REVIEW OF RANDOM VARIABLES AND EXPECTED VALUES

1. Random Variables

Let's say you measure the tensile strength of 20 steel rods and get $x_1 = 280\ \mathrm{N/mm}$, $x_2 = 305\ \mathrm{N/mm}$, … , $x_{20} = 298\ \mathrm{N/mm}$. It is impossible to predict the exact tensile strength of the 21st steel rod before you actually test it, but, based on past experience, you can make statements about the likelihood of its value.

Any quantity whose value cannot be precisely predicted ahead of actually measuring it is called a random variable (or probabilistic quantity). In the previous example, x is a random variable.

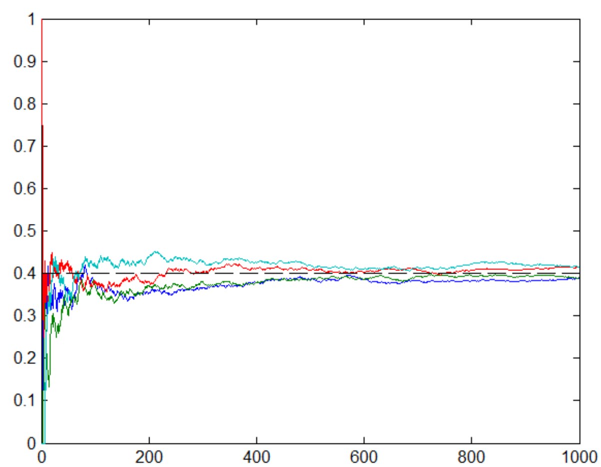
Each of the x_i listed above is called a sample point. The measurement of a sample point (or more generally, the observation of a random phenomenon) is called a trial.

The set of all possible outcomes of a trial is called the sample space, denoted S. Some examples are: a coin $S = \{\text{heads, tails}\}$, a die $S = \{1,2,3,4,5,6\}$, steel tensile strength $S = \{x \mid 250 \le x \le 320\ \mathrm{N/mm}\}$ (assuming it is physically impossible to produce a steel sample outside this range). We can say that a random variable assigns a numerical value to a trial. Technically, it is a function from the sample space to the real or complex numbers.

An event is a subset of the sample space to which a probability can be assigned (technically, the event must be a measurable subset of the sample space). Some examples for rolling a die are: $E = \{1\}$, $E = \{x \mid x \text{ is even}\}$, and $E = \{x \mid x \le 4\}$.

If the probability of an event is P, then in a long series of trials, the fraction of instances where E occurs is approximately P. Note that this definition of probability is non-constructive; you cannot exactly determine the probability of an event in finite time.

The following figure shows the fraction of instances of an event with probability $P = 0.4$. Only after an infinite number of trials can the actual probability be calculated.



This is illustrated by Bernoulli's Law of Large Numbers:

$$\lim_{N\to\infty} P\left\{\left|\frac{N_E}{N} - P(E)\right| \geq \varepsilon\right\} = 0, \text{ for any } \varepsilon > 0$$

which states that the probability that the fraction of occurrences of an event is different from its probability decreases to 0 as the number of trials goes to infinity.

Probabilities must satisfy the following three axioms:

1. For each event E, $0 \leq P(E)$.

2. $P(S) = 1$, where S is the entire sample space.

3. If $E_1, E_2, \ldots$ are <u>mutually exclusive</u>, meaning $i \neq j \Rightarrow E_i \cap E_j = \varnothing$, then $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

2. Probability distributions of a single variable

The <u>cumulative distribution function (CDF)</u> of a real-valued random variable x is defined by $F_x(\xi) = P(x \leq \xi)$. The CDF has the following properties:

1. F_x is non-decreasing, hence the term "cumulative" – as ξ increases it "accumulates" the probability that $x \leq \xi$.

2. F_x is right-continuous. Discontinuities occur if there is a finite probability that x can have a specific value, for example the value on the side of a die.

3. $\lim_{\xi\to-\infty} F_x(\xi) = 0$

4. $\lim_{\xi\to\infty} F_x(\xi) = 1$

The <u>probability density function (PDF)</u> of a real-valued random variable x is defined by

$$p_x(\xi) = \left.\frac{dF_x(u)}{du}\right|_{u=\xi} = \lim_{\Delta\xi\to 0} \frac{F_x(\xi+\Delta\xi) - F_x(\xi)}{\Delta\xi} = \lim_{\Delta\xi\to 0} \frac{P(\xi < x \leq \xi + \Delta\xi)}{\Delta\xi}$$

(The last equation indicates why the term "density" is used.)

The PDF has the following properties:

1. $F_x(\xi) = \int_{-\infty}^{\xi} p_x(u)du$, where u is a dummy variable.

2. $\int_{-\infty}^{\infty} p_x(u)du = F_x(\infty) = 1$, where u is a dummy variable. Hence, the total area under the curve of p_x is 1.

3. Probability distributions of two variables

Often we want to estimate the probability that two (or more) random variables each satisfy a certain condition on each. Some examples of such events are: a steel rod has yield strength greater than 275 MPa and ultimate strength greater than 500 MPa, the first die roll is 3 and the second is odd, two accelerometers both read less than 0.1g simultaneously, and one accelerometer reads less than 0.1g at two different specified times.

The joint cumulative distribution function (joint CDF) of two real-valued random variables x and y is defined by $F_{x,y}(\xi,\eta) = P(x \le \xi \text{ and } y \le \eta) = P(\{(x,y) \mid x \le \xi\} \cap \{(x,y) \mid y \le \eta\})$. (Technically, the sample space in this case is the Cartesian product of the individual sample spaces.) The joint CDF has the following properties:

1. $F_{x,y}$ is non-decreasing and right-continuous in each variable.

2. $\lim_{\xi \to -\infty} F_{x,y}(\xi,\eta) = \lim_{\eta \to -\infty} F_{x,y}(\xi,\eta) = 0$

3. $\lim_{\xi \to \infty} \lim_{\eta \to \infty} F_{x,y}(\xi,\eta) = 1$

4. $\lim_{\eta \to \infty} F_{x,y}(\xi,\eta) = F_x(\xi)$ and $\lim_{\xi \to \infty} F_{x,y}(\xi,\eta) = F_y(\eta)$

The joint probability density function (joint PDF) of two real-valued random variables x and y is defined by

$$p_{x,y}(\xi,\eta) = \left. \frac{\partial^2 F_{x,y}(u,v)}{\partial u \partial v} \right|_{(u,v)=(\xi,\eta)}$$

The joint PDF has the following properties:

1. $F_{x,y}(\xi,\eta) = \int_{-\infty}^{\xi} \int_{-\infty}^{\eta} p_{x,y}(u,v)dvdu$, where u and v are dummy variables.

2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{x,y}(u,v)dvdu = F_{x,y}(\infty,\infty) = 1$, where u and v are dummy variables. Hence, the total area under the curve of $p_{x,y}$ is 1.

3. $\int_{-\infty}^{\xi}\int_{-\infty}^{\infty} p_{x,y}(u,v)dvdu = F_x(\xi)$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\eta} p_{x,y}(u,v)dvdu = F_y(\eta)$

4. $p_x(\xi) = \int_{-\infty}^{\infty} p_{x,y}(u,v)dv$ and $p_y(\eta) = \int_{-\infty}^{\infty} p_{x,y}(u,v)du$

4. Expected values

Let x be a random variable, and let f be a real or complex function. Then $y = f(x)$ is also a random variable if $P(f(x) \le \eta)$ is well defined for each η. This is the case if f is a Borel function. (A sufficient condition is that f has only a finite number of discontinuities.)

The expected value or mean of x is defined by

$$E[x] = \int_{-\infty}^{\infty} \xi p_x(\xi)d\xi$$

> If x only takes on a finite number of values, say $x_1,\ldots,x_N$ with probabilities $p_1,\ldots,p_N$, respectively, then
>
> $$p_x(\xi) = p_1\delta(\xi - x_1) + \cdots + p_N\delta(\xi - x_N)$$
>
> where $\delta(\cdot)$ is the Dirac delta function, which indicates that all the "mass" of the PDF is located at the discrete points $x_1,\ldots,x_N$. Thus,
>
> $$E[x] = \int_{-\infty}^{\infty} [\xi p_1\delta(\xi - x_1) + \cdots + \xi p_N\delta(\xi - x_N)]d\xi = x_1 p_1 + \cdots + x_N p_N$$
>
> which is the usual arithmetic mean.

$E[\cdot]$ is called the expectation operator. The mean is also written $E[x] \equiv \mu_x \equiv \bar{x}$.

If $y = f(x)$, then $E[y] = \int_{-\infty}^{\infty} \eta p_y(\eta)d\eta = \int_{-\infty}^{\infty} f(\xi)p_x(\xi)d\xi = E[f(x)]$.

Here are some common expected values:

mean square value: $E[x^2] = \int_{-\infty}^{\infty} \xi^2 p_x(\xi)d\xi$

variance: $\mathrm{var}(x) \equiv \sigma_x^2 \equiv E[(x-\bar{x})^2] = \int_{-\infty}^{\infty} (\xi - \bar{x})^2 p_x(\xi)d\xi = \overline{x^2} - \bar{x}^2$ (note that the variance is equal to the mean square value if the mean of the random variable is 0)

standard deviation: $\sigma_x = \sqrt{\sigma_x^2}$ (hence the notation σ_x^2 for variance)

covariance: $\operatorname{cov}(x,y) \equiv \sigma_{xy} \equiv E[(x-\bar{x})(y-\bar{y})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\xi-\bar{x})(\eta-\bar{y})p_{x,y}(\xi,\eta)d\eta d\xi = E[xy]-\overline{xy}$.
The covariance is a measure of the correlation between x and y. If $\sigma_{xy}>0$ then above average values of x occur roughly at the same time as above average values of y (i.e. above average values of x and y are correlated). If $\sigma_{xy}<0$ then above average values of x occur roughly at the same time as below average values of y, and vice versa.

correlation coefficient: $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$, which has the following properties

1. $-1 \le \rho_{xy} \le 1$

2. If $\rho_{xy} = \pm 1$ then $y-\bar{y} = c(x-\bar{x})$ where c is a constant and $\operatorname{sgn}(c) = \operatorname{sgn}(\rho_{xy})$. This means that x and y are linearly related.

3. If $\rho_{xy} = 0$, then x and y are uncorrelated. This means σ_x^2 cannot be reduced by subtracting a linear function of y, and σ_y^2 cannot be reduced by subtracting a linear function of x.

5. Estimating statistics from samples

In most scenarios, the underlying probability distribution (CDF or PDF) of a random variable is unknown, and so the defining formulas for mean, variance, etc. given in section 4 cannot be used directly. However, it is reasonable to assume that the distribution of a large number of samples is representative of the underlying distribution of the variable.

An estimator is a process or formula for computing an unknown (or un-computable) parameter using a finite set of samples.

Suppose a random variable x has a mean $\bar{x}$. It is impossible to compute $\bar{x}$ exactly because $p_x(\xi)$ is unknown and cannot be computed in finite time (recall Bernoulli's Law of Large Numbers). Instead, the mean can be estimated using the following formula (or estimator):

$$\hat{\bar{x}} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where the x_i are the samples of x. In general, $\hat{\bar{x}} \ne \bar{x}$, i.e. it is very unlikely that the N samples will happen to have the exact mean of the underlying distribution. However, one can prove that $E\left[\hat{\bar{x}}\right] = \bar{x}$. This indicates that this estimator is unbiased, meaning that it is equally likely to produce an estimate of the mean that is greater than the actual mean as to produce an estimate of the mean that is less than the actual mean.

One would expect that the larger N is, the better the estimate of the mean is. Indeed, one can derive the following formula:

$$\mathrm{var}\left(\hat{\bar{x}}\right) = E\left[\left(\hat{\bar{x}} - E\left[\hat{\bar{x}}\right]\right)^2\right] = \frac{\sigma_x^2}{N}$$

In words, this result states that the variance of the estimated mean equals the true variance of x divided by the number of samples. Thus, the more samples you take into consideration, the smaller the variance of the estimated mean, indicating that the estimated mean will generally be closer to the true mean. This result also indicates that if the underlying distribution has a large variance, then more samples are needed to estimate the mean to the same level of confidence.

In addition to estimating the mean of a random variable, one may also wish to estimate the variance of a random variable. Here is an unbiased estimate of the variance of x:

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \hat{\bar{x}}\right)^2$$

This estimate is unbiased because $E\left[\hat{\sigma}_x^2\right] = \sigma_x^2$. One would expect that the larger N is, the better the estimate of the variance is. Indeed, one can derive the following formula:

$$\mathrm{var}\left(\hat{\sigma}_x^2\right) = E\left[\left(\hat{\sigma}_x^2 - E\left[\hat{\sigma}_x^2\right]\right)^2\right] = \frac{1}{N}\left(\sigma_x^4 - \frac{N-3}{N-1}\left(\sigma_x^2\right)^2\right)$$

where

$$\sigma_x^4 = E\left[\left(x - \bar{x}\right)^4\right] = \int_{-\infty}^{\infty}\left(\xi - \bar{x}\right)^4 p_x(\xi)\,d\xi$$

This formula indicates that $\mathrm{var}\left(\hat{\sigma}_x^2\right) \to 0$ as $N \to \infty$. Thus, the more samples you take into consideration, the smaller the variance of the estimated variance, indicating that the estimated variance will generally be closer to the true variance.