S21CSE471: Statistical Methods in AI **Spring 2021**

Basics of Machine Learning and Statistics

Prepared by: 2020201086 2020201091 2020701030 2020201079

1 Central Tendency

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations and is called the central tendency.

1.1 Measures of Central Tendency

1.1.1 Mean

The mean of a variable is defined as the sum of the observations divided by the number of observations.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1}$$

1.1.2 Median

The median is the middle most item that divides the group into two equal parts, one part comprising all values greater, and the other, all values less than that item.

1.1.3 Mode

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it. It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

2 Random Variable

Given a random experiment with sample space S, a random variable X is a set function that assigns one and only one real number to each element s that belongs in the sample space S. The set of all possible values of the random variable X, denoted x, is called the support, or space, of X.

2.1 Discrete Random Variable

A random variable X is discrete if there are a finite number of possible outcomes of X, or there are a countably infinite number of possible outcomes of X.

2.2 Continuous Random Variable

A random variable X is continuous if possible values comprise either a single interval on the number line or a union of disjoint intervals

3 Expectation

The expectation of a random variable is its average value, where each value is weighted according to the probability that it comes up. The expectation is also called the expected value or the mean of the random variable and is denoted by μ.
Let X be a discrete random variable that takes values x_1, x_2, . . . , x_n with probabilities $p(x_1)$, $p(x_2)$, ..., $p(x_n)$. The expected value of X is defined by,

$$E(X) = \sum_{i=1}^{n} x_i p(x_i) \tag{2}$$

Let X be a continuous random variable with range $[a, b]$ and probability density function $f(x)$. The expected value of X is defined by,

$$E(X) = \int_a^b x f(x) dx \tag{3}$$

3.1 Properties of Expectation

- If X and Y are random variables on a sample space Ω then, $E(X + Y) = E(X) + E(Y)$
- If a and b are constants then, $E(aX + b) = aE(X) + b$

4 Variance

The variance is a measure of how spread out the distribution of a random variable is. Let X be a random variable with mean μ. The variance of X is

$$Var(X) = E[(X - \mu)^2] \tag{4}$$

4.1 Properties of Variance

- If X and Y are independent then, $Var(X + Y) = Var(X) + Var(Y)$
- For constants a and b, $Var(aX + b) = a^2 Var(X)$
- $Var(X) = E(X^2)$-$E(X)^2 = E(X^2)$-μ^2

5 Covariance

The covariance is a measure of the joint variability of two random variables.If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, the covariance is negative.
The covariance of X and Y is defined as,

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{5}$$

5.1 Properties of Covariance

- $Cov(X,X) = Var(X)$
- If X and Y are independent then, $Cov(X,Y) = 0$
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(aX,Y) = aCov(X,Y)$
- $Cov(X{+}c,Y) = Cov(X,Y)$
- $Cov(X{+}Y,Z) = Cov(X,Z){+}Cov(Y,Z)$

6 Standard Deviation

The standard deviation is the square root of the variance. It is denoted by σ.

$$\sigma(X) = \sqrt{Var(X)} \tag{6}$$

7 Uniform Distribution

A random variable that takes on each possible value with the same probability is said to be uniform. If the sample space is $1, 2, 3, ..., n$, then the uniform distribution has a PDF of the form

$$f_n : 1, 2, ..., n \rightarrow [0, 1] \tag{7}$$

where,

$$f_n(k) = 1/n \tag{8}$$

8 Special Distributions

From a practical perspective, we can think of a distribution as a function that describes the relationship between observations in a sample space. For example, we may be interested in the age of humans, with individual ages representing observations in the domain, and ages 0 to 125 the extent of the sample space. The distribution is a mathematical function that describes the relationship of observations of different heights.

8.1 Geometric Distribution

A random variable X is said to be a geometric random variable with parameter p, shown as $X \sim Geometric(p)$, if its PMF is given by

$$P_X(k) = \begin{cases} p(1-p)^{k-1} & \text{for } k = 1, 2, 3, ... \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$.

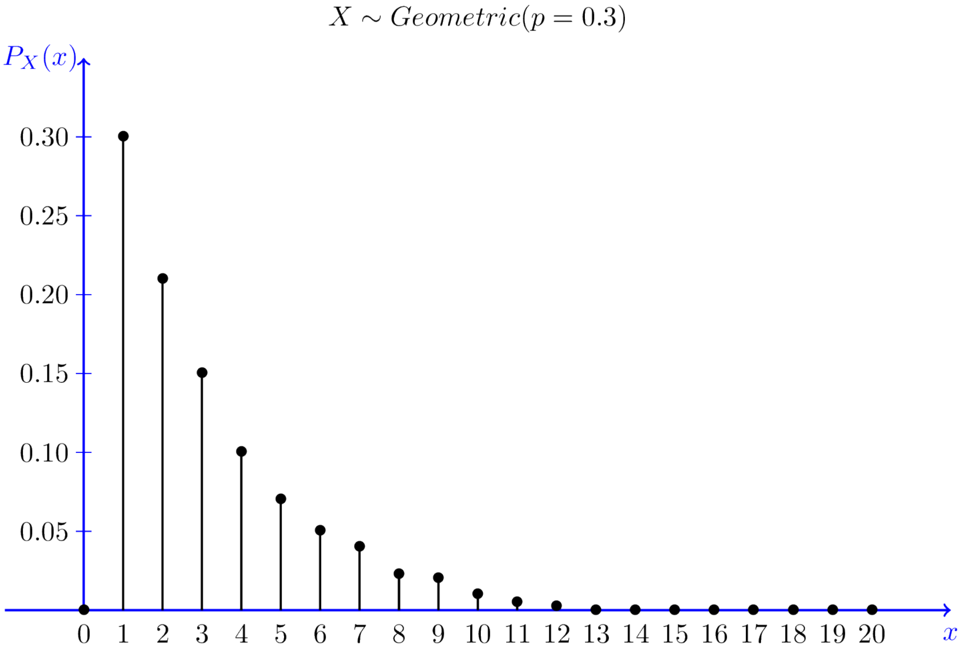


Figure 1: PMF for $Geometric(0.3)$ Random Variable.

8.2 Poisson Distribution

Let X be a discrete random variable that can take on the values $R_X = 0, 1, 2, ...$ such that the probability function of X is given by:

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where λ is a given positive constant. This distribution is called the Poisson distribution (after S. D. Poisson, who discovered it in the early part of the nineteenth century), and a random variable having this distribution is said to be Poisson distributed.

Figures below show $Poisson(\lambda)$ PMF for $\lambda = 1, \lambda = 5, and\ \lambda = 10$ respectively.

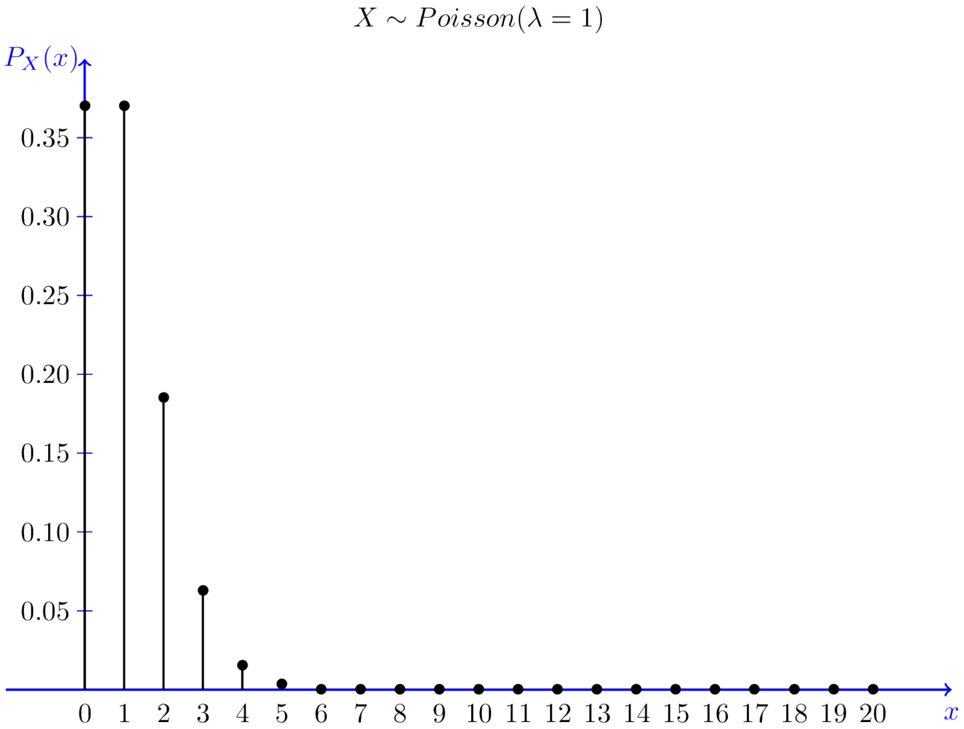


Figure 2: PMF of a $Poisson(1)$ random variable.

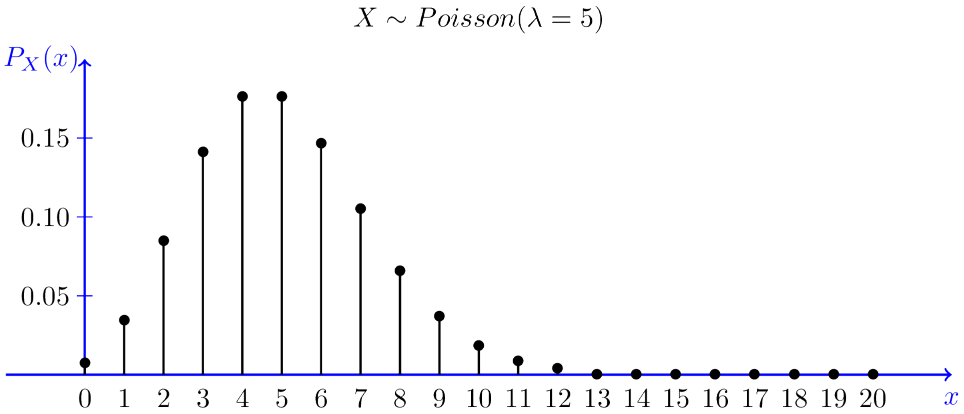


Figure 3: PMF of a $Poisson(5)$ random variable.

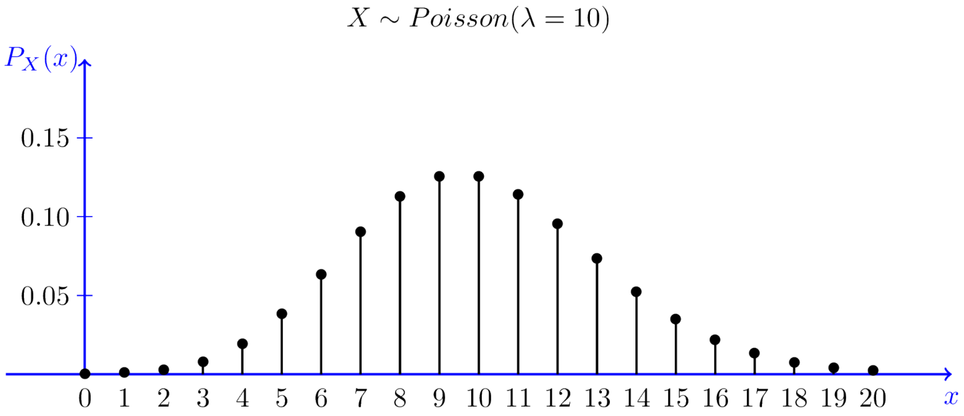


Figure 4: PMF of a $Poisson(10)$ random variable.

Poisson as an approximation for binomial

The Poisson distribution can be viewed as the limit of binomial distribution. Suppose $X \sim Binomial(n, p)$ where n is very large and p is very small. In particular, assume that $\lambda = np$ is a positive constant. We show that the PMF of X can be approximated by the PMF of a $Poisson(\lambda)$ random variable. The importance of this is that Poisson PMF is much easier to compute than the binomial. Let us state this as a theorem.

Theorem :

Let $X \sim Binomial(n, p = \frac{\lambda}{n})$ where $\lambda > 0$ is fixed. Then for any $k \in \{0, 1, 2, ...\}$, we have

$$\lim_{n\to\infty} P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Proof : We have,

$$\lim_{n\to\infty} P_X(k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \lambda^k \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{1}{n^k}\right) \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} \left(\left[\frac{n(n-1)(n-2)...(n-k+1)}{n^k}\right] \left[\left(1 - \frac{\lambda}{n}\right)^n\right] \left[\left(1 - \frac{\lambda}{n}\right)^{-k}\right] \right)$$

Note that, for a fixed k, we have

$$\lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-k+1)}{n^k} = 1$$

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1$$

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

Thus, we conclude

$$\lim_{n\to\infty} P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

8.3 Gaussian Distribution

The normal distribution is by far the most important probability distribution. One of the main reasons for that is the Central Limit Theorem. The Central Limit Theorem states that if you add a large number of random variables, the distribution of the sum will be approximately normal under certain conditions. The importance of this result comes from the fact that many random variables in real life can be expressed as the sum of a large number of random variables and, by the Central Limit Theorem, we can argue that distribution of the sum should be normal.

A continuous random variable Z is said to be a standard normal (standard Gaussian) random variable, shown as $Z \sim N(0,1)$, if its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \qquad \text{for all } z \in \mathbb{R}.$$

The $\frac{1}{\sqrt{2\pi}}$ is there to make sure that the area under the PDF is equal to one.

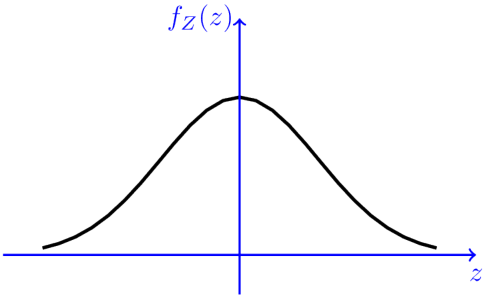


Figure 5: PDF of a standard Gaussian random variable.

Let us find the mean and variance of the standard gaussian distribution. Consider a function $g(u) : \mathbb{R} \to \mathbb{R}$. If $g(u)$ is an odd function, i.e., $g(-u) = -g(u)$, and $|\int_0^\infty g(u)du| < \infty$, then

$$\int_{-\infty}^{\infty} g(u)du = 0$$

CDF of standard gaussian

To find the CDF of the standard normal distribution, we need to integrate the PDF function. In particular, we have

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left\{-\frac{u^2}{2}\right\} du$$

This integral does not have a closed form solution. Nevertheless, because of the importance of the normal distribution, the values of $F_Z(z)$ have been tabulated and many calculators and software packages have this function. We usually denote the standard gaussian CDF by ϕ.

The CDF of the standard normal distribution is denoted by the ϕ function:

$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du$$

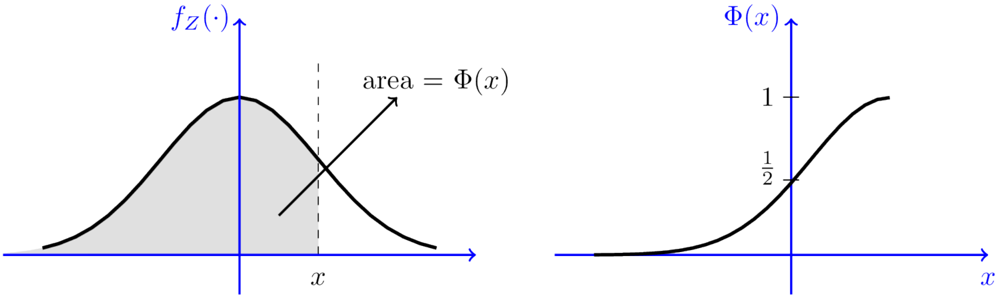


Figure 6: The ϕ function(CDF of standard gaussian).

If X is a normal random variable with mean μ and variance σ^2, i.e, $X \sim N(\mu, \sigma^2)$, then

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

$$F_X(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

$$P(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

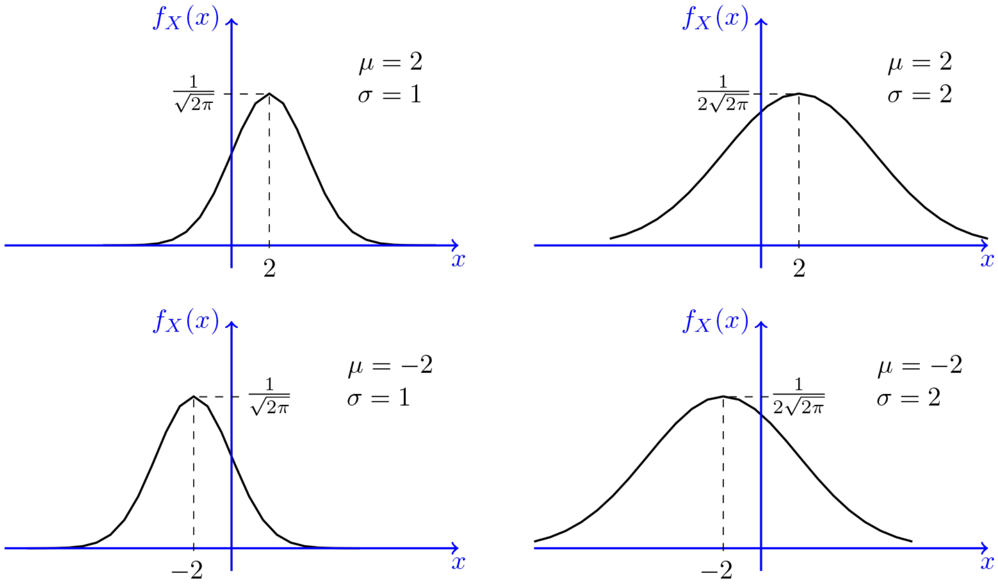


Figure 7: PDF for Gaussian Distribution.

8.4 Exponential Distribution

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events.

A continuous random variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim \mathit{Exponential}(\lambda)$, if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

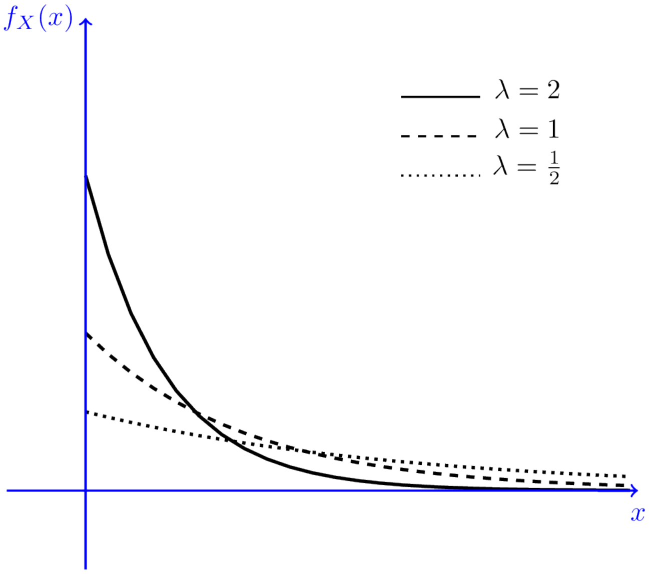


Figure 8: PDF for Exponential Random Variable.

It is convenient to use the unit step function defined as

$$u(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

so we can write the PDF of an *Exponential*(λ) random variable as

$$f_X(x) = \lambda e^{-\lambda x} u(x)$$

Let us find its CDF, mean and variance. For $x > 0$, we have

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$

So we can express the CDF as

$$F_X(x) = \big(1 - e^{-\lambda x}\big) u(x)$$

Let $X \sim \textit{Exponential}(\lambda)$. We can find its expected value as follows, using integration by parts:

$$\begin{aligned} EX &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \\ &= \frac{1}{\lambda} \Big[-e^{-y} - y e^{-y} \Big]_0^\infty \\ &= \frac{1}{\lambda}. \end{aligned}$$

Now let's find $Var(X)$. We have

$$
\begin{aligned}
EX^2 &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\
&= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy \\
&= \frac{1}{\lambda^2} \Big[-2e^{-y} - 2ye^{-y} - y^2 e^{-y} \Big]_0^\infty \\
&= \frac{2}{\lambda^2}.
\end{aligned}
$$

Thus, we obtain

$$\text{Var}(X) = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

9 The Multivariate Gaussian Distribution

A vector-valued random variable $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mu \in R_n$ and covariance matrix $\Sigma \in S^n_{++}$ if its probability density function is given by

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

S^n_{++} is the space of symmetric positive definite $n \times n$ matrices

9.1 Relationship to univariate Gaussians

Recall that the density function of a univariate normal (or Gaussian) distribution is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\},$$

Here, the argument of the exponential function, $\exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$, is a quadratic function of the variable x. Furthermore, the parabola points downwards, as the coefficient of the quadratic term is negative. The coefficient in front, $\frac{1}{\sigma\sqrt{2\pi}}$, is a constant that does not depend on x; hence, we can think of it as simply a "normalization factor" used to ensure that

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} = 1$$

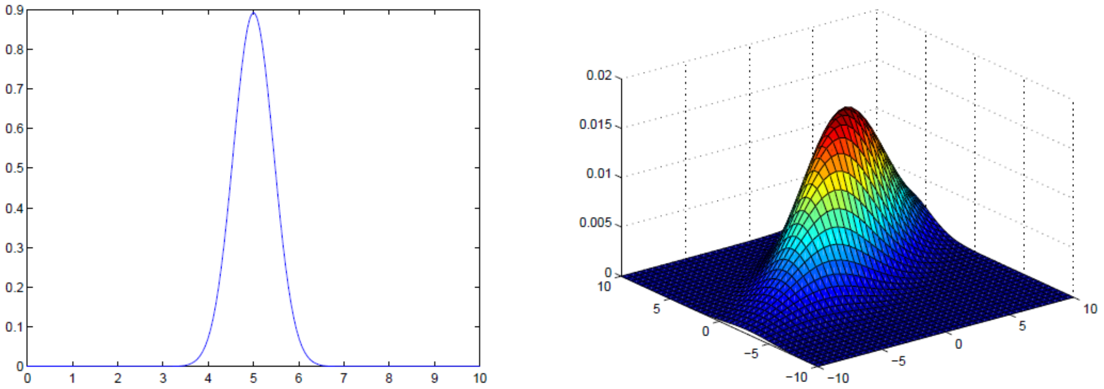


Figure 9: The figure on the left shows a univariate Gaussian density for a single variable X. The figure on the right shows a multivariate Gaussian density over two variables X1 and X2.

In the case of the multivariate Gaussian density, the argument of the exponential function, $\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$, is a quadratic form in the vector variable x. Since Σ is positive definite, and since the inverse of any positive definite matrix is also positive definite, then for any non-zero vector z, $z^T\Sigma^{-1}z > 0$. This implies that for any vector $x \neq \mu$,

$$(x-\mu)^T\Sigma^{-1}(x-\mu) > 0$$

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) < 0$$

Like in the univariate case, you can think of the argument of the exponential function as being a downward opening quadratic bowl. The coefficient in front has an even more complicated form than in the univariate case. However, it still does not depend on x, and hence it is again simply a normalization factor used to ensure that

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx_1dx_2\cdots dx_n = 1$$

9.2 The covariance matrix

The concept of the covariance matrix is vital to understanding multivariate Gaussian distributions. Recall that for a pair of random variables X and Y , their covariance is defined as

$$Cov[X,Y] = E[(X-E[X])(Y-E[Y])] = E[XY] - E[X]E[Y].$$

When working with multiple variables, the covariance matrix provides a succinct way to summarize the covariances of all pairs of variables. In particular, the covariance matrix, which we usually denote as Σ, is the $n \times n$ matrix whose $(i,j)th$ entry is $Cov[X_i, X_j]$.

9.2.1 The diagonal covariance matrix case

To get an intuition for what a multivariate Gaussian is, consider the simple case where $n = 2$, and where the covariance matrix Σ is diagonal, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

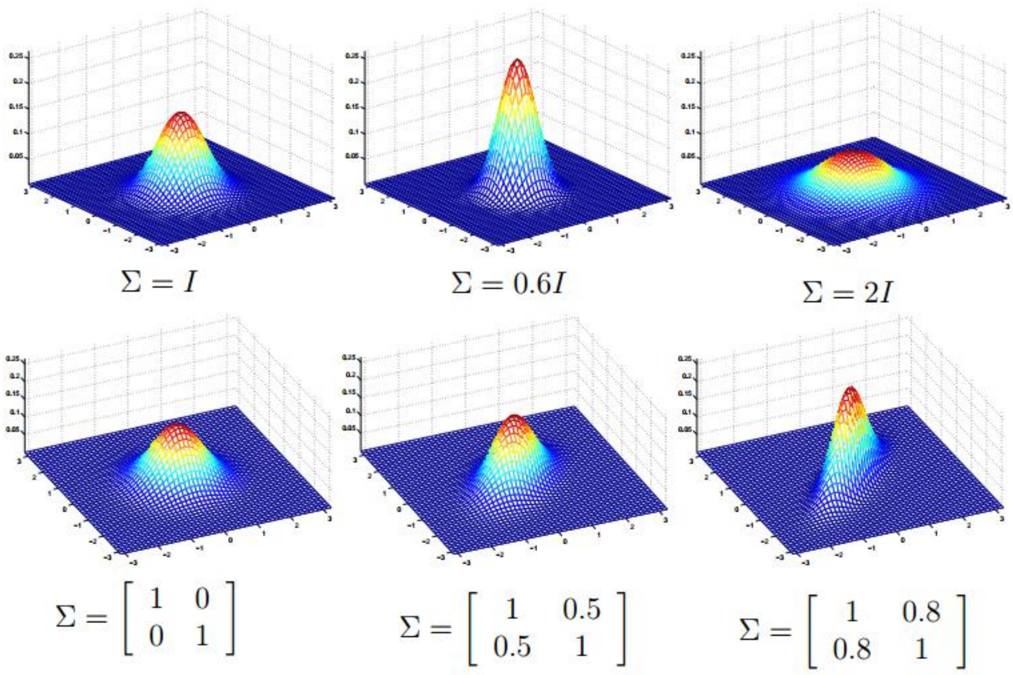
In this case, the multivariate Gaussian density has the form,

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right) \\ &= \frac{1}{2\pi \left(\sigma_1^2 \cdot \sigma_2^2 - 0 \cdot 0\right)^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right) \end{aligned}$$

where we have relied on the explicit formula for the determinant of a 2×2 matrix, and the fact that the inverse of a diagonal matrix is simply found by taking the reciprocal of each diagonal entry. Continuing,

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2}(x_1 - \mu_1) \\ \frac{1}{\sigma_2^2}(x_2 - \mu_2) \end{bmatrix} \right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \end{aligned}$$

The last equation we recognize to simply be the product of two independent Gaussian den sities, one with mean μ_1 and variance σ_1^2 and the other with mean μ_2 and variance σ_2^2 More generally, one can show that an n-dimensional Gaussian with mean $\mu \in R_n$ and diagonal covariance matrix $\Sigma = diag(\sigma_1^2, \sigma_2^2, .., \sigma_n^2)$ is the same as a collection of n independent Gaussian random variables with mean μ_i and variance σ_i^2, respectively.



$\Sigma = I$ $\Sigma = 0.6I$ $\Sigma = 2I$

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

Figure 10: Gaussian distribution for different covariance matrix

10 Basics of Probability

Probability theory allows us to describe and analyze random phenomena i.e. events or experiments whose outcomes we cannot predict with certainty. One way to think about randomness is that it is a way of expressing what we don't know. Let us consider a random experiment of flipping a fair coin. Perhaps if we knew the force the coin was flipped with, the initial orientation of the coin, the impact point between finger and coin, the turbulence of the air, the surface smoothness of the table the coin lands on, the material characteristics of the coin and the table, etc, we would be able to definitively predict the outcome of the flip. However, in the absence of all this information, we cannot predict the outcome of the coin flip. When we say that something is random, we are saying that our knowledge about the outcome is limited so we cannot be sure what will happen.

10.1 Random Experiment

A random experiment is a process by which we observe something uncertain. Rolling a die is a random experiment because we don't know the result.

- **Outcome:** It is the result of a random experiment.
- **Sample Space:** It is the set of all possible outcomes.
- **Event:** It is a subset of the sample space.

Some examples of random experiments and their sample spaces:

- Tossing a coin. Sample Space $S = \{H, T\}$
- Rolling a die. $S = \{1, 2, 3, 4, 5, 6\}$

10.2 Probability

A probability measure $P(A)$ is assigned to an event A, the value of which lies between 0 and 1 and denotes how likely the event is. If $P(A)$ is close to 0 then the event is highly unlikely. On the other hand, if $P(A)$ is close to 1 then event A is very likely to occur.

Axioms of Probability

1. For any event $A, P(A) \geq 0$
2. Probability of the sample space S is $P(S) = 1$
3. If $A_1, A_2, A_3, ...$ are disjoint events then
   $P(A_1 \cup A_2 \cup A_3 ...) = P(A_1) + P(A_2) + P(A_3) + ...$

In a finite sample space S, where all outcomes are equally likely, the probability of any event A can be found by

$$P(A) = \frac{|A|}{|S|}$$

Inclusion Exclusion Principle:

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

In general, given n events $A_1, A_2, ..., A_n$, we have

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P\Big(\bigcap_{i=1}^{n} A_i\Big)$$

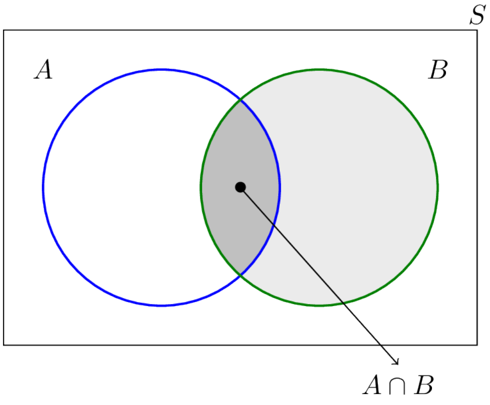
Example 1.1: In a presidential election, there are four candidates. Call them A, B, C, and D. Based on our polling analysis, we estimate that A has a 20 percent chance of winning the election, while B has a 40 percent chance of winning. What is the probability that A or B win the election?

Solution: Notice that the events that {A wins }, {B wins }, {C wins } and {D wins } are disjoint since more than one of them cannot occur at the same time. For example, if A wins, then B cannot win. From the third axiom of probability, the probability of the union of two disjoint events is the summation of individual probabilities. Therefore,

$$\begin{aligned} P(\text{A wins or B wins}) &= P\big(\{\text{A wins}\} \cup \{\text{B wins}\}\big) \\ &= P(\{\text{A wins}\}) + P(\{\text{B wins}\}) \\ &= 0.2 + 0.4 \\ &= 0.6 \end{aligned}$$

10.3 Conditional Probability

Conditional probability asks the question that if we obtain additional information, how should we update the probabilities of events? For example, suppose that 23% of the days are rainy. Thus if we pick a random day, the probability that it rains on that day is 23 %. Now suppose we know that it is cloudy on a particular day. How does having this extra piece of information change the probability for it to rain on that day, i.e. what is the probability that it rains **given that** it is cloudy? If C is the event that it is cloudy and event R is the event that it rains, then the probability that it rains given it is cloudy, is given by $P(R|C)$



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Figure 11: Venn Diagram for conditional probability

If A and B are two events in a sample space S, then the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ when } P(B) > 0$$

10.4 Independence

Let A be the event that it rains tomorrow, and suppose that $P(A) = \frac{1}{3}$. Also suppose that a fair coin is tossed such that B is the probability that it lands heads up. We have $P(B) = \frac{1}{2}$. What is $P(A|B)$? You are right, $P(A|B) = P(A) = \frac{1}{3}$. The result of the coin toss does not have anything to do with tomorrow's weather. Thus, no matter if B happens or not, the probability of A should not change.

Two events A and B are independent if and only if $P(A \cap B) = P(A).P(B)$. Thus,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(A)P(B)}{P(B)}$$

$$= P(A)$$

10.5 Law of Total Probability

If $B_1, B_2, B_3, ...$ is a partition of the sample space S, then for any event A we have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

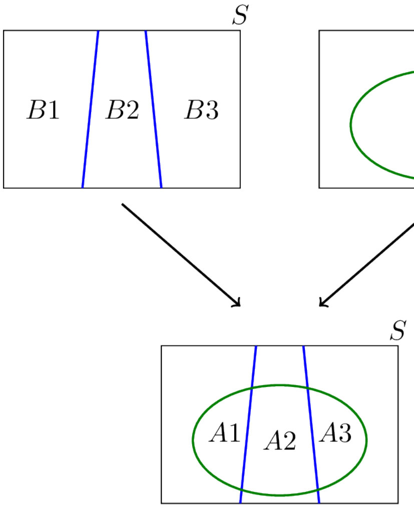


Figure 12: Law of total probability

Using a Venn diagram, we can pictorially see that

$$A_1 = A \cap B_1$$
$$A_2 = A \cap B_2$$
$$A_3 = A \cap B_3$$

Here, A_1, A_2 and A_3 form a partition of the set A.

$$P(A) = P(A_1) + P(A_2) + P(A_3)$$

10.6 Baye's Rule

Suppose that we know $P(A|B)$, but we are interested in the probability $P(B|A)$.

- For any two events A and B, where $P(A) \neq 0$, we have

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

- If $B_1, B_2, B_3, ...$ form a partition of the sample space S, and A is any event with $P(A) \neq 0$, we have

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Example 1.2: A certain disease affects about 1 out of 10,000 people. There is a test to check whether the person has the disease. The test is quite accurate. In particular, we know that

- The probability that the test result is positive (suggesting the person has the disease), given that the person does not have the disease, is only 2 percent.
- The probability that the test result is negative (suggesting the person does not have the disease), given that the person has the disease, is only 1 percent.

A random person gets tested for the disease and the result comes back positive. What is the probability that the person has the disease?

Solution: Let D be the event that the person has the disease, and let T be the event that the test result is positive. We know

$$P(D) = \frac{1}{10,000}$$
$$P(T|D^c) = 0.02$$
$$P(T^c|D) = 0.01$$

Using Baye's rule,

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)}$$
$$= \frac{(1-0.01)\times 0.0001}{(1-0.01)\times 0.0001 + 0.02\times(1-0.0001)}$$
$$= 0.0049$$

Another way to think about the problem is using a tree diagram:

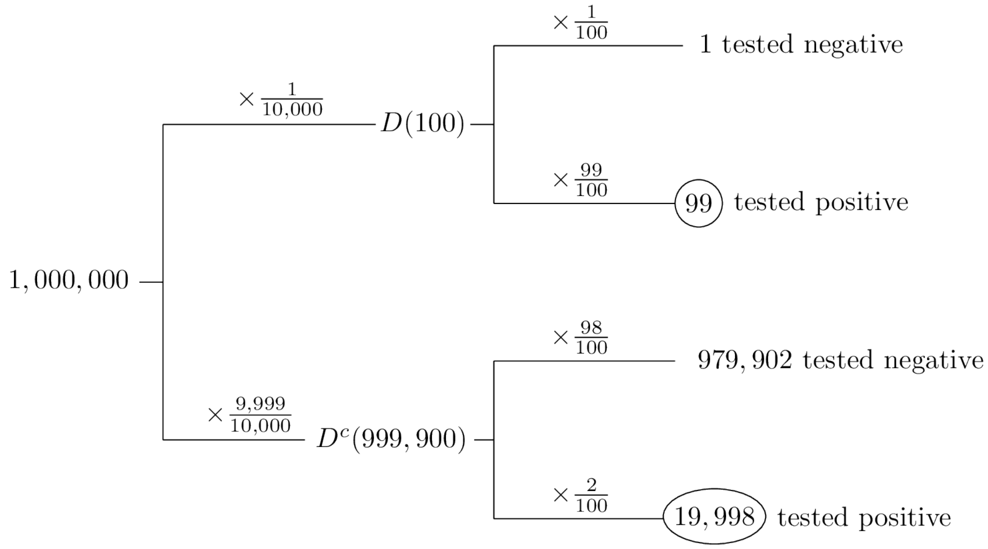


Figure 13: Tree diagram for Example 1.2

Suppose 1 million people get tested for the disease. Out of the one million people, about 100 of them have the disease, while the other 999,900 do not have the disease. Out of the 100 people who have the disease 1000.99 = 99 people will have positive test results. However, out of the people who do not have the disease 999,9000.02 = 19998 people will have positive test results. Thus in total there are 19998 + 99 people with positive test results, and only 99 of them actually have the disease. Therefore, the probability that a person from the "positive test result" group actually have the disease is

$$P(D|T) = \frac{99}{19998+99} = 0.0049$$

References

[1] http://www.probabilitycourse.com

[2] http://cs229.stanford.edu/section/gaussians.pdf

[3] https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/

[4] https://en.wikipedia.org/wiki/Covariance