

Introduction to probability

Defⁿ: Experiment

An experiment is any activity from which an outcome, measurement or result is obtained. When the outcomes cannot be predicted with certainty then the expt. is a random experiment.

Defⁿ: Basic Outcomes and Sample Space

Each possible outcome of a random experiment is called a basic outcome. The set of all possible basic outcomes for a given expt. is called the sample space of the expt.

We use O_i denotes the i^{th} possible outcome

$$S = \{O_1, O_2, \dots, O_n\}$$

e.g. throw a coin $S = \{H, T\}$

throw a dice $S = \{1, 2, 3, 4, 5, 6\}$

Defⁿ: An event is a specific collection of basic outcomes, i.e. a set contains one or more basic outcomes from the sample space.

e.g. Throw a dice,

A is an event denote odd[#] $\{1, 3, 5\}$

B " " " denote eva[#] $\{2, 4, 6\}$

Basic rules of probability:

(a) probability of a basic outcome:

For each basic outcome O_i , $0 \leq P(O_i) \leq 1$.

(b) A : event

$A = \{O_1, O_2, \dots, O_k\}$ where O_i $i=1, 2, \dots, k$
are k different basic outcomes.

$$P(A) = \sum_{i=1}^k P(O_i) = P(O_1) + P(O_2) + \dots + P(O_k)$$

(c) For any event A , $0 \leq P(A) \leq 1$.

(d) Let $S = \{O_1, \dots, O_n\}$ be the sample space of an experiment, $\text{prob}(S) = 1$.

e.g. ① dice $S = \{1, 2, 3, 4, 5, 6\}$

$$A = \{\text{all odd \#}\} = \{1, 3, 5\}$$

$$P(A) = P(1) + P(3) + P(5)$$

Assume the dice is unbiased, i.e. $P(i) = \frac{1}{6}$, $i=1, \dots, 6$

Then we have $P(A) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

② Quality control divisions of an electronic firm tests a product for defects. Basic outcomes are $O_1 = \text{defective}$, $O_2 = \text{good}$

$$S = \{O_1, O_2\} = \{\text{defective, good}\}$$

Suppose the past experience indicates that 90% of the products produced are good. Show that the sample space has a probability of 1.

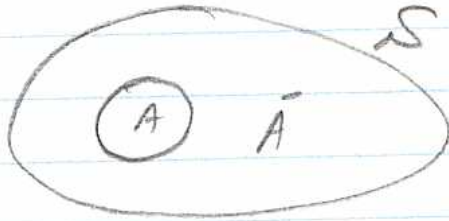
$$P(\text{defective}) = 0.1, \quad P(\text{good}) = 0.9$$

$$P(S) = P(\text{defective}) + P(\text{good}) = 1.0$$

Probability of Compound events:

Defⁿ: Complement of an event

Let A denote an event in sample space S .
The complement of A denoted by \bar{A} represents the event composed of all basic outcomes in that S that do NOT belong to A .



e.g. $S = \{1, 2, 3, 4, 5, 6\}$ of throwing a dice

$A = \{ \text{odd } \# \}$, $\bar{A} = \{ \text{even } \# \}$

$$P(A) + P(\bar{A}) = 1.0$$

Defⁿ: Union of two events

A, B are two events in sample space S

$A \cup B$ is an event that composed of all basic outcomes in S that belong to at least one of the two events A or B . Hence $A \cup B$ occurs if either A or B (or both) occurs.

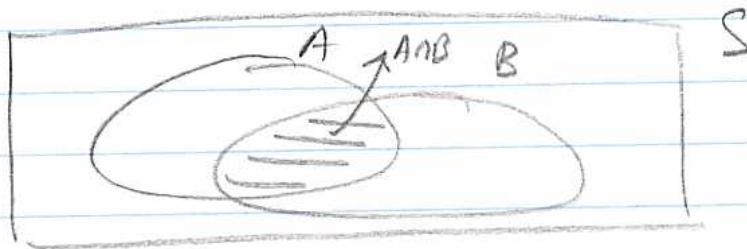
Defⁿ: Intersection of 2 events

$A \cap B$ is an event that composed of all basic outcomes that belong to both A and B.

Hence $A \cap B$ occurs if Both A and B occur.

Additive Law of probability

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



e.g. At a university, all 1st year students must take Chemistry and Mathematics

15% fails Chemistry

12% fail Mathematics

5% fail BOTH.

Suppose a first year student is picked at random. Find probability that the student failed at least one of the two courses.

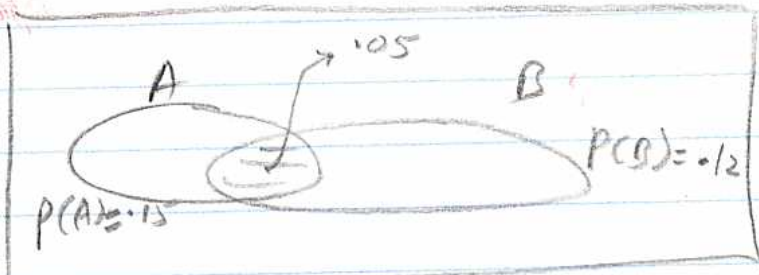
Sol:

$A = \{ \text{person fails in chemistry} \}$

$B = \{ \text{person fails in math.} \}$

$A \cap B = \{ \text{person fails in both} \}$

$A \cup B = \{ \text{person fail math, chem, or both} \}$



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

$$= 0.15 + 0.12 - 0.05 = 0.22$$

e.g. A small bank has 1000 customers having both chequing & saving accounts.

Average monthly balances:

		\bar{A} Saving account	A	
		≤ 1000	> 1000	
Chequing account	\bar{B} ≤ 500	300	400	700
	B > 500	200	100	300
		500	500	1000

A : saving account > 1000

B : chequing account > 500

Find $P(A)$, $P(B)$, $P(A \cup B)$, $P(\bar{A} \cup \bar{B})$, $P(A \cap B)$

Sol.:

$$P(A) = \frac{500}{1000} = 0.5, \quad P(B) = \frac{300}{1000} = 0.3$$

$$P(A \cap B) = \frac{100}{1000} = 0.1$$

$$P(A \cup B) = 0.5 + 0.3 - 0.1 = 0.7$$

$$P(\bar{A} \cup \bar{B}) = \frac{300 + 200 + 400}{1000} = 0.9$$

Defⁿ: Mutually Exclusive Events

Two events A and B are mutually exclusive events if they do not have any outcome in common and, consequently, cannot occur simultaneously. It follows, therefore, that

$$P(A \text{ and } B) = P(A \cap B) = P(\emptyset) = 0 \text{ for}$$

mutually exclusive events A and B.

By the addition law of probability, we have the following:

$$\begin{aligned} P(A \text{ or } B) &= P(A \cup B) = P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - 0 \\ &= P(A) + P(B) \text{ if } A \text{ \& } B \end{aligned}$$

are mutually exclusive.

e.g. A box contains 3 red, 4 green & 5 white balls. One ball is picked at random. What is the prob. that it will be red or white?

Soln: R: A red ball is picked

W: A white ball is picked

The events R, W are mutually exclusive, because if a ball is red it cannot be white & vice versa.

Therefore $P(R \cap W) = 0$.

Conditional Probability: In many applications, we may be interested in determining the prob. that some event A occurs given that some other event B has already occurred. This prob. is called a conditional probability, if the prob. of one event varies depend on whether 2nd event has occurred, these two events are said to be dependent.

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

	\sum_{col}	N
P		
F		

e.g. $A = \{ \text{college degree} \}$
 $B = \{ \text{making} \geq 50K \text{ per year} \}$

$P\{ \text{making} \geq 50K \text{ per year} \} \neq P\{ \text{a person makes} \geq 50K \text{ per year} \mid \text{given that the person is college graduate} \}$

i.e. $P\{ \geq 50K \text{ per year} \} \neq P\{ \geq 50K \text{ per year} \mid \text{college graduate} \}$

Joint prob. & Marginal prob.

Defⁿ: A joint probability shows the probability that an observation will possess two (or more) characteristics simultaneously. Every joint probability must a number $\in [0, 1]$ and the sum of all joint prob. is 1.0.

Defⁿ: Marginal probability shows the proportion of observations that possess any specific characteristic.

e.g. Admission status to a graduate school of a university.

		Admitted	Rejected	
Gender	M	3800	4700	8500
	F	1600	2400	4000
		5400	7100	12500

table 1

Joint prob. table

	Admitted	Rejected	Marginal prob.
M	0.304	0.376	0.680
F	0.128	0.192	0.320
Marginal prob.	0.432	0.568	1.000

Gender discrimination ??

table 2

$$P\{\text{Male} \cap \text{Rejected}\} = \frac{4700}{12500} = 0.376$$

$$P\{\text{Female} \cap \text{Rejected}\} = \frac{2400}{12500} = 0.192$$

Marginal probability

$$\begin{aligned} P\{\text{admitted}\} &= P\{\text{Adm} \cap \text{male}\} + P\{\text{Adm} \cap \text{Female}\} \\ &= 0.304 + 0.128 = 0.432 \end{aligned}$$

$$P\{\text{Rejected} | \text{male}\} = \frac{P\{\text{rejected} \cap \text{male}\}}{P(\text{male})} = \frac{0.376}{0.680} = 0.553$$

or just simply looked at table 1

$$P\{\text{Rejected} | \text{male}\} = \frac{4700}{8500} = 0.553$$

$$P\{\text{rejected} | \text{Female}\} = \frac{2400}{4000} = 0.60$$

Independent events

A & B are independent $\Leftrightarrow P(A \cap B) = P(A)P(B)$

e.g. 30% of new employees hired in a company quit within 1 year. Suppose 2 employees are hired and the first employee's behavior is independent of 2nd employee's behavior

$$P\{\text{both quit}\} = P(A) \cdot P(B) = 0.3 \times 0.3 = 0.09$$

Sensitivity and Specificity

		Disease	
		Present	Absent
Test	+ve	a	b
	-ve	c	d

$$\text{sensitivity} = \text{True +ve} = \frac{a}{a+c} = P\{\text{+ve} \mid \text{Disease}\}$$

$$\text{specificity} = \text{True -ve} = \frac{d}{d+b} = P\{\text{-ve} \mid \text{No Disease}\}$$

$$PV^+ = \text{Predictive value +ve} = P\{\text{Disease present} \mid \text{test +ve}\} = \frac{a}{a+b}$$

$$PV^- = \text{Predictive value -ve} = P\{\text{Disease absent} \mid \text{test -ve}\} = \frac{d}{c+d}$$

Defⁿ: Sensitivity is defined as the probability of testing positive if the disease is truly present.

Defⁿ: Specificity is defined as the probability of testing negative if the disease is truly absent.

Altering the criterion of positivity or abnormality will influence both the sensitivity and specificity of the test. Lowering or making less stringent the criterion of positivity will mean that more people who actually have the disease will be test positive (increased sensitivity), but so will a number of people who do not have the disease (decreased specificity). Conversely, making the criterion more stringent will mean that a greater proportion of those who test negative will actually not have the disease (increased specificity), but a larger number of true cases will also be missed (decreased sensitivity).

e.g. If in a hypertension screening program the criterion of positivity for dbp we set low, at 88 mm Hg very few people with actual hypertension would be missed but many normotensive individuals will be falsely labeled as hypertensive. Thus this test would be very sensitive but non-specific. If on the other hand, the criterions of positivity were set higher, at 100 mm Hg for example, exactly the opposite

would occur, in that a higher proportion of those hypertension would test -ve, but so would a greater proportion of persons with the disease.

Any decision regarding specific criteria for acceptable levels of sensitivity and specificity in a given situation involves weighing the consequences of leaving cases undetected (false negatives) against erroneously classifying healthy persons as diseased (false positives).

e.g.	Test	Disease		Total
		present	absent	
	+ve	900 (.009)	4950 (.0495)	5850
	-ve	100 (.001)	94050 (.9405)	94150
		1000 (.01)	99000 (.99)	100000

$$\text{sensitivity} = P\{+ve \mid \text{Disease}\} = \frac{P(+ve \cap \text{Disease})}{P(\text{Disease})}$$

$$= \frac{.009}{.01} = 90\%$$

$$\text{specificity} = P\{-ve \mid \text{NO Disease}\} = \frac{P(-ve \cap \text{ND})}{P(\text{ND})}$$

$$= \frac{.9405}{.99} = 95\%$$

$$PV^+ = P\{D \mid \text{test +ve}\} = \frac{P(D \cap \text{test +ve})}{P(\text{test +ve})}$$

$$= \frac{.009}{.0585} = 15.4\%$$

$$PV^- = P\{\overset{ND}{\text{no disease}} \mid \text{test -ve}\} = \frac{0.9405}{0.9415} = 99.9\%$$

ODD RATIO

Dependent Variable	Risk Present	Factor Absent
	Success	a
Failure	c	d
TOTAL	a+c	b+d
Proportion of success (success rate)	$\frac{a}{a+c}$	$\frac{b}{b+d}$

The Odd ratio (OR, Cornfield 1951) is the ratio of two odds, namely the odds of success when the risk factor is present and the odds of success when the risk factor is absent.

In terms of probabilities,

$$OR = \frac{\left[\frac{P(\text{success}|\text{Present})}{P(\text{failure}|\text{Present})} \right]}{\left[\frac{P(\text{success}|\text{absent})}{P(\text{failure}|\text{absent})} \right]}$$

from the table $P(\text{success}|\text{Present})$ and $P(\text{failure}|\text{Present})$

may be estimated by $\frac{a}{a+c}$, $\frac{c}{a+c}$ respectively.

Similarly, $P(\text{success}|\text{absent})$ and $P(\text{failure}|\text{absent})$

may be estimated by $\frac{b}{b+d}$ and $\frac{d}{b+d}$ respectively.

$$\therefore OR = \frac{(a/c)}{(b/d)} = \frac{ad}{bc}$$

which is also referred to as the cross-product ratio.

An odd ratios of 1 indicates that the odds of success is the same no matter whether the risk factor is present or not. An odd ratio greater than 1 means that the odds in favor of success is higher when the risk factor is present, and therefore there is a positive association between the risk factor and success. Similarly, an odd ratio less than 1 signifies a negative association between the risk factor & success. The interpretation should not be totally based on the point estimate. A confidence interval is always more meaningful, just like in other estimation procedure.

e.g. In the study of the response rate of 71 leukemia patients, age is considered as one of the possible risk variables. The following 2×2 table is constructed:

	Age < 50	Age \geq 50	TOTAL
Response	27 (69%)	10 (31%)	37
Non response	12	22	34
TOTAL	39	32	71

3.14

The estimated odds ratio, is

$$OR = \frac{27 \times 22}{10 \times 12} = 4.95$$

The data show that the odds in favor of response are almost five times higher in patients under 50 yrs. of age than in patients at least 50 yrs. old.

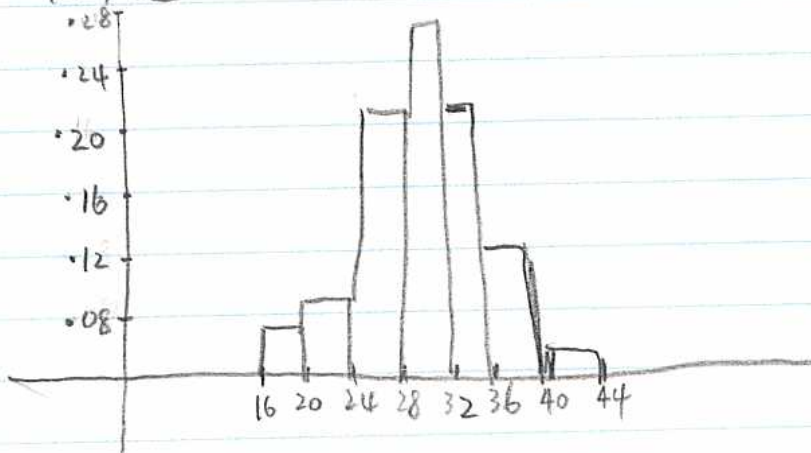
$$PV^- = P\{D \mid \text{test -ve}\} = \frac{P(D \cap \text{test -ve})}{P(\text{test -ve})}$$

$$= \frac{0.9405}{.9415} = 99.9\%$$

Representing probabilities by areas

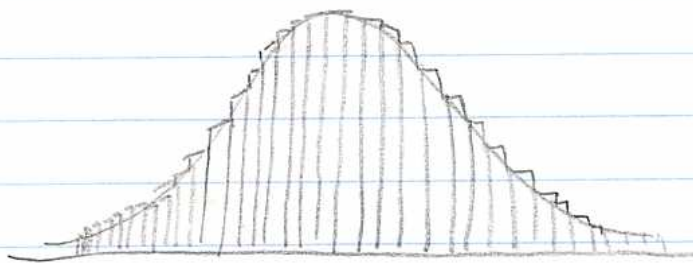
We select a random sample of 100 individuals and record the time required for each individual to perform a certain task. We have the following table:

Time	Freq.	relative freq.
16 - under 20	6	.06
20 - under 24	9	.09
24 - under 28	22	.22
28 - under 32	27	.27
32 - under 36	21	.21
36 - under 40	12	.12
40 - under 44	3	.03
TOTAL	100	1.00



Now suppose we take a sample of 100,000 individuals then construct a relative frequency distribution with a ~~very~~ large # of very narrow classes, say .01 or 0.001. The graph of the resulting relative frequency distribution will contain a large # of very narrow rectangles.

As # of observations become very large & the class intervals become very narrow, the shape of relative freq. histogram approaches a smooth curve.



For continuous variables, probability are measured by areas under the density function. The probability that the random variable X falls in any particular interval (a, b) is the area under the curve between points a and b .

Defⁿ: Density functions

Let a smooth curve represent the prob. dist. of a continuous r.v. X , & let the smooth curve be represented in math. notation by the function $f(x)$. The function $f(x)$ is called the density function of the continuous random variable X where x represents a specific value of the random variable X .

3.17
~~3.14~~

Characteristic of Density function

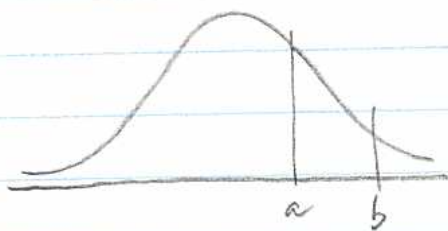
All density functions must satisfy the following 2 requirements:

(1) The curve must never fall below the horizontal axis. That is

$$f(x) \geq 0 \quad \forall x$$

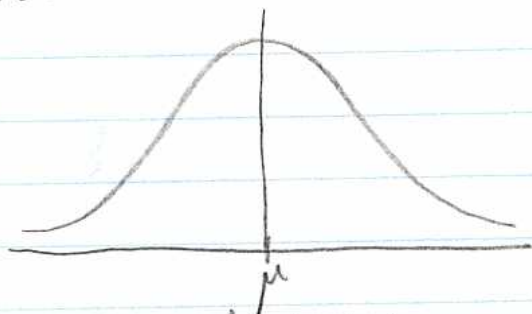
(2) The total area between the curve and the horizontal axis must be 1.

$$\int_{-\infty}^{\infty} f(x) dx = 1$$



$$P\{a \leq X \leq b\} = P\{a < X < b\}$$

Normal distribution



Normal distribution, first used in 1733 by Antoine de Moivre (1667-1745) is the most important statistical distribution. It is sometimes called Gaussian distribution in honour of Carl F. Gauss (1777-1855) a famous German mathematician who did extensive work with it.

The normal curve is represented by the density function.

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

$$\pi : 3.14159,$$

$$e = 2.71828$$

x is a specific value of random variable $X \sim N(\mu, \sigma^2)$

Characteristics.

- (1) Curve bell shaped & symmetric about the value $x = \mu$.
- (2) Curve extends from $-\infty \rightarrow \infty$
- (3) Area under the curve is 1
- (4) Curve is always above x -axis
- (5) mean, mode and median = μ

Empirical rule

When a distribution is approximately normal,
68% of values $\in (\bar{x} - s.d., \bar{x} + s.d.)$

95% of values $\in (\bar{x} - 2sd, \bar{x} + 2sd)$

99% of values $\in (\bar{x} - 3sd, \bar{x} + 3sd)$

3.19

Standardized scores

sample : $\bar{x} = \frac{\sum x_i}{n}$, $S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

population : $\mu = \frac{\sum x_i}{N}$, $\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$

Observations in which the standard deviation is the unit of measurement are called standard scores. The standard score or z-score corresponds to a particular observation in a set of data is simply the number of standard deviation units that the observation is away from the mean, with a -ve value indicating that the observation is below the mean and a +ve value indicating the observation is above the mean.

The standard score z_i from an observation x_i from a sample with a mean of \bar{x} and s.d. of s ,

$$z_i = \frac{x_i - \bar{x}}{s}$$

The standard score z_i from an observation x_i from a population with a mean μ and a s.d. of σ ,

$$z_i = \frac{x_i - \mu}{\sigma}$$

e.g. Here is a random sample of size 6, the scores are:

Subj	X	$x - \bar{x}$	Z
1	3	-7	-1.11
2	6	-4	-.63
3	7	-3	-.47
4	9	-1	-.16
5	15	5	.79
6	20	10	1.58
Sum	60	0	0
mean	10	0	0
s.d.	6.32	6.32	1