CEE 697M: Big Data and Machine Learning for Engineers
Lecture 1b: Probability

Jimi Oke

UMass Amherst
College of Engineering

Feb 10, 2023

Outline

Random variables

A random variable is a function that uniquely maps events in a sample space to the set of real numbers.

A random variable X may be:

- *Discrete*
- *Continuous*
- *Mixed* (probability defined over both discrete and range of continuous values)

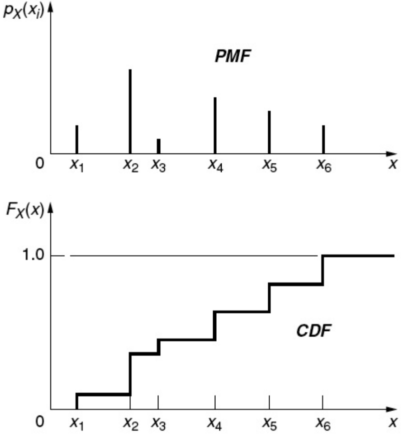
Probability mass function (PMF)

The PMF is given by

$$p_X(x_i) \equiv P(X = x_i) \quad \forall x \qquad (1)$$

CDF of discrete random variable

$$F_X(x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p_X(x_i)$$



The probability masses in a PMF sum up to 1.

Probability density function (PDF)

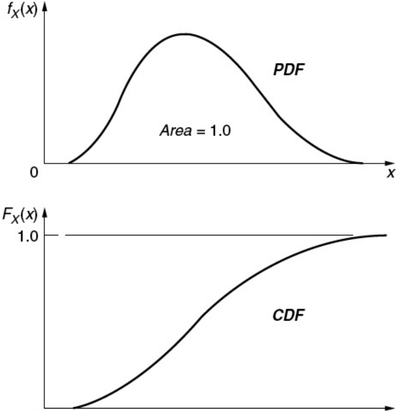
The PDF is denoted $f_X(x)$ such that the probability of X in the interval $(a, b]$ is:

$$P(a < X \leq b) = \int_a^b f_X(x)dx \quad (2)$$

CDF of continuous random variable

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} f_X(\tau)d\tau \end{aligned}$$

It follows that the PDF is the derivative of the CDF:



The total area under a PDF is 1.

Central values

These include the mean, median and mode.

- Mean: weighted average (by probability of occurence) or expected value

$$\mathbb{E}(X) = \mu_X = \sum_i x_i p_X(x_i) \quad \text{discrete case} \tag{4}$$

$$\mathbb{E}(X) = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx \quad \text{continuous case} \tag{5}$$

Generalized expectation

The mathematical expectation can be defined for a function g of random variable X:

$$\mathbb{E}[g(X)] = \sum_i g(x_i) p_X(x_i) \quad \text{discrete case} \tag{6}$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad \text{continuous case} \tag{7}$$

Measures of dispersion

Variance

In discrete case:

$$\mathbb{V}(X) = \sum_i (x_i - \mu_X)^2 p_X(x_i) \tag{8}$$

In continuous case:

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx \tag{9}$$

Expanding both equations results in:

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu_X^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{10}$$

Measures of dispersion (cont.)

Standard deviation

The standard deviation is convenient as it has the same unit as the random variable:

$$\sigma_X = \sqrt{\mathbb{V}(X)} \tag{11}$$

Coefficient of variation

The COV gives the deviation relative to the mean. It is unitless.

$$\delta_X = \frac{\sigma_X}{\mu_X} \tag{12}$$

Mean of a linear function

For a continuous random variable X, the mean is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx \tag{13}$$

Now, given that $Z = aX + bY$, then the mean of Z is

$$\begin{aligned}
\mathbb{E}(Z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (aX + bY) f_{X,Y} dxdy \\
&= a\int_{-\infty}^{\infty} x f_X(x) dx + b\int_{-\infty}^{\infty} y f_Y(y) dy \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y)
\end{aligned}$$

Variance of a linear function

We also recall the variance of an r.v. X:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] \tag{14}$$

Thus, for $Z = aX + bY$:

$$\begin{aligned}
\mathbb{V}(Z) &= \mathbb{E}[((aX + bY) - (a\mu_X + b\mu_Y))^2] \\
&= \mathbb{E}[(a(X - \mu_X) + b(Y - \mu_Y))^2] \\
&= \mathbb{E}[a^2(X - \mu_X)^2 + 2ab(X - \mu_X)(Y - \mu_Y) + b^2(Y - \mu_Y)] \\
&= a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2abCov(X, Y)
\end{aligned}$$

Moments

The *m*-th order moment of a distribution is given by:

$$\mathbb{E}(X^m) = \begin{cases} \sum_i x_i^m \cdot p_X(x_i) & \text{(discrete)} \\ \int x^m \cdot f_X(x)dx & \text{(continuous)} \end{cases} \tag{15}$$

- *m*-th central moment: $\mathbb{E}[(X - \mu_X)^m]$
- Normalized *m*-th central moment: $\left(\frac{\mathbb{E}[(X-\mu_X)^m]}{\sigma^m}\right)$

Examples

- **Mean**: first moment, $\mathbb{E}(X)$
- **Variance**: second central moment, $\mathbb{E}[(X - \mu_X)^2]$
- **Skewness**: normalized third central moment, $\left(\frac{\mathbb{E}[(X-\mu_X)^3]}{\sigma^3}\right)$

Bernoulli distribution

Let X be an event with only two outcomes $\{1{,}0\}$. And let the probability of the event be given by:

$$p(X) = \theta, \quad 0 \leq \theta \leq 1$$

And $p(X = 1) = \theta$ and $p(X = 0) = 1 - \theta$. X is said to be Bernoulli distributed:

$$X \sim \mathrm{Ber}(\theta) \tag{16}$$

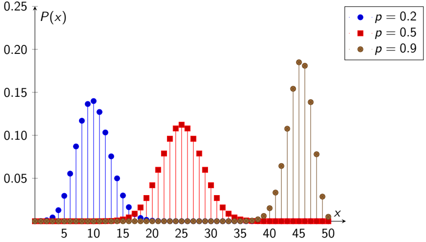
The PMF is then given by:

$$\mathrm{Ber}(x|\theta) := \theta^{x}(1-\theta)^{1-x} \tag{17}$$

Binomial distribution

Given a Bernoulli sequence with X random number of occurrences of an event, N trials and θ the probability of occurrence of each event:

- $X \sim \mathrm{Bin}(N, \theta)$
- PMF: $P(X = x) := \mathrm{Bin}(x|N, \theta) := \binom{N}{x} p^x (1-\theta)^{N-x}, \quad x = 0, 1, 2, \ldots, N$
- CDF: $F_X(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{N}{k} \theta^k (1-\theta)^{N-k}$
- Mean: $\mathbb{E}(X) = N\theta$
- Variance: $\mathbb{V}(X) = N\theta(1-\theta)$



Bernoulli, binomial, categorical and multinomial

- The Bernoulli distribution is a special case of the binomial distribution with $N = 1$
- The categorical distribution is generalization of the Bernoulli to more than two outcomes for a single trial (e.g. set of labels $x \in \{1, ..., C\}, C > 2$):

$$\mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta}) := \prod_{c=1}^{C} \theta_c^{x_c} \tag{18}$$

  where $\boldsymbol{x}$ is a one-hot vector (e.g. (1,0,0,0) for class 1 of four classes)
- The multinomial distribution generalizes the categorical distribution for multiple trials:

$$\mathscr{M}(\boldsymbol{x}|N, \boldsymbol{\theta}) := \binom{N}{N_1 \dots N_C} \prod_{c=1}^{C} \theta_c^{N_c} \tag{19}$$

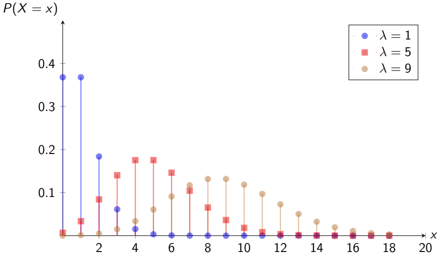
Poisson distribution

- The Poisson distribution is used to model the probability that a number of independent events occur within a fixed time interval (or within a finite space)
- Such events are described as Poisson processes
- The PMF of a Poisson random variable with **rate parameter λ** is given by:

$$P(X = x) := \mathrm{Poiss}(x|\lambda) := \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \geq 0 \tag{20}$$

- The mean and variance of a Poisson random variable are equal:

$$\mathbb{E}(X) = \mathbb{V}(X) = \lambda \tag{21}$$



Gaussian distribution

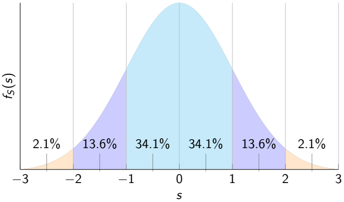
The PDF of a Gaussian (normal) distribution $X \sim \mathcal{N}(\mu, sigma^2)$ is given by:

$$\mathcal{N}(x|\mu,\sigma^2) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{22}$$

where μ is the mean and σ^2 is the variance.

$$P(a < X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{23}$$

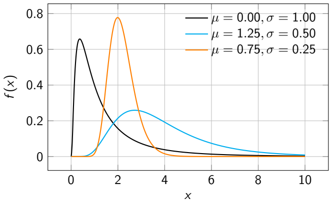
where Φ is the CDF of the standard normal distribution ($N(0,1)$).



Lognormal distribution

A random variable X that is lognormally distributed with the parameters μ and σ^2 (denoted $X \sim \mathscr{L}\mathscr{N}(\mu, \sigma^2)$) has the PDF:

$$\mathscr{L}\mathscr{N}(x|\mu, \sigma^2) = \frac{1}{(\sigma x)\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x) - \mu}{\sigma}\right)^2\right] \quad x \geq 0 \tag{24}$$

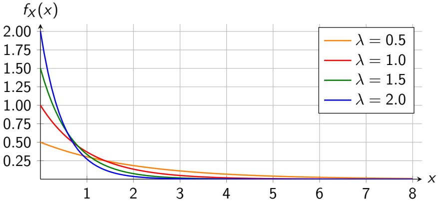


CDF: $F_X(x) = P(X \leq x) = \Phi((\ln(x) - \mu)/\sigma)$
Mean: $\mathbb{E}(X) = e^{\left(\mu + \frac{1}{2}\sigma^2\right)}$
Variance: $\mathbb{V}(X) = (e^{\sigma^2} - 1)e^{(2\mu + \sigma^2)}$

Exponential distribution

A random variable X exponentially distributed with parameter λ has the PDF:

$$\mathrm{Exp}(x|\lambda) = \lambda e^{-\lambda x} \qquad x > 0 \tag{25}$$



CDF:

$$F_X(x) \quad = \quad P(X \leq x) = 1 - e^{-\lambda x}, \qquad x > 0 \tag{26}$$

Mean:

$$\mathbb{E}(X) = 1/\lambda \tag{27}$$

Variance:

$$\mathbb{V}(X) = 1/\lambda^2 \tag{28}$$

Covariance and correlation

Recall that the variance of an r.v. X is given by:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \tag{29}$$

Then given two r.v.'s X and Y, the *covariance* measures the strength of the linear relationship between them.

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{30}$$

Correlation coefficient

This is the normalized covariance

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}} \tag{31}$$

Useful results

If $\boldsymbol{x}$ is a D-dimensional r.v., then its **covariance matrix** is defined as:

$$\mathrm{Cov}[\boldsymbol{x}] := \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T\right] := \boldsymbol{\Sigma} \tag{32}$$

This implies that:

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{33}$$

The covariance of a linear transformation is defined as:

$$\mathrm{Cov}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathrm{Cov}[\boldsymbol{x}]\boldsymbol{A}^T \tag{34}$$

The cross-covariance between two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is:

$$\mathrm{Cov}[\boldsymbol{x}\boldsymbol{y}] = \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T\right] \tag{35}$$

Uncorrelated does not imply independent

The following scatterplots indicate pairs of variables with various correlation values.

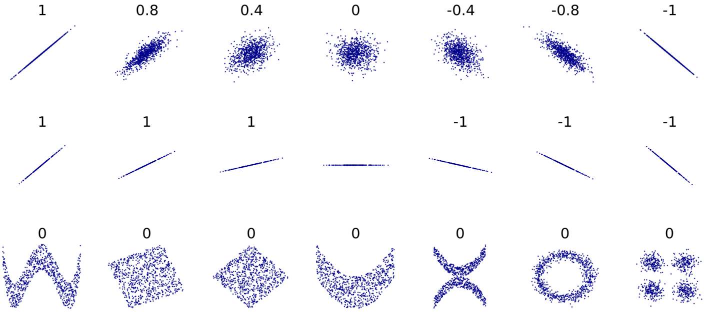


Figure: Source:
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Note that some with 0 correlation still have functional dependence (but non-linear).

Correlation does not imply causation

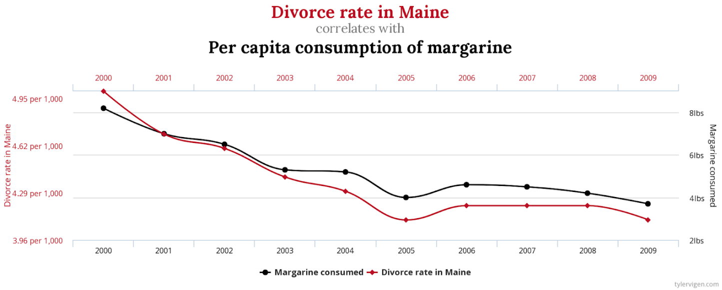


Figure: Source: `https://sitn.hms.harvard.edu/flash/2021/when-correlation-does-not-imply-causation-why-your-gut-microbes-may-not-yet-be-a-silver-bullet-to-all-your-problems/`

Visit `https://www.tylervigen.com/spurious-correlations` for more examples.

Simpson's paradox

Trends appearing in different groups may be reversed or disappear when groups are combined

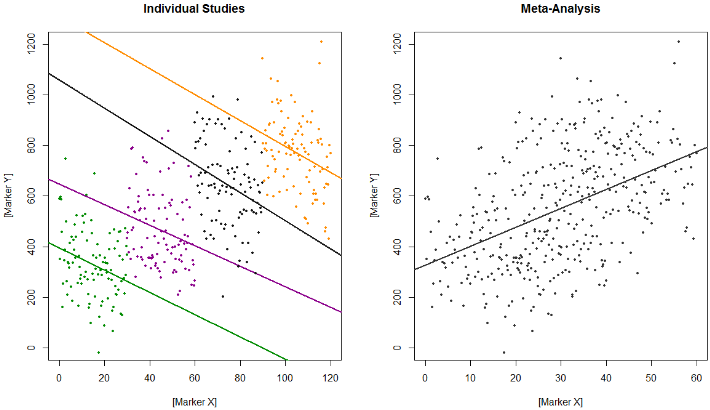


Figure: Source: `https://rinterested.github.io/statistics/simpsons_paradox.html`

Joint distributions

Given two random variables X and Y:

Discrete case

The joint PMF is:

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j) \tag{36}$$

The CDF is:

$$F_{X,Y}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{X,Y}(x_i, y_j) \tag{37}$$

Continuous case

The joint probability is given by:

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y) dy dx \tag{38}$$

Conditional distributions of continuous random variables

Recall the definition of conditional probability (multiplication rule):

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{39}$$

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{40}$$

Similarly, for two continuous r.v.'s, the conditional PDF of X given Y is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{41}$$

Joint PDF and CDF of two variables

The joint PDF is given by:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \tag{42}$$

While the joint CDF is given by:

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)dydx \tag{43}$$

Marginal distributions of continuous random variables

Recall the theorem of total probability:

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i) \tag{44}$$

Similarly, the marginal PDFs from a joint distribution of two continuous r.v.'s X and Y is given as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \tag{45}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \tag{46}$$

Multivariate normal distribution (MVN)

The MVN PDF is given by:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{47}$$

where:

- $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}] \in \mathbb{R}^D$ is the mean vector
- $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{x}]$ is the $D \times D$ covariance matrix:

$$\mathrm{Cov}[\boldsymbol{x}] := \mathbb{E}\left[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T\right] \tag{48}$$

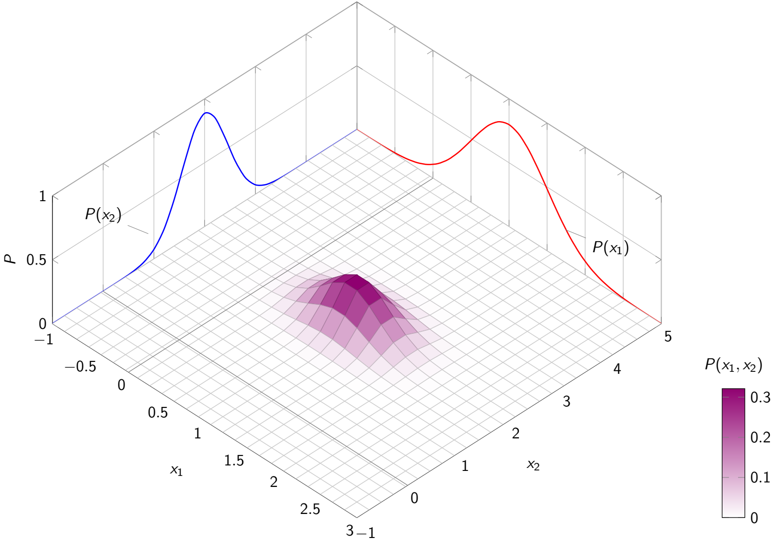
In 2D:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{49}$$

where ρ is the correlation coefficient.

Bivariate MVN

Marginal distributions: $P(x_1)$ and $P(x_2)$; Joint distribution: $P(x_1, x_2)$.



Reading

- PMLI 1, 2, 3
- PMLCE 1, 3, 4