Random models and probabilities

A *model* is a (simplified) map of reality.

Observing quantities from a model is achieved by performing an *experiment*.

Deterministic model:	Identical outcome if experiment is repeated
Random model:	Outcomes might be different if the experiment is repeated

Components of a random model:

outcome:	ω
sample space:	set Ω consisting of all possible values that ω can attain
event:	subset $B \subset \Omega$
probability:	probability of an event is denoted by $P(B) \in \mathbb{R}$

Basic properties of probabilities

Axioms of probability:

1. $P(B) \geq 0$ for any $B \subset \Omega$
2. $P(\Omega) = 1$
3. If B and C are disjoint events (i.e., $B \cap C = \emptyset$), then

   $P(B \cup C) = P(B) + P(C)$

Conclusions:

1. $P(B) \leq 1$ for any $B \subset \Omega$
2. $P(\bar{B}) = 1 - P(B)$ for any $B \subset \Omega \quad (\bar{B} := \Omega \setminus B)$
3. $P(B \cup C) = P(B) + P(C) - P(B \cap C)$ for all $B, C \subset \Omega$

Conditional probability

Assume C is an event with $P(C) > 0$. Then, the conditional probability of B given C is defined as

$$P(B \mid C) := \frac{P(B \cap C)}{P(C)}$$

Example: Suppose that a fair die is rolled once. Let $C :=$ number is less or equal to 3, $B :=$ number is odd. Then,

$$\begin{aligned} P(B) &= P(\{1,3,5\}) = 1/2, \\ P(C) &= P(\{1,2,3\}) = 1/2, \\ P(B \cap C) &= P(\{1,3\}) = 2/6 \end{aligned}$$

$$\Rightarrow P(B \mid C) = \frac{2/6}{1/2} = \frac{2}{3} \neq P(B)$$

Conditional probabilities: Penetrances

Suppose that the susceptibility to a certain disease depends on the genotype at one diallelic locus with alleles D and d. Then, the *penetrances* of the disease are the conditional probabilities that an individual is affected given the genotype:

$$\begin{aligned} f_{DD} &:= P(\text{“affected”} \mid DD) \\ f_{Dd} &:= P(\text{“affected”} \mid Dd) \\ f_{dd} &:= P(\text{“affected”} \mid dd) \end{aligned}$$

$f_{DD} = f_{Dd} = 1$, $f_{dd} = 0$: fully penetrant dominant mode of inheritance

$f_{DD} = 1$, $f_{Dd} = f_{dd} = 0$: fully penetrant recessive mode of inheritance

Independent events

Two events B and C are *independent* if

$$P(B \cap C) = P(B) \cdot P(C).$$

For B with $P(B) > 0$, this is equivalent to

$$P(C) = P(C \mid B) = \frac{P(B \cap C)}{P(B)},$$

i.e., the probability of C equals the conditional probability of C given B. In other words, B and C are independent if the occurrence of B does not influence the occurrence of C and vice versa.

Hardy-Weinberg equilibrium

Example (Hardy-Weinberg equilibrium):

Consider a locus with alleles $A_1, \ldots, A_s$. Let $p_j := P(A_j)$ denote the population frequency of allele A_j. Further, let g_{ij} denote the probability that a randomly chosen individual possesses genotype A_iA_j. Under the assumption that the alleles inherited by the mother (ma) and father (fa) are independent, it follows that

$$\begin{aligned} g_{ii} &= P((ma = A_i) \cap (fa = A_i)) = p_i^2, \\ g_{ij} &= P((ma = A_i) \cap (fa = A_j)) + P((ma = A_j) \cap (fa = A_i)) \\ &= 2p_ip_j. \end{aligned}$$

If the genotype probabilities satisfy these formulae, we have Hardy-Weinberg equilibrium.

Law of total probability

Let $C_1, \ldots, C_k$ be a disjoint decomposition of the sample space Ω (i.e., $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\bigcup_{j=1}^{k} C_j = \Omega$). Then, for every event B,

$$\begin{aligned} P(B) &= P(B \cap (\bigcup_{j=1}^{k} C_j)) \\ &= P(\bigcup_{j=1}^{k} (B \cap C_j)) \\ &= \sum_{j=1}^{k} P(B \cap C_j) \\ &= \sum_{j=1}^{k} P(B \mid C_j) \cdot P(C_j) \end{aligned}$$

Example: Prevalence under HWE

Consider a diallelic disease locus with alleles D and d. Let p denote the frequency of allele D. Let f_{DD}, f_{Dd}, and f_{dd} denote the penetrances of the disease. Let C_{DD}, C_{Dd}, and C_{dd} denote the event that a randomly picked individual has genotype DD, Dd, and dd, respectively. Then,

$$\begin{aligned} K_P := P(\text{“affected”}) &= P(\text{“affected”} \mid C_{DD}) \cdot P(C_{DD}) \\ &\quad + P(\text{“affected”} \mid C_{Dd}) \cdot P(C_{Dd}) \\ &\quad + P(\text{“affected”} \mid C_{dd}) \cdot P(C_{dd}) \\ &= f_{DD} \cdot p^2 + f_{Dd} \cdot 2p(1-p) + f_{dd} \cdot (1-p)^2 \end{aligned}$$

Bayes' Theorem

Let $C_1, \ldots, C_k$ be a disjoint decomposition of the sample space Ω. Then, for any event B, and for $i = 1, \ldots, k$

$$P(C_i \mid B) = \frac{P(B \mid C_i) \cdot P(C_i)}{P(B)} = \frac{P(B \mid C_i) \cdot P(C_i)}{\sum_{j=1}^{k} P(B \mid C_j) \cdot P(C_j)}$$

Example: Genotype distribution in affecteds

With the notation of the previous example, it follows that

$$
\begin{aligned}
P(C_{DD} \mid \text{“affected”}) &= \frac{f_{DD} \cdot p^2}{K_P} \\
P(C_{Dd} \mid \text{“affected”}) &= \frac{f_{Dd} \cdot 2p(1-p)}{K_P} \\
P(C_{dd} \mid \text{“affected”}) &= \frac{f_{dd} \cdot (1-p)^2}{K_P}
\end{aligned}
$$

Exercise:

Show that the genotype distribution in affected individuals is in Hardy-Weinberg equilibrium if and only if $f_{Dd} = \sqrt{f_{DD} \cdot f_{dd}}$.

Random variable

A *random variable* (r.v.) X is a function of the outcome ω in a random experiment. X represents that part of the outcome which can be observed or the part which is of current interest.

Examples:

1. Rolling a die once:

   $X :=$ number of the die ($X \in \{1, \ldots, 6\}$)

2. Rolling a die ten times:

   - $X :=$ number of throws resulting in a “6” ($X \in \{0, \ldots, 10\}$)
   - $X :=$ sum of all ten throws ($X \in \{10, \ldots, 60\}$)

3. $X :=$ Genotype of an individual at some diallelic locus

   ($X \in \{DD, Dd, dd\}$)

Discrete random variable

A random variable X is *discrete* if the set of possible values x_i is countable, i.e., can be arranged in a (possibly infinite) sequence x_1, x_2,. . . .

Suppose the random variable X is discrete. The *probability function* is defined by

$$x_i \rightarrow P(X = x_i) \quad \text{for} \quad i = 1, 2, \ldots$$

Example: Binomial distribution

A sequence of n random experiments is conducted. In each single experiment, only two different outcomes are possible: “success” or “failure”. Let $p \in [0, 1]$ denote the probability of “success” in a single experiment. Further, assume that the outcome of one single experiment is not influenced by the outcomes of all other single experiments. Let X denote the total number of successes in n experiments. Then, $X \in \{0, \ldots, n\}$ and

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for} \quad k = 0, \ldots, n.$$

This equation gives the probability function of the binomial distribution Bin(n, p).

Continuous random variable

A random variable X is *continuous* if a function $x \to f_X(x)$ exists such that

$$P(b < X \leq c) = \int_b^c f_X(x)dx \;\; \text{for all } b, c \in \mathbb{R} \;\; \text{with } b < c.$$

$f_X(x)$ is the *probability density function* of X.

Example (Uniform distribution):

Let $b, c \in \mathbb{R}$ and $b < c$. The random variable X is said to have a uniform distribution on the interval $[b, c]$ if its probability density function is given by

$$f_X(x) = \begin{cases} 0 & \text{for } x < b \\ 1/(c-b) & \text{for } b \leq x \leq c \\ 0 & \text{for } x > c \end{cases}$$

Notation: $X \sim U(b, c)$

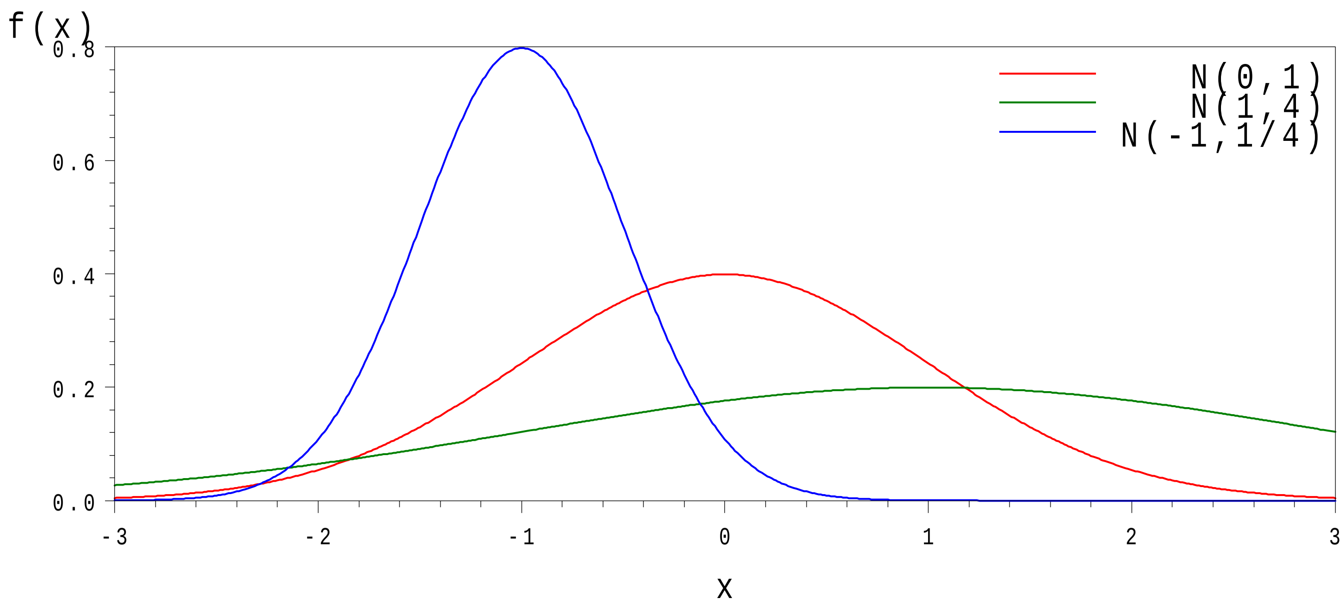
Example: Normal distribution

The random variable X is said to have a *normal distribution* if its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ for } x \in \mathbb{R}.$$

Notation: $X \sim N(\mu, \sigma^2)$

Special case: $\mu = 0$, $\sigma = 1 \Rightarrow$ *standard normal distribution* $N(0, 1)$

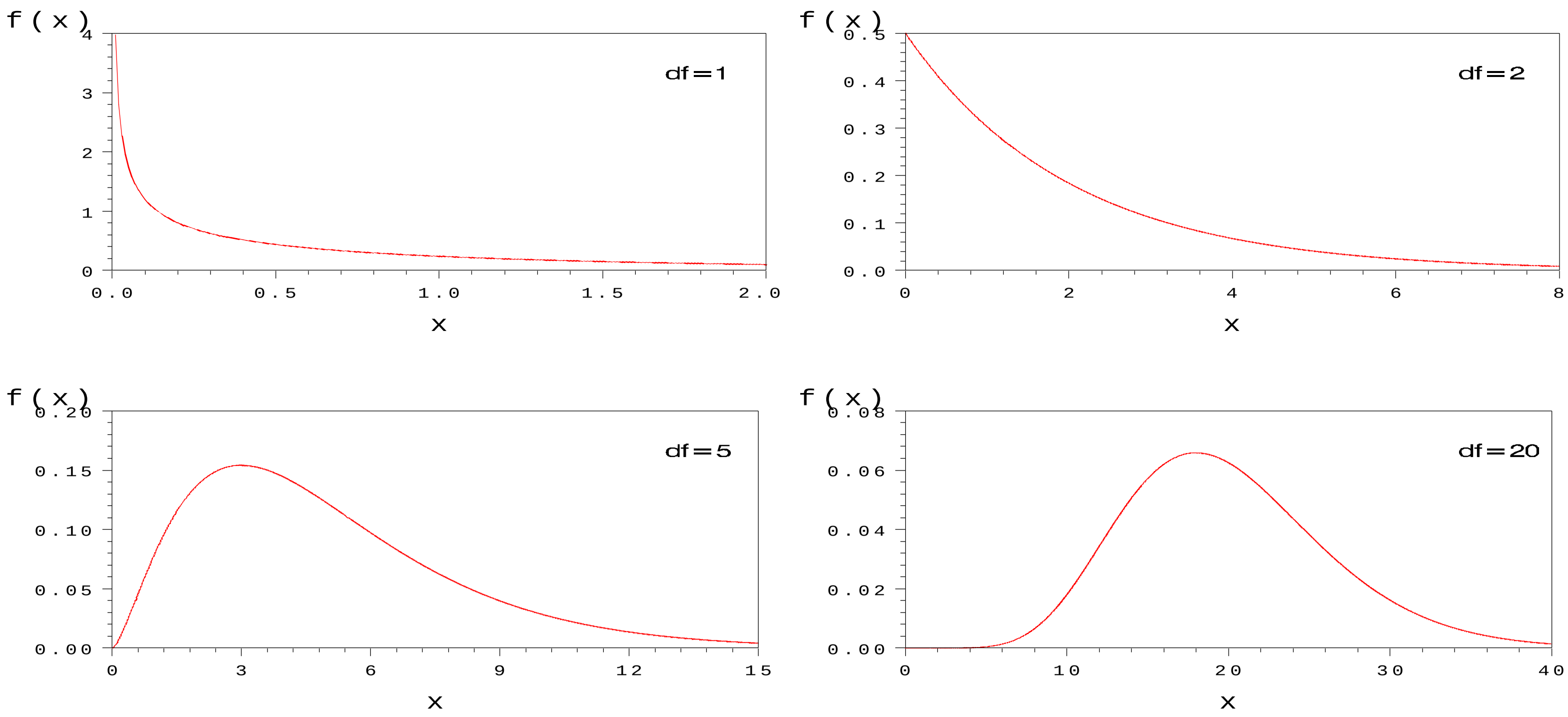


Example: χ^2 distribution

The random variable X is said to have a χ^2 *distribution* if its probability density function is given by

$$f_X(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} \, x^{n/2-1} \exp(-x/2) \text{ for } x > 0.$$

Notation: $X \sim \chi^2_n$



Distribution function

The *cumulative distribution function* (cdf) of a real-valued random variable X is defined as

$$F_X(x) = P(X \leq x), \; x \in \mathbb{R}.$$

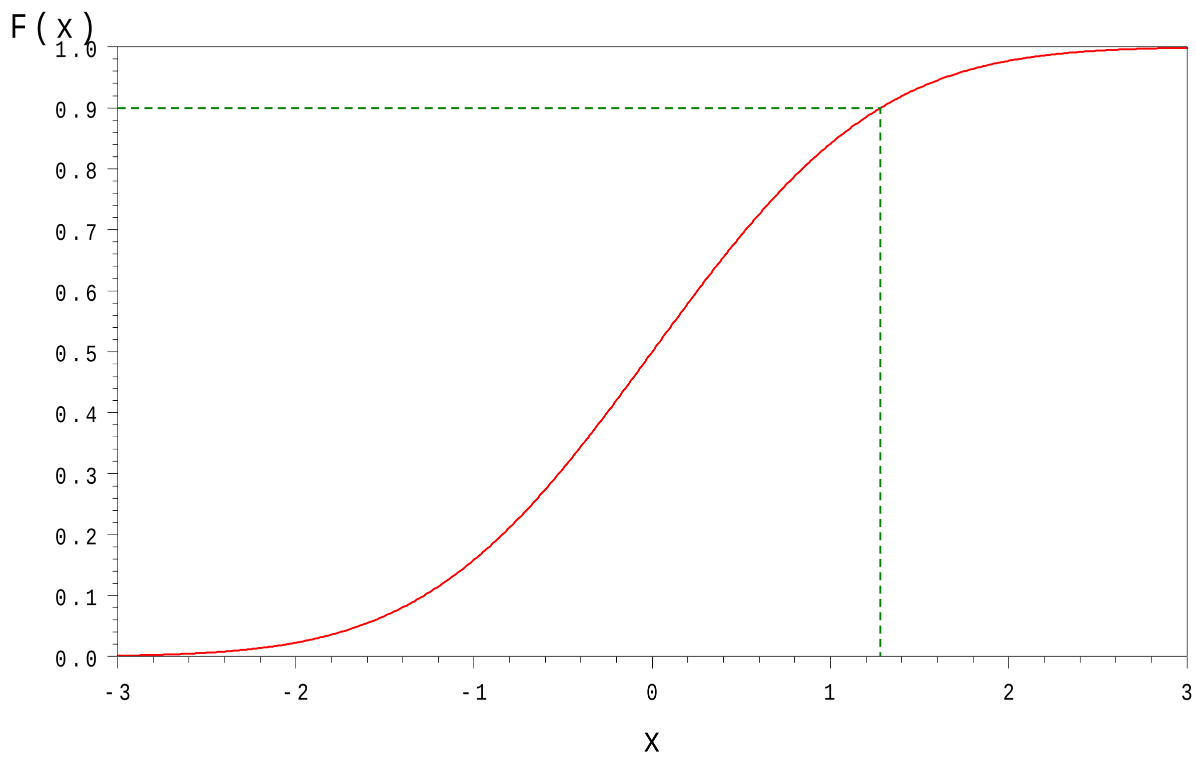
Special cases:

- discrete r.v.: $F_X(x) = \sum_{y \leq x} P(X = y)$
- continuous r.v.: $F_X(x) = \int_{-\infty}^{x} f_X(y) \, dy$

Quantile

Let $\alpha \in (0, 1)$. The α-*quantile* of the distribution of a continuous random variable X is defined as that number x_α with $F_X(x_\alpha) = \alpha$.

Example: Distribution function of $N(0, 1)$ and $x_{0.9} = 1.28155$



Conditional distribution

Suppose that X and Y are random variables, $x \in \mathbb{R}$ and $P(X = x) > 0$.

- If Y is discrete, then

$$y \to P(Y = y \mid X = x) := \frac{P((Y = y) \cap (X = x))}{P(X = x)}$$

  is the conditional probability function of Y given $X = x$.

- If Y is continuous and if there is a function $y \to f_{Y|X}(y \mid x)$ such that

$$\begin{aligned} P(b < Y \leq c \mid X = x) \quad &:= \quad \frac{P(b < Y \leq c \cap X = x)}{P(X = x)} \\ &= \quad \int_b^c f_{Y|X}(y \mid x)\, dy \text{ for all } b < c, \end{aligned}$$

  then $f_{Y|X}(y \mid x)$ is the conditional density function of Y given $X = x$.

Independent random variables

Let X and Y be two random variables with distribution function F_X and F_Y, respectively. Let

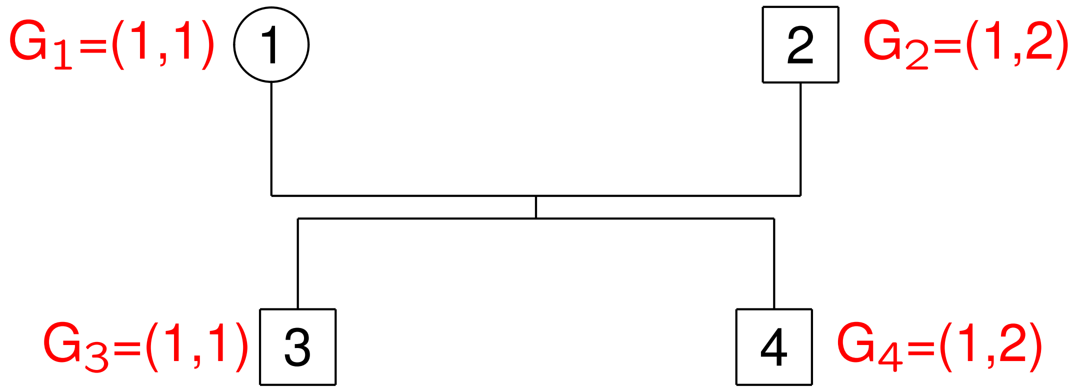
$$F_{(X,Y)}(x,y) := P((X \leq x) \cap (Y \leq y))$$

denote the distribution function of the joint distribution of (X,Y). Then, X and Y are *independent* if

$$F_{(X,Y)}(x,y) = F_X(x) \cdot F_Y(y) \text{ for all } x, y \in \mathbb{R}.$$

Example

Consider a diallelic marker locus with alleles “1” and “2”. Let p denote the population frequency of allele “1”. Now, consider the following pedigree:



What is the probability of observing these marker genotypes in a family with two parents and two children, i.e.,

$$P((G_1 = (1,1)) \cap (G_2 = (1,2)) \cap (G_3 = (1,1)) \cap (G_4 = (1,2))) = ?$$

Example (continued)

Step 1:

$$P((G_1 = (1,1)) \cap (G_2 = (1,2)) \cap (G_3 = (1,1)) \cap (G_4 = (1,2)))$$

$$= P(G_1 = (1,1)) \cdot P(G_2 = (1,2) \mid G_1 = (1,1))$$

$$\cdot P(G_3 = (1,1) \mid (G_1 = (1,1)) \cap (G_2 = (1,2)))$$

$$\cdot P(G_4 = (1,2) \mid (G_1 = (1,1)) \cap (G_2 = (1,2)) \cap (G_3 = (1,1)))$$

Step 2:

random mating $\Rightarrow P(G_2 = (1,2) \mid G_1 = (1,1)) = P(G_2 = (1,2))$

Step 3:

given parental genotypes, the genotypes of the children are independent

$$\Rightarrow P(G_4 = (1,2) \mid (G_1 = (1,1)) \cap (G_2 = (1,2)) \cap (G_3 = (1,1)))$$

$$= P(G_4 = (1,2) \mid (G_1 = (1,1)) \cap (G_2 = (1,2)))$$

Example (continued)

Step 4:

given parental genotypes, the genotype of the child is determined by

Mendelian segregation

$$\Rightarrow P(G_3 = (1,1) \mid (G_1 = (1,1)) \cap (G_2 = (1,2))) = 0.5,$$

$$P(G_4 = (1,2) \mid (G_1 = (1,1)) \cap (G_2 = (1,2))) = 0.5$$

Step 5:

assuming HWE in the parental generation

$\Rightarrow P(G_1 = (1,1)) = p^2$ and $P(G_2 = (1,2)) = 2 \cdot p \cdot (1-p)$

Step 1–5:

$$P((G_1 = (1,1)) \cap (G_2 = (1,2)) \cap (G_3 = (1,1)) \cap (G_4 = (1,2)))$$

$$= p^2 \cdot 2 \cdot p \cdot (1-p) \cdot 1/2 \cdot 1/2 = p^3 \cdot (1-p)/2$$

Expectation of a random variable

The *expected value* of a random variable X is defined as

- $\mathrm{E}(X) := \sum_i x_i \cdot P(X = x_i)$, if X is discrete
- $\mathrm{E}(X) := \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$, if X is continuous

Examples:

1. $X \sim$Bin(n, p): $\mathrm{E}(X) = \sum_{i=0}^{n} i \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} = n \cdot p$
2. $X \sim U(b, c)$: $\mathrm{E}(X) = \int_b^c x \cdot \frac{1}{c-b}\, dx = \frac{1}{c-b} \cdot \left[\frac{x^2}{2}\right]_{x=b}^{x=c} = \frac{c+b}{2}$
3. $X \sim N(\mu, \sigma^2)$: $\mathrm{E}(X) = \mu$

Variance of a random variable

The *variance* of a random variable X is $\mathrm{Var}(X) := \mathrm{E}\left[(X - \mathrm{E}(X))^2\right]$, i.e.,

- $\mathrm{Var}(X) := \sum_i (x_i - \mathrm{E}(X))^2 \cdot P(X = x_i)$, if X is discrete
- $\mathrm{Var}(X) := \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 \cdot f_X(x)\, dx$, if X is continuous

Examples:

1. $X \sim \mathrm{Bin}(n, p)$:

   $$\mathrm{Var}(X) = \sum_{i=0}^{n} (i - n \cdot p)^2 \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} = n \cdot p \cdot (1-p)$$

2. $X \sim U(b, c)$: $\quad \mathrm{Var}(X) = \int_b^c \left(x - \frac{b+c}{2}\right)^2 \cdot \frac{1}{c-b}\, dx = \frac{(c-b)^2}{12}$

3. $X \sim N(\mu, \sigma^2)$: $\mathrm{Var}(X) = \sigma^2$

Covariance and correlation coefficient

The *covariance* between two random variables X and Y is given by

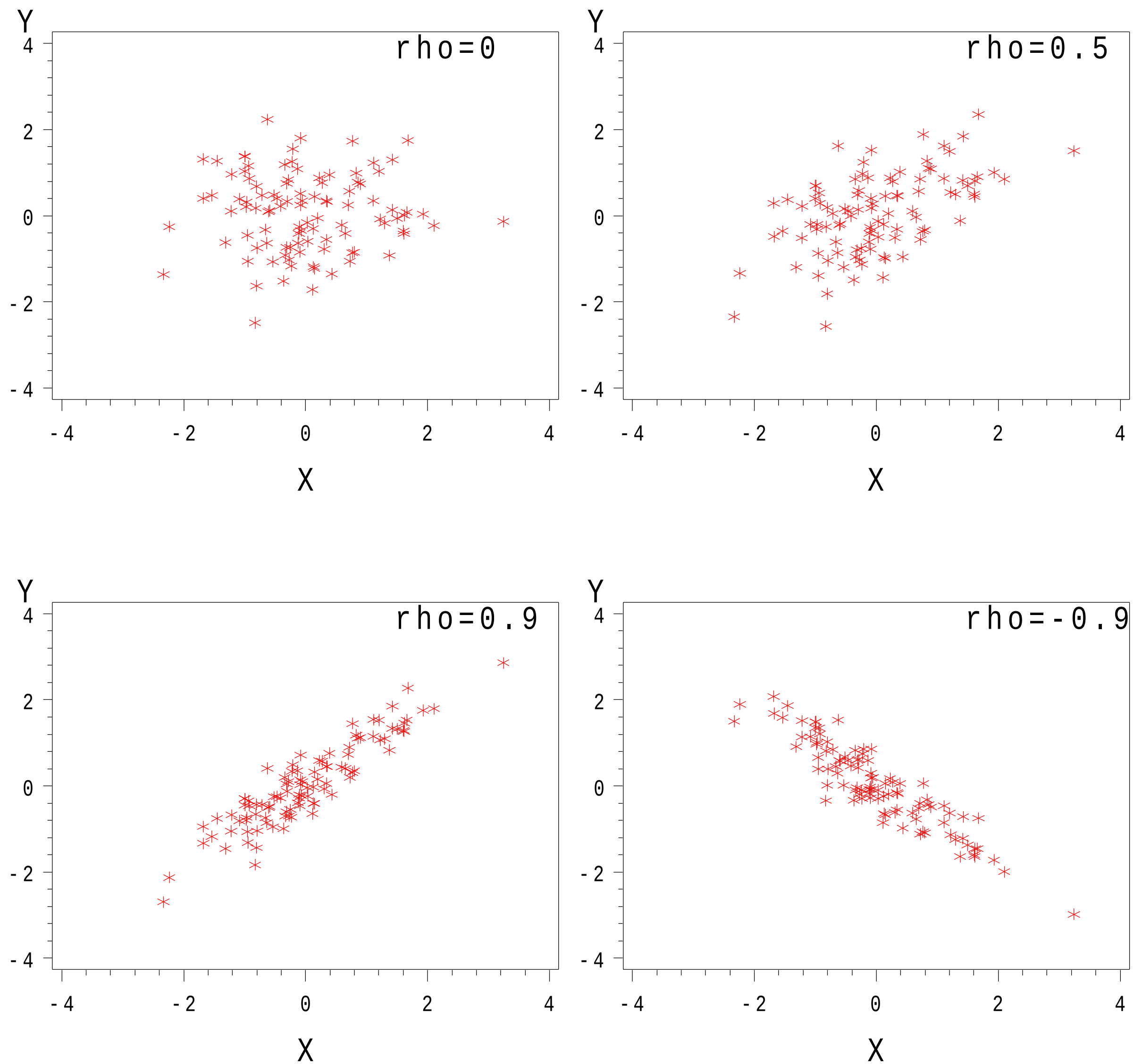
$$\mathrm{Cov}(X, Y) := \mathrm{E}\left[(X - \mathrm{E}(X)) \cdot (Y - \mathrm{E}(Y))\right],$$

and the *correlation coefficient* between X and Y is defined as

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}.$$

Example: Covariance and correlation coefficient

Plots of 100 pairs (X, Y), when both X and $Y \sim N(0, 1)$



Independence and correlation

Two random variables X and Y are said to be *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$.

If X and Y are independent, then X and Y are uncorrelated. The converse is generally not true, i.e.,

$$X \text{ and } Y \text{ uncorrelated} \not\Rightarrow X \text{ and } Y \text{ independent.}$$

However, if X and Y both have a normal distribution and are uncorrelated, then X and Y are independent.

Example: Dependence and uncorrelation

Exercise:

Assume that $P((X=-1)\cap(Y=0)) = P((X=0)\cap(Y=1)) = P((X=0)\cap(Y=-1)) = P((X=1)\cap(Y=0)) = 1/4$.

Show that (i) X and Y are uncorrelated, but (ii) X and Y are dependent.

(Hint: For (ii), show that $F_{(X,Y)}(-1,-1) \neq F_X(-1)\cdot F_Y(-1)$.)

Properties of $\mathrm{E}(X)$, $\mathrm{Var}(X)$, $\mathrm{Cov}(X)$, and $\rho(X)$

Let X and Y be random variables and $b, c, d, e \in \mathbb{R}$. Then,

- $\mathrm{E}(bX + c) = b \cdot \mathrm{E}(X) + c$
- $\mathrm{Var}(bX + c) = b^2 \cdot \mathrm{Var}(X)$
- $\mathrm{Cov}(bX + c, dY + e) = b \cdot d \cdot \mathrm{Cov}(X, Y)$
- if $b, d > 0$, then $\rho(bX + c, dY + e) = \rho(X, Y)$
- $\rho(X, Y) \in [-1, 1]$ and
  - ⋆ $\rho(X, Y) = 1$ if and only if $Y = bX + c$ for some $b > 0$
  - ⋆ $\rho(X, Y) = -1$ if and only if $Y = bX + c$ for some $b < 0$

Example: Standardizing a random variable

Let X be a random variable with $\mathrm{Var}(X) > 0$. Let

$$Z := \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var} X}}.$$

Then, Z is called the *standardized random variable* corresponding to X.

Since

$$Z = \underbrace{\frac{1}{\sqrt{\mathrm{Var} X}}}_{=:b} \cdot X - \underbrace{\frac{\mathrm{E}(X)}{\sqrt{\mathrm{Var} X}}}_{=:c},$$

it follows that $\mathrm{E}(Z) = b \cdot \mathrm{E}(X) - c = 0$ and $\mathrm{Var}(Z) = b^2 \cdot \mathrm{Var}(X) = 1$.

Expected value and variance for sums of r.v.'s

Let X, Y, Z, and W be random variables. Then,

$$\begin{aligned} \mathrm{E}(X+Y) &= \mathrm{E}(X)+\mathrm{E}(Y), \\ \mathrm{Var}(X+Y) &= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\cdot\mathrm{Cov}(X,Y), \\ \mathrm{Cov}(X+Y,Z+W) &= \mathrm{Cov}(X,Z)+\mathrm{Cov}(X,W)+\mathrm{Cov}(Y,Z)+\mathrm{Cov}(Y,W) \end{aligned}$$

If X and Y are uncorrelated, then

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$$

Example: Two rules for variance and covariance

Let X and Y be random variables. Then,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{E}\left[X^2 - 2 \cdot X \cdot \mathrm{E}(X) + [\mathrm{E}(X)]^2\right] \\
&= \mathrm{E}(X^2) - \mathrm{E}[2 \cdot X \cdot \mathrm{E}(X)] + [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - 2 \cdot [\mathrm{E}(X)]^2 + [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}\left[(X - \mathrm{E}(X)) \cdot (Y - \mathrm{E}(Y))\right] \\
&= \mathrm{E}\left[X \cdot Y - X \cdot \mathrm{E}(Y) - \mathrm{E}(X) \cdot Y + \mathrm{E}(X) \cdot \mathrm{E}(Y)\right] \\
&= \mathrm{E}(X \cdot Y) - \mathrm{E}(X) \cdot \mathrm{E}(Y)
\end{aligned}$$