**Fitting time series to a Rayleigh distribution**

**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 1](#_Toc124417459)

[2. Common statistical distributions used in radio 6](#_Toc124417460)

[3. The Rayleigh distribution 7](#_Toc124417461)

[4. The exponential distribution 11](#_Toc124417462)

[5. Example: Fitting a series to the Rayleigh distribution 12](#_Toc124417463)

[6. The chi-square test 16](#_Toc124417464)

[7. Further work 21](#_Toc124417465)

[8. References 21](#_Toc124417466)

[9. Software Supplied 22](#_Toc124417467)

[10. ANNEX. The Chi-square distribution and goodness of fit test 23](#_Toc124417468)

## Mobile communications propagation basics

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

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's antenna.

Depending on the BS/access point height, cells of larger or smaller size can be created. The classical cellular environment with tall masts above rooftops gives rise 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.

Man-made structures [Lee] such as buildings or small houses in suburban areas, 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 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 power decay, two main effects are present in mobile propagation: *shadowing* and *multipath*.

We can identify three different rates of change in the received signal as a function of the distance (spatial variations) between BS 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. Time variations, also take place. However, unless we are dealing with fixed links where we cannot observe the spatial variations, time variations tend to be less significant. 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 is considerably reduced.

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 will 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. All signal levels are referred to that we would have in LOS conditions and where no multipath is present.

*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.rigth). Note that the received signal oscillates around levels well below that of the direct, LOS signal. 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 later.

|  |
| --- |
| rice and rayleigh |

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

The received field strength or the received voltage may be represented in the time domain, , or in the traveled distance domain, . Figure 2 shows a typical mobile communications scenario with MS driving away from BS along a radial route so that the link profile is the same as that of the route 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. Note that, superposed, we will find the fast Rayleigh variations, not shown.

|  |
| --- |
| F 1-3 modified |

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

For carrying out propagation channel measurements, the mobile speed, , should 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, , it is straightforward to make the conversion between the signal represented in the time domain,, and the signal represented in the (traveled) distance domain, .

Variable 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 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.

|  |
| --- |
| slow and fast |

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

Generally, the received signal variations, or , may be broken down, in a more or less artificial way, into two components [Lee],

- the slow or long-term variations: or ;and

- the fast or short-term variations: or .

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 may be very small. In this way, reasonably large numbers are used in the calculations.

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

|  |  |
| --- | --- |
| or, alternatively, | (xxx) |

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

|  |  |
| --- | --- |
| or, alternatively, | (xxxx) |

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 equal to some tens of wavelengths,

|  |  |
| --- | --- |
|  | (1.14) |

Typically, lengths of to are used [Lee]. For example, for the 2 GHz band(,) the averaging length would be. The average value, ,computed for a given route position,, is usually called the *local mean* at .

In dense urban areas, it has been observed experimentally [Lee] that the slow variations of the received signal, that is, the variations of the local mean, , 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, .

The length, , 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 sections of route, we need a combination of distributions, e.g. the Suzuki distribution [Suzuki]: Rayleigh-plus-Lognormal.

Over longer route sections ranging from 50 m or 100 m to even 1 km, the variations of the local mean are generally studied. This extended surface is usually called a *larger area*. Typically, standard propagation models do not attempt to predict the fast signal variations. Instead they predict the mean, , the standard deviation (or *location variability*), , of the local mean variations (normal distribution in dB) and the correlation distance, , within the *larger area*.

Before low-pass filtering, the very slow variations due to the radio path range (also called **path loss**) from BS must be removed. The **free-space loss**, (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 (Figure 4), i.e.,

|  |  |
| --- | --- |
|  | (YYYY) |

where indicates "proportional to". Parameter is the difference in received power expressed in dB at distances and . 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.

|  |
| --- |
| F 1-5 modified |

Figure 4 Received signal decay with distance: and laws.

These variations are first quite steep while, later, they show gentler decays for greater distances. For example, using Eq. YYY, 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 power law (as in free space conditions) but, rather, it presents a larger exponent. Signal decay is usually modeled by a law, i.e., being proportional to the distance risen to the power of ***.*** The values of are typically somewhere near 4, i.e., 40 dB/decade (Figure 4). Well known models such as that of Hata [4] predict exponents close to 4.

The path-loss expressions normally provided by propagation models (Chapter II.3) are of the form

|  |  |
| --- | --- |
|  | (xxx) |

where and are dependent on the frequency and a number of other factors as listed below. Parameter is the loss at a *reference distance*, in this case, 1 km, and 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 and . 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*, Δ*h*, or *roughness*,σt);

- the type of *land usage* (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.,

|  |  |
| --- | --- |
|  | (1.19) |

The *path loss* is defined as the loss between isotropic antennas (0 dB gain) for a given antenna separation, . Isotropic antennas do not exist in practice but are commonly used in link budget calculations since they allow the definition of the loss independently of the antennas used in the actual link. Then, when computing an actual 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 [4] use a different value. Other models use the so-called *plane-earth* *loss* () as their reference.

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 usage in the vicinity of MS: urban, suburban, rural, open, woodland, etc.

## Common statistical distributions used in radio

Here, we provide a brief reminder of some very basic statistical definitions. A random variable, , has a **probability density function** (pdf), , where is a non-negative, if

|  |  |
| --- | --- |
|  | (xxx) |

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

|  |  |
| --- | --- |
|  | (xxx) |

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

|  |  |
| --- | --- |
|  | (xxx) |

Furthermore, if is continuous at , then

|  |  |
| --- | --- |
|  | (xxx) |

Intuitively, one can think of as being the probability of falling within the **infinitesimal interval** .

A pdf fulfils the following property,

|  |  |
| --- | --- |
|  | (xxx) |

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

|  |  |
| --- | --- |
|  | (xxx) |

another is *mean square value*

|  |  |
| --- | --- |
|  | (xxx) |

The square root of is called the rms value (root mean square).

Finally, the variance of is

|  |  |
| --- | --- |
|  | (xxx) |

where σ is the *standard deviation*, and the **median value**, is given by

|  |  |
| --- | --- |
|  | (xxx) |

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.

## 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

|  |  |
| --- | --- |
|  | (xxx) |

where is a voltage This distribution has a single parameter, its *mode* or *modal value*, . Other related parameters are given in Table 1 as a function of the mode. In this chapter, we use 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 I.4.7 for .

|  |
| --- |
| Rayleigh parameters mode 1 |
| **Figure 5.** |

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.,

|  |  |
| --- | --- |
|  | (xxx) |

**Table 1.** Rayleigh distribution parameters as a function of its mode

|  |  |
| --- | --- |
| Mode |  |
| Median |  |
| Mean |  |
| RMS value |  |
| Standard deviation |  |

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,

|  |  |
| --- | --- |
|  | (xxx) |
|  | (xxx) |
|  | (xxx) |
|  | (xxx) |

where 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.

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

|  |  |
| --- | --- |
|  | (xxx) |

where is the *mean*.

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

|  |  |
| --- | --- |
|  | (xxx) |

where is the *rms value*.

We are interested in discussing 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 a phasor sum of the various multipath contributions, where each term corresponds to one such contribution, that is,

|  |  |
| --- | --- |
|  | (xxx) |

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 (or any other RF component of the signal.) In this case, the result is a time-varying complex valued time-series, , that is,

|  |  |
| --- | --- |
|  | (xxx) |

where is a complex voltage (base band equivalent), and where we have the summation of multipath contributions with time-varying complex amplitudes and time-varying delays .

The magnitude of term , i.e., 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.

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

The envelope of the base band equivalent voltage is Rayleigh distributed (we will see this somewhere else.) The same occurs if instead of working with complex base band signals, we could also work with the RF (pass-band) voltages. The relationship between a signal in the low-pass and RF domains is as follows,

|  |  |
| --- | --- |
|  | (xxx) |

which becomes

|  |  |
| --- | --- |
| **VERIFICAR ESTO** | (xxx) |

This is illustrated in Figure 6. Note how is real valued while 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 has been generated and the description of the corresponding script are discussed somewhere else on this site. To help with the illustration we have forced a very high Doppler by using a high terminal speed and a very low carrier frequency. Note the carrier in blue 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, . Now, to deal with convenient magnitudes (not very small voltage values), we can normalize so that we get reasonably high values.

A good way of performing the sough after normalization is use as reference the average power, . If we assume a measured received time-series from a transmitted a CW at , the **instantaneous** received power is given by

|  |  |
| --- | --- |
|  | (XXX) |

In this case, we have assumed an impedance . A typical impedance value would be . In this case,

|  |  |
| --- | --- |
|  | (XXX) |

We can perform the following normalization,

|  |  |
| --- | --- |
|  | (XXX) |

We can also work with normalized voltages defined as follows,

|  |  |
| --- | --- |
|  | (XXX) |

We want to verify whether follows a Rayleigh distribution with pdf

|  |  |
| --- | --- |
|  | (xxx) |

Given the normalization we have used, , it turns out that , and thus,

|  |  |
| --- | --- |
|  | (xxx) |

We can also analyze the behavior of the normalized power, . In case the voltage follows a Rayleigh distribution, the power follows an **exponential distribution** (see section below.)

The CDF for becomes

|  |  |
| --- | --- |
|  | (xxx) |

Finally, we illustrate how the Rayleigh pdf and CDF look like with the normalization contention we have taken. These are illustrated in Figure XXX. Note how the modal value is no longer unity, value that coincides now with the rms value.



Figure 6 Rayleigh distribution normalized with respect to the rms level (rms = 1).

RMSvalue = 1.0

modalvalue = RMSvalue/sqrt(2);

medianvalue = RMSvalue\*sqrt(log(2));

meanvalue = RMSvalue\*sqrt(pi)/2;

standarddeviation = RMSvalue\*sqrt(1-pi/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 . In this case, is exponentially distributed with a pdf is given by

|  |  |
| --- | --- |
|  | (xxx) |

where is the mean value of the distribution. Incidentally, is also its standard deviation of the distribution (Table XXXX). Since the normalization was done in such a way that , then the pdf simplifies to

|  |  |
| --- | --- |
|  | (xxx) |

And the CDF is given by

|  |  |
| --- | --- |
|  | (xxx) |

For , we get

|  |  |
| --- | --- |
|  | (xxx) |

Table XXXX. Statistics of the Exponential distribution

|  |  |
| --- | --- |
| Mean |  |
| Median |  |
| Mode | 0 |
| Standard deviation |  |
| rms value | XXXXX |



Figure 7 Exponential distribution: pdf and CDF plots (script Exponential\_pdf\_cdf)

## Example: Fitting a series to the Rayleigh distribution

File **RayleighSeries.mat** (Figure I.4.6) 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 difference is that, in this case we assume a very high signal-to-noise (SNR) ratio in such a way that there is not a distortion from the ideal cae.

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 a two-column matrix which includes, in the first column, the time axis (variable **time\_axis**) in seconds and, in the second column, the power in dBm (variable **PdBm**). The time axis is sampled with an interval , corresponding to a sampling frequency .

TotSimTime = 100.0; % seconds

f = 2000e6; % carrier frequency (Hz)

V = 20; % MS speed (m/s)

F = 8; % Sampling: Fraction of wavelength

sigma = 1/sqrt(2); % to achieve rms = 1 series

DIV\_BRCHS = 1; % No of diversity branches FOR LATER TESTS

|  |
| --- |
|  |

Figure 8 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 series is plotted in dBm in Figure I.4.6. What we want is to model the corresponding **normalized voltage**. A matched load resistance, , of 50 Ω is assumed, thus, the instantaneous power can be computed by . The power and the voltage are linked through . 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. 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 Figures XXX-YYY. Note the original voltages are actually very small in the order of . When normalized they become magnitudes easier to handle.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figure 9 (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 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 intervals 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 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 bint to 100 bins. This is a good approach for computing CDFs while for pdfs we need to use a smaller number to prevent 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** frequency is a count of the number of occurrences, not a probability. To convert those probabilities, we divide the counts 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 and theoretical pdfs are 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**

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) and 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 below the pdf curve by assimilating it to a rectangle. A more correct way is performing the integration properly, In this case, it is easy since it means subtracting exponential functions,

|  |  |
| --- | --- |
|  | (xxx) |

where, in this case, to have an rms value equal to 1, we need to make .

Then we plotted the pdf curve with the histogram to show that there are different things and cannot be compared directly. Finally, an approximation of the integral in Eq. XXX is performed approximately by multiplying the height of the pdf by the bin width. The obtained results are shown in Figure BBB.

|  |  |
| --- | --- |
|  |  |

Figure 10 (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 **fetExponential**). The normalized power series and the results are illustrated in Figures XXXX-XXX. 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;

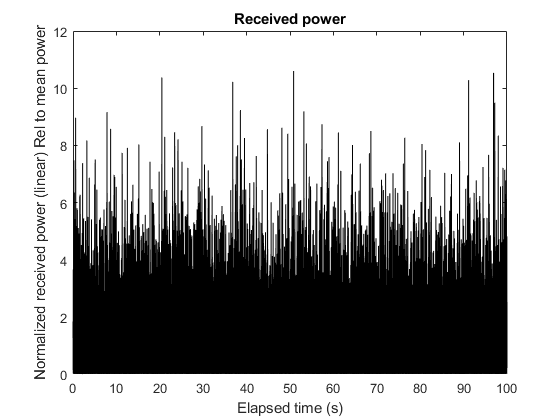


Figure 11 Normalized received power time series

|  |  |
| --- | --- |
|  |  |

Figure 12 (left) Measured and theoretical CDFs of the normalized power. (right) Measured and theoretical pdfs of the normalized power.

## The chi-square test

Script **fitRayChiTest** goes a step further from what we did in script **fitRayleigh**. We repeat most of what we already did 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**.

|  |
| --- |
| CHI210 |

Figure 13 Threshold in chi-square test is selected so that  [7]

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** [7] [wikipedia] (see also the Annex at the end.)

From the Annex, there are two basic elements in the chi-square test [7]. The chi-square distribution models the squared differences between a theoretical pdf and an experimental one. In Figure XXX we illustrate the pdf of a chi-square distribution. We can compare 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.

Second, this measure has to be compared with a**threshold** whichis 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 xxxx.

Below we enumerate which are the steps to be followed [7] [wikipedia], for performing this test:

1.- We partition the sample space, , into the union of disjoint intervals/bins.

2.- Then, we compute the probability, , that an outcome falls in the -th interval under the assumption that follows the proposed distribution. Thus, if we have repetitions of the experiment, the expected number of outcomes in the -th interval is .

3.- The chi-square measure, , is defined as the weighted difference between the observed number of outcomes, , that fall in the -th interval, and the expected number, , i.e.,

|  |  |
| --- | --- |
|  | (1.46) |

4.- If the fit is good, then will be small and the null hypothesis, i.e., that the measured data follows a given theoretical distribution, will not be rejected. If however, is too large, that is, if , where is the **threshold** for significance level , the null hypothesis is rejected.

The chi-square test is based on the fact that, for large , the random variable follows a chi-square distribution with degrees of freedom.

The **threshold**, , can be computed by finding the point at which , (Figure XXX), where is a chi-square random variable with **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 III. The number of DoFs is is equal to the number of intervals used () or bins minus one.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table III.** Thresholds for significance levels 1% and 5%, and different degrees of freedom | | | | | |
|  | 5% | 1% |  | 5% | 1% |
| 1  2  3  4  5  6  7  8  9  10  11 | 3.84  5.99  7.81  9.49  11.07  12.59  14.07  15.51  16.92  18.31  19.68 | 6.63  9.21  11.35  13.28  15.09  16.81  18.48  20.09  21.67  23.21  24.76 | 12  13  14  15  16  17  18  19  20  25  30 | 21.03  22.36  23.69  25.00  26.30  27.59  28.87  30.14  31.41  37.65  43.77 | 26.22  27.69  29.14  30.58  32.00  33.41  34.81  36.19  37.57  44.31  50.89 |

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 I.4.10.

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**

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 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: since we only extracted one parameter for a one-parameter distribution (Rayleigh.)

From **Table III** 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 , we get **alpha = 43.3564** %.

|  |
| --- |
|  |

Figure 14 Time series and Rayleigh histograms

In case, we barely passed or failed the test, we can repeat it again what we can do is to 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 .

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

|  |  |
| --- | --- |
| , | (xxx) |

here is the mathematical expectation, \* indicates complex conjugate, and 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.,

|  |  |
| --- | --- |
|  | (xxx) |

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 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 I.4.10) how, at a time lag of 2 samples the correlation coefficient goes below . As an extra precaution, we have chosen a spacing of 5 samples and performed a decimation of the original series.

|  |
| --- |
|  |

Figure 15 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 I.4.11. 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 16 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 I.4.12) and we get a value for is now equal to , which is a little be lower which, in case it is necessary would help pass the test. Now, the value corresponding to the we got is **alpha = 51.628 %**.

|  |
| --- |
|  |

Figure 17 Time series and mean=1 Rayleigh histograms. Decimated series.

## Further work

We propose the reader to repeat the above process and analyze the data contained in file **UnknownSeries.mat**.

## References

[HERFON] J.M. Hernando & F. Pérez-Fontán. *An Introduction to Mobile Communications Engineering*. Artech House, 1999.

[Wiley modeling]

[Tranter] William H. Tranter, K. Sam 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.

[ITU] RECOMMENDATION ITU-R P.1057-1. Probability distributions relevant to radiowave propagation modelling. (1994-2001). International Telecommunication Union. Radiocommunication Sector

[Lee] W.C.Y. Lee. *Mobile Communications Design Fundamentals. Wiley Series in Telecommunications and Signal Processing*. John Wiley & Sons, Ltd, Chichester, UK, 1993.

[Hata] M. Hata. Empirical formula for propagation loss in land mobile radio services. *IEEE Trans. Veh. Tech*., **29**(3), 1980, 317–325.

[Okumura]

[LeonG] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering*, Second Edition (International Edition). Addison-Wesley, 1994.

[Sklar] B. Sklar. *Rayleigh Fading Channels. Mobile Communications Handbook* (Ed. S.S. Suthersan). CRC Press, 1999.

[Suzuki],

[wikipedia] http://en.wikipedia.org/wiki/Chi-squared\_distribution (20131002)

[Gree99] L. J. Greenstein, D. G. Michelson, and V. Erceg. Moment-Method Estimation of the Ricean -Factor. IEEE COMMUNICATIONS LETTERS, VOL. 3, NO. 6, JUNE 1999 175-176

[APPLIED] S.C. Chapra. Applied Numerical Methods with MATLAB for Engineers and Scientists. 3rd edition. McGraw-Hill, 2012

[regression] R.C. Rumpf. Computational Science: Computational Methods in Engineering. Curve Fitting to Polynomials & Interpolation/Extrapolation. https://empossible.net/academics/emp4301\_5301/ (accessed 20200114)

Spring 2018. University of Texas at El Paso. College of Engineering. Department of Electrical and Computer Engineering.

[1st edition]

## 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** |
| **Rayleigh\_pdf\_cdf.m** |
| **Rayleighpdf.m** |
| **Rayleighpdfrms.m** |
|  |
|  |

-

|  |
| --- |
| **SCRIPTS** |
| **plotChiCCDFs** |
| **plotChiPDFs** |
| **fitRayleigh** |
| **fitRayChiTest** |
| **Rayleigh\_pdf\_cdf** |
| **Exponential\_pdf\_cdf** |

Additionally, the following time series are supplied:

|  |
| --- |
| **SERIES** |
| **RayleighSeries.mat** |
| **UnknownSeries.mat** |
|  |
|  |

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

This test involves the validation of a hypothesis, , **the null hypothesis**, which states that the distribution we are testing fits the data. The **alternative hypothesis**, , 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 independent, standardized (zero mean and unit standard deviation) normal random variables, their sum of their squares is given by

|  |  |
| --- | --- |
|  | (A-1) |

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

|  |  |
| --- | --- |
|  | (A-2) |

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

|  |  |
| --- | --- |
|  | (A.3) |

when , is a positive integer, and , and where is the Gamma function, that is,

|  |  |
| --- | --- |
|  | (A-4) |

The shape of the chi-square pdf as a function of its parameter , is presented in Figure A.1.

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

|  |  |
| --- | --- |
|  |  |
| Figure A.1. Representation of various chi-square pdfs as a function of the DoF (using script **plotChiPDFs**) | Figure A.2. 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, , that is, the various measurements (in our example, the experimental time series). We partition the sample space, , into the union of disjoint intervals/bins.

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

|  |  |
| --- | --- |
|  | (xxx) |

where and are the lower and upper limit of interval/bin [Wikipedia]. Thus, if we have repetitions of the experiment (number of samples in our time-series), the expected number of outcomes in the -th interval would be ;

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

|  |  |
| --- | --- |
|  | (xxx) |

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

4.-The final step is verifying whether the test parameter, , 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 should be small. The null hypothesis, i.e., that the measured data follows a given theoretical distribution, will be rejected if is too large, that is, if , where is the *threshold* for **significance level**, .

As said, the chi-square test is based on the fact that, for large , the random variable follows a chi-square distribution with degrees of freedom. The threshold, , can be computed by finding the point at which , where is a chi-square random variable with **degrees of freedom**, DoF.

The thresholds for the 1% and 5% levels of significance and different degrees of freedom are given in Table I.4.2. The number of DoFs is , that is, the number of intervals or bins minus one. It is recommended that, if is the number of parameters extracted from the data (e.g., mean, standard deviation, etc.), then is better approximated by a chi-square distribution with 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 by a chi-square distribution