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

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

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, $P(S) = 1$.

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. $P(i) = \frac{1}{6}$, $i = 1 \cdots$

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

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.

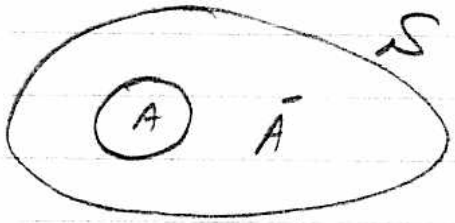
$P(\text{defective}) = 0.1$, $P(\text{good}) = 0.9$

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

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

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

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

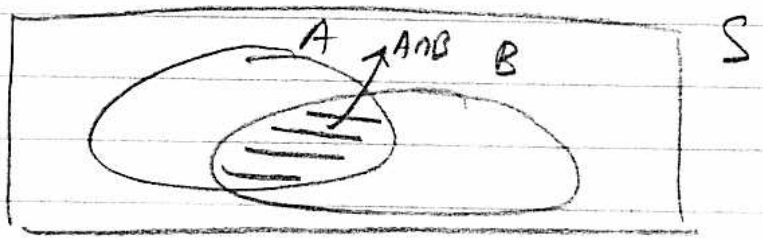
Def^n: Union of two events

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.

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

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



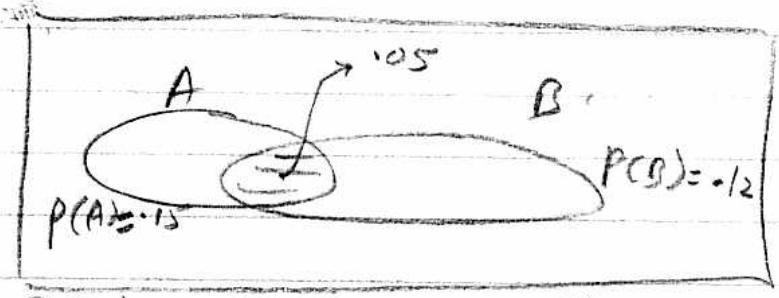
e.g. At a university, all 1st year students must take Chemistry and Mathematics

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:

$A = \{$ 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

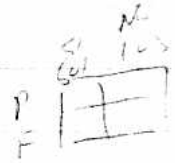
W: A white ball is picked

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 applications, 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 depending on whether 2nd event has occured, these two events are said to be dependent

1. 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 year 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

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

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

$= 0.304 + 0.128 = 0.432$

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

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

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

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 hypertensive 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

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

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

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

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

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

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

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

DD RATIO

	Risk Factor Present	Risk Factor Absent
Dependent Variable		
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 $\frac{b}{}$ and $\frac{d}{}$ respectively

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

which is also refered 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 2x2 table is constructed:

	Age <50		Age ≥50	TOTAL
Response	27	(69%)	10 (31%)	37
Non response	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.

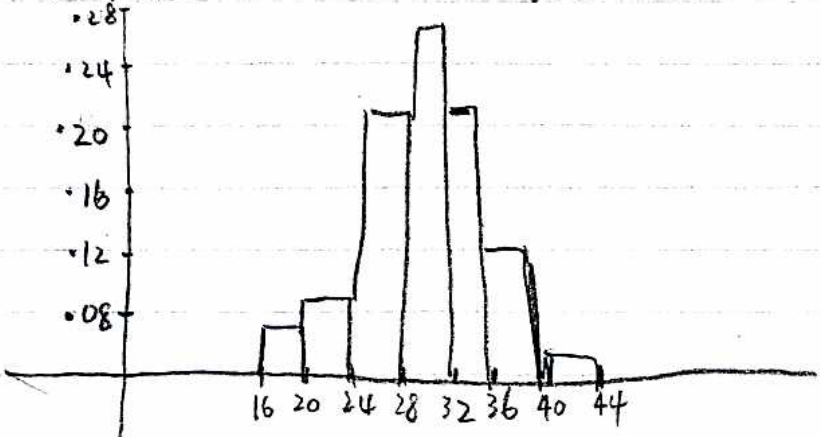
$$PV^- = P\{D \mid \text{test } -ve\} = \frac{P(D \cap \text{test } -ve)}{P(\text{test } -ve)}$$

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

Representing probabilities by areas

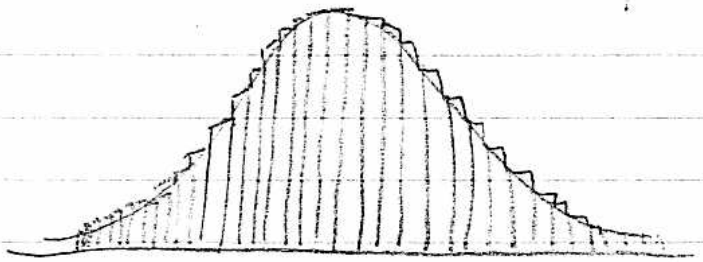
We select a random sample of 100 individuals and record the time required for each individual to perform a certain task. We have the following table:

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

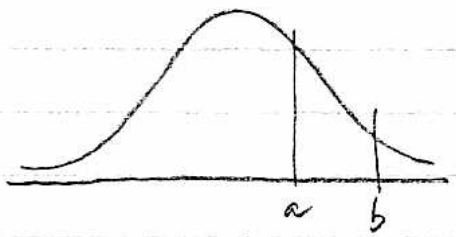
All density functions must satisfy the following 2 requirements.

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

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

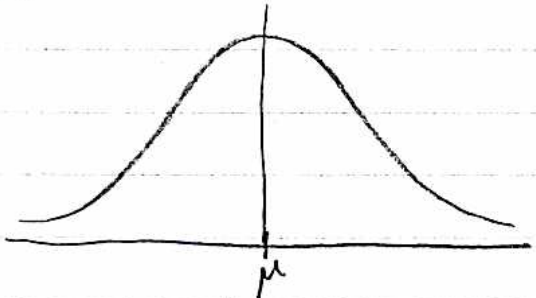
(2) The total area between the curve and the horizontal axis must be 1.

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



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

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 : 3.14159,
e = 2.71828
x is a specific value of random variable $X \sim N(\mu,$

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	"	"	"	"
subj 6	1.58	s.d.	above	the	mean

Standard scores are frequently used to obtain comparability of observations 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

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

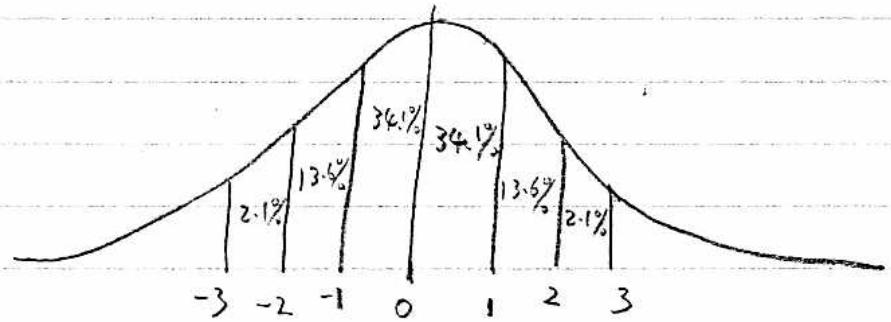
In the new distibution, the mean is 50 and s.d. is 10
a score of 75 is still 2.5 s.d. above the mean

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.

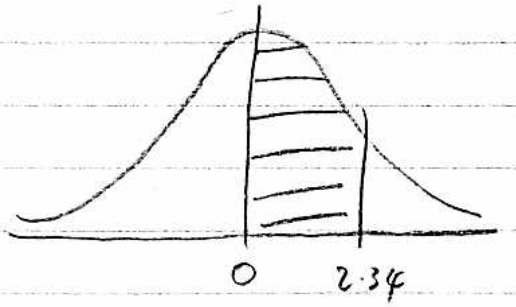
The standard Normal distribution & its area



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 $P\{0 \leq Z \leq 2.34\}$.



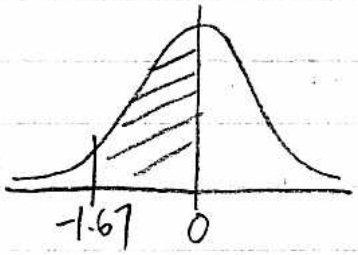
Move down the left-side of the table to the row value of 2.3. Now move across to the hundredth column headed by the digit 4.

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

S.N.



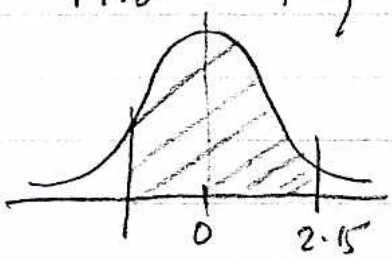
We want to find the shaded area. Because the normal distribution is symmetric, the area between -1.67 & 0 is the same as area between 0 and 1.67.

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

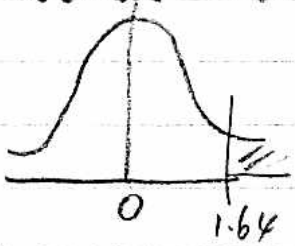


$P\{-1.21 \leq Z \leq 2.15\}$

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

$= 0.3869 + 0.4842 = 0.8711$

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

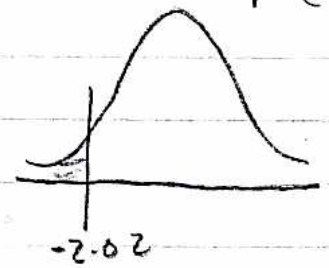


$P(Z > 1.64) = 0.5 - P(0 \leq Z \leq 1.64)$

$= 0.5 - 0.4495 = 0.0505$

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

Find $P(Z \leq -2.02)$

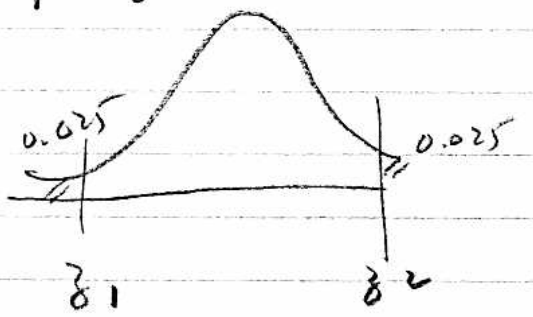


$$P(0 \leq Z \leq 2.02) = P(-2.02 \leq Z \leq 0) = 0.4783$$

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

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

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

$$= 0.2295$$

Chapter 4

Sampling distribution

Statistics is the science of answering questions on the basis of empirical data. Often, the question is such that it is impossible to obtain all of the relevant observations (the population) and a conclusion must be drawn on the basis of a sample of observations.

Sampling refers to the process of selecting a sample from a population. From the sample, we calculate statistics (numerical summaries of the sample) which used as estimates of the corresponding parameters (numerical summaries of the population).

The nature of Sampling Distribution

Sampling involves the selection of a sample ~~sample~~ from which statistics are calculated to estimate the corresponding parameters of the population. Once an estimate have been obtained, it is reasonable to ask how close the estimate is to the true value. What is the probability that estimate differs from the unknown parameter by less than so & so many units.

Def^{n}: A sampling distribution of a statistics is a description of the probabilities with which various values or set of values of the statistic occur.

In other words, a sampling distribution is a particular kind of distribution, namely, a probability distribution in which the random variable is itself a statistic (numerical description of a sample). Thus, we may inquire about the sampling distribution of sample standard deviation, sampling distribution of the sample mean or the sampling distribution of other statistic. In each case, the sampling distribution relates the possible values of the statistic to their prob. of occurance.

At this point you may object: How can we have a distribution of something like a sample mean, where in practice just one sample whose mean has just one value? How can we have a distribution of the sample mean or any other statistics? Perhaps it is partly the terminology that is confusing here. A distribution of the sample mean is actually a distribution of the possible values of the sample mean. While it is true that a given sample provides only one sample mean, it must be remembered that the given sample occured by chance, each one of which would have yielded its own sample mean.

4.3

9.2 • Sampling Distribution of the Sample Mean

It is important to distinguish between the distribution of the population of values X and the sampling distribution of the sample mean $\bar{X}$. The following example utilizes sampling without replacement and illustrates how to obtain the sampling distribution of $\bar{X}$.

Example 9.3

The Sampling Distribution of $\bar{X}$: The data in Table 9.1 show 200 random digits. Consider these values to be a population. The population mean for these values is $\mu = 4.62$. Table 9.2 shows the frequency and relative frequency distributions for the data in Table 9.1, and Figure 9.3 shows the relative frequency histogram.

TABLE 9.1
200 random digits

1	2	6	5	1	6	1	6	4	6	1	1	7	6	9	7	5	1	0	9
8	6	9	9	6	9	7	6	6	9	2	5	7	5	7	3	2	5	3	5
0	7	1	2	2	7	6	7	6	3	8	1	7	6	9	7	4	4	3	6
0	2	6	3	0	7	2	3	1	0	4	5	0	4	9	1	8	0	2	9
0	7	4	6	9	4	2	3	4	1	9	8	1	7	3	7	9	2	6	0
3	6	7	3	7	9	8	8	6	3	7	7	2	4	0	7	6	2	5	1
0	0	6	5	4	6	4	6	8	8	0	9	3	4	3	7	0	2	7	8
6	7	3	3	1	9	8	7	2	9	8	2	8	6	1	5	4	3	7	1
7	6	6	1	0	9	4	9	3	4	7	2	7	4	8	4	4	1	2	4
0	5	6	1	0	5	3	7	5	0	9	5	9	3	8	0	1	4	8	5

TABLE 9.2 **Distribution of data in Table 9.1**

X	0	1	2	3	4	5	6	7	8	9
Frequency	19	20	17	19	20	15	27	28	15	20
Relative frequency	.095	.100	.085	.085	.100	.075	.135	.140	.075	.100

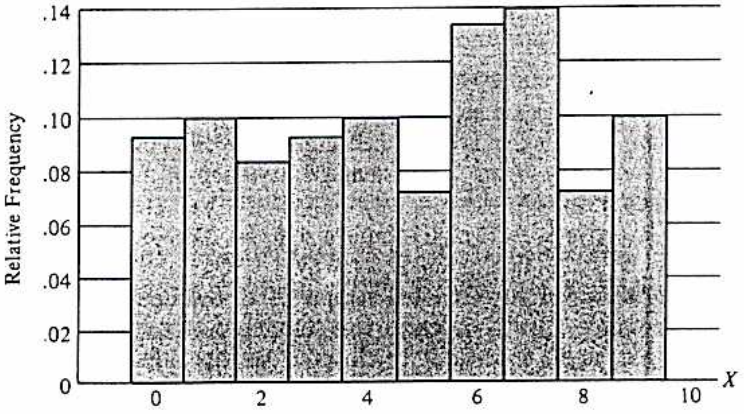


FIGURE 9.3
Relative frequency histogram of data in Table 9.1.

In Table 9.1 the values $x_1, x_2, \ldots, x_{20}$ are in the first row; $x_{21}, x_{22}, \ldots, x_{40}$ are in the second row; and so forth. To get information about the sampling distribution of the sample mean, we took 80 different samples of size $n = 10$ and calculated the sample mean $\bar{x}$ for each sample. The first sample, which contained the first 10 observations in the first row of Table 9.1, had a sample mean of 3.8. The second sample contained the last 10 observations in the first row of Table 9.1 and had a sample mean of 4.6. An additional 78 samples of size 10 were obtained, and the sample mean was calculated for each sample. The 80 values of $\bar{X}$ obtained in this fashion are shown in Table 9.3. Note that if we sample without replacement, ${}_{200}C_{10}$ different samples of size 10 can be selected from Table 9.1. We have chosen 80 of these possible samples to illustrate that $\bar{X}$ is a random variable with a value that varies from sample to sample. Table 9.4 shows the relative frequency distribution of the 80 sample means, and Figure 9.4 shows the relative frequency histogram of these 80 sample means.

**ple means
les of size 10
Table 9.1**

3.8	4.6	7.5	4.4	4.1	5.5	2.4	4.2	4.0	5.2
6.0	4.1	4.7	4.3	5.5	4.5	4.9	4.3	3.2	5.2
2.5	4.8	5.4	3.8	3.0	7.1	4.5	6.2	4.5	4.3
5.5	4.5	5.1	4.9	5.7	4.8	4.3	2.4	4.3	4.8
3.3	5.1	5.1	4.3	4.0	4.4	5.0	6.2	4.0	4.3
4.2	4.1	5.0	6.4	4.8	5.6	5.0	4.3	2.7	4.8
3.5	5.7	4.6	3.2	4.7	5.1	3.5	4.8	4.8	4.9
4.9	6.2	5.0	3.4	4.5	5.0	5.5	5.2	4.4	3.5

**n of 80 sample
Table 9.3**

$\bar{X}$	Frequency	Relative Frequency
2.0 to under 2.5	2	.0250
2.5 to under 3.0	2	.0250
3.0 to under 3.5	5	.0650
3.5 to under 4.0	5	.0650
4.0 to under 4.5	19	.2375
4.5 to under 5.0	20	.2500
5.0 to under 5.5	13	.1625
5.5 to under 6.0	7	.0875
6.0 to under 6.5	5	.0625
6.5 to under 7.0	0	.0000
7.0 to under 7.5	2	.0250
Total	80	1.0000

Figure 9.3 shows that the relative frequency distribution of the population of data is approximately a uniform distribution. This result is reasonable because if the digits are truly random, each digit should be equally likely with a relative frequency of .1.

4.5

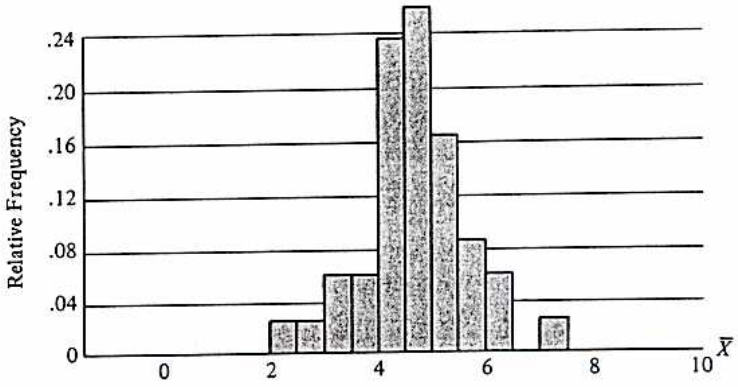


FIGURE 9.4
Relative frequency histogram of 80 sample means in Table 9.3.

Figure 9.4 shows the distribution of $\bar{X}$ values. This distribution is approximately bell shaped, and most of the values of $\bar{X}$ are clustered around 4.62, which is the mean of the original population of data. Even though the population is approximately uniformly distributed, the sampling distribution of $\bar{X}$ is approximately normal. It can be shown that this approximation to the normal distribution improves as the sample size increases. For example, if we had taken samples of size $n = 30$ rather than $n = 10$, the approximation to normality would have been more exact. ■

Mean and Variance of the Random Variable $\bar{X}$

The mean and variance of the sampling distribution of $\bar{X}$ depend on the mean and variance of the population and on the sample size n. The formulas for the mean and variance of the sampling distribution of $\bar{X}$ are given in the accompanying box.

Formulas for the Mean and Variance of $\bar{X}$

Let $x_1, x_2, \ldots, x_n$ denote a random sample of size n selected from a population having mean μ_x (or μ) and variance σ_x^2 (or σ^2). Let $\bar{X}$ denote the sample mean.

1. The mean value, or expected value, of $\bar{X}$ is
$$E(\bar{X}) = \mu_x = \mu$$
2. If the population is infinitely large or if sampling is done with replacement, then the variance of $\bar{X}$ is
$$\text{Var}(\bar{X}) = \sigma_x^2/n$$
and the standard deviation of $\bar{X}$ is given by $\sigma_x/\sqrt{n}$.
3. If sampling is done *without replacement from a finite population* containing N elements, then the variance of $\bar{X}$ is
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

The value $[(N - n)/(N - 1)]$ in the formula for calculating the variance when sampling is done without replacement is called the *finite population correction factor*

where N is the population size and n is the sample size. If the population size N is much larger than the sample size n, this correction factor is close to 1; in such cases it makes little difference whether we sample with or without replacement. Unless explicitly noted, we will assume henceforth that sampling is done from an extremely large population so that this correction factor can be ignored.

According to the theorem, the standard deviation of the $\bar{X}$ values is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. This shows that $\sigma_{\bar{X}}$ will be small if σ is small or if n is large. If the population standard deviation σ is small, there is little variation in the population and we would expect sample means obtained from different samples to be highly concentrated. If the sample size is large, we would expect the different estimates $\bar{X}$ obtained from different samples to be highly concentrated about the mean μ.

Effect of a Small Population Standard Deviation: An employment agency wants to estimate the mean hourly wage earned by high school students who work as cashiers at supermarkets. The agency hires several pollsters and asks each one to obtain a random sample of $n = 25$ students. Because most students earn between \$3.35 and \$5.00 per hour, the population standard deviation is relatively small, say, $\sigma = \$1.00$. Because the students' wages are highly concentrated and do not exhibit much variation, the estimates obtained by the different pollsters should not exhibit much variation either. That is, we would expect the values of $\bar{X}$ obtained by each pollster to be quite close to one another. Thus, if the population has a small standard deviation, the sampling distribution of sample means will also be highly concentrated. ■

Effect of a Large Population Standard Deviation: The employment agency in Example 9.4 is also interested in estimating the mean annual income earned by lawyers. Again pollsters are hired, and each is told to obtain a random sample of $n = 25$ lawyers. The population mean is, say, \$70,000 per year. However, the standard deviation of this population is very large, say $\sigma = \$20{,}000$, because the annual incomes are widely dispersed about the mean. As a result, the sample mean obtained by one pollster could be quite different from the sample mean obtained by another. We would not necessarily expect the values of $\bar{X}$ obtained by the pollsters to be quite so close to one another. If the population has a large standard deviation, the sampling distribution of the sample mean can also have a large amount of dispersion. ■

The following example illustrates the point that the variance of the sampling distribution of $\bar{X}$ depends on the sample size n. If the sample size is large, then the sampling distribution of $\bar{X}$ will be highly concentrated about the population mean.

Effect of Sample Size on the Variance of $\bar{X}$: Several economists are interested in estimating the mean annual income of accountants. Suppose the unknown population mean is $\mu = \$50{,}000$ and the population standard deviation is $\sigma = \$10{,}000$. Now suppose one economist takes a random sample of $n = 4$ employees. Because this sample size is very small, the estimate obtained by the economist could be quite different from the true population mean. For example, the standard deviation of the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n} = 10{,}000/\sqrt{4} = \$5000$. From the Empirical Rule, approximately 68% of the data lie within 1 standard deviation of the mean if the distribution is approximately bell shaped, so only about 68% of the potential estimates of the population mean will be within \$5000 of the true mean. Since the remaining 32% of the estimates will be

more than \$5000 from the mean, the probability is approximately .32 that the sampling error will be more than \$5000. An estimate of the mean income that is in error by more than \$5000 could be useless in many economic studies. Because the sample size is very small, the sampling distribution of $\bar{X}$ is widely dispersed about the population mean.

Now suppose another economist takes a random sample of $n = 1600$ employees. Because this sample size is quite large, we would expect to obtain a very good estimate of the population mean. In this case, the standard deviation of the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n} = \$10{,}000/\sqrt{1600} = \$250$. From the Empirical Rule, the probability is approximately .95 that the sampling error $|\bar{X} - \mu|$ will be less than 2 standard deviations, namely \$500, if the distribution is approximately bell shaped. Thus, in approximately 95% of the potential samples of size 1600, the estimated sample mean will be within \$500 of the true population mean. An estimate of the mean income that is in error by less than \$500 would probably be quite useful in most economic studies. When the sample size is large, the sampling distribution of the sample mean is highly concentrated about the population mean.

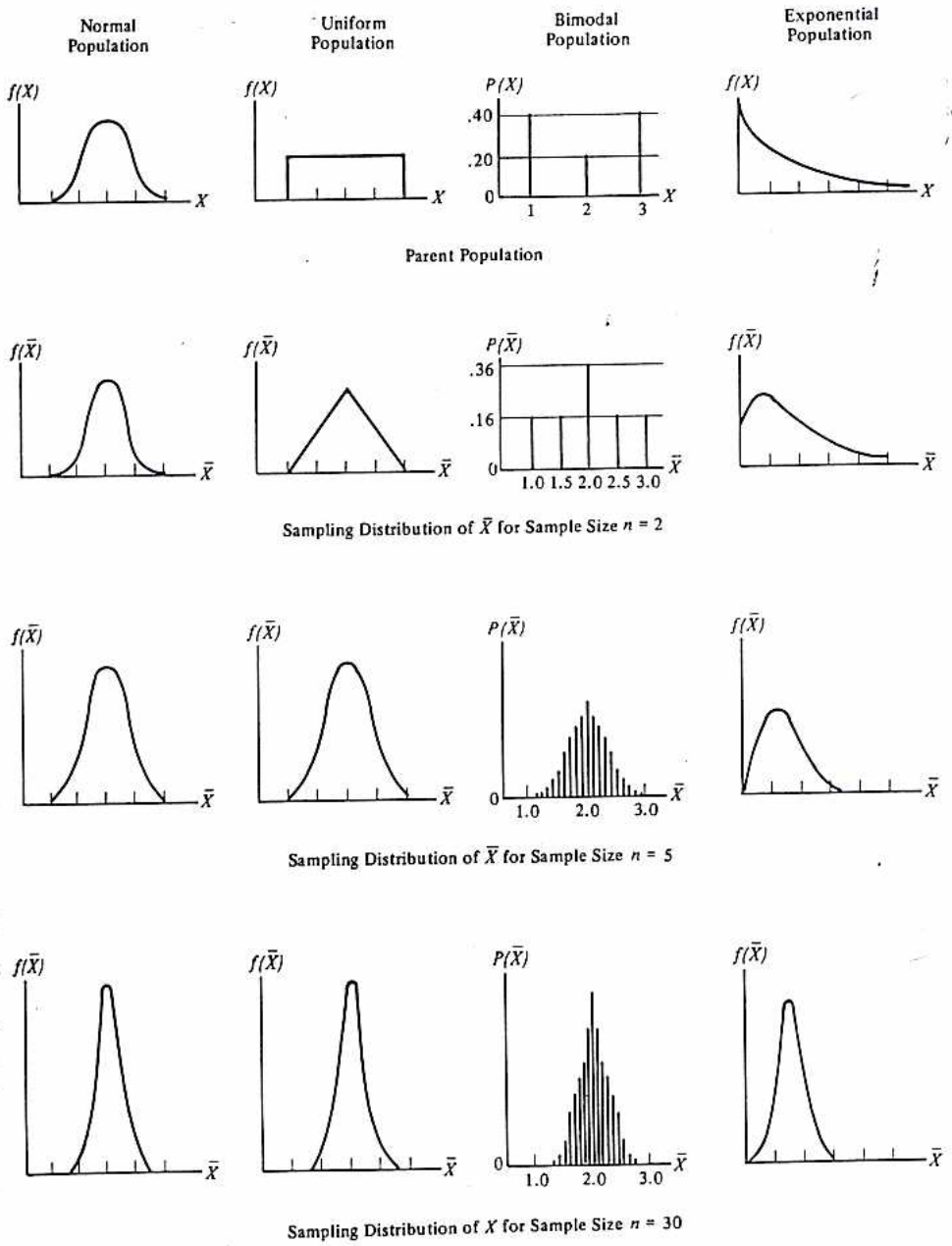
Graphical Representation of the Distribution of $\bar{X}$

The graphs in Figure 9.5 show that the sampling distribution of $\bar{X}$ is centered at the population mean and becomes more concentrated as the sample size increases. Also note that as the sample size gets larger, the distribution of $\bar{X}$ begins to approach normality regardless of the shape of the original distribution.

Exercises for Section 9.2

1. Suppose the average IQ of all college seniors is 125 and the variance is 144. You take a random sample of 100 students and calculate the sample mean IQ ($\bar{X}$). The sample mean $\bar{X}$ is a random variable having a probability distribution. What is the mean, variance, and standard deviation of this distribution?
2. A city council is studying the tourist trade. During a certain year, the mean expenditure per tourist was $\mu = \$500$ with variance 8100. The city takes a random sample of 10 tourists and calculates their average expenditure $\bar{X}$. What is the mean, variance, and standard deviation of this distribution?
3. In Exercise 2 the city takes a random sample of 1000 tourists and calculates the mean expenditure. Now what is the mean, variance, and standard deviation of the distribution of $\bar{X}$? In this case would you be very confident that your value of $\bar{X}$ is very close to μ?
4. You take a random sample of size $n = 10$ and obtain $x_1 = 8$, $x_2 = 6$, $x_3 = 5$, $x_4 = 5$, $x_5 = 8$, $x_6 = 7$, $x_7 = 6$, $x_8 = 2$, $x_9 = 4$, and $x_{10} = 6$. Calculate the sample mean and sample standard deviation.
5. A person takes a random sample of 10 college seniors to obtain their IQs and gets the values $x_1 = 110$, $x_2 = 145$, $x_3 = 150$, $x_4 = 130$, $x_5 = 120$, $x_6 = 105$, $x_7 = 130$, $x_8 = 105$, $x_9 = 125$, and $x_{10} = 120$. Use the formulas in Chapter 4 to calculate the sample mean and sample variance. Use the formula S^2/n to estimate the variance of $\bar{X}$.

Sampling distribution of $\bar{X}$ for $n = 2$, 5, and 30.



6. Five randomly selected tourists are asked how much they spent on their vacation. We obtain $x_1 = \$600$, $x_2 = \$800$, $x_3 = \$400$, $x_4 = \$1000$, and $x_5 = \$500$. Calculate the sample mean and sample variance. Estimate the variance of $\bar{X}$ using the estimator S^2/n.

7. If the population is approximately normal, the variance of the sample median equals $(\pi/2)(\sigma^2/n)$. If a sample median is to have a standard deviation equal to that of a sample mean (from the same population), how much bigger must the sample size be?

The Central Limit Theorem (C.L.T.)

The Central Limit Theorem, one of the most important theorems in statistics, is the reason that the normal distribution is the most important of all probability distributions.

C. L. T. :

Let $x_1, x_2, \ldots x_n$ denote a random sample selected from a population having mean μ and variance σ^2. Let $\bar{X}$ denote the sample mean. If n is large, then $\bar{X}$ has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

The C.L.T. would be of little practical value if the approximation was applicable only if n were extremely large, but in fact it works in most cases, even when n is fairly small. A common rule of thumb is that the theorem applies as n gets up to 20 or 30.

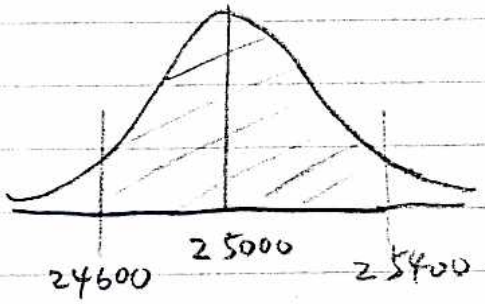
If the original population follows the normal distribution, then the distribution of $\bar{X}$ is exactly a normal distribution, regardless of the sample size. The C.L.T. is important because it states that if the original population is NOT Normal, the distribution of $\bar{X}$ will be approximately normal (provided $n \geq 30$).

e.g. Application of the C.L.T. : Suppose the Bureau of Labour Statistics wants to estimate the mean starting salary of newly graduate

chemical engineers. Suppose the population mean is actually $\mu = \$25\,000$ and the s.d. is $\sigma = \$2000$. In order to estimate μ, the bureau takes a random sample of 100 recently graduated chemical engineers. Find the probability the sample mean will be within \$400 of the true mean.

Soln: We have $n = 100$, $\mu = 25000$, $\sigma = 2000$. Because the sample size > 30, we assume C.L.T holds. The sampling distribution of $\bar{x}$ is approximately normal with mean $\mu = \$25000$ & s.d. $= \frac{2000}{\sqrt{100}} = 200$. If $\bar{x}$ is within \$400 of the population mean, then $\bar{x}$ will be between 24600 & 25400.

Thus we seek $P\{24600 \leq \bar{x} \leq 25400\}$



$$P\{24600 \leq \bar{x} \leq 25400\}$$

$$= P\left\{\frac{24600-25000}{200} \leq z \leq \frac{25400-25000}{200}\right\}$$

$$= P\{-2 \leq z \leq 2\} = .4772 + .4772$$

$$= .9544$$

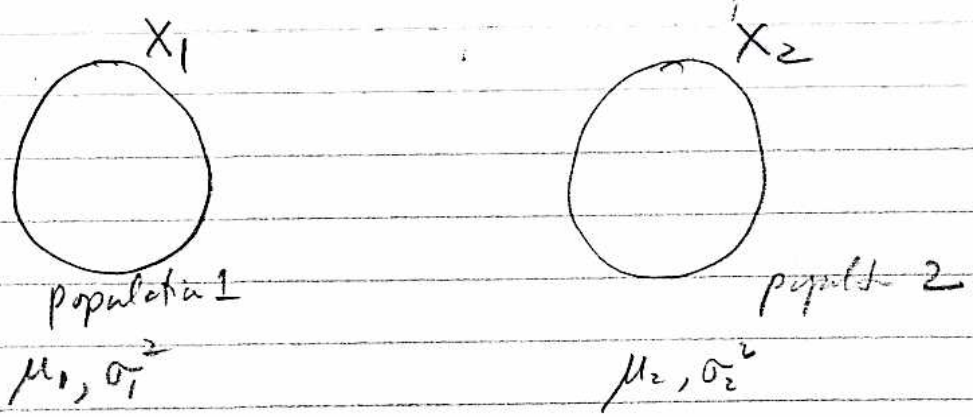
Thus, if the Bureau of Labor Statistics takes a random sample of 100 chemical engineers, the prob. is 0.9544 that the sample mean will be within \$400 of the true pop. mean.

Sampling distribution of the difference between two sample means

Frequently we are interested in determining if the mean of one population is equal to the mean of another.

e.g. (a) Is the mean height of male the same as the mean height of female.

(b) Is the mean number of cavities for children who use Brand A toothpaste the same as mean # of cavities for children who use Brand B.



To make comparisons, we take <u>independent</u> random sample of n_1 observations from group 1 & n_2 obs. from population 2 and then calculate $\bar{x}_1, \bar{x}_2$.

$\bar{X}_1, \bar{X}_2$ are random variables with values that vary from sample to sample. $\bar{X}_1 - \bar{X}_2$ is also a random variable with values vary from sample to sample.

Now, we have 5 different prob distribut—
$X_1, X_2, \bar{X}_1, \bar{X}_2, \bar{X}_1 - \bar{X}_2$

If n_1, n_2 is large enough ($n_1, n_2 \geq 30$), we

have

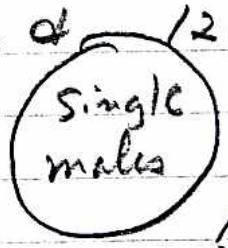
$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

and $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

e.g A financial loan officer claims that the mean monthly payment for credit cards is \$80 with a variance of 1400 for single females and \$80 with a variance of 1320 for single males. We take a random sample of 100 single females & 120 single ma

$\mu_1 = 80$
$\sigma_1^2 = 1400$
$n_1 = 100$



$\mu_1 = 80$
$\sigma_2^2 = 1320$
$n_2 = 120$

What is the probability that the sample mean for females will be at least \$5 higher than the sample mean for males?

Soln: by CLT

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{1400}{100} + \frac{1320}{120}\right)$$

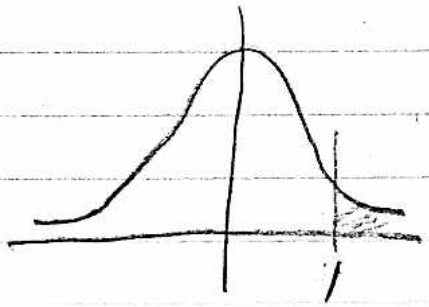
$$= N(0, 25)$$

We seek the prob that $\bar{X}_1$ is at least \$5 more than $\bar{X}_2$.

i.e. we seek

$$P\{\bar{X}_1 - \bar{X}_2 \geq 5\} = P\left\{\frac{\bar{X}_1 - \bar{X}_2 - 0}{5} \geq \frac{5-0}{5}\right\}$$

$$= P\{Z > 1\} = 0.1587$$



Sampling Distribution of the sample proportion

There are numerous problems in real life where we want to know the proportion of items in a population that posses a certain characteristic

e.g. (1) TV producers wants to know what proportion of viewers of a certain program have incomes exceeding $40k per year

(2) proportion of lung cancer patient out of a smoking population.

In each case we have to estimate the parameter p, the proportion of the population having the characteristic of interest.

Let x denote the # of observations in the sample that possess the characteristic of interest. We use the sample proportion $\hat{p} = \frac{x}{n}$ to estimate the population proportion p.

If $np \geq 5$, $nq \geq 5$, $X \sim N(np, npq)$

under the same constraints, the sample distribution of $\hat{p}$ is approximately normal

i.e. $\hat{p} = \frac{X}{n} \sim N\left(p, \frac{pq}{n}\right)$

i.e. $\text{mean}(\hat{p}) = p$, $\sigma^2(\hat{p}) = \frac{pq}{n}$

e.g. TV producer claims that 30% of all homes in a city have cable TV. To test this claim, a random sample of 400 homes is examined.

(a) What is the probability that the sample proportion is less than 0.25?

(b) What is the probability that the sample proportion is bigger than 0.20?

(c) Suppose the sample proportion is 0.18 Would this make you doubt the producer's claim?

Solution:

$$\sigma = \sqrt{\frac{0.3 \times 0.7}{400}} = 0.02291$$

(a) $$P\{\hat{p} < 0.25\} = P\left\{\frac{\hat{p} - 0.3}{0.02291} \leq \frac{0.25 - 0.3}{0.02291}\right\}$$

$$= P\{Z \leq -2.18\} = 0.0146$$

(b)

$$P\{\hat{p} > 0.20\} = P\left\{\frac{\hat{p}-0.3}{0.02291} > \frac{0.20-0.3}{.02291}\right\}$$

$$= P(Z > -4.36) = 1$$

(c) Yes, there is virtually NO chance of it being below 0.20.

e.g. A medical journal states that the probability of surviving open heart surgery is .9. To check this claim, you examine the records of a sample of 200 patients who had open heart surgery. In the sample, 77% survived.

(a) If the journal's claim is correct, find the probability that the proportion of survivors is 77% or less

(b) Does the result in part (a) make you doubt the journal's claim? Explain.

Soln: $\sigma = \sqrt{\frac{(.9)(.1)}{200}} = 0.02121$

(a)

$$P\{\hat{p} \leq 0.77\} = P\left\{\frac{\hat{p}-0.9}{0.02121} \leq \frac{0.77-0.9}{0.02121}\right\}$$

$$= P\{Z \leq -6.128\} = 0$$

(b) Yes, there is NO chance of it being below 0.77.