Chapter 4

Statistical Decision Making

In previous chapters, the focus was on representing and analyzing the elements of measurement. In the remaining chapters, our goal is to describe the processes of decision-making and measurement-system evaluation. Because objects and errors are naturally stochastic, decisions and evaluations are inherently statistical. An introductory review of probability and random processes is offered in Appendix B. Readers are urged to at least skim through that material to familiarize yourself with notation before proceeding.

4.1 Bayesian methods

A Bayesian decision maker considers prior knowledge about the state of an object along with any measurements from the object when making a decision about the state of the object. These are the elements of medical decision making that are based on patient history and test measurements.

Let the event space $\mathtt{S}$ be a population of N cancer patients that all have tumors. It is important to define the event space so that probabilities can be correctly computed. Let W (with) be the event that a patient has a metastatic tumor and the complement W^c be the event that a patient has a non-metastatic tumor. P (positive) is the event that a patient tests positive for metastatic disease regardless of the true tumor type, and ins complement P^c is the event a patient tests negative.

All N patients in $\mathtt{S}$ have a tumor and each has a test result. The number of patients with a metastatic tumor (regardless of test results) is N_W, and the probability of finding such

Patient Condition

Test Results

State	w/ Disease	w/o Disease	Total #
Positive	N_{TP}	N_{FP}	N_P
Negative	N_{FN}	N_{TN}	N_{P^c}
Total #	N_W	N_{W^c}	N

Numbers

1) $N_W = N_{TP} + N_{FN}$
2) $N_{W^c} = N_{FP} + N_{TN}$
3) $N_P = N_{TP} + N_{FP}$
4) $N_{P^c} = N_{FN} + N_{TN}$
5) $N = N_W + N_{W^c} = N_P + N_{P^c}$
6) Disease Prevalence $= N_W / N$
7) Sensitivity $= N_{TP} / N_W$
8) Specificity $= N_{TN} / N_{W^c}$
9) PPV $= N_{TP} / N_P$
10) NPV $= N_{TN} / N_{P^c}$

Probabilities

a) $\Pr(W) = N_W / N$ = Prevalence
b) $\Pr(W^c) = N_{W^c} / N$
c) $\Pr(P) = N_P / N$
d) $\Pr(P^c) = N_{P^c} / N$
e) $\Pr(P \mid W) = N_{TP} / N_W$ = Sensitivity
f) $\Pr(P \mid W^c) = N_{FP} / N_{W^c}$ = 1-Specificity
g) $\Pr(W \mid P) = N_{TP} / N_P$ = PPV
h) $\Pr(W \mid P^c) = N_{FN} / N_{P^c}$
i) $\Pr(P^c \mid W) = N_{FN} / N_W$
j) $\Pr(P^c \mid W^c) = N_{TN} / N_{W^c}$ = Specificity
k) $\Pr(W^c \mid P) = N_{FP} / N_P$
l) $\Pr(W^c \mid P^c) = N_{TN} / N_{P^c}$ = NPV

Figure 4.1: (left) Truth table in terms of numbers of cases and associated probabilities. Table columns describe the states of the patient and rows describe the state of test results. N_W is the number of patients with disease and N_{W^c} is the number of patients without disease. N_P is the number of patients testing positive and N_{P^c} is the number of patients testing negative. Of course, $N = N_W + N_{W^c} = N_P + N_{P^c}$. N_{TP} is the number of true positive results; i.e., patients with disease that also test positive. N_{FP} is the number of false positive results. N_{TN} is the number of true negative results. N_{FN} is the number of false negative results. PPV is the positive predictive value and NPV is the negative predictive value. (right) A list of classification numbers and probabilities.

a patient in the study is $\Pr(W) = N_W/N$. The number of patients that test positive for a metastatic lesion (regardless of their true tumor type) is N_P, and the probability of finding such a patient in the study is $\Pr(P) = N_P/N$. Hopefully you are catching on to the combinations that each generate a classification number and probability $\Pr(\cdot)$.

From the definitions in Fig 4.1 and Eq (B.5), the marginal probability that a patient in the study has a positive test result, $\Pr(P)$, may be found from the sum of two conditional probabilities,

$$\Pr(P) = \Pr(P|W)\Pr(W) + \Pr(P|W^c)\Pr(W^c) = \frac{N_{TP}}{N_W}\frac{N_W}{N} + \frac{N_{FP}}{N_{W^c}}\frac{N_{W^c}}{N} = \frac{N_P}{N} \ .$$

The probabilities labeled (e)-(l) in Fig 4.1 may not all be obvious until you think carefully about them. Some are given special names, like sensitivity, because they often arise when evaluating a diagnostic test. I will point out a few of them found throughout the diagnostic testing literature. Note that these follow from the rules of probability discussed in Appendix B.

4.1.1 Terminology

- *True-positive fraction* (TPF) is the sensitivity of a binary test. It is the probability of obtaining a positive test result given that the patient has the disease, $\Pr(P|W) = N_{TP}/N_W$.

- *True-negative fraction* (TNF) is the specificity of a binary test. It is the probability of obtaining a negative test result given that the patient does not have the disease, $= \Pr(P^c|W^c) = N_{TN}/N_{W^c}$.

- *False-positive fraction* (FPF) is labeled 1-specificity when it is used as the abscissa (horizontal axis) for the ROC curve. Beginning with 1-TNF = $1 - N_{TN}/N_{W^c} = (N_{W^c} - N_{TN})/N_{W^c} = N_{FP}/N_{W^c}$. It equals the probability of obtaining a positive test result given that the patient does not have the disease.

- *False-negative fraction* (FNF) equals the probability of obtaining a negative test result given that the patient has the disease, $\Pr(P^c|W) = N_{FN}/N_W$.

- It is also true that

$$\begin{aligned} \text{TNF} + \text{FPF} &= 1 = \Pr(P^c|W^c) + \Pr(P|W^c) \\ \text{TPF} + \text{FNF} &= 1 = \Pr(P|W) + \Pr(P^c|W) \ . \end{aligned}$$

- *Positive predictive value* (PPV) is the probability that a patient has a positive test result and is actually positive (e.g., has a malignant tumor) divided by the probability a patient has a positive test result. $\text{PPV} = \frac{\Pr(PW)}{\Pr(P)} = \Pr(W|P)$.

- *Negative predictive value* (NPV) is the probability that a patient has a negative test result and is actually negative (e.g., has a benign tumor) divided by the probability a patient has a negative test result. $\text{NPV} = \frac{\Pr(P^cW^c)}{\Pr(P_c)} = \Pr(W^c|P^c)$.

Let's apply the definitions summarized in Fig 4.1 to an example that requires a decisions. Although it might not always be obvious, engineers and physicians are both decision makers that manage risk by balancing competing factors. Bayesian approaches offer guidelines for optimizing that balance.

Example 4.1.1. Diagnostic test evaluation.

We are asked to evaluate two tests that both sample patient tissues to detect metastatic tumors. Our data are derived from a clinical study on a population of N cancer patients. The result of each test performed on a patient is a binary indication – YES or NO – of whether that tumor is metastatic.

The prior probability of metastasis (prevalence of disease) in this population is 50%, i.e., $\Pr(W) = 0.50$ *and therefore* $\Pr(W^c) = 1 - \Pr(W) = 0.50$. *For* <u>*standard test A*</u>*, we are given that the sensitivity was measured to be* $\Pr(P|W) = 0.90$ *and the false-positive probability was measured to be* $\Pr(P|W^c) = 0.01$. *In the latter case, one out of 100 non-metastatic tumors will incorrectly be called metastatic.*

We are asked to compare method A above with a <u>*new blood test B*</u> *that uses optical absorption to detect tracer amounts of a blood-born protein specific to metastasis. Test B is much cheaper and offers increased sensitivity to* $\Pr(P|W) = 0.99$. *Unfortunately, test B also has a larger false-positive probability,* $\Pr(P|W^c) = 0.10$. *Which test has the better positive predictive value?*

To solve the problem, we must find a form of the PPV equation that includes the values given above. In this case, we have $\Pr(W)$, $\Pr(P|W)$, $\Pr(P|W^c)$. *So we expand the PPV equation to include these terms:*

$$\mathrm{PPV}_A = \Pr(W|P) = \frac{\Pr(P|W)\Pr(W)}{\Pr(P|W)\Pr(W) + \Pr(P|W^c)\Pr(W^c)} = \frac{(0.9)(0.5)}{(0.9)(0.5) + (0.01)(0.5)} = 0.989 \ .$$

Therefore 989 people out of 1000 are correctly diagnosed by test A. For test B, the sensitivity and false positive rates change:

$$\mathrm{PPV}_B = \frac{(0.99)(0.5)}{(0.99)(0.5) + (0.1)(0.5)} = 0.908 \ ,$$

or 908 out of 1000 are correctly diagnosed by test B. We find that increases in the false positive rate are very influential. Test A has a greater PPV because it gives a lower false positive rate. Provided a binary test has some sensitivity to metastasis, the PPV is maximized for any sensitivity provided the false positive rate is zero. That fact is pretty obvious once you think about what PPV tells us. It is the number of correct positive diagnoses divided by the number of patients who test positive.

How do the results change if the prior probability of having a metastatic tumor is reduced from 50% to 1%. This change is representative of changing the event space from a diagnostic test applied to patients with known tumors to a screening procedure applied to the general public. Notice the tests have not changed but the population being tested has changed. *The PPV for test A used to screen a general population is*

$$\mathrm{PPV}_A = \frac{(0.9)(0.01)}{(0.9)(0.01) + (0.01)(0.99)} = 0.476 \ .$$

For test B,

$$\mathrm{PPV}_B = \frac{(0.99)(0.01)}{(0.99)(0.01) + (0.1)(0.99)} = 0.091 \ .$$

As the disease prevalence in the population falls, both tests give lower PPVs. Clearly error rates and prevalence are very influential in the interpretation of test performance via PPV as a performance metric.

Bayes formula shows us how decisions about hypotheses can be optimized by combining measurements on patients with prior knowledge about the patient populations. Bayesian reasoning has major applications in statistically-based image reconstruction algorithms for data acquired from projections, a topic we will not discuss here except in one respect.

4.1.2 Posterior probability

The PPV equation, given by $\Pr(W|P)$, is a posterior probability. It is the probability of of W (patient is sick) given the evidence supplied by the test P. It can be written several ways to provide insights into its value. Applying Eq (B.5),

$$\Pr(W|P) = \frac{\Pr(P|W)\Pr(W)}{\Pr(P|W)\Pr(W) + \Pr(P|W^c)\Pr(W^c)} \;=\; \Pr(P|W)\frac{\Pr(W)}{\Pr(P)}\,.$$

In words related to the cancer-detection example above, the terms are described as

$$\text{Posterior Probability} = \text{Likelihood Function} \times \text{Prior Probability}$$

The likelihood function, $\Pr(P|W)$, is the sensitivity of the test, and hence a measure of the value of the data. I will call the prior probability the ratio of disease prevalence and the probability of a positive test result, i.e., $\Pr(W)/\Pr(P)$, which is everything we know about the situation *before* we make a measurement. The posterior probability, $\Pr(W|P)$, is what we really want to know – it tells us the chance that patient is sick given the evidence of clinical tests and prior histories. If we change the population, either by changing the prevalence of disease or the FPF, there will be an effect on how we should interpret data via the likelihood when making a diagnosis, as we found in Example 4.1.1.

Posterior probabilities can be used in all types of decision-making from patient diagnosis to parameter estimation and image reconstruction [27].

4.1.3 Accuracy

Accuracy is perhaps the most familiar metric of test performance because it is used in daily life to make statements about correctness of a measurement. Accuracy also has a

technical definition: it is the total number of correct test responses divided by the total number in the population, $(N_{TP} + N_{TN})/N$. Using probabilities, we can say accuracy is the probability of a diseased patient AND a positive test result (this becomes the intersection of the sets, $\Pr(WP)$, when written as a probability) plus the probability of a non-diseased patient AND a negative test result, $\Pr(W^c P^c)$. Let's see if we can get these two statements to equal by applying the definitions in Fig 4.1 and the rules of probability from Appendix B. In its many forms,

$$\begin{aligned}
\text{Accuracy} &= \Pr(WP) + \Pr(W^c P^c) \\
&= \Pr(P|W)\Pr(W) + \Pr(P^c|W^c)\Pr(W^c) \\
&= \frac{N_{TP}}{N_W}\frac{N_W}{N} + \frac{N_{TN}}{N_{W^c}}\frac{N_{W^c}}{N} = \frac{N_{TP} + N_{TN}}{N} \\
&= \text{sensitivity} \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})\ .
\end{aligned} \tag{4.1}$$

We need to be suspicious of accuracy as a reliable quality metric because of the way it ignores prevalence. Although prevalence shows up in Eq (4.1), ultimately $\Pr(W)$ and $\Pr(W^c)$ are eliminated from the equation.

A more concrete example of the suspicious nature of accuracy as an evaluation metric is found by imagining a completely worthless diagnostic test that simply decides everyone tested is negative regardless of the data. Hopefully you agree this is a worthless test. Do we get an intuitive result from calculations of accuracy all the time? The answer is NO. In the first case, assume the prevalence of the disease in the population to be tested is $\Pr(W) = 0.50$. Using the third line of Eq (4.1), we find an accuracy of $(0 + 0.50)N/N = 0.50$, which is intuitive but nevertheless disturbing that a worthless test is still 50% accurate! Now let the prevalence fall to $\Pr(W) = 0.01$. The accuracy becomes $(0 + 0.99)N/N = 0.99$, which suggests high performance. Accuracy is most intuitive near $\Pr(W) \simeq 0.5$. Fine, is there an alternative?

Yes. The gold standard for evaluating the performance of any binary tests, such as those often applied in medical diagnosis, is Receiver-Operating Characteristic (ROC) analysis. The *area under the ROC curve (AUC)* ranges between 0.5 (worthless test) and 1.0 perfect test, and therefore is often used to define accuracy in a manner that is not biased by prevalence. In fact, it can be shown that AUC is the accuracy defined above but averaged over all possible prevalence values. First let's review the fundamentals of hypothesis testing for statistical decision making.

4.2 Binary hypothesis testing

Measurements are made to inform decisions, so the ultimate assessment of measurement-system performance is the quality of the decisions that result once observers consider the data. In this section, we examine binary decisions as in §4.1 where the question was whether a lesion was malignant or benign. In example 4.1.1, two tests were evaluated by comparing PPV statistics estimated from prior knowledge of disease prevalence (prior) and patient data (likelihood). Now we discuss hypothesis testing that informs decisions by asking whether a new data point "belongs" to a class of objects characterized by its class distribution.

4.2.1 A diagnostic problem

A physician is tasked with deciding if a tumor is dangerous enough to warrant risky aggressive treatment. This particular tumor has several diagnostic features suggesting it could be either benign or malignant. Equivocation from a lack of convincing evidence is a good reason to order further tests, and so an ultrasonic pulsed-Doppler study is ordered. Enhanced blood flow in the lesion is a sign of a malignancy. Pulsed-Doppler techniques measure red blood cell (RBC) velocity based on changes in echo patterns observed between sequential pulse transmissions characteristic of blood flow or perfusion. We will view velocity measurements as a univariate normal random variable.

The true blood velocity in the lesion is θ_0, and our estimate of blood velocity is $\hat{\theta}$. Obviously, we want measurements that provide $\hat{\theta} = \theta_0$, so our analysis begins with a model of how data are acquired and processed to give velocity estimates. Let's model the tumor as a wide-sense stationary (WSS) random variable with fixed glandular tissue scatterers and moving RBC scatterers using the object function $f(\mathbf{x}, \theta_0(\mathbf{x}, t))$. This f is a function of both space via $\mathbf{x}$ and true RBC velocity via θ_0, which itself is a function of space and time. In a later section, we will model this process to make it less abstract.

The echo signal from this object function is found using a LSI acquisition operator $\boldsymbol{\mathcal{H}}$,

$$g(t) = \boldsymbol{\mathcal{H}}\{f(\mathbf{x}, \theta_0(\mathbf{x}, t))\} + e(t) \quad \text{expressed in discrete form using} \quad \mathbf{g} = \mathbf{Hf} + \mathbf{e}\,. \tag{4.2}$$

The function $e(t)$ is acquisition noise. Like $\mathbf{g}$, $\mathbf{e}$ is an $M \times 1$ noise vector, $\mathbf{f}$ is $N \times 1$ and so $\mathbf{H}$ is $M \times N$. Noise samples are uncorrelated with each other and independent of the signal, $\mathbf{Hf}$. Noise samples are drawn from a zero-mean normal distribution, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Since both object and noise vectors are multivariate normal (MVN) processes, so is $g(t)$ despite being filtered by measurement matrix $\mathbf{H}$.

Next, we pass $\mathbf{g}$ through a display-stage operator that estimates velocity; i.e.,

$\hat{\theta} = \boldsymbol{\mathcal{O}}_D\{\mathbf{g}\} = \boldsymbol{\mathcal{O}}_D\{\mathbf{H}\,\mathbf{f} + \mathbf{e}\}$. $\boldsymbol{\mathcal{O}}_D$ is the operator that converts echo signals into the blood velocity estimates that are displayed for the observer/physician. To be effective, the blood velocity estimator must separate the blood signal from the non-moving scatters and the noise. We will discuss $\boldsymbol{\mathcal{O}}_D$ later in the chapter, but for now we see it as a way to convert echo signals into velocity estimates, both are random variables. Note that the echo-data model has changed since the discussion in earlier chapters by including object movement and acquisition noise.

Although we want estimates $\hat{\theta}$ that equal the true velocity θ_0, we understand there will be variability from estimation uncertainty. Some level of variability must be tolerated provided it does not significantly degrade our clinical assessments. Estimation variability, defined as $|\hat{\theta} - \theta_0| > 0$, occurs for many reasons, including the variability in true blood velocity within patient anatomy, adjacent tissue movements and acquisition noise. The measurement data combines these effects, which masks diagnostic information related to our task of estimating blood flow and leads to decision errors.

Errors

We can decide if some level of estimation error is acceptable by statistically testing the hypothesis that $\hat{\theta} = \theta_0$ assuming we have a model of estimate distributions. Assume velocity θ is a normal random variable, as illustrated in Fig 4.2, parameterized by the true velocity θ_0 and variance σ^2,

$$p(\theta; \theta_0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(\theta-\theta_0)^2/2\sigma^2} \ . \tag{4.3}$$

The hypothesis to be tested is that estimate $\hat{\theta}$ is a sample from the $p(\theta)$ distribution of Eq (4.3). Alternatively, we may test whether $(\hat{\theta} - \theta_0) = 0$ within some bounds defined by σ^2. The notation used here corresponds to that of Eq (B.10) by equating r.v. θ with x.

Let α be the probability of erroneously rejecting the null hypothesis (type I error) when in fact it is true.[1] You can see from the shaded regions in the pdf plot of Fig 4.2 that α depends on the thresholds set for accepting or rejecting the null hypothesis; the threshold values are constants $\theta_{1-\alpha/2}$ and $\theta_{\alpha/2}$. The associated probabilities $\Pr(\cdot)$, cumulative distributions $P(\cdot)$, and pdfs $p(\cdot)$ are

[1] Type I errors are false positives. To understand the terminology, suppose $\hat{\theta}$ belongs to $p(\theta)$ but has a value that falls outside the region of acceptance. Because it falls outside the acceptance region, a decision to reject the null hypothesis is a "positive" result and, in this case, a false-positive result.

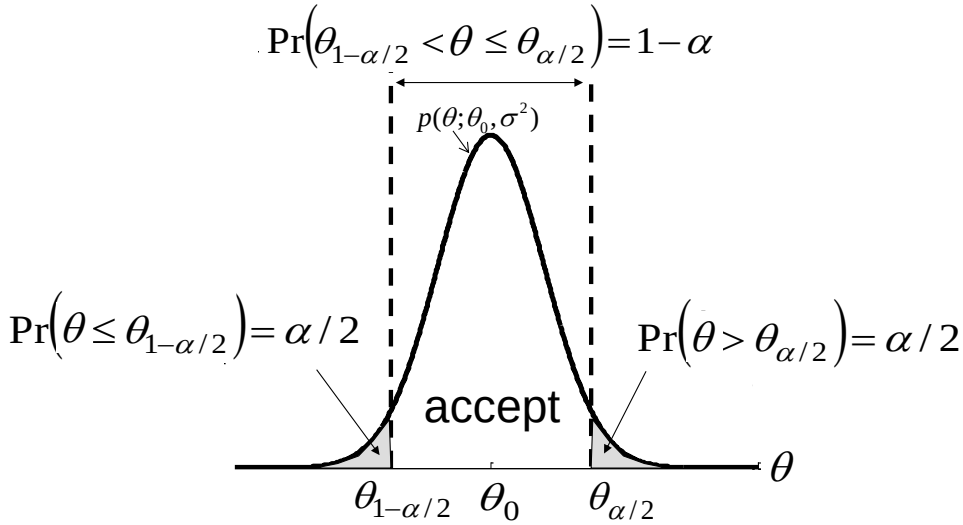


Figure 4.2: Illustration of the acceptance region for normally distributed estimates of blood velocity, θ. A test sample $\hat{\theta}$ that belongs to the $p(\theta;\theta_0,\sigma^2)$ distribution is expected to fall outside the acceptance region (shaded area) to give a type I (false positive) error with probability α.

$$
\begin{aligned}
\Pr(\theta > \theta_{\alpha/2}) &= 1 - P_\theta(\theta_{\alpha/2}) = \int_{\theta_{\alpha/2}}^{\infty} d\theta\, p(\theta) \triangleq \alpha/2 \;=\; \text{right shaded area} \\
\Pr(\theta < \theta_{1-\alpha/2}) &= P_\theta(\theta_{1-\alpha/2}) = \int_{-\infty}^{\theta_{1-\alpha/2}} d\theta\, p(\theta) = \alpha/2 \;=\; \text{left shaded area} \\
\Pr(\theta > \theta_{\alpha/2}) + \Pr(\theta < \theta_{1-\alpha/2}) &= \alpha \;=\; \text{probability of a type I error.} \qquad (4.4) \\
\Pr(\theta_{1-\alpha/2} < \theta \le \theta_{\alpha/2}) &= P_\theta(\theta_{\alpha/2}) - P_\theta(\theta_{1-\alpha/2}) \\
&= \int_{\theta_{1-\alpha/2}}^{\theta_{\alpha/2}} d\theta\, p(\theta) = 1-\alpha = \text{prob correctly accepting null hypoth}
\end{aligned}
$$

Let's assume the patient being examined has a benign tumor and no blood flow enhancement. Then $1-\alpha$ is the true negative fraction, TNF. As we saw from the Bayesian discussion in §4.1, TNF + FPF = 1 and so α = FPF in this example. Of course, if $\hat{\theta}$ does not belong to $p(\theta)$ and its value falls outside the acceptance region, we would score that decision as a true positive.

Selection of decision thresholds, $\theta_{\alpha/2}$, are not absolute. The are set based on the level of risk the decision maker is willing to assume given the cost of making errors. We refer to the example above as a two-sided hypothesis test because errors can be made with measurements that are larger and smaller than θ_0. However, in our example, we should select a one-sides test because the malignant condition is found only when the velocity is

greater than normal and not less than normal.

If the set threshold is increased (moved away from θ_0), we reduce α making the test more specific (higher true negative fraction) but less sensitive to malignancies (lower true positive fraction). In contrast, lowering the threshold (moved toward θ_0) increases α, making the test more sensitive and less specific. We have not specified the enhanced flow range because we don't know where it is for each patient. We only know that "enhanced flow" is more likely to indicate a malignant lesion. Since we know the distribution of normal flows for a population, experience and prior information about the patient can help us decide where to set the threshold. This is why physicians get the big bucks!

The example in this section involves one known distribution of benign lesions. We decide that estimates falling within the acceptance region belongs to this lesion class and those falling outside do not – a binary Yes/No decision. Alternatively, we may know the distributions of two classes of data, say breast fibroadenomas (benign lesions) and infiltrating ductal carcinomas (malignant lesions), and are asked to classify test data as belonging to one of the two classes. This is a two-hypothesis binary decision.

4.3 Two-hypothesis binary decisions

Assume we now have the distributions for two classes of blood velocity data measured from patient lesions. Let the negative hypothesis, H_0, represent benign lesions as in the last section. Its pdf $p(\theta|H_0)$ is the probability density of velocity θ conditioned on the patient having a benign lesion. We also have $p(\theta|H_1)$ as the pdf for θ conditioned on the alternate hypothesis, H_1, that patients have a malignant lesion. No other possibilities are considered at this time. Assume we have equal representation of patients from both hypotheses and the pdfs have nonequal means but equal variances as in Fig 4.3. That is, $p(\theta|H_0) = p(\theta; \theta_0, \sigma^2)$, $p(\theta|H_1) = p(\theta; \theta_1, \sigma^2)$, and $\Delta\theta = \theta_1 - \theta_0$ where $\theta_1 \geq \theta_0$. The diagnostic value of the test is related to the magnitude of $\Delta\theta/\sigma$, which is a signal-to-noise-like value. The distribution overlap shows there will be decision errors; in fact, two types of errors are possible.

With two hypotheses, each having a distribution, a decision threshold θ_t must be set. As shown in Fig 4.3, the decision will be D_0 for a patient with estimate $\hat{\theta} < \theta_t$. D_0 is a decision that the patient belongs to group H_0, which we base on test result $\hat{\theta}$ being below the set threshold. Patients for whom $\hat{\theta} > \theta_t$ are classified as belonging to the group from H_1, and we decide D_1 when $\hat{\theta} > \theta_t$. The binary decision function is to choose

$$D_j(\hat{\theta}; \theta_t) \;\text{ for }\; j = \text{step}(\hat{\theta} - \theta_t) \;,$$

where we know that $\text{step}(x) = 0$ when $x < 0$ and $\text{step}(x) = 1$ when $x > 0$. Human

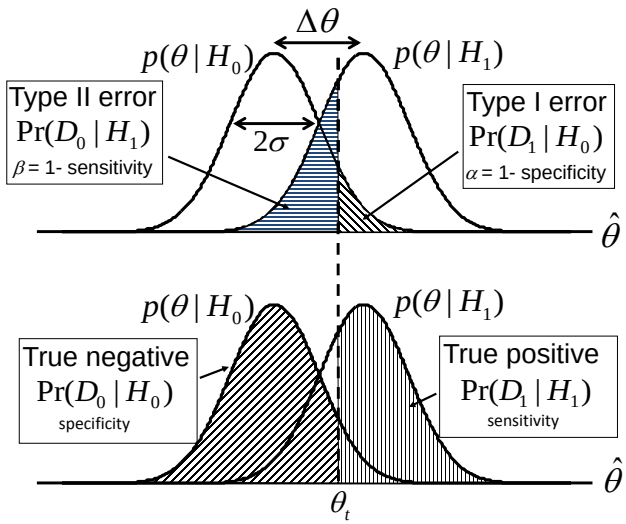


Figure 4.3: Conditional probability density functions $p(\theta|H_i)$ under two hypotheses H_0 and H_1 are plotted. The normal pdfs have different means and equal variances σ^2. (Top) For threshold θ_t, the integral of $p(\theta|H_0)$ from θ_t to ∞ is the probability of a type I error (false positive, α). The integral of $p(\theta|H_1)$ from $-\infty$ to θ_t is the probability of a type II error (false negative, β). Shaded areas in the bottom graphs yield probabilities that we consider the sensitivity and specificity of the test.

decision makers also follow an algorithm, although we are not completely sure of the details.

For two classes of patients, there are two possible decisions and four outcomes as illustrated in Fig 4.3, which are quantified by the following probabilities, cdfs, and pdfs.

$$\begin{aligned}
\Pr(D_0|H_0) \quad &: \quad P(\theta_t|H_0) = \int_{-\infty}^{\theta_t} d\theta\, p(\theta|H_0) \qquad \text{true negative decision probability} \\
\Pr(D_1|H_0) \quad &: \quad 1 - P(\theta_t|H_0) = \int_{\theta_t}^{\infty} d\theta\, p(\theta|H_0) \qquad \text{false positive, type I error, } \alpha \qquad (4.5) \\
\Pr(D_0|H_1) \quad &: \quad P(\theta_t|H_1) = \int_{-\infty}^{\theta_t} d\theta\, p(\theta|H_1) \qquad \text{false negative, type II error, } \beta \\
\Pr(D_1|H_1) \quad &: \quad 1 - P(\theta_t|H_1) = \int_{\theta_t}^{\infty} d\theta\, p(\theta|H_1) \qquad \text{true positive decision probability}
\end{aligned} \tag{4.6}$$

You know it is an error probability being calculated when the index on the decision and hypothesis are not the same. Between these equations and Fig 4.3 you should have a clear mental image of probabilities associated with decisions. Also, from the first axiom

of probability,

$$\Pr(D_0|H_0) + \Pr(D_1|H_0) \;=\; 1 \;=\; \Pr(D_0|H_1) + \Pr(D_1|H_1)\ .$$

In words, we have the specificity plus the probability of a type I error equals one. Also, the sensitivity plus the probability of a type II error equals one. This reminds us that by setting thresholds we necessarily make tradeoffs for fixed class distributions. Measurement quality is what determines the distributions, since precise measurements minimize the σ parameters and maximize measurement sensitivity for a task via the difference between means, $\Delta\theta$. The best decisions begin with high-quality measurements that generates the greatest $\Delta\theta$ and smallest $\sigma_t^2 heta$. Further, the best decision makers seek to know these distributions so they can assess the risks of errors for the task conditions as they set appropriate thresholds.

4.3.1 Figures of merit

Risk assessment is tricky business because of the large number of factors that must be considered for each patient. For example, a radiologist viewing a breast mass in a mammogram might dismiss the finding as insignificant if the patient is 30 years old with no family history of breast cancer and no other indications. The radiologist might order a followup image in a year to be sure. However, the same radiologist viewing essentially the same mass in a 60-year-old patient with a family history and genetic markers for breast cancer may be quite concerned and immediately order a biopsy procedure. How can we assess the technology without bringing to bear all the factors that determine the decision threshold for an individual patient? There are statistics we can measure for use as a scalar figure of merit (FOM) that include the task and the measurement instruments but not patient specifics that imply a decision threshold.

4.3.2 Detectability index

If we know the distributions for two classes of patient data and these distributions are approximately normal, we can say the best-performing measurements have the largest separation of means with respect to the deviations about the means, Fig 4.3. One common FOM is the detectability index [28],

$$d'^2 = \frac{(\Delta\theta)^2}{\sigma^2} \quad 0 \le d'^2 \le \infty \quad \text{for equal-variance normal distributions.} \tag{4.7}$$

When $d' = 0$, the distributions overlap for equal variance because $p(\hat{\theta}|H_0) = p(\hat{\theta}|H_1)$; this test has no discriminability. Discrimination increases with d'^2 by increasing the difference

between means or decreasing the population variances. These features are controlled by measurement instrument properties and population selection.

If the bi-normal distributions have unequal variances then

$$d'^2 = \frac{2(\Delta\theta)^2}{\sigma_0^2 + \sigma_1^2} \quad \text{for unequal-variance normal distributions.} \tag{4.8}$$

If the distributions are non-normal then d' cannot be fully trusted in the sense that performance is not always monotonic with d'. Also, d' offers no insights about how to account for different levels of disease prevalence in the population. Remember the problems we had with accuracy earlier in the chapter?

4.3.3 Receiver operating characteristic analysis

Receiver operating characteristic (ROC) curves provide an objective measure of task performance for any measurement system that leads to a binary decision. An ROC curve requires measurements of TPF = sensitivity = $p(D_1|H_1)$ and measurements of FPF = 1-specificity = $p(D_1|H_0) = \alpha$ for all values of θ_t. Using Fig 4.3, we see that both class distributions are integrated to the right of θ_t to arrive at (TPF(θ_t), FPF(θ_t)) pairs. An ROC curve is a plot of TPF as a function of FPF for $-\infty \leq \theta_t \leq \infty$.

This process is illustrated in Fig 4.4. Beginning on the right side of the decision axis θ (position 1) where TPF is small and FPF is even smaller we estimate TPF and FPF values. This process is repeated as θ_t is swept right to left over θ. Both axes of an ROC curve are bounded by 0 and 1, which are the minimum and maximum values of any probability function.

If the distributions completely overlap, then θ is a worthless test for that task. The ROC curve in that extreme situation is a diagonal line and the area under the ROC curve, or AUC = 0.5. Conversely, if the two distributions are completely separable, then θ is perfectly discriminating. The ROC curve for this happy extreme is given by the function step(1-specificity) so that AUC = 1.0. Consequently, $0.5 \leq \text{AUC} \leq 1.0$ is used as a performance metric when evaluating diagnostic devices. Values for AUC > 0.85 are considered in the range of acceptable diagnostic performance, although the criterion varies widely with the application. AUC is the probability that randomly-drawn values of data θ, given H_1, will be larger than those given H_0.

We can compute detectability from AUC, although this quantity is labeled d_a and not d' described in Eqs (4.7) and (4.8). d_a refers to detectability found from AUC and is given by [1],

$$d_a = 2\text{erf}^{-1}(2\text{AUC} - 1) \ , \tag{4.9}$$

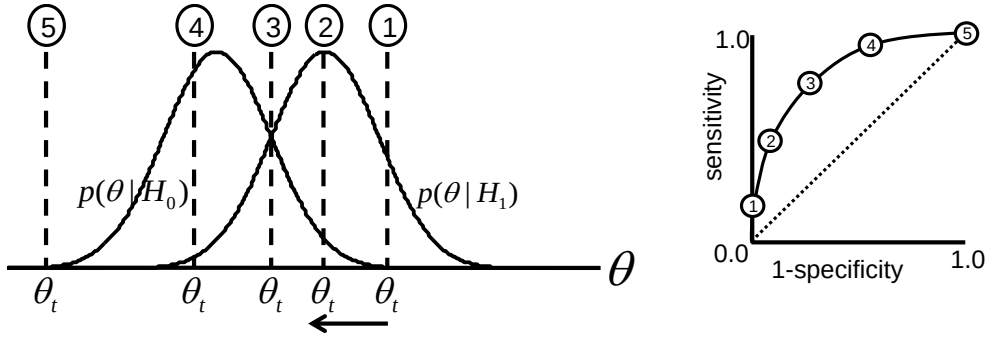


Figure 4.4: (left) Illustration of how threshold θ_t is varied over decision variable θ from right to left. At each θ_t, TPF and FPF are estimated and plotted against each other to form the ROC curve (right).

where $\text{erf}^{-1}(\cdot)$ is the inverse error function. d_a can always be found whenever an ROC curve is available, whereas d' is meaningful only for normally-distributed decision variables $\theta|H_0$ and $\theta|H_1$.

We prefer d_a estimates over AUC when we wish to estimate the efficiency of one method relative to another. For example, the efficiency of method 2 relative to method 1 is given by, [29, 28]

$$\eta = \frac{d_{a,2}^2}{d_{a,1}^2} \ . \tag{4.10}$$

ROC analysis is the gold standard for performance assessment. Standards of practice in research and regulatory agencies like the FDA often require ROC analysis as rigorous evidence of task performance. It is well worth your time to investigate ROC techniques if faced with the need to rigorously demonstrate performance or to compare performance among competing methods.

Fortunately, the late Professor Charles Metz at the Kurt Rossmann Laboratories within the University of Chicago has a website offering free of charge a vast array of ROC software via

`http://www-radiology.uchicago.edu/krl/roc_soft6.htm`

If you need to use ROC analysis for a study, I highly recommend that you investigate this resource. There are many details that I have not touch upon but that must be addressed in practice. Primarily these are details related to estimates of the errors in AUC, power

calculations and other important metrics that are difficult to compute correctly when correlations exist among the data used in the study. For example, if human observers are making decisions and they each view the same images, the responses are correlated and correct error bar calculations must take those correlations into account. There are alternative ROC techniques that allow for parameter uncertainties, resampling methods, and other options. So read the documentation and associated literature carefully before using the software.

4.4 Central limit theorem

We have been assuming that the decision variables are normally distributed, and I'm sure some readers are wondering why we can make that assumption. This invokes a very powerful fundamental theorem in probability theory.

The *central limit theorem* states that the distribution of the sum of a large number of independent, identically distributed variables will be approximately normal regardless of the underlying distribution of those variables. This theorem magnifies the importance of the normal distribution especially when modeling measurements composed of the summation of many random variables.

In observer-performance experiments, observers (human or computational) view a large number of data samples, e.g., images or time series, as they generate the random variable we call their decision. Provided the task required of the observer includes "many" i.i.d. samples, we can be reasonably sure the decision variable will be normally distributed. Exceptions to this situation include those where the data are from two separable classes instead one, so the decision for that hypothesis are bimodal.

Texts show proofs of the central limit theorem, which are interesting to follow for the insights provided. However, I find numerical demonstrations are quite instructive about how quickly a distribution of summed data converges to a normal distribution. I ask you to examine this aspect in the homework.

4.5 Statistical properties of acquired and displayed data

The previous sections in this chapter introduced ideas related to statistical decision making. This sections describes how statistical properties of data change at different stages of the measurement process. Assuming readers have examined Appendix B, especially the sections on functions of univariate §B.18.1 and multivariate §B.18.2 random

Appendix B

B. Review of Probability and Random Processes

B.1 Introduction

You might be surprised to hear that experts are sharply divided on the definition of "probability". Despite the randomness it describes and a dual definition, probability theory is an exact science built on fundamental principles, a few of which we describe in this appendix. The challenge is always to interpret the rules and then correctly apply them to each problem.

Frequentists define probability as the frequency of event occurrence that is estimated experimentally. For them, data tell the whole story. In contrast, *Bayesians* view probability as the degree of belief in a state. They might define a prior probability before taking data and then update that belief with a posterior probability after viewing some experimental data. As you might imagine, each view has its strengths and weaknesses. We shall remain agnostic here, using each definition as it suits us.

Let's introduce the frequentist position with an example. Assume a sample of biological media is exposed to light. The absorption of light by the sample is *stochastic*, meaning it is randomly determined. Intuition about this particular stochastic process builds as we measure the distribution of light absorption. First a little formal notation.

Let ω indicate a molecule is present that could absorb a photon. Then $\Omega = \{\omega\}$ defines the *sample space* of size N for the light-absorption experiment, i.e., there are N absorbing molecules in the medium. Let E be an event that a photon is absorbed. E is

within a subset of Ω, and $n(E)$ is the number of molecules in the *event space* $\mathtt{S} = \{E\}$. Finally, Pr is a measure of the probability of the events, the chance a photon will be absorbed. The triplet $(\Omega, \mathtt{S}, \Pr)$ is a particular type of *measure space* called the *probability space* of the problem.

Definition B.1.1. *The probability of E in probability space $(\Omega, \mathtt{S}, \Pr)$ is*

$$\Pr(E) = \lim_{N\to\infty} \frac{n(E)}{N} \ .$$

The presence of absorbing molecules is defined by Ω, a list of opportunities for a photon to encounter an absorbing molecule is defined by $\mathtt{S}$, and the likelihood that such opportunities occur are defined by Pr. This very simple equation is packed with information about the physics of the experiment, which is why many find it intuitive if not always practical. There can be questions about whether the ratio converges in the limit using experimental data, a discussion we leave to probability theorists. With these definitions, we can define the probability axioms.

B.2 Probability axioms

Definition B.2.1. .

Axiom 1 $\Pr(E) \in \mathbb{R}, \Pr(E) \geq 0$ *(probabilities are nonnegative real numbers)*

Axiom 2 $\Pr(\mathtt{S}) = 1$ *(one of the events is certain to occur)*

Axiom 3 $\Pr\left(\bigcup_{i=1}^{M} E_i\right) = \sum_{i=1}^{M} \Pr(E_i)$ *(discussed below)*

If we assume $E_1, E_2, \ldots, E_M$ are mutually exclusive events, e.g., a photon with wavelength λ_1 cannot also have wavelength λ_2, then the third axiom states that the probability of at least one of these events occurring is the sum of their respective probabilities.

'$E \cup F$' denotes the union of events (E or *F where $\cup \equiv$ 'or'), while '$\bigcup_{i=1}^{M} E_i$' denotes the union of M events, E_1* or *E_2* or *$\ldots$* or *E_M. Also 'EF' denotes the intersection of events (E* and *F where $\cap \equiv$ 'and').*

Example B.2.1. *A fair die with 6 sides generates events $E_1 = \{1\}, E_2 = \{2\} \ldots$ with probability $\Pr(E_i) = 1/6$. From axioms 3, since the events are mutually exclusive (each roll can only generate one value), we can find the probability of rolling an odd number using $\Pr(\{\text{odd}\}) = \Pr(\{1, 3, 5\}) = 1/6 + 1/6 + 1/6 = 1/2$.*

B.3 Consequences of the axioms

- If E^c is the complement of event E (E^c means 'not E'), then

$$\Pr(\mathtt{S}) = 1 = \Pr(E \cup E^c) = \Pr(E) + \Pr(E^c) \text{ so that } \Pr(E^c) = 1 - \Pr(E) \ .$$

- If event E is contained in event F, then $\Pr(E)$ is less than or equal to $\Pr(F)$, i.e.,

$$\text{if} \quad E \subset F \ , \ \text{ then } \ \Pr(E) \leq \Pr(F) \ .$$

- The probability of the union of two events $E \cup F$ that are not mutually exclusive is

$$\Pr(E \cup F) = \Pr(E \cup E^c F) = \Pr(E) + \Pr(E^c F) \ .$$

  We need an expression for the last term. For that, note that $F = EF \cup E^c F$ (See the illustration in Fig B.1). From Axiom 3,

$$\Pr(F) = \Pr(EF) + \Pr(E^c F) \ .$$

  Combining the two equations above,

$$\Pr(E \cup F) = \Pr(E) + \Pr(F) - \Pr(EF) \ . \tag{B.1}$$

  This is the expression used when events are mutually dependent.

Example B.3.1. *Let E be the event that a patient is sick with heart disease, F be the event a patient is sick with cancer, and EF be the event a patient is sick with both diseases, i.e., $E \cap F$. Further, $n(E) = 5$ patients have heart disease out of $N = 9$ total patients, $n(F) = 5$ patients have cancer, and $n(EF) = 1$ patient has both diseases. Clearly* $\Pr(EF) = 1/9$. *Eq (B.1) gives the intuitive probability that patients in this event space that are sick with either disease is one,* $\Pr(E \cup F) = 5/9 + 5/9 - 1/9 = 1$. *The last term subtracts the probability of patients with both diseases, which avoid double counting.*

B.4 Conditional probability

Why is it that dice seem to offer the simplest examples? Let E be the event that the sum of two throws of a die is 7 and F is the event that the first throw was 3. The probability that E occurs given that (or conditional upon) event F has occurred is defined as $\Pr(E|F)$.

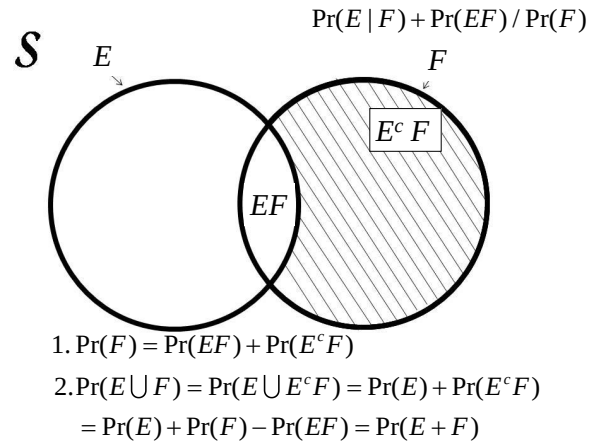


Figure B.1: An illustration of a non-mutually-exclusive event space S from §B.3.

It's easy to see that $\Pr(F) = 6/36 = 1/6$ because $\{F\} = \{(3,1)(3,2)(3,3)(3,4)(3,5)(3,6)\}$ has six possibilities and the sample space S includes $6^2 = 36$ possibilities Also $\{E\} = \{(1,6)(6,1)(2,5)(5,2)(3,4)(4,3)\}$, and so $\Pr(E) = 6/36 = 1/6$. Looking at the set $\{E\}$, the intersection EF is just one event, $(3,4)$. Hence $\Pr(EF) = 1/36$. Also $\Pr(E|F) = 1/6$, which is found by limiting the choices not to the whole event space but only to those event in the set $\{F\}$ where there is just one of six possibilities where the dice sum to give a 7.

B.4.1 $\Pr(EF)$ versus $\Pr(E|F)$

To emphasize the difference between these quantities, consider a different example illustrated quantitatively in Fig B.2. Follow along by counting squares. This event space has $N = 56$ elements on a 7×8 grid. E and F are events in two shaded regions that overlap (not mutually exclusive). We see that $\Pr(E) = \Pr(F) = 16/56$. Also $\Pr(E^c) = \Pr(F^c) = (56 - 16)/56$ and $\Pr(EF) = 4/56$. Finally

$$\Pr(E|F) = \frac{\Pr(EF)}{\Pr(F)} = (4/56)/(16/56) = 1/4 \ , \tag{B.2}$$

which is defined only when $\Pr(F) \neq 0$.

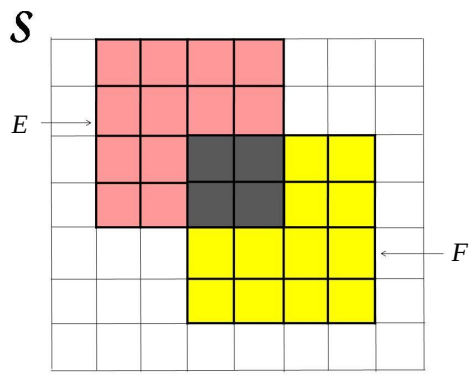


Figure B.2: Graphical representation of an event space.

B.5 Bayes' formula

From Eq (B.2), $\Pr(EF) = \Pr(E|F)\ \Pr(F)$ and $\Pr(FE) = \Pr(F|E)\ \Pr(E)$. Since $\Pr(EF) = \Pr(FE)$, then

$$\begin{aligned} \Pr(E|F)\Pr(F) &= \Pr(F|E)\Pr(E) \\ \Pr(E|F) &= \frac{\Pr(F|E)\Pr(E)}{\Pr(F)} \ , \end{aligned} \tag{B.3}$$

which is Bayes' formula. Decomposing $\Pr(F)$,

$$\Pr(F) = \Pr(FE) + \Pr(FE^c) \ , \tag{B.4}$$

and applying Bayes' formula, we find

$$\Pr(F) = \Pr(F|E)\Pr(E) + \Pr(F|E^c)\Pr(E^c) \ . \tag{B.5}$$

Generalizing to include all event possibilities within $\mathtt{S}$, viz., $E_1, E_2, \ldots, E_N$,

$$\Pr(F) = \sum_{i=1}^{N} \Pr(F|E_i)\Pr(E_i) \ .$$

Combining this result with Eq (B.3) yields a more general form of Bayes formula,

$$\Pr(E_j|F) = \frac{\Pr(F|E_j)\Pr(E_j)}{\sum_{i=1}^{N}\Pr(F|E_i)\Pr(E_i)} \ . \tag{B.6}$$

B.6 Independence

Events E and F are independent if

$$\Pr(EF) = \Pr(E)\Pr(F) \ .$$

Of course, if E and F are independent and mutually exclusive then the equation above still holds but equals zero. N events are independent if

$$\Pr\left(\prod_i E_i\right) = \prod_i \Pr(E_i) \ .$$

From Eq (B.3), we have for non-mutually exclusive but independent E, F that $\Pr(E|F) = \Pr(EF)/\Pr(F) = \Pr(E)\Pr(F)/\Pr(F) = \Pr(E)$, where conditioning makes no difference.

B.7 Distribution functions

Let event X be a random variable (r.v.) defined below and x one of the possible outcomes or *realizations* of X. Let X be defined for all $-\infty < x < \infty$; i.e., $X \in \mathbb{R}^1$. The *cumulative distribution function, cdf,* is

$$P(x) = \Pr(X \leq x)$$

Notice that P and Pr are different. The later quantity is a probability of an event defined as a random variable X, while the former is an accumulation of those probabilities over some r.v. range.
Properties of cumulative distributions:

- $P(x)$ is nondecreasing; if $a < x$ then $P(a) \leq P(x)$.
- $\lim_{x\to\infty} P(x) = 1$
- $\lim_{x\to-\infty} P(x) = 0$

B.8 Discrete random variables

If r.v. X can take on at most a countable number of values, then X is a *discrete r.v.* A discrete random variable is a measurement made on a sample contained in the sample

space, i.e., on $\omega \in \Omega$. For example, if a photon at wavelength λ is absorbed by a molecule able to absorb it, i.e., one in the subset of Ω that we label as event E, we say $X(E) = 1$, otherwise $X(E) = 0$. The absorption attribute of the molecule in this case is a yes/no label. In general, $X(E) = a$, where a is the value of the attribute. It can be a real number like the molecular weight of the molecule or a label as in this example. Valid random variables are *consistent* in the sense that $E = X^{-1}(a) \in \mathtt{S}$. That is, for a measurement to be a valid random variable on $(\Omega, \mathtt{S}, \Pr)$, we must have $X^{-1}(a)$ defined as an event in $\mathtt{S}$ so that its probability is defined by Pr. The mean of r.v. X is therefore

$$\text{mean of } X = \sum_{a \in X(\Omega)} a \Pr(X^{-1}(a)) \ .$$

We will discuss the mean of a r.v. later in the discussion of moments of distributions.

B.8.1 Probability mass function

For X discrete, we define the *probability mass function* (pmf) $p_X(x_n)$ or simply $p(x)$ or $p(n)$, as

$$p(x) = \Pr(X = x) \ .$$

Careful! I use $p(x)$ for the pmf here and then $p(x)$ later for the probability density function (pdf) in the continuous r.v. discussion below. The exact meaning depends on context, so there should be no confusion. The pmf can be thought of as a sampled subset of the pdf in some cases. Properties of probability mass function:

- $p(x)$ is positive for at most a countable number of values of x; i.e.,
  $p(x_n) \geq 0 \ \ n = 1, 2, \ldots$
  $p(x) = 0$ for all other values of x.
- $\sum_{n \in \mathtt{S}} p(x_n) = 1$.
- $$P(x) = \sum_{x'_n = -\infty}^{x} p(x'_n) \ .$$

See examples of $p(x_n)$ and $P(x)$ in Fig B.3 for a Poisson process, which is described in §B.8.2. These curves were generated using MATLAB functions (left) `p=poisspdf(x,10);` and (right) `P=poisscdf(x,10);` Another useful function is `R=poissrnd(lambda,M,N);` that generates a $M \times N$ matrix of "uncorrelated" Poisson random numbers using parameter λ. Sorry for the notational overlap, but the symbol λ is used to represent many different things.

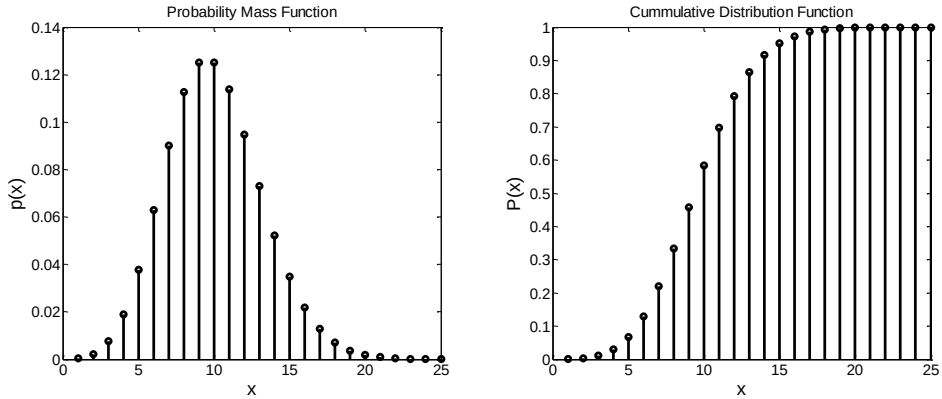


Figure B.3: (left) Probability mass function $p(x)$ and (right) cumulative distribution function $P(x)$ for a discrete Poisson random variable X, where the range of the variable is integers between $1 \leq x \leq 25$ and $\lambda = 10$. Notice $p(x)$ is slightly asymmetric. Also the mean is not at the peak value.

B.8.2 Poisson random variable

Let X be a Poisson r.v. in the space of non-negative integers $x = n$ with unitless parameter $\lambda > 0$. Switching to n to emphasize the discrete nature of this r.v.,[1]

$$p_X(n) = \Pr(X = n) = e^{-\lambda}\frac{\lambda^n}{n!} \quad \text{for } \; n = 0, 1, 2, \ldots \; \text{ and all } n \in \mathtt{S}. \tag{B.7}$$

A Poisson distribution is completely specified by its one parameter, λ. Operator notation $\mathcal{P}(\lambda)$ can also be applied to denote a Poisson process. $\mathcal{P}(\lambda)$ is used when we want to indicate the random process without giving specifics. The dependence of $p(n)$ on λ for a Poisson process is shown in Fig B.4.

Let's check to be sure $p(n)$ sums to one as it must to represent $\Pr(X = n)$. From probability axiom #2,

$$P(\infty) = \sum_{n=0}^{\infty} p(n) = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} = e^{-\lambda}\,e^{\lambda} = 1\ .$$

Examining Fig B.4 we note that as λ increases the pmf broadens as the peak decreases, such that the area is always one.

[1]We read $p_X(x) = \Pr(X = x)$ as the pmf of r.v. X evaluated at the specific value $X = x$. In the discussion below, we will drop the subscript X for simplicity, i.e., $p_X(x) \triangleq p(x)$. Explicit notation returns where needed for clarity.

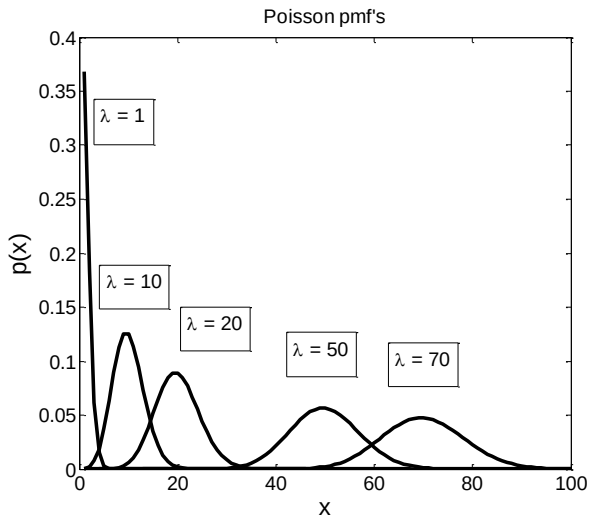


Figure B.4: Poisson probability mass functions for for various values of λ. Notice that the r.v.s are constrained to be positive integers (despite the continuous curves plotted) and that pmf symmetry increases with λ.

Poisson is the appropriate process for describing many common stochastic processes. It is widely used to model photon noise (microscopy, x-rays, nuclear imaging) since photons are positive, countable, discrete-number events. Other common phenomena modeled as a Poisson process include

- the number of cells in unit volume of tissue
- number of radioactive decays per second for an isotope sample
- number of x-ray photons absorbed in a detector area per second.

Evans (Ch 26 [6]) illustrates the principles underlying a Poisson process by deriving its frequency distribution from consideration of radioactive nuclear decay. I'll modify his example to consider biological cell proliferation. A goal of the following is to enumerate assumptions required to derive the pmf and show how they enter the derivation.

Example B.8.1. *Let event X be the number of cells that undergo division in a culture sample during measurement time interval T, where* $\mathtt{S} \in$ *positive integers. We assume that*

1. *the rate of cellular proliferation is the same for all cells in an experiment*
2. *each cell proliferates independently of the others*

3. *cell proliferation rate is constant over the measurement time. That generally means the reproductive lifetime of cells should be much longer than the observation time of the experiment*

4. *the number of cells and the observation time intervals are both large to obtain statistical averages that approach the ensemble.*

If positive constant λ' *is the average rate of cell division for a given culture, then* $\lambda' T$ *is the probability that a cell will divide in duration* T*, viz.,* $\Pr(X(T) = 1) = \lambda' T$*. Reducing the time interval to the infinitesimal* dt*,* $\Pr(X(dt) = 1) = \lambda' dt$ *and we can say that* $\lambda' dt \ll 1$*. Also the probability of observing two or more cells dividing during* dt *becomes much less than that of observing one division; i.e.,* $\Pr(X(dt) = 1) \gg \Pr(X(dt) = 2) \ldots$*. Therefore it is a very good approximation to set the probability of observing no cells dividing during interval* dt *as*

$$\Pr(X(dt) = 0) = 1 - \Pr(X(dt) = 1) = 1 - \lambda' dt \ .$$

From the third axiom of probability, the chance of finding n *cells dividing during* $t + dt$*, i.e.,* $\Pr(X(t + dt) = n)$*, may be expressed as a combination of the probabilities of finding* $n - 1$ *divisions in time* t *and one division during* dt*, i.e.,* $\Pr(X(t) = n - 1) \times \Pr(X(dt) = 1)$*, or* n *divisions during* t *and no divisions during* dt*, i.e.,* $\Pr(X(t) = n) \times \Pr(X(dt) = 0)$*. Thus*

$$\begin{aligned}
\Pr(X(t + dt) = n) &= \Pr(X(t) = n)\ \Pr(X(dt) = 0) + \Pr(X(t) = n - 1)\ \Pr(X(dt) = 1) \\
&= \Pr(X(t) = n)(1 - \lambda' dt) + \Pr(X(t) = n - 1)\, \lambda' dt \\
\frac{\Pr(X(t + dt) = n) - \Pr(X(t) = n)}{dt} &= \lambda' \big(\Pr(X(t) = n - 1) - \Pr(X(t) = n) \big) \\
\frac{d \Pr(X(t) = n)}{dt} &= \lambda' \left(\Pr(X(t) = n - 1) - \Pr(X(t) = n) \right) \ .
\end{aligned}$$

The solution to this first-order differential equation is

$$\Pr(X(t) = n) = p_X(n) = \frac{(\lambda' t)^n}{n!} e^{-\lambda' t} \ , \tag{B.8}$$

which is easily verified by substituting Eq (B.8) into the equation above it. The expressions for the Poisson pmf of Eqs (B.7) and (B.8) differ in several important ways. First, λ' *is not the unitless parameter* λ*; it is a rate constant and has the units of time*$^{-1}$*. In fact,* $\lambda \leftrightarrow \lambda' t$*. Second, the r.v.* X *in Eq (B.8) describes a time-dependent random process. The associated distributions of* X *in Eqs (B.7) and (B.8) are essentially the same except that the time dependence in Eqs (B.8) can be important for interpretation. Remember that* $x_n = n$ *are specific values that random variable* $X(t)$ *can take on, so each is a function of the deterministic independent variable* t*.*

A reason to go through the Poisson pmf derivation is to reinforce that many of the analysis tools used in research are strictly valid only under the stated assumptions. We must be careful to ensure the assumptions hold for our problem before putting our faith in what that tool is telling us. In the cell-growth problem above, if the observation time of the experiment is on the order of or greater than the lifetime of the cells, we violate the third condition necessary for our cell proliferation experiment to be modeled by a Poisson random process. Turns out that we can remedy the situation by adding a cell-death term in the differential equation. Ultimately, it is up to you to decide how egregious any violation really is when modeling your experiment as a Poisson r.v. Assumptions should always be carefully reviewed before models are adopted to avoid significant errors.

B.8.3 Mean of a Poisson random variable

The mean value of r.v. X, also called the expected value or the first moment, is found by applying the expectation operator to the random process. For a Poisson process,

$$\begin{aligned} \mathcal{E}X(t) &= \sum_{n:p(n)>0} n(t)\,p(n(t)) = \sum_{n=0}^{\infty} ne^{-\lambda}\frac{\lambda^n}{n!} = \sum_{n=1}^{\infty} ne^{-\lambda}\frac{\lambda^n}{n!} \\ &= \lambda e^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^{(n-1)}}{(n-1)!} = \lambda e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!} = \lambda e^{-\lambda}e^{\lambda} = \lambda\ , \end{aligned} \tag{B.9}$$

where we applied a change of variable, $m = n - 1$. The mean of a Poisson r.v. is parameter λ. For this discrete random process, the expectation operator is defined as

$$\mathcal{E} \triangleq \sum_{n:p(n)>0} p(n(t))$$

which is 'applied' to the r.v. X as expressed through its elements, $x_n = n$. Note that the sum is over X and not over t. The random process $X(t) = \{n(t)\}$ is a function of the nonrandom variable t.

The variance of the distribution can be found using

$$\mathrm{var}X(t) = \mathcal{E}\left\{\left(X - \mathcal{E}X\right)^2\right\} = \lambda\ ,$$

which you will be asked to show in a homework problem.

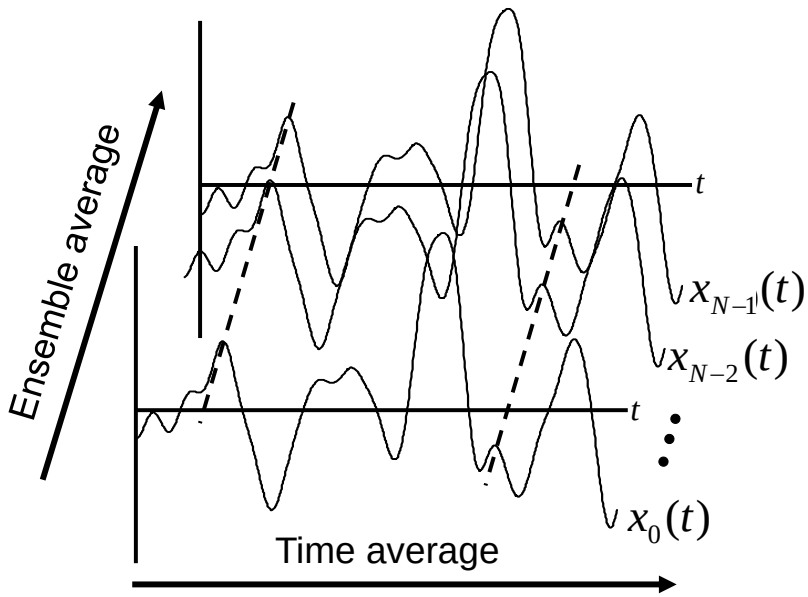


Figure B.5: Illustration of an ensemble of N random functions, $\{x_n(t)\}$.

B.9 Ensembles

An ensemble of waveforms results from repeating an experiment under the same conditions. If you are taking an ensemble average over a variable object, then all experimental variables are held constant and the experiment is repeated N times for different objects in the ensemble to give $\{x_N(t)\}$ shown in Fig B.5. If you wish to take an ensemble average over experimental variables like noise, then the object is held constant and the experiment is repeated N times to give $\{x_N(t)\}$. All deterministic components of the waveform are assumed to be identically reproduced with each experiment, while the random components adopt a new realization from some underlying random process. Of course, there can be systematic (nonrandom) errors, but we'll ignore those for now. The waveforms in Fig B.5 are similar because most of the waveform energy is from a deterministic signal component; i.e., the signal-to-noise ratio (SNR) is relatively large. Yet there are minor differences caused by additive noise, in this example. You may know from experience that it is wise to repeat each experiment several times and combine findings before reaching conclusions. It is necessary to estimate the uncertainty in a result, and indicate that information with error bars, if you wish to statistically compare data obtained under two situations. Building en ensemble of results by repeating experiments helps us (a) evaluate whether the deterministic components of the waveform

can be reproduced except for random error, and (b) obtain the data needed for noise reduction through signal averaging.

Eq (B.9) defines exactly what is meant by "finding the expected value of the Poisson r.v. $X(t)$." Yet the equation may not be very intuitive. The expectation operator $\mathcal{E}$ picks a point in time, multiplies the value of the r.v. at that time by the pmf for that value, and then sums products over the ensemble of waveforms but only for values *at that instant of time*. Contrast that situation with a time averaging procedure. One waveform is selected, each value is multiplied[2] by its pmf, and the results are summed over time but perhaps only for that one waveform. Ensemble averaging allows us to see if the pmf of the r.v. is a function of time, in which case its mean will also be a function of time. When we cannot obtain the repeated measurements required to compute an ensemble average, we settle for a time average. We will see below that waveforms from an *ergodic process* exhibit the property that ensemble averages approximately equal time averages.

B.10 Continuous random variables

X is a continuous r.v. if there exists a nonnegative function p that can be defined for all $x \in \mathbb{R}$ with the property that for any set of real numbers B,

$$\Pr(X \in B) = \int_B dx\, p(x) \ .$$

Because X is continuous, $p(x)$ has the units of $[x]^{-1}$, which is a reason to call it a *probability density function, pdf* for continuous r.v. X. The equation above states that the probability that X will be in subspace B can be obtained by integrating the pdf over B. Remember our earlier discussion about the strange behavior of delta functions? Well pdfs evaluated at specific values of the continuous variable $X = x$ also respond most predictably when placed inside integrals.

The *cumulative distribution function, cdf* of the continuous r.v. X is $P(x)$. It may be expressed using several different notations

$$P(x) = \Pr(X \in (-\infty, x)) = \Pr(X \leq x) = \int_{-\infty}^{x} dx'\, p(x') \qquad \text{for } \infty < x < \infty \ .$$

Differentiating gives

$$\left.\frac{dP(x')}{dx'}\right|_{x'=x} = p(x) \ .$$

[2]When we don't know the pmf, we often assume the random components of the time-series are statistically independent and equally likely so that $x_n(t) = 1/N$ where N is the number of points in the waveform.

Let's examine the probability of X over the small continuous interval dx centered at $X = x$,

$$\Pr(x - dx/2 \le X \le x + dx/2) = \int_{x-dx/2}^{x+dx/2} dx'\, p(x') \simeq dx\, p(x) \ .$$

In words, the probability of X over the interval dx around x is $dx\, p(x)$. Consequently, the physical interpretation of the $p(x)$ is that of a *measure* of how likely we are to find the continuous r.v. X *near* x. When summed over dx, the pdf times the interval, $dx\, p(x)$, is similar to the pmf, $p(x_n) = p(n)$ in Eq B.7. For the discrete r.v. $\Pr(X = x_n) = p(x_n)$, and so, like probability, the pmf is unitless. For the continuous r.v. $\Pr(X = x) = \int_x^{x+dx} dx'\, p(x') = dx\, p(x)$, and so the units of pdf are the same as those of x^{-1}.

Properties associated with continuous random variables:

$$\begin{aligned}
\Pr(X \in (-\infty, \infty)) &= \int_{-\infty}^{\infty} dx\, p(x) = 1 \ , \\
\Pr(a \le X \le b) &= \int_a^b dx\, p(x) = P(b) - P(a) \ , \\
\text{and when } \ a = b \quad \Pr(X = a) &= \int_a^a dx\, p(x) = 0 \ .
\end{aligned}$$

Analogous to the discrete r.v. case

$$\Pr(X < a) = \Pr(X \le a) = P(a) = \int_{-\infty}^{a} dx\, p(x) \ .$$

Let's examine a strange little problem to become familiar with these functions.

Example B.10.1. *Let X be a continuous r.v. with pdf*

$$p(x) = \begin{cases} a(x - x^2) + b & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find a and b given that $p_{min} = p(-1) = 0$.

Note that $\int_{-\infty}^{\infty} dx\, p(x) = 1 = \int_{-1}^{1} dx\, p(x)$. Therefore

$$\begin{aligned}
\int_{-1}^{1} dx\, a(x - x^2) + b &= 1 \\
\left[a\frac{x^2}{2} - a\frac{x^3}{3} + bx \right]_{x=-1}^{1} &= 1 \\
a &= \frac{3}{2}(2b - 1) \ .
\end{aligned}$$

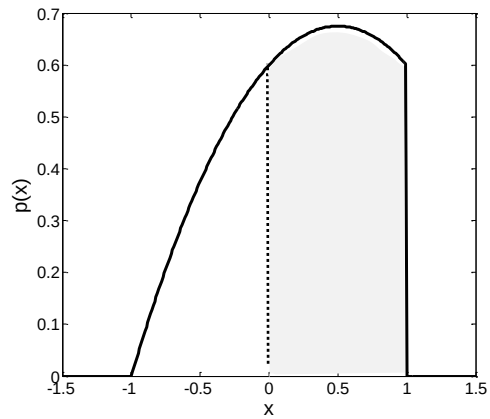


Figure B.6: Illustration of $p(x)$ from Example B.10.1. The shaded area is $\Pr(X \geq 0) = 0.65$ and the unshaded area is $P(0) = \Pr(X < 0) = 0.35$.

To find b, we use the fact that $p(-1) = 0$. *Consequently,*

$$\begin{aligned} p(-1) &= \left[\frac{3}{2}(2b-1)(x-x^2)+b\right]_{x=-1} = 0 \\ &= \frac{3}{2}(2b-1)(-2)+b = 0 \\ b &= 3/5 \ . \end{aligned}$$

Therefore, $a = 3/10$, $b = 3/5$, *and*

$$p(x) = \begin{cases} \frac{3}{10}(x-x^2)+\frac{3}{5} & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(b) Find $\Pr(X \geq 0)$.

$$\Pr(X \geq 0) = \int_0^1 dx\, p(x) = \int_0^1 dx\, \left[\frac{3}{10}(x-x^2)+\frac{3}{5}\right] = \frac{13}{20} = 0.65 \ .$$

Notice that the scale constant a and offset constant b are needed to satisfy the probability axioms of positivity and integration to one over the event space (Fig B.6). Occasionally, it is convenient to think of measurement data as a probability or probability density. In that case, like the example above, we need to normalize measurements so they behave as probabilities must.

Notational Comparison Summary

Same symbols mean different things depending on whether X is a DRV or CRV

Discrete Random Variables	Continuous Random Variables
$\Pr(X)$	$\Pr(X)$
$P(x) = P_X(x_n) = P(n) \triangleq \Pr(X \le x_n)$	$P(x) = P_X(x) \triangleq \Pr(X < x)$
$p(x) = p_X(x_n) = p(n) \triangleq \Pr(X = x_n)$	$p(x) = p_X(x) = dP(x')/dx'\|_{x'=x}$
$= P(n) - P(n-1)$	$dx\, p(x) = \Pr(x - dx/2 \le X \le x + dx/2) = dP(x)$
$\Pr(X \in (-\infty, \infty)) = P(x = \infty) = \sum p(n) = 1$	$\Pr(X \in (-\infty, \infty)) = P(x = \infty) = \int dx\, p(x) = 1$
$\Pr(X = a) = \sum_{n=a}^{a} p(n) = p(a)$	$\Pr(X = a) = \int_a^a dx\, p(x) = 0$

B.10.1 Univariate normal random variable

The best-known example of a continuous r.v. follows a normal (or Gaussian) distribution. The univariate normal probability density function (pdf) is [8]

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{for} -\infty < x < \infty \,. \tag{B.10}$$

To specify the pdf we might write $p_X(x; \mu, \sigma^2)$ where everything following the semicolon are constants. These details ensure readers understand that X is the random variable, that x is one realization of X and that the two parameters of this distribution are (μ, σ^2). This level of detail is usually understood, such that the pdf above can be specified simply by the operator $\mathcal{N}(\mu, \sigma^2)$. The scale factor preceding the exponential function is necessary to have the pdf integrate to one. Also the units are the inverse of the r.v. units as required. For example, if x has units of time then $p(x)$ has units of temporal frequency.

Let's be sure the pdf integrates to one. Changing the variable to $y = (x - \mu)/\sigma$, we obtain $dy = dx/\sigma$ and have the same limits. Consequently,

$$dI = dy\, p(y) = \frac{1}{\sqrt{2\pi}}\, dy\, e^{-y^2/2} \,.$$

The integral of the function above, call it I, is easier if we instead compute I^2.

$$I^2 = \int_{-\infty}^{\infty} dI \int_{-\infty}^{\infty} dI' = \int_{-\infty}^{\infty} dy\, e^{-y^2/2} \int_{-\infty}^{\infty} dy'\, e^{-y'^2/2} = \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dy'\, e^{-(y^2+y'^2)/2} \,.$$

Transforming to polar coordinates, $y' = r\cos\theta$, $y = r\sin\theta$, $dy\, dy' = r\, dr\, d\theta$, and changing the integration limits as required gives

$$I^2 = \int_0^{2\pi} d\theta \int_0^{\infty} dr\, r e^{-r^2/2} = -2\pi e^{-r^2/2}\Big|_0^{\infty} = 2\pi \,.$$

Since $I = \sqrt{2\pi}$, we scale the exponential function by $1/\sqrt{2\pi}$ so ensure unit area. You really need to know that

$$\int_{-\infty}^{\infty} dy\, e^{-ay^2} = \sqrt{\pi/a}$$

and be able to perform a change of variable.

The *mean* of the normal r.v. is given by its first moment,

$$\begin{aligned}
\text{mean}(X) &\triangleq \mathcal{E}X = \int_{-\infty}^{\infty} dx\, x\, p(x) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, x\, e^{-(x-\mu)^2/2\sigma^2} \\
&= \frac{1}{\sigma\sqrt{2\pi}} \left[\int_{-\infty}^{\infty} dx\, (x-\mu)\, e^{-(x-\mu)^2/2\sigma^2} + \mu \int_{-\infty}^{\infty} dx\, e^{-(x-\mu)^2/2\sigma^2} \right] \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} dy\, y\, e^{-y^2/2\sigma^2} + \mu \int_{-\infty}^{\infty} dx\, p(x) \\
&= \mu\ .
\end{aligned}$$

The first integral is zero because the product of the exponential and y is antisymmetric. The second integral is one given the axioms of probability.

The *variance* of a normal r.v. is given by its second central moment. Applying the change of variable $y = (x-\mu)/\sigma$, we have

$$\begin{aligned}
\text{var}(X) &\triangleq \mathcal{E}(X-\mu)^2 = \int_{-\infty}^{\infty} dx\, (x-\mu)^2\, p(x) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dy\, y^2\, e^{-y^2/2}
\end{aligned}$$

Applying integration by parts[3] where $u = y$, $du = dy$, $dv = dy\, y\, \exp(-y^2/2)$, and $v = -\exp(-y^2/2)$,

$$\begin{aligned}
\text{var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \left[-ye^{-y^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} dy\, e^{-y^2/2} \right] \\
&= \sigma^2 \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dy\, e^{-y^2/2} \right] \\
&= \sigma^2\ .
\end{aligned}$$

[3] $uv = \int d(uv) = \int u\, dv + \int v\, du$. If u and v are both functions of t, then $u(t)v(t) = \int u(t)\, \dot{v}(t)\, dt + \int v(t)\, \dot{u}(t)\, dt$. If we switch to definite integration, we obtain the expression for integration by parts, $\int_a^b u(t)\, \dot{v}(t)\, dt = u(t)v(t)|_a^b - \int_a^b v(t)\, \dot{u}(t)\, dt$.

The normal pdf is a two-parameter distribution, where the two parameters just so happen to equal the mean and variance computed from the first two moments. That is not always the case. We will use $\mathrm{var}(X)$ wherever possible, instead of σ^2, to refer to variance. Simply stating $p(x) = \mathcal{N}(\mu, \sigma^2)$ provide all the information necessary to specify this distribution.

Variance is the *central* second moment of a distribution $\mathcal{E}(X-\mu)^2$. Its relationship to the *non-central* second moment $\mathcal{E}X^2$ is

$$\begin{aligned}
\mathrm{var}(X) &= \mathcal{E}(X-\mu)^2 = \mathcal{E}\{X^2 - 2\mu X + \mu^2\} \\
&= \int_{-\infty}^{\infty} dx\,(x^2 - 2\mu x + \mu^2)p(x) \\
&= \int_{-\infty}^{\infty} dx\,x^2\,p(x) - 2\mu \int_{-\infty}^{\infty} dx\,x\,p(x) + \mu^2 \int_{-\infty}^{\infty} dx\,p(x) \\
&= \mathcal{E}X^2 - 2\mu\mu + \mu^2 \\
&= \mathcal{E}X^2 - \mu^2\ .
\end{aligned}$$

$\mathcal{E}X^2$ is the mean-squared value of X; it equals variance only for zero-mean r.v.'s.

In MATLAB , `z=randn(100,1);` gives a 100×1 column vector of *standard normal* r.v.'s. The standard normal pdf has parameters $\mu = 0$ and $\sigma^2 = 1$. It is related to an arbitrary normal pdf through the relations,

$$\begin{aligned}
\text{standard normal r.v. } Z\text{:}\quad P(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} dy\,e^{-y^2/2}\ , \qquad \text{where } p(z) = \mathcal{N}(0,1) \\
\text{non-standard normal r.v. } X\text{:}\quad P(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} dy\,e^{-(y-\mu)^2/2\sigma^2}\ , \quad \text{where } p(x) = \mathcal{N}(\mu,\sigma^2)\ .
\end{aligned}$$

The conversion between variables is $x = \sigma z + \mu$. Therefore in MATLAB , `x=s*randn(M,N)+m` generates a 2-D array of normal random values having mean m and variance s^2 in this example.

The standard normal distributions of Fig B.7 were generated using

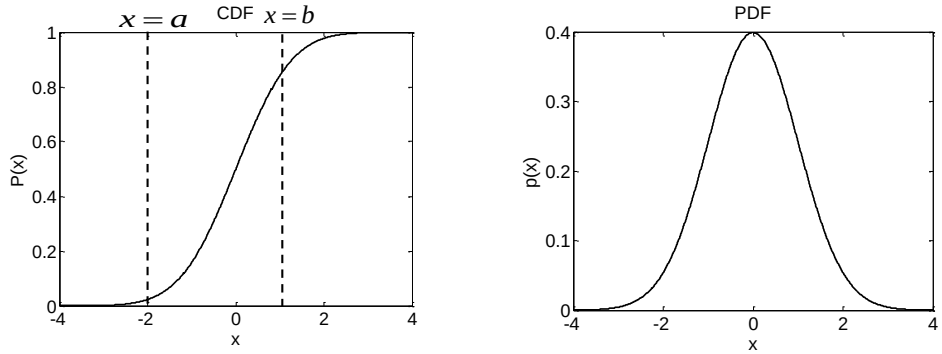
```
x=-4:0.01:4;y=normcdf(x,0,1);plot(x,y);z=normpdf(x,0,1);figure;plot(x,z)
```

You can also use the more general `cdf` and `pdf` functions in MATLAB and then just fill in your choice of distribution.

The cdf for a standard normal distribution is sometimes given a special symbol $\Phi(z)$,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} dy\,e^{-y^2/2}\ .$$

$$P(b) - P(a) = \Pr(a \le X \le b) = \text{the probability that r.v. } X \text{ lies between } a \text{ and } b$$



$$P(x) = \int_{-\infty}^{x} dx'\, p(x') \qquad\qquad p(x) = \left.\frac{dP(x')}{dx'}\right|_{x=x'}$$

Figure B.7: The standard normal cdf $P(x)$ and pdf $p(x)$.

Note that z has the units of y and σ. We know without calculating that $\Phi(0) = 0.5$. We can also find the area between $\pm n\sigma$ for $n = 1, 2, 3...$ (see Fig B.8) using `phi=normcdf(n,0,1)-normcdf(-n,0,1)` to find

$$\begin{aligned}
\Pr(|X| \le 1\sigma) &= \Phi(1) - \Phi(-1) = 0.683 \\
\Pr(|X| \le 2\sigma) &= \Phi(2) - \Phi(-2) = 0.954 \\
\Pr(|X| \le 3\sigma) &= \Phi(3) - \Phi(-3) = 0.997\ .
\end{aligned}$$

This tells us something about confidence intervals for normally-distributed measurements. Each time you make a measurement, there is a 68% chance that the next measurement will fall within $\pm\sigma$ and a 95% chance it will fall with $\pm 2\sigma$, etc.

In MATLAB $\Phi(x)$ may be found from error functions,

$$\mathrm{erf}(x) = \frac{2}{\pi} \int_0^x dy\, e^{-y^2}\ ,$$

however, you will need to apply a change of variables. The expression is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} dy\, e^{-y^2/2} = \frac{1}{2}\left[1 + \mathrm{erf}(z/\sqrt{2})\right]\ , \quad z \in \mathbb{R}\ .$$

which you can obtain from the MATLAB function `Phi=normcfd(x,m,s)`. We will leave the normal distribution for awhile so we can introduce other general concepts needed before discussing multivariate distributions.

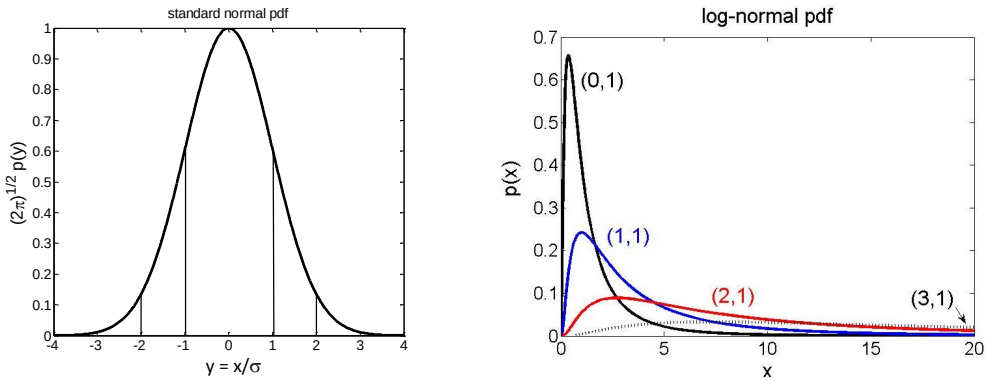


Figure B.8: (left) Standard normal pdf. Vertical lines show $\pm n\sigma$ for $n = 1, 2, 3$. (right) Log-normal pdf for parameters (μ, σ).

B.10.2 Log-normal pdf

Asymmetric and long-tailed, the *log-normal* distribution curve describes phenomena like the spread of blood pressures in the adult population and epidemic curves for some infectious diseases. These can be explored in MATLAB using the general function `Y = pdf(NAME,X,A,B)`, where `NAME='logn'`, and, e.g., variable range `X = 0:0.01:20` and parameters `A=0` (μ) and `B=1` (σ). The log-normal pdf is

$$p_Y(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) , \qquad x > 0 . \tag{B.11}$$

Examples are plotted in Fig B.8 for four different values of μ and $\sigma = 1$. Since $\ln x$ must be unitless, so must parameters μ and σ. The mean and variance of the log-normal distribution are

$$\begin{aligned} \mathcal{E}Y &= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ \mathrm{var}(Y) &= \exp\left(2(\mu + \sigma^2)\right) - \exp(2\mu + \sigma^2) . \end{aligned}$$

Here is an example where the parameters compose but are not equal to the mean and variance of the distribution.

B.11 What is a statistic?

A *statistic* is a number (sometimes a matrix of numbers) that characterizes properties of a random process. Statistics are summary metric on which reliable decisions can be

based. For example, core body temperature is a test statistic that varies depending on several factors. The mean normal temperature in the population is 98.6^oF, the variance is $(0.9^o\text{F})^2$, and when body temperature exceeds 102^oF it is time to seek medical assistance.

First-order statistics, which are found from moments of a distribution, describe properties of individual random variables. Examples in common use are mean, variance, skewness, and kurtosis found from the first four statistical moments, respectively. *Second-order statistics*, also found from moments, describe properties of pairs of random variables. Common examples include covariance matrices and power spectral densities as described in Chapter 4.

The full probability distribution will completely characterize a random process. For a distribution to apply, we must find good reasons for believing a phenomenon is represented by a particular distribution, as we did in §B.8.2. There is a significant body of experimental data to suggest various physical processes are well represented by Poisson or normal variables. Since those random processes are one (λ for Poisson) and two (μ and σ for normal) parameter distributions, the corresponding physical processes may be completely specified by just one or two parameters that we can estimate from the first few statistical moments. Other distributions are characterized by more than two parameters and may require us to measure more moments to estimate them. Nevertheless, a reasonable summary of a random process can often be obtained from the first two moments of first- and second-order statistics, which we now describe.

B.11.1 First-order moments

Definition B.11.1. *The m-th* non-central moment *of a distribution for r.v. X that generates a first-order statistic is the expected value of the m-th power of X:*

$$\mathcal{E}X^m = \int_{-\infty}^{\infty} dx\, x^m\, p(x)\ . \tag{B.12}$$

The distribution mean is $\mathcal{E}X$, the first moment at $m = 1$. The m-th central moment *is*

$$\mathcal{E}\{(X - \mathcal{E}X)^m\} = \int_{-\infty}^{\infty} dx\, (x - \mathcal{E}X)^m\, p(x)\ . \tag{B.13}$$

For example, the variance is the second central moment, $\text{var}_X = \mathcal{E}(X - (\mathcal{E}X))^2$. *Central moments summarize distribution properties centered about the mean value.*

Sample ensemble statistics

In practice, we are always presented with data samples acquired from experiments. Here we get a little more specific than Eqs (B.12) and (B.13).

From the illustration of data in Fig B.5, we see there are N sample waveforms at each fixed point in time t that yield discrete data values, $x_n(t)$. Therefore, Eq (B.12) yields a first-moment expression in terms of a sum,

$$\mathcal{E}X(t) = \sum_{-\infty}^{\infty} x_n(t)\, p(x_n(t)) \ , \qquad \text{POPULATION ENSEMBLE MEAN} \tag{B.14}$$

and is explicitly a function of the independent variables, in this case time. Also $p(x_n(t)) = dx\, p(x(t))$. In practice, there will be $N < \infty$ waveforms, so the sum is reduced to 0 to $N-1$. If there is no reason to expect a value observed at t in one waveform is any more likely than another, it is reasonable to assume the pdf is uniform, $p(x_n(t)) = 1/N$ for $0 \le n \le N-1$. Eq (B.14) gives the familiar approximate time-varying first moment,

$$\bar{x}(t) = \frac{1}{N} \sum_{n=0}^{N-1} x_n(t) \ . \qquad \text{SAMPLE ENSEMBLE MEAN}$$

This expression converges to the population mean for a uniform distribution in Eq (B.14) as $N \to \infty$. If samples are not equally probable and the distribution is known, we replace the uniform pdf with one more appropriate.

Applying the same reasoning, *sample ensemble variance* is $s^2_{e,X} = \sum_n (x_n(t) - \mathcal{E}X(t))^2/N$. When the sample mean $\bar{x}(t)$ estimated from the same data is used in place of the population mean $\mathcal{E}X(t)$, we reduce N by one,

$$\widehat{\text{var}}_X(t) = \frac{1}{N-1} \sum_{n=0}^{N-1} \left(x_n(t) - \bar{x}(t)\right)^2 \ , \qquad \text{SAMPLE ENSEMBLE VARIANCE}$$

and is also a function of time.

Sample temporal statistics

Say we do not have a waveform ensemble, only a single waveform $x(t)$. We are stuck computing time averages and hoping they approximate the corresponding ensemble averages.

We first estimate the pdf of the sampled function, $p_s(x(t))$ using the theorems of §2.7.

$$p_s(x(t)) = \boldsymbol{\mathcal{S}}^\dagger \boldsymbol{\mathcal{S}} p(x(t)) = \sum_{n=-\infty}^{\infty} \Pr(X = x(t))\delta(x(t) - x(n'T)) \quad \text{for } \; 0 \le n' \le N' - 1 \; .$$

Note that index n indicates the waveform at a fixed time and index n' indicates the time sample along a given waveform; i.e., $x_n[n']$, but in this example we have one waveform so we eliminate the n index. From Eq (B.12), time-average moments for the sampled waveform are

$$\begin{aligned}
\langle X^m \rangle &= \int_{-\infty}^{\infty} dx(t)\, x^m(t)\, p_s(x(t)) = \int_{-\infty}^{\infty} dx(t)\, x^m(t) \sum_{n'=-\infty}^{\infty} \Pr(X = x(t))\delta(x(t) - x(n'T)) \\
&= \sum_{n'=-\infty}^{\infty} \int_{-\infty}^{\infty} dx(t)\, x^m(t) \Pr(X = x(t))\delta(x(t) - x(n'T)) \\
&\simeq \sum_{n'=0}^{N'-1} x^m[n'] \Pr(X = x[n']) = \sum_{n'=0}^{N'-1} x^m[n']\, p(x[n']) \; . \qquad \text{(B.15)}
\end{aligned}$$

$x[n']$ is shorthand for $x(n'T)$. The last line is an approximation because $N' < \infty$ values of $p(x[n'])$ are included in the sum.

Assuming uniform sampling, $p(x[n']) = 1/N'$, the sample mean found using Eq (B.15) is

$$\langle x_n \rangle = \frac{1}{N'} \sum_{n'=0}^{N'-1} x_n[n'] \; , \qquad \text{SAMPLE TEMPORAL-AVERAGE MEAN}$$

which is constant over time. The subscript n reminds us that this mean applies to the data in the nth time series. Compare this expression with that for the sample ensemble mean above and look closely at the differences. The *sample temporal-average variance* is

$$s_{x_n}^2 = \frac{1}{N'-1} \sum_{n=0}^{N'-1} \big(x_n[n'] - \langle x \rangle\big)^2 \; , \qquad \text{SAMPLE TEMPORAL-AVERAGE VARIANCE}$$

which also only applies to the nth waveform. In §B.14 below, we will discuss situations when time-averaged moments $\langle x_n^m \rangle$ and ensemble-averaged moments $\mathcal{E}X^m(t)$ are expected to yield the same results.

B.11.2 Vector forms

We may write the continuous-time waveforms from the ensemble of X at a fixed point in time as a $1 \times N$ row vector, $\mathbf{x}(t) = (x_0(t), \ldots, x_n(t) \ldots, x_{N-1}(t))$, where each vector

element is a whole waveform drawn from the ensemble. If each waveform is then sampled in time to form a time series where $t = n'T$, $\mathbf{x}(t)$ becomes a data matrix with elements having two indices. Each row of the matrix describes the ensemble at one time, $\mathbf{x}(t = n'T) = (x_0[n'], \ldots, x_n[n'] \ldots, x_{N-1}[n'])$ that can also be written as $(x_{n'0}, \ldots, x_{n'n}, \ldots, x_{n'N-1})$. The whole $N' \times N$ matrix is

$$\mathbf{x}(t) = \begin{pmatrix} x_{00} & x_{01} & \cdots & x_{0n} & \cdots & x_{0N-1} \\ x_{10} & x_{11} & & & & \\ \vdots & & \ddots & & & \vdots \\ x_{n'0} & x_{n'1} & \cdots & x_{n'n} & \cdots & x_{n'N-1} \\ \vdots & & & & \ddots & \\ x_{N'-1,0} & & \cdots & & & x_{N'-1,N-1} \end{pmatrix}, \tag{B.16}$$

where columns are time series from the ensemble. Summing over rows and dividing by N, we find the sample ensemble mean vector $\bar{\mathbf{x}}(t) = (\bar{x}[0], \ldots, \bar{x}[n'] \ldots, \bar{x}[N'-1])$. Summing over columns and dividing by N', we find the sample time-averaged mean vector $\langle \mathbf{x}_n \rangle = \langle x_0 \rangle, \ldots, \langle x_n \rangle, \ldots, \langle x_{N-1} \rangle)$, which varies for each of the N time series.

First-order moments consider the statistical properties of individual samples. Second-order moments consider statistical properties of samples two at a time via products, which reveals correlations that exist among samples.

B.11.3 Second-order moments

The components of *covariance matrix* $\mathbf{K}$ are second-order statistics because they are the ensemble averages of two-sample products. To find $\mathbf{K}$, we begin with the time series given by the nth column of Eq (B.16), $\mathbf{x}_n = (x_{0n}, \ldots, x_{n'n}, \ldots x_{N'-1n})^t$, where the first index indicates time samples via $n' = t/T$ for $0 \leq n' \leq N'-1$. Subtracting the sample ensemble mean from $\mathbf{x}_n$ gives
$\mathbf{y}_n = (\mathbf{y}_{0n}, \ldots, \mathbf{y}_{n'n}, \ldots, \mathbf{y}_{N'-1n})^t = \mathbf{x}_n - \bar{\mathbf{x}} = (x_{0n} - \bar{x}_0, \ldots, x_{n'n} - \bar{x}_{n'}, \ldots x_{N'-1n} - \bar{x}_{N'-1})^t$.
The sample covariance matrix is the expected value of the outer product of $\mathbf{y}$ with itself,

$$\mathbf{K}_Y = \mathcal{E}\{\mathbf{y}\mathbf{y}^t\} = \begin{pmatrix} \mathcal{E}y_0^2 & \cdots & \mathcal{E}y_0 y_{n'} & \cdots & \mathcal{E}y_0 y_{N'-1} \\ \vdots & \ddots & & & \\ \mathcal{E}y_{n'} y_0 & & \mathcal{E}y_{n'}^2 & & \vdots \\ \vdots & & & \ddots & \\ \mathcal{E}y_{N'-1} y_0 & & \cdots & & \mathcal{E}y_{N'-1}^2 \end{pmatrix}. \tag{B.17}$$

The expected value is over all n in the ensemble, so only the temporal n' index labels the matrix elements. Diagonal elements of the covariance matrix are the variances at each

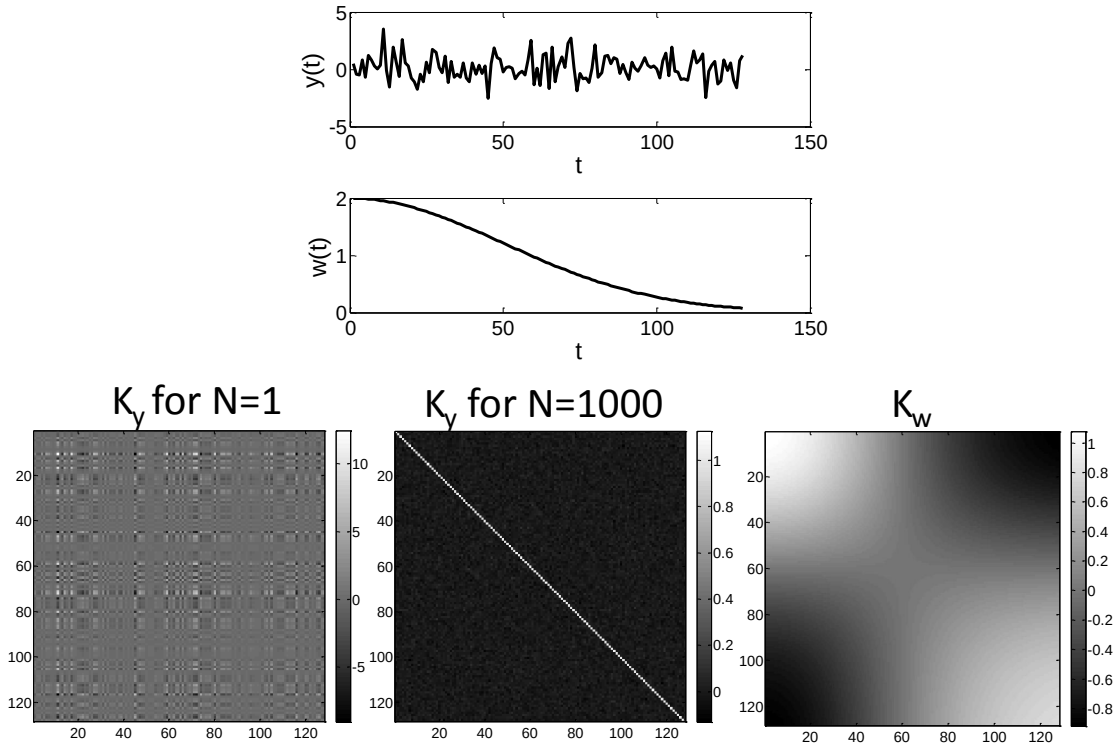


Figure B.9: (top plots) One $N' = 128$-pt zero-mean random sequence $y(t)$ from $\mathcal{N}(0,1)$ and a $N' = 128$-pt deterministic Gaussian function $w(t)$. (bottom row from left to right) The $N' \times N'$ covariance matrix estimate of y for an ensemble of $N = 1$, for an ensemble of $N = 1000$, and the covariance matrix for the deterministic $w(t)$.

time, $\mathrm{var}(x_{n'}) = \mathcal{E}y_{n'}^2$, and the off-diagonal elements are the covariances, e.g., $\mathrm{cov}(n', 0) = \mathcal{E}y_{n'}y_0$. The covariance matrix generalizes the notion of variance for every-possible pair of samples. The covariance matrix is square, symmetric/Hermitian ($\mathbf{K} = \mathbf{K}^\dagger$), and positive semi-definite, i.e., $\mathbf{y}^\dagger\mathbf{K}\mathbf{y} \geq 0$. Therefore covariance matrices are diagonalizable, invertible, and have positive real eigenvalues. (See Appendix A).

Fig B.9 provides numerical examples of Eq (B.17) in image form. At the top of the figure, I show random $y(t)$ and deterministic $w(t)$ time series. The matrix on the lower left is what is found for an uncorrelated random time series for only one realization (no ensemble averaging, $N' = 128, N = 1$). This provides a very poor estimate of $\mathbf{K}_y$. The center matrix is a better estimate as the ensemble averages is over $N = 1000$ time series, each $N' = 128$ points long. (Notice the ensemble average of matrix elements is required, not that of the time series!) The result closely approximates the true covariance matrix, which in this example is diagonal $\mathbf{K}_y = \mathbf{I}$ because data generation ensures that all variances on the main diagonal $\mathcal{E}\mathbf{y}_{n'}^2 = 1$. A diagonal covariance matrix tells us that elements of the time series are uncorrelated. As we will see in the next section, equal diagonal elements indicate a stationary random process.

There is a simple way to generate covariance matrices, including ensemble averaging, using matrix multiplications. Construct a data matrix as shown in Eq (B.16) where there are N columns in matrix $\mathbf{x}$ that are each a recorded time series of length N'. Matrix $\mathbf{x}$ has size $N' \times N$. Subtracting the mean to find matrix $\mathbf{y}$, then $\mathbf{K}_y = \mathbf{y}\mathbf{y}^t$. The 128×128 examples of $\mathbf{K}_y$ in Fig B.9 were generated using

```
Np=128;N=1000;x=randn(Np,N);K=x*x'/N;
imagesc(K);colormap(gray);axis square; colorbar
```

for a standard-normal random-number generator `randn` that produces zero-mean samples. Varying the number of columns via N changes the size of the ensemble.

In Fig B.9, lower right, we see the covariance matrix of the deterministic Gaussian function, $\mathbf{K}_w$, which is highly correlated in time and requires no ensemble averaging. Bright regions in $\mathbf{K}_w$ are positively correlated while dark regions are negatively correlated.

B.12 Stationary random processes

Definition B.12.1. *Assume random process $X(t)$. Let $p_X(x[0], \ldots, x[n'], \ldots, x[N'-1])$ be the joint pdf for samples at times $t = n'T$, $0 \leq n' \leq N'-1$. Then $X(t)$ is a stationary*

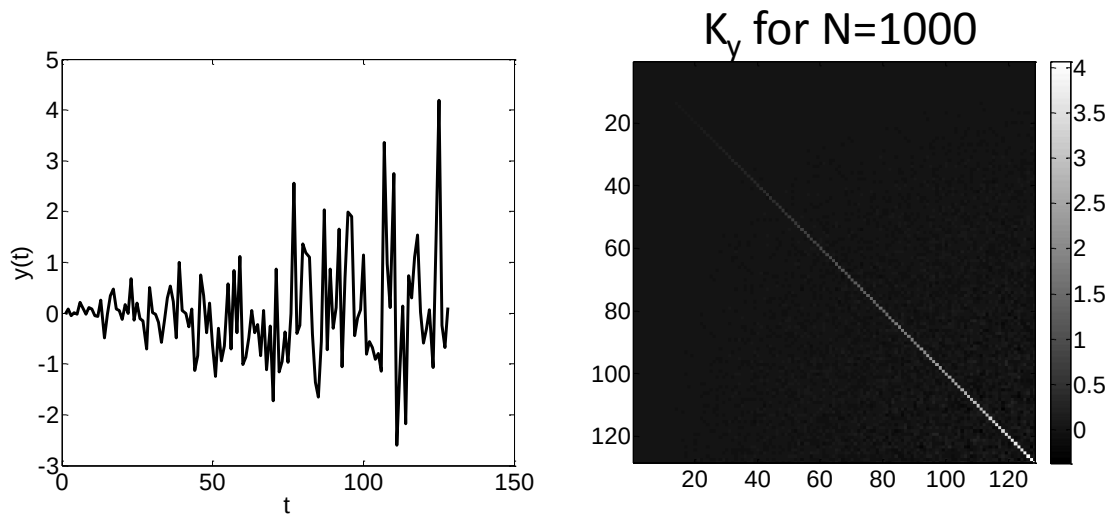


Figure B.10: A plot of a nonstationary, uncorrelated random time series and its covariance matrix for an ensemble $N = 1000$.

random process if

$$p_X(x[0], \ldots, x[n'], \ldots, x[N'-1]) = p_X(x[0+\ell], \ldots, x[n'+\ell], \ldots, x[N'-1+\ell]) \ .$$

That is, if we shift the time for each sample in the pdf by a constant amount ℓ and find the distribution does not change, the process is strictly stationary*. The equation must hold for any value of integer ℓ provided $x[n'+\ell] \in X(t)$.*

The function $y(t)$ shown in Fig B.9 is stationary since each value in the time series was drawn from the same random number generator without reference to time. Consequently, any shift in the time axis produces the same distribution. This implies that all moments of strictly stationary random processes are time invariant.

It is difficult to measure stationarity of a process other than measuring a few of the moments to see if they change as the time axis is shifted. If we know the *first two moments* of a process are time invariant, i.e., the mean $\mathcal{E}X(t) = \mathcal{E}X$ and covariance matrix $\mathbf{K}_X(t) = \mathbf{K}_X$ are not functions of t, the process is said to be *wide-sense stationary*, WSS. Wide-sense stationary random processes have covariance matrices with a Toeplitz structure; e.g., $\mathbf{K}_Y$ for $N = 1000$ in Fig B.9.

The covariance matrix for an uncorrelated normal random process that is not WSS is shown in Fig B.10. Here the parameter σ increases with time. That increase is clearly

seen in the time series. This covariance matrix is not Toeplitz, although it is still square, Hermitian, and positive semidefinite. The code used to generate this nonstationary covariance matrix is

```
K=zeros(128);r=2/128:2/128:2;
for j=1:1000;xp=randn(128,1);x=r'.*xp;K=K+x*x';end
imagesc(K/1000);colormap gray;axis square;colorbar
```

B.13 Covariance and correlation for stationary processes

Consider the random process $X(t)$ as a continuous function of time. Its autocovariance function, defined over the time range $0 \le t, \tau \le T_t$, is

$$\begin{aligned} K_X(t, t-\tau) &= \mathcal{E}\left\{(x(t) - \bar{x}(t))(x(t-\tau) - \bar{x}(t-\tau))\right\} \qquad \text{(B.18)} \\ &= \int_{x\in X} dx(t) \int_{x\in X} dx(t-\tau)\Big(x(t) - \bar{x}(t)\Big)\Big(x(t-\tau) - \bar{x}(t-\tau)\Big)\, p_X(x(t), x(t-\tau)) \ , \end{aligned}$$

where we measure the ensemble statistics of X two points at a time, at $x(t)$ and $x(t-\tau)$. For a stationary process, $\bar{x}(t) = \bar{x}$ and $p_X(x(t), x(t-\tau)$ is independent of t. Applying the linearity property of the ensemble operator to Eq (B.18), we find

$$\begin{aligned} K_X(\tau) &= \mathcal{E}\{x(t)\, x(t-\tau)\} - \mathcal{E}\{x(t)\}\bar{x} - \bar{x}\mathcal{E}\{x(t-\tau)\} + \bar{x}^2 \;=\; \mathcal{E}\{x(t)x(t-\tau)\} - \bar{x}^2 \\ &= \left[\int_{x\in X} dx(t) \int_{x\in X} dx(t-\tau)\left(x(t)\, x(t-\tau)\right) p_X(x(t), x(t-\tau))\right] - \bar{x}^2 \\ &= R_X(\tau) - \bar{x}^2 \ . \end{aligned}$$

$R_X(\tau)$ is the autocorrelation function for stationary process X. For stationary processes X and Y, the auto- and cross-covariance functions are

$$\begin{aligned} K_X(\tau) &= R_X(\tau) - \bar{x}^2 \\ K_Y(\tau) &= R_Y(\tau) - \bar{y}^2 \\ K_{XY}(\tau) &= R_{XY}(\tau) - \bar{x}\bar{y} \end{aligned}$$

where the cross-correlation function is

$$R_{XY}(\tau) = \int_{x\in X} dx(t) \int_{y\in Y} dy(t-\tau)\left(x(t)\, y(t-\tau)\right) p_{XY}(x(t), y(t-\tau))$$

B.14 Ergodic random process

Consider the WSS random processes $X(t)$ and $Y(t)$.

Definition B.14.1. *An ergodic process is a stationary process in which ensemble averages equal time averages. For example, $X(t)$ will be a first-order ergodic process if its ensemble-average mean $\bar{x}$ and its time-average mean $\langle x_n \rangle$ are equal. $X(t)$ is a second-order ergodic process if the ensemble correlation function $R_X(\tau)$ may be represented by a time average correlation $\phi_X(\tau)$ from Eq (1.16). The advantages of being able to show a process is ergodic is that you may substitute averages over the independent variables, usually time and/or space, in place of ensemble averages that may be more difficult to obtain experimentally.*

Specifically, consider $x_n(n'T)$ as a single realization of the ensemble $X(t)$. The subscript n denotes the waveform in the ensemble and n' indicates the time sample along any waveform. $X(t)$ is first-order ergodic if

$$\begin{aligned} \bar{x} &= \langle x_n \rangle \quad \text{or} \\ \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_n(n'T) &= \lim_{N'\to\infty} \frac{1}{N'} \sum_{n'=0}^{N'-1} x_n(n'T) \ . \end{aligned}$$

$X(t)$ and $Y(t)$ are second-order ergodic processes if, e.g.,

$$\begin{aligned} R_X(\tau) &= \phi_X(\tau, n) \quad \text{or} \\ \int_{x\in X} dx(t) \int_{x\in X} dx(t-\tau) \big(x(t)\, x(t-\tau)\big)\, p_X(x(t), x(t-\tau)) &= \lim_{T_t\to\infty} \frac{1}{T_t} \int_0^{T_t} dt\, x_n(t)\, x_n(t-\tau) \\ \text{and the same holds for} \quad R_{XY}(\tau) &= \phi_{XY}(\tau, n) \ . \end{aligned}$$

All ergodic processes are stationary, but not all stationary processes are ergodic. Also it is possible for a process to be first-order ergodic and second-order nonergodic as shown in the example below.

Example B.14.1. *Consider random process $G(t)$, where the nth waveform realization is*

$$g_n(t) = A_n \sin(2\pi t + \theta_n) \ .$$

The amplitude A_n and phase θ_n are both uniformly-distributed random variables that vary with the nth waveform. Examples of four such waveforms are shown in Fig B.11a.

Ideally, $G(t)$ is first-order ergodic process because the time-averaged and ensemble-averaged means are both zero in the limits, which for time averaging is found for

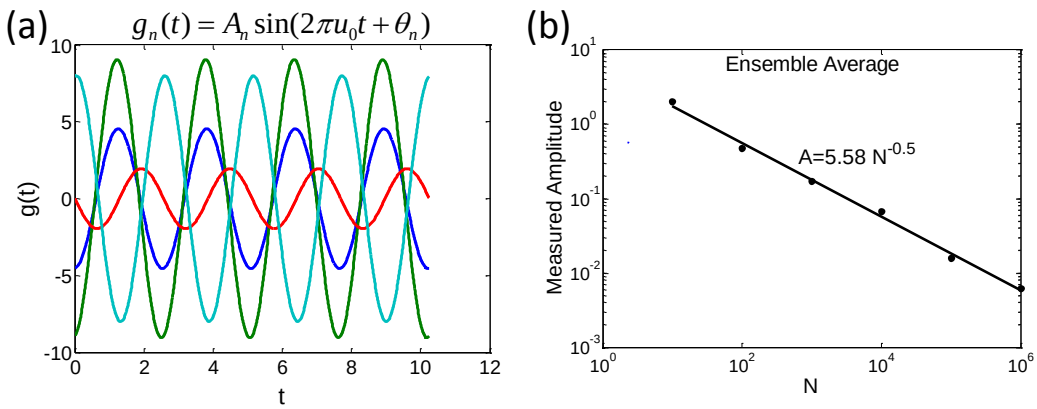


Figure B.11: (a) Four sine waves $g_n(t)$ with random amplitude and phase. (b) Plot of the ensemble mean maximum amplitude as a function of ensemble size N. $\bar{g}(t) \to 0$ as $N \to \infty$ in a manner $\propto 1/\sqrt{N}$.

any integer number of sine-wave cycles. However, there are practical considerations when considering this statement using experimental data. While the time-average mean is zero for every waveform, the ensemble-averaged mean converges to zero but slowly as N increases at a rate proportional to $1/\sqrt{N}$. This is shown in Fig B.11b. It does not mean $G(t)$ is first-order nonergodic if only 100 waveforms are available. It does mean that moments computed using 100 or fewer waveforms are much better obtained using time averaging for this first-order ergodic process. It is difficult to make strong statements about ergodicity from measurements where few data samples are available. Ergodicity arguments are usually made analytically and not numerically.

We can analytically show that $G(t)$ is not second-order ergodic. We already computed a very similar problem in Example 1.7.1 but there the amplitude and phase were constant. Adapting that result for random amplitude and phase, we find

$$\phi_G(\tau, n) = \frac{A_n^2}{2} \cos(2\pi u_0 \tau) \ ,$$

which is a function of the exact waveform used via the index n. In comparison, $R_G(\tau)$ is not a function of the waveform, so $G(t)$ is not second-order ergodic. Hence, if you need to compute the covariance matrix, for example, you need to obtain many realizations of waveforms and cannot use time-averaged estimates.

If the waveform amplitude is constant, even if the phase remains random, then $\phi_G(\tau, n) = \frac{A^2}{2} \cos(2\pi u_0 \tau) = R_G(\tau)$ and so $G(t)$ is both first- and second-order ergodic.

B.15 Jointly-distributed random variables

In this section, we discuss probabilities of two or more random variables. The joint cumulative distribution function (cdf) of variables X and Y is

$$\Pr(X \le x, Y \le y) = P_{XY}(x,y) = \int_{-\infty}^{y} dy' \int_{-\infty}^{x} dx'\, p_{XY}(x',y') \qquad \text{for} \quad -\infty < x, y < \infty \; .$$

As in the univariate case,

$$\Pr(x < X < x{+}dx, y < Y < y{+}dy) = \int_{y}^{y+dy} dy' \int_{x}^{x+dx} dx'\, p(x',y') \simeq dx\, dy\, p(x,y) = \partial^2 P(x,y) \; ,$$

which established a relationship among probability, cdf, and pdf.

The *marginal cdf*, $P_X(x)$, is found from the joint bivariate distribution using

$$\begin{aligned} P_{XY}(x, y=\infty) &= \Pr(-\infty < X \le x, -\infty < Y \le \infty) \\ &= \int_{-\infty}^{x} dx' \int_{-\infty}^{\infty} dy' p_{XY}(x',y') = \int_{-\infty}^{x} dx'\, p_X(x') = P_X(x) \; . \end{aligned}$$

Notice the integration limits in the equation above. We can apply a similar procedure to find $P_Y(y)$. The *marginal probability density functions* are

$$\begin{aligned} p(x) = \frac{dP(x)}{dx} &= \int_{-\infty}^{\infty} dy' p(x,y') \\ p(y) = \frac{dP(y)}{dy} &= \int_{-\infty}^{\infty} dx' p(x',y) \end{aligned}$$

There is no reason to stop at two variables. Multivariate r.v.s can be compactly expressed as $N \times 1$ column vectors $\mathbf{X}$. The corresponding cumulative distribution function is $P(\mathbf{x}) = \Pr(X_1 \le x_1, \ldots, X_N \le x_N)$ and the associated density is $p(\mathbf{x}) = \partial^N P(\mathbf{x}) / \partial x_1 \ldots \partial x_N$. Also

$$\begin{aligned} p(\mathbf{x}) &= \int_{-\infty}^{\infty} d\mathbf{y}\, p(\mathbf{x},\mathbf{y}) \\ p(\mathbf{x}|\mathbf{y}) &= p(\mathbf{x},\mathbf{y})/p(\mathbf{y}) \; . \end{aligned}$$

If random variable vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$. Of course, the elements of each vector may still be dependent. Furthermore, when the vector elements are independent, $p(\mathbf{x}) = p(x_1)p(x_2)\ldots p(x_N)$.

B.16 Multivariate normal density

Returning to our discussion of normal random processes in §B.10.1, we can expand Eq (B.10) to an N-dimensional multivariate normal (MVN) probability density for the real vector $\mathbf{x} = (x_0, \ldots, x_{n'}, \ldots, x_{N'-1})^t$. That is, we are treating elements of the vector as normal random variables that may be coupled as explained by the covariance matrix, $\mathbf{K}$. Because this is a normal r.v., parameters $\mu_{n'}$ and $\sigma^2_{n'}$ are the mean and variance of the n'th element.

The pdf for a MVN process may be written as

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}) = [(2\pi)^N \det \mathbf{K}]^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] , \tag{B.19}$$

where the population mean may be written as $\boldsymbol{\mu}_X = \mathcal{E}\mathbf{X} = (\mu_0 \ldots \mu'_n \ldots \mu_{N'-1})^t$ and the covariance matrix $\mathbf{K} \triangleq \mathbf{K}_X$ is

$$\begin{aligned}
\mathbf{K} &= \mathcal{E}\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\} \qquad (B.20)\\
&= \mathcal{E}\left\{\begin{pmatrix} x_0-\mu_0 \\ x_n-\mu_1 \\ \vdots \\ x_{N'-1}-\mu_{N'-1} \end{pmatrix} \begin{pmatrix} (x_0-\mu_0) & (x_1-\mu_1) & \ldots & (x_{N'-1}-\mu_{N'-1}) \end{pmatrix}\right\} \\
&= \begin{pmatrix} \sigma^2_{00} & \sigma^2_{01} & \cdots & \sigma^2_{0,N'-1} \\ \sigma^2_{10} & \sigma^2_{11} & \cdots & \sigma^2_{1,N'-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{N'-1,0} & \sigma^2_{N'-1,1} & \cdots & \sigma^2_{N'-1,N'-1} \end{pmatrix}
\end{aligned}$$

where σ^2_{ij} are variances when $i = j$ and covariances when $i \neq j$. We can write the MVN pdf compactly using simply $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$.

If $\mathbf{K}$ is diagonal, $\{x_{n'}\}$ are uncorrelated. An example of a diagonal covariance matrix for a stationary process is given in Fig B.9 and for a nonstationary process in Fig B.10. For diagonal $\mathbf{K}$ and a stationary MVN process, the exponent is Eq (B.19) reduces to

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu}) =$$

$$\frac{1}{2}\left((x_0-\mu_0)\cdots(x_{N'-1}-\mu_{N'-1})\right) \begin{pmatrix} 1/\sigma^2_{00} & & & 0 \\ & 1/\sigma^2_{11} & & \\ & & \ddots & \\ 0 & & & 1/\sigma^2_{N'-1,N'-1} \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ \vdots \\ x_N-\mu_N \end{pmatrix}$$

so that Eq (B.19) becomes

$$p(\mathbf{x}) = \prod_{n'=0}^{N'-1} \frac{1}{\sigma_{n'}\sqrt{2\pi}} e^{-(x_{n'}-\mu_{n'})^2/2\sigma_{n'}^2}$$
$$= [(2\pi)^{N'/2}\sigma_0 \cdots \sigma_{N'-1}]^{-1} \exp\left[-\frac{1}{2}\sum_{n'=0}^{N'-1}(x_{n'}-\mu_{n'})^2/\sigma_{n'}^2\right] . \qquad \text{(B.21)}$$

Being able to diagonalize the covariance matrix greatly simplifies the MVN expressions. As we saw in Appendix A, square, symmetric, positive-semi-definite matrices, like $\mathbf{K}$, are diagonalizable, although approximate forms are used to manage very large matrices.

B.17 Maximum likelihood estimation

Random errors of real-valued measurements $\mathbf{g}$ are often modeled as random processes with known or assumed probability distributions. Random processes are fully characterized by distribution parameters or moments if their parent distributions are known. If we analyze the physics of the problem and find that errors are normally distributed, we may become interested in estimating the mean and covariance,

$$\mathcal{E}\{\mathbf{g}\} = \int_{-\infty}^{\infty} d\mathbf{g}\, \mathbf{g}\, p(\mathbf{g}) = \boldsymbol{\mu} \quad \text{and} \quad \mathcal{E}\{(\mathbf{g}-\boldsymbol{\mu})(\mathbf{g}-\boldsymbol{\mu})^t\} = \int_{-\infty}^{\infty} d\mathbf{g}\, (\mathbf{g}-\boldsymbol{\mu})(\mathbf{g}-\boldsymbol{\mu})^t\, p(\mathbf{g}) = \mathbf{K} .$$

Even if the parent distributions can be assumed, the associated parameters often cannot be assumed. They must be estimated from the *sample moments* given a limited about of measurement data.

The *maximum likelihood* (ML) approach provides a method for estimating distribution parameters. A likelihood function $\mathcal{L}(\theta|\mathbf{g}) = p(\mathbf{g}|\theta)$ looks like a probability but is not exactly because it is formed from a specific data set $\mathbf{g}$ that is drawn from a parent population. It asks what value of parameter θ associated with the assumed parent distribution is most likely responsible for that specific data. We will use this method to estimate the sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ from a set of N measurements $\mathbf{g}$ assumed drawn from an independent identically-distributed (i.i.d.) normal random process. The pdf for this distribution is given by Eq (B.21), where the means for each sample are equal to each other as are their variances.

Since the log-likelihood, $\ln \mathcal{L}(\theta|\mathbf{g})$, is monotonic with $\mathcal{L}(\theta|\mathbf{g})$ and easier to analyze, we will maximize the log-likelihood function by taking its derivative with respect to θ and setting

the result equal to zero. Using Eq (B.21) where the parameter of interest is $\theta = \mu$,

$$\begin{aligned}
\left.\frac{\partial \ln \mathcal{L}(\mu|\mathbf{g})}{\partial \mu}\right|_{\mu=\hat{\mu}} &= \frac{\partial \ln p(\mathbf{g}|\mu)}{\partial \mu} = -\frac{\partial}{\partial \mu}\left[\frac{N'}{2}\ln[(2\pi)\sigma^2] + \frac{1}{2\sigma^2}\sum_{n'=0}^{N'-1}(g_{n'}-\mu)^2\right]_{\mu=\hat{\mu}} = 0 \\
&= -\frac{1}{\sigma^2}\left[\sum_{n'=0}^{N'-1} g_{n'} - N'\hat{\mu}\right] = 0 \ ,
\end{aligned}$$

and therefore

$$\hat{\mu} \triangleq \bar{g} = \frac{1}{N'}\sum_{n'=0}^{N'-1} g_{n'} \ . \tag{B.22}$$

The ML estimate of mean given a specific set of data $\mathbf{g}$ is the sample ensemble mean, $\bar{g}$. As we found in §B.14, if we can show the process is also ergodic, then time averaged can be used in place of ensemble averages for these samples.

The ML estimate of variance is found in a similar way. Because $\mathbf{g}$ is an i.i.d normal r.v., $\widehat{\mathrm{var}}(\mathbf{g})(t) = \sigma^2$. Therefore,

$$\begin{aligned}
\left.\frac{\partial \ln \mathcal{L}(\sigma|\mathbf{g})}{\partial \sigma}\right|_{\sigma=\hat{\sigma}} &= \frac{\partial \ln p(\mathbf{g}|\sigma)}{\partial \sigma} = -\frac{\partial}{\partial \sigma}\left[\frac{N'}{2}\ln[(2\pi)\sigma^2] + \frac{1}{2\sigma^2}\sum_{n'=1}^{N'-1}(g_n-\mu)^2\right] = 0 \\
&= -\left[\frac{N'}{\sigma} - \frac{1}{\sigma^3}\sum_{n'=1}^{N'-1}(g_{n'}-\mu)^2\right]_{\sigma=\hat{\sigma},\ \mu=\hat{\mu}} = 0 \\
\hat{\sigma}^2 &= \frac{1}{N'}\sum_{n'=1}^{N'-1}(g_{n'}-\bar{g})^2 \ .
\end{aligned} \tag{B.23}$$

This expression should look familiar; it is the sample ensemble variance. Examining further, we complete the square from Eq (B.23) and substitute $\Delta_{n'} = \mu - g_{n'}$ to find

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{N'}\left[\sum_{n'=0}^{N'-1} g_{n'}^2 - 2\bar{g}\sum_{n'=0}^{N'-1} g_{n'} + \bar{g}^2\right] = \frac{1}{N'}\left[\sum_{n'=0}^{N'-1} g_{n'}^2 - \frac{1}{N'}\sum_{m'=0}^{N'-1} g_{m'}\sum_{n'=0}^{N'-1} g_{n'}\right] \\
&= \frac{1}{N'}\sum_{n'=0}^{N'-1}(\mu-\Delta_{n'})^2 - \frac{1}{N'^2}\sum_{n'=0}^{N'-1}\sum_{m'=0}^{N'-1}(\mu-\Delta_{n'})(\mu-\Delta_{m'}) \\
\mathcal{E}\{\hat{\sigma}^2\} &= \mathcal{E}\left\{\frac{1}{N'}\sum_{n'=0}^{N'-1}(\mu-\Delta_{n'})^2 - \frac{1}{N'^2}\sum_{n'=0}^{N'}(\mu-\Delta_{n'})^2\right\} = \sigma^2 - \frac{1}{N'}\sigma^2 = \frac{N'-1}{N'}\sigma^2 \ .
\end{aligned}$$

The last line uses $\mathcal{E}\{\Delta_{n'}\} = 0$, $\mathcal{E}\{\Delta_{n'}^2\} = \sigma^2$ and the fact that the samples are statistically independent to show the ML estimate of sample variance is biased. Of course, as

$N' \to \infty$, the bias is negligible. We can accept the bias from the ML estimate or use the unbiased estimate,

$$\frac{N'}{N'-1}\hat{\sigma}^2 = \frac{1}{N'-1}\sum_{n'=0}^{N'-1}(g_{n'} - \bar{g})^2 \ .$$

The ML estimates are $\theta = (\bar{g}, \hat{\sigma}^2)$. The mean estimate is unbiased but the variance estimate is biased because of the loss of one degree of freedom when using sample mean $\bar{g}$ in place of μ. Bias is a consequence of using one estimate of a parameter in the estimation of another, but we normally don't have a choice.

B.17.1 Mean-squared error

We can establish a relationship between sample variance, Eq (B.23), and squared bias, $b^2(g) = (\bar{g} - \mu)^2$, through comparisons with the mean-squared error (MSE). While variance describes the random error or *precision* of the measurement,

$$\text{MSE} = \frac{1}{N'}\sum_{n'=0}^{N'-1}(g_{n'} - \mu)^2$$

describes the systematic error or *accuracy* of the measurement, so both have important applications in measurement assessment. Beginning with the above expression for MSE and completing the square,

$$\begin{aligned}
\text{MSE} &= \frac{1}{N'}\sum_{n'=0}^{N'-1}(g_{n'} - \mu)^2 = \left[\frac{1}{N'}\sum_{n'=0}^{N'-1} g_{n'}^2\right] - 2\bar{g}\mu + \mu^2 + \bar{g}^2 - \bar{g}^2 \\
&= \frac{1}{N'}\left[\sum_{N'=0}^{N'-1} g_{n'}^2 - N'\bar{g}^2\right] + \bar{g}^2 - 2\bar{g}\mu + \mu^2 = \frac{1}{N'}\sum_{n'=0}^{N'-1}(g_{n'} - \bar{g})^2 + (\bar{g} - \mu)^2 \\
&= \hat{\sigma}^2 + b^2(g) \qquad \text{(B.24)} \\
\text{RMSE} &= \sqrt{\text{MSE}} = \sqrt{\hat{\sigma}^2 + b^2(g)} \ .
\end{aligned}$$

The mean-square error equals the sample variance plus the squared bias, which means it includes both random and systematic errors as illustrated in the example in Fig B.12. RMSE is the root-mean-squared error.

B.17.2 ML estimation in correlated normally-distributed data

When the elements of data vector $\mathbf{g}$ are independent and identically distributed, as in §B.17, we were able to reduce Eq (B.19) to Eq (B.21) because the covariance matrix for

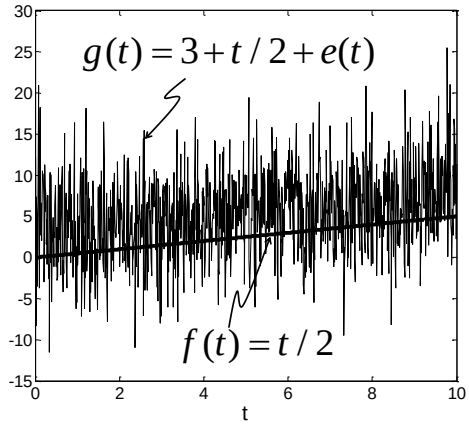


Figure B.12: Object function $f(t)$ represents the population mean for this data set. Data measured from the object, $g(t)$, includes significant additive white-Gaussian noise ($\mathcal{N}(0, 5^2)$) and the temporally-constant bias term shown. From Eq (B.24), MSE $= \hat{\sigma}^2 + b^2(g) = 25 + 9 = 36$. While the standard deviation describing random error is 5, the RMS error is larger, 6, because of the systematic error.

that data is naturally diagonal and all the variances are equal, $\mathbf{K} = \sigma^2\mathbf{I}$. Also $p(\mathbf{g})$ is simply the product of pdfs, $\prod_{n'} p(g_{n'})$. This special situation might occur when $\mathbf{g}$ is a deterministic signal to which white Gaussian noise is added (see Fig B.12). We now consider what happens to the ML estimates in the more general case of correlated normal data.

If $\mathbf{g}$ is real, its covariance matrix is real and symmetric and thus diagonalized by a unitary matrix with columns consisting of eigenvectors of $\mathbf{K}$, like the Fourier matrix in Eq (2.30). Hence, the exponent of Eq (B.19) is diagonalized using $\mathbf{K}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\dagger$ to find (see §8.3 in [1])

$$(\mathbf{g} - \bar{\mathbf{g}})^t\mathbf{K}^{-1}(\mathbf{g} - \bar{\mathbf{g}}) = \sum_{n'=0}^{N'-1} \frac{\Delta\beta_{n'}^2}{\lambda_{n'}} \ .$$

For vector $\Delta\mathbf{g} = \mathbf{g} - \bar{\mathbf{g}}$, $\Delta\boldsymbol{\beta} = \mathbf{Q}^\dagger\Delta\mathbf{g}$ is the discrete Karhunen-Loeve expansion of the modified data vector $\Delta\mathbf{g}$. The product generates an $N' \times 1$ vector of uncorrelated coefficients, $\Delta\boldsymbol{\beta} = [\Delta\beta_0 \ \ldots \ \Delta\beta_{N'-1}]^t$. $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues, $\mathrm{diag}(\mathbf{\Lambda}) = \lambda_0 \ \ldots \ \lambda_{N'-1}$. These quantities allow the expression of $\Delta\mathbf{g}^t\mathbf{K}^{-1}\Delta\mathbf{g}$ as the sum of uncorrelated coefficients. We will have more to say about Karhunen-Loeve expansions in Chapter 4.

The ML estimates for the specific data set $\mathbf{g}$ are

$$\begin{aligned}\hat{\boldsymbol{\mu}} \triangleq \bar{\mathbf{g}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{g} \\ \hat{\mathbf{K}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{g}-\bar{\mathbf{g}})(\mathbf{g}-\bar{\mathbf{g}})^t \ .\end{aligned}$$

The algebra required to obtain this result is shown elsewhere [11] and the result is pretty intuitive, so it is not repeated here. *Remember that* $\mathbf{g}$ *is a specific data set, an* $N' \times 1$ *vector of measurement values, and not the general random variable, which makes these results ML estimates of distribution parameters.*

Definition B.17.1.

Estimators *are mathematical expressions or algorithms that input data and output a statistic that represents a parameter associated with a statistical model of the data.*

Efficient Estimator *is the one that yields the smallest MSE for a specific parameter.*

Consistent Estimators *yield estimates that approach the true value asymptotically. For example, the ML estimator for sample mean is consistent since*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} g_n = \mu \ .$$

The ML estimator of sample variance is biased. $\hat{\sigma}^2$ *is not an efficient estimator but it is consistent since* $\lim_{N\to\infty}\mathcal{E}\{\hat{\sigma}^2\} = \sigma^2$*. The unbiased estimate of sample variance is an efficient estimator.*

Summary

- $\mathcal{N}(\mu, \sigma^2)$ univariate normal pdf; $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ MVN pdf
- $\mathcal{E}X(t)$ population mean; $\mathcal{E}X$ for a stationary process
- $\bar{x}(t)$ sample ensemble mean; $\bar{x}$ for stationary process
- $\widehat{\mathrm{var}}_x(t)$ sample ensemble variance; $\widehat{\mathrm{var}}_x$ for stationary process
- $\langle x_n \rangle$ sample temporal-average mean for nth waveform realization
- $s^2_{x_n}$ sample temporal-average variance for nth waveform realization

B.18 Functions of random variables

Often we are interested in functions of a r.v. Up to this point, we have been discussing random processes that generate $X(t)$, which is a random function of nonrandom variable t. In this section, we expand the discussion to include functions of the random function, i.e., $f(X(t))$. This subject comes up in measurements of all types. For example in photon imaging, the number of photons falling on a detector, X, is a Poisson r.v. We can defend that assumption by arguing the absorption of photons by a detector meets the four criteria listed in §B.8.2, where 'cell proliferation' in that discussion is replaced by 'photon accumulation.'

While the statistics of photons falling on a sensor is Poisson, the sensor and associated instrumentation can influence the data to change the statistical properties of the recording. How can we account for that influence? Since we cannot measure events X directly, we must understand how a Poisson process is passed through functions or operators that model instruments that give signals we can measure. In Chapter 1 we might have modeled the process as a linear system with multiplicative and additive noise sources; perhaps something like $g(t) = \mathcal{H}\{f(t)[1 + \sqrt{X(t)}]\} + e(t)$, where X is a Poisson process describing quantum noise[4] from the variability in photon number over time, and e is a normal process representing additive electronic noise.

B.18.1 Univariate probability transformations

We begin more simply. Let $X(t)$ be a time-varying continuous r.v. with elements $\{x(t)\}$, and let $Y(t) = f(X(t))$ be a one-to-one monotonic transformation of $X(t)$ (see Fig B.13), where one value of $x(t)$ is associated with one value of $y(t)$. Therefore the inverse $x = f^{-1}(y)$ is well defined. Suppose the probability in region Δx of X is approximated by the probability in region Δy of Y. In that case,

$$p_Y(y)\Delta y = p_X(x)\Delta x \ . \tag{B.25}$$

For these conditions, the linear relationship between the two r.v.s $\Delta y \simeq |dy/dx|\Delta x$ from the geometry of Fig B.13 is reasonable, where derivative $|dy/dx|$ is the Jacobian of the transformation. Combining this result with Eq (B.25) yields

$$p_Y(y) = p_X(x)\frac{\Delta x}{\Delta y} = p_X(x)\left|\frac{dx}{dy}\right| = p_X(f^{-1}(y))\left|\frac{df^{-1}(y)}{dy}\right| \ . \tag{B.26}$$

The last form of Eq (B.26) is a bit strange, and yet it reminds us that we need to express the original pdf and the Jacobian in terms of the new variable $y(t)$. Examples illustrate

[4] Quantum noise X is a multiplicative Poisson process where changes in variance follow changes in mean.

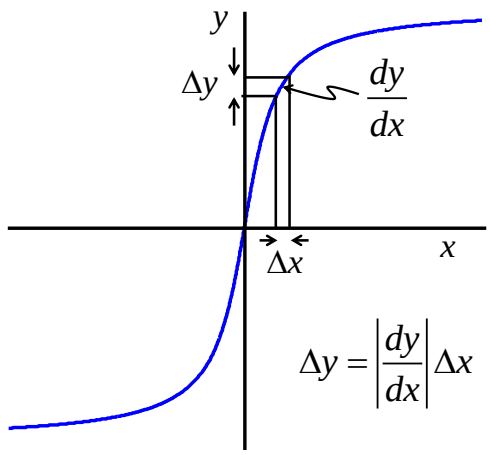


Figure B.13: Example of an invertible probability transformation between r.v.s X and Y.

use of Eq (B.26).

Example B.18.1. *Suppose the input r.v. X is described by a one-sided exponential density, $p_X(x) = \exp(-x)$ for $x > 0$. The challenge is to find $p_Y(y)$ where $y = f(x) = ax + b$. Computing the terms in the last form of Eq (B.26), $x = f^{-1}(y) = (y-b)/a$ and $|df^{-1}(y)/dy| = |a|^{-1}$, we find*

$$\begin{aligned} p_Y(y) &= p_X((y-b)/a)|a|^{-1} \\ &= \frac{1}{|a|}e^{-(y-b)/a} \quad \text{for} \quad (y-b)/a \geq 0 \ \text{ and } \ a \neq 0 \ . \end{aligned}$$

The range is just $y > b$. We state $(y-b)/2 \geq 0$ because we originally have $x > 0$. Examples for two different values of (a, b) are shown in Fig B.14 (left).

Transformation that are not functions (one-to-one) do not have well-defined inverses $x = f^{-1}(y)$, although they may be well defined when considered piecewise. In that case, by partitioning $f(x)$ into M monotonic segments we find Eq (B.26) extends to

$$p_Y(y) = \sum_{m=1}^{M} p_X(f_m^{-1}(y)) \left| \frac{df_m^{-1}(y)}{dy} \right| \ . \tag{B.27}$$

Example B.18.2. *Let $y = Ax^2$ where $x = f^{-1}(y) = +\sqrt{y/A}$ for $x > 0$ and $-\sqrt{y/A}$ for $x \leq 0$. This is not a one-to-one transformation, so we need to partition X into halves at the origin. We find that $|df^{-1}(y)/dy| = (2\sqrt{Ay})^{-1}$ for all y except at the origin. Therefore*

$$p_Y(y) = \frac{p_X(\sqrt{y}) + p_X(-\sqrt{y})}{2\sqrt{y}} \ . \tag{B.28}$$

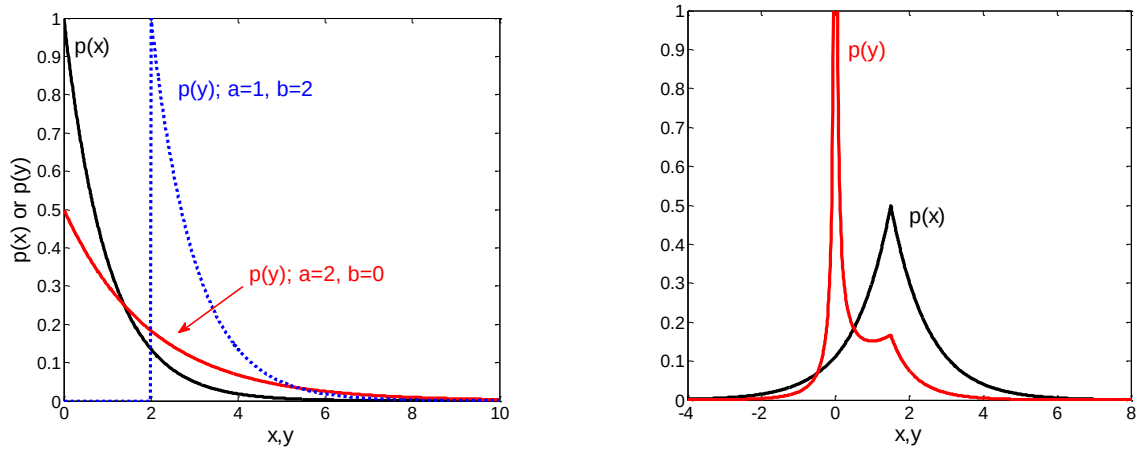


Figure B.14: (left) Illustration of Example B.18.1 where r.v. X follows a unit exponential and the output is through $y = ax + b$ for $a = 2, 1$ and $b = 0, 2$. (right) Results of Example B.18.2 where $A = 1$ and $a = 1.5$.

Selecting the two-sided exponential pdf $p_X(x) = 0.5 \exp(-|x - a|)$ *and applying this to Eq (B.28), we find the results in Fig B.14 (right). You see hints of* $p_X(x)$ *in* $p_Y(y)$ *but the* $(Ay)^{-1/2}$ *scaling term tends to dominate near the origin.*

B.18.2 Functions of multivariate random variables

In practice, we are most interested in functions of multivariate r.v.s. This would include situations where you are analyzing time series or image data. Extending the univariate discussion from §B.18.1, we have $\mathbf{Y} = f(\mathbf{X})$ where $\mathbf{X}$ and $\mathbf{Y}$ are column vectors of random variables having dimensions $N \times 1$ and $M \times 1$, respectively. For the current discussion, assume $M = N$ and $\mathbf{X} = f^{-1}(\mathbf{Y})$ exists. Then, expanding Eq (B.26) and using $\mathbf{x}(\mathbf{y}) \triangleq f^{-1}(\mathbf{y})$,

$$p_Y(\mathbf{y}) = p_X(\mathbf{x}(\mathbf{y}))/|\det(\partial\mathbf{y}/\partial\mathbf{x})|, \tag{B.29}$$

where

$$\det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) = \begin{vmatrix} \frac{\partial y_0}{\partial x_0} & \frac{\partial y_0}{\partial x_1} & \cdots & \frac{\partial y_0}{\partial x_{N-1}} \\ \frac{\partial y_1}{\partial x_0} & \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_{N-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{M-1}}{\partial x_0} & \frac{\partial y_{M-1}}{\partial x_1} & \cdots & \frac{\partial y_{M-1}}{\partial x_{N-1}} \end{vmatrix}.$$

Also notice that the properties of determinants give $|\det(\partial\mathbf{x}/\partial\mathbf{y})| = |\det(\partial\mathbf{y}/\partial\mathbf{x})^{-1}| = 1/|\det(\partial\mathbf{y}/\partial\mathbf{x})|$, which was used in Eq (B.29). Determinant $\det(\partial\mathbf{y}/\partial\mathbf{x})$ is the Jacobian of the transformation.

Example B.18.3. *x_0 and x_1 are independent, standard normal random variables with identical pdfs; in the developing jargon, they are i.i.d. MVN. Their joint density is*

$$p_X(\mathbf{x}) = p_X(x_0, x_1) = p(x_0)p(x_1) = \frac{1}{2\pi} \exp\left(-(x_0^2 + x_1^2)/2\right) \ .$$

We are interested in two new variables $y_0 = x_0 + x_1$ and $y_1 = x_0 - x_1$. Find $p_Y(\mathbf{y})$.

The Jacobian from Eq (B.29) is

$$\det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) = \det\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = -2 \ .$$

Also $\mathbf{x}(\mathbf{y})$ is $x_0 = (y_0 + y_1)/2$ and $x_1 = (y_0 - y_1)/2$. Therefore

$$p_Y(\mathbf{y}) = \frac{1}{4\pi} \exp(-(y_0^2 + y_1^2)/4) \ .$$

Note that y_0 and y_1 are separable and thus independent, just like x_0 and x_1.

B.19 Moment-generating functions

Moments are important summary measures of probability distributions and are often much easier to estimate from sample data than the distribution itself. For random variable X and complex nonrandom variable s, the moment-generating function is

$$\begin{aligned} M_X(s) &= \mathcal{E} e^{sX} && \text{(B.30)} \\ &= \int_{-\infty}^{\infty} dx \; e^{sx} \, p_X(x) && \text{for continuous random variables} \\ &= \sum_{x_n : p(x_n) > 0} e^{sx_n} \, p_X(x_n) && \text{for discrete random variables .} \end{aligned}$$

$s = \sigma + i\Omega$ is a frequency-like complex variable[5] with units of $[x]^{-1}$. Statistical moments are found by successive differentiation of M with respect to s followed by setting $s = 0$:

$$\begin{aligned} M_X^{(1)}(0) &= \left[\frac{d}{ds} \mathcal{E} e^{sX}\right]_{s=0} = \mathcal{E}\left[\frac{d}{ds} e^{sX}\right]_{s=0} = \mathcal{E}X \\ M_X^{(m)}(0) &= \mathcal{E}\left[\frac{d^m}{ds^m} e^{sX}\right]_{s=0} = \mathcal{E}X^m \ . \end{aligned}$$

---

[5] Actually s is the Laplace frequency variable whose imaginary part is the Fourier frequency variable $\Omega = 2\pi u$. The moment-generating function has the form of a 2-sided, conjugate-Laplace transform of the density/mass function for a distribution. That is, $M_X(s) = \mathcal{E}e^{sX} = \int_{-\infty}^{\infty} dx \; p_X(x) \, e^{sx} = \mathcal{L}\{p_X(x)\}_{s \to -s} = \mathcal{L}^* p_X(x)$.

Superscript $^{(m)}$ in parentheses identifies the number of derivatives taken with respect to s, while X^m denotes X raised to the mth power. We can reverse the order of differentiation and expectation (most of the time) because both are linear operators. Another view is to consider the power series expansion of the exponent, $e^x = \sum_{n=0}^{\infty} x^n/n!$,

$$\begin{aligned} M_X(s) &= \mathcal{E}e^{sX} = \mathcal{E}\left\{\sum_{k=0}^{\infty}\frac{s^k}{k!}X^k\right\} = \sum_{k=0}^{\infty}\frac{s^k}{k!}\mathcal{E}X^k \\ M_X^{(m)}(0) &\triangleq \left.\frac{d^m M_X}{ds^m}\right|_{s=0} = \left.\sum_{k=m}^{\infty}\frac{s^{k-m}}{(k-m)!}\mathcal{E}X^m\right|_{s=0} = \mathcal{E}X^m \ . \end{aligned}$$

Applying this method, let's compute moments from the Poisson and standard normal distributions.

Example B.19.1. *Find the first two moments and the variance of Poisson distribution with parameter λ*

$$p_X(n;\lambda) = \Pr(X = n) = \lambda^n e^{-\lambda}/n!$$

using moment generating functions via Eq (B.30).

$$\begin{aligned} M_X(s) &= \mathcal{E}e^{sX} = \sum_{n=0}^{\infty} e^{sn}\frac{\lambda^n}{n!}e^{-\lambda} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^s)^n}{n!} \\ &= e^{-\lambda}\exp(\lambda e^s) = \exp(\lambda(e^s-1)) \\ M_X^{(1)}(0) &= \lambda e^s \exp(\lambda(e^s-1))|_{s=0} = \lambda \\ M_X^{(2)}(0) &= \left[\lambda e^s\exp(\lambda(e^s-1)) + \lambda^2 e^{2s}\exp(\lambda(e^s-1))\right]_{s=0} = \lambda^2+\lambda \end{aligned}$$

Of course, $\mathrm{var}(X) = \mathcal{E}X^2 - (\mathcal{E}X)^2 = \lambda$. *This is one way to do one of the homework problems. You need to find another way to do the homework problem.*

Example B.19.2. *Find the first three moments of a standard normal distribution*

$$p_Z(z;0,1) = \mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$$

using its moment generating function.

$$\begin{aligned} M_Z(s) &= \mathcal{E}e^{sZ} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dz\, e^{sz}e^{-z^2/2} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dz\, e^{-(z^2-2sz)/2} \\ &= \frac{e^{s^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} d(z-s)\, e^{-(z-s)^2/2} = e^{s^2/2} \qquad \text{(B.31)} \\ M_Z^{(1)}(0) &= \left. s\, e^{s^2/2}\right|_{s=0} = 0 \\ M_Z^{(2)}(0) &= \left[(s^2+1)\, e^{s^2/2}\right]_{s=0} = 1 \\ M_Z^{(3)}(0) &= \left[(s^3+3s)\, e^{s^2/2}\right]_{s=0} = 0 \ . \end{aligned}$$

We find that the mean and variance is 0 and 1 as expected. The scaled, central, third moment is called skewness *in statistics. It is $\mathcal{E}(X-\mu)^3/\sigma^3$ generally or $\mathcal{E}Z^3$ for the standard normal process where $Z = (X-\mu)/\sigma$. The third moment indicates symmetry of the pdf about the mean. Since normal distributions are symmetric, the skewness measure is zero. In fact, the symmetry of the standard normal pdf sets all the odd moments to zero.*

If we can find $M(s)$, we can generate any moment of $p(x)$ or $p[n]$.

B.20 Characteristic functions

The *characteristic function* of X can be obtained from the moment-generating function using the relation $C_X(\Omega) = M_X(s)_{s=i\Omega}$, where $\Im\{s\} = \Omega = 2\pi u$. Just as $M(s)$ is the conjugate-Laplace transform of $p(x)$, $C(\Omega)$ is the conjugate-Fourier transform of $p(x)$,

$$C_X(\Omega) = \int_{-\infty}^{\infty} dx\, p_X(x)\, e^{i\Omega x} = \left[\int_{-\infty}^{\infty} dx\, p_X(x)\, e^{-i\Omega x}\right]^* = \mathcal{F}^*\{p_X(x)\} = \mathcal{E}e^{i\Omega X} \ . \quad \text{(B.32)}$$

The characteristic function may be thought of as the conjugate FT of $p_X(x)$ or equivalently as the expected value of a complex exponential involving X. Context dictates whether you wish to consider the pdf $p_X(x)$ or the Fourier basis $\exp(i\Omega x)$ as the kernel of the transformation.

Example B.20.1. *From the moment-generating function computed in Example B.19.2, compute the characteristic function for $Z \sim \mathcal{N}(0,1)$.*

$$C_Z(\Omega) = M_Z(s)_{s=i\Omega} = e^{-\Omega^2/2} \ .$$

Find $C_X(\Omega)$ for $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$\begin{aligned} C_X(\Omega) &= \mathcal{E}e^{i\Omega X} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, e^{-(x-\mu)^2/2\sigma^2}\, e^{i\Omega x} \\ &= e^{i2\pi u\mu}\, e^{-2\pi^2 u^2 \sigma^2} = e^{i\Omega\mu - \Omega^2\sigma^2/2} \ . \end{aligned}$$

The FT of a Gaussian function from §2.7 was used here. The complex conjugate of that result gave the equation above.

The characteristic function for a discrete, integer-valued r.v. $X[n]$ that is periodic over the time range T_0 is given by a variation of the Fourier series expression,

$$C_X(\Omega) = \mathcal{E}e^{i\Omega X} = \sum_n p_X(n)\, e^{i2\pi un} \ .$$

$C_X(\Omega)$ are interpreted as coefficients that properly summed reconstruct the original pdf where

$$p(n) = \Pr(X = n) = \frac{2}{T_0} \int_{-T_0/2}^{T_0/2} du\, C_X(\Omega)\, e^{-i2\pi un} \ .$$

The variation is that complex conjugates of the Fourier operators are involved. The sum is over all values of n in $\mathtt{S}$ for which $p(n) > 0$.

An important application of characteristic functions is in describing the pdf of the sum of probability density (or mass) functions.

Example B.20.2. *Let X and Y are independent random variables with densities $p_X(x)$ and $p_Y(y)$. If $Z = X + Y$, find $p_Z(z)$ in terms of the other two pdfs.*

First, we find the general expression. Beginning with the cdf for Z and noting $Y = Z - X$,

$$\begin{aligned} P_Z(z) &= \Pr(X + Y \le z) = \int_{-\infty}^{\infty} dx\, p_X(x) \int_{-\infty}^{z-x} dy\, p_y(y) \\ &= \int_{-\infty}^{\infty} dx\, p_X(x) P_Y(z - x) \ . \end{aligned}$$

Differentiating with respect to Z, the cdfs are converted into pdfs,

$$p_Z(z) = \int_{-\infty}^{\infty} dx\, p_X(x) \frac{dP_Y(z-x)}{dz} = \int_{-\infty}^{\infty} dx\, p_X(x)\, p_Y(z - x) \ . \tag{B.33}$$

Hence the pdf of the sum of two independent r.v.s is given by the convolution of the component pdfs, $p_Z(z) = [p_X * p_Y](z)$.

Now, perhaps, you can see the value of characteristic functions. To find $p_Z(z)$, first find $C_X(\Omega)$ and $C_Y(\Omega)$, take their product $C_Z(\Omega) = C_X(\Omega)C_Y(\Omega)$, and convert $C_Z(\Omega) \to p_Z(z)$ using Fourier transforms. For example,

$$\begin{aligned} p_X(x) &= \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \qquad \text{and} \qquad p_Y(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2) \\ C_X(u) &= \exp(-2\pi^2 u^2) \qquad \text{and} \qquad C_Y(u) = \exp(-2\pi^2 u^2) \\ C_Z(u) &= C_X(u) C_Y(u) = \exp(-4\pi^2 u^2) \\ p_Z(z) &= \frac{1}{2\sqrt{\pi}} \exp(-z^2/4) \ . \end{aligned}$$

The last line requires that you know the inverse conjugate Fourier transform of a Gaussian function,

$$\left(\mathcal{F}^{-1}\right)^* \{\exp(a^2 u^2)\} = \frac{\sqrt{\pi}}{a} \exp(-\pi^2 x^2/a^2) \ ,$$

which is also the inverse transform. It is very helpful to remember that

$$\mathcal{F}\left\{\frac{1}{\sqrt{2\pi}}\exp\big(-(t-t_0)^2/2\sigma^2\big)\right\} = \exp(-i2\pi t_0 u)\exp(-2\pi^2\sigma^2u^2)\ .$$

You can see that for N i.i.d. normal r.v.s, X_j, the pdf of their sum is

$$p_Z(z) = \frac{1}{\sqrt{2\pi N}}\exp(-z^2/2N)\ , \qquad \text{for}\ \ Z = \sum_{j=1}^{N} X_j\ .$$

The sum of independent normal random variables is another normal random variable.

4.6 Chapter 4 Problems

1. A blood test is 95% effective at detecting the HIV virus when a patient is infected. However, the test has a 1% false positive rate for healthy persons that are tested. (a) If 0.5% of the population has the disease, what is the positive predictive value, PPV? (b) How do the results change if the sensitivity increases to 100% and prevalence in the population falls to 0.1%? (c) For PPV to remain high, is it more important for the test to be highly sensitive or highly specific?

2. (a) Find an expression for $\Pr(X > 2)$ for the Poisson random variable X.

   (b) Find $\Pr(X > 2)$ for $\lambda = 2$.

   (c) Find λ such that $\Pr(X > 2) = 0.5$.

3. Let $p_{XY}(x, y)$ be a standard bi-normal pdf where X and Y are independent.

   (a) Find $p_{R\Theta}(r, \theta)$ where $r^2 = x^2 + y^2$ and $\theta = \tan^{-1}(y/x)$.

   (b) Find the marginal densities $p_R(r)$ and $p_\Theta(\theta)$ from part a.

   (c) Are the marginal densities independent? Why?

4. Let X and Y be independent Poisson r.v.s, $\mathcal{P}_X(\lambda_x)$ and $\mathcal{P}_Y(\lambda_y)$. Show that if $Z = X + Y$ then $\mathcal{P}_Z(\lambda_z) = \mathcal{P}_Z(\lambda_x + \lambda_y)$. Do not use characteristic functions for this problem. It helps to review Example B.20.2 and to know the binomial formula,

$$\begin{pmatrix} N \\ n \end{pmatrix} = \frac{N!}{n!(N-n)!} \qquad \text{and} \quad (a+b)^N = \sum_{n=0}^{N} \begin{pmatrix} N \\ n \end{pmatrix} a^n b^{N-n} \ .$$

5. Derive the variance for a Poisson process.

6. You measure a patient's arterial pO_2 value (partial pressure of oxygen, sometimes paO_2) to be 90 mm Hg. The patient does not appear to be in distress but you can't remember the normal range so you look it up. According to the chart, the healthy range is >10.5 kPa. Oh, oh.

   - Is the patient OK? Why?
   - Let θ represent pO_2. Is is a normally-distributed random variable in the healthy adult population given by $\mathcal{N}(\theta_0, \sigma_\theta^2)$, where $\theta_0 = 13.33$ kPa. Also let the threshold for taking emergency action be $\theta_t < 10.5$ mmHg, which accounts for 95% of normal adults. You know some patients will be OK and yet fall below that threshold. What must parameter σ_θ be for θ_t to account for 95% of the normal adult population?

7. You are asked to test the Central Limit Theorem using numerical methods. (a) Start with a MATLAB uniform random variable routine and generate a matrix of 1000×10000 univariate samples. Histogram the first row of 10000 samples. Then sum the first two rows and histogram the resulting 10000 summed values. Repeat for the first 10 rows and for all 1000 rows. Use a MATLAB routine to run a test for normality. In a 2x2 subplot, show me histograms and best-fit normal pdf curves. (b) Repeat part (a) using a log-normal r.v. Also plot the 2x2 matrix of histograms and fits. What can you conclude about differences in convergence?

8. (a) Simulate an observer study by drawing many normally-distributed samples using `randn` from two distributions. They have equal variance $\sigma_1^2 = \sigma_0^2 = \sigma^2$ but the means are separated according to $d' = \Delta\theta/\sigma = 1.5$. Histogram the results, and use the histograms to generate an ROC curve. Plot the two histograms together on the same axes. Also plot the ROC curve and find the AUC. (b) Repeat (a) with the same parameters except use $p(\theta|H_1)$ where $\sigma_1 = 2\sigma_0$ from (a). Plot these new distributions, the ROC curve and give the AUC. What is different about the ROC curve for part (b) that results when the distributions have unequal variances?

9. A Cauchy random variable is an example of a single-parameter, continuous r.v. It is an even function with a shape somewhat like a normal distribution. Its pdf is given by

$$p(x|\lambda) = \mathcal{C}(\lambda) = \frac{\lambda/\pi}{\lambda^2 + x^2} \ .$$

(a) Find the cdf for X that follows a Cauchy pdf when $\lambda = 1$. Plot $p(x)$ and $P(x)$ on the same graph.

(b) You enter the Illinois λ lottery by selecting a number $0 \leq \lambda \leq 10$. Notice there are an infinite number of numbers to select in this range! The winner is then chosen from the r.v. X that follows a Cauchy density with parameter λ. If $|X| > 1$, you win the lottery and 100 million dollars so you can forget about all this school nonsense. What value of λ should you pick to maximize your chances of winning?