Some frequently used mathematical functions, formulas and conceptions

Statistics and probability

In statistics we apply probability to draw conclusions from data. This can be messy and usually involves as much art as science.

Experiment

A repeatable procedure with well defined possible outcomes.

Sample space

the set of all possible outcomes. it is usually denoted by S.

Event

a subset of sample space

Probability function

A function giving the probability for each outcome.

Probability density

A continuous distribution of probabilities.

Random variable

A random numerical outcome.

Mean, Median, and Mode

Mean, Median, and Mode

- The distribution function $F(x)$ or the density $f(x)$ (or pmf $p(x_i)$) completely characterizes the behavior of a random variable X.
- Often, we need a more concise description such as a single number or a few numbers, instead of an entire function.
- Quantities most often used to describe a random variable X are
  - the **expectation** or the **mean**, $E[X]$.
  - the **median**, any number x such that $P(X < x) \leq 1/2$ and $P(X > x) \geq 1/2$ and
  - the **mode**, any number x for which $f(x)$ or $p(x_i)$ attains its maximum.
- The mean, median, and mode are often called **measures of central tendency** of a random variable X.

Regression Analysis

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').

Least squares

The method of least squares is a standard approach in regression analysis to approximate the solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

Let Y be the outcome vector, X is the matrix by stacking all the x vector.

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad X = \begin{bmatrix} — & \vec{x}_1 & — \\ & \vdots & \\ — & \vec{x}_N & — \end{bmatrix}$$

Usually their is no consistent solution w for Y=Xw. But we can work out w that minimize the following square error. In this sense, the w is an optimal solution to work out the problem.

$$\text{squared error} = ||Y - X\vec{w}||^2$$

and the optimal solution w is:

$$\vec{w} = (X^\top X)^{-1}(X^\top Y)$$

Probability density function

A probability density function for X is real valued function f which satisfies

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

for all a, b that belongs to R.

Cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Also

$$\frac{d}{dx}F(x) = f(x)$$

Chain rule formula

$$F'(x) = f'(g(x))g'(x)$$

another form of chain rule

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

Failure rate in the unit of time

$$\lambda = \frac{n}{N * \Delta t}$$

n is the number of failure component

N is the number of total observed components

delta t is the observing time period

MTBF: mean time between failures

$$MTBF = \frac{1}{\lambda}$$

Reliability

$$R(t) = e^{-\lambda t}$$

The mean of discrete random variables

$$\mu = \frac{1}{n}(a_1 + \ldots + a_n) = \frac{1}{n}\sum_{i=1}^{n} a_i$$

Variance of discrete random variables

$$\sigma^2 := \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)^2$$

Covariance

Covariance is a measure of how much two random variables vary together. Suppose X and Y are random variables with means uX and uY. the covariance of X and Y is defined as

$$\operatorname{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

Properties of covariance

1. $\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X,Y)$ for constants a, b, c, d.
2. $\mathrm{Cov}(X_1+X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$.
3. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
4. $\mathrm{Cov}(X, Y) = E(XY) - \mu_X \mu_Y$.
5. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$ for any X and Y.
6. If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.

Correlation

The units of covariance $\mathrm{Cov}(X, Y)$ are 'units of X times units of Y'. This makes it hard to compare covariances: if we change scales then the covariance changes as well. Correlation is a way to remove the scale from the covariance.

Definition: The correlation coefficient between X and Y is defined by

$$\mathrm{Cor}(X,Y) = \rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \, \sigma_Y}.$$

Properties of correlation

1. ρ is the covariance of the standardizations of X and Y.
2. ρ is dimensionless (it's a ratio!).
3. $-1 \le \rho \le 1$. Furthermore,
   $\rho = +1$ if and only if $Y = aX + b$ with $a > 0$,
   $\rho = -1$ if and only if $Y = aX + b$ with $a < 0$.

The expectation of a random variable X with probability density function f(x)

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

And the variance of X

$$\operatorname{Var}(X) \quad := \quad \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx$$

And variance formula alternative

$$\operatorname{Var}(X) \quad = \quad \int_{-\infty}^{\infty} x^2 f(x) dx - E(X)^2$$

Two core conceptions about probability statistics

The frequentist definition sees probability as the long-run expected frequency of occurrence. P(A) = n/N, where n is the number of times event A occurs in N opportunities.

The Bayesian view of probability is related to degree of belief. It is a measure of the plausibility of an event given incomplete knowledge.

Conditional probability

The conditional probability of A given B

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

Multiplication Rule

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Bayes theorem

$$P(\mathcal{H} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{H}) P(\mathcal{H})}{P(\mathcal{D})}$$

alternatively

$$P(\text{hypothesis is true} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis is true}) \cdot P(\text{hypothesis is true})}{P(\text{data})}$$

Likelihood

$$L(\theta) = f(x_1, .., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Maximum likelihood estimate

$$L(\hat{\theta}_{MLE}) \geq L(\theta), \quad \forall \theta$$

Loglikelihood

$$l(\theta\,; x) = \log L(\theta\,; x)$$

More exactly,

$$\begin{aligned} l(\theta\,; x) &= \log f(x\,; \theta) \\ &= \log \prod_{i=1}^{n} f(x_i\,; \theta) \\ &= \sum_{i=1}^{n} \log f(x_i\,; \theta) \\ &= \sum_{i=1}^{n} l(\theta\,; x_i). \end{aligned}$$

Confidence and Confidence Interval

Just as $\bar{x}$ is normally distributed about μ, $\hat{\theta}$ is approximately normally distributed about θ in large samples.

This property is called the "asymptotic normality of the MLE," and the technique of forming confidence intervals is called the "asymptotic normal approximation." This method works for a wide variety of statistical models, including all the models that we will use in this course.

The asymptotic normal 95% confidence interval for a parameter θ has the form

$$\hat{\theta} \pm 1.96 \frac{1}{\sqrt{-l''(\hat{\theta}; x)}}, \tag{2}$$

where $l''(\hat{\theta}; x)$ is the second derivative of the loglikelihood function with respect to θ, evaluated at $\theta = \hat{\theta}$.

IID or iid

Two random variables {\displaystyle X}
and {\displaystyle Y}
are **i.i.d.** if they are independent *and* identically distributed, i.e. if and only if

$$\begin{aligned} F_X(x) &= F_Y(x) && \forall x \in I \\ F_{X,Y}(x,y) &= F_X(x) \cdot F_Y(y) && \forall x, y \in I \end{aligned} \qquad \textbf{(Eq.1)}$$

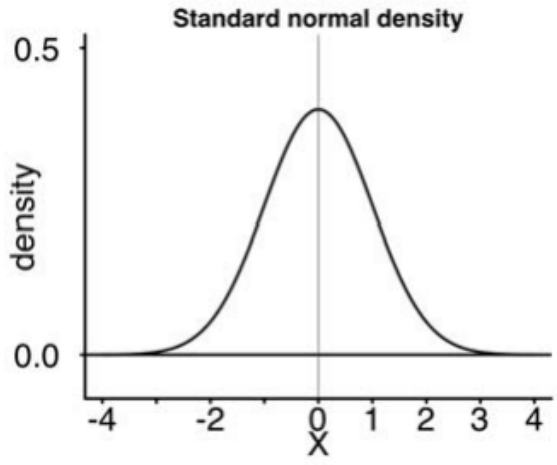
Binomial Distribution

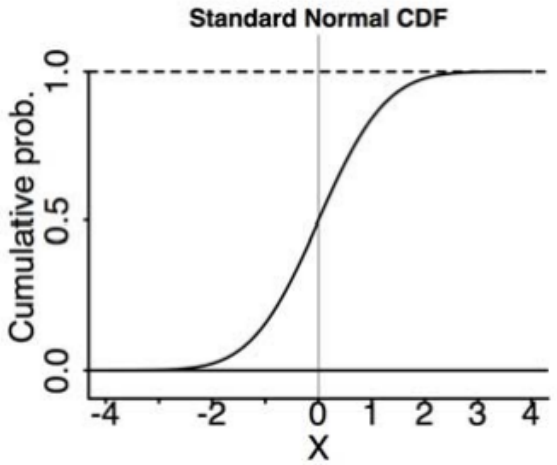
$$b(n,p,k) = \binom{n}{k} p^k q^{n-k}$$

Normal distribution

Also called Gauss distribution

1. Parameters: μ, σ.
2. Range: $(-\infty, \infty)$.
3. Notation: normal(μ, σ^2) or $N(\mu, \sigma^2)$.
4. Density: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$.
5. Distribution: $F(x)$ has no formula, so use tables or software such as `pnorm` in R to compute $F(x)$.
6. Models: Measurement error, intelligence/ability, height, averages of lots of data.





B(n, p) and N(0, 1)

If x is a random variable with distribution B(n, p), then for sufficiently large n, the following random variable has a standard normal distribution:

$$z = (x - \mu)/\sigma \sim N(0,1)$$

$$\mu = np \qquad \sigma^2 = np(1-p)$$

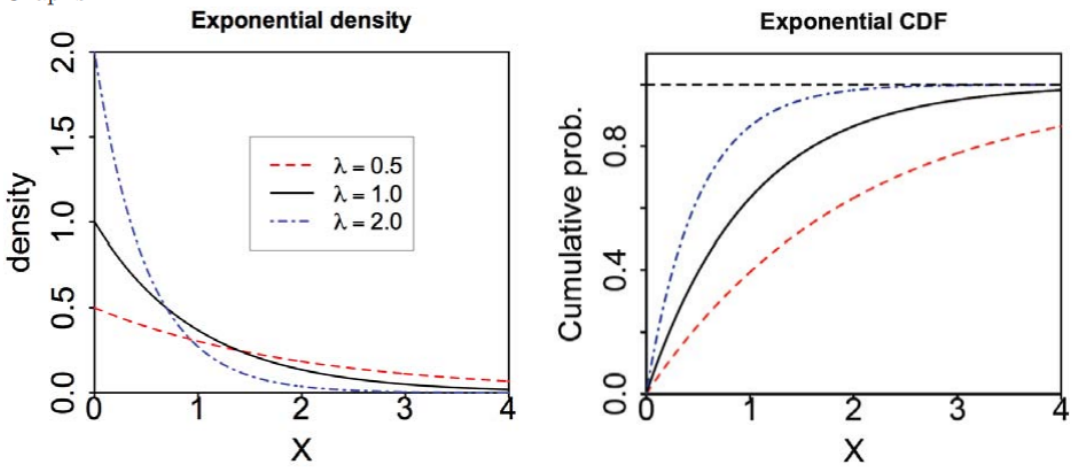
Exponential distribution

1. Parameter: λ.
2. Range: $[0, \infty)$.
3. Notation: exponential(λ) or exp(λ).
4. Density: $f(x) = \lambda e^{-\lambda x}$ for $0 \leq x$.
5. Distribution: (easy integral)

$$F(x) = 1 - e^{-\lambda x} \text{ for } x \geq 0$$

6. *Right tail distribution*: $P(X > x) = 1 - F(x) = e^{-\lambda x}$.
7. Models: The waiting time for a continuous process to change state.

Graphs:



ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

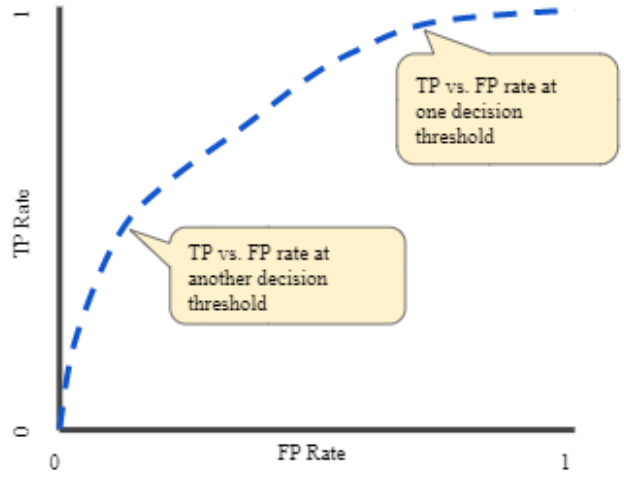
$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (**FPR**) is defined as follows:

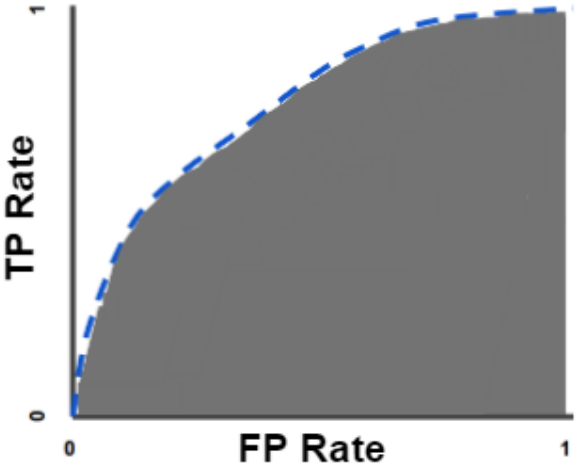
$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus

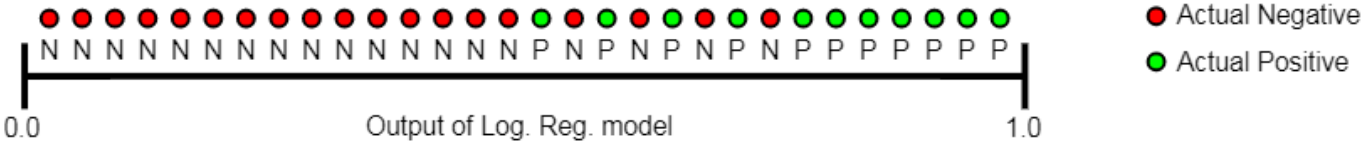
increasing both False Positives and True Positives. The following figure shows a typical ROC curve.



AUC: Area Under the ROC Curve



AUC provides an aggregate measure of performance across all possible classification thresholds.

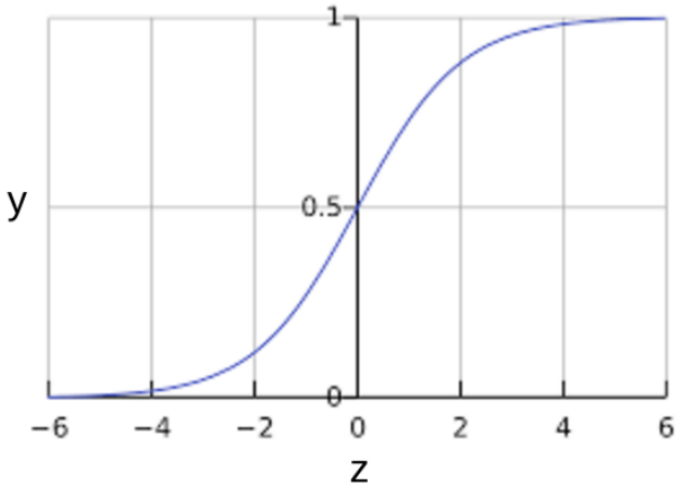


AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example.

Sigmoid activation also sigmoid function

$$y = \frac{1}{1 + e^{-z}}$$

Here z represents the output of the linear layer of a deep learning model trained with logistic regression, then sigmoid(z) will yield a value (a probability) between 0 and 1.

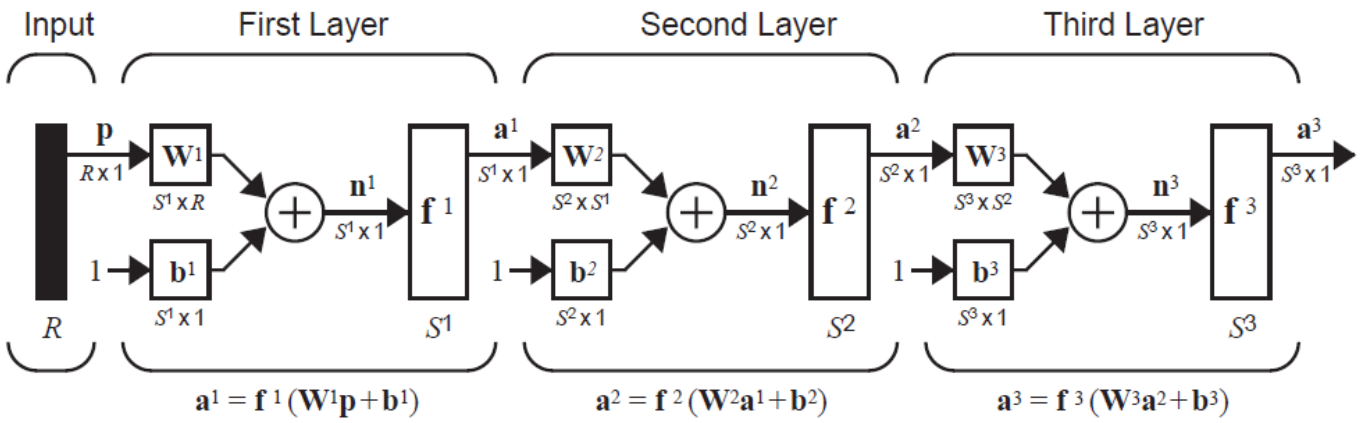


Softmax function

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$$

Softmax is often used in neural networks, to map the non-normalized output of a network to a probability distribution over predicted output classes.

Deep Neural network



References

	subject	address
1	statistics and probability	https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/

2	formula editor	http://atomurl.net/math/
3	frequent distributions	http://people.stern.nyu.edu/adamodar/New_Home_Page/StatFile/statdistns.htm
4	binomial distribution	https://stattrek.com/online-calculator/binomial.aspx
5	distribution summary	http://www.stat.tamu.edu/~twehrly/611/distab.pdf
6	life analysis	http://reliawiki.org/index.php/Life_Data_Analysis_Reference_Book
7	weibull	https://weibull.com/hotwire/issue7/relbasics7.htm https://weibull.com/hotwire/issue8/relbasics8.htm http://reliawiki.org/index.php/The_Weibull_Distribution
8	ROC and AUC	https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
9	Sigmoid activation also sigmoid function	https://developers.google.com/machine-learning/crash-course/logistic-regression/calculating-a-probability
10	Maintenance Decision Model P379	https://books.google.com/books?id=aYmrWtaVdkQC&lpg=PA384&ots=Wuk7QWT-1t&dq=NFF%20ratio&pg=PA379#v=onep
11	Least Squares and its solution derivation	http://pillowlab.princeton.edu/teaching/statneuro2018/slides/notes03b_LeastSquaresRegression.pdf
12	Confidence and Confidence Interval	http://personal.psu.edu/abs12/stat504/Lecture/lec3_4up.pdf