

Lab 3

Binomial and Normal Distributions

Binomial Distribution (Sec. 3.8, p. 102)

Many experiments or studies result in a binary outcome, that is, only two outcomes are possible. For example, a treatment worked or didn't work, a patient went into remission or didn't go into remission. Even when there are many categories or a range of values for an outcome, we can partition the result into a binary response. For example, patients are less than 60 or at least 60 years old, an illness is either stages A, B, or C, or is in stages D or E.

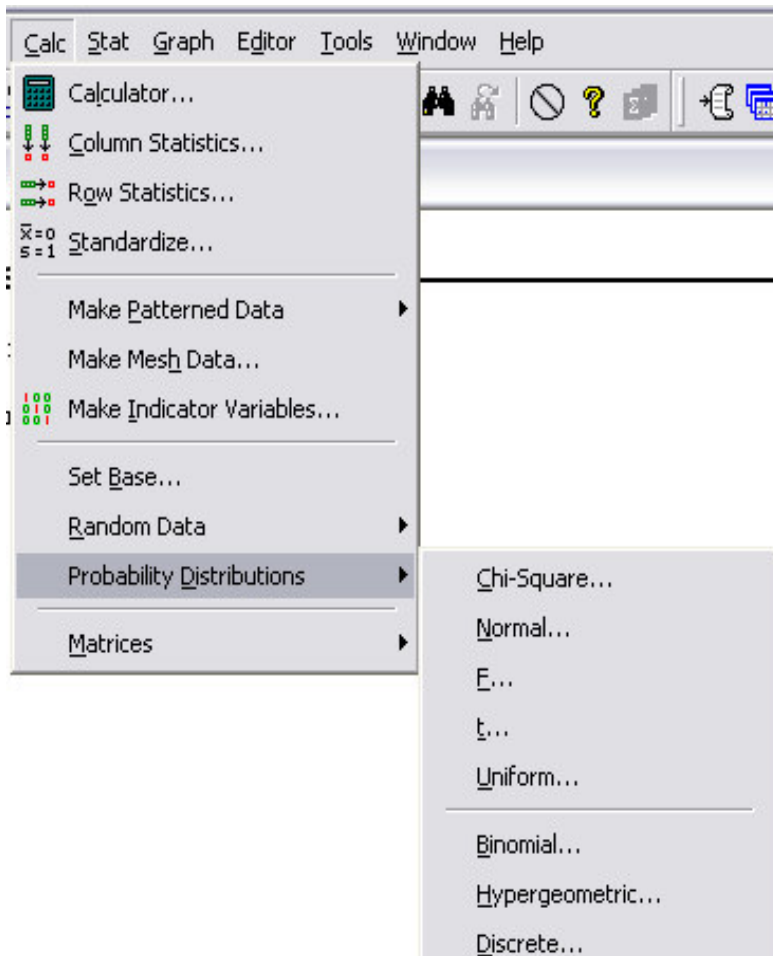
Independent-Trials Model

The important conditions for using the binomial distribution are that of **n independent trials**, the probability of success p is the same for each trial. (Choose the outcome of interest to be a success and the other to be a failure.)

The binomial distribution counts the number of success, each with probability p , in n independent trials. This is analagous to flipping a coin n times and counting the number of heads, where a head has probability p of landing side up (showing the obverse side).

See Example 3.45 on p. 106.

We can access Minitab's probability calculator from the Calc menu. We will be using both the Binomial and Normal distributions.



In example 3.45 we are interested in the Binomial distribution with $n=5$ trials, and the probability of success $p=.39$ of selecting a mutant (oh, just great). We want to know the probabilities associated with selecting a certain number of mutants from the 5 samples (trials) we randomly select.

Choosing Binomial from the menu shown above, we get the binomial distribution window, shown below.

↓	C1	C2
	Mutants	Probability
1	0	0.08445963
2	1	0.269994
3	2	0.345238
4	3	0.220726
5	4	0.070560
6	5	0.009022
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		

without specifying Optional storage (leaving that field blank) it outputs to the session window:

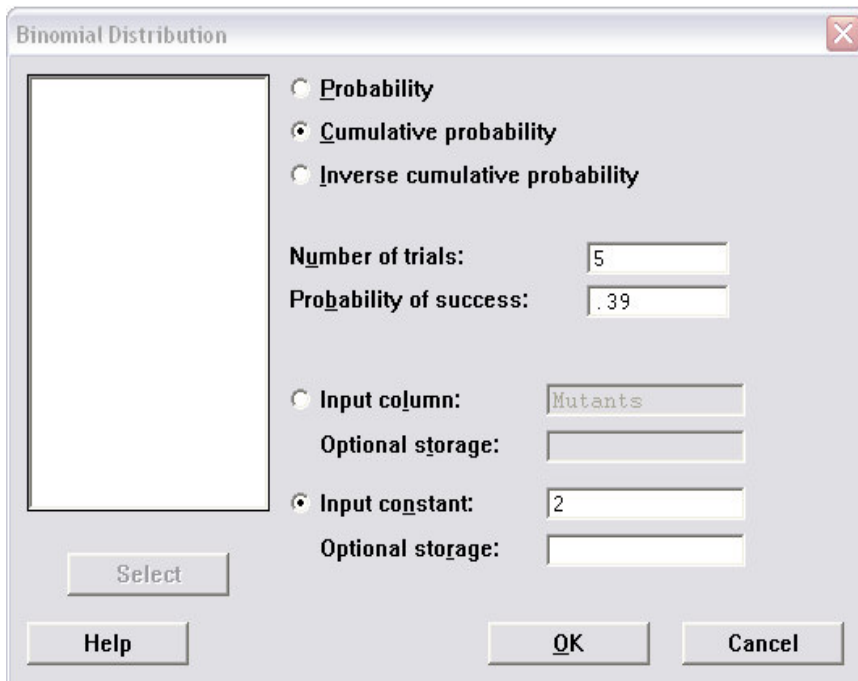
Probability Density Function

Binomial with $n = 5$ and $p = 0.39$

x	P(X = x)
0	0.084460
1	0.269994
2	0.345238
3	0.220726
4	0.070560
5	0.009022

This result corresponds with the table of Minitab output on p. 107.

It is sometimes useful to calculate the cumulative probability. The cumulative probability is the “area to the left” of a certain value. For example, the cumulative probability for $x=2$ is the probability of getting no more than 2 mutant (that is, 2 or fewer). In the table above, that is simply the sum of the probabilities of getting 0, 1, or 2 mutants. By selecting Cumulative probability instead, the result is given below. A quick check shows that this is the sum of the first three probabilities above.



Cumulative Distribution Function

Binomial with $n = 5$ and $p = 0.39$

x	$P(X \leq x)$
2	0.699692

What is the probability that more than 2 mutants are selected?

Because the total probability is 1, we can simply subtract the probability of selecting 2 or fewer from 1.

$$\Pr(Y > 2) = 1 - \Pr(Y \leq 2) = 1 - 0.699692 = .300308.$$

This calculation is not a great time-saver when $n=5$, since we're still only adding up 3 probabilities (for $Y=3, 4, 5$). But, when n is large, say $n=20$, now we would need to add 18 probabilities (for $Y=3, 4, \dots, 19, 20$). In this case, it is much easier to subtract from 1 the cumulative probability up to $Y=2$ (that is, $Y=0, 1, 2$). Homework 2 has such an example.

See example 3.50 for a situation where the binomial distribution does not apply since the trials are not independent, and the probability of success (contracting chickenpox) is not the same for each child since it is much greater for the other children once one child contracts it.

Normal Distribution (Sec. 3.8, p. 102)

The normal distribution is the most important statistical distribution (or density), both because it approximates other distributions and because it is the distribution of means (governed by the central limit theorem). Here we concentrate on an introduction, standardization, calculating areas under the distribution corresponding to probabilities of observing an outcome in a given interval, and obtaining percentiles using the inverse cumulative density function.

Introduction

The normal distribution is a bell-shaped, symmetric, unimodal density curve. It is characterized by its mean and standard deviation. Figure 4.7 on p. 123 give an example of three normal distributions plotted together. Normal distributions all have the same shape. The mean (μ) determines the location of the center of the distribution. The standard deviation (σ) determines its spread.

Standardization

A valuable property allows any normal density to be standardized to a normal distribution with mean 0 and standard deviation 1 (the standard normal density). Figure 4.8 shows a normal density with two horizontal scales. The top scale is in terms of the natural units centered at the mean μ , and indicated are distances from the mean in standard deviations σ . The bottom scale shows the standardized scale which is in terms of standard deviations and centered at 0. Notice that the standardized scale simply centers the original density at 0, then makes the units in terms of standard deviations.

The formula to convert between the natural scale Y and the standardized scale Z is shown on p. 124, and is $Z = (Y - \mu) / \sigma$.

Areas under the curve

For much of the rest of the course we will be interested in areas under the normal density (or it's cousin the t density), particularly because p -values are such areas when we conduct hypothesis tests.

Areas under the normal density can not be calculated directly from a formula, so these values are tabulated for lookup in Appendix Table 3 on p. 675, or can be calculated by software such as Minitab. Because all normal densities can be standardized, it is the standard normal density which is tabulated. The areas are values of the cumulative density function, which just means they are areas to the left of a particular value.

Reading Table 3.

For example, the areas to the left of a few values of z are given below:

$$\Pr(Z < 0) = .5$$

$$\Pr(Z < -1) = .1587$$

$$\Pr(Z < -1.55) = .0606$$

$$\Pr(Z < 1.55) = .9394 = 1 - \Pr(Z < -1.55) = 1 - .0606$$

$$\Pr(Z < -1.96) = .0250$$

Using Minitab for area calculations.

A single value:

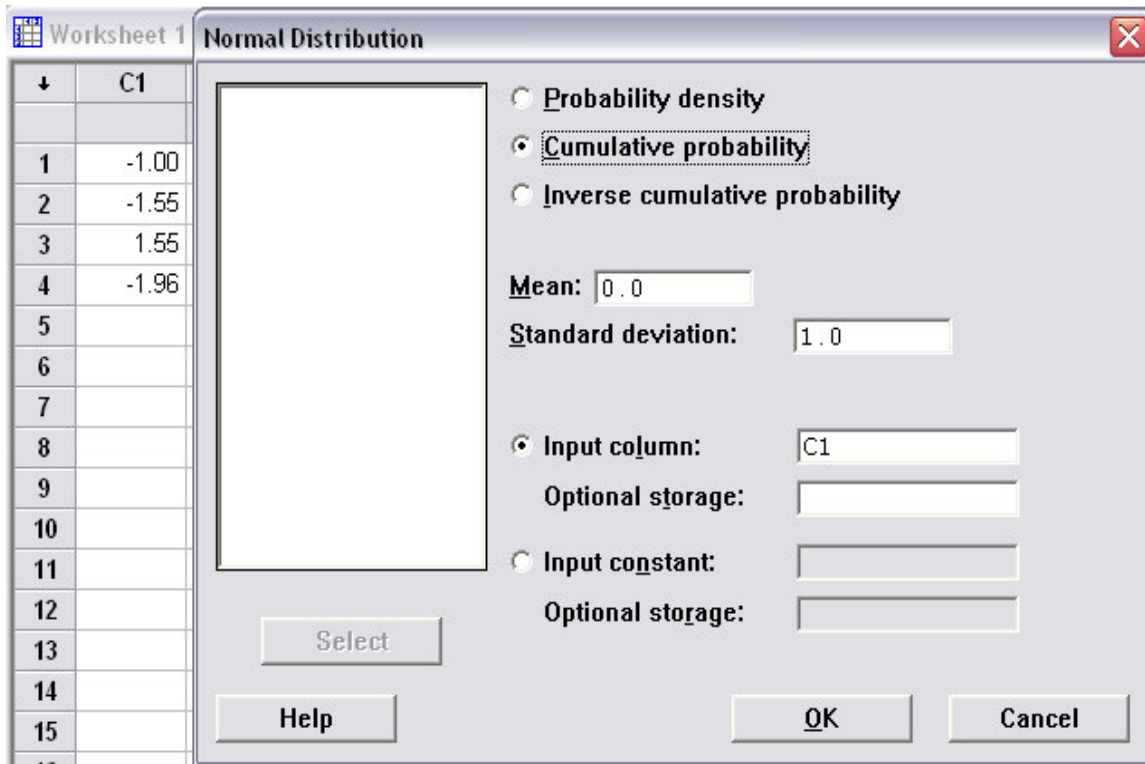
The screenshot shows the 'Normal Distribution' dialog box in Minitab. It has a title bar with a close button. On the left is a large empty rectangular area. To its right are three radio buttons: 'Probability density', 'Cumulative probability' (which is selected), and 'Inverse cumulative probability'. Below these are two input fields: 'Mean: 0.0' and 'Standard deviation: 1.0'. Further down are three more input fields: 'Input column:', 'Optional storage:', and 'Input constant: -1'. Below the 'Input constant' field is another 'Optional storage:' field. At the bottom left is a 'Select' button. At the bottom center are 'Help', 'OK', and 'Cancel' buttons.

Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1

x	P (X <= x)
-1	0.158655

Multiple values from a column:



Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1

x	P(X ≤ x)
-1.00	0.158655
-1.55	0.060571
1.55	0.939429
-1.96	0.024998

Because the normal density is symmetric about its mean, the standard normal density is symmetric about 0. Therefore, the area to the left of z is the same as the area to the right of $-z$. For example, the area to the left of 1 (.8413) is the same as the area to the right of -1 ($1 - .1587 = .8413$). Calculation of areas to the right are done by subtracting the area to the left from the total area of 1.

We are also interested in areas in an interval. If we are interested in the area under the standard normal density from 1 to 2, we can calculate this using the difference of areas to the left. That is $\Pr(1 < Z < 2) = \Pr(Z < 2) - \Pr(Z < 1) = .9772 - .8413 = .1359$.

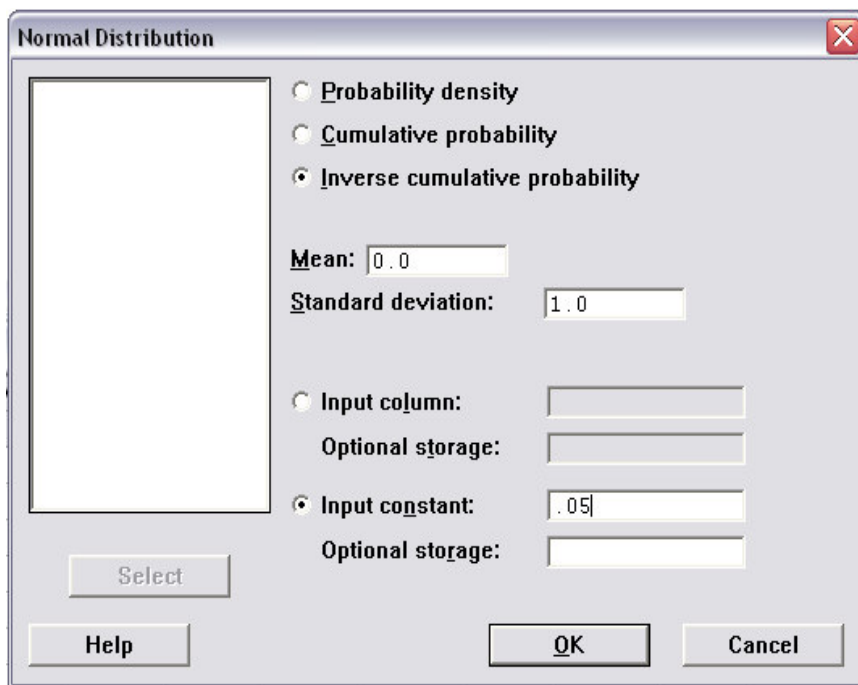
A nice rule of thumb to remember is the 68-95-99.7 rule which states that the areas 1, 2, and 3 standard deviations about the mean are 68%, 95%, and 99.7% of the area, respectively. Therefore, if you know the mean and standard deviation, you can quickly say where the 95% of the central area is. For example, if we know the lengths of a population of herring follow a normal distribution with mean length 54.0 mm and standard deviation 4.5, then we know 95% of the fish will be within 2 standard deviations

of the mean, so in the range $54 \pm 2*4.5$ or from 45 to 63 mm. So observing a fish outside this range is somewhat unlikely (an idea we will later use of hypothesis testing).

Percentiles

Sometimes we are interested in the value of Y (or Z) that give us a certain area to the left under the normal density. This is called the inverse cumulative density function because we are going from an area to a value of Y (or Z). This can be done from Table 3 approximately, but we'll do it in Minitab.

For example, the value from a normal density with mean 0 and standard deviation 1 to left of which is area .05 is given below as -1.64485.



Inverse Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1

P (X <= x)	x
0.05	-1.64485