Probability & Statistics

Teemu Weckroth, September 7, 2023

De Morgan's Laws

$$(A \cup B)^C = A^C \cap B^C$$
$$(A \cap B)^C = A^C \cup B^C$$

Probability function

Let Ω be a finite sample space. A probability function P assigns to each event A in Ω a number $P(A)$ in $[0, 1]$ such that:

1. $P(\Omega) = 1$
2. $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint

The number $P(A)$ is called the probability of A

Addition and complement rule

For any two events A and B we have

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

For any event A we have

$$P(A^C) = 1 - P(A)$$

Conditional probability

The conditional probability of A given C is defined as

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

provided that $P(C) > 0$

Multiplication rule

For any events A and C it hold that

$$P(A \cap C) = P(A \mid C)P(C)$$

Independence equivalence

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

Law of total probability

Let A and C be two events. We have

$$P(A) = P(A \mid C)P(C) + P(A \mid C^C)P(C^C)$$

Suppose we have disjoint events $C_1, C_2, ..., C_m$ such that $C_1 \cup C_2 \cup ... \cup C_m = \Omega$. For any event A we have

$$P(A) = P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + ... + P(A \mid C_m)P(C_m)$$

Bayes' rule

Suppose the events $C_1, C_2, ..., C_m$ are disjoint and fill up the sample space Ω. Then the conditional probability of C_i given the same event A is

$$P(C_i \mid A) = \frac{P(A \mid C_i)P(C_i)}{P(A \mid C_1)P(C_1) + ... + P(A \mid C_m)P(C_m)}$$

Discrete random variable

Let Ω be a sample space. A discrete random variable is a function $X : \Omega \to \mathbb{R}$ that takes on a finite number of values $a_1, a_2, ..., a_n$ or an infinite number of values $a_1, a_2, a_3, ...$

Probability mass function

The probability mass function p of a discrete random variable X is the function $p : \mathbb{R} \to [0, 1]$ defined by

$$p(a) = P(X = a) \text{ for } -\infty < a < \infty$$

Distribution function

The distribution function F of a discrete random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(a) = P(X \leq a) \text{for } -\infty < a < \infty$$

Binomial coefficient

The binomial coefficient $\binom{n}{k}$ gives the number of combinations of k objects from a set of n objects:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Bernoulli distribution

A random variable X has a Bernoulli distribution if X only takes on the values 0 and 1 with probabilities

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$X \sim \text{Ber}(p)$

Binomial distribution

A random variable X has a binomial distribution with parameters n and p if X can take on the values $k = 0, 1, ..., n$ with probabilities

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$X \sim \text{Bin}(n, p)$

Geometric distribution

A random variable X has a geometric distribution with parameter p if X can take on the values $k = 1, 2, 3, ...$ with probabilities

$$P(X = k) = p \cdot (1-p)^{k-1}$$

$X \sim \text{Geo}(p)$

Poisson distribution

A random variable X has a Poisson distribution with parameter μ if X can take on the values $k = 0, 1, 2, ...$ with probabilities

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$$

$X \sim \text{Pois}(\mu)$
$Y \approx \text{Pois}(np)$ for $Y \sim \text{Bin}(n, p)$ with large n and small np

Uniform distribution

A continuous random variable X has a uniform distribution on the interval $[\alpha, \beta]$ if its probability density function f is given by

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{for } x \notin [\alpha, \beta] \end{cases}$$

$X \sim \text{U}(\alpha, \beta)$

Exponential distribution

A continuous random variable X has an exponential distribution with parameter λ if its probability density function f is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

$X \sim \text{Exp}(\lambda)$

Pareto distribution

A continuous random variable X has a pareto distribution with parameter $\alpha > 0$ if its probability density function f is given by

$$f(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}} & \text{for } x \geq 1 \\ 0 & \text{for } x < 1 \end{cases}$$

$X \sim \text{Par}(\alpha)$

Normal distribution

A continuous random variable X has a normal distribution with parameters μ and σ^2 if its probability density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$X \sim \text{N}(\mu, \sigma^2)$

Standard normal distribution

If $\mu = 0$ and $\sigma^2 = 1$, the distribution $\text{N}(0, 1)$ is called the standard normal distribution

Quantiles

Let X be a continuous random variable and $0 \leq p \leq 1$. The pth quantile or the $100p$th percentile of the distribution of X is the smallest number q_p such that

$$F_X(q_p) = P(X \leq q_p) = p$$

Expectation of a random variable

Expectation $E[X]$ of a discrete random variable X is defined as the number

$$\text{E}[X] = \sum_{a_i} a_i \cdot P(X = a_i)$$

Expectation $E[X]$ of a continuous random variable X with pdf f is given by

$$\text{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Change-of-variable formula

Let X be a RV and $g: \mathbb{R} \to \mathbb{R}$ a function
If X is discrete:

$$E[g(X)] = \sum_i g(a_i)P(X = a_i)$$

If X is continuous:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Variance

Variance of a random variable X is defined as

$$\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

Standard deviation of a random variable X is defined as

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Change-of-units formula

For any random variable X and any real values r and s it holds that

$$E[rX + s] = rE[X] + s$$

$$\mathrm{Var}(rX + s) = r^2\mathrm{Var}(X)$$

Jensen's inequality

Let g be a convex function and let X be a random variable. Then

$$g(E[X]) \le E[g(x)]$$

Let g be a concave function and let X be a random variable. Then

$$g(E[X]) \ge E[g(X)]$$

Transforming normal RVs

Suppose $X \sim N(\mu, \sigma^2)$, then the RV $rX + s$ also has a normal distribution:

$$rX + s \sim N(r\mu + s, r^2\sigma^2)$$

Every normally distributed RV can also be transformed into a standard normal RV:

$$\text{if } X \sim N(\mu, \sigma^2), \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Joint probability mass function

Let X and Y be two discrete RVs. The joint probability mass function of X and Y is the function defined by $p: \mathbb{R}^2 \to [0, 1]$

$$p(a, b) = P(X = a, Y = b) \text{ for all } a \text{ and } b$$

From joint to marginal, take sum of rows and columns:

$$p_X(x) = \sum_y p(x, y)$$

$$p_Y(y) = \sum_x p(x, y)$$

Joint distribution function

Let X and Y be two RVs. The joint distribution function F of X and Y is the function $F: \mathbb{R}^2 \to [0, 1]$ defined by

$$F(a, b) = P(X \le a, Y \le b) \text{ for all } a \text{ and } b$$

Joint density function

Let X and Y be two continuous RVs. The joint density function f of X and Y is the function $f: \mathbb{R}^2 \to \mathbb{R}$ such that

$$P(a_1 \le X \le b_1, a_2 \le Y \le b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x, y)dxdy$$

Marginal density function

Let f be the joint density function of X and Y. Then the marginal densities of X and Y can be found as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$$

Expectations of a function of two RVs

Let X and Y be random variables and let $g: \mathbb{R}^2 \to \mathbb{R}$ be a function
If X and Y are discrete with values $a_1, a_2, \dots$ and $b_1, b_2, \dots$ respectively, then

$$E[g(X, Y)] = \sum_i \sum_j g(a_i, b_j)P(X = a_i, Y = b_j)$$

If X and Y are continuous with joint probability density function f, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$$

Covariance

Let X and Y be two RVs. The covariance between X and Y is

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

If $\mathrm{Cov}(X, Y) > 0$, then X and Y are positive correlated
If $\mathrm{Cov}(X, Y) < 0$, then X and Y are negatively correlated
If $\mathrm{Cov}(X, Y) = 0$, then X and Y are not correlated
Let X and Y be two RVs. Then always

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

If X and Y are uncorrelated, then

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

Let X and Y be two RVs. Then

$$\mathrm{Cov}(rX + s, tY + u) = rt\mathrm{Cov}(X, Y)$$

for all values r, s, t, and u

Correlation

Let X and Y be two RVs. The correlation coefficient is

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

if $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$. Else $\rho(X, Y) = 0$

Expected number of events

$E[M(0, 1)]$ = expected number of events in interval of unit length = intensity of the process λ

$$E[M(a, b)] = np = (b - a)\frac{n}{b - a}p = (b - a)E[M(0, 1)] = \lambda(b - a)$$

$N(a, b) \sim \mathrm{Pois}(\lambda(b - a))$

Sum of two independent discrete RVs

Let X and Y be two independent discrete RVs. The pmf of $Z = X + Y$ satisfies

$$p_Z(c) = \sum_j p_X(c - b_j)p_Y(b_j)$$

where the sum runs over all possible values b_j of Y
If $X \sim \mathrm{Bin}(n, p)$, $Y \sim \mathrm{Bin}(m, p)$ and X and Y are independent, then $X + Y \sim \mathrm{Bin}(n + m, p)$

Sum of two independent continuous RVs

Let X and Y be two independent continuous RVs. The pdf of $Z = X + Y$ satisfies

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y)f_Y(y)dy$$

If $X \sim N(\mu, \sigma^2)$, $Y \sim N(\nu, \tau^2)$ and X and Y are independent, then $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$

Independent and identically distributed sequence of random events

$X_1, X_2, \dots, X_n$ all have the same distribution and are independent
Same expectations $E[X_i] = \mu$
Same variances $\mathrm{Var}(X_i) = \sigma^2$

$$E[\bar{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Rules of the variance

$$\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$$

For independent random variables:

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

Variances of the average

If $X_1, X_2, \dots, X_n$ is an i.i.d. sequence, $\mathrm{Var}(X_i) = \sigma^2$, then

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

Chebyshev's inequality

Suppose $E(X) = \mu$, $\mathrm{Var}[X] = \sigma^2$. Then

$$P(|X - \mu| \ge a) \le \frac{\sigma^2}{a^2}$$

Taking $a = k\sigma$,

$$P(|X - \mu| \ge k\sigma) \le \frac{\sigma^2}{(k\sigma)^2} = \frac{1}{k^2}$$

Law of large numbers

Let $X_1, \dots X_n$ be an i.i.d. sequence with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then for any $\epsilon > 0$

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

Central limit theorem

Let $X_1, ..., X_n$ be an i.i.d. sequence with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. For $n \geq 1$, let Z_n be defined by

$$Z_n = \frac{X_n - \mu}{\sigma/\sqrt{n}}$$

Then for any number a it holds that

$$P(Z_n \leq a) \to P(Z \leq a) \text{ as } n \to \infty$$

where Z has a $\mathrm{N}(0,1)$ distribution
In other words: for large n, Z_n, $\bar{X}_n$, and $\sum X_i$ all approximately have a normal distribution

$$Z_n \overset{d}{\approx} \mathrm{N}(0,1)$$

$$\bar{X}_n \overset{d}{\approx} \mathrm{N}(\mu, \sigma^2/n)$$

$$\sum_{i=1}^{n} X_i \overset{d}{\approx} \mathrm{N}(n\mu, n\sigma^2)$$

Histogram

1. Divide the range of the data into intervals
2. Determine the height

$$\text{Height on } B_i = \frac{\#B_i}{n \cdot |B_i|}$$

$$\text{Area on } B_i = \frac{\#B_i}{n}$$

Empirical distribution function

The value of the empirical distribution function at a point x is equal to the fraction of datapoints that is smaller than or equal to x

$$F_n(x) = \frac{\#x_i \leq x}{n}$$

Standard deviation as a measure of variation

The sample variance of a dataset is defined as

$$S_n^2 = \frac{1}{n-1}((x_1 - \bar{x}_n)^2 + ... + (x_n - \bar{x}_n)^2)$$

The sample standard deviation is defined as

$$S_n = \sqrt{S_n^2}$$

Median Absolute Deviation as measure of variation

The median of absolute deviation (MAD) of a dataset is defined as

$$\mathrm{MAD} = \mathrm{Med}(|x_1 - m_n|, ..., |x_n - m_n|)$$

Five-number summary

1. Minimum
2. Lower quartile
3. Median
4. Upper quartile
5. Maximum

Quartiles and their computation

Let $x_1, ..., x_n$ be a dataset. For any $p \in [0,1]$ the pth empirical quantile is the number $q_n(p)$ such that a proportion p of the dataset is less than $q_n(p)$ and a proportion $1-p$ is larger than $q_n(p)$
Let $x_1, ..., x_n$ be an ordered dataset. Compute

$$p(n+1) = k + \alpha,$$

where k is the integer part of $p(n+1)$ and α is its decimal part. Then

$$q_n(p) = x_k + \alpha(x_{k+1} - x_k)$$

Random sample and statistical model

A random sample is a collection of RVs $X_1, X_2, ..., X_n$ that have the same probability distribution and are mutually independent.

Model distribution

The probability distribution of each random variable from a random sample is called the model distribution
The random variable $h(X_1, ..., X_n)$ depending only on the random sample $X_1, ..., X_n$ is called a sample statistics

Estimators

An estimate t is a value that depends only on the data

$$t = h(x_1, x_2, ..., x_n)$$

An estimator is a random variable that gives the value of an estimate calculated from a random sample $X_1, X_2, ..., X_n$

$$T = h(X_1, X_2, ..., X_n)$$

Unbiased estimators

An unbiased estimator is an estimator T for the parameter λ such that $\mathrm{E}[T] = \lambda$ for all values of λ

Sampling distribution

Let $T = h(X_1, X_2, ..., X_n)$ be an estimator based on a random sample $X_1, X_2, ..., X_n$. The probability distribution of T is called the sampling distribution of T.

Unbiased estimators for mean and variance

Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with finite expectation μ and variance σ^2. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

are unbiased estimators of μ and σ^2

Mean squared error

Let T be an estimator for the parameter θ. The mean squared error of T is defined as

$$\mathrm{MSE}(T) = \mathrm{E}[(T-\theta)^2] = \mathrm{Var}(T) + (\mathrm{E}[T] - \theta)^2$$

Efficiency

Let T_1 and T_2 be two estimators for the same parameter. If

$$\mathrm{MSE}(T_2) < \mathrm{MSE}(T_1),$$

we say that T_2 is more efficient than T_1

General coefficient intervals

Let X_i have a distribution dependent on the parameter θ. Suppose sample statistics $L_n = g(X_1, ..., X_n)$ and $U_n = h(X_1, ..., X_n)$ exist such that

$$P(L_n < \theta < U_n) = \gamma \text{ for some } 0 < \gamma < 1 \text{ and all } \theta$$

Then, given a realization $x_1, ..., x_n$ of the variables $X_1, ..., X_n$ and $l_n = g(x_1, ..., x_n)$ and $u_n = h(x_1, ..., x_n)$, the interval

$$(l_n, u_n)$$

is a $100\gamma\%$ confidence interval for θ

Critical values of the normal distribution

The critical value of a standard normal distribution is the real number z_p such that

$$P(Z \geq z_p) = p$$

where $Z \sim \mathrm{N}(0,1)$

Confidence interval for the mean of a normal distribution; variance known

Suppose $X_1, ..., X_n$ are independent and normally distributed with parameters μ and σ^2. Then

$$P(\bar{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Confidence interval for the mean of a normal distribution; variance unknown

Suppose $X_1, ..., X_n$ are independent and normally distributed with parameters μ and σ^2. If the dataset $x_1, ..., x_n$ is a realization of the random sample $X_1, ..., X_n$ and $\gamma = 1 - \alpha$, then

$$(\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}})$$

is called a $100\gamma\%$ confidence interval for μ

Three steps of hypothesis testing

1. Formulate H_0 and H_1
2. Do the experiment
3. Calculate whether results justify rejecting H_0

DO NOT REJECT H_0	REJECT H_0
Insufficient evidence to support H_1	H_1 true beyond reasonable doubt

Test statistic

Suppose the data are modelled as a realization of random variables X_i. A test statistic is any sample statistic

$$T = h(X_1, ..., X_n)$$

whose numerical value is used to decide whether we reject H_0

Tail probabilities

Give a test statistic T, a left tail probability is $P(T \leq t)$ for some t. A right tail probability is $P(T \geq t)$
The $p-$value is the probability, given H_0 is true, of an event at least as extreme as the observations in the direction which provides evidence for H_1
The $p-$value reflects how improbable the observed value t is under H_0: small $p-$values are bad for the null

Significance level and critical region

The significance level α is the largest acceptable probability of committing a type I error

Suppose we test H_0 against H_1 by means of the test statistic T. The set of values for T for which we reject H_0 in favour of H_1 is called the critical region
Values on the boundary of this region are called critical values

t-test statistic

The $t-$test statistic is defined as

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1)$$

Normal samples

Let $X_1, ..., X_n$ be a sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution. To test the null hypothesis $H_0 : \mu = \mu_0$, we define the $t-$test statistic T as

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

Then the distribution of this statistic under $H_0 : \mu = \mu_0$ is

$$T \sim t(n-1)$$

To perform a $t-$test for samples from normal data with unknown variance at significance level α:

1. Formulate the hypotheses
2. Compute the value of the $t-$test statistic
3. Compare this value with the critical values $t_{n-1,\alpha/2}$ or $t_{n-1,\alpha}$ depending on two-sided/one-sided test
4. Decide whether to reject the null hypothesis

Large samples

Let $X_1, ..., X_n$ be a sample from an unknown distribution. For large n, the distribution of the studentized mean can be approximated by the standard normal distribution
To perform a $t-$test for samples large samples from non-normal data at significance level α:

1. Formulate the hypotheses
2. Compute the value of the $t-$test statistic
3. Compare this value with the critical values $z_{\alpha/2}$ or z_α depending on two-sided/one-sided test
4. Decide whether to reject the null hypothesis