

STATISTICAL REVIEW

This is an overview of some statistical definitions and procedures that you should be aware of.

1. BASIC DEFINITIONS

Start with a triple (Ω, \mathcal{B}, P) where Ω is the set of objects, \mathcal{B} is a collection of “measurable” subsets of objects, and P is the probability measure which is a mapping from the domain \mathcal{B} to $[0, 1]$. The measure P satisfies certain properties:

- $P(\Omega) = 1$
- $0 \leq P(A) \leq 1$ for all $A \in \mathcal{B}$
- $P(\bigcup A_n) = \sum_n P(A_n)$ for any (countable) collection of pairwise disjoint sets A_n .

It’s actually the last property that requires that we specify the collection of sets \mathcal{B} because if Ω is an uncountably infinite collection of objects (like the real numbers), then the last property is not consistent with allowing the domain of P to be the collection of all possible subsets of Ω (i.e. 2^Ω). Notice, however, that there is no problem with taking \mathcal{B} to be the power set of Ω if Ω is countable.

Most of statistics takes place in a derived probability space which is found by introducing a *random variable* (which is neither random nor a variable).

1.1. Random variables.

Definition. A *random variable* is a map from Ω to the real line \mathbb{R} : $X : \Omega \rightarrow \mathbb{R}$.

A random variable is often denoted with a capital letter or with a bold-face letter, or some other distinguishing topographical style. This mapping creates an induced triple $(\mathbb{R}, \mathcal{B}_X, P_X)$ where \mathcal{B}_X is the collection of images of the sets in \mathcal{B} under X and P_X is the induced probability whose domain is \mathcal{B}_X . Notice, that we require that \mathcal{B}_X is a particular subset of the so-called **Borel field**, which is the smallest collection of sets of real numbers containing all of the open sets and having the three properties.

- If A and B are in the field, then $A \cup B$ is also.
- If A is in the field, then \tilde{A} (the complement of A) is also
- If each A_i in a countably (infinite) collection of sets is in the field, then $\bigcup A_i$ is also. (Notice this is different than the first property because we require that an *infinite* union of sets be in the field).

These technical requirements allow us to define a probability measure P_X which satisfies the properties given above. Notice that \mathcal{B}_X contains all of open and closed intervals. In addition, it contains half-infinite intervals

like $(-\infty, x]$. Thus, we can define for all x the function

$$F_X(x) = P_X((-\infty, x]) = \text{Prob}\{X(a) \leq x\}.$$

Using the notion of inverse images this can be related to the original probability measure

$$F_X(x) = \text{Prob}\{a \in X^{-1}((-\infty, x])\} = P(A).$$

So, $F_X(x)$ is a well-defined function for each x .

Definition. The *Cumulative Distribution Function (CDF)* is a map from the real numbers to the interval $[0, 1]$ defined by

$$F_X(x) = \text{Prob}\{X(a) \leq x\}.$$

Notice that the rate-of-change of this function with respect to x tells us something about which values of X are more likely to happen. We call this derivative, the probability density function.

Definition. The *Probability Density Function (PDF)* is the derivative of the CDF:

$$p_X(x) = \frac{dF_X(x)}{dx}.$$

We allow this function to contain delta functions to handle discrete (i.e. countable) object spaces Ω . Notice, then that

$$F_X(x) = \int_{-\infty}^x p_X(t) dt$$

and we have to integrate the PDF to get a probability measure. The integral of the PDF over any interval gives us the probability of X mapping to values in that interval:

$$P_X(A) = \int_A p_X(x) dx.$$

The expectation operator is an important tool particularly in defining moments. Intuitively, it defines an average of a function with respect to the PDF of a random variable.

Definition. The *expected-value* of a function $g(X)$ of a random variable is the integral of the PDF with respect to this function:

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) p_X(x) dx.$$

Notice, that the expectation operator E is a linear operator. This is a useful observation which can help you avoid unnecessary integral calculations.

It is often useful to describe a random-variable in terms of the moments of its PDF. We define two types of moments: non-central and central-moments.

Definition. *Non-central moments* of a random-variable are the expected value of X^n for a non-negative integer n :

$$\mu'_n = E \{X^n\} = \int_{-\infty}^{\infty} x^n p_X(x) dx.$$

Notice that $\mu'_0 = 1$. The first non-central moment is called the *mean* of the distribution. $\mu'_1 = E \{X\} = \mu$.

Definition. *Central moments* describe the expected value of the distance from the mean to some power. Central moments give a measure of how significant the mean really is.

$$\mu_n = E \{(X - \mu)^n\} = \int_{-\infty}^{\infty} (x - \mu)^n p_X(x) dx.$$

Again, $\mu_0 = 1$ and $\mu_1 = 0$. The first useful central moment is μ_2 . This is called the variance of the random variable and sometimes denoted $\sigma^2 \equiv \mu_2$. The variance gives a measure of how far the random variable is from the mean on average. It is easy to show that central-moments can be written as algebraic sums of non-central moments using the binomial expansion and linearity of the expectation operator:

$$\begin{aligned} \mu_n &= E \{(X - \mu)^n\} \\ &= \sum_{k=0}^n \frac{n!}{k!(n-k)!} E \{X^{n-k}\} (-1)^k \mu^k \\ &= \sum_{k=0}^n \frac{(-1)^k n!}{k!(n-k)!} \mu^k \mu'_{n-k}. \end{aligned}$$

In particular.

$$\begin{aligned} \mu_2 &= \mu'_2 - \mu^2 \\ E \{(X - E \{X\})^2\} &= E \{X^2\} - E^2 \{X\}. \end{aligned}$$

1.2. Random Vectors. In signal processing applications, rarely is a single random-variable enough. In an image, for example, each pixel in the image is usually represented as a random-variable. A finite collection of random variables is called a random-vector

$$\mathbf{X} = [X_1, X_2, \dots, X_N]^T.$$

The random-vector can be described by the joint probability density function

$$p_{\mathbf{X}}(x_1, \dots, x_n) = p_{\mathbf{X}}(\mathbf{x}).$$

This function has a similar interpretation as in the one-dimensional case: it describes where in N -dimensional space the vector \mathbf{X} is most likely to map to, and integrals (over all the variables) give the probability that X mapped to that interval:

$$P_X(A) = \int_A p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Here A is a collection of N -dimensional intervals. If integration takes place over fewer than N variables, a probability density function in the other variables (the marginals) is obtained:

$$p_{x_i}(x_i) = \int p_{\mathbf{X}}(\mathbf{x}) d\tilde{\mathbf{x}}_i$$

where $\tilde{\mathbf{x}}_i = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N]$.

With more than one random variable to consider, the idea of *conditional probability* becomes important. This concept allows us to think of some of the random-variables as fixed and consider the statistics of the other random-variables.

Definition. The *Conditional* probability density function is defined as the ratio of the joint-density function and the marginal density functions. If \mathbf{X}_i is a subset of the random vector \mathbf{X} and \mathbf{X}_j is the remaining random-variables in \mathbf{X} then

$$p_{\mathbf{X}_j|\mathbf{X}_i}(\mathbf{x}_j | \mathbf{x}_i) p_{\mathbf{X}_i}(\mathbf{x}_i) = p_{\mathbf{X}}(\mathbf{x}).$$

Conditional PDF's tell us how much uncertainty is left in \mathbf{X}_j if \mathbf{X}_i is known. From an estimation perspective, we can set \mathbf{X}_j to be what we are trying to find and \mathbf{X}_i to be the measurements. Then, a very strong case can be made for considering the conditional probability density function as *the answer* to our estimation question. We will explore this idea more below. Notice, that we can interchange the roles of \mathbf{X}_j and \mathbf{X}_i and get

$$p_{\mathbf{X}_i|\mathbf{X}_j}(\mathbf{x}_i | \mathbf{x}_j) p_{\mathbf{X}_j}(\mathbf{x}_j) = p_{\mathbf{X}}(\mathbf{x}).$$

Thus, we quickly arrive at *Bayes rule*:

$$p_{\mathbf{X}_j|\mathbf{X}_i}(\mathbf{x}_j | \mathbf{x}_i) = \frac{p_{\mathbf{X}_i|\mathbf{X}_j}(\mathbf{x}_i | \mathbf{x}_j) p_{\mathbf{X}_j}(\mathbf{x}_j)}{p_{\mathbf{X}_i}(\mathbf{x}_i)}.$$

Definition. Random vectors (variables) \mathbf{X}_i and \mathbf{X}_j are *independent* if

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{X}_i}(\mathbf{x}_i) p_{\mathbf{X}_j}(\mathbf{x}_j)$$

where $\mathbf{X}^T = [\mathbf{X}_i^T, \mathbf{X}_j^T]$.

Notice that if two random-variables are independent then the conditional probability density function is the same as the marginal density function (thus conditioning on an independent random-variable doesn't change the uncertainty we have about a particular random variable).

The expectation operator is defined in a similar fashion to the one-dimensional case as the integral of the joint-probability density function with the function:

$$E\{g(\mathbf{X})\} = \int g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Analogously we define conditional expectations to be the same integral but with respect to the conditional probability density function:

$$E\{g(\mathbf{X}) | \mathbf{X}_i\} = \int g(\mathbf{x}) p_{\mathbf{X}_j|\mathbf{X}_i}(\mathbf{x}_j|\mathbf{x}_i) d\mathbf{x}_j.$$

In general $g(\mathbf{X})$ is a function of all the random-variables but two-special cases deserve particular mention. Let X_m and X_n be two random variables in the vector \mathbf{X} , the cross-correlation matrix of the vector \mathbf{X} is the matrix formed by

$$R_{mn} = E\{X_m X_n\}.$$

The covariance matrix is computed using zero-mean random-variables

$$C_{mn} = E\{(X_m - E\{X_m\})(X_n - E\{X_n\})\}.$$

If $C_{mn} = 0$, the two variables are said to be *uncorrelated*. Notice that

$$C_{mn} = R_{mn} - E\{X_m\}E\{X_n\}.$$

A complex-valued random-vector is one with a real and imaginary part

$$\mathbf{Z} = \mathbf{X} + j\mathbf{Y}.$$

Certain kinds of complex-valued random vectors can be well-described in terms of a slightly modified correlation matrix

$$R_{mn} = E\{Z_m Z_n^*\} = E\{X_m X_n\} + E\{Y_m Y_n\} - j[E\{X_m Y_n\} - E\{X_n Y_m\}]$$

1.3. Common density functions. The two most common density functions you will deal with are Gaussian and Poisson. Gaussian density functions are used extensively in signal processing applications primarily because they make the mathematics tractable. In addition, the Gaussian density function satisfies some

useful properties: (1) it is almost always the result of adding together a large number of independent random variables (central-limit-theorem) — so if your noise process involves many independent contributing factors, a Gaussian random variable often models this well. (2) if we are only concerned about first and second moments (means and variances) in our data and result, then assuming the random variable to be Gaussian adds the “least” “information” about the underlying probability density function. So, while often an approximation, the Gaussian model is a good one in many situations.

The multivariate Gaussian PDF of a length N random-vector is

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} \sqrt{|\mathbf{C}|}} \exp \left[-\frac{1}{2} (\mathbf{x} - E\{\mathbf{X}\})^T \mathbf{C}^{-1} (\mathbf{x} - E\{\mathbf{X}\}) \right].$$

If the random-variables in the random-vector are independent, then $\mathbf{C} = \sigma_i^2 \mathbf{I}$ and this reduces to

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi \prod_{i=1}^N \sigma_i^2}} \exp \left[-\frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (x_i - E\{X_i\})^2 \right].$$

Another, important density function which is defined for discrete values of x is the Poisson distribution

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \sum_n \delta(x - n).$$

The delta functions just allow us to treat x as a continuous variable. If it is remembered that x only takes on integer values, then the comb can be dropped. We will only use, the multivariate case for independent random-variables

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_i e^{-\lambda_i} \frac{\lambda_i^{x_i}}{x_i!}.$$

1.3.1. *Confidence (concentration) ellipsoids.* For a Gaussian random vector the random-variable

$$r_C^2 = (\mathbf{x} - E\mathbf{X})^T \mathbf{C}^{-1} (\mathbf{x} - E\mathbf{X})$$

is a chi-squared random variable with N degrees of freedom (where N is the number of variables in \mathbf{X}). The probability that $r_C^2 \leq r^2$ gives is calculated using the CDF of the chi-squared distribution evaluated at r^2 . This gives the probability that \mathbf{x} lies within the ellipsoid defined by the equation

$$(\mathbf{x} - E\mathbf{X})^T \mathbf{C}^{-1} (\mathbf{x} - E\mathbf{X}) = r^2.$$

Note that the axes of the ellipsoid have length $1/\lambda_i$ where λ_i is the i^{th} eigenvalue of the matrix \mathbf{C}^{-1} .

1.4. **Complex-valued random variables.** We can always deal with complex-valued random vectors by considering them as the sum of two real-valued random vectors so that $\mathbf{Z} = \mathbf{X} + j\mathbf{Y}$. The statistics of this

random vector can be described in terms of the joint density function

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}).$$

If \mathbf{X} and \mathbf{Y} are jointly normal, then \mathbf{Z} is also normal. We can always write $\mathbf{P}^T = [\mathbf{X}^T, \mathbf{Y}^T]$ and consider the real-valued random-vector \mathbf{P} . The density function is then described by the matrix

$$\mathbf{C}_{\mathbf{P}} = E\{\mathbf{P}\mathbf{P}^T\} = \begin{bmatrix} \mathbf{C}_{\mathbf{X}\mathbf{X}} & \mathbf{C}_{\mathbf{X}\mathbf{Y}} \\ \mathbf{C}_{\mathbf{Y}\mathbf{X}} & \mathbf{C}_{\mathbf{Y}\mathbf{Y}} \end{bmatrix}$$

which can be thought of as the covariance matrix of the (real-valued) vector \mathbf{P} .

In certain cases, it is also possible to describe the density function in terms of a correlation matrix of \mathbf{Z} defined (for a zero-mean vector) as

$$\mathbf{C}_{\mathbf{Z}} = \frac{1}{2}E\{\mathbf{Z}\mathbf{Z}^H\}.$$

(Notice the 1/2 in the previous definition — this allows the PDF to look similar to the real-valued case). If $\mathbf{C}_{\mathbf{X}\mathbf{X}} = \mathbf{C}_{\mathbf{Y}\mathbf{Y}}$ and $\mathbf{C}_{\mathbf{X}\mathbf{Y}} = -\mathbf{C}_{\mathbf{Y}\mathbf{X}} (= -\mathbf{C}_{\mathbf{X}\mathbf{Y}}^T)$, then the distribution of (zero-mean) \mathbf{Z} is determined completely in terms of this correlation matrix as

$$p_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^N |\mathbf{C}_{\mathbf{Z}}|} \exp\left\{-\frac{1}{2}\mathbf{z}^H \mathbf{C}_{\mathbf{Z}}^{-1} \mathbf{z}\right\}$$

where N is the number of elements in each of \mathbf{X} and \mathbf{Y} .

1.5. Fundamental Theorem of Statistics. The fundamental theorem of statistics is used to determine the PDF of a random vector that can be written as some function of another random-vector. So, if $\mathbf{Y} = \mathbf{h}(\mathbf{X})$ where \mathbf{X} is a random vector with known PDF, the fundamental theorem of statistics shows how to determine the PDF of \mathbf{Y} . The idea is simple. Suppose A is an arbitrary collection of sets of real vectors \mathbf{y} . The probability of A is

$$\int_A p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}.$$

Because $\mathbf{Y} = \mathbf{h}(\mathbf{X})$, this can be written as an integral over \mathbf{X} . Suppose for a moment, that \mathbf{h} is one-to-one (and therefore invertible), then we can write this integral over \mathbf{y} in terms of an integral over \mathbf{x} using the PDF of \mathbf{X} .

$$\int_{\mathbf{h}^{-1}(A)} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

Now, apply the change of variables theorem to write this last expression as an integral over \mathbf{y} :

$$\int_A p_{\mathbf{X}}(\mathbf{h}^{-1}(\mathbf{y})) \left| \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|^{-1} d\mathbf{y}.$$

where $\left| \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|$ is the determinant of the derivative matrix of the transformation \mathbf{h} (also called the Jacobian).

Because this is true for arbitrary A we must have

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{h}^{-1}(\mathbf{y})) \left| \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|^{-1} \Bigg|_{\mathbf{h}^{-1}(\mathbf{y})}.$$

If \mathbf{h} is not one-to-one, so that there are multiple vectors \mathbf{x} which map to the same \mathbf{y} , then we have to sum over all of these to get the total probability. Thus for a given \mathbf{y} define as $\mathbf{h}_i^{-1}(\mathbf{y})$ the i^{th} root of the equation $\mathbf{y} = \mathbf{h}(\mathbf{x})$. Then,

$$p_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^{N(\mathbf{y})} p_{\mathbf{X}}(\mathbf{h}_i^{-1}(\mathbf{y})) \left| \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|^{-1} \Bigg|_{\mathbf{h}_i^{-1}(\mathbf{y})}.$$

where $N(\mathbf{y})$ is the number of roots of the equation $\mathbf{h}(\mathbf{x}) = \mathbf{y}$ at \mathbf{y} .

For example, suppose X is a random-variable with a Gaussian distribution with mean 0 and variance σ^2 . Suppose we want to find the distribution of $Y = X^2$. Thus, $h(x) = x^2$ and $h'(x) = 2x$. Also $h^{-1}(y) = \pm\sqrt{y}$ if $y \geq 0$ and there are no (real) roots if $y < 0$. Using the fundamental theorem of statistics we see that

$$\begin{aligned} p_Y(y) &= \frac{p_X(\sqrt{y})}{2\sqrt{y}} u(y) + \frac{p_X(-\sqrt{y})}{2\sqrt{y}} u(y) \\ &= \frac{1}{\sigma\sqrt{2\pi y}} \exp\left(-\frac{y}{2\sigma^2}\right) u(y). \end{aligned}$$

where $u(y)$ is a step function.

Notice, that if \mathbf{h} is a linear (or affine) transformation, then if \mathbf{X} is Gaussian then \mathbf{Y} is also Gaussian.