Probability measure is a function that assigns a number between 0 and 1 to "nice" sets in the sample space, subject to the axioms of probability (WMS p.30 - below).

Suppose S is a sample space associated with an experiment. To every event A in S (A is a subset of S), we assign a number, $P(A)$, called the *probability* of A, so that the following axioms hold:

Axiom 1: $P(A) \geq 0$.

Axiom 2: $P(S) = 1$.

Axiom 3: If $A_1, A_2, A_3, \ldots$ form a sequence of pairwise mutually exclusive events in S (that is, $A_i \cap A_j = \emptyset$ if $i \neq j$), then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i).$$

A random variable (rv) is a function from the sample space to the real line.

(Cumulative) distribution function (cdf) F of a rv X is defined as $F(x) = Pr(X \leq x)$ for every $x \in \mathbb{R}$.

Properties

- F is nondecreasing
- $F(x) \in [0, 1]$ for every valid x
- $F(\infty) = 1$.

Q: Why?

\infty

A rv X is called *discrete* iff there are at most *countably* many values $x_1, x_2, \ldots$ such that $\sum_{i=1}^{n} Pr(X = x_i) = 1$.

Aside: a set is *countable* if it can be enumerated ("counted") using positive integers.

The function defined as $f(x_i) = Pr(X = x_i)$ is called the *probability mass function (pmf)* of the rv X.

Q: How does the cdf of any discrete rv look like?

Some standard discrete distributions/rvs (WMS)

Distribution	Probability Function	Mean	Variance	Moment-Generating Function
Binomial	$p(y) = \binom{n}{y} p^y (1-p)^{n-y};$ $y = 0, 1, \ldots, n$	np	$np(1-p)$	$[pe^t + (1-p)]^n$
Geometric	$p(y) = p(1-p)^{y-1};$ $y = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
Hypergeometric	$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}};$ $y = 0, 1, \ldots, n$ if $n \le r$, $y = 0, 1, \ldots, r$ if $n > r$	$\frac{nr}{N}$	$n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$	
Poisson	$p(y) = \frac{\lambda^y e^{-\lambda}}{y!};$ $y = 0, 1, 2, \ldots$	λ	λ	$\exp[\lambda(e^t - 1)]$

A rv X is called *continuous* if $Pr(X = x) = 0$ for every $x \in \mathbb{R}$.

Q: How does the cdf of any continuous rv look like?

If X is a continuous rv and has differentiable cdf F, then the *probability density function (pdf)* of X is defined as

$$f(x) = \frac{\partial}{\partial x} F(x).$$

Some standard continuous distributions/rvs (WMS)

Distribution	Probability Function	Mean	Variance	Moment-Generating Function
Uniform	$f(y) = \frac{1}{\theta_2 - \theta_1}; \theta_1 \le y \le \theta_2$	$\frac{\theta_1 + \theta_2}{2}$	$\frac{(\theta_2 - \theta_1)^2}{12}$	$\frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$
Normal	$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{1}{2\sigma^2}\right)(y-\mu)^2\right]$ $-\infty < y < +\infty$	μ	σ^2	$\exp\left(\mu t + \frac{t^2\sigma^2}{2}\right)$
Exponential	$f(y) = \frac{1}{\beta} e^{-y/\beta}; \quad \beta > 0$ $0 < y < \infty$	β	β^2	$(1 - \beta t)^{-1}$
Gamma	$f(y) = \left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right] y^{\alpha-1} e^{-y/\beta};$ $0 < y < \infty$	$\alpha\beta$	$\alpha\beta^2$	$(1 - \beta t)^{-\alpha}$

Indicator function h of a set or event A is defined as $h(x) = \mathbb{I}(x \in A)$, where

$$\mathbb{I}(x \in A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A. \end{cases}$$

Multivariate cdf and pmf

If $X_1, X_2, \ldots, X_n$ are rvs, then their *joint cdf* is defined as

$$F(x_1, x_2, \ldots, x_n) = Pr\left(\bigcap_{i=1}^{n} [X_i \le x_i]\right)$$

for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

If $X_1, X_2, \ldots, X_n$ are *discrete* rvs, then their *joint pmf* is defined as

$$f(x_1, x_2, \ldots, x_n) = Pr\left(\bigcap_{i=1}^{n} [X_i = x_i]\right)$$

for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Connections between multivariate cdf and pmf

	y_1		
y_2	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

Connection between multivariate pdf and cdf

If $X_1, X_2, \ldots, X_n$ are *continuous* rvs with differentiable joint cdf F, then their *joint pdf* is defined as

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}$$

for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Q: How to obtain the joint cdf from the joint pdf?

Exercise: Find the joint cdf of X_1 and X_2 if the joint pdf is $f(x_1, x_2) = e^{-x_1} e^{-x_2} \cdot \mathbb{I}\{x_1 > 0\} \cdot \mathbb{I}\{x_2 > 0\}$.

Marginal cdf/pmf/pdf of X_1 is just the univariate cdf/pmf/pdf of X_1. This wording is used when the cdf/pmf/pdf of X_1 is obtained from the joint cdf/pmf/pdf of $X_1, X_2, \ldots, X_n$.

Marginal cdf from the joint cdf

Exercise: in the previous exercise, find the marginal cdf of X_1 from the joint cdf of X_1 and X_2.

Marginal pmf from the joint pmf

Exercise: find the marginal pmf of Y_1 if the joint pmf of Y_1 and Y_2 is given in the table below.

	y_1		
y_2	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

Marginal pdf from the joint pdf

Exercise: find the marginal pdf of X_1 if the joint pdf of X_1 and X_2 is $f(x_1, x_2) = e^{-x_1} e^{-x_2} \cdot \mathbb{I}\{x_1 > 0\} \cdot \mathbb{I}\{x_2 > 0\}$.

Conditional pdf/pmf f_C of $X_{k+1}, X_{k+2}, \ldots X_n$ given $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$ is defined as

$$f_C(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{f_J(x_1, \ldots, x_k, \ldots, x_n)}{f_M(x_1, \ldots, x_k)},$$

assuming the denominator is positive. Here, f_J is the joint pdf/pmf of $X_1, \ldots, X_n$ and f_M is the joint pdf/pmf of $X_1, \ldots, X_k$.

If $n = 2$ and $k = 1$, this becomes $f_C(x_2|x_1) = f_J(x_1, x_2)/f_M(x_1)$.

Q: Where does this definition come from?

Statistical Independence

The rvs $X_1, X_2, \ldots, X_n$ with respective marginal cdfs $F_1, F_2, \ldots, F_n$ and joint cdf F are *mutually independent* iff

$$F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i)$$

for every $x_1, x_2, \ldots, x_n \in \mathbb{R}$. This is often referred to as *factorization criterion for independence*.

Recall

Σ-notation: $\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n$.

Π-notation: $\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdot \cdots \cdot a_n$.

Factorization criterion for independence via pdfs

Let $X_1, X_2, \ldots, X_n$ be continuous rvs with respective marginal pdfs f_i and joint pdf f. Show that

$$\begin{array}{rcll} F(x_1, x_2, \ldots, x_n) & = & \prod_{i=1}^n F_i(x_i) & \quad \text{if and only if} \\ f(x_1, x_2, \ldots, x_n) & = & \prod_{i=1}^n f_i(x_i). & \end{array}$$

Mutual independence vs pairwise independence

The rvs $X_1, X_2, \ldots, X_n$ with respective marginal cdfs $F_1, F_2, \ldots, F_n$ and joint cdf F are *pairwise independent* iff

$$Pr(X_i \leq x_i, X_j \leq x_j) = Pr(X_i \leq x_i) \cdot Pr(X_j \leq x_j)$$

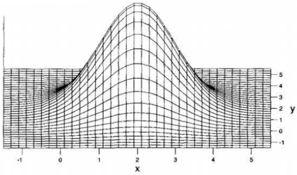
for every pair $i \neq j$ and all $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Q: Which is stronger, mutual or pairwise independence?

Bivariate normal distribution

```
p(x) = const * exp(-0.5 * (x-mu)'*inv(Sigma)*(x-mu))
```

Bivariate Normal pdfs



$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}\right].$$

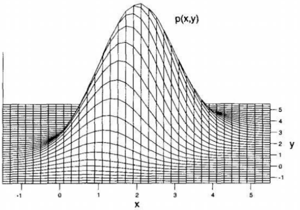


Figure B.4.1 A plot of the bivariate normal density $p(x, y)$ for $\mu_1 = \mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$, $\rho = 0$ (top) and $\rho = 0.5$ (bottom).

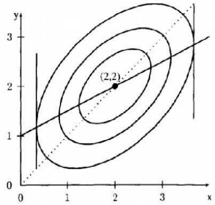


Figure B.4.2 25%, 50%, and 75% probability level-curves, the regression line (solid line), and major axis (dotted line) for the $\mathcal{N}(2, 2, 1, 1, 0.5)$ density.

Some characteristics of random variables/vectors

1. joint, marginal and conditional cdfs
2. joint, marginal and conditional pdfs or pmfs (if the rvs are continuous or discrete)
3. moments - if they are well-defined ("exist"); e.g., expectation

   functions of moments; e.g., variance, covariance, correlation
4. moment-generating function (mgf) - if it is defined in a neighborhood of zero (i.e., "exists").
5. quantile function (for a rv but not a vector); loosely, this can be thought of as the inverse of the cdf

Population moments

Let X be a rv with a pdf or pmf f and let k be a positive integer. The *k-th (population) moment* of X, denoted by $\mathrm{E}\left(X^k\right)$, is defined as follows:

If X is a discrete rv and

$$\sum_{x \in supp(f)} |x|^k f(x) < \infty, \qquad \text{then} \qquad \mathrm{E}\left(X^k\right) = \sum_{x \in supp(f)} x^k f(x).$$

If X is a continuous rv and

$$\int_{-\infty}^{\infty} |x|^k f(x) dx < \infty, \qquad \text{then} \qquad \mathrm{E}\left(X^k\right) = \int_{-\infty}^{\infty} x^k f(x) dx.$$

Important: in general, rvs need not have all or any moments to "exist" (i.e., be well-defined).

Example: non-existence of moments

Consider a *Cauchy* distribution with pdf

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2}, \quad \text{defined for all } x \in \mathbb{R}.$$

(Mathematical) expectation (aka *expected value*) of a rv X is $\mathrm{E}(X)$, the first (population) moment, provided it is well-defined.

Convention/shortcut: let's denote by $\mathrm{E}\left(|X|^k\right)$ the k-th moment of the rv $|X|$, the absolute value of X, assuming that the expectation is finite.

Proposition: Let p be a positive real number. If $\mathrm{E}(|X|^p) < \infty$, then $\mathrm{E}(|X|^q) < \infty$ for every $q \in [0, p]$.

Expectations of functions of a single or multiple rvs

Let $X_1, X_2, \ldots, X_n$ be rvs be continuous rvs with the joint pdf f. In the discrete case, replace the integrals by sums.

Let g be any function in n variables. Then $Y = g(X_1, \ldots, X_n)$ is a rv. The expectation of Y can be computed as

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

assuming that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |g(x_1, \ldots, x_n)| f(x_1, \ldots, x_n) dx_1 \cdots dx_n < \infty.$$

Some special cases:

- $g(x_1, \ldots, x_n) = x_1^k$ for a positive integer k (for moments)
- $g(x_1, \ldots, x_n) = x_1 + \cdots + x_n$ (to compute expectation of a sum)
- $g(x_1, \ldots, x_n) = x_1 x_2$ (for covariances).

Properties of Expected Values (Expectations), I

Let $X_1, X_2, \ldots, X_n$ be rvs with well-defined expectations $\mathrm{E}(X_1), \mathrm{E}(X_2), \ldots, \mathrm{E}(X_n)$. Let $b_1, b_2, \ldots, b_n$ be any constants. Assume that the rvs X_i are continuous (have pdfs). In the discrete case, replace the integrals by sums (left as an exercise).

Linearity of expectations:

$\mathrm{E}(b_1 X_1 + b_2 X_2) = b_1 \mathrm{E}(X_1) + b_2 \mathrm{E}(X_2)$

0. $\mathrm{E}(b_1) = b_1$.

1. $\mathrm{E}(b_1 X_1) = b_1 \mathrm{E}(X_1)$.

Properties of Expected Values (Expectations), II

2. $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$**.**

Properties **3 – 5** below follow from properties 0, 1 and 2.

Properties of Expected Values (Expectations), III

3. $\mathrm{E}\left(b_1 + X_1\right) = b_1 + \mathrm{E}\left(X_1\right)$.

4. $\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}\left(X_i\right)$, n is finite.

5. $\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \, \mathrm{E}\left(X_i\right)$.

Covariance

Assume that $\mathrm{E}(|X_1|)$, $\mathrm{E}(|X_2|)$ and $\mathrm{E}(|X_1X_2|)$ are all finite. Then the *covariance* between X_1 and X_2 is defined as

$$\mathrm{Cov}(X_1, X_2) = E(X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2)).$$

An alternative equivalent definition is

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2).$$

Symmetry: notice $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$.

Main property: bilinearity of covariances

Proposition: Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be rvs with well-defined covariances $Cov(X_i, Y_j)$ for every i and j. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ be any constants. Then

$$Cov(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j).$$

Variance is a special case of covariance

Variance of a rv. If $X_1 = X_2$, then
$Cov(X_1, X_2) = Cov(X_1, X_1) = \mathrm{E}\left(X_1^2\right) - \left(\mathrm{E}\left(X_1\right)\right)^2 = Var(X)$.

Q: Is there a difference between $\mathrm{E}\left(X^2\right)$ and $(\mathrm{E}\left(X\right))^2$?

Variance of a linear combination of rvs

Random variables X_1 and X_2 are said to be *uncorrelated* if
$Cov(X_1, X_2) = 0$.

Correlation

Correlation coefficient (loosely, *correlation*) between rvs X_1 and X_2 that have a finite second moment is defined as

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1) \cdot Var(X_2)}}.$$

Let $\rho = Corr(X_1, X_2)$. If $\rho = 0$, the rvs are *uncorrelated*.

Q: Why is $|\rho| \leq 1$?

```
 Var(t*X1 +X2) >= 0    ==> |rho| <= 1
i.e., we want the quadratic equation
t^2*Var(X1) +2*t*Cov(X1,X2) + Var(X2) = 0,
to have either 1 real root or no real roots;
else for some t, Var(t*X1 +X2) < 0.
Now examine the conditions on the coefficients
in order for the equationo to have <= 1 real roots.
```

Independence and correlation

Let X_1 and X_2 be rv's with the joint cdf $F_{1,2}$, joint pdf/pmf $f_{1,2}$, marginal cdf's F_1 and F_2 and marginal pdf's/pmf f_1, f_2.
Recall that *X_1 and X_2 are independent* if and only if
$F_{1,2}(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$ if and only if
$f_{1,2}(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$ for every x_1, x_2.

Independence of X_1 and X_2 does not guarantee existence of moments.
However if $\mathrm{E}\left(X_1^2\right) < \infty$ and $\mathrm{E}\left(X_2^2\right) < \infty$ and X_1 and X_2 are independent, we have

$$\mathrm{E}\left(X_1 \cdot X_2\right) = \mathrm{E}\left(X_1\right) \mathrm{E}\left(X_2\right) \Rightarrow \mathrm{Cov}\left(X_1, X_2\right) = 0,$$

i.e., X_1 and X_2 are uncorrelated.

$\mathrm{Cov}\left(X_1, X_2\right) = 0$ *does not imply that X_1 and X_2 are independent.*