- If $\beta_1 < 0$, the distribution is negatively skewed

Definition 38. *Coefficient of kurtosis:* $\beta_1 = \frac{\mu_4}{\mu_2^2}$

- If $\beta_2 > 3$, the curve is leptokurtic
- If $\beta_2 < 3$, the curve is platykurtic
- If $\beta_2 = 3$, the curve is mesokurtic

Example 47. *Find the coefficient of skewness and coefficient of kurtosis, if the pdf of*

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & otherwise \end{cases}$$

Definition 39. r^{th} *raw moment about origin*

- $\mu_r^{'} = \begin{cases} \sum x^r f(x) & Xisdiscrete \\ \int x^r f(x)dx & Xiscontinuous \end{cases}$
- $M_X(t) = E(e^{tX}) = \begin{cases} \sum e^{tx} f(x) & Xisdiscrete \\ \int e^{tx} f(x)dx & Xiscontinuous \end{cases}$

Example 48. *Properties of Moment generating function:*

- $M_{cX}(t) = M_X(e^{ct})$
- $M_{X-a}(t) = e^{-at} M_X(t)$
- $\frac{d^r(M_X(t))}{dt^r} = \mu_r^{'}$
- $\mu_2 = var(x)$
- $\mu_2 = \mu_2^{'} - (\mu_1^{'})^2$
- $mean = \mu_1^{'}$

Example 49. *Find MGF for* $f(x) = \begin{cases} \frac{2}{3} & x = 1 \\ \frac{1}{3} & x = 2 \end{cases}$ *Find Mean and variance.*

Example 50. *Find MGF for* $f(x) = \begin{cases} \frac{1}{3} & -1 < x < 2 \\ 0 & otherwise \end{cases}$ *Find Mean and variance.*

Definition 40. *Correlation is used to give information about the relation between x and y. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative.*

- $|r| \leq 1$

- *r=0 i.e, no correlation*

- *r=-1 i.e, perfect negative correlation*

- *r=1 i.e, perfect positive correlation*

Definition 41. *Line of regression of Y on X is given by* $y - \bar{y} = r\dfrac{\sigma_Y}{\sigma_X}(x - \bar{x})$, *where r is correlation coefficient and σ is standard deviation.*

Definition 42. *Line of regression of Y on X is given by* $x - \bar{x} = r\dfrac{\sigma_X}{\sigma_Y}(y - \bar{y})$

Definition 43. *Angle between two lines of regression* $\tan(\theta) = \dfrac{1 - r^2}{r}\left(\frac{\sigma_Y \sigma_X}{{\sigma_X}^2 + {\sigma_Y}^2}\right)$

- $r = 0, \Rightarrow \theta = 90$, *line of regression are perpendicular to each other*

- $r = \pm 1, \Rightarrow \theta = 0 or \pi$, *line of regression are parallel or coincide*

  Definition 44. *Regression coeff of Y on X:*

  $r\dfrac{\sigma_Y}{\sigma_X} = b_{YX} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

- $r\dfrac{\sigma_X}{\sigma_Y} = b_{XY} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$

Example 51. *From the following data, find*

(1) two regression eqns

(2) Coeff of correlation between the marks in economics and stats

(3) The most likely marks in stats when marks in economics is 30

X	*1*	*2*	*3*	*5*	*9*
Y	*2*	*3*	*4*	*1*	*5*

Example 52. *The two lines of regression are* $8x - 10y + 66 = 0$ *and* $40x - 18y - 214 = 0$. *Find mean of X and Y and correlation coefficient.*

Example 53. *The following data are* $\bar{x} = 910, \bar{y} = 18, \sigma_X = 38, \sigma_y = 2, r = 0.6$. *Find the line of regression and obtain the value of X when $Y = 20$*

Example 54. *The regression equation of x and y is* $3y - 5x + 108 = 0$. *If* $\bar{y} = 44, Var(X) = \frac{9}{16}Var(Y)$. *Find mean of X and correlation coefficient.*

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variable, with the effect of the rest of the variables eliminated.

If there are three variables X_1, X_2 *and* X_3, there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote by $r_{12.3}$, that is, the coefficient of partial correlation X_1 *and* X_2 keeping X_3 constant, it is calculated as

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}, \quad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}, \; r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

The Binomial Distribution

A. It would be very tedious if, every time we had a slightly different problem, we had to determine the probability distributions from scratch. Luckily, there are enough similarities between certain types, or families, of experiments, to make it possible to develop formulas representing their general characteristics.

For example, many experiments share the common element that their outcomes can be classified into one of two events, e.g. a coin can come up heads or tails; a child can be male or female; a person can die or not die; a person can be employed or unemployed. These outcomes are often labeled as "success" or "failure." Note that there is no connotation of "goodness" here - for example, when looking at births, the statistician might label the birth of a boy as a "success" and the birth of a girl as a "failure," but the parents wouldn't necessarily see things that way. The usual notation is

p = probability of success,
q = probability of failure = 1 - p.

Note that p + q = 1. In statistical terms, **A Bernoulli trial is each repetition of an experiment involving only 2 outcomes**.

We are often interested in the result of *independent, repeated bernoulli trials*, i.e. the number of successes in repeated trials.

1. *independent* - the result of one trial does not affect the result of another trial.

2. *repeated* - conditions are the same for each trial, i.e. p and q remain constant across trials. Hayes refers to this as a <u>stationary process</u>. If p and q can change from trial to trial, the process is <u>nonstationary</u>. The term <u>identically distributed</u> is also often used.

B. A binomial distribution gives us the probabilities associated with independent, repeated Bernoulli trials. **In a binomial distribution the probabilities of interest are those of receiving a certain number of successes, r, in n independent trials each having only two possible outcomes and the same probability, p, of success.** So, for example, using a binomial distribution, we can determine the probability of getting 4 heads in 10 coin tosses.

How does the binomial distribution do this? Basically, a two part process is involved. First, we have to determine the probability of one possible way the event can occur, and then determine the number of different ways the event can occur. That is,

P(Event) = (Number of ways event can occur) * P(One occurrence).

Suppose, for example, we want to find the probability of getting 4 heads in 10 tosses. In this case, we'll call getting a heads a "success." Also, in this case, n = 10, the number of successes is r = 4, and the number of failures (tails) is n – r = 10 – 4 = 6. One way this can occur is if the first 4 tosses are heads and the last 6 are tails, i.e.

S S S S F F F F F F

The likelihood of this occurring is

P(S) * P(S) * P(S) * P(S) * P(F) * P(F) * P(F) * P(F) * P(F) * P(F)

More generally, if p = probability of success and q = 1 – p = probability of failure, the probability of a specific sequence of outcomes where there are r successes and n-r failures is

$$p^r q^{n-r}$$

So, in this particular case, p = q = .5, r = 4, n-r = 6, so the probability of 4 straight heads followed by 6 straight tails is $.5^4.5^6$ = 0.0009765625 (or 1 out of 1024).

Of course, this is just one of many ways that you can get 4 heads; further, because the repeated trials are all independent and identically distributed, each way of getting 4 heads is equally likely, e.g. the sequence S S S S F F F F F F is just as likely as the sequence S F S F F S F F S F. So, we also need to know how many different combinations produce 4 heads.

Well, we could just write them all out…but life will be much simpler if we take advantage of two counting rules:

1. The number of different ways that N distinct things may be arranged in order is

N! = (1)(2)(3)...(N-1)(N), (where 0! = 1).

An arrangement in order is called a permutation, so that the total number of permutations of N objects is N!. The symbol N! is called N factorial.

EXAMPLE. Rank candidates A, B, and C in order. The possible permutations are: ABC ACB BAC BCA CBA CAB. Hence, there are 6 possible orderings. Note that 3! = (1)(2)(3) = 6.

NOTE: Appendix E, Table 6, p. 19 contains a Table of the factorials for the integers 1 through 50. For example, 12! = 4.79002 * 10^8. (Or see Hayes Table 8, p. 947). Your calculator may have a factorial function labeled something like x!

2. The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is

$$\frac{N!}{r!(N-r)!} = \binom{N}{r} = \binom{N}{N-r}$$

We refer to this as "N choose r." Sometimes the number of combinations is known as a *binomial coefficient*, and sometimes the notation ${}_NC_r$ is used. So, in the present example,

$$\binom{N}{r} = \binom{10}{4} = \frac{N!}{r!(N-r)!} = \frac{10!}{4!(10-4)!} = \frac{10*9*8*7}{4*3*2*1} = \frac{5040}{24} = 210$$

Note that, for 10!, I stopped once I got to 7; and I didn't write out 6! in the denominator. This is because both numerator and denominator have 6! in them, so they cancel out. So, there are 210 ways you can toss a coin 10 times and get 4 heads.

EXAMPLE. Candidates A, B, C and D are running for office. Vote for two.

The possible choices are: AB AC AD BC BD CD, i.e. there are 6 possible combinations. Confirming this with the above formula, we get

$$\frac{N!}{r!(N-r)!} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{12}{2} = 6$$

EXAMPLE. There are 100 applicants for 3 job openings. The number of possible combinations is

$$\frac{N!}{r!(N-r)!} = \binom{100}{3} = \frac{100!}{3!97!} = \frac{100*99*98}{3*2} = \frac{970{,}200}{6} = 161{,}700$$

Again, note that, if you didn't take advantage of 97! appearing on both top and bottom, you'd have a much lengthier calculation.

See Appendix E, Table 7, page 20 for ${}_N C_r$ values for various values of N and r. (Or see Hayes, Appendix E, Table IX, p. 948). Your calculator may have a function labeled nCr or something similar.

C. So putting everything together now: we know that any specific sequence that produces 4 heads in 10 tosses has a probability of 0.0009765625. Further, we now know that there are 210 such sequences. Ergo, the probability of 4 heads in 10 tosses is 210 * 0.0009765625 = 0.205078125.

We can now write out the complete formula for the binomial distribution:

In sampling from a stationary Bernoulli process, with the probability of success equal to p, the probability of observing exactly r successes in N independent trials is

$$\binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}$$

Once again, N choose r tells you the number of sequences that will produce r successes in N tries, while $p^r q^{N-r}$ tells you what the probability of each individual sequence is.

To put it another way, the random variable X in a binomial distribution can be defined as follows:

Let Xi = 1 if the ith bernoulli trial is successful, 0 otherwise. Then,

$X = \Sigma X_i$, where the X_i's are independent and identically distributed (iid).

That is, X = the # of successes. Hence**, Any random variable X with probability function given by**

$$p(X = r; N, p) = \binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}, \quad X = 0,1,2,\ldots,N$$

is said to have a binomial distribution with parameters N and p.

EXAMPLE. In each of 4 races, the Democrats have a 60% chance of winning. Assuming that the races are independent of each other, what is the probability that:

a. The Democrats will win 0 races, 1 race, 2 races, 3 races, or all 4 races?
b. The Democrats will win at least 1 race
c. The Democrats will win a majority of the races

SOLUTION. Let X equal the number of races the Democrats win.

a. Using the formula for the binomial distribution,

$$\binom{4}{0} p^0 q^{4-0} = \frac{4!}{0!(4-0)!} .60^0 .40^4 = .40^4 = .0256,$$

$$\binom{4}{1} p^1 q^{4-1} = \frac{4!}{1!(4-1)!} .60^1 .40^3 = 4 * .60 * .40^3 = .1536,$$

$$\binom{4}{2} p^2 q^{4-2} = \frac{4!}{2!(4-2)!} .60^2 .40^2 = 6 * .60^2 .40^2 = .3456,$$

$$\binom{4}{3} p^3 q^{4-3} = \frac{4!}{3!(4-3)!} .60^3 .40^1 = 4 * .60^3 .40^1 = .3456,$$

$$\binom{4}{4} p^4 q^{4-4} = \frac{4!}{4!(4-4)!} .60^4 .40^0 = .60^4 = .1296$$

b. P(at least 1) = P(X ≥ 1) = 1 - P(none) = 1 - P(0) = .9744. Or, P(1) + P(2) + P(3) + P(4) = .9744.

c. P(Democrats will win a majority) = P(X ≥ 3) = P(3) + P(4) = .3456 + .1296 = .4752.

EXAMPLE. In a family of 11 children, what is the probability that there will be more boys than girls? Solve this problem WITHOUT using the complements rule.

SOLUTION. You could go through the same tedious process described above, which is what most students did when I first asked this question on an exam. You would compute P(6), P(7), P(8), P(9), P(10), and P(11).

Or, you can look at Appendix E, Table II (or Hays pp. 927-931). Here, both Hayes and I list binomial probabilities for values of N and r from 1 through 20, and for values of p that range from .05 through .50.

Thus, on page E-5, we see that for N = 11 and p = .50,

P(6) + P(7) + P(8) + P(9) + P(10) + P(11) = .2256 + .1611 + .0806 + .0269 + .0054 + .0005 = .50.

NOTE: Understanding the tables in Appendix E can make things a lot simpler for you!

EXAMPLE. [WE MAY SKIP THIS EXAMPLE IF WE RUN SHORT OF TIME, BUT YOU SHOULD STILL GO OVER IT AND MAKE SURE YOU UNDERSTAND IT]

Use Appendix E, Table II, to once again solve this problem: In each of 4 races, the Democrats have a 60% chance of winning. Assuming that the races are independent of each other, what is the probability that:

a. The Democrats will win 0 races, 1 race, 2 races, 3 races, or all 4 races?
b. The Democrats will win at least 1 race
c. The Democrats will win a majority of the races

SOLUTION. It may seem like you can't do this, since the table doesn't list p = .60. However, all you have to do is redefine success and failure. Let success = P(opponents win a race) = .40. The question can then be recast as finding the probability that

a. The opponents will win 4 races, 3 races, 2 races, 1 race, or none of the races?
b. The opponents will win 0, 1, 2, or 3 races; or, the opponents will not win all the races
c. The opponents will not win a majority of the races

We therefore look at page E-4 (or Hayes, p. 927), N = 4 and p = .40, and find that

a. P(4) = .0256, P(3) = .1536, P(2) = .3456, P(1) = .3456, and P(0) = .1296.
b. P(0) + P(1) + P(2) + P(3) = 1 - P(4) = .9744
c. P(1) + P(0) = .3456 + .1296 = .4752

In general, for p > .50: To use Table II, substitute 1 - p for p, and substitute N - r for r. Thus, for p = .60 and N = 4, the probability of 1 success can be found by looking up p = .40 and r = 3.

D. *Mean of the binomial distribution*. Recall that, for any discrete random variable, $E(X) = \Sigma xp(x)$. Therefore, $E(X_i) = \Sigma xp(x) = 0 * (1 - p) + 1 * p = p$, that is, the mean of any

bernoulli trial is p, the probability of success. By applying our theorems for expectations, we find that

$$\mathbf{E(X) = E(X_1 + X_2 + ... + X_N) = E(X_1) + E(X_2) + ... + E(X_N) = Np.}$$

NOTE: Hayes provides an alternative proof in section 4.8; I think my proof is much simpler.

EXAMPLE. If we toss a fair coin 3 times (N = 3, p = .50), we expect to get 1.5 heads.

E. *Variance of the binomial distribution.* Given our above definitions, note that, for any Bernoulli trial, $X_i^2 = X_i$. Thus, $V(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1 - p) = pq$. Because trials are independent, we find that

$$\mathbf{V(X) = V(X_1) + V(X_2) + ... + V(X_N) = Npq.}$$

EXAMPLE. If you toss a fair coin 3 times (N = 3, p =.50, q = .50) the variance of the expected number of heads is 3 * .50 * .50 = .75.

NOTE. These last two points mean that the mean and variance of the binomial distribution are dependent on only two parameters, N and p.

WARNING: The symbol X gets used many different ways in statistics. In this case, X = the sum of all the Bernoulli trials, but in other instances it might refer to an individual Bernoulli trial.

F. *Shape of the binomial distribution.* When p = .5, the binomial distribution is symmetrical - the mean and median are equal. Even when p <> .5, the shape of the distribution becomes more and more symmetrical the larger the value of N. This is very important, because the binomial distribution can quickly become unwieldy - as we will later see, there are approximations to the binomial that can be much easier to use when N is large.

g. A MORE COMPLETE LISTING OF THE COUNTING RULES FOR PERMUTATIONS AND COMBINATIONS (OPTIONAL).

Here is a more extensive set of counting rules that can be useful for various problems in probability. They aren't essential for our immediate purposes, so we probably won't go over them in class unless we have extra time. But, these aren't very hard, and they may come in handy for you some day, so I am including them here.

1. NUMBER OF POSSIBLE SEQUENCES FOR N TRIALS. Suppose that a series of N trials were carried out, and that on each trial any of K events might occur. Then the following rule holds:

If any one of K mutually exclusive and exhaustive events can occur on each of N trials, then there are K^n different sequences that may result from a set of such trials.

EXAMPLE. If you toss a die once, any of 6 numbers can show up (K = 6). Ergo, if you toss it 3 times, any of $6^3 = 216$ sequences are possible (e.g., 111, 342, 652, etc.). [Your calculator probably has a y^x function or something similar; on mine, I press 6, then y^x, then 3, then =.]

2. SEQUENCES. Sometimes the number of possible events in the first trial of a series is different from the number possible in the second, the second different from the third, etc. That is, $K_1 \neq K_2 \neq K_3$, etc. Under such conditions,

If $K_1,...,K_N$ are the numbers of distinct events that can occur on trials 1,..., N in a series, then the number of different sequences of N events that can occur is $(K_1)(K_2)...(K_N)$.

EXAMPLE. Two occupations and three religions yield 6 combinations of occupation and religion. Tossing a coin (2 outcomes) and tossing a die (6 outcomes) yield 12 possible outcomes.

Note that, when $K_i = K$ for all i, then rule 1 becomes a special case of rule 2. Note also that, when $K_1 = 1$, $K_2 = 2,..., K_N = N$, then rule 3 becomes a special case of rule 2.

3. PERMUTATIONS. A rule of extreme importance in probability computations concerns the number of ways that objects may be arranged in order.

The number of different ways that N distinct things may be arranged in order is
$$N! = (1)(2)(3)...(N-1)(N), \text{ (where } 0! = 1).$$

An arrangement in order is called a permutation, so that the total number of permutations of N objects is N!. The symbol N! is called N factorial. The notation ${}_NP_N$ is also sometimes used for N!, for reasons which will be clear in a moment.

EXAMPLE. Rank candidates A, B, and C in order. The possible permutations are:
ABC ACB BAC BCA CBA CAB

There are 6 possible orderings. Note that 3! = (1)(2)(3) = 6.

NOTE: Appendix E, Table 6, p. 19 contains a Table of the factorials for the integers 1 through 50. For example, 12! = 4.79002 * 10^8. (Or see Hayes Table 8, p. 947). Your calculator may have a factorial function labeled something like x!

3B. PERMUTATIONS OF SIMILAR OBJECTS.

Suppose we have N objects, N_1 alike, N_2 alike,..., N_k alike ($\Sigma N_i = N$). Then, the number of ways of arranging these objects is

$$\frac{N!}{N_1!\,N_2!\ldots N_k!}$$

EXAMPLE. We have 4 balls, 2 red and 2 blue. The possible ways of arranging the balls are BBRR, BRBR, BRRB, RRBB, RBRB, RBBR, or 6 ways altogether. To confirm that there are 6 ways,

$$\frac{N!}{N_1!\,N_2!\ldots N_k!} = \frac{4!}{2!2!} = 6$$

EXAMPLE. If we have 6 balls, 2 red, 2 blue, and 2 green, the number of possible ways of arranging them is

$$\frac{N!}{N_1!\,N_2!\ldots N_k!} = \frac{6!}{2!2!2!} = \frac{720}{8} = 90$$

NOTE: I add this rule to Hayes's list because many of the other rules are special cases of it. When $N_i = 1$ for all i, Rules 3 (Permutations) and 3B are equivalent. When $N_i = 1$ for i = 1 to r, and $N_{r+1} = N - r$, then rules 4 (Ordered combinations) and 3B are the same. When $N_1 = r$ and $N_2 = N - r$, Rules 5 (Combinations) and 3B are the same.

4. ORDERED COMBINATIONS; or, PERMUTATIONS OF N OBJECTS TAKEN r AT A TIME.

Sometimes it is necessary to count the number of ways that r objects might be selected from among some N objects in all ($r \le N$). Further, each different arrangement of the r objects is considered separately.

The number of ways of selecting & arranging r objects from among N distinct objects is

$$\frac{N!}{(N-r)!}$$

Verbally, we refer to this as ordered combinations of N things taken r at a time. The notation ${}_NP_r$ is sometimes used, and may appear on your calculator using similar notation. Note also that ${}_NP_N = N!$.

EXAMPLE. Candidates A, B, C, and D are running for office. Indicate your first and second choice.

There are 12 possible choices:

AB BA AC CA AD DA BC CB BD DB CD DC.

To confirm this, using counting Rule #4, we get

$$\frac{N!}{(N-r)!} = \frac{4!}{(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)} = 12$$

EXAMPLE. There are 100 applicants for 3 job openings. Indicate your first, second, and third choices.

If you try to list all the possible permutations, you will be busy for a very long time. If you try to calculate 100! and 97! by hand, you'd better plan on being in grad school a long time. And, even a calculator with a factorial function may not help you; mine just produces an error message when I give it numbers this big. If, on the other hand, you take advantage of the fact that most terms in the numerator and denominator cancel out, e.g.

$$\frac{N!}{(N-r)!} = \frac{100!}{(100-3)!} = \frac{(100)(99)(98)(97)(96)...(1)}{(97)(96)...(1)} = 100*99*98 = 970{,}200$$

you will find that it is pretty easy to figure out the number of ordered combinations.

NOTE: You could also use Rule 3B. People are divided into 4 categories: the first best ($N_1 = 1$), the 2nd best ($N_2 = 1$), the 3rd best ($N_3 = 1$) and all the rest ($N_4 = 97$). Hence, rule 3B yields 100!/(1!1!1!97!) = 100!/97!

5. *COMBINATIONS*. Often, we are not interested in the order of events, but only in the number of ways r things could be selected from N things, irrespective of order. From rule 4, we know that the total number of ways of selecting r things from N and ordering them is N!/(N-r)! From rule 3, we know that each set of r objects has r! orderings. Therefore,

The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is

$$\frac{N!}{r!(N-r)!} = \binom{N}{r} = \binom{N}{N-r}$$

We refer to this as "N choose r." Sometimes the number of combinations is known as a *binomial coefficient*, and sometimes the notation ${}_NC_r$ is used. Note that the number of permutations is r! times larger than the number of combinations.

EXAMPLE. Candidates A, B, C and D are running for office. Vote for two.

The possible choices are:

AB AC AD BC BD CD, i.e. there are 6 possible combinations. Confirming this with rule #5, we get

$$\frac{N!}{r!(N-r)!} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{12}{2} = 6$$

EXAMPLE. There are 100 applicants for 3 job openings. The number of possible combinations is

$$\frac{N!}{r!(N-r)!} = \binom{100}{3} = \frac{100!}{3!97!} = \frac{970{,}200}{6} = 161{,}700$$

NOTE: You could also use rule 3B. People are divided into 2 groups: the 3 best ($N_1 = 3$) and all the rest ($N_2 = 97$). Hence, rule 3B yields 100!/(3!97!)

See Appendix E, Table 7, page 20 for ${}_NC_r$ values for various values of N and r. (Or see Hayes, Appendix E, Table IX, p. 948). Your calculator may have a function labeled nCr or something similar.

6. Binomial distribution. **In sampling from a stationary Bernoulli process, with the probability of success equal to p, the probability of observing exactly r successes in N independent trials is**

$$\binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}$$

7. Binomially distributed variables. **Let Xi = 1 if the ith bernoulli trial is successful, 0 otherwise. If $X = \Sigma X_i$, where the X_i's are independent and identically distributed (iid), then X has a binomial distribution, and E(X) = Np, V(X) = Npq.**

SUMMARY OF HIGHLIGHTS

1. NUMBER OF POSSIBLE SEQUENCES FOR N TRIALS. **If any one of K mutually exclusive and exhaustive events can occur on each of N trials, then there are K^n different sequences that may result from a set of such trials.**

2. SEQUENCES. **If $K_1,...,K_N$ are the numbers of distinct events that can occur on trials 1,..., N in a series, then the number of different sequences of N events that can occur is $(K_1)(K_2)...(K_N)$.**

3. PERMUTATIONS **The number of different ways that N distinct things may be arranged in order is**

N! = (1)(2)(3)...(N-1)(N), (where 0! = 1).

3B. PERMUTATIONS OF SIMILAR OBJECTS. **Suppose we have N objects, N_1 alike, N_2 alike,..., N_k alike ($\Sigma N_i = N$). Then, the number of ways of arranging these objects is**

$$\frac{N!}{N_1!\,N_2!...N_k!}$$

4. ORDERED COMBINATIONS; or, PERMUTATIONS OF N OBJECTS TAKEN r AT A TIME. **The number of ways of selecting & arranging r objects from among N distinct objects is**

$$\frac{N!}{(N-r)!}$$

5. COMBINATIONS. **The total number of ways of selecting r distinct combinations of N objects, irrespective of order, is**

$$\frac{N!}{r!(N-r)!} = \binom{N}{r} = \binom{N}{N-r}$$

6. Binomial distribution. **In sampling from a stationary Bernoulli process, with the probability of success equal to p, the probability of observing exactly r successes in N independent trials is**

$$\binom{N}{r} p^r q^{N-r} = \frac{N!}{r!(N-r)!} p^r q^{N-r}$$

7. Binomially distributed variables. **Let Xi = 1 if the ith bernoulli trial is successful, 0 otherwise. If $X = \Sigma X_i$, where the X_i's are independent and identically distributed (iid), then X has a binomial distribution, and E(X) = Np, V(X) = Npq.**

5 BINOMIAL DISTRIBUTION

Objectives

After studying this chapter you should

- be able to recognise when to use the binomial distribution;
- understand how to find the mean and variance of the distribution;
- be able to apply the binomial distribution to a variety of problems.

Note: Statistical tables can be found in many books and are also available online.

5.0 Introduction

'Bi' at the beginning of a word generally denotes the fact that the meaning involves 'two' and binomial is no exception. A random variable follows a binomial distribution when each trial has exactly **two possible outcomes**. For example, when Sarah, a practised archer, shoots an arrow at a target she either hits or misses each time. If X is 'the number of hits Sarah scores in ten shots', then the probabilities associated with 0, 1, 2, ..., 10 hits can be expected to follow a particular pattern, known as the **binomial distribution**.

5.1 Finding the distribution

You have already met this type of distribution in Chapter 4, as can be seen in the following example.

Example

Ashoke, Theo and Sadie will each visit the local leisure centre to swim on one evening next week but have made no arrangement between themselves to meet or go on any particular day. The random variable X is 'the number of the three who go to the leisure centre on Wednesday'. Find the probability distribution for X.

Solution

The probabilities of 0, 1, 2 or 3 people going on Wednesday can be found by using the tree diagram method covered in Section 1.5.

The following tree diagram shows probabilities for how many go on Wednesday.

X	Prob.
3	$\frac{1}{7}\times\frac{1}{7}\times\frac{1}{7}$
2	$\frac{1}{7}\times\frac{1}{7}\times\frac{6}{7}$
2	$\frac{1}{7}\times\frac{6}{7}\times\frac{1}{7}$
1	$\frac{1}{7}\times\frac{6}{7}\times\frac{6}{7}$
2	$\frac{6}{7}\times\frac{1}{7}\times\frac{1}{7}$
1	$\frac{6}{7}\times\frac{1}{7}\times\frac{6}{7}$
1	$\frac{6}{7}\times\frac{6}{7}\times\frac{1}{7}$
0	$\frac{6}{7}\times\frac{6}{7}\times\frac{6}{7}$

You can see that

$$P(X=3)=\left(\frac{1}{7}\right)^3, \quad P(X=0) = \left(\frac{6}{7}\right)^3$$

and $$P(X=2)=\frac{1}{7}\times\frac{1}{7}\times\frac{6}{7} + \frac{1}{7}\times\frac{6}{7}\times\frac{1}{7} + \frac{6}{7}\times\frac{1}{7}\times\frac{1}{7}$$

$$= 3\times\left(\frac{1}{7}\right)^2\left(\frac{6}{7}\right)$$

$$P(X=1)=\frac{1}{7}\times\frac{6}{7}\times\frac{6}{7} + \frac{6}{7}\times\frac{1}{7}\times\frac{6}{7} + \frac{6}{7}\times\frac{6}{7}\times\frac{1}{7}$$

$$= 3\times\left(\frac{1}{7}\right)\left(\frac{6}{7}\right)^2.$$

This gives the table below:

x	0	1	2	3
$P(X=x)$	$\left(\frac{6}{7}\right)^3$	$3\left(\frac{1}{7}\right)\left(\frac{6}{7}\right)^2$	$3\left(\frac{1}{7}\right)^2\left(\frac{6}{7}\right)$	$\left(\frac{1}{7}\right)^3$
	$=\frac{216}{343}$	$=\frac{108}{343}$	$=\frac{18}{343}$	$=\frac{1}{343}$

The method used in the example above can be extended to a fourth person so that there will be sixteen branches to cover all the possibilities as shown in the diagram below, with X now 'the number of the four people who go on Wednesday'.

.

The resulting probability distribution is

x	0	1	2	3	4
$P(X=x)$	$\left(\frac{6}{7}\right)^4$	$4\left(\frac{1}{7}\right)\left(\frac{6}{7}\right)^3$	$6\left(\frac{1}{7}\right)^2\left(\frac{6}{7}\right)^2$	$4\left(\frac{1}{7}\right)^3\left(\frac{6}{7}\right)$	$\left(\frac{1}{7}\right)^4$

The fractions are as you might expect. For instance, looking at

$$P(X=2)=6\left(\frac{1}{7}\right)^2\left(\frac{6}{7}\right)^2,$$

since you are interested in having two present and the probability for each is $\frac{1}{7}$, the $\left(\frac{1}{7}\right)^2$ is explained, and as two are not to attend this produces the $\left(\frac{6}{7}\right)^2$.

Explain the reason for the coefficient 6.

This has come from the second of the tree diagrams: there are six branches corresponding to two being present. This is because there are six ways of writing down two $\left(\frac{1}{7}\right)$s and two $\left(\frac{6}{7}\right)$s in a row and each produces a branch.

The six ways are shown below:

$$\frac{1}{7}\times\frac{1}{7}\times\frac{6}{7}\times\frac{6}{7},\quad \frac{1}{7}\times\frac{6}{7}\times\frac{1}{7}\times\frac{6}{7},\quad \frac{1}{7}\times\frac{6}{7}\times\frac{6}{7}\times\frac{1}{7},$$

$$\frac{6}{7}\times\frac{1}{7}\times\frac{1}{7}\times\frac{6}{7},\quad \frac{6}{7}\times\frac{1}{7}\times\frac{6}{7}\times\frac{1}{7},\quad \frac{6}{7}\times\frac{6}{7}\times\frac{1}{7}\times\frac{1}{7}.$$

There are four fractions to write down, two of each type, and the number of different ways of combining these is six. You don't want to draw a tree diagram every time so another method can be developed. For example, having ten people going to the leisure centre would need a tree with $2^{10} = 1024$ branches.

If you can produce the fractions in the probability distribution then all that is needed is a way of getting the right numbers to put with them. With three people the numbers were 1, 3, 3, 1 and with four they were 1, 4, 6, 4, 1.

You might recognise these as being rows of **Pascal's Triangle**, shown opposite.

Pascal's Triangle

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

1 9 36 84 126 126 84 36 9 1

1 10 45 120 210 252 210 120 45 10 1

For three to be present you need $\left(\frac{1}{7}\right)^3$ and for seven to stay away the term is $\left(\frac{6}{7}\right)^7$. When writing down these fractions, how many different ways are there of combining them? According to Pascal's Triangle there are 120, so

$$P(X=3) = 120\left(\frac{1}{7}\right)^3\left(\frac{6}{7}\right)^7$$

which is easier to calculate than drawing 1024 branches! The number of ways of choosing 3 from 10 is often written

$$\binom{10}{3} \quad \text{or} \quad {}^{10}C_3,$$

so $$p(X=3) = \binom{10}{3}\left(\frac{1}{7}\right)^3\left(\frac{6}{7}\right)^7$$

and the probability distribution is

$$P(X=x) = \binom{10}{x}\left(\frac{1}{7}\right)^x\left(\frac{6}{7}\right)^{10-x} \quad x = 0,\ 1,\ \ldots,\ 10.$$

Whilst the values needed can easily be read off Pascal's Triangle, there is an even easier way of working out the coefficients given in terms of factorials. Note that $n! = n(n-1)\ldots 2.1$ and, for example,

$$\begin{aligned}\binom{10}{3} &= \frac{10!}{7!\,3!}\\ &= \frac{10\times 9\times 8\times(7\times 6\times 5\times 4\times 3\times 2\times 1)}{(7\times 6\times 5\times 4\times 3\times 2\times 1)\times(3\times 2\times 1)}\\ &= \frac{10\times 9\times 8}{3\times 2\times 1} \qquad \left(\text{since } \frac{10!}{7!} = 10\times 9\times 8\right)\\ &= 120.\end{aligned}$$

Activity 1

Check the values of

$$\binom{10}{1},\ \binom{10}{2},\ \binom{10}{4} \text{ and } \binom{10}{5}$$

by using Pascal's Triangle and the factorial formula.

It should also be noted that most calculators have the facility to produce the coefficients. Another notation used for the number of different ways of combining three from ten is ${}^{10}C_3$, and the majority of calculators have a key labelled with one of ${}^{n}C_r$ or $\binom{n}{r}$. Your instruction book will tell you how to use this function.

In general, when the probability of success is p (instead of $\frac{1}{7}$), and the experiment is repeated n independent times, the probabilitry distribution for the number of successes is given by

$$\boxed{P(X=x) = \binom{n}{x}p^x(1-p)^{n-x} \qquad x = 0,\ 1,\ \ldots\ldots n}$$

The notation $X \sim B(n,p)$ is often used.

Activity 2 E.S.P. test

Take five cards numbered from one to five. Seat two people back to back and give the cards to one of them. This person selects a card at random and the other participant tries to identify it.

This is done five times and repeated with other pairs of people. Record the results.

Now, if X is 'the number predicted correctly out of five attempts', the probability distribution is

$$P(X=x) = \binom{5}{x}(0.2)^x(0.8)^{5-x} \quad \text{for } x = 0,\ 1,\ \ldots\ 5.$$

So, for example, the probability of getting one correct is given by

$$\begin{aligned} P(X=1) &= \binom{5}{1}(0.2)(0.8)^4 \\ &= 5(0.2)\,(0.8)^4 \\ &= 0.4096. \end{aligned}$$

Hence for 20 people, you would expect

$$20 \times 0.4096 = 8.192 \approx 8$$

to get one correct. A table of expected frequencies is shown opposite.

Number correct (x)	$p\,(x)$	Expected frequencies $20\ p\,(x)$
0	0.3277	6.554
1	0.4096	8.192
2	0.2048	4.096
3	0.0512	1.024
4	0.0064	0.128
5	0.0032	0.064

Compare your observed frequencies from Activity 2.

How closely do they match the expected frequencies?

You may want to discard results from anyone who you feel has not co-operated, perhaps saying the same number every time.

Example

If X is binomially distributed with 6 trials and a probability of success equal to $\frac{1}{4}$ at each attempt, what is the probability of:

(a) exactly 4 successes (b) at least one success?

Solution

This question can be rewritten in the following way.

If $X \sim B\left(6, \frac{1}{4}\right)$, what is: (a) $P(X=4)$ (b) $P(X \geq 1)$?

(a) $$P(X=4) = \binom{6}{4}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^2$$

$$= 15 \times \frac{1}{256} \times \frac{9}{16}$$

$$= \frac{135}{4096} = 0.033 \quad (\text{to 3 d.p.}).$$

(b) $$P(X \geq 1) = 1 - P(X=0)$$

$$= 1 - \left(\frac{3}{4}\right)^6$$

$$= 1 - \frac{729}{4096}$$

$$= \frac{3367}{4096} = 0.822 \quad (\text{to 3 d.p.}).$$

Note that tables giving cumulative binomial probabilities are given in the Appendix (p 253) and these can be used where appropriate.

Example

When an unbiased coin is tossed eight times what is the probability of obtaining:

(a) less than 4 heads (b) more than five heads?

Solution

H = number of heads $\Rightarrow H \sim B(8,\ 0.5)$

(a) Using the appropriate table in the Appendix you can simply write down the answer,

$$P(H<4) = P(H \leq 3)$$

$$= 0.3633.$$

Alternatively,

$$P(H \le 3) = P(H=0) + P(H=1) + P(H=2) + P(H=3)$$

$$= \left(\frac{1}{2}\right)^8 + \binom{8}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^7 + \binom{8}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^6 + \binom{8}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^5$$

$$= \left(\frac{1}{2}\right)^8 + 8\left(\frac{1}{2}\right)^8 + 28\left(\frac{1}{2}\right)^8 + 56\left(\frac{1}{2}\right)^8$$

$$= 93\left(\frac{1}{2}\right)^8$$

$$= \frac{93}{256}$$

$$= 0.3633 \text{ (to 4 d.p.)}.$$

(b) $P(H > 5) = 1 - P(H \le 5)$

$= 1 - 0.8555$ (from the table)

$= 0.1445$.

or

$$P(H > 5) = P(H=6) + P(H=7) + P(H=8)$$

$$= \binom{8}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 + \binom{8}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^8$$

$$= 28\left(\frac{1}{2}\right)^8 + 8\left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^8$$

$$= 37\left(\frac{1}{2}\right)^8$$

$$= \frac{37}{256}$$

$$= 0.1445 \text{ (to 4 d.p.)}.$$

Discuss why it is true that $\binom{8}{2}$ is the same as $\binom{8}{6}$.

Will it always be true that $\binom{n}{r} = \binom{n}{n-r}$?

Exercise 5A

Where decimals are used give answers correct to 3 significant figures.

1. If $X \sim B\left(6, \frac{1}{3}\right)$ find:
   (a) $P(X=2)$ (b) $P(X<2)$ (c) $P(X \geq 1)$.
2. If $X \sim B(10, 0.3)$ find:
   (a) $P(X=9)$ (b) $P(X=0)$ (c) $P(X \leq 5)$.
3. A regular tetrahedron has three white faces and one red face. It is rolled four times and the colour of the bottom face is noted. What is the most likely number of times that the red face will end downwards?
4. If the probability that I get a lift to work on any morning is 0.6 what is the probability that in a working week of five days I will get a lift only twice?
5. When a consignment of pens arrives at the retailer's, ten of them are tested. The whole batch is returned to the wholesaler if more than one of those selected is found to be faulty. What is the probability that the consignment will be accepted if 2% of the pens are faulty?

5.2 The mean and variance of the binomial distribution

If you play ten games of table tennis against an opponent who, from past experience, you know only has a $\frac{1}{5}$ chance of winning a game with you, how many games do you expect him to win?

Most people would reply 'two' and would argue that since the opponent wins on average $\frac{1}{5}$ of the games he can expect to be successful in $\frac{1}{5} \times 10 = 2$.

Another way of writing this would be to say, if $X \sim B\left(10, \frac{1}{5}\right)$, what is the value of $E(X)$? The answer then is $E(X) = 10 \times \frac{1}{5} = 2$.

In general, if $X \sim B(n, p)$, then the expected value of X is given by

$$\boxed{E(X) = np}$$

The formal proof of this result requires some work from pure maths, and in particular uses the result that

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

Note that it will **not** be examined in the AEB Statistics paper.

Now $E(X) = \sum_{0}^{n} x\,p(x)$

$$= 0\binom{n}{0}q^n + 1\binom{n}{1}pq^{n-1} + 2\binom{n}{2}p^2q^{n-2}$$

$$+ 3\binom{n}{3}p^3q^{n-3} + \ldots + n\binom{n}{n}p^n, \text{ where } q = 1-p$$

$$= \frac{n!}{1!(n-1)!}pq^{n-1} + \frac{2\times n!}{2!(n-2!)}p^2q^{n-2}$$

$$+ \frac{3\times n!}{3!(n-3)!}p^3q^{n-3} + \ldots + np^n$$

$$= np\left\{\frac{(n-1)!}{1!(n-1)!}q^{n-1} + \frac{(n-1)!}{1!(n-2)!}pq^{n-2}\right.$$

$$\left. + \frac{(n-1)!}{2!(n-3)!}p^2q^{n-3} + \ldots + p^{n-1}\right\}$$

$$= np\left\{q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{2}q^{n-3}p^2 + \right.$$

$$\left. \ldots + p^{n-1}\right\}$$

$$= np(p+q)^{n-1}, \quad \text{using the Binomial Theorem}$$

$$= np, \text{ since } p+q=1.$$

The variance of X is given by

$$\boxed{V(X) = npq}$$

The proof is even more complex than the analysis above, so it is set in the next Activity, which is optional.

*Activity 3

Using the formula

$$V(X) = E\left(X^2\right) - \left[E(X)\right]^2$$

show that

$$V(X) = npq.$$

*Activity 4

Use a computer package designed to show graphs of binomial distributions for different values of n and p to look at a variety of binomial distributions. In particular, identify the most likely outcome by picking out the tallest bar and see, for example, how much more spread out are the outcomes for B (40, 0.5) than B (10, 0.5).

Example

A biased die is thrown thirty times and the number of sixes seen is eight. If the die is thrown a further twelve times find:

(a) the probability that a six will occur exactly twice;

(b) the expected number of sixes;

(c) the variance of the number of sixes.

Solution

(a) Let X be defined by 'the number of sixes seen in twelve throws'.

Then $X \sim B\left(12, p\right)$ where $p = \dfrac{8}{30} = \dfrac{4}{15}$.

Since $X \sim B\left(12, \frac{4}{15}\right)$,

$$P(X=2) = \binom{12}{2}\left(\frac{4}{15}\right)^2\left(\frac{11}{15}\right)^{10}$$

$$= \frac{66 \times 4^2 \times 11^{10}}{15^{12}}$$

$$= 0.211 \quad \text{(to 3 d.p.).}$$

(b) $E(X) = np = 12 \times \frac{4}{15} = 3.2$.

(c) $V(X) = npq = 12 \times \frac{4}{15} \times \frac{11}{15} = 2\frac{26}{75} = 2.347$ (to 3 d.p.).

Example

A random variable X is binomially distributed with mean 6 and variance 4.2. Find $P(X \le 6)$.

Solution

Since X is a binomial distribution,

$$\text{mean} = np = 6,$$

$$\text{variance} = npq = 4.2.$$

Dividing, $q = \frac{4.2}{6} = 0.7$

and so $p = 1 - q = 0.3$.

This gives $0.3n = 6$

$\Rightarrow$ $n = 20$.

Hence $X \sim B(20, 0.3) \Rightarrow P(X \le 6) = 0.6080$ (from tables).

Activity 5 Binomial quiz

Ask your fellow students, and anyone else who will participate, whether the following statements are 'true' or 'false'.

1. *The Portrait of a Lady* was written by Henry James.
2. Psalms is the 20th book of the Old Testament.
3. The equatorial diameter of Mercury is about 3032 miles.
4. Mankoya is a place in Zambia.
5. 'The Potato Eaters' is a painting by Cezanne.
6. The Battle of Sowton was fought in 1461.

Make a frequency table to show the number of correct answers out of six for those asked.

Is there any evidence from your results that people really know some or all of the answers?

If they are just guessing, the number of correct answers, C say,

should follow a binomial distribution,

$$C \sim B\left(6, \tfrac{1}{2}\right).$$

Work out $P(C=0)$, $P(C=1)$, …., $P(C=6)$ and multiply them by the number of people asked to get the frequencies with which you would expect 0, 1,...., 6 correct answers to occur. Draw a diagram to show how your observed and expected frequencies compare.

Note that tables of **cumulative** binomial probabilities are available. For example, $B(10,\ 0.1)$ has entries as shown opposite.

x	$P(X \le x)$
0	0.3487
1	0.7361
2	0.9298
3	0.9872
4	0.9984
5	0.9999
6	1.0000

So if X is distributed $B(10, 0.1)$, then $P(X \le 3) = 0.9872$.

But
$$\begin{aligned} P(X=3) &= P(X \le 3) - P(X \le 2) \\ &= 0.9872 - 0.9298 \\ &= 0.0574. \end{aligned}$$

This figure may be inaccurate in the last digit as it has come from two numbers which have both been rounded. Using tables usually saves having to do several calculations and the benefit is considerable in cases such as $Y \sim B(30,\ 0.1)$ when the value of quantities like $P(Y \le 8)$ are needed.

Exercise 5B

1. On average a bowler takes a wicket every eight overs. What is the probability that he will bowl ten overs without succeeding in getting a wicket?
2. How many times must an unbiased coin be tossed so that the probability that at least one tail will occur is at least 0.99?
3. The random variable X has a binomial distribution $B(11, p)$. If $P(X=8) = P(X=7)$ find the value of p.
4. 100 families each with three children are found to have the following number of boys.

Number of boys	0	1	2	3
Frequency	13	34	40	13

   (a) Find the probability that a single baby born is a boy.

   (b) Calculate the number of families with three children you would expect to have two boys in a sample of 100 using your value from (a).
5. A multiple choice test has twenty questions and five possible answers for each one with only one correct per question. If X is 'the number of questions answered correctly' give:

   (a) the distribution of X;

   (b) the mean and variance of X;

   (c) the probability that a student will achieve a pass mark of 10 or more purely by guessing.
6. Investigate your results from the 'drawing pins' question in Exercise1B to see if X, the number of times a pin finishes point upwards in 10 trials, follows a binomial distribution.
7. The probability that a student will pass a maths test is 0.8. If eighteen students take the test, give the distribution of X, 'the number of students who pass', and find its most likely value.

Correlation

If the change in one variable affects a change in the value of the other variable, the variables are said to be correlated.

- **Coefficient of correlation**

  It is a numerical measure of linear relationship of two variables x and y, it is denoted by $r(x, y)$.

That is, $r(x,y) = \dfrac{\dfrac{\sum xy}{n} - \bar{x}.\bar{y}}{\sqrt{\dfrac{\sum x^2}{n} - (\bar{x})^2}.\sqrt{\dfrac{\sum y^2}{n} - (\bar{y})^2}}$ where $\bar{x} = \dfrac{\sum x}{n}$, $\bar{y} = \dfrac{\sum y}{n}$.

Problem

1. Calculate the correlation coefficient for the following data

x	1	3	5	8	9
y	3	4	8	10	12

x	y	xy	x^2	y^2
1	3	3	1	9
3	4	12	9	16
5	8	40	25	64
8	10	80	64	100
9	12	108	81	144
10	11	110	100	121
36	**48**	**353**	**280**	**454**

Now, $\bar{x} = \dfrac{\sum x}{n} = \dfrac{36}{6} = 6$, $\bar{y} = \dfrac{\sum y}{n} = \dfrac{48}{6} = 8$

$$r(x,y) = \frac{\dfrac{\sum xy}{n} - \bar{x}.\bar{y}}{\sqrt{\dfrac{\sum x^2}{n} - (\bar{x})^2}.\sqrt{\dfrac{\sum y^2}{n} - (\bar{y})^2}} = \frac{\dfrac{353}{6} - 48}{\sqrt{\dfrac{280}{6} - 36}.\sqrt{\dfrac{454}{6} - 64}} = 0.97$$

Practice Problem

1. Find the correlation coefficient for the following data:

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Solution: $r(x,y) = 0.603$

- **Rank Correlation Coefficient:**

Rank correlation coefficient, is given by

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i = (x_i - y_i)$, that is difference between the ranks.

Problem

1. The rankings of ten students in two subjects A and B are as follows:

A	3	5	8	4	7	10	2	1	6	9
B	6	4	9	8	1	2	3	10	5	7

Find the rank correlation coefficient.

Solution:

A(x_i)	B(y_i)	$d_i = (x_i - y_i)$	d_i^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
		0	$\Sigma d_i^2 = 214$

The rank correlation coefficient is $r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(214)}{10(10^2-1)} = -0.297$ =

Note: When we have repeated ranks, we have to use correction factor $= \frac{m(m^2-1)}{12}$

Where m represents the number of times the data repeated.

2. The following table gives the number of units rejected by two operators X and Y in 8 inspections:

X	15	20	28	12	40	60	20	80
Y	40	30	50	30	20	10	30	60

Obtain rank correlation coefficient between X and Y with respect to the quality of the product.

Solution:

X	Y	Ranks in X (x_i)	Ranks in Y (y_i)	$d_i = (x_i - y_i)$	d_i^2
15	40	2	6	-4	16
20	30	3.5	4	-0.5	0.25
28	50	5	7	-2	4
12	30	1	4	-3	9
40	20	6	2	4	16
60	10	7	1	6	36
20	30	3.5	4	-0.5	0.25
80	60	8	8	0	0
					$\sum d_i^2 = 81.5$

In X series 20 repeated twice, correction factor $= \dfrac{m(m^2-1)}{12} = \dfrac{2(2^2-1)}{12} = \dfrac{1}{2}$

In Y series 30 repeated thrice, correction factor $= \dfrac{m(m^2-1)}{12} = \dfrac{3(3^2-1)}{12} = 2$

Therefore, $r(X,Y) = 1 - \dfrac{6\left[81.5 + \frac{1}{2} + 2\right]}{8(8^2-1)} = 0$

Practice Problem:

1. The marks secured by recruits in the selection test X and in the proficiency test Y are given below:

Serial No.	1	2	3	4	5	6	7	8	9
X	10	15	12	17	13	16	24	14	22
Y	30	42	45	46	33	34	40	35	39

Calculate the rank correlation coefficient.
Solution:0.4

2. Find the rank correlation coefficient for the following.

X	92	89	87	86	86	77	71	63	53	50
Y	86	83	91	77	68	85	52	82	37	57

Solution: 0.727

3. Following are the marks obtained by 10 students in a class in two tests.

Students	A	B	C	D	E	F	G	H	I	J
Test 1	70	68	67	55	60	60	75	63	60	72
Test 2	65	65	80	60	68	58	75	63	60	70

Find the rank correlation for tied observations.
Solution: r = 0.68.

21

The Exponential Distribution

From Discrete-Time to Continuous-Time:

In Chapter 6 of the text we will be considering Markov processes in continuous time. In a sense, we already have a very good understanding of continuous-time Markov chains based on our theory for discrete-time Markov chains. For example, one way to describe a continuous-time Markov chain is to say that it is a discrete-time Markov chain, except that we explicitly model the times between transitions with continuous, positive-valued random variables and we explicity consider the process at any time t, not just at transition times.

The single most important continuous distribution for building and understanding continuous-time Markov chains is the exponential distribution, for reasons which we shall explore in this lecture.

The Exponential Distribution:

A continuous random variable X is said to have an Exponential(λ) distribution if it has probability density function

$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases},$$

where $\lambda > 0$ is called the *rate* of the distribution.

In the study of continuous-time stochastic processes, the exponential distribution is usually used to model the *time until something happens in the process*. The mean of the Exponential(λ) distribution is calculated using integration by parts as

$$\begin{aligned} E[X] &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \lambda \left[\frac{-xe^{-\lambda x}}{\lambda}\bigg|_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x} dx \right] \\ &= \lambda \left[0 + \frac{1}{\lambda}\frac{-e^{-\lambda x}}{\lambda}\bigg|_0^\infty \right] \\ &= \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda}. \end{aligned}$$

So one can see that as λ gets larger, the thing in the process we're waiting for to happen tends to happen more quickly, hence we think of λ as a rate.

As an exercise, you may wish to verify that by applying integration by parts twice, the second moment of the Exponential(λ) distribution is given by

$$E[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} = \ldots = \frac{2}{\lambda^2}.$$

From the first and second moments we can compute the variance as

$$\mathsf{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

The Memoryless Property:

The following plot illustrates a key property of the exponential distribution. The graph after the point s is an exact copy of the original function. The important consequence of this is that the distribution of X conditioned on $\{X > s\}$ is *again exponential*.

Figure 21.1: The Exponential Function e^{-x}

To see how this works, imagine that at time 0 we start an alarm clock which will ring after a time X that is exponentially distributed with rate λ. Let us call X the *lifetime* of the clock. For any $t > 0$, we have that

$$P(X > t) = \int_t^\infty \lambda e^{-\lambda x} dx = \lambda \left. \frac{-e^{-\lambda x}}{\lambda} \right|_t^\infty = e^{-\lambda t}.$$

Now we go away and come back at time s to discover that the alarm has not yet gone off. That is, we have observed the event $\{X > s\}$. If we let Y denote the *remaining* lifetime of the clock given that $\{X > s\}$, then

$$\begin{aligned} P(Y > t | X > s) &= P(X > s + t | X > s) \\ &= \frac{P(X > s + t, X > s)}{P(X > s)} \\ &= \frac{P(X > s + t)}{P(X > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t}. \end{aligned}$$

But this implies that the remaining lifetime after we observe the alarm has not yet gone off at time s has the same distribution as the original lifetime X. The really important thing to note, though, is that this implies that the distribution of the remaining lifetime *does not depend on* s. In fact, if you try setting X to have *any other* continuous distribution, then ask what would be the distribution of the remaining lifetime after you observe $\{X > s\}$, the distribution will depend on s.

This property is called the *memoryless* property of the exponential distribution because I don't need to remember when I started the clock. If the distribution of the lifetime X is Exponential(λ), then if I come back to the clock at any time and observe that the clock has not yet gone off, regardless of when the clock started I can assert that the distribution of the time till it goes off, starting at the time I start observing it again, is Exponential(λ). Put another way, given that the clock has currently not yet gone off, I can forget the past and still know the distribution of the time from my current time to the time the alarm will go off. The resemblance of this property to the Markov property should not be lost on you.

It is a rather amazing, and perhaps unfortunate, fact that the exponential distribution is the only one for which this works. The memoryless property is like enabling technology for the construction of continuous-time Markov chains. We will see this more clearly in Chapter 6. But the exponential distribution is even more special than just the memoryless property because it has a second enabling type of property.

Weibull Distribution

Weibull Distribution

Definition

A random variable X is said to have a **Weibull distribution** with parameters α and β ($\alpha > 0, \beta > 0$) if the pdf of X is

$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Weibull Distribution

Definition

A random variable X is said to have a **Weibull distribution** with parameters α and β ($\alpha > 0, \beta > 0$) if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Remark:

1. The family of Weibull distributions was introduced by the Swedish physicist Waloddi Weibull in 1939.

Weibull Distribution

Definition

A random variable X is said to have a **Weibull distribution** with parameters α and β ($\alpha > 0, \beta > 0$) if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Remark:

1. The family of Weibull distributions was introduced by the Swedish physicist Waloddi Weibull in 1939.
2. We use $X \sim \text{WEB}(\alpha, \beta)$ to denote that the rv X has a Weibull distribution with parameters α and β.

Weibull Distribution

Weibull Distribution

Remark:

Weibull Distribution

Remark:

3. When $\alpha = 1$, the pdf becomes

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

which is the pdf for an exponential distribution with parameter $\lambda = \frac{1}{\beta}$. Thus we see that the exponential distribution is a special case of both the gamma and Weibull distributions.

Weibull Distribution

Remark:
3. When $\alpha = 1$, the pdf becomes

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

which is the pdf for an exponential distribution with parameter $\lambda = \frac{1}{\beta}$. Thus we see that the exponential distribution is a special case of both the gamma and Weibull distributions.
4. There are gamma distributions that are not Weibull distributios and vice versa, so one family is not a subset of the other.

Weibull Distribution

Weibull Distribution

Weibull Distribution

Weibull Distribution

Weibull Distribution

Weibull Distribution

Proposition

Let X be a random variable such that $X \sim WEI(\alpha, \beta)$. Then

$$E(X) = \beta\Gamma\left(1 + \frac{1}{\alpha}\right) \quad \textit{and} \quad V(X) = \beta^2\left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right\}$$

The cdf of X is

$$F(x; \alpha, \beta) = \begin{cases} 1 - e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Weibull Distribution

Weibull Distribution

Example:
The shear strength (in pounds) of a spot weld is a Weibull distributed random variable, $X \sim WEB(400, 2/3)$.

a. *Find* $P(X > 410)$.

Weibull Distribution

Example:
The shear strength (in pounds) of a spot weld is a Weibull distributed random variable, $X \sim WEB(400, 2/3)$.

a. *Find* $P(X > 410)$.
b. *Find* $P(X > 410 \mid X > 390)$.

Weibull Distribution

Example:
The shear strength (in pounds) of a spot weld is a Weibull distributed random variable, $X \sim WEB(400, 2/3)$.

a. *Find* $P(X > 410)$.
b. *Find* $P(X > 410 \mid X > 390)$.
c. *Find* $E(X)$ *and* $V(X)$.

Weibull Distribution

Example:
The shear strength (in pounds) of a spot weld is a Weibull distributed random variable, $X \sim WEB(400, 2/3)$.

a. *Find* $P(X > 410)$.
b. *Find* $P(X > 410 \mid X > 390)$.
c. *Find* $E(X)$ *and* $V(X)$.
d. *Find the* $95th$ *percentile.*

Weibull Distribution

Weibull Distribution

In practical situations, $\gamma = \min(X) > 0$ and $X - \gamma$ has a Weibull distribution.

Weibull Distribution

In practical situations, $\gamma = \min(X) > 0$ and $X - \gamma$ has a Weibull distribution.

Example (Problem 74):

Let $X =$ the time (in 10^{-1} weeks) from shipment of a defective product until the customer returns the product. Suppose that the minimum return time is $\gamma = 3.5$ and that the excess $X - 3.5$ over the minimum has a Weibull distribution with parameters $\alpha = 2$ and $\beta = 1.5$.

a. *What is the cdf of X?*

Weibull Distribution

In practical situations, $\gamma = \min(X) > 0$ and $X - \gamma$ has a Weibull distribution.

Example (Problem 74):

Let $X =$ the time (in 10^{-1} weeks) from shipment of a defective product until the customer returns the product. Suppose that the minimum return time is $\gamma = 3.5$ and that the excess $X - 3.5$ over the minimum has a Weibull distribution with parameters $\alpha = 2$ and $\beta = 1.5$.

a. *What is the cdf of X?*
b. *What are the expected return time and variance of return time?*

Weibull Distribution

In practical situations, $\gamma = \min(X) > 0$ and $X - \gamma$ has a Weibull distribution.

Example (Problem 74):

Let $X =$ the time (in 10^{-1} weeks) from shipment of a defective product until the customer returns the product. Suppose that the minimum return time is $\gamma = 3.5$ and that the excess $X - 3.5$ over the minimum has a Weibull distribution with parameters $\alpha = 2$ and $\beta = 1.5$.

a. *What is the cdf of X?*
b. *What are the expected return time and variance of return time?*
c. *Compute $P(X > 5)$.*

Weibull Distribution

In practical situations, $\gamma = \min(X) > 0$ and $X - \gamma$ has a Weibull distribution.

Example (Problem 74):

Let $X =$ the time (in 10^{-1} weeks) from shipment of a defective product until the customer returns the product. Suppose that the minimum return time is $\gamma = 3.5$ and that the excess $X - 3.5$ over the minimum has a Weibull distribution with parameters $\alpha = 2$ and $\beta = 1.5$.

a. *What is the cdf of X?*
b. *What are the expected return time and variance of return time?*
c. *Compute $P(X > 5)$.*
d. *Compute $P(5 \leq X \leq 8)$.*

Lognormal Distribution

Lognormal Distribution

Definition

A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution. The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with parameters μ and σ is

$$f(x; \mu, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & x \leq 0 \\ 0 & x < 0 \end{cases}$$

Lognormal Distribution

Definition

A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution. The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with parameters μ and σ is

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & x \leq 0 \\ 0 & x < 0 \end{cases}$$

Remark:

1. We use $X \sim \text{LOGN}(\mu, \sigma^2)$ to denote that rv X have a lognormal distribution with parameters μ and σ.

Lognormal Distribution

Definition

A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution. The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with parameters μ and σ is

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & x \leq 0 \\ 0 & x < 0 \end{cases}$$

Remark:

1. We use $X \sim \text{LOGN}(\mu, \sigma^2)$ to denote that rv X have a lognormal distribution with parameters μ and σ.
2. Notice here that the parameter μ is not the mean and σ^2 is not the variance, i.e.

$$\mu \neq E(X) \quad \text{and} \quad \sigma^2 \neq V(X)$$

Lognormal Distribution

Lognormal Distribution

Lognormal Distribution

Lognormal Distribution

Proposition

If $X \sim LOGN(\mu, \sigma^2)$, *then*

$$E(X) = e^{\mu+\sigma^2/2} \quad \textit{and} \quad V(X) = e^{2\mu+\sigma^2} \cdot (e^{\sigma^2} - 1)$$

The cdf of X *is*

$$\begin{aligned} F(x; \mu, \sigma) &= P(X \leq x) = P[\ln(X) \leq \ln(x)] \\ &= P\left(Z \leq \frac{\ln(x) - \mu}{\sigma}\right) = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right) \qquad x \leq 0 \end{aligned}$$

where $\Phi(z)$ *is the cdf of the standard normal rv* Z.

Lognormal Distribution

Lognormal Distribution

Example (Problem 115)
Let I_i be the input current to a transistor and I_0 be the output current. Then the current gain is proportional to $\ln(I_0/I_i)$. Suppose the constant of proportionality is 1 (which amounts to choosing a particular unit of measurement), so that current gain $= X = \ln(I_0/I_i)$. Assume X is normally distributed with $\mu = 1$ and $\sigma = 0.05$.

Lognormal Distribution

Example (Problem 115)
Let I_i be the input current to a transistor and I_0 be the output current. Then the current gain is proportional to $\ln(I_0/I_i)$. Suppose the constant of proportionality is 1 (which amounts to choosing a particular unit of measurement), so that current gain $= X = \ln(I_0/I_i)$. Assume X is normally distributed with $\mu = 1$ and $\sigma = 0.05$.

a. *What is the probability that the output current is more than twice the input current?*

Lognormal Distribution

Example (Problem 115)
Let I_i be the input current to a transistor and I_0 be the output current. Then the current gain is proportional to $\ln(I_0/I_i)$. Suppose the constant of proportionality is 1 (which amounts to choosing a particular unit of measurement), so that current gain $= X = \ln(I_0/I_i)$. Assume X is normally distributed with $\mu = 1$ and $\sigma = 0.05$.

a. *What is the probability that the output current is more than twice the input current?*
b. *What are the expected value and variance of the ratio of output to input current?*

Lognormal Distribution

Example (Problem 115)
Let I_i be the input current to a transistor and I_0 be the output current. Then the current gain is proportional to $\ln(I_0/I_i)$. Suppose the constant of proportionality is 1 (which amounts to choosing a particular unit of measurement), so that current gain $= X = \ln(I_0/I_i)$. Assume X is normally distributed with $\mu = 1$ and $\sigma = 0.05$.

a. *What is the probability that the output current is more than twice the input current?*
b. *What are the expected value and variance of the ratio of output to input current?*
c. *What value r is such that only 5% chance we will have the ratio of output to input current exceed r?*

Beta Distribution

Beta Distribution

Definition

A random variable X is said to have a **beta distribution** with parameters α, β(both positive), A, and B if the pdf of X is

$$f(x;\alpha,\beta,A,B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \left(\frac{x-A}{B-A}\right)^{\alpha-1} \cdot \left(\frac{B-x}{B-A}\right)^{\beta-1} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

The case $A = 0, B = 1$ gives the **standard beta distribution**.

Beta Distribution

Definition

A random variable X is said to have a **beta distribution** with parameters α, β(both positive), A, and B if the pdf of X is

$$f(x;\alpha,\beta,A,B) = \begin{cases} \frac{1}{B-A} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \left(\frac{x-A}{B-A}\right)^{\alpha-1} \cdot \left(\frac{B-x}{B-A}\right)^{\beta-1} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

The case $A = 0, B = 1$ gives the **standard beta distribution**.

Remark: We use $X \sim \text{BETA}(\alpha, \beta, A, B)$ to denote that rv X has a beta distribution with parameters α, β, A, and B.

Beta Distribution

Beta Distribution

Proposition

If $X \sim BETA(\alpha, \beta, A, B)$, *then*

$$E(X) = A + (B - A) \cdot \frac{\alpha}{\alpha + \beta} \quad \textit{and} \quad V(X) = \frac{(B - A)^2 \alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Beta Distribution

Beta Distribution

Beta Distribution

Beta Distribution

Example (Problem 127)
An individual's credit score is a number calculated based on that person's credit history which helps a lender determine how much he/she should be loaned or what credit limit should be established for a credit card. An article in the `Los Angeles Times` *gave data which suggested that a beta distribution with parameters* $A = 150, B = 850, \alpha = 8, \beta = 2$ *would provide a reasonable approximation to the distribution of American credit scores.*
[`Note: credit scores are integer-valued`*].*

Beta Distribution

Example (Problem 127)
An individual's credit score is a number calculated based on that person's credit history which helps a lender determine how much he/she should be loaned or what credit limit should be established for a credit card. An article in the `Los Angeles Times` *gave data which suggested that a beta distribution with parameters* $A = 150, B = 850, \alpha = 8, \beta = 2$ *would provide a reasonable approximation to the distribution of American credit scores.*
[`Note: credit scores are integer-valued`*].*

a. *Let* X *represent a randomly selected American credit score. What are the mean value and standard deviation of this random variable? What is the probability that* X *is within 1 standard deviation of its mean value?*

Beta Distribution

Example (Problem 127)
An individual's credit score is a number calculated based on that person's credit history which helps a lender determine how much he/she should be loaned or what credit limit should be established for a credit card. An article in the `Los Angeles Times` *gave data which suggested that a beta distribution with parameters* $A = 150, B = 850, \alpha = 8, \beta = 2$ *would provide a reasonable approximation to the distribution of American credit scores.*
[`Note: credit scores are integer-valued`*].*

a. *Let X represent a randomly selected American credit score. What are the mean value and standard deviation of this random variable? What is the probability that X is within 1 standard deviation of its mean value?*

b. *What is the approximate probability that a randomly selected score will exceed 750 (which lenders consider a very good score)?*

Introduction To Statistics and Data Analysis

- 1980-21st century-inordinate attention to improving the quality
- Japanese - 'industrial miracle'
- Statistical methods and thinking
- Scientific data or Gathering of information(Computer software, Food products,Medicine)
- Difference between Scientific information and inferential statistics
- Product density of material differ from manufacturing Process (Batch to batch)
- Statistical data$\rightarrow$ analyze $\rightarrow$ quality (Special instrument or Medicine)
- Variability in Scientific data
- Product density $\Leftrightarrow$ target
- Information is gathered(Population) $\rightarrow$ samples
- Scientific data must be limited
- 100 items are sampled and 10 items found defective.

Measure of Central tendency or Central location:

To describe a whole set of data with a single value that represents the middle or center of its distribution.

- Mode(Most commonly occurring value of the distribution)
- Median(Middle value of the distribution if its in order)
- Mean

Definition 1. Sample Mean:
Let observations in the same are $x_1, x_2 \ldots x_n$, *the sample mean defined by* $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

Note:Mean=$\dfrac{1}{N}\sum\limits_{i=1}^{n} f_i x_i$ where $N = \sum\limits_{i=1}^{n} f_i$.

Example 1. *Find the arithmetic mean of the following frequency distribution*

x	*1*	*2*	*3*	*4*	*5*	*6*	*7*
frequency	*5*	*9*	*12*	*17*	*14*	*10*	*6*

Definition 2. *Let A be an assumed mean. Arithmetic mean of the distribution is defined as mean=*$A + \frac{1}{N}\sum\limits_{i=1}^{n} f_i d_i$

Example 2. *Calculate the arithmetic mean of the marks from the following table:*

Marks	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
No. of students	*12*	*18*	*27*	*20*	*17*	*6*

In case of grouped or continuous frequency distribution, the arithmetic is reduced to still greater extent by $d_i = \frac{x_i - A}{h}$, h common magnitude of class interval.

Definition 3. *Let A be an assumed mean. Arithmetic mean of the distribution is defined as mean=*$A + \frac{h}{N}\sum\limits_{i=1}^{n} f_i d_i$

Example 3. *Calculate the arithmetic mean of the marks from the following table:*

Class interval	*0-8*	*8-16*	*16-24*	*24-32*	*32-40*	*40-48*
frequency	*8*	*7*	*16*	*24*	*15*	*7*

Definition 4. Median *Median of a distribution is the value of the variable which divides into two equal parts.It is the value which exceeds and is exceeded by same number of observations.*

Example 4.
- *Find median of 25,20,15,35,18.*
- *Find median of 8,20,50,25,15,30.*

Definition 5. Median for Discrete frequency distribution:

- $\dfrac{N}{2}$, *where* $N = \sum\limits_{i} f_i$
- *See the (less than) cumulative frequency just greater than* $\frac{N}{2}$.
- *The corresponding value of* x *is median.*

Example 5.

Wages	*20-30*	*30-40*	*40-50*	*50-60*	*60-70*
No. of labours	*3*	*5*	*20*	*10*	*5*

Find the median.

Example 6. *Obtain the median for the following frequency distribution:*

Wages	*20-30*	*30-40*	*40-50*	*50-60*	*60-70*
No.of lab	*3*	*5*	*20*	*10*	*5*

Median=$l + \frac{h}{f}\left(\frac{N}{2} - c\right)$ where

- l lower limit of median class
- f frequency of median class
- h magnitude of median class
- c cf of the class preceding the median class.

Example 7. *An incomplete frequency distribution is given as follows*

Variable	*10-20*	*20-30*	*20-40*	*40-50*	*50-60*	*60-70*	*70-80*	*Total*
No.of lab	*12*	*30*	◯	*65*	◯	*25*	*18*	*229*

Given that median value is 46, determine the missing frequencies using the median formula.

Example 8. *Then Median and Mode of the following wage distribution are known to be Rs 33.5 and Rs 32 respectively. Find the values of f_3, f_4, f_5.*

Wages	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*	*60-70*	*Total*
frequency	*4*	*16*	f_3	f_4	f_5	*6*	*4*	*230*

Definition 6.

- *for D.F.D, mode value of x, corresponding to maximum frequency.*
- *For C.F.D, mode value is $l + \dfrac{h(f_1 - f_0)}{2f_1 - f_0 - f_2}$ where l is lower limit, h is magnitude, f_1 is frequency modal.*

Example 9.

0-10	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
5	*6*	*7*	*12*	*28*	*20*

Definition 7. Geometric mean: *Geometric mean of observations $x_1, x_2, \ldots x_n$ is $(x_1 x_2 \ldots x_n)^{\frac{1}{n}}$.*

Definition 8. GM for frequency distribution: *If $x_1, x_2 \ldots x_n$ occur with frequency $f_1, f_2 \ldots f_n$ respectively, then GM=$({x_1}^{f_1}{x_2}^{f_2} \ldots {x_n}^{f_n})^{\frac{1}{N}}$ where N is sum of frequency.*

Definition 9. *HM*=$\dfrac{n}{\frac{1}{x_1} + \frac{1}{x_2} \ldots \frac{1}{x_n}}$

Definition 10. HM for frequency distribution *HM*=$\dfrac{N}{\frac{f_1}{x_1} + \frac{f_2}{x_2} \ldots \frac{f_n}{x_n}}$

Definition 11. First Quartile

Chapter 100

Tests for One Proportion

Introduction

The *One-Sample Proportion Test* is used to assess whether a population proportion (*P1*) is significantly different from a hypothesized value (*P0*). This is called the hypothesis of *inequality*. The hypotheses may be stated in terms of the proportions, their difference, their ratio, or their odds ratio, but all four hypotheses result in the same test statistics.

For example, suppose that the current treatment for a disease cures 62% of all cases. A new treatment method has been proposed and studied. In a sample of 80 subjects with the disease that were treated with the new method, 63 were cured. Do the results of this study support the claim that the new method has a higher response rate than the existing method?

This procedure calculates sample size and statistical power for testing a single proportion using either the exact test or other approximate z-tests. Exact test results are based on calculations using the binomial (and hypergeometric) distributions. Because the analysis of several different test statistics is available, their statistical power may be compared to find the most appropriate test for a given situation.

This procedure has the capability for computing power using both the normal approximation and binomial enumeration for all tests. Some sample size programs use only the normal approximation to the binomial distribution for power and sample size estimates. The normal approximation is accurate for large sample sizes and for proportions between 0.2 and 0.8, roughly. When the sample sizes are small or the proportions are extreme (i.e. less than 0.2 or greater than 0.8) the binomial calculations are much more accurate.

Binomial Model

A binomial variable should exhibit the following four properties:

1. The variable is binary --- it can take on one of two possible values.
2. The variable is observed a known number of times. Each observation or replication is called a Bernoulli trial. The number of replications is *n*. The number of times that the outcome of interest is observed is *r*. Thus *r* takes on the possible values 0, 1, 2, ..., *n*.
3. The probability, *P*, that the outcome of interest occurs is constant for each trial.
4. The trials are independent. The outcome of one trial does not influence the outcome of the any other trial.

A binomial probability is calculated using the formula

$$b(r;n,P)=\binom{n}{r}P^r(1-P)^{n-r}$$

© NCSS, LLC. All Rights Reserved.

where

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

The Hypergeometric Model

When samples are taken without replacement from a population of known size, *N*, the *hypergeometric* distribution should be used in place of the binomial distribution. The properties of a variable that is distributed according to the hypergeometric distribution are

1. The variable is binary--it can take on one of two possible values.
2. The variable is observed a known number of times. Each observation or replication is called a Bernoulli trial. The number of replications is *n*. The number of times that the outcome of interest is observed is *r*. Thus *r* takes on the possible values 0, 1, 2, ..., *n*.
3. The total number of items is *N*. The proportion of items with the characteristic of interest is *P*.

The hypergeometric probability of obtaining exactly *r* of *n* items with the characteristic of interest is calculated using

$$h(r;N,n,P) = \frac{\binom{NP}{r}\binom{N-NP}{n-r}}{\binom{N}{n}}$$

Note that the quantity *NP* is rounded to the nearest integer.

Hypothesis Testing

Parameterizations of the Proportions

There are several ways to specify the proportions under the null and the alternative hypotheses. The most direct is to simply give values for P0 and P1. However, it is often more meaningful to specify P0 and then specify the alternative as the difference, the ratio, or the odds ratio. The value of P1 is calculated from these values.

Mathematically, these alternative parameterizations are

Parameter	**Computation**
Difference	$\delta = P1 - P0$
Ratio	$\phi = P1 / P0$
Odds Ratio	$\psi = \frac{O1}{O0} = \frac{P1/(1-P1)}{P0/(1-P0)}$

© NCSS, LLC. All Rights Reserved.

Difference

The (risk) difference, $\delta = P1 - P0$, is perhaps the most direct method of comparison between the two proportions. This parameter is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One interpretation difficulty occurs when the event of interest is rare. If a difference of 0.001 is reported for an event with a baseline probability of 0.40, we would dismiss this as being trivial. That is, there is usually little interest in a treatment that decreases the probability from 0.400 to 0.399. However, if the baseline probability of a disease is 0.002, a 0.001 decrease in the disease probability would represent a reduction of 50%. The interpretation depends on the baseline probability of the event.

Ratio

The (risk) ratio, $\phi = P1 / P0$, gives the relative change in the probability of the outcome under each of the hypothesized values. This parameter is direct and easy to interpret. To compare the ratio with the difference, examine the case where P0 = 0.1437 and P1 = 0.0793. One should consider which number is more enlightening, the difference of -0.0644, or the ratio of 55.18%. In many cases, the ratio communicates the change in proportion in a manner that is more appropriate than the difference.

Odds Ratio

Chances are usually communicated as long-term proportions or probabilities. In betting, chances are often given as odds. For example, the odds of a horse winning a race might be set at 10-to-1 or 3-to-2. Odds can easily be translated into probability. An odds of 3-to-2 means that the event is expected to occur three out of five times. That is, an odds of 3-to-2 (1.5) translates to a probability of winning of 0.60.

The odds of an event are calculated by dividing the event risk by the non-event risk. Thus the odds are

$$O1 = \frac{P1}{1 - P1} \text{ and } O0 = \frac{P0}{1 - P0}$$

For example, if P1 is 0.60, the odds are 0.60/0.4 = 1.5. Rather than represent the odds as a decimal amount, it is re-scaled into whole numbers. Thus, instead of saying the odds are 1.5-to-1, we say they are 3-to-2.

Thus, another way to compare proportions is to compute the ratio of their odds. The odds ratio of two proportions is

$$\psi = \frac{O1}{O0} = \frac{P1/(1 - P1)}{P0/(1 - P0)}$$

Test Statistics

Many different test statistics have been proposed for testing a single proportion. Most of these were proposed before computers or hand calculators were widely available. Although these legacy methods are still presented in textbooks, their power and accuracy should be compared against modern exact methods before they are adopted for serious research. To make this comparison easy, the power and significance of several tests of a single proportion are available in this procedure.

Exact Test

The test statistic is *r*, the number of successes in *n* trials. This test should be the standard against which other test statistics are judged. The significance level and power are computed by enumerating the possible values of *r*, computing the probability of each value, and then computing the corresponding value of the test statistic. Hence the values that are reported in the output for these tests are exact, not approximate.

© NCSS, LLC. All Rights Reserved.

Z-Tests

Several *z* statistics have been proposed that use the central limit theorem. This theorem states that for large sample sizes, the distribution of the *z* statistic is approximately normal. All of these tests take the following form:

$$z = \frac{p - P0}{s}$$

Although these *z* tests were developed because the distribution of *z* is approximately normal in large samples, the actual significance level and power can be computed exactly using the binomial distribution.

We include four *z* tests which are based on two methods for computing *s* and whether a continuity correction is applied.

Z-Test using S(P0)

This test statistic uses the value of P0 to compute *s*.

$$z_1 = \frac{p - P0}{\sqrt{P0(1 - P0)/n}}$$

Z-Test using S(P0) with Continuity Correction

This test statistic is similar to the one above except that a continuity correction is applied to make the normal distribution more closely approximate the binomial distribution.

$$z_2 = \frac{(p - P0) + c}{\sqrt{P0(1 - P0)/n}}$$

where

$$c = \begin{cases} \frac{-1}{2n} & \text{if } p > P0 \\ \frac{1}{2n} & \text{if } p < P0 \\ 0 & \text{if } |p - P0| < \frac{1}{2n} \end{cases}$$

Z-Test using S(Phat)

This test statistic uses the value of *p* to compute *s*.

$$z_3 = \frac{p - P0}{\sqrt{p(1 - p)/n}}$$

Z-Test using S(Phat) with Continuity Correction

This test statistic is similar to the one above except that a continuity correction is applied to make the normal distribution more closely approximate the binomial distribution.

$$z_4 = \frac{(p - P0) + c}{\sqrt{p(1 - p)/n}}$$

© NCSS, LLC. All Rights Reserved.

where

$$c = \begin{cases} \dfrac{-1}{2n} & \text{if } p > P0 \\ \dfrac{1}{2n} & \text{if } p < P0 \\ 0 & \text{if } |p - P0| < \dfrac{1}{2n} \end{cases}$$

Power Calculation

Normal Approximation Method

Power may be calculated for one-sample proportions tests using the normal approximation to the binomial distribution. This section provides the power calculation formulas for the various test statistics available in this procedure. In each case, power is presented for the lower and upper one-sided hypothesis tests and for the two-sided hypothesis test. In the equations that follow, $\Phi()$ represents the standard normal cumulative distribution function, and z_α represents the value that leaves α in the upper tail of the standard normal distribution. All power values are evaluated at $P = P1$.

Exact Test

Power for the two-sided hypothesis test of $H0: P = P0$ vs. $H1: P \neq P0$ is

$$Power_{\text{ET, Two-Sided}} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha/2}\sqrt{P0(1-P0)FPC}}{\sqrt{P1(1-P1)FPC}}\right) + 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_{\alpha/2}\sqrt{P0(1-P0)FPC}}{\sqrt{P1(1-P1)FPC}}\right)$$

where $FPC = (N - n)/(N - 1)$ if the population size, N, is finite, and $FPC = 1$ if the population size is infinite.

Power for the lower one-sided hypothesis test of $H0: P \geq P0$ vs. $H1: P < P0$ is

$$Power_{\text{ET, Lower One-Sided}} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha}\sqrt{P0(1-P0)FPC}}{\sqrt{P1(1-P1)FPC}}\right)$$

Power for the upper one-sided hypothesis test of $H0: P \leq P0$ vs. $H1: P > P0$ is

$$Power_{\text{ET, Upper One-Sided}} = 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_{\alpha}\sqrt{P0(1-P0)FPC}}{\sqrt{P1(1-P1)FPC}}\right)$$

Z Test using S(P0)

Power for the two-sided hypothesis test of $H0: P = P0$ vs. $H1: P \neq P0$ is

$$Power_{\text{ZS(P0), Two-Sided}} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha/2}\sqrt{P0(1-P0)}}{\sqrt{P1(1-P1)}}\right) + 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_{\alpha/2}\sqrt{P0(1-P0)}}{\sqrt{P1(1-P1)}}\right)$$

Power for the lower one-sided hypothesis test of $H0: P \geq P0$ vs. $H1: P < P0$ is

$$Power_{ZS(P0),\ Lower\ One-Sided} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha}\sqrt{P0(1-P0)}}{\sqrt{P1(1-P1)}}\right)$$

© NCSS, LLC. All Rights Reserved.

Power for the upper one-sided hypothesis test of $H0: P \leq P0$ vs. $H1: P > P0$ is

$$Power_{ZS(P0),\ Upper\ One-Sided} = 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_\alpha\sqrt{P0(1-P0)}}{\sqrt{P1(1-P1)}}\right)$$

Z Test using S(P0) with Continuity Correction

Power for the two-sided hypothesis test of $H0: P = P0$ vs. $H1: P \neq P0$ is

$$Power_{\mathrm{ZS(P0)CC,\ Two-Sided}} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha/2}\sqrt{P0(1-P0)} - c}{\sqrt{P1(1-P1)}}\right) + 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_{\alpha/2}\sqrt{P0(1-P0)} + c}{\sqrt{P1(1-P1)}}\right)$$

where $c = 1/2\sqrt{n}$ if $|P1 - P0| < 1/2n$ otherwise $c = 0$.

Power for the lower one-sided hypothesis test of $H0: P \geq P0$ vs. $H1: P < P0$ is

$$Power_{ZS(P0)CC,\ Lower\ One-Sided} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_\alpha\sqrt{P0(1-P0)} - c}{\sqrt{P1(1-P1)}}\right)$$

Power for the upper one-sided hypothesis test of $H0: P \leq P0$ vs. $H1: P > P0$ is

$$Power_{ZS(P0)CC,\ Upper\ One-Sided} = 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_\alpha\sqrt{P0(1-P0)} + c}{\sqrt{P1(1-P1)}}\right)$$

Z Test using S(Phat)

Power for the two-sided hypothesis test of $H0: P = P0$ vs. $H1: P \neq P0$ is

$$Power_{\mathrm{ZS(P1),\ Two-Sided}} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha/2}\sqrt{P1(1-P1)}}{\sqrt{P1(1-P1)}}\right) + 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_{\alpha/2}\sqrt{P1(1-P1)}}{\sqrt{P1(1-P1)}}\right)$$

Power for the lower one-sided hypothesis test of $H0: P \geq P0$ vs. $H1: P < P0$ is

$$Power_{ZS(P1),\ Lower\ One-Sided} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_\alpha\sqrt{P1(1-P1)}}{\sqrt{P1(1-P1)}}\right)$$

Power for the upper one-sided hypothesis test of $H0: P \leq P0$ vs. $H1: P > P0$ is

$$Power_{ZS(P1),\ Upper\ One-Sided} = 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_\alpha\sqrt{P1(1-P1)}}{\sqrt{P1(1-P1)}}\right)$$

Z Test using S(Phat) with Continuity Correction

Power for the two-sided hypothesis test of $H0: P = P0$ vs. $H1: P \neq P0$ is

$$Power_{\mathrm{ZS(P1)CC,\ Two-Sided}} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_{\alpha/2}\sqrt{P1(1-P1)} - c}{\sqrt{P1(1-P1)}}\right) + 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_{\alpha/2}\sqrt{P1(1-P1)} + c}{\sqrt{P1(1-P1)}}\right)$$

where $c = 1/2\sqrt{n}$ if $|P1 - P0| < 1/2n$ otherwise $c = 0$.

Power for the lower one-sided hypothesis test of $H0: P \geq P0$ vs. $H1: P < P0$ is

$$Power_{ZS(P1)CC,\ Lower\ One-Sided} = \Phi\left(\frac{\sqrt{n}(P0 - P1) - z_\alpha\sqrt{P1(1-P1)} - c}{\sqrt{P1(1-P1)}}\right)$$

© NCSS, LLC. All Rights Reserved.

Power for the upper one-sided hypothesis test of $H0: P \leq P0$ vs. $H1: P > P0$ is

$$Power_{ZS(P1)CC,\ \ Upper\ One-Sided} = 1 - \Phi\left(\frac{\sqrt{n}(P0 - P1) + z_\alpha\sqrt{P1(1-P1)} + c}{\sqrt{P1(1-P1)}}\right)$$

Steps to Calculate Power using Binomial Enumeration of All Possible Outcomes

The power calculation procedure is as follows. Let P represent the true probability that an item selected at random from a population will have a characteristic of interest.

1. State the Hypotheses

H0 is the null hypothesis that the proportion is P0.

H1 is the alternative hypothesis that the proportion is P1.

Three sets of statistical hypotheses may be formulated:

1. $H0: P = P0$ versus $H1: P = P1 \neq P0$. This hypothesis results in a two-tailed test.
2. $H0: P \leq P0$ versus $H1: P = P1 > P0$. This hypothesis results in a one-tailed test.
3. $H0: P \geq P0$ versus $H1: P = P1 < P0$. This hypothesis results in a one-tailed test.

2. Find the Critical Value

For an upper-tailed test with a given sample size find the critical value, Pc, based on the binomial (or hypergeometric) distribution, so that the probability of rejecting H0 when H0 is true is equal to a specified significance level, α .

3. Evaluate the Sample

Select a sample of *n* items from the population and compute the sample proportion, p = r / n. If p > Pc then reject the null hypothesis that P = P0 in favor of an alternative hypothesis that P = P1 > P0.

4. Calculate the Power

The power is the probability of rejecting H0 when the true proportion is P1. That is, the power is the probability that p > Pc calculated from a binomial (or hypergeometric) distribution in which P = P1.

Similar steps are used for the lower-tail and two-tailed tests.

© NCSS, LLC. All Rights Reserved.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Design tab. For more information about the options of other tabs, go to the Procedure Window chapter.

Design Tab

The Design tab contains the parameters associated with this test such as the proportions, sample sizes, alpha, and power.

Solve For

Solve For

This option specifies the parameter to be solved for using the other parameters. The parameters that may be selected are *Power*, *Sample Size*, and *Effect Size*.

Power Calculation

Power Calculation Method

Select the method to be used to calculate power. When the sample sizes are reasonably large (i.e. greater than 50) and the proportions are between 0.2 and 0.8 the two methods will give similar results. For smaller sample sizes and more extreme proportions (less than 0.2 or greater than 0.8), the normal approximation is not as accurate so the binomial calculations may be more appropriate.

The choices are

- **Binomial Enumeration**
  Power for each test is computed using binomial enumeration of all possible outcomes when n ≤ Max n for Binomial Enumeration (otherwise, the normal approximation is used). Binomial enumeration of all outcomes is possible because of the discrete nature of the data.

- **Normal Approximation**
  Approximate power for each test is computed using the normal approximation to the binomial distribution.

Actual alpha values are only computed when “Binomial Enumeration” is selected.

Max n for Binomial Enumeration

Only shown when Power Calculation Method = “Binomial Enumeration”

When *n* is less than or equal to this value, power is calculated using the binomial distribution and enumeration of all possible outcomes. This is possible because of the discrete nature of the data. “Actual Alpha” values are only computed when binomial power calculations are made.

When *n* is greater than this value, the normal approximation to the binomial is used when calculating power.

© NCSS, LLC. All Rights Reserved.

Test

Alternative Hypothesis

Specify whether the alternative hypothesis of the test is one-sided or two-sided. If a one-sided test is chosen, the hypothesis test direction is chosen based on whether P1 is greater than or less than P0.

- **Two-Sided Hypothesis Test**
  $H0: P = P0$ vs. $H1: P \neq P0$
- **One-Sided Hypothesis Tests**
  Upper: $H0: P \leq P0$ vs. $H1: P > P0$
  Lower: $H0: P \geq P0$ vs. $H1: P < P0$

Test Type

Specify the type of test that will be used in reporting. Note that *C.C.* is short for *Continuity Correction*. This refers to the adding or subtracting 1/(2*n*) to (or from) the numerator of the z-value to bring the normal approximation closer to the binomial distribution.

N (Population Size)

Only shown when Test Type = "Exact Test"

Enter the total number of items in the population from which the sample of *n* items is selected. Enter *Infinite* to signify an infinite population so that no correction factor is applied. An *infinite* population is generally one in which the number in the population is large and unknown.

Note that *N* must be greater than *n*.

When samples are drawn from a very large (infinite) population, calculations are based on the binomial distribution.

When samples are drawn from a population of known size, specified here as *N*, calculations are based on the hypergeometric distribution.

Power and Alpha

Power

This option specifies one or more values for power. Power is the probability of rejecting a false null hypothesis, and is equal to one minus Beta. Beta is the probability of a type-II error, which occurs when a false null hypothesis is not rejected.

Values must be between zero and one. Historically, the value of 0.80 (Beta = 0.20) was used for power. Now, 0.90 (Beta = 0.10) is also commonly used.

A single value may be entered here or a range of values such as *0.8 to 0.95 by 0.05* may be entered.

Alpha

This option specifies one or more values for the probability of a type-I error. A type-I error occurs when a true null hypothesis is rejected. For this procedure, a type-I error occurs when you reject the null hypothesis of equal proportions when in fact they are equal.

Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. You should pick a value for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

© NCSS, LLC. All Rights Reserved.

Note that because of the discrete nature of the binomial distribution, the alpha level rarely will be achieved exactly.

A single value may be entered here or a range of values such as *0.05 to 0.2 by 0.05* may be entered.

Sample Size

n (Sample Size)

This option specifies the total number of observations in the sample. Values must be integers greater than one.

You may enter a single value or a range of values such as *10, 50, 100* or *10 to 100 by 10*.

Effect Size

Input Type

Indicate what type of values to enter to specify the effect size. Regardless of the entry type chosen, the test statistics used in the power and sample size calculations are the same. This option is simply given for convenience in specifying the effect size.

P0 (Null Proportion)

Enter a value (or range of values) for the population proportion under the null hypothesis, P0. This is the baseline proportion, the proportion that exists in the general population. The proportion estimated from the data will be compared to this value by the statistical test.

Proportions must be between zero and one.

You may enter a single value or a range of values such as *0.1 0.2 0.3* or *0.1 to 0.9 by 0.1*.

P1 (Alternative Proportion)

Enter a value (or range of values) for the value of the binomial proportion at which the power is calculated. This is labeled P1 on the screen. Power calculations assume that this is the true value of the proportion. This value cannot be equal to P0. When solving for effect size, specify the search direction.

(P1–P0) Alternative Difference

This option implicitly specifies the value of P1 (the proportion at which the power is calculated) by explicitly specifying the difference. The difference is used with P0 to calculate the value of P1 using the formula, P1 = diff + P0.

Since P1 is a proportion, the difference must be between –P0 and 1 – P0. The difference cannot be equal to zero. A single value or a range of values may be entered here. When solving for effect size, specify the search direction.

(P1/P0) Alternative Ratio

This option implicitly specifies the value of P1 (the proportion at which the power is calculated) by explicitly specifying the ratio. The ratio is used with P0 to calculate the value of P1 using the formula, P1 = (P0) x (ratio).

Since P1 is a proportion, the ratio must be between 0 and 1 / P0. The ratio cannot be equal to one. A single value or a range of values may be entered here. When solving for effect size, specify the search direction.

(O1/O0) Alternative Odds Ratio

This option implicitly specifies the value of P1 (the proportion at which the power is calculated) by explicitly specifying the odds ratio. Since P1 is a proportion, the odds ratio must be greater than zero. The odds ratio cannot be equal to one. A single value or a range of values may be entered here. When solving for effect size, specify the search direction.

© NCSS, LLC. All Rights Reserved.

Example 1 – Finding the Power

Suppose 50% of patients with a certain type of cancer survive five years. Researchers have developed a new treatment to increase the percentage of individuals that survive five years. Although the researchers do not know the true percentage of patients that will survive with the new treatment, they would like to examine the power that is achieved if the percentage under the new treatment is 60%. The power will be determined for trials with sample sizes of 50, 100, 200, 300, 500, or 800 and a significance level of 0.05. For comparative purposes, the power is also to be calculated for alternative proportions of 55% and 65%.

This is an example of a *historically controlled* trial. *Historically controlled* means that no control group is formed for the current study. Instead, the rates reported from previous studies or that are known to exist in the general population are used. Because of the many advantages that occur when an actual control group is used, historically controlled trials should only be used when a control group is either impossible to obtain or unethical.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**Two-Sided**
Test Type	**Z-test using S(P0)**
Alpha	**0.05**
n (Sample Size)	**50 100 200 300 500 800**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.55 0.60 0.65**

© NCSS, LLC. All Rights Reserved.

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If |Z| >
0.10766	50	0.5000	0.5500	0.0500	0.0500	1.9600
0.28905	50	0.5000	0.6000	0.1000	0.0500	1.9600
0.56717	50	0.5000	0.6500	0.1500	0.0500	1.9600
0.16879	100	0.5000	0.5500	0.0500	0.0500	1.9600
0.51632	100	0.5000	0.6000	0.1000	0.0500	1.9600
0.86220	100	0.5000	0.6500	0.1500	0.0500	1.9600
0.29202	200	0.5000	0.5500	0.0500	0.0500	1.9600
0.81229	200	0.5000	0.6000	0.1000	0.0500	1.9600
0.99164	200	0.5000	0.6500	0.1500	0.0500	1.9600
0.40951	300	0.5000	0.5500	0.0500	0.0500	1.9600
0.93763	300	0.5000	0.6000	0.1000	0.0500	1.9600
0.99965	300	0.5000	0.6500	0.1500	0.0500	1.9600
0.60931	500	0.5000	0.5500	0.0500	0.0500	1.9600
0.99483	500	0.5000	0.6000	0.1000	0.0500	1.9600
1.00000	500	0.5000	0.6500	0.1500	0.0500	1.9600
0.80863	800	0.5000	0.5500	0.0500	0.0500	1.9600
0.99992	800	0.5000	0.6000	0.1000	0.0500	1.9600
1.00000	800	0.5000	0.6500	0.1500	0.0500	1.9600

* Power was computed using the normal approximation method.

References
Chow, S. C., Shao, J., and Wang, H. 2008. Sample Size Calculations in Clinical Research, Second Edition. Chapman & Hall/CRC. Boca Raton, Florida.
Fleiss, J. L., Levin, B., and Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.
Lachin, John M. 2000. Biostatistical Methods. John Wiley & Sons. New York.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, Mass.
Ryan, Thomas P. 2013. Sample Size Determination and Power. John Wiley & Sons. Hoboken, New Jersey.
Zar, Jerrold H. 2010. Biostatistical Analysis (Fifth Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

Report Definitions
Power is the probability of rejecting the null hypothesis when it is false. It should be close to one.
n is the size of the sample drawn from the population. To conserve resources, it should be as small as possible.
P0 is the value of the population proportion under the null hypothesis.
P1 is the value of the population proportion under the alternative hypothesis.
P1-P0 is the difference to be detected by the study.
Alpha (significance level) is the probability of rejecting the null hypothesis when it is true. It should be small.
Reject H0 If... gives the critical value(s) for the test.

Summary Statements
A sample size of 50 achieves 10.766% power to detect a difference (P1-P0) of 0.0500 using a two-sided Z-test that uses S(P0) to estimate the standard deviation with a significance level (alpha) of 0.0500. These results assume that the population proportion under the null hypothesis (P0) is 0.5000.

This report shows the values of each of the parameters, one scenario per row. The values from this table are plotted in the chart below.

© NCSS, LLC. All Rights Reserved.

Chart Section

These plots show the relationship between power, sample size, and P1 in this example. We note that 80% power is achieved with a sample size of about 200 when P1 is 0.60, which was the specific value of interest.

© NCSS, LLC. All Rights Reserved.

Example 2 – Finding the Sample Size

Continuing with Example 1, suppose you want to study the impact of various choices for P1 on sample size. Using a significance level of 0.05 and 90% power, find the sample size when P1 is 0.55, 0.60, 0.65, 0.70, 0.75, and 0.80. Assume that an two-tailed Z-test using S(P0) will be used.

In this example we'll show you how to setup the calculation by inputting proportions, differences, ratios, and odds ratios. In all cases, you'll see that the sample sizes are exactly the same. The only difference is in the way P1 is specified.

Setup (Proportions)

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 2a** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**Two-Sided**
Test Type	**Z-test using S(P0)**
Power	**0.90**
Alpha	**0.05**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.55 to 0.80 by 0.05**

Output (Proportions)

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If |Z| >
0.90011	1047	0.5000	0.5500	0.0500	0.0500	1.9600
0.90055	259	0.5000	0.6000	0.1000	0.0500	1.9600
0.90120	113	0.5000	0.6500	0.1500	0.0500	1.9600
0.90286	62	0.5000	0.7000	0.2000	0.0500	1.9600
0.90249	38	0.5000	0.7500	0.2500	0.0500	1.9600
0.90321	25	0.5000	0.8000	0.3000	0.0500	1.9600

* Power was computed using the normal approximation method.

This report shows the sample sizes corresponding to various values of P1. Notice that a sample size of only 25 is needed to detect the difference between 0.5 and 0.8, but a sample size of 1047 is needed to detect the difference between 0.50 and 0.55.

© NCSS, LLC. All Rights Reserved.

Setup (Differences)

The setup for differences is exactly the same as that for proportions except for the following two inputs on the Design tab. You may then make the appropriate entries as listed below, or open **Example 2b** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Input Type	**Differences**
P1 (Alternative Difference)	**0.05 to 0.3 by 0.05**

Output (Differences)

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If |Z| >
0.90011	1047	0.5000	0.5500	0.0500	0.0500	1.9600
0.90055	259	0.5000	0.6000	0.1000	0.0500	1.9600
0.90120	113	0.5000	0.6500	0.1500	0.0500	1.9600
0.90286	62	0.5000	0.7000	0.2000	0.0500	1.9600
0.90249	38	0.5000	0.7500	0.2500	0.0500	1.9600
0.90321	25	0.5000	0.8000	0.3000	0.0500	1.9600

* Power was computed using the normal approximation method.

This report shows the sample sizes corresponding to various values of P1 entered indirectly as differences.

Setup (Ratios)

The setup for ratios is exactly the same as that for proportions except for the following two inputs on the Design tab. You may then make the appropriate entries as listed below, or open **Example 2c** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Input Type	**Ratios**
P1 (Alternative Ratio)	**1.1 to 1.6 by 0.1**

© NCSS, LLC. All Rights Reserved.

Output (Ratios)

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Ratio P1/P0	Alpha	Reject H0 If |Z| >
0.90011	1047	0.5000	0.5500	1.1000	0.0500	1.9600
0.90055	259	0.5000	0.6000	1.2000	0.0500	1.9600
0.90120	113	0.5000	0.6500	1.3000	0.0500	1.9600
0.90286	62	0.5000	0.7000	1.4000	0.0500	1.9600
0.90249	38	0.5000	0.7500	1.5000	0.0500	1.9600
0.90321	25	0.5000	0.8000	1.6000	0.0500	1.9600

* Power was computed using the normal approximation method.

This report shows the sample sizes corresponding to various values of P1 entered indirectly as ratios.

Setup (Odds Ratios)

The setup for odds ratios is exactly the same as that for proportions except for the following two inputs on the Design tab. You may then make the appropriate entries as listed below, or open **Example 2d** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Input Type	**Odds Ratios**
P1 (Alternative Odds Ratio)	**1.2222 1.5 1.8571 2.3333 3 4**

Output (Odds Ratios)

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Odds Ratio O1/O0	Alpha	Reject H0 If |Z| >
0.90006	1047	0.5000	0.5500	1.2222	0.0500	1.9600
0.90055	259	0.5000	0.6000	1.5000	0.0500	1.9600
0.90118	113	0.5000	0.6500	1.8571	0.0500	1.9600
0.90285	62	0.5000	0.7000	2.3333	0.0500	1.9600
0.90249	38	0.5000	0.7500	3.0000	0.0500	1.9600
0.90321	25	0.5000	0.8000	4.0000	0.0500	1.9600

* Power was computed using the normal approximation method.

This report shows the sample sizes corresponding to various values of P1 entered indirectly as odds ratios.

© NCSS, LLC. All Rights Reserved.

Example 3 – Investigating the Saw-Tooth Power Function

After releasing the first version of ***PASS***, we received many inquiries about the strange shape of the relationship between sample size and power computed using binomial enumeration of all possible outcomes when testing a single proportion using the exact test. This example will show why this strange shape occurs.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Binomial Enumeration**
Max n for Binomial Enumeration	**10000**
Alternative Hypothesis	**Two-Sided**
Test Type	**Exact Test**
N (Population Size)	**Infinite**
Alpha	**0.05**
n (Sample Size)	**51 to 60 by 1**
Input Type	**Proportions**
P0 (Null Proportion)	**0.60**
P1 (Alternative Proportion)	**0.70**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Exact Test
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Target Alpha	Actual Alpha*	Reject H0 If R ≤ | R ≥
0.29656	51	0.6000	0.7000	0.1000	0.0500	0.0443	23|38
0.26688	52	0.6000	0.7000	0.1000	0.0500	0.0328	23|39
0.23931	53	0.6000	0.7000	0.1000	0.0500	0.0348	24|40
0.31244	54	0.6000	0.7000	0.1000	0.0500	0.0371	24|40
0.28284	55	0.6000	0.7000	0.1000	0.0500	0.0379	25|41
0.25494	56	0.6000	0.7000	0.1000	0.0500	0.0281	25|42
0.32774	57	0.6000	0.7000	0.1000	0.0500	0.0417	26|42
0.29806	58	0.6000	0.7000	0.1000	0.0500	0.0314	26|43
0.27013	59	0.6000	0.7000	0.1000	0.0500	0.0327	27|44
0.34228	60	0.6000	0.7000	0.1000	0.0500	0.0354	27|44

* Power and actual alpha were computed using binomial enumeration of all possible outcomes.

© NCSS, LLC. All Rights Reserved.

This report shows the values of each of the parameters, one scenario per row. The symbol, *R*, stands for the number of items with the characteristic of interest out of the *n* items sampled. Hence, for the scenario presented on the first line, an exact test does not exist for these parameters at the target alpha of 0.05. The closest that can be achieved is an alpha of 0.0443. In this case, we would reject the null hypothesis in any sample of size 51 in which the count of individuals with the characteristic of interest is less than or equal to 23 or greater than or equal to 38.

The values from this table are plotted in the chart below.

Chart Section

Notice that the power decreases as *n* increases from 51 to 52 and continues to decrease as *n* increases to 53. Usually, the power increases as the sample size increases.

To understand why this happens, look at the last column and at the Actual Alpha column. Note that at *n* = 51, the actual alpha is 0.0443 and at *n* = 52, the actual alpha has decreased to 0.0328. Remember that as alpha decreases, power decreases as well. Hence, increasing the sample size from 51 to 52 was not enough to counterbalance the effect on power of a decrease in alpha from 0.04428 to 0.03281. Hence, the power drops from 0.29656 to 0.26688.

This phenomenon usually occurs for relatively small values of *n*.

© NCSS, LLC. All Rights Reserved.

Example 4 – Step-by-Step Calculations

In this example, we will take you step by step through the calculations necessary to compute the power of a specific scenario using binomial enumeration of all possible outcomes. We will set *n* = 10, P0 = 0.50, P1 = 0.80, and alpha = 0.05. We will compute the power of the two-tailed test.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Binomial Enumeration**
Max n for Binomial Enumeration	**10000**
Alternative Hypothesis	**Two-Sided**
Test Type	**Exact Test**
N (Population Size)	**Infinite**
Alpha	**0.05**
n (Sample Size)	**10**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.8**
Reports Tab	
Power Decimals	**6**
Alpha Decimals	**6**
Proportions Decimals	**6**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Exact Test
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Target Alpha	Actual Alpha*	Reject H0 If R ≤ | R ≥
0.375814	10	0.500000	0.800000	0.300000	0.050000	0.021484	1|9

* Power and actual alpha were computed using binomial enumeration of all possible outcomes.

We will now proceed through the calculations necessary to compute this power value.

© NCSS, LLC. All Rights Reserved.

We first construct a table of binomial probabilities for n = 10 and p = 0.5 using the formula

$$b(r;10,0.5) = \binom{10}{r} 0.5^r (1-0.5)^{10-r}$$

Binomial Probabilities for *n* = 10 and *p* = 0.5

R	Prob(r = R)	Cumulative Total	1 – Cumulative Total
0	0.000977	0.000977	0.999023
1	0.009766	0.010742	0.989258
2	0.043945	0.054688	0.945313
3	0.117188	0.171875	0.828125
4	0.205078	0.376953	0.623047
5	0.246094	0.623047	0.376953
6	0.205078	0.828125	0.171875
7	0.117188	0.945313	0.054688
8	0.043945	0.989258	0.010742
9	0.009766	0.999023	0.000977
10	0.000977	1.000000	0.000000

When we construct a two-tailed test, we split the alpha value evenly between the two tails. Hence, we place $\alpha / 2$ (or 0.025) in each tail. Moving down from the top, we find that the cumulative probability is 0.010742 for R = 1 and 0.054688 for R = 2. Since 0.054688 is greater than 0.025, we adopt R = 1 as our lower rejection value. Likewise, we find that R = 9 is the upper rejection value.

Our testing strategy is

1. Draw a sample of 10 items and count the number with the characteristic of interest. Call this value r.
2. If r = 0, 1, 9, or 10, reject the null hypothesis that $p = 0.5$ in favor of the alternative hypothesis that $p \neq 0.5$.

Now, to compute the power for P1 = 0.8, we must compute another table of binomial probabilities, this time for p = 0.8 using the formula.

$$b(r;10,0.8) = \binom{10}{r} 0.8^r (1-0.8)^{10-r}$$

Binomial Probabilities for *n* = 10 and *p* = 0.8

R	Prob(r = R)	Cumulative Total	1 - Cumulative Total
0	0.000000	0.000000	1.000000
1	0.000004	0.000004	0.999996
2	0.000074	0.000078	0.999922
3	0.000786	0.000864	0.999136
4	0.005505	0.006369	0.993631
5	0.026424	0.032793	0.967207
6	0.088080	0.120874	0.879126
7	0.201327	0.322200	0.677800
8	0.301990	0.624190	0.375810
9	0.268435	0.892626	0.107374
10	0.107374	1.000000	0.000000

The power is the probability of rejecting the null hypothesis. This occurs when r = 0, 1, 9, or 10. From the above table, we compute the power as 0.000000 + 0.000004 + 0.268435 + 0.107374 = 0.375813. This matches the calculated power value as displayed in the results above to within rounding error.

© NCSS, LLC. All Rights Reserved.

Example 5 – Comparing Test Statistics

One important decision that must be made before conducting the experiment is to decide which of the available test statistics to use. This procedure makes it easy to make this comparison. The alternative proportion will be set to 0.60 and the sample sizes will be 10, 11, 12, 25, 50, and 70. We'll also calculate exact power using binomial enumeration.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 5** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Binomial Enumeration**
Max n for Binomial Enumeration	**10000**
Alternative Hypothesis	**Two-Sided**
Test Type	**Exact Test**
N (Population Size)	**Infinite**
Alpha	**0.05**
n (Sample Size)	**10 11 12 25 50 70**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.6**
Reports Tab	
Show Comparative Reports	**Checked**
Comparative Plots Tab	
Show Comparative Plots	**Checked**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results and Plots

Power Comparison of Five Different Tests for One Proportion
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

n	P0	P1	Target Alpha	Exact Test Power	Z-Test S(P0) Power	Z-Test S(P0)C Power	Z-Test S(P) Power	Z-Test S(P)C Power
10	0.5000	0.6000	0.0500	0.04804	0.04804	0.04804	0.17958	0.17958
11	0.5000	0.6000	0.0500	0.03097	0.12484	0.03097	0.12484	0.12484
12	0.5000	0.6000	0.0500	0.08625	0.08625	0.08625	0.24060	0.08625
25	0.5000	0.6000	0.0500	0.15476	0.15476	0.15476	0.15476	0.15476
50	0.5000	0.6000	0.0500	0.23706	0.33613	0.23706	0.33613	0.23706
70	0.5000	0.6000	0.0500	0.36009	0.36009	0.36009	0.45495	0.36009

Note: Power was computed using binomial enumeration of all possible outcomes.

© NCSS, LLC. All Rights Reserved.

Actual Alpha Comparison of Five Different Tests for One Proportion
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

n	P0	P1	Target Alpha	Exact Test Alpha	Z-Test S(P0) Alpha	Z-Test S(P0)C Alpha	Z-Test S(P) Alpha	Z-Test S(P)C Alpha
10	0.5000	0.6000	0.0500	0.0215	0.0215	0.0215	0.1094	0.1094
11	0.5000	0.6000	0.0500	0.0117	0.0654	0.0117	0.0654	0.0654
12	0.5000	0.6000	0.0500	0.0386	0.0386	0.0386	0.1460	0.0386
25	0.5000	0.6000	0.0500	0.0433	0.0433	0.0433	0.0433	0.0433
50	0.5000	0.6000	0.0500	0.0328	0.0649	0.0328	0.0649	0.0328
70	0.5000	0.6000	0.0500	0.0414	0.0414	0.0414	0.0722	0.0414

Note: Actual alpha was computed using binomial enumeration of all possible outcomes.

Chart Section

An examination of the first report shows that for most sample sizes, the power is different for at least one of the tests. Also, notice that the exact test always has the minimum power in each row. This would lead us discard this test statistic. However, consider the second report, which shows the actual alpha level (the target was 0.05) for each test. By inspecting corresponding entries in both tables, we can see that whenever a test statistic achieves a better power than the exact test, it also exceeds the target alpha. For example, look at the powers for n = 12. The z- test using s(p hat) has an unusually large power of 0.24060. This is a much larger power than the exact test's value of 0.08625. However, note that the actual alpha level for this test is 0.1460, which is much higher than the target of 0.05 and the actual alpha values of the other tests, which are 0.0386.

We still would conclude that indeed, the exact test is consistently the best test since it always achieves a significance level that is less than the target value.

© NCSS, LLC. All Rights Reserved.

Example 6 – Comparing Power Calculation Methods

Continuing with Example 5, let's see how the results compare if we were to use approximate power calculations instead of power calculations based on binomial enumeration.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 6** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Binomial Enumeration**
Max n for Binomial Enumeration	**10000**
Alternative Hypothesis	**Two-Sided**
Test Type	**Exact Test**
N (Population Size)	**Infinite**
Alpha	**0.05**
n (Sample Size)	**10 11 12 25 50 70**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.6**
Reports Tab	
Show Power Detail Report	**Checked**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results and Plots

Power Detail Report for Testing One Proportion using the Exact Test
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

				Normal Approximation		Binomial Enumeration	
n	P0	P1	P1-P0	Power	Alpha	Power	Alpha
10	0.5000	0.6000	0.1000	0.09180	0.0500	0.04804	0.0215
11	0.5000	0.6000	0.1000	0.09657	0.0500	0.03097	0.0117
12	0.5000	0.6000	0.1000	0.10135	0.0500	0.08625	0.0386
25	0.5000	0.6000	0.1000	0.16486	0.0500	0.15476	0.0433
50	0.5000	0.6000	0.1000	0.28905	0.0500	0.23706	0.0328
70	0.5000	0.6000	0.1000	0.38504	0.0500	0.36009	0.0414

Notice that the approximate power values consistently overestimate the power, particularly for small sample sizes. As the sample size increases, the approximate power values come nearer to the power values from binomial enumeration.

© NCSS, LLC. All Rights Reserved.

Example 7 – Determining the Power after Completing an Experiment

A group of researchers is studying the effects of a new diet on cholesterol levels in high-risk patients. The researchers had hypothesized that the cholesterol level would be reduced to a safe level in more than 70% of subjects following the new diet. They are confident that the proportion will be no less than 0.70. To test this one-sided hypothesis, they randomly sampled 200 individuals with dangerously high cholesterol and put them on the new diet. After the period of the study, the researchers determined that 150 of the 200 patients sampled (75%) had reduced their cholesterol level while on the new diet. Statistical analysis using the z-test with S(P0) and an alpha level of 0.05, however, resulted in failure to reject the null hypothesis that the proportion is 0.70. The researchers desire now to compute the power of their study for true proportions ranging from 0.71 to 0.80.

Note that a range of proportions is considered for power calculations instead of just 0.75, the sample proportion found in the experiment. While it is tempting to use the sample proportion as the true proportion in post-experiment power calculations, it is more informative to review a range of possible alternatives representing practically significant differences from the null value.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 7** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**One-Sided**
Test Type	**Z-test using S(P0)**
Alpha	**0.05**
n (Sample Size)	**200**
Input Type	**Proportions**
P0 (Null Proportion)	**0.70**
P1 (Alternative Proportion)	**0.71 to 0.80 by 0.01**

© NCSS, LLC. All Rights Reserved.

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: One-Sided (H0: P ≤ P0 vs. H1: P > P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If Z >
0.08859	200	0.7000	0.7100	0.0100	0.0500	1.6449
0.14713	200	0.7000	0.7200	0.0200	0.0500	1.6449
0.22899	200	0.7000	0.7300	0.0300	0.0500	1.6449
0.33404	200	0.7000	0.7400	0.0400	0.0500	1.6449
0.45709	200	0.7000	0.7500	0.0500	0.0500	1.6449
0.58780	200	0.7000	0.7600	0.0600	0.0500	1.6449
0.71268	200	0.7000	0.7700	0.0700	0.0500	1.6449
0.81900	200	0.7000	0.7800	0.0800	0.0500	1.6449
0.89872	200	0.7000	0.7900	0.0900	0.0500	1.6449
0.95064	200	0.7000	0.8000	0.1000	0.0500	1.6449

* Power was computed using the normal approximation method.

Chart Section

Power ranges from 0.08859 for a true proportion of 0.71 to 0.95064 for a true proportion of 0.80.

© NCSS, LLC. All Rights Reserved.

Example 8 – Finding the Power using Ratios

Suppose that only 5% of patients with an aggressive type of cancer respond to the standard treatment. Designers have found a new treatment which could be widely used if the percentage of patients responding is at least 0.5 times greater than the proportion responding to the standard treatment, i.e. P1 = 1.5(P0), or in terms of ratios, P1/P0 = 1.5. What power will be achieved for trials with sample sizes of 200, 300, 500, or 800 and a significance level of 0.05? For comparative purposes, also calculate the power for alternative ratios of 1.25 and 1.75.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 8** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Binomial Enumeration**
Max n for Binomial Enumeration	**10000**
Alternative Hypothesis	**Two-Sided**
Test Type	**Exact Test**
N (Population Size)	**Infinite**
Alpha	**0.05**
n (Sample Size)	**200 300 500 800**
Input Type	**Ratios**
P0 (Null Proportion)	**0.05**
P1 (Alternative Proportion)	**1.25 1.50 1.75**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Exact Test
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Target Alpha	Actual Alpha*	Reject H0 If R ≤ | R ≥
0.12473	200	0.0500	0.0625	1.2500	0.0500	0.0328	3|17
0.33215	200	0.0500	0.0750	1.5000	0.0500	0.0328	3|17
0.58605	200	0.0500	0.0875	1.7500	0.0500	0.0328	3|17
0.13148	300	0.0500	0.0625	1.2500	0.0500	0.0328	7|24
0.40189	300	0.0500	0.0750	1.5000	0.0500	0.0328	7|24
0.70620	300	0.0500	0.0875	1.7500	0.0500	0.0328	7|24
0.21380	500	0.0500	0.0625	1.2500	0.0500	0.0395	15|36
0.62478	500	0.0500	0.0750	1.5000	0.0500	0.0395	15|36
0.90715	500	0.0500	0.0875	1.7500	0.0500	0.0395	15|36
0.35086	800	0.0500	0.0625	1.2500	0.0500	0.0420	27|53
0.84317	800	0.0500	0.0750	1.5000	0.0500	0.0420	27|53
0.98822	800	0.0500	0.0875	1.7500	0.0500	0.0420	27|53

* Power and actual alpha were computed using binomial enumeration of all possible outcomes.

© NCSS, LLC. All Rights Reserved.

This report shows the values of each of the parameters, one scenario per row. Because of the discrete nature of the binomial distribution, the stated (Target) alpha is usually greater than the actual alpha. Hence, we also show the Actual Alpha along with the rejection region. The values from this table are plotted in the chart below.

Charts Section

These plots show the relationship between power and P1/P0 in this example. We note that 80% power is achieved with a sample size of about 720 when P1/P0 is 1.50, which was the specific ratio of interest.

© NCSS, LLC. All Rights Reserved.

Example 9 – Validation of Sample Size Calculation for the Z-Test with S(P0) using Ryan (2013)

Ryan (2013) page 106 gives the results of a sample size calculation for a one-sided Z-test with S(P0). When P0 = 0.50, P1 = 0.6, power = 0.8, and alpha = 0.05, Ryan (2013) reports a required sample size of 153.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 9** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**One-Sided**
Test Type	**Z-test using S(P0)**
Power	**0.8**
Alpha	**0.05**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.6**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0)
Alternative Hypothesis: One-Sided (H0: P ≤ P0 vs. H1: P > P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If |Z| >
0.80125	153	0.5000	0.6000	0.1000	0.0500	1.6449

* Power was computed using the normal approximation method.

PASS calculated the same sample size of 153 as Ryan (2013).

© NCSS, LLC. All Rights Reserved.

Example 10 – Validation of Effect Size Calculation for the Z-Test using S(P0) with C.C. using Fleiss, Levin, and Paik (2003)

Fleiss, Levin, and Paik (2003) page 34 gives the results of an effect size calculation for a one-sided Z-test using S(P0) with Continuity Correction. When n = 100, P0 = 0.06, power = 0.9, and alpha = 0.05, Fleiss, Levin, and Paik (2003) reports that the detectable value for P1 is less than 0.01.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 10** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**One-Sided**
Test Type	**Z-test using S(P0) with C.C.**
Power	**0.90**
Alpha	**0.05**
n (Sample Size)	**100**
Input Type	**Proportions**
P0 (Null Proportion)	**0.06**
P1 (Alternative Proportion)	**Search < P0**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(P0) and C.C.
Alternative Hypothesis: One-Sided (H0: P ≥ P0 vs. H1: P < P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If Z <
0.90000	100	0.0600	0.0060	-0.0540	0.0500	-1.6449

* Power was computed using the normal approximation method.

PASS calculated the detectable P1 to be 0.006, which is less than 0.01 as reported by Fleiss, Levin, and Paik (2003).

© NCSS, LLC. All Rights Reserved.

Example 11 – Validation of Sample Size Calculation for the Z-Test with S(Phat) using Chow, Shao, and Wang (2008)

Chow, Shao, and Wang (2008) page 87 gives the results of a sample size calculation for a two-sided Z-test with S(Phat). When P0 = 0.30, P1 = 0.5, power = 0.8, and alpha = 0.05, Chow, Shao, and Wang (2008) reports a required sample size of 49.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 11** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**Two-Sided**
Test Type	**Z-test using S(Phat)**
Power	**0.80**
Alpha	**0.05**
Input Type	**Proportions**
P0 (Null Proportion)	**0.30**
P1 (Alternative Proportion)	**0.50**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Z-Test with S(Phat)
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If |Z| >
0.80743	50	0.3000	0.5000	0.2000	0.0500	1.9600

* Power was computed using the normal approximation method.

PASS calculates the same sample size to be 50, one higher than reported in Chow, Shao, and Wang (2008). The difference is due to rounding. **PASS** uses full precision but the book uses rounded values. If you calculate the power for n=49, you'll see that the actual power is 0.79956, which is slightly less than the required 0.8.

© NCSS, LLC. All Rights Reserved.

Example 12 – Validation of Binomial Enumeration Power Calculation for the Exact Test using Zar (2010)

Zar (2010) page 539 gives the results of a power analysis for a one-sided binomial exact test. When *n* = 12, P0 = 0.50, P1 = 0.833, and alpha = 0.05, Zar (2010) reports a power of 0.68.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 12** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Binomial Enumeration**
Max n for Binomial Enumeration	**10000**
Alternative Hypothesis	**One-Sided**
Test Type	**Exact Test**
N (Population Size)	**Infinite**
Alpha	**0.05**
n (Sample Size)	**12**
Input Type	**Proportions**
P0 (Null Proportion)	**0.50**
P1 (Alternative Proportion)	**0.833**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Exact Test
Alternative Hypothesis: One-Sided (H0: P ≤ P0 vs. H1: P > P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Target Alpha	Actual Alpha*	Reject H0 If R ≥
0.67624	12	0.5000	0.8330	0.3330	0.0500	0.0193	10

* Power and actual alpha were computed using binomial enumeration of all possible outcomes.

PASS calculated the power to be 0.67624, which agrees with the Zar (2010) value of 0.68.

© NCSS, LLC. All Rights Reserved.

Example 13 – Validation of Approximate Power Calculation for Exact Test using Zar (2010)

Zar (2010) page 541 gives the results of a power analysis for a two-sided binomial exact test. When *n* = 50, P0 = 0.50, P1 = 0.4 (the book has a typo and states this to be 0.5, but uses 0.4 in calculations), and alpha = 0.05, Zar (2010) reports a power of 0.29.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the **Tests for One Proportion** procedure window by expanding **Proportions**, then **One Proportion**, then **Test (Inequality)**, and then clicking on **Tests for One Proportion**. You may then make the appropriate entries as listed below, or open **Example 13** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Power**
Power Calculation Method	**Normal Approximation**
Alternative Hypothesis	**Two-Sided**
Test Type	**Exact Test**
Alpha	**0.05**
n (Sample Size)	**50**
Input Type	**Proportions**
P0 (Null Proportion)	**0.5**
P1 (Alternative Proportion)	**0.4**

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results

Numeric Results for Testing One Proportion using the Exact Test
Alternative Hypothesis: Two-Sided (H0: P = P0 vs. H1: P ≠ P0)

Power*	n	Proportion Given H0 P0	Proportion Given H1 P1	Difference P1-P0	Alpha	Reject H0 If R ≤ | R ≥
0.28905	50	0.5000	0.4000	-0.1000	0.0500	17|33

* Power was computed using the normal approximation method.

PASS calculated the power to be 0.28905, which agrees with the Zar (2010) value of 0.29.

© NCSS, LLC. All Rights Reserved.

- *Find* $\frac{N}{4}$
- *CF*$> \frac{N}{4}$
- *corresponding value of* x *is First quartile*

Definition 12. Third Quartile

- *Find* $\frac{3N}{4}$
- *CF*$> \frac{3N}{4}$
- *corresponding value of* x *is Third quartile*

Definition 13. *Second quartile=Median*

Find Quartiles for the frequency distribution

x	0	1	2	3	4	5
f	2	1	3	2	1	4

Definition 14.

- $Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$
- $Q_1 = l + \frac{h}{f}\left(\frac{N}{4} - c\right)$
- $Q_3 = l + \frac{h}{f}\left(\frac{3N}{4} - c\right)$
- *Quartile deviation*$=\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Example 10. *Find quartile deviation for the following continuous distribution*

class	*0-10*	*10-20*
Frequency	*8*	*20*

Definition 15. *Let* $x_1, x_2, \ldots x_n$ *are observations. Then*

- $\overline{x}$*=Mean=*$\frac{\sum x_i}{n}$
- *Mean deviation about mean=* $\frac{\sum |x_i - \overline{x}|}{n}$

Example 11. *Find mean deviation about the mean for 18,20,12,14,19,22,26,16,19,24.*

Definition 16. *For frequency distribution, MD is defined as*

- $\overline{x} = mean = \frac{\sum f_i x_i}{\sum f_i}$
- *MD*$=\dfrac{\sum f_i |x_i - \overline{x}|}{\sum f_i}$

Find Mean deviation about the mean for

x	10	11	12	13	14
f	3	12	18	12	3

Definition 17. *For continuous frequency distribution, MD is defined as*

- $\overline{x} = mean = A + \frac{\sum f_i d_i}{\sum f_i}, d_i = \frac{x_i - A}{h}$
- $MD = \dfrac{\sum f_i |x_i - \overline{x}|}{\sum f_i}$

Example 12.

class	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*
f	*2*	*1*	*2*	*1*	*1*

Definition 18. Sample Median:

Let observations in the same are $x_1, x_2 \ldots x_n$, *the sample median defined by* $\widetilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ is even} \end{cases}$

Example 13. *The measurements were recorded for the drying time, in hours, of certain brand of latex paint. 3.4, 2.5,4.8,2.9,3.6,2.8. Assume the measurements are simple random sample.*

- *what is the sample size for the above sample?*
- *Calculate Sample mean and median?*
- *Plot the data by way of dot plot*

Definition 19. Frequency distribution: *Organize and present frequency counts so that the information can be interpreted more easily.*

Example 14.

Age	*Frequency*
54	*3*
55	*1*
56	*1*
57	*2*
58	*2*
60	*2*

In case of frequency distribution, $Mean = \dfrac{\sum fx}{\sum f}$

Example 15. *If 5,8,6,2 occur with frequencies 3,2,4,1 resp. Find the arithmetic mean*

Definition 20.

For Raw distribution $x_1, x_2, \ldots x_n$, *S.D*$=\sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$

If A is assumed mean, $d = x - A$, *S.D*$=\sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$

For frequency distribution, S.D$=h\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{n}\right)^2}$

Coefficient of variation $=\frac{SD}{mean} * 100$

Example 16.

1. *Find SD of mean and SD of 60,47,17,43,30*
2. *The weekly salaries of a group of employee given in the following data. Find SD and coefficient of variation and discuss the behavior of the curve*

salary	*75*	*80*	*85*	*90*	*95*	*10*
No. of Persons	*3*	*7*	*18*	*12*	*6*	*4*

3. *Find SD for the continuous frequency distribution*

Wages	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
No.	*2*	*1*	*3*	*2*	*1*

4. *For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores 43 and 35 were*

missed as 34 and 53 respectively.Find corrected mean and standard deviation to the corrected figures

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- skewness tells you the amount and direction of skew (departure from horizontal symmetry)
- Moment of coefficient if skewness $g_1 = \frac{m_3}{{m_2}^{\frac{3}{2}}}$
- If skewness is positive, the data are positively skewed or skewed right, that is the right tail of the distribution is longer than the left.
- If skewness is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer.
- If skewness is less than -1 or greater than +1, the distribution is highly skewed.
- If skewness is between -1 and - or between + and +1, the distribution is moderately skewed.
- If skewness is between - and +, the distribution is approximately symmetric.
- kurtosis tells you how tall and sharp the central peak is, relative to a standard bell curve.
- Moment of coefficient if skewness $\beta_2 = \frac{m_4}{{m_2}^2}$
- Measure the peakness of the normal distribution curve is mesokurtic. $\beta_2 = 3$
- A distribution with kurtosis < 3 is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A distribution with kurtosis > 3 is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

- Observations obtained by repeating several times

- Cannot be predicted

Definition 21. Sample Space(S): *Set of all possible outcomes of a statistical experiment*

Definition 22. Sample point or element: *Each outcome in the sample space*

Example 17.
- $S = \{H, T\}$
- $S_1 = \{1, 2, 3, 4, 5, 6\}$

- $S_2 = \{even, odd\}$
- *An experiment consists of flipping a coin and then flipping it a second time if a head occurs otherwise a die is tossed once.*

Definition 18. Event: *Subset of a sample space*

Example 12. *Let S=$\{t|t > 0\}$ where t is life time of certain electronic components and A=$\{t|0 < t < 5\}$*

- S={laptop, table, mouse, pen, wire}
- Complement, intersection, Mutually exclusive or disjoint, union
- How many sample points are there in a sample space when a pair of dice is thrown once?

Definition 19. *A probability of an event A is the sum of the weights of all sample points in A such that $P(S) = 1, P(A) > 0, P(\phi) = 0$*

Example 13. *A coin is tossed twice. What is the probability that at least 1 head occurs?*

Theorem 1. *If A and B are two events, then*
$P(AUB) = P(A) + P(B) - P(A \cap B)$

- $P(AUB) = P(A) + P(B)$ if A and B are mutually exclusive
- $P(A') = 1 - P(A)$

Example 14. *What is the probability of getting a total 7 or 11 when a pair of fair dice is tossed?*

Definition 20. *A random variable is a function that associates a real number with each element in the sample space.*

Example 15. *Two balls are drawn in succession without replacement from an urn containing 4 red balls and 3 black balls. The possible outcomes and the values y of the random variable Y, Y is number of red balls.*

- Random variables with 0 and 1 are possible values, then X is called Bernoulli random variable.

Definition 21. *If X is a Random variable which can take a finite or countably number of values*

- Tossing a coin for two times
- Tossing a coin until head appears

Definition 22. *Probability of an event A: $P(A) = \frac{n}{N}$ where n is number of outcomes corre-*

sponding to A, N is differently equally likely outcomes

The set of ordered pairs $(x, f(x))$ is called probability function, probability mass function or probability distribution if

- $f(x) \geq 0$
- $\sum f(x) = 1$

- Prob mass function of Binomial distribution
- Prob mass function of Poisson distribution
- Mean=$\sum_{i=1}^{n} p_i x_i$
- Variance $\sigma^2 = \sum_{i=1}^{n} p_i {x_i}^2 - \mu^2$
- Mean=np
- Variance=npq
- Frequency of the Binomial distribution= $N \times P(r)$

Example 16. *Find the probability distribution of the number of the heads in two tosses of a coin.*

Example 17. *Two cards are drawn successively with replacement from a well shuffled pack of 52 cards. Find the probability distribution of the number of kings.*

Example 18. *Find the probability distribution of the number of sixes in three tosses of a dice.*

Example 19. *A urn contains 4 white and 3 red balls. Find the probability distribution of the red balls in three draws with replacement, from the urn. Also find its mean and variance.*

Example 20. *A shipment of 6 television sets contains 2 defective sets. A hotel makes a random purchase 3 of the sets. If X is number of defective sets purchased by hotel. Find the probability distribution.*

Example 21. *A random variable X assume 4 values with probabilities* $\frac{1+3x}{4}, \frac{1-x}{4}, \frac{1+2x}{4}, \frac{1-4x}{4}$. *Find condition on x, such that there values represent the probability function of x.*

Example 22. *If the random variable X takes the values of 1,2,3,4 such that $2P(X = 1) = 3P(X = 2) = P(X = 3) = 5P(X = 4)$. Find the probability distribution function of X.*

Definition 23. *Binomial distribution is a discrete probability distribution which is obtained*

when the probability p of the happening of the event is same in all trials

Example 23. *In a Binomial distribution, the mean and standard deviations are 12 and 2 respectively. Find n and p.*

Example 24. *Using Binomial distribution, find the probability of rolling atmost 2 sixes in 5 rolls of a dice.*

Example 25. *An experiment succeeds twice as often as its fails. what is the probability that in next five trials there will be*

- *three successes*
- *atleast three successes*

Definition 24. *Cumulative distribution Function* $F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$

Example 26. *A random variable X has the following probability distribution*

x	*-2*	*-1*	*0*	*1*	*2*	*3*
P(x)	*0.1*	*k*	*0.2*	*2k*	*0.3*	*3k*

- *Find k*
- $P(X < 2)$
- $P(-2 < X < 2)$
- *Find the cumulative distributive of X*
- *Evaluate mean of X*

Definition 25. Continuous Probability distributions:
$P(a < X < b) = \int_a^b f(x)dx$

Definition 26. *Probability density function:*
The function $f(x)$ *is a probability density function for the continuous random variable* X *defined on* $\mathbb{R}$, *if*

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a < X < b) = \int_a^b f(x)dx$

Example 27. *Suppose that the error in the reaction temp in C, for a controlled lab experiment is a continuous random variable* X *having the probability density function* $f(x) =$

$$\begin{cases} \frac{x^2}{3} & -1 < x < 2 \\ 0 & otherwise \end{cases}$$

- *verify f is density function*
- *Find* $P(0 < X \leq 1)$
- *Find CDF of f*

Definition 27.
- *Mean of X is* $E(X) = \sum xP(X = x)$
- *Variance of X is* $E(X^2) - (E(X))^2$

Example 28. *A random variable X has the following probability distribution*

x	0	1	2	3	4	5	6	7
$P(x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2 + k$

- *Find k*
- $P(\frac{1.5<X<4.5}{X>2})$
- *The smallest value of* λ *for which* $P(X \leq \lambda) > \frac{1}{2}$

Example 29. *The probability of an infinite discrete distribution is given by* $P(X = j) = \frac{1}{2^j}, j = 1, 2, \ldots$. *Verify that the total probability is 1 and find mean and variance of the distribution. Find*

- *P(X is even),*
- *P(X* ≥ 5*),*
- *P(X is divisible by 3)*

Example 30. *If* $P(X = r) = k(1 - a)^{r-1}, 0 < a < 1$ *where* $r = 1, 2, 3, \ldots$. *find k, mean and variance of X*

Example 31. *A continuous random variable X has a pdf* $f(x) = kx^2e^{-x}, x \geq 0$. *Find k, mean and variance.*

Example 32. *The distribution function of a random variable X is given by* $F(x) = 1 - (1 + x)e^{-x}, x \geq 0$. *find the density function, mean and variance of X*

Example 33. *The CDF of a continuous random variable X is given by*
$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 < x < \frac{1}{2} \\ 1 - \frac{3}{25}(3 - x)^2 & \frac{1}{2} \leq x < 3 \\ 1 & x \geq 3 \end{cases}$ *Find the PDF of X and evaluate* $P(|X| \leq 1)$ *and* $P(\frac{1}{3} \leq X < 4)$ *using both PDF and CDF*

Multiple Regression

If the number of independent variables in a regression model is more than one, then the model is called as multiple regression. In fact, many of the real-world applications demand the use of multiple regression models.

A sample application is as stated below:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

where Y represents the economic growth rate of a country, X_1 represents the time period, X_2 represents the size of the populations of the country, X_3 represents the level of employment in percentage, X_4 represents the percentage of literacy, b_0 is the intercept and b_1, b_2, b_3 and b_4 are the slopes of the variables X_1, X_2, X_3 and X_4 respectively. In this regression model, X_1, X_2, X_3 and X_4 are the independent variables and Y is the dependent variable.

Regression Model with Two Independent Variables using Normal Equations:

Suppose the number of independent variables is two, then $Y = b_0 + b_1X_1 + b_2X_2$.
Normal equations are

$$\Sigma Y = nb_0 + b_1\Sigma X_1 + b_2\Sigma X_2$$
$$\Sigma YX_1 = b_0\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2$$
$$\Sigma YX_2 = b_0\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2$$

where n is the total number of combinations of observations. The solution to the above set of simultaneous equations will form the results for the coefficients b_0, b_1 and b_2 of the regression model.

Example 1: The annual sales revenue (in crores of rupees) of a product as a function of sales force (number of salesmen) and annual advertising expenditure (in lakhs of rupees) for the past 10 years are summarized in the following table. Fit a least squares regression model.

Annual sales revenue Y	20	23	25	27	21	29	22	24	27	35
Sales force X_1	8	13	8	18	23	16	10	12	14	20
Annual advertising expenditures X_2	28	23	38	16	20	28	23	30	26	32

Solution: Let the regression model be $Y = b_0 + b_1X_1 + b_2X_2$
where Y is the annual sales revenue; X_1 is the sales force; X_2 is the annual advertising expenditures.

Y	X_1	X_2	X_1^2	X_2^2	X_1X_2	YX_1	YX_2
20	8	28	64	784	224	160	560

Scanned by CamScanner

23	13	23	169	529	229	299	529
25	8	38	64	1444	304	200	950
27	18	16	324	256	288	486	432
21	23	20	529	400	460	483	420
29	16	28	256	784	448	464	812
22	10	23	100	529	230	220	506
24	12	30	144	900	360	288	720
27	14	26	196	676	364	378	702
35	20	32	400	1024	640	700	1120
253	**142**	**264**	**2246**	**7326**	**3617**	**3678**	**6751**

Substituting the required values in the normal equations, we get the following simultaneous equations

$$10\,b_0 + 142\,b_1 + 264\,b_2 = 253$$

$$142\,b_0 + 2246\,b_1 + 3617\,b_2 = 3678$$

$$264\,b_0 + 3617\,b_1 + 7326\,b_2 = 6751$$

The solution to the above set of simultaneous equation is $b_0 = 5.1483$, $b_1 = 0.6190$ and $b_2 = 0.4304$.

Therefore, the regression model is $Y = 5.1483 + 0.6190X_1 + 0.4304X_2$.

Scanned by CamScanner

R Lab Session : Part 2
(Solutions)

To see a review of how to start R, look at the beginning of Lab1
Lab1 http://www-stat.stanford.edu/ epurdom/RLab.htm

Probability Calculations

The following examples demonstrate how to calculate the value of the cumulative distribution function at (or the probability to the left of) a given number.

- **Normal(0,1) Distribution :**

```
> x <- c(-2,-1,0,1,2)
> x
[1] -2 -1  0  1  2
> pnorm(x)
[1] 0.02275013 0.15865525 0.50000000 0.84134475 0.97724987
```

- **Binomial(n, p) Distribution :**

```
> x <- c(0,1,2,5,8,10,15,20)
> pbinom(x,size=20,prob=.2)
[1] 0.01152922 0.06917529 0.20608472 0.80420779 0.99001821 0.99943659 0.99999999
[8] 1.00000000
```

- **Poisson(λ) Distribution :**

```
> x <- c(0,1,2,5,8,10,15,20)
> ppois(x,6)
[1] 0.002478752 0.017351265 0.061968804 0.445679641 0.847237494 0.957379076
[7] 0.999490902 0.999998545
```

Exercise : Calculate the following probabilities :

1. Probability that a normal random variable with mean 22 and variance 25

   (i) lies between 16.2 and 27.5
   ```
   pnorm(27.5,22,sd=5)-pnorm(16.2,22,sd=5)
   [1] 0.7413095
   ```
   (ii) is greater than 29 `1-pnorm(29,22,sd=5)`
   ```
   [1] 0.08075666
   ```
   (iii) is less than 17 `pnorm(17,22,sd=5)`
   ```
   [1] 0.1586553
   ```
   (iv) is less than 15 or greater than 25 `pnorm(15,22,sd=5)+1-pnorm(25,22,sd=5)`
   ```
   [1] 0.3550098
   ```

2. Probability that in 60 tosses of a fair coin the head comes up

   (i)

20,25 or 30 times

```
sum(dbinom(c(20,25,30),60,prob=0.5))
[1] 0.1512435
```

(ii)

less than 20 times

```
pbinom(19,60,prob=0.5)
[1] 0.0031088
```

(iii)

between 20 and 30 times `pbinom(30,60,prob=0.5)-pbinom(20,60,prob=0.5)`

```
[1] 0.5445444
```

3.

A random variable X has Poisson distribution with mean 7. Find the probability that

(i)

X is less than 5 `less or equal is:`

```
> ppois(5,7)
[1]0.3007083
less than is
> ppois(4,7)
[1]0.1729916
```

(ii)

X is greater than 10 (strictly)

```
> 1-ppois(10,7) [1] 0.0985208
```

(iii)

X is between 4 and 16 `> ppois(16,7)-ppois(3,7) [1] 0.9172764`

Quantiles

The following examples show how to common the quantiles of some common distributions for a given probability (or a number between 0 and 1).

- **Normal(0,1) Distribution :**

```
> y <- c(.01,.05,.1,.2,.5,.8,.95,.99)
> qnorm(y,mean=0,sd=1)
[1] -2.3263479 -1.6448536 -1.2815516 -0.8416212  0.0000000 0.8416212  1.6448536
[8]  2.3263479
```

- **Binomial(n,p) Distribution :**

```
> y <- c(.01,.05,.1,.2,.5,.8,.95,.99)
> qbinom(y,size=30,prob=.2)
[1]  1  3  3  4  6  8 10 11
```

- **Poisson(λ) Distribution :**

```
> y <- c(.01,.05,.1,.2,.5,.8,.95,.99)
> qpois(y,6)
[1]  1  2  3  4  6  8 10 12
```

Random Variable generation

The following examples illustrate how to generate random samples from some of the well-known probability distributions.

- **Normal(μ, σ^2) Distribution :**

  The first sample is from $N(0,1)$ distribution and the next one from $N(5,1)$ distribution.

```
> z <- rnorm(10)
> z
[1] -0.90361592 -1.96522764 -1.35107949 -0.10846423 0.29756634  1.40831606
[7] -0.07844737  1.40575257 -0.97511415 -0.33418299
```

  If you would like to see how the distribution of the sample points looks like ....

```
> w <- rnorm(1000,mean=5,sd=1)
> hist(w)
```

- **Binomial(n, p) Distribution :**

```
> k <- rbinom(20,size=5,prob=.2)
> k
[1] 1 2 0 1 0 0 0 2 0 1 0 0 0 0 0 2 4 1 1 1
```

- **Poisson(λ) Distribution :**

```
> x <- rpois(20,6)
> x
[1]  2  8  7  5  5  5  3  8  5  5  1  8  5  5  5  4 10  7  3  4
```

Exercise (Advanced) : Generate 500 samples from Student's t distribution with 5 degrees of freedom and plot the historgam. (Note: t distribution is going to be covered in class). The corresponding function is **rt** .
`hist(rt(500,5),40)`

Density Plots

- **Plotting the probability density function (pdf) of a Normal distribution :**

```
> x11()
> x <- seq(-4.5,4.5,.1)
> normdensity <- dnorm(x,mean=0,sd=1)
> plot(x,normdensity,type="l")
```

- **Plotting the probablity mass function (pmf) of a Binomial distribution :**

```
> par(mfrow=c(2,1))
> k <- c(1:30)
> plot(k,dbinom(k,size=30,prob=.15),type="h")
> plot(k,dbinom(k,size=30,prob=.4),type="h")
> par(mfrow=c(1,1))
```

- **Discrete Probabilities** For a discrete random variable, you can use the probability mass to find $P(X = k)$

```
> dbinom(3,size=10,prob=0.5)
[1] 0.1171875
```

** Note the distinction between the continuous (Normal) and the discrete (Binomial) distrubtions.

Exercise : Plot the probability mass functions for the Poisson distribution with mean 4.5 and 12 respectively. Do you see any similarity of these plots to any of the plots above? If so, can you guess why ?

Exercise : Recreate the probabilities that Professor Holmes did in class (Bin(5,.4)) [You can do it in 1 command!] How would you get the expected counts?

Q-Q plot

R has two different functions that can be used for generating a Q-Q plot. Use the function **qqnorm** for plotting sample quantiles against theoretical (population) quantiles of standard normal random variable.

Example :

```
> stdnormsamp <-rnorm(100,mean=0,sd=1)
> normsamp <- rnorm(100,mean=5,sd=1)
> binomsamp <-rbinom(100,size=20,prob=.25)
> poissamp <- rpois(100,5)

> par(mfrow=c(2,2))

> qqnorm(stdnormsamp,main="Normal Q-Q plot : N(0,1) samples")
> qqline(stdnormsamp,col=2)
> qqnorm(normsamp,main="Normal Q-Q plot : N(5,1) samples")
> qqline(normsamp,col=2)
> qqnorm(binomsamp,main="Normal Q-Q plot : Bin(20,.25) samples")
> qqline(binomsamp,col=2)
> qqnorm(poissamp,main="Normal Q-Q plot : Poisson(5) samples")
> qqline(poissamp,col=2)
```

Note: Systematic departure of points from the Q-Q line (the red straight line in the plots) would indicate some type of departure from normality for the sample points.

Use of function **qqplot** for plotting sample quantiles for one sample against the sample quantiles of another sample

Example :

```
> par(mfrow=c(2,1))

> qqplot(stdnormsamp,normsamp,xlab = "Sample quantiles : N(0,1) samples",
+ ylab = "Sample quantiles : N(5,1) samples")
> qqplot(stdnormsamp,binomsamp,xlab = "Sample quantiles : N(0,1) samples",
+ ylab = "Sample quantiles : Bin(20,.25) samples")
```

Exercise : Generate 100 samples from Student's t distribution with 4 degrees of freedom and generate the qqplot for this sample.
`qqnorm(rt(100,df=4))`Generate another sample of same size, but now from a t distribution with 30 degrees of freedom and generate the q-q plot. Do you see any difference ?
`qqnorm(rt(100,df=30))`

It should be evident to you that the t distribution is very far from normal, and the 30 degrees of freedom t is indistinguishable from Normal.

Susan Holmes 2004-10-31

Example 34. *If* $p(x) = \begin{cases} xe^{-\frac{x^2}{2}} & x \geq 0 \\ 0 & otherwise \end{cases}$

- *Show that* $p(x)$ *is a PDF of a continuous random variable*
- *Find its distribution function*

Example 35. *A continuous random variable* X *that can assume any value between* $x = 2$ *and* $x = 5$ *has density function given by* $f(x) = k(1 + x)$. *Find* $P(X < 4)$

Example 36. *A continuous random variable has a PDF* $f(x) = 3x^2, 0 \leq x \leq 1$. *Find* a *and* b *such that*

- $P(X \leq a) = P(X > a)$
- $P(X > b) = 0.05$

Definition 28. *Conditional density of* X *given* Y

- $f(x|y) = \frac{f(x,y)}{f_Y(y)}$

Definition 29. *Let* X *and* Y *are two discrete random variables, the probability distribution for the simultaneous occurrence can be represented by a function* $f(x, y)$
18-wheeler is to have its tires serviced and X *is number of miles these tires have been driven and* Y *is number of tires have been replaced.*

Definition 30. *The function* $f(x, y)$ *is joint probability distribution or probability mass function of the discrete random variables* X *and* Y *if*

- $f(x, y) \geq 0$
- $\sum\sum f(x, y) = 1$
- *f(x,y)=P(X=x, Y=y)*

Example 37. *Three balls drawn at random without replacement from a box containing 2 white, 3 red and 4 black balls. If* X *denote the number of white balls drawn and* Y *denote the number of red balls drawn. Find the joint prob distribution. For the bivariate probability distribution of* (x, y) *given below, find* $P(X \leq 1), P(Y \leq 3), P(X = 1|Y = 2), P(X \leq 1, Y \leq 3), P(X \leq 1|Y \leq 3), P(Y \leq 3|X \leq 1), P(X + Y \leq 4)$

Example 38. *A privately owned business operates both a drive in facility and walk in facility. On a randomly selected day, Let* X *and* Y *be the propositions of the time that the drive in and walk in facilities are in use. Let* $f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y) & 0 \leq x, y \leq 1 \\ 0 & elsewhere \end{cases}$

- *Verify* f *is joint PDF*

- *Find* $P([X,Y] \in A)$ *where* $A = \{(x,y)|0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$

Definition 31. *(For discrete case) The marginal distributions of* X *alone and of* Y *alone are* $g(x) = \sum_y f(x,y)$ *and* $h(y) = \sum_x f(x,y)$

Definition 32. *(For Continuous case) The marginal density of* X *alone and of* Y *alone are* $g_X(x) = g(x) = \int_{-\infty}^{\infty} f(x,y)dy$ *and* $h_Y(y) = h(y) = \int_{-\infty}^{\infty} f(x,y)dx$

Definition 33. *Conditional Probability*

- $P(B|A) = \frac{P(A \cap B)}{P(B)}$
- $P(X = x|Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$

Example 39. *For the bivariate probability distribution of* (x,y) *given below, find* $P(X \leq 1), P(Y \leq 3), P(X \leq 1, Y \leq 3), P(X \leq 1|Y \leq 3), P(Y \leq 3|X \leq 1), P(X + Y \leq 4)$

$X\|Y$	1	2	3	4	5	6
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$

Example 40. *The joint density for the random variables* (X,Y) *where* X *is the unit temperature change and* Y *is the proportion of spectrum shift that a certain atomic particle produces is* $f(x,y) = \begin{cases} 10xy^2 & 0 < x < y < 1 \\ 0 & elsewhere \end{cases}$

- *Find the marginal distributions and conditional density* $f(y|x)$
- *Find the probability that the spectrum shifts more than half of the total observations, given that the temperature is increased by 0.25 unit*

Example 41. *The joint PMF of* (x,y) *is given by* $P(x,y) = k(2x+3y), x = 0,1,2; y = 1,2,3.$ *Find the marginal distribution*

- *conditional distribution of* X *given* $Y = 1$
- *conditional distribution of* Y *given* $X = 2$

Definition 34. *Let* X, Y *be two random variables with joint probability distribution* $f(x,y)$ *and marginal distributions* $g(x)$ *and* $h(y)$. X *and* Y *are statistically independent if* $f(x,y) = g(x)h(y)$ *for all*x, y

Definition 35.
- $E(X) = \sum xP(X = x)$
- $E(XY) = \sum_x \sum_y xyP(X = x, Y = y)$

- $Cov(X,Y) = E(XY) - E(X)E(Y)$
- *correlation coefficient of X and Y is* $\frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$

Example 42. *The joint PMF of X and Y is*

$X \setminus Y$	-1	-1
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$

Find the coefficient of X and Y.

Example 43. *The joint pdf of the RV* (x, y) *is given by* $f(x, y) = kxye^{-x^2-y^2}, x, y > 0$. *Find the value of k and prove that X and Y are independent.*

Example 44. *A coin is tossed until a head appears. Find the expected number of tosses.*

Example 45. *An unbiased coin is tossed until a head appears. Find the expected number of tails preceding the first head*

Definition 36.
- *var(ax)=*a^2*var(x)*
- *E(ax+b)=aE(x)+b*
- *Raw moments about the origin* $\mu_r^{'} = \int x^r f(x)dx$
- $\mu_1^{'} = E(X) = \bar{x}$
- $\mu_2^{'} = E(X^2)$
- *Raw moments about the A* $\mu_r^{'} = \int (x - A)^r f(x)dx$
- *Central moments* $\mu_r = E(X - \bar{x})^r$
- $\mu_0 = 1$
- $\mu_1 = 0$
- $\mu_2 = var(x)$

Example 46. *Relation between raw and central moments*

- $\mu_1 = 0$
- $\mu_2 = \mu_2^{'} - (\mu_1^{'})^2$
- $\mu_3 = \mu_3^{'} - 3\mu_2^{'}\mu_1^{'} + 2(\mu_1^{'})^3$

Definition 37. *Coefficient of skewness:* $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

- If $\mu_3 < 0$, β_1 is associated with a negative sign
- If $\mu_3 > 0$, β_1 is associated with a positive sign
- If $\beta_1 = 0$, the distribution is symmetric
- If $\beta_1 > 0$, the distribution is positively skewed