Introduction to probability

Def^n: 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.

Def^n: 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 $S = \{1, 2, 3, 4, 5, 6\}$

Def^n: 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 even$^{\#}$ $\{2, 4, 6\}$

Basic rules of probability:

(a) probability of a basic outcome:

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

(b) A: event

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

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

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

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

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

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

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

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

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

② Quality control division of an electronic firm tests a product for defects. Basic outcomes are O_1 = defective, O_2 = good

$S = \{O_1, O_2\} = \{$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.

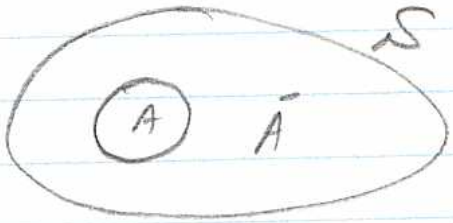
$P(\text{defective}) = 0.1$, $P(\text{good}) = 0.9$

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

Probability of compound events:

Def^n: 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

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

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

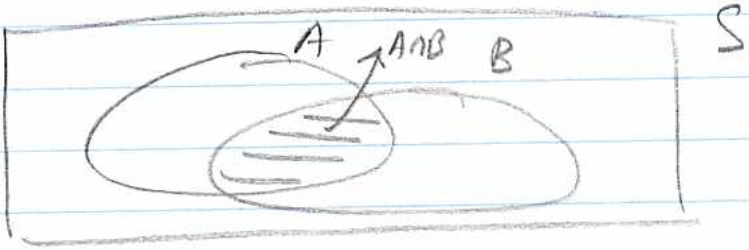
Def^n: Union of two events

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.

Def^n: 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

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



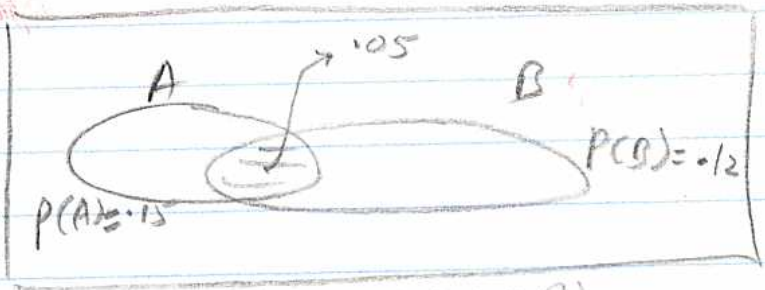
e.g. At a university, all 1st year students must take Chemistry and Mathematics

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 = { person fails in chemistry }
B = { person fails in Math. }
$A \cap B$ = { person fails in both }
$A \cup B$ = { person fail math, chem, or both }



$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$= 0.15 + .12 - .05 = 0.22$

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

Average monthly balances:

		$\bar{A}$ Saving account ≤ 1000	A > 1000	
$\bar{B}$ Chequing account	≤ 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.:

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

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

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

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

Def^n: 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

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

mutually exclusive events A and B.

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

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

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

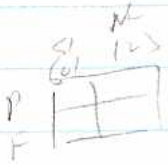
W: A white ball is picked

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 application, we may be interested in determining the prob. that some event A occurs given that some other event B has already occured. This prob. is called a conditional probability. If the prob. of one event varies depend on whether 2nd event has occured, these two events are said to be depend

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



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

$P\{$ making $\geq 50k$ per year$\} \neq P\{$ a person makes $\geq 50k$ per yr 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.

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

Def^n: Marginal probability shows the proportion of observation that possess any specific characteristic.

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

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

table 1

Joint prob. table

	Admitted	Rejected	Marginal prob.
M	.304	.376	.680
F	.128	.192	.320
Marginal prob.	.432	.568	1.000

table 2

Gender discrimination ??

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

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

Marginal probability

$$P\{\text{admitted}\} = P\{\text{Adm} \cap \text{male}\} + P\{\text{Adm} \cap \text{Female}\}$$

$$= 0.304 + 0.128 = 0.432$$

$$P\{\text{Rejected} \mid \text{male}\} = \frac{P\{\text{rejected} \cap \text{male}\}}{P(\text{male})} = \frac{0.376}{.680} = .55$$

or just simply looked at table 1

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

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

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 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
		$a+c$	$b+d$

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

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

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

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

Def^{n}: Sensitivity is defined as the probability of testing positive if the disease is truely present.

Def^{n}: Specificity is defined as the probability of testing negative if the disease is truely 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 normatensive individuals will be falsely labeled as hypertensive. Thus this test would be very sensitive but non-specific. If on the other hand, the criterion 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 (i.e. false negatives) against erroneously classifying healthy persons as diseased (false positives).

e.g.

Test	Disease present	Disease absent	Total
+ve	900 (.009)	4950 (.0495)	5850
-ve	100 (.001)	94050 (.9405)	94150
	1000 (.01)	99 000 (.99)	100 000

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

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

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

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

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

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

$$PV^- = P\{ND \mid \text{test -ve}\} = \frac{0.9405}{0.9415} = 99.9\%$$

DD RATIO

	Risk Factor Present	Factor Absent
Dependent Vaiable		
Success	a	b
Failure	c	d
TOTAL	a+c	b+d
Proportion of success (success rate)	$a/a+c$	$b/b+d$

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,

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

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.

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

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

An odd ratio 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 x 2 table is constructed:

	Age <50		Age ≥50	TOTAL
Response	27	(69%)	10 (31%)	37
Nonresponse	12		22	34
TOTAL	39		32	71

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.

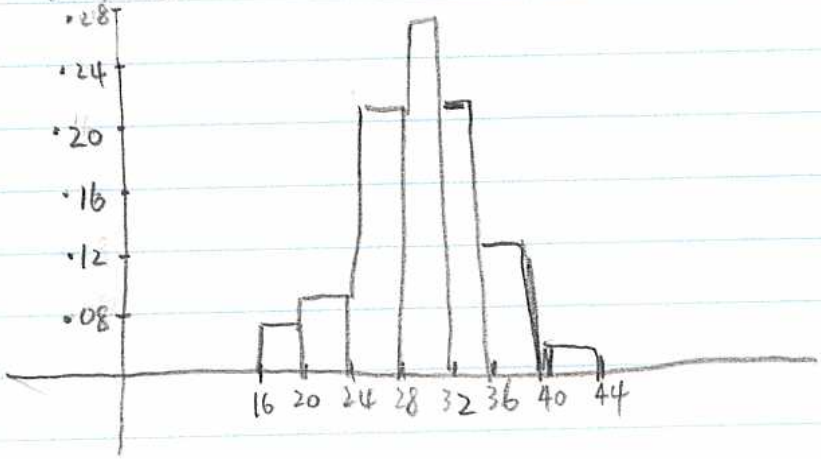
$$PV^- = P\{D \mid \text{test } -ve\} = \frac{P(D \cap \text{test } -ve)}{P(\text{test } -ve)}$$

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

Representing probabilities by areas

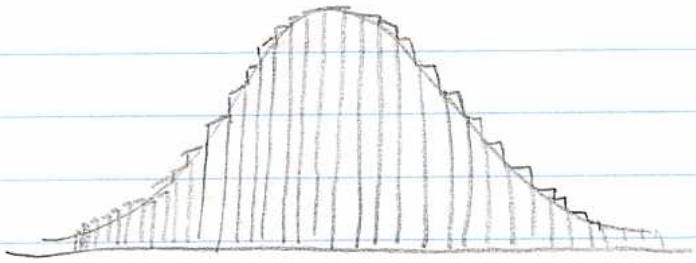
We select a random sample of 100 individuals and record the time required for each individual to perform a certain task. We have the following table:

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 rectangle.

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.

Def^{n}: Density function
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.

Characteristic of Density function

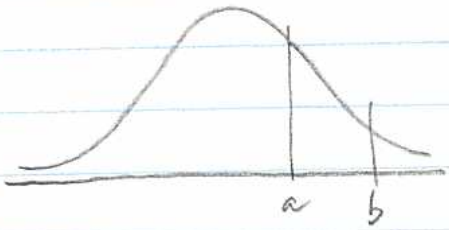
All density functions must satisfy the following 2 requirements:

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

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

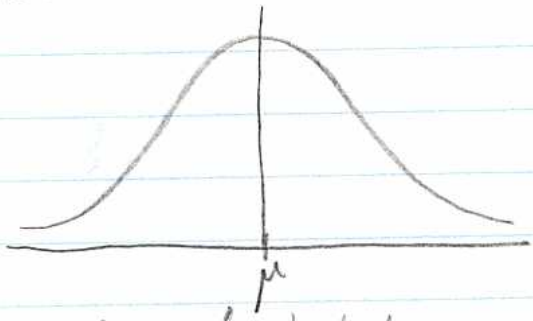
(2) The total area between the curve and the horizontal axis must be 1.

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



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

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.

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

π : 3.1415-9,
e = 2.71828
x is a specific value of random variable $X \sim N(\mu, \sigma^2)$

Characteristic.
(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 $= \mu$

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 $\in (\bar{x} - 3sd, \bar{x} + 3sd)$

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

subj = 1	1.11	s.d. below the mean
subj = 2	.63	" " " "
s.j 6	1.58	s.d. above the mean

Standard scores are frequently used to obtain comparability of observation by different procedures.

e.g.

1st yr. Calculus exam. scores of 100 students $\mu = 65$, $\sigma = 8$

" " algebra " " of the same 100 stds $\mu = 52$, $\sigma = 3$

student-tak got Calculus 58, z-score $= \frac{58-65}{8} = -.875$

student-tak got algebra 55, z-score $= \frac{55-52}{3} = 1$

Clearly tak did much poorly in Calculus than Algebra relative to the group of student taking the exam, although this is NOT reflected in the original marks assigned.

Because z-scores results in -ve, some people prefer to transform them into other distribution.
One distribution that has been widely used is one with a mean of 50 and a s.d. of 10
Such transformed standard score are generally called T-scores. To convert a z-score to a T-score, use $T = 10z + 50$

e.g. with a z-score of 2.5, the T-score would be

$$T = 10 \times 2.5 + 50 = 75$$

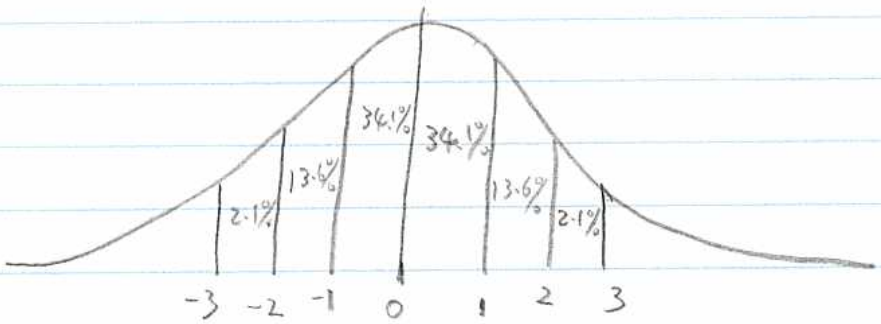
In the new distibution, the mean is 50 and s.d. is 10
a score of 75 is still 2.5 s.d. above the mean.

Standard Normal Distribution

A random variable is said to have the standard Normal Distribution if it has the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ denoted as $N(0, 1)$.

It is common to denote the standard Normal Distribution by Z rather than X.

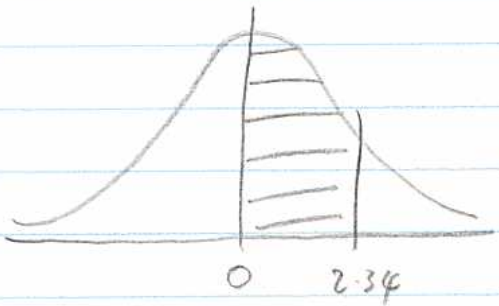
The standard Normal distribution & its area



Calculating Area under the standard Normal Curve

e.g. Area between a mean & a +ve value under the Standard Normal Curve:

Find area under the Standard normal curve between 0 and 2.34. i.e Find $P\{0 \leq Z \leq 2.34\}$.

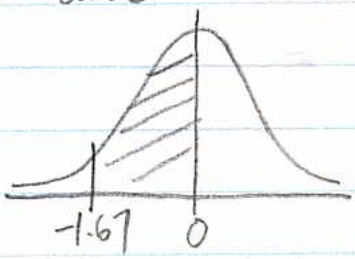


Move down the left-side of the table to the row value of 2.3. Now move across to the hundredth column headed by the digit 4. The number found at the intersection of this row and the column is .4904 which represents the area between 0 and 2.34. Thus we have $P(0 \leq Z \leq 2.34) = .4904$

e.g. Area between the mean and a -ve value under the standard Normal curve:

Find area under the standard normal curve between 0 and -1.67

Sol.



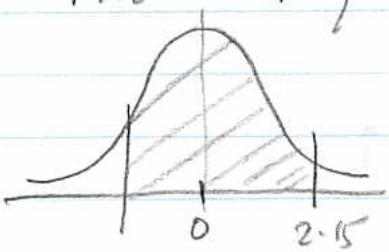
We want to find the shaded area. Because the normal distribution is symmetric, the area between -1.67 & 0 is the same as area between 0 and 1.67.

From table $P(0 \leq Z \leq 1.67) = 0.4525$

$\therefore P(-1.67 \leq Z \leq 0) = 0.4525$.

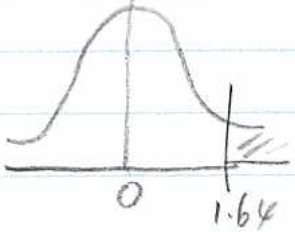
e.g. Area between a -ve and a +ve value under standard Normal Curve $Z \sim N(0,1)$

Find $P\{-1.21 \leq Z \leq 2.15\}$.



$P\{-1.21 \leq Z \leq 2.15\}$
$= P\{-1.21 \leq Z \leq 0\} + P\{0 \leq Z \leq 2.15\}$
$= 0.3869 + 0.4842 = 0.8711$

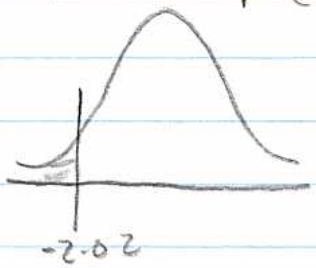
e.g. Area in the right-hand tail of the standard Normal curve. Find $P(Z > 1.64)$



$P(Z > 1.64) = 0.5 - P(0 \leq Z \leq 1.64)$
$= 0.5 - 0.4495 = 0.0505$

e.g. Area in the left hand tail of the standard normal curve.

Find $P(Z \leq -2.02)$

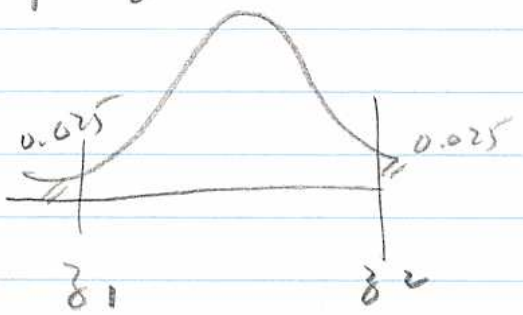


$$P(0 \leq Z \leq 2.02) = P(-2.02 \leq Z \leq 0) = 0.4783$$

$$\therefore P(Z \leq -2.02) = 0.5 - 0.4783 = 0.0217.$$

e.g. Find a z-score associated with a specific area

Find the values of z_1 and z_2 such that the area of right of z_2 is 0.025 and area the left of $z_1 = 0.025$



From table,

$z_2 = 1.96$,

$z_1 = -1.96$

e.g. Area under $N(\mu, \sigma^2)$

$X \sim N(10, 25)$

Find $P\{12 \leq X \leq 16\} = P\left\{\frac{12-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{16-\mu}{\sigma}\right\}$

$$= P\left\{\frac{12-10}{5} \leq z \leq \frac{16-10}{5}\right\} = P\{0.4 \leq z \leq 1.2\}$$

$$= 0.2295$$