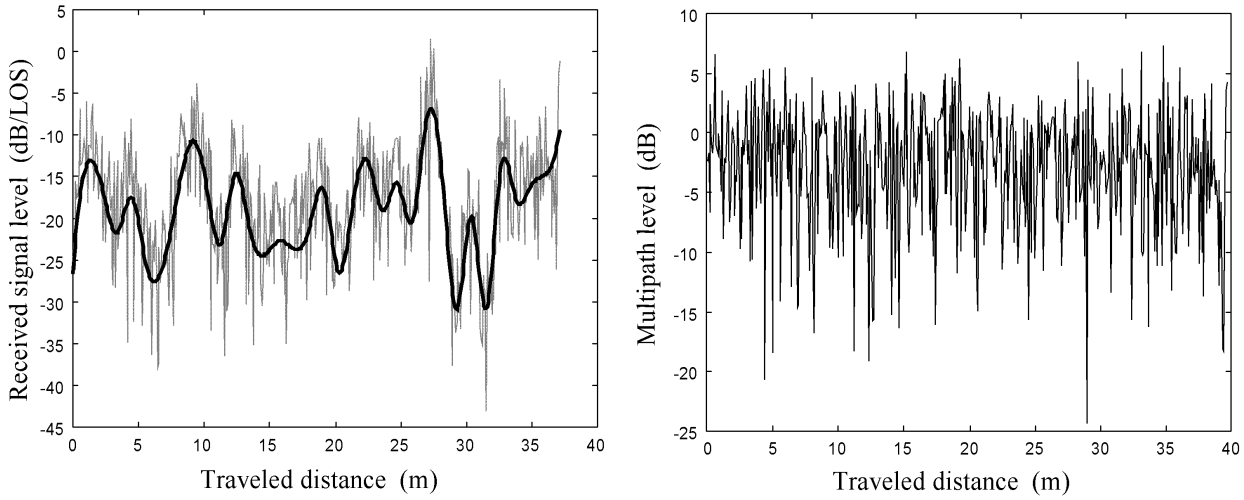


Figure 3 Overall and slow variations, and fast variations after removing the slow variations [3]

Generally, the received signal variations, $r(t)$ or $r(x)$, may be broken down, in a more or less artificial way, into two components [2],

- the slow or long-term variations: $m(t)$ or $m(x)$; and
- the fast or short-term variations: $r_0(t)$ or $r_0(x)$.

Remember that the above variables represent **voltages** or **electric fields**, not powers or power densities. In many cases, some sort of normalization is eventually made on those variables since their actual values are very small (e.g. micro volts). In this way, reasonable magnitudes around unity are used in the calculations.

The received signal may, therefore, be described as the product of these two terms,

$$r(t) = m(t)\, r_0(t) \text{ or, alternatively, } r(x) = m(x)\, r_0(x) \tag{1}$$

when expressed in linear units. In dB, the products become additions, i.e.,

$$R(t) = M(t) + R_0(t) \text{ or, alternatively, } R(x) = M(x) + R_0(x) \tag{2}$$

Note, for example, that $R(t) = 20 \log r(t)$ and so on.

With this approach, we are assuming that the fast variations are superposed on the slow variations. Figure 3.left illustrates a time-series where the slow variations are also plotted. The figure (plot on the right) also shows the fast variations after removing (filtering out) the slow variations. The slow variations can be extracted from the overall variations through low-pass filtering by, for example, computing a **running mean**. This is equivalent to calculating the signal average for the samples within a route section of length $2L$ equal to some tens of wavelengths.

As will be shown later, we preform our normalization with respect to the average local power $\bar{p}(x_i)$ where x_i indicates the center position of the averaged samples. In this case

$$\bar{p}(x_i) = \frac{\sum_{k=-N}^{N} p_{i+k}}{2N+1} \quad \text{with} \quad r_{i-N} \cdots r_i \cdots r_{i+N} \in x_i - L < x < x_i + L \tag{3}$$

Note that the above operation, a running mean, is equivalent to a low pass FIR (Finite Impulse Response) filter with $2N+1$ identical coefficients of magnitude $1/(2N+1)$.

Then we convert the average power to $\bar{p}(x_i)$ to "average" voltage actually "reference voltage, by making

$$v_{\mathrm{Ref}}(x_i) = \sqrt{\bar{p}(x_i)}.) \tag{4}$$

Typically, lengths of 10λ to 40λ are used [2]. For example, for the 2 GHz band ($\lambda = 0.15$ m,) the averaging length would be $2L \approx 3-6$ m. The average value, $\bar{p}(x_i)$, computed for a given route position, x_i, is usually called the **local mean** at x_i.

In dense urban areas, it has been observed experimentally [2] that the slow variations of the received signal, that is, the variations of the local mean, $\bar{p}(x_i)$ and of $\boldsymbol{v}_{\mathbf{Ref}}(x_i)$, follow a **lognormal distribution** (to be presented elsewhere) when expressed in linear units (V, V/m, ...) or, alternatively, a **normal distribution** when expressed in logarithmic units, $M(x_i) = 10\log\bar{p}(x_i) = 20\log\boldsymbol{v}_{\mathbf{Ref}}(x_i)$. Note how, with our normalization, $M(x_i)$ is the same whether we work with voltages or powers.

The length, $2L$, of route considered for the computation of the local mean, i.e., used to separate out the fast from the slow variations, is usually called a **small area** or **local area**. It is within a small area where the fast variations of the received signal are studied since they can be described there with well-known distributions e.g., Rayleigh. Over longer distances, the Rayleigh parameter varies, i.e., it is no longer is constant.

Over longer sections of route, we need a combination of distributions, e.g. the Suzuki distribution [4]: Rayleigh plus Lognormal.

Over longer route sections ranging from 50 m or 100 m to even 1 km, we can characterize the variations of the local mean. This extended surface is usually called a **larger area** or sometimes a **sector**. Typically, standard propagation models do not attempt to predict the fast signal variations. Instead they predict the mean, $\overline{M}(x)$, the average power expressed in dB units and the standard deviation (or **location variability**), σ_L, of $M(x_i)$, that is the two parameters of a Gaussian distribution. Further, we can characterized the spatial variability of $M(x_i)$, more specifically its rate of change, by means of the **correlation distance**, ρ_L, within the **larger area**.

Before low-pass filtering, the very slow variations due to the radio path range (also called **path loss**) must be removed. The **free-space loss**, L_{FS}(dB), is a very common model for the range-dependent loss. The free-space loss gives rise to a distance decay in the received power following an inverse power law with exponent $n = 2$ (Figure 4), i.e.,

$$p_1 \propto \frac{1}{d_1^2} \quad \text{and} \quad p_2 \propto \frac{1}{d_2^2}, \qquad \text{and, in dB,} \qquad \Delta_p = 10\log\frac{p_2}{p_1} = 20\log\frac{d_1}{d_2} \tag{5}$$

where $\propto$ indicates "proportional to". Parameter Δ_p is the difference in received power expressed in dB at distances d_1 and d_2. The above expressions show a 20 dB/decade (20 dB decrease when the distance is multiplied by 10) or 6 dB/octave (6 dB decrease when the distance is doubled) distance decay rate.

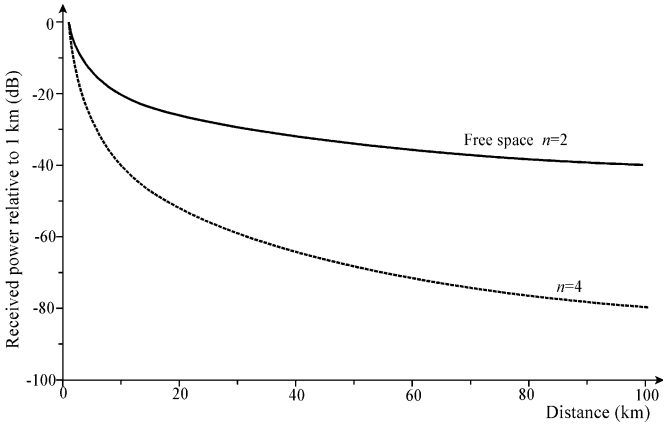


Figure 4 Received signal decay with distance: $n = 2$ and $n = 4$ laws.

These variations are first quite steep while, later, they show a gentler decay for greater distances. For example, variations are first quite steep while, later, they show a gentler decay for greater distances. For example, using **¡Error! No se encuentra el origen de la referencia.**), from km 1 to km 2, a 6 dB decrease takes place. However, the same 6 dB reduction is observed from km 10 to km 20.

It has been experimentally verified that, in typical large cell mobile propagation paths, the signal's distance decay does not follow a $n = 2$ power law (as in free space conditions) but, rather, it presents a higher exponent. Signal decay is usually modeled by a $l \propto d^n$ law, i.e., l being proportional to the distance risen to the power of n**.** The values of n are typically somewhere near 4, i.e., 40 dB/decade (Figure 4). Well known models such as that of Hata [5] predict exponents close to 4.

The path-loss expressions normally provided by propagation models are of the form

$$L\ (\mathrm{dB}) = A + B\log(d_{\mathrm{km}})\ = A + 10n\log(d_{\mathrm{km}}) \tag{6}$$

where A and n are dependent on the frequency and a number of other factors as listed below. Parameter A is the loss at a **reference distance**, in this case, 1 km, and n is the propagation decay law.

Several factors, apart from the frequency and the distance that influence path loss are taken into consideration by existing propagation models affecting the expressions for A and n. These factors are:

- the height of the MS/UE antenna;
- the height of the BS relative to the surrounding terrain (**effective height**);
- the **terrain irregularity** (sometimes called **undulation**, $\Delta\mathrm{h}$, or **roughness**, σ_t);
- the type of **land use** (clutter) in the surroundings of MS: urban, suburban, rural, open, etc.

When calculating the path loss, all such factors must be taken into account, i.e.,

$$L\ (\mathrm{dB}) = L_{\mathrm{Ref}} + L_{\mathrm{Terrain}} + L_{\mathrm{Clutter}} \tag{7}$$

The **path loss** is defined between isotropic antennas (with 0 dB gain) for a given distance, d. Isotropic antennas do not exist in practice but are commonly used in **link budget** calculations since they allow the definition of the propagation loss independently of the antennas used in the actual link. Then, when computing the link budget, the actual antenna gains must be introduced in the calculations.

The **path loss** is normally made up of three main components: a *reference loss*, typically the free-space loss, although some models like Hata's [5] use a different value. Other models use the so-called **plane-earth loss** ($n = 4$) as their reference. There is a fascicle in this site where we simulate how empirical propagation models are developed.

The second component is the loss due to **terrain irregularity** and, finally, the third component is the loss due to the **local clutter** or **local environment** where the additional loss will very much depend on the land use in the vicinity of MS: urban, suburban, rural, open, woodland, etc.

2. Common statistical distributions used in radio

Here, we provide a brief reminder of some very basic statistical definitions. A random variable, X, has a **probability density function** (pdf), f, which is non-negative and fulfils that

$$\mathrm{Prob}(a \le X \le b) = \int_a^b f(x)dx \tag{8}$$

Its associated cumulative distribution function (CDF) of X, F, is

$$F(x) = \text{Prob}(-\infty \leq X \leq x) = \int_{-\infty}^{x} f(u)du \tag{9}$$

that is, it gives the probability of a given value x not being exceeded. Sometimes, it is more convenient to use the complementary CDF, CCDF, which gives the probability that a given value x is exceeded, i.e.,

$$\text{Prob}(X < x) = 1 - F(x) \tag{10}$$

Furthermore, if f is continuous at x, then

$$f(x) = \frac{d}{dx}F(x) \tag{11}$$

We can picture the product $f(x)dx$ as the probability of X falling within the **infinitesimal interval** $[x,\ x\ +\ dx]$. A pdf fulfils the following property,

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{12}$$

that is, the area under the pdf curve is unity.

Other parameters are the statistical **moments** of x, one is the **mean** is given by

$$m_1 = \int_{-\infty}^{\infty} xf(x)dx = \bar{x} \tag{13}$$

another is **mean square value**

$$m_2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \overline{x^2} \tag{14}$$

The square root of m_2 is called the **rms value** (**root mean square**). Finally, the variance of x is

$$\sigma^2 = m_2 - m_1^2 \tag{15}$$

where σ is the **standard deviation**, and the **median value**, $\tilde{x}$, is given by

$$\int_{-\infty}^{\tilde{x}} f(x)dx = \frac{1}{2} \tag{16}$$

that is, the value not exceeded 50% of the cases.

Another parameter is the **mode** or most probable value, located at the maximum of the pdf. In Figure 5 a Rayleigh distribution pdf and CDF are illustrated together with its moments.

3. The Rayleigh distribution

Signal variations caused by multipath, in the case where the direct signal is assumed to be totally blocked, are usually represented by a Rayleigh distribution when expressed in units of voltage. In addition, if the voltage is Rayleigh distributed then the associated power follows an *exponential distribution*. The *probability density function*, pdf, of the Rayleigh distribution is given by

$$f(r) = \frac{r}{q^2}\exp\left(-\frac{r^2}{2q^2}\right) \qquad \text{for } r \geq 0 \tag{17}$$

where r is a voltage This distribution has a single parameter, its *mode* or *modal value*, q. Other related parameters are given in Table 1 as a function of the mode. In this chapter, we use q to designate the modal value in this distribution.

Script **`Rayleigh_pdf_cdf`** is provided, which was used for plotting the Rayleigh pdf and CDF in Figure 5 for $q = 1$.

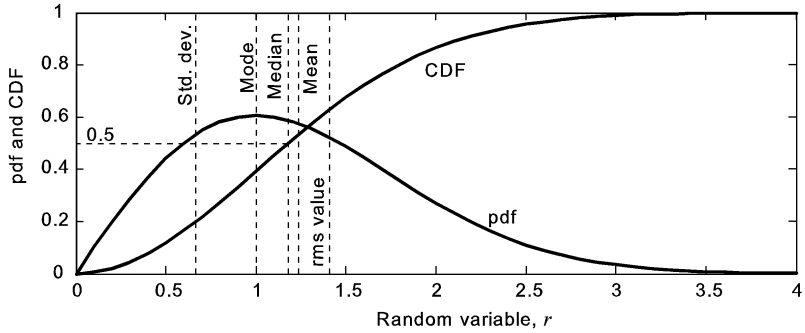


Figure 5 Probability density function and cumulative distribution function for a Rayleigh distribution with σ=1. Generated with `Rayleigh_pdf_cdf`.

By integrating the pdf, the **cumulative distribution function**, CDF, can be obtained, i.e.,

$$\mathrm{CDF}(R) = \mathrm{Prob}(r \le R) = \int_0^R f(r)dr = 1 - \exp\left(-\frac{R^2}{2q^2}\right) \tag{18}$$

Table 1 Rayleigh distribution parameters as a function of its mode [6]

Mode	q
Median	$q\sqrt{2\ln 2} = 1.18q$
Mean	$q\sqrt{\pi/2} = 1.25q$
RMS value	$q\sqrt{2} = 1.41q$
Standard deviation	$q\sqrt{2-\pi/2} = 0.655q$

The CDF is very useful when computing *outage probabilities* and *link margins*. The CDF gives the probability that a given signal level is not exceeded. If this level is the system's *operation threshold*, this provides us with the probability that the signal level is equal or below such threshold, i.e., the *outage probability*. Knowing the CDF, adequate *fade margins* can also be set up.

The parameters in Table 1 are defined as follows,

$$\mathrm{mean}(r) = E\{r\} = \bar{r} = \int_{-\infty}^{\infty} rf(r)dr = \int_0^{\infty} rf(r)dr = q\sqrt{\pi/2} = 1.2533q \tag{19}$$

$$r_{\mathrm{rms}} = \sqrt{E\{r^2\}} = \sqrt{\overline{r^2}} = \sqrt{\int_0^{\infty} r^2 f(r)dr} = \sqrt{2q^2} \tag{20}$$

$$\mathrm{variance}(r) = E\{r^2\} - (E\{r\})^2 = q^2(4-\pi)/2 = 0.4292q^2 \tag{21}$$

$$1 - \exp\left(-\frac{\tilde{r}^2}{2q^2}\right) = 0.5, \qquad \text{thus,} \qquad \mathrm{median}(r) = \tilde{r} = \sqrt{2q^2\ln 2} = 1.1774q \tag{22}$$

where $E\{\cdot\}$ is the expectation operator.

For completeness, we show how the Rayleigh distribution can be expressed in terms of some of its parameters, other than its mode [7].

As a function of **the mean** the pdf and CDF are as follows,

$$f(r) = \frac{\pi\, r}{2\bar{r}^2}\exp\left(-\frac{\pi\, r^2}{4\bar{r}^2}\right) \quad \text{for } r \geq 0 \quad \text{and} \quad \mathrm{P}(R) = 1 - \exp\left(-\frac{\pi\, R^2}{4\bar{r}^2}\right) \tag{23}$$

where $\bar{r}$ is the **mean** of the distribution.

Put now as a function of the **mean square value**, r^2, the pdf and CDF have the form,

$$f(r) = \frac{2r}{\overline{r^2}}\exp\left(-\frac{r^2}{\overline{r^2}}\right) \quad \text{for } r \geq 0 \quad \text{and} \quad P(r \leq R) = 1 - \exp\left(-\frac{R^2}{\overline{r^2}}\right) \tag{24}$$

where $\sqrt{\overline{r^2}}$ is the **rms value**.

Now we briefly discuss in what context does the Rayleigh distribution normally appears in the frame of a mobile communications link. The usual case is that we are assuming a narrowband signal, here represented by a single RF frequency, that is, a continuous wave, CW.

The low-pass equivalent time-varying **channel frequency response**, CFR, for a wireless channel subjected to Rayleigh multipath propagation is the result the phasor sum of the various multipath contributions, where each term corresponds to one ray [8], that is,

$$\tilde{H}(t,f) = \sum_{n=0}^{N} \tilde{a}_n(t)\exp\{-j2\pi f\tau_i(t)\} \tag{25}$$

In the case of a CW (which could be used to approximate a narrowband signal transmission), we are only interested in the time behavior of the channel at and around the carrier, f_0 (or any other RF component of the signal). In this case, the result is a time-varying complex valued time-series, $\tilde{v}_{BB}(t)$, that is,

$$\tilde{v}_{BB}(t) = \tilde{H}(t,f_0) = \sum_{n=0}^{N} \tilde{a}_n(t)\exp\{-j2\pi f_0\tau_i(t)\} \tag{26}$$

where $\tilde{v}_{BB}$is a complex voltage (base band equivalent), and where we have the summation of N multipath contributions with time-varying complex amplitudes $\tilde{a}_n(t)$ and time-varying delays $\tau_i(t)$.

The magnitude of term $\tilde{v}_{BB}(t)$, i.e., $r(t) = |\tilde{v}_{BB}(t)|$ can be modeled using a Rayleigh distribution. Normally, for convenience, this series is normalized with respect to one of its parameters: mode, rms, mean, etc. as we discussed earlier.

Thus, above, we showed the case where we normalized the magnitude with respect to the modal value, i.e., $q = 1$ (Figure 5). We also showed different representations of this distribution in terms of the mean and the rms values.

If the envelope $r(t)$ of the base band equivalent voltage, $\tilde{v}_{BB}(t)$, is Rayleigh distributed, the same occurs if, instead of working with complex base band signals, we work with the RF (pass-band) voltages. The relationship between a signal in the low-pass and RF domains is as follows,

$$v_{\mathrm{RF}}(t) = \frac{1}{2}\tilde{v}_{BB}(t)\exp(j2\pi f_0 t) + \frac{1}{2}\tilde{v}^*_{BB}(t)\exp(-j2\pi f_0 t) \tag{27}$$

which becomes

$$v_{\mathrm{RF}}(t) = |\tilde{v}_{BB}(t)|\cos(2\pi f_0 t + \phi(t)) = r(t)\cos\{2\pi f_0 t + \phi(t)\} \tag{28}$$

This is illustrated in Figure 6. Note how $v_{\mathrm{RF}}(t)$ is real valued while $\tilde{v}_{BB}(t)$ is complex valued.

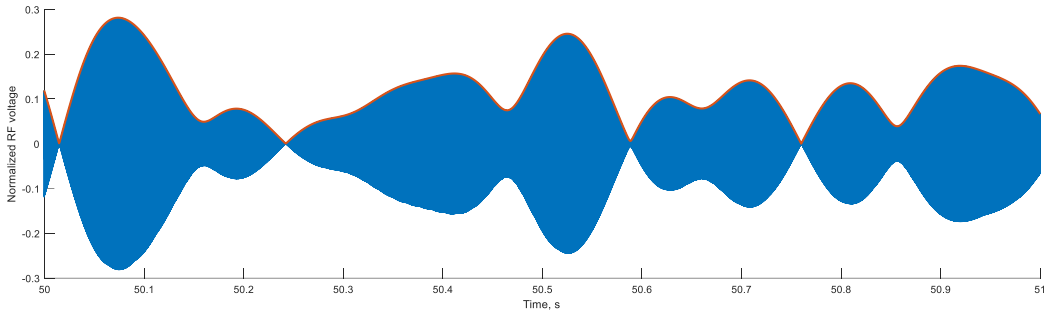


Figure 6 Illustration of a CW RF signal affected by Rayleigh fading. Note that the fading is much slower than the RF oscillations. Here this difference has been reduced for ease of representation purposes.

The way Figure 6 was generated and the description of the corresponding script are discussed somewhere else on this site. To help with the illustration we have chosen a very high Doppler by using a high terminal speed and a very low carrier frequency. Note the carrier, in blue, is much higher in variation rate than the Rayleigh fading, in red.

The term we are interested in modeling is, again, the same as in the baseband case, that is, $r(t) = |\tilde{v}_{BB}(t)|$. Now, to deal with convenient magnitudes around unity (not very small voltage values), we can normalize $r(t)$ so that we get reasonably high values.

A good way of performing the sough after normalization is to use as reference the average power, $\bar{p}_{\mathrm{RF}}$. If we assume a measured received time-series from a transmitted a CW at f_0, the **instantaneous** received power is given by

$$p_{\mathrm{RF}}(t) = v(t)^2/2 \tag{29}$$

In this case, we have assumed an impedance $R = 1$. A typical impedance value would be $R = 50\ \Omega$. In that case,

$$p_{\mathrm{RF}}(t) = \frac{v(t)^2}{2R} \tag{30}$$

We can perform the following normalization,

$$p_{\mathrm{norm}}(t) = \frac{p_{\mathrm{RF}}(t)}{\bar{p}_{\mathrm{RF}}} \tag{31}$$

We can also work with normalized voltages defined as follows,

$$v_{\mathrm{norm}}(t) = \sqrt{p_{\mathrm{norm}}(t)} \tag{32}$$

We want to verify whether $v_{\mathrm{norm}}(t)$ follows a Rayleigh distribution with pdf

$$f(v_{\mathrm{norm}}(t)) = \frac{v_{\mathrm{norm}}(t)}{q^2} \exp\left(-\frac{v_{\mathrm{norm}}(t)^2}{2q^2}\right) \tag{33}$$

Given the normalization we have used, $\overline{p_{\mathrm{norm}}}(t) = 1$, it turns out that $2q^2 = 1$, and thus,

$$f\big(v_{\mathrm{norm}}(t)\big) = 2\, v_{\mathrm{norm}}(t) \exp(-v_{\mathrm{norm}}(t)^2) \tag{34}$$

We can also analyze the behavior of the normalized power, $p_{\mathrm{norm}}(t)$. In case the voltage follows a Rayleigh distribution, the power follows an **exponential distribution** (see section below.)

The CDF for $2q^2 = 1$ becomes

$$P(v_{\mathrm{norm}} \leq U) = 1 - \exp(-U^2) \tag{35}$$

Finally, we illustrate how the Rayleigh pdf and CDF look like with the normalization contention we have taken.

These are illustrated in Figure 7. Note how the modal value is no longer unity, value that coincides now with the rms value. A summary of Rayleigh distribution parameters for this normalization is provided in Table 2.

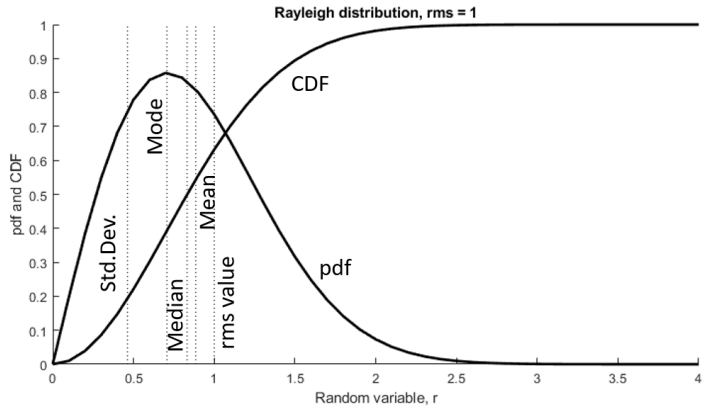


Figure 7 Rayleigh distribution normalized with respect to the rms level (rms = 1).

Table 2 Rayleigh distribution parameters as a function of its rms value

Mode	$s/\sqrt{2}$
Median	$s\sqrt{\ln 2}$
Mean	$s\sqrt{\pi}/2$
RMS value	s
Standard deviation	$s\sqrt{1-\pi/4}$

A note required at this point. We have calculated the power of the RF (pass band) signal and we have carried out a normalization for the magnitude (voltage) of the RF signal. The normalization would also be possible with respect to the complex, base band signal. In this case, the signal power is double that of the RF signal [8], however, since we normalize with respect to the average power, we arrive at the same Rayleigh distributed signal magnitude. We show this somewhere else on this site, where we also introduce the noise both at base band and RF.

4. The exponential distribution

Associated to the Rayleigh distribution is the exponential distribution. We can again come back to the normalization convention utilized above and study the distribution of $p_{\mathrm{norm}}(t) = v_{\mathrm{norm}}^2(t)$. In this case, p_{norm} is exponentially distributed with a pdf is given by

$$f(p_{\mathrm{norm}}) = \frac{1}{\Delta}\exp\left(-\frac{p_{\mathrm{norm}}}{\Delta}\right) \quad \text{for } p_{\mathrm{norm}} \geq 0 \tag{36}$$

where Δ is the mean value of the distribution. Incidentally, Δ is also its standard deviation of the distribution (Table 3). Since the normalization was done in such a way that $\Delta = 1$, then the pdf simplifies to (Figure 8)

$$f(p_{\mathrm{norm}}) = \exp(-p_{\mathrm{norm}}) \quad \text{for } p_{\mathrm{norm}} \geq 0 \tag{37}$$

And the CDF is given by

$$\mathrm{Prob}(p_{\mathrm{norm}} \leq U) = F(U) = 1 - \exp\left(-\frac{U}{\Delta}\right) \tag{38}$$

For $\Delta = 1$, we get

$$\mathrm{Prob}(p_{\mathrm{norm}} \leq U) = F(U) = 1 - \exp(-U) \tag{39}$$

Table 3 Statistics of the Exponential distribution

Mean	Δ
Median	$\Delta \ln 2$
Mode	0
Standard deviation	Δ
RMS value	$\Delta\sqrt{2}$

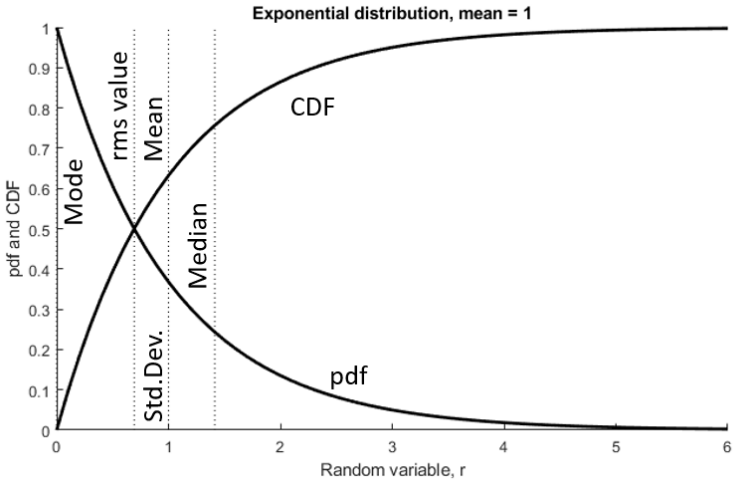


Figure 8 Exponential distribution: pdf and CDF plots (script `Exponential_pdf_cdf`)

5. Example: Fitting a series to the Rayleigh distribution

File **`RayleighSeries.mat`** (Figure 9) is supplied for analysis. It corresponds to a simulated signal (in dBm) assumed to be received under multipath conditions. This series, even though simulated, could just as well be a measured one. The only unrealistic feature is that we assume a very high signal-to-noise (SNR) ratio in such a way that there is not a distortion from the ideal case.

One possible way of recording a measured series could be in dB units, e.g., dBm (dB relative to 1 mW) as is the case here. This series could correspond to a recording with a spectrum analyzer or field strength meter. In other cases, the measured series could be given in terms of analog to digital converter (ADC) units which must be translated into voltage or power units. We are assuming here a CW transmission at 2 GHz.

File **`RayleighSeries`** contains two columns, the first is the time axis (variable **`time_axis`**) in seconds and the second, the power in dBm (variable **`PdBm`**). The time axis is sampled with an interval $t_s = 9.3750 \times 10^{-4}$ s, corresponding to a sampling frequency $f_S = 1/t_s = 1.0667$ kHz.

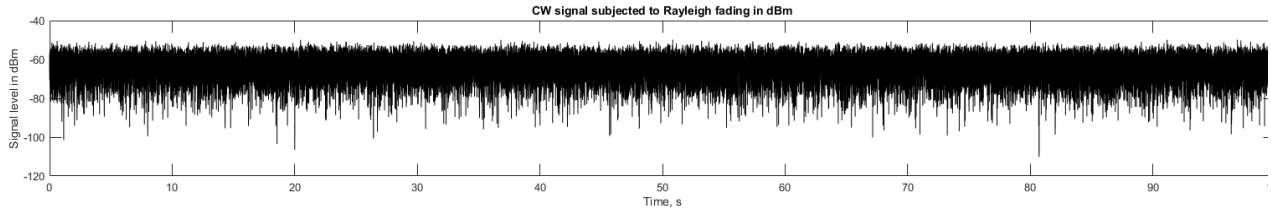


Figure 9 Representation of `RayleighSeries` processed with script `fitRayleigh`

What we will do in this example (script **`fitRayleigh`**) is analyze the contents of the file by computing its histogram (approximation of its pdf) and its sample CDF. Then, we will verify whether the series provided fits a Rayleigh distribution. The verification will be done very roughly by superposing the measured and theoretical pdfs and CDFs. We will quantify the comparison in a section below.

The series is plotted in dBm in Figure 9. What we want is to study the corresponding **normalized voltage**. A

matched load resistance, R, of 50 Ω is assumed. The instantaneous power can be computed by $p(\mathrm{W}) = 10^{P(\mathrm{dBm})/10}/1000$. The power and the voltage are linked through $v = \sqrt{2Rp}$. This voltage represents the instantaneous variations in the received signal envelope. It is like an amplitude modulation of the carrier's amplitude due to multipath propagation.

We start by loading in memory the file and reading the variables contained,

```
load RayleighSeries
timeAxis = time_axis;        % timeAxis in s
P = PdBm;                    % P in dBm
```

We follow the normalization steps discussed above. We create the following variables

```
p = 10.^(P/10);    % now p is in mW
p = p/1000;        % now p is in W

p_mean = mean(p);
p_norm = p/p_mean;      % normalize the power wrt its mean value

v = sqrt(2*50*p);       % Voltage in V
```

We do not use variable **v** but compute the normalized voltage as follows,

```
v_norm = sqrt(p_norm);
```

Making some verifications, we get

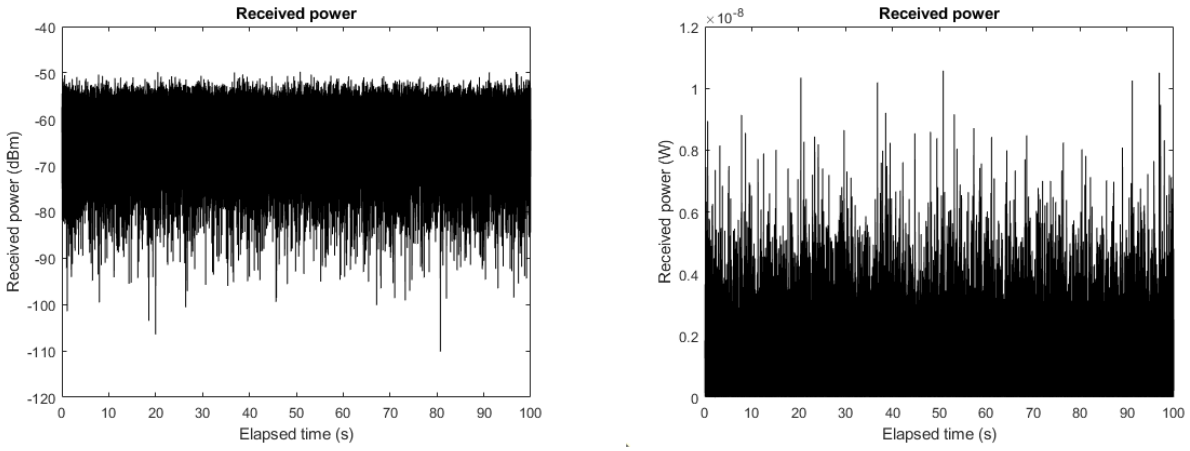
```
mean(p) =  9.9710e-10    % W
10*log10(mean(p)*1e3) = -60.0126 dBm
```

Note how we extracted the mean power value which turned out to be equal to **9.9710e-10** W and its value in logarithmic units is -60.0126 dBm. This value is useful when we compute link budgets. Also note that the average power was computed in the linear power domain and then, converted to dBm. Note that the obtained value is not the mean value of the series in dBm.

Over verifications regard the normalization process, i.e.,

```
mean(p_norm) = 1
mean(v_norm) = 0.8861
mean(v_norm.^2) = 1
```

Some of these intermediate results are reproduced in Figure 10. Note the original voltages are actually very small in the order of μV. When normalized they become magnitudes easier to handle around one.



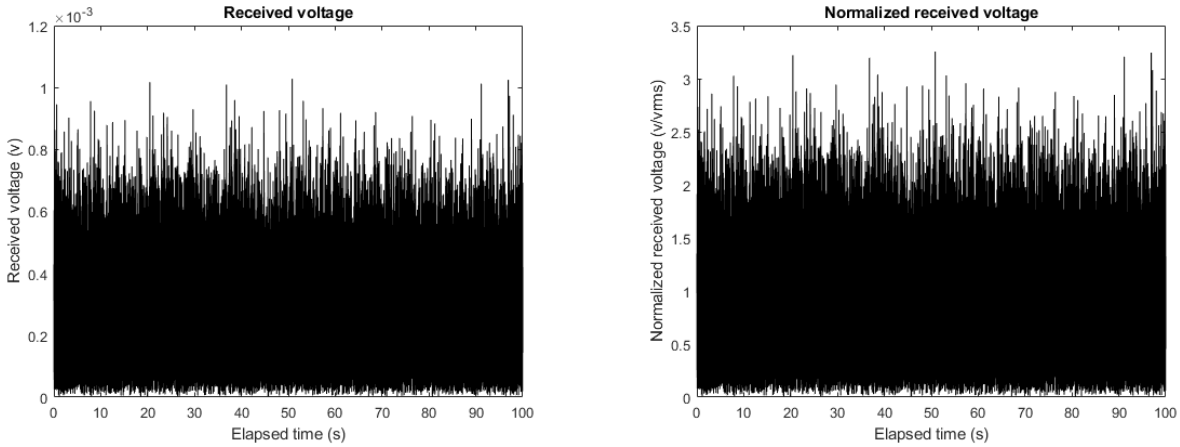


Figure 10 (top-left) Received signal power in logarithmic units (dBm). (top-right) Received signal power in linear power units (W). (bottom-left) Received signal converted to voltage (V). (bottom-right) Received voltage `v_norm` normalized with respect to its rms value.

Once the various magnitudes have been properly normalized, we can go on to obtain their experimental distributions and compare them with the theoretical ones. One important tool to be used when computing histograms is Matlab function **`hist`**. This function is no longer recommended by Matlab and the use of a more sophisticated function, **`histogram`**, is encouraged. Here we stick to using it.

This function splits the range of values of an input series into a number of separate, non-adjacent **intervals** (**bins**) and counts the number of occurrences in each. Then another interesting function is **`cumsum`** which computes the accumulative sum of a series. The sample pdf and CDF are computed using our function **`fpdfCDFbins`** while the theoretical CDF is computed using the formula (24) or (35) presented earlier for the Rayleigh distribution as a function of its rms value,

```
[pdfX, pdfY, CDFx,CDFy, stepp] = ...
    fpdfCDFbins(v_norm, 100);    % compute experimental pdf and CDF

rms = rmsValue;
CDFyTheoretical = 1 - exp(-CDFx.^2/rms^2);
```

To illustrate the computation of the pdf and CDF we list function **`fpdfCDFbins`** where we have set the number of bins to **`nBins`**. Usually a large number such as 100 is advisable. This is a good approach for computing CDFs while for pdfs we need to use a smaller number (e.g. 20) to prevent the appearance of bins with small number of occurrences next to others with many more. In the function we have the following code:

```
[a,b] = hist(z,nBins);
a = a/length(z);
a = [0 a 0];
step = b(2)-b(1);
pdfX = [0.0 b b(end)+step];
CDFx = pdfX;

pdfY = a;
CDFy = cumsum(a);
```

Note that the output of **`hist`** is a count of the number of occurrences, not a probability. To convert those into probabilities, we divide the counts by the total number of samples. The other output of **`hist`** is a vector with the **bin centers**. The **bin width** is given by the difference between the equally spaced bin centers. Finally, we perform the cumulative sum. We should start the curve at 0 and end at 1.

The experimental is computed again using function **`fpdfCDFbins`** but now with only 20 bins,

```
[pdfX, pdfY, CDFx,CDFy, step] = ...
    fpdfCDFbins(v_norm, 20); % compute experimental pdf and CDF
```

and compared with the theoretical pdf.

In order to compare the experimental pdf from the histogram with the theoretical pdf, we need to compare equivalent things. Since the histogram gives the number of counts in an interval (bin) which we then convert to probability while the theoretical distribution gives the probability in an infinitesimal interval, we have to perform the conversion below,

```
pdfRayTheor = (2*pdfX/rms^2).*exp(-pdfX.^2/rms^2);

pdfApprox = pdfRayTheor*step;
```

The above approximation performs a rough integration of the area below the pdf curve by assimilating it to a rectangle for each bin width.

A more correct way is performing the integration properly. In this case, it is easy since it means subtracting exponential functions,

$$\text{Prob}(R_1 < r < R_2) = \int_{R_1}^{R_2} f(r)dr = \exp\left(-\frac{R_2^2}{2q^2}\right) - \exp\left(-\frac{R_1^2}{2q^2}\right) \tag{40}$$

where, in this case, to have an rms value equal to 1, we need to make $q = 1/\sqrt{2}$.

We plotted the theoretical pdf curve together with the histogram to show that they are different things and cannot be compared directly. Finally, an approximation of the integral in (40) is performed approximately by multiplying the height of the pdf by the bin width as discussed above. The obtained results are shown in Figure 11.

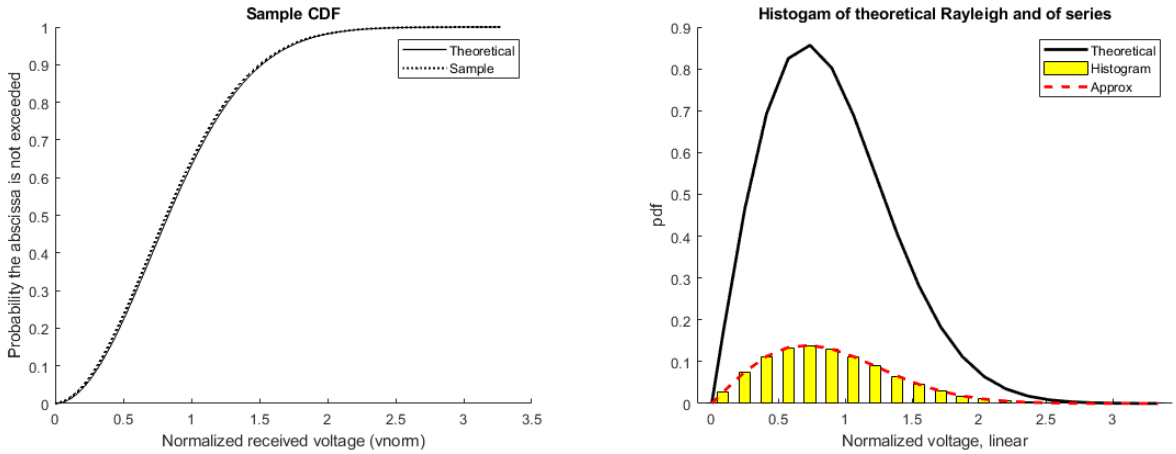


Figure 11 (left) Experimental and theoretical CDFs. (right) Theoretical pdf, and histogram and modified theoretical pdf.

We can also verify that the power is exponentially distributed. We can go ahead and repeat the procedure for the normalized power as follows (implemented in script **`fitExponential`**). The normalized power series and the results are illustrated in Figure 12 and Figure 13. As for the highlights of the code, we reproduce below some important steps which are very similar to those in the previous Rayleigh exercise. So for the CDF we used

```
[pdfX, pdfY, CDFx,CDFy, stepp] = ...
    fpdfCDFbins(p_norm, 100); % compute experimental pdf and CDF

CDFyTheoretical = 1 - exp(-CDFx/mean_p_norm);
```

while for the pdf we used less bins in the histogram, that is,

```
[pdfX, pdfY, CDFx,CDFy, step] = ...
    fpdfCDFbins(p_norm, 20); % compute experimental pdf and CDF
```

```
pdfRayTheor = (1/mean_p_norm).*exp(-pdfX/mean_p_norm);

pdfApprox = pdfRayTheor*step;
```

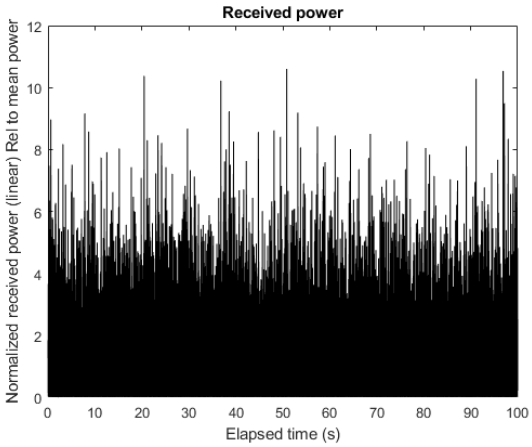


Figure 12 Normalized received power time series.

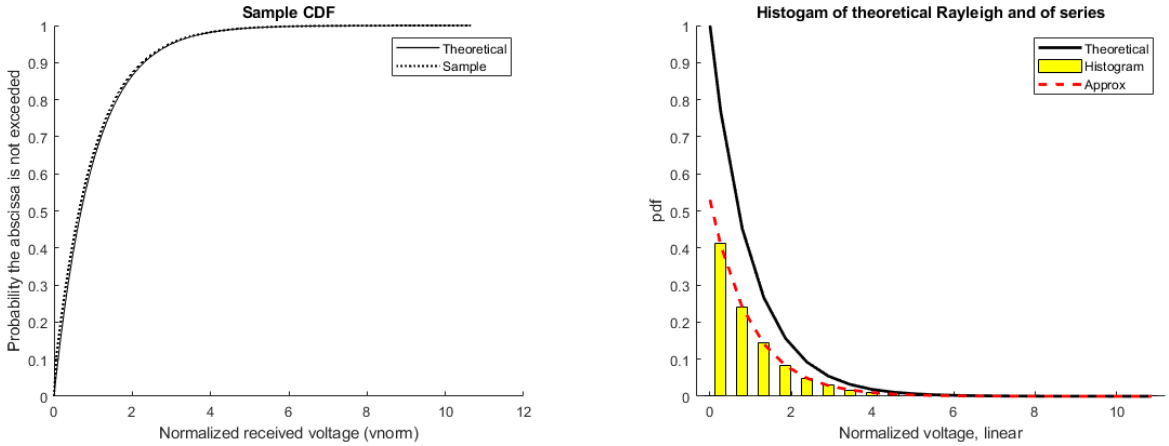


Figure 13 (left) Measured and theoretical CDFs of the normalized power. (right) Measured and theoretical pdfs of the normalized power.

6. The chi-square test

Script **`fitRayChiTest`** goes a step further from what we did in script **`fitRayleigh`**. We repeat most of what was done but the difference is that, before, we validated the fit on a qualitative way by comparing the experimental and theoretical pdfs and CDFs. We now try to perform the verification in an objective way. The purpose is whether to reject or not the premise that the series under analysis can be described by a Rayleigh distribution.

Here we continue with the previous Rayleigh example where our input series is contained in **`RayleighSeries.mat`**.

As said, from a visual comparison between the measured and the theoretical CDFs and histograms, it is clear that the agreement is quite good. Now, we want to **quantify how good the fit is**. This can be achieved by means of the **chi-square goodness-of-fit test** [9] (see also the Annex at the end.)

From the Annex, there are two basic elements in the chi-square test [9]. The chi-square distribution models the squared differences between a theoretical pdf and an experimental one. In Figure 14 we illustrate the pdf of a chi-square distribution. We can compare the sum of the mentioned squared differences to a threshold. If this sum is small, we can say the test is passed, if the sum is high, the test is failed, see details below.

First, we must define a **measure** of the difference between the values observed experimentally and the values that would be expected if the proposed pdf were correct.

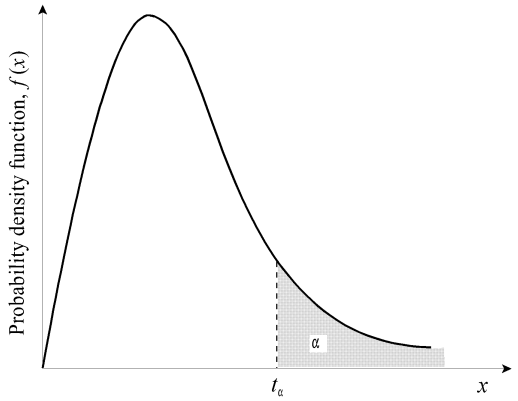


Figure 14 Threshold in chi-square test is selected so that $\mathrm{Prob}(X \geq t_\alpha) = \alpha$ [9]

Second, this measure has to be compared with a **threshold** which is set as a function of the so-called **significance level** of the test. Usually this level is established at 1% or 5%. Note these levels correspond to the right tail of the chi-square distribution, Figure 14.

Below we enumerate which are the steps to be followed [9] [10], for performing this test:

1.- We partition the sample space, S_X, into the union of K disjoint intervals/bins.
2.- Then, we compute the probability, b_k, that an outcome falls in the k-th interval under the assumption that X follows the proposed distribution. Thus, if we have n repetitions of the experiment, the expected number of outcomes in the k-th interval is $m_k = nb_k$.
3.- The chi-square measure, D^2, is defined as the weighted difference between the observed number of outcomes, N_k, that fall in the k-th interval, and the expected number, m_k, i.e.,

$$D^2 = \sum^{K} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \sum_{k=1}^{K} \frac{(N_k - m_k)^2}{m_k} \tag{41}$$

4.- If the fit is good, then D^2 will be small and the null hypothesis, i.e., that the measured data follows a given theoretical distribution, will not be rejected. If however, D^2 is too large, that is, if $D^2 \geq t_\alpha$, where t_α is the **threshold** for significance level α, the null hypothesis is rejected.

The chi-square test is based on the fact that, for large n, the random variable D^2 follows a chi-square distribution with $K-1$ degrees of freedom.

The **threshold**, t_α, can be computed by finding the point at which $\mathrm{Prob}(Y \geq t_\alpha) = \alpha$, (Figure 14), where Y is a chi-square random variable with $K-1$ **degrees of freedom**, DoF, see Annex.

The **thresholds** for the 1% and 5% levels of significance and different **degrees of freedom** are given in Table 4. The number of DoFs is $K-1$ is equal to the number of intervals used (K) or bins minus one.

Moreover, It is recommended that, if r is the number of parameters extracted from the data (e.g., mean, standard deviation, etc.), then D^2 is better approximated by a chi-square distribution with $K-r-1$ degrees of freedom. Each estimated parameter decreases the degrees of freedom by one.

In **`fitRayChiTest`** we have performed this test. First, we chose the number of bins to be used in the histogram, in this case **`Nbins = 10`**. Histograms in this case are calculated in terms of number of counts in each bin instead of dividing by the total number of counts as we did in previous sections, Figure 15.

The steps in the script were the following,

```
Nbins = 10;
[HvnormY, HvnormX] = hist(v_norm, Nbins);
step = HvnormX(2) - HvnormX(1);
end1 = HvnormX - step/2;
end2 = HvnormX + step/2;

HvnormYtheoretical = ...
    exp(-(end1.^2)./(rms^2))-exp(-(end2.^2)./(rms^2));

HvnormYtheoretical = ...
    HvnormYtheoretical*length(v_norm);  % convert to counts
```

Table 4 Thresholds for significance levels 1% and 5%, and different degrees of freedom

K	5%	1%	K	5%	1%
1	3.84	6.63	12	21.03	26.22
2	5.99	9.21	13	22.36	27.69
3	7.81	11.35	14	23.69	29.14
4	9.49	13.28	15	25.00	30.58
5	11.07	15.09	16	26.30	32.00
6	12.59	16.81	17	27.59	33.41
7	14.07	18.48	18	28.87	34.81
8	15.51	20.09	19	30.14	36.19
9	16.92	21.67	20	31.41	37.57
10	18.31	23.21	25	37.65	44.31
11	19.68	24.76	30	43.77	50.89

Note how we convert the theoretical histogram, i.e., the area under the pdf within the bin (a probability) to a frequency or count by multiplying by **`length(v_norm)`** which is parameter n in the step-by-step procedure presented above, that is, the number of repetitions of the experiment.

Then we go on to calculate **`D2`** and the number of degrees of freedom: $10 - 1 - 1 = 8$ since we only extracted one parameter for a one-parameter distribution (Rayleigh.)

From Table 4 we can read out the thresholds for 5% and 1% significance levels, those thresholds are 15.51 and 20.09, respectively.

```
D2 = sum((HvnormYtheoretical - HvnormY).^2./HvnormYtheoretical)

df = (Nbins - 1) - 1; % reduce DOFs by 1 as we extracted rms value from data
```

From the above calculations we obtain **`D2=7.999`**. This means that the test is **passed** both for 5% significance and for 1%. Using

```
alpha = 1 - gammainc(0.5*D2,0.5*df) % significance level
```

we can figure out the actual point in the chi-square pdf we are at by checking out the value of α, we get **`alpha = 43.3564`** %. That is the value of **`D2`** we get leads to an area to the right equal to 0.43, much higher than that corresponding to both thresholds, 0.05 (5%) and 0.01 (1%).

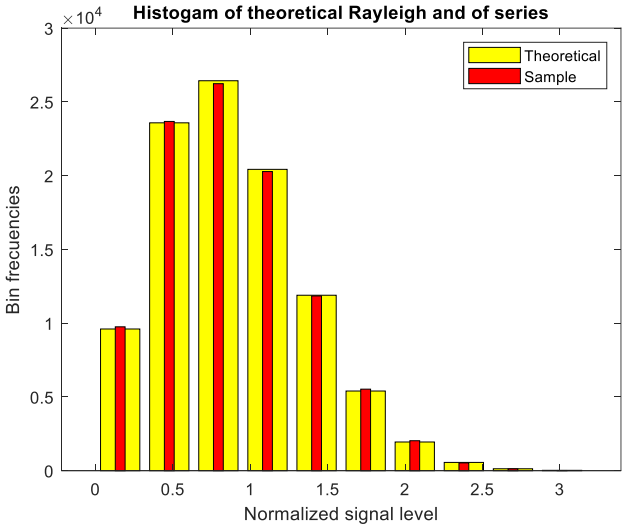


Figure 15 Time series and $\boldsymbol{q} = \mathbf{1}$ Rayleigh histograms

This is not the case here, but in the event that we barely passed or failed the test, we could repeat it again under more favorable conditions. What we can do is remove those samples which are correlated to neighboring ones. Keeping them would actually mean that we are using the same sample several times, depending on the correlation time/distance in the series. Theoretically, the samples used are required to be **independent**. We will be approximating this requirement by using **uncorrelated** samples.

Removal of correlated samples can be performed by simple decimation. A measure of the correlation time/distance can be carried out by computing the **auto-covariance** of the signal and checking the sample spacing required to go from a correlation level of 1 for null sample separations to somewhere under a value of $1/e \approx 0.3679$.

Briefly we provide here the mathematical definition of the cross-covariance, implemented in Matlab's function **`xcov`**,

$$\phi_{xy}(m) = E\{(x[n+m] - \bar{x})(y[n] - \bar{y})^*\} \tag{42}$$

here $E\{\cdot\}$ is the mathematical expectation, * indicates complex conjugate, $\bar{x} = E(x)$ and m is a sample shift. The cross-covariance is computed for all possible negative and positive shifts between the two series.

To normalize the covariance, the above expression is divided by the product of variances, i.e.,

$$\rho_{xy}(m) = \frac{E\{(x[n+m] - \bar{x})(y[n] - \bar{y})^*\}}{\mathrm{var}(x)\ \mathrm{var}(y)} \tag{43}$$

which we will call **correlation coefficient** and is provided by Matlab by adding **`'coeff'`** to the list of inputs to **`xcov`**. In the case both series are the same, we get the auto-covariance. To cover all possible shifts, the result is a series of $2N - 1$ values. The normalized auto-covariance will have a maximum of one for zero shift, right in the center of the resulting plot.

After computing the auto-covariance of **`RayleighSeries`**, we can see (Figure 16) how, at a time lag of 2 samples the correlation coefficient goes below $1/e$. As an extra precaution, we have chosen a spacing of 5 samples and performed a decimation of the original series.

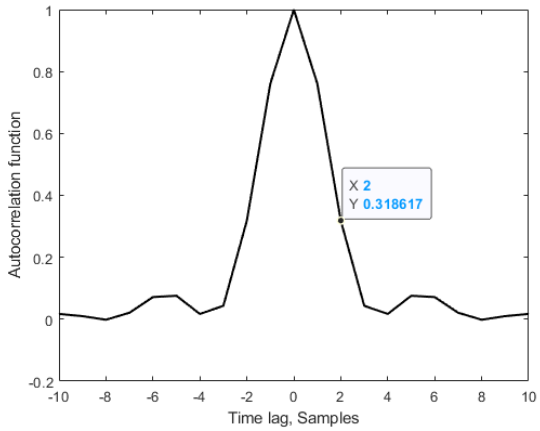


Figure 16 Normalized autocorrelation function of original time series.

Thus, we decimate the series by a factor of 5. This can be achieved by taking one in every five samples in the series, i.e.,

```
v_norm_dec = v_norm(1:5:length(v_norm))
```

To prove that we have not lost any information, we plot again the CDFs of both the original and the decimated series in Figure 17. We have used a double logarithmic scale plot (**`loglog`**) to make the differences more apparent, especially at the lower tail of the distribution. From the figure, it is clear that the distribution has not changed after decimation.

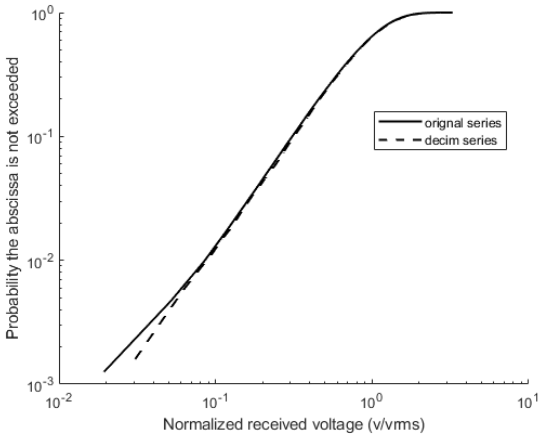


Figure 17 CDFs of original and decimated series in double-logarithmic scale.

Thus, we proceed to perform the chi-square test on the decimated series. We now repeat the test on the decimated series (**Figure 18**) and we get a value for D^2 is now equal to 7.19, which is a little be lower which, in case it was necessary, it would help pass the test. Now, the α value corresponding to the D^2 we got is **`alpha = 51.628 %`**.

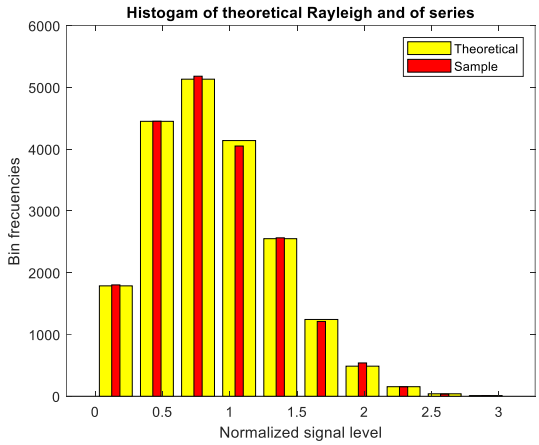


Figure 18 Time series and mean=1 Rayleigh histograms. Decimated series.

7. Further work

We propose the reader to repeat the above process and analyze the data contained in file **`UnknownSeries.mat`**. This series is illustrated in Figure 19. For full disclosure, this new series has been created with a Rice model (see elsewhere on this site). The Rayleigh distribution is a special case of the Rice distribution. We have selected a set of parameters (Carrier to Multipath power ratio, parameter, $K = -5.0$ dB) that make the series quite close to a Rayleigh series. The question is, is it enough to pass the test.

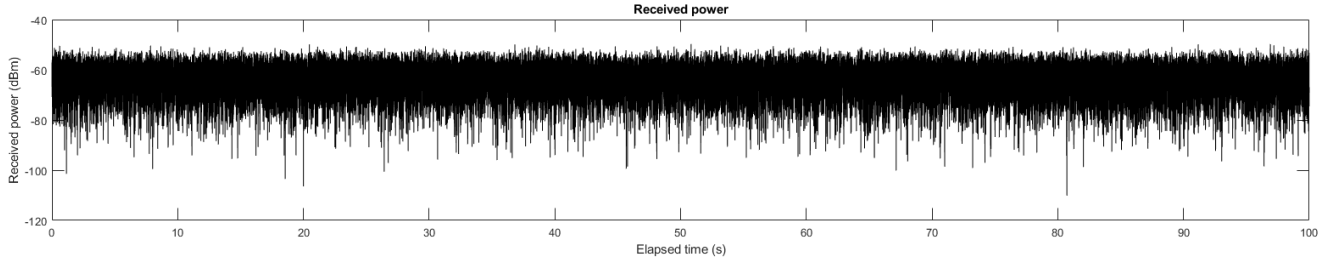


Figure 19 Series provided to repeat the study

8. References

[1] J.M. Hernando & F. Pérez-Fontán. An Introduction to Mobile Communications Engineering. Artech House, 1999.
[2] W.C.Y. Lee. Mobile Communications Design Fundamentals. Wiley Series in Telecommunications and Signal Processing. John Wiley & Sons, Ltd, Chichester, UK, 1993.
[3] F.Pérez-Fontán and P.Mariño Espiñeira. Modelling the Wireless Propagation Channel: A Simulation Approach with MATLAB. Wiley. 2008
[4] H.Suzuki. A Statistical Model for Urban Radio Propagation. IEEE Transactions on Communications, Vol. Com-25, No. 7, July 1977, Pp. 673-680
[5] M. Hata. Empirical formula for propagation loss in land mobile radio services. IEEE Trans. Vehicular Technology, 29(3), 1980, 317–325.
[6] Rec. ITU-R P.1057-6. Probability distributions relevant to radiowave propagation modelling. 2019. International Telecommunication Union. Radiocommunication Sector.
[7] J.D.Parsons. The Mobile Radio Propagation Channel, Second Edition. 2nd Edition. John Wiley & Sons, Ltd. 2000
[8] W.H. Tranter, K.S. Shanmugan, Theodore S. Rappaport, and Kurt L. Kosbar. Principles of Communication Systems Simulation with Wireless Applications, Prentice Hall, Professional Technical Reference, 2004, IBSN: 0-13-494790-8.
[9] A. Leon-Garcia. Probability and Random Processes for Electrical Engineering, Second Edition (International Edition). Addison-Wesley, 1994.
[10] http://en.wikipedia.org/wiki/Chi-squared_distribution (20230116)

9. Software Supplied

In this section, we provide a list of functions and scripts, developed in MATLAB®, implementing the various projects and theoretical introductions mentioned in this chapter. They are the following:

FUNCTIONS
fpdfCDFbins.m
RayleighCDF.m
RayleighCDFrms.m
Rayleighpdf.m
Rayleighpdfrms.m

-

SCRIPTS
Exponential_pdf_cdf
fitExponential
fitRayleigh
fitRayChiTest
plotChiCCDFs
plotChiPDFs
Rayleigh_pdf_cdf

Additionally, the following time series are supplied:

SERIES
RayleighSeries.mat
UnknownSeries.mat

10. ANNEX. The Chi-square distribution and goodness of fit test

This test involves the validation of a hypothesis, H_0, **the null hypothesis**, which states that the distribution we are testing fits the data. The **alternative hypothesis**, H_1, states that the chosen distribution does for fit the data.

The test is based on the central limit theorem where instead of having numerous independent random variables which, when added together, yield a Gaussian distribution, what we have is multiple random variables squared. In this case the resulting random variable converges to a Chi-Square distribution.

Assuming $Z_1, \ldots, Z_K$ independent, standardized (zero mean and unit standard deviation) normal random variables, their sum of their squares is given by

$$Y = \sum_{i=1}^{K} Z_i^2 \tag{44}$$

The resulting random variable follows a **chi-squared distribution** with k **degrees of freedom**, DoF. Its pdf is given by

$$f(x) = \frac{x^{(k-2)/2}\, e^{-x/2}}{2^{k/2}\, \Gamma(k/2)} \qquad x > 0 \tag{45}$$

where k is a positive integer and Γ is the Gamma function. The chi-square distribution is associated or is a special case of the Gamma distribution given below [10],

$$f(x) = \frac{\lambda(\lambda x)^{\alpha-1}\, e^{-\lambda x}}{\Gamma(\alpha)} \qquad x > 0 \text{ and } \alpha > 0, \lambda > 0$$

(46)

when $\alpha = k/2$, k is a positive integer, and $\lambda = 1/2$, and where Γ is the Gamma function, that is,

$$\Gamma(x) = \int_0^\infty e^{-t}\ t^{x-1}\, dt$$

(47)

The shape of the chi-square pdf as a function of its parameter k, is presented in Figure 20.

We used Matalb's built-in function **`gammainc`.**

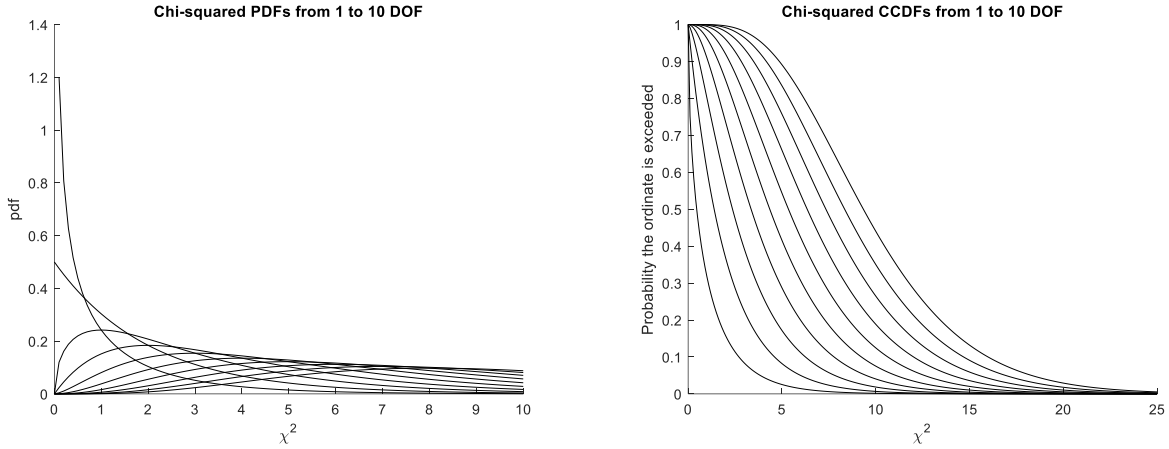


Figure 20 (left) Representation of various chi-square pdfs as a function of the DoF (using script `plotChiPDFs`). Representation of various chi-square CDFs as a function of the DoF (using script `plotChiCCDFs`).

The **chi-squared goodness of fit test** verifies whether a data set fits a given theoretical distribution. The data is split into **mutually exclusive** events.

Four steps are followed in the chi-square test:

1.- We start off with a sample, X, that is, the various measurements (in our example, the experimental time series). We partition the sample space, S_X, into the union of K disjoint intervals/bins.

2.- Then we compute the probability, b_k, that an outcome falls in the k-th interval under the assumption that the data follows the proposed distribution. This information is obtained from the assumed theoretical pdf, f_{th}, as

$$b_k = \int_{L_{k-}}^{L_{k+}} f_{\text{th}}(u)du$$

(48)

where L_{k-} and L_{k+} are the lower and upper limit of interval/bin k [10]. Thus, if we have n repetitions of the experiment (number of samples in our time-series), the expected number of outcomes in the k-th interval would be $m_k = nb_k$;

3.- Test parameter D^2 is defined as the averaged, weighted difference between the observed number of outcomes, N_k, that fall in the k-th interval (number of counts in the histogram in bin k), and the expected number, m_k, i.e.,

$$D^2 = \sum^{K} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \sum_{k=1}^{K} \frac{(N_k - m_k)^2}{m_k} \tag{49}$$

where D^2 is the test statistic which should follow a chi-square distribution, O_i is an observed frequency, and E_i is the expected (theoretical) frequency.

4.-The final step is verifying whether the test parameter, D^2, is large or small in comparison with a threshold obtained from the chi-square distribution.

It is clear that if the fit is good, then the value of D^2 should be small. The null hypothesis, i.e., that the measured data follows a given theoretical distribution, will be rejected if D^2 is too large, that is, if $D^2 \geq t_\alpha$, where t_α is the *threshold* for **significance level**, α.

As said, the chi-square test is based on the fact that, for large n, the random variable D^2 follows a chi-square distribution with $K-1$ degrees of freedom. The threshold, t_α, can be computed by finding the point at which $\text{Prob}(Y \geq t_\alpha) = \alpha$, where Y is a chi-square random variable with $K-1$ **degrees of freedom**, DoF.

The thresholds for the 1% and 5% levels of significance and different degrees of freedom are given in Table 4. The number of DoFs is $K-1$, that is, the number of intervals or bins minus one. It is recommended that, if r is the number of parameters extracted from the data (e.g., mean, standard deviation, etc.), then D^2 is better approximated by a chi-square distribution with $K-r-1$ degrees of freedom. Each estimated parameter decreases the degrees of freedom by one.

It is also recommended that we perform the test in such a way (choose the number of bins) that the expected number of outcomes in each interval be at least five or more. This will improve the accuracy of approximating the CCDF of D^2 by a chi-square distribution