

I. Statistics

Contents

1	What statistics is for	2
2	Descriptive statistics	2
2.1	The representation of data	2
2.2	Measures of location	2
2.3	Measures of spread	2
3	The normal distribution	2
3.1	The basic model	2
3.2	Applying the model	4
3.3	Working backwards	6
3.4	The standardized normal variable	7

1 What statistics is for

“lies, damn lies, and statistics”

2 Descriptive statistics

2.1 The representation of data

frequency distribution

discrete v continuous variables: grouped data

histogram

cumulative frequency diagram

2.2 Measures of location

mode

median

mean

2.3 Measures of spread

range

quartile range

variance, standard deviation

3 The normal distribution

3.1 The basic model

in the previous section we were considering empirical data, i.e. measurements, and representing them in various ways: graphically, in a histogram, or by calculating a few values, notably the mean and the standard deviation; we shall now investigate a particular statistical model that has been found to work very well in a wide range of situations: this model will enable us to calculate the distribution of values from just the mean and the standard deviation;

what do we mean by ‘a model’?

a model is a simplified representation of a complex real situation, which preserves certain essential aspects of the real situation, and can therefore be used to make plans and predictions about the real situation;

- a road map is a model of the landscape: it preserves the ways in which the real roads intersect and has a scale so that we can estimate distances; often colour is used to represent the width of the road; but other details, such as hills and other cars on the road are

not represented; an old map is still a model, but not a good one;

- all of physics depends on the use of models: a theory makes simplifying assumptions – that there is no friction, for instance, or that the gravitational field of the earth is homogeneous, or that springs are perfectly elastic; we can then apply our mathematics to these models and derive (approximate) predictions of the results of future experiments;

the normal distribution is a model of how measurements of certain variables are distributed; it is an excellent model for certain variables, and completely inappropriate for others;

- it is an excellent model for the distribution of the heights of people in a large population, or their weights, or the thickness of bolts produced by a machine, or the number of letters per line in a novel; note that these are mostly continuous distributions, in which the variable can take any value (in a certain range);

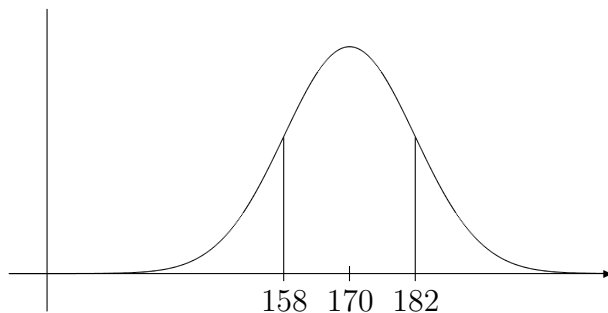
e.g. when we consider the heights of people, we would find that most people have a height somewhere near the average, and that few people are much taller or much shorter than the average;

- it is inappropriate for such variables as the scores when we roll a die many times, or the page numbers of a book; note that these are typically discrete distributions, in which the variable can only take certain separate values: we cannot score $\sqrt{21}$ on a die, and there is no page numbered 92.3 in any book;

e.g. when a die is rolled 120 times, say, we will get about 20 Ones, 20 Twos, and so on – we do not get a higher proportion of values in the middle!

the histogram of the normal distribution has a characteristic bell-shape curve;

e.g. if the heights of the students at the College have mean $m = 170$ cm and standard deviation $s = 12$ cm, then the histogram of the distribution of heights is that on the right;



note that as in any histogram, it is not the height of the graph that matters but the area; since the whole area under the curve represents all the students, we can estimate from the diagram that about 70% of students have a height between 158 cm and 182 cm, i.e. between $m - s$ and $m + s$;

to obtain the same graph on your calculator, enter:

```
[Y=] Y1= 10* [2nd+Distr 1] normalpdf(X,170,12)
[Window] 120, 210, 10, -0.1, 0.5, 0.1 [Graph]
```

so the function `normalpdf()` takes up to three arguments: the variable, the mean and the standard deviation – but we will not have much occasion to use it; (a convention that I will

use throughout is to specify the `[Window]` for a graph by giving the values of `Xmin`, `Xmax`, `Xscl`, `Ymin`, `Ymax`, `Yscl` in that order;)

notation:

when we found the mean and standard deviation of an observed distribution, the letters we used were m and s ; however, when we use the mean and standard deviation in a theoretical model, like the normal distribution, the convention is to use the Greek letters corresponding to m and s , which are μ (pronounced ‘myu’) and σ (‘sigma’);

$X \sim N(\mu, \sigma)$ is a short-hand way of writing that the variable x is normally distributed with mean μ and standard deviation σ ;

3.2 Applying the model

using the calculator it is easy to find areas under the normal distribution curve: if a variable is normally distributed, with mean μ and standard deviation σ , we can find the proportion of observations that lie between two values by using the function `normalcdf()`; this function takes up to four arguments: the bottom and the top values, the mean μ and the standard deviation σ ; (if you enter only the first three arguments, it will use $\sigma = 1$, and if only the first two, it will use $\mu = 0$, $\sigma = 1$; these are called the ‘default values’;)

e.g. consider the distribution of students’ heights x , with $\mu = 170$ cm and $\sigma = 12$ cm;

then the proportion of students that have a height between 160 cm and 175 cm is

$$p(160 \leq x \leq 175) =$$

$$\begin{aligned} & \cdot \quad [2\text{nd}+\text{Distr } 2] \quad \text{normalcdf}(160,175,170,12) = .459 \\ & = 45.9\%; \end{aligned}$$

and the proportion of students that are taller than 180 cm is

$$p(x \geq 180) =$$

$$\begin{aligned} & \cdot \quad [2\text{nd}+\text{Distr } 2] \quad \text{normalcdf}(180,230,170,12) = .202 \\ & = 20.2\%; \text{ note that in this case no upper limit was given, but the calculator does expect one, so we used a very large value for the upper limit – a value of } \mu + 5\sigma \text{ is safe;} \end{aligned}$$

similarly, the proportion of students that are shorter than 155 cm is

$$p(x \leq 155) =$$

$$\begin{aligned} & \cdot \quad [2\text{nd}+\text{Distr } 2] \quad \text{normalcdf}(110,155,170,12) = .106 \\ & = 10.6\%; \text{ this time there was no lower limit, so we used } \mu - 5\sigma; \end{aligned}$$

when one has to make almost the same entry in the calculator repeatedly, as in the last example, it saves time to use `2nd+Entry`: the last line that was entered is displayed again, and one can then go back over it and make the necessary changes; pressing that key combination repeatedly brings back earlier lines that are still stored in memory;

in the above conditions, it does not matter if we write “less than or equal” or “less than”, since the proportion of people that are precisely of a certain height is zero: even in a large

population, nobody is 155.0000... cm tall; or to put it differently, in the histogram, the area above one x -value is zero;

the proportion of observations in a certain range will be a number between 0 and 1, or 0% and 100%; we can use this to also calculate the expected number of observations in that range: if on average 0.1 = 10% of a box of lightbulbs are broken, then in a box of 60 lightbulbs we must expect $60 \cdot 0.1 = 6$ lightbulbs to be broken;

e.g. If the heights of 340 students are normally distributed with mean 170 cm and standard deviation 12 cm, how many of those students would you expect to be between 175 and 190 cm tall?

$$\mu = 170, \sigma = 12, N = 340$$

$$\text{number of students } n(175 \leq x \leq 190) = N \cdot p(175 \leq x \leq 190) =$$

$$. \quad 340 * [2\text{nd}+\text{Distr } 2] \text{ normalcdf}(175,190,170,12) = 98.8$$

(while there will never be a group of 98.8 students, an expected number of students can involve a fraction;)

unbiased estimates:

in most cases when one uses the normal distribution model, the mean μ and standard deviation σ of the whole population are not known, and we can only find the mean m and standard deviation s of a sample of observations; let n be the sample size, i.e. the number of observations in the sample;

we then use the m and s_n of the sample to estimate the μ and σ of the whole population; while it is indeed the case that the best estimate for μ is m , it turns out that the best ('unbiased') estimate for σ is not s_n but a slightly larger value $s_{n-1} = s_n \cdot \sqrt{n/(n-1)}$; note that if n is a large number, $s_{n-1} \approx s_n$;

when using **1-Var Stats**, the calculator gives two values for the standard deviation: the smaller value is the sample standard deviation s_n , the larger one is the unbiased estimate s_{n-1} of the population standard deviation – it is this latter value that should be used for the normal distribution model;

e.g. A machine fills bags of sugar, to a nominal weight of 1000 g. Eight filled bags are selected at random and are found to have weights of 1003, 1001, 1004, 998, 1002, 1000, 1003 and 1002 g. Find the mean and the standard deviation of the weights in this sample accurate to 3 decimal places.

Assuming the weights of bags to be normally distributed, use the unbiased estimates of the population mean and standard deviation to determine what proportion of bags will have a weight below 1000 g. If a day's production consists of 4000 bags, how many of these would be expected to have less than the nominal weight?

sample mean and standard deviation:

$$. \quad [\text{Stat } 1] \text{ enter values into L1 } [\text{Stat Calc } 1] \text{ 1-Var Stats } [2\text{nd}+\text{L1}] \text{ L1} =$$

$$m = 1001.625, s = 1.798$$

use unbiased estimates: $\mu = 1001.625$, $\sigma = 1.923$ (also from the calculator)

$p(x \leq 1000) =$

. [2nd+Distr 2] normalcdf(990,1000,1001.625,1.923) = 0.199
= 19.9%

$N = 4000$: $n(x \leq 1000) = 4000 \cdot 0.199 = 796$ bags below nominal weight;

3.3 Working backwards

in all the problems up to now, we were given the mean and the standard deviation of a normal distribution, and a range of values, and we had to find the proportion of values in that range; i.e. we were given all the required arguments of the calculator's `normalcdf()`-function;

in this section we will be given the proportion(s) and have to solve for one of the arguments; since this will require use of the equation solver, we will start with a simple example:

e.g. solve $x^2 = 5$:

. [Math 0 ↑] eqn:0=X^2-5 [Enter] X=2 [Alpha+Solve] X=2.24

note that

- the equation solver requires the equation to be entered in the form $0 = \dots$; of course any equation can be rewritten in that form;
- pressing [Enter] only enters the equation, it does not solve it: to solve it, a rough estimate or first approximation to the solution must be entered (– in this case one solution is close to 2, the other close to -2);
- when the value shown is a solution, a little square is displayed in front of $X = \dots$;

in the next three examples we solve for different arguments of the function `normalcdf()`:

e.g. a variable is normally distributed with mean 17 and standard deviation 3; find the critical value c such that 90% of all values are less than it;

$X \sim N(17, 3)$, $p(x \leq c) = 0.9$

. [Math 0 ↑] eqn:0= [2nd+Distr 2] normalcdf(0,X,17,3)-0.9

. [Enter] X=20 [Alpha+Solve] X=20.8

so $c = 20.8$;

(in this case, as in some others, if the rough estimate is too far away from the solution, the calculator may take a long time or not find the solution at all; a rough graph may often help to choose a first approximation;)

this problem can also be solved using another of the calculator's functions:

if $p(x \leq c) = p$, then `normalcdf()`: $c \rightarrow p$ but `invNorm()`: $p \rightarrow c$;

e.g. alternative method:

$X \sim N(17, 3)$, $p(x \leq c) = 0.9$

. [2nd+Distr 3] invNorm(0.9,17,3) = 20.8

so $c = 20.8$;

e.g. a variable is normally distributed with standard deviation 0.3, and 20% of all values are greater than 2.5; find the mean μ ;

$$X \sim N(\mu, 0.3), \quad p(x \geq 2.5) = 0.2$$

. eqn:0=normalcdf(2.5,5,X,0.3)-0.2 [Enter] X=2 [Alpha+Solve] X=2.25
so $\mu = 2.25$;

or:

$$X \sim N(\mu, 0.3), \quad p(x \geq 2.5) = 0.2$$

since the first argument of `invNorm()` is the proportion of values that are less than the result of the function: $p(x \leq 2.5) = 1 - 0.2 = 0.8$

. eqn:0=invNorm(0.8,X,0.3)-2.5 [Enter] X=2 [Alpha+Solve] X=2.25
so $\mu = 2.25$;

e.g. a variable is normally distributed with mean 45, and $\frac{1}{3}$ of all values are less than 40; find the standard deviation σ ;

$$X \sim N(45, \sigma), \quad p(x \leq 40) = \frac{1}{3}$$

. eqn:0=normalcdf(0,40,45,X)-1/3 [Enter] X=7 [Alpha+Solve] X=11.6

or:

. eqn:0=invNorm(1/3,45,X)-40 [Enter] X=7 [Alpha+Solve] X=11.6

so $\sigma = 11.6$;

3.4 The standardized normal variable

when we use the normal distribution model, the mean μ and standard deviation σ can have any value (except that of course $\sigma \geq 0$), so there are infinitely many such distributions; but sometimes it is useful to switch to a standard normal distribution, with mean 0 and standard deviation 1;

e.g. copy and complete the following table, using the statistical functions of the calculator if necessary:

						mean	s.d.
values x_r	2	4	6	6	7	$m =$	$s = 4$
$x_r - m$	-3	-1				0	
$z_r = (x_r - m)/s$							

as in this example, if a variable x is normally distributed with mean μ and standard deviation σ , then the standardized variable $z = (x - \mu)/\sigma$ will also be normally distributed, with mean 0 and standard deviation 1, i.e $Z \sim N(0, 1)$;

using a normal distribution table:

the most common normal distribution table lists the proportions $p(z \leq c)$ of values of the standardized variable z that are less than c , for $0 \leq c \leq 4$, and can therefore be used instead of a calculator;

e.g. consider again the distribution of students' heights x , with mean $\mu = 170$ cm and standard deviation $\sigma = 12$ cm;

$$p(x \leq 190) = p\left(z \leq \frac{190 - 170}{12}\right) \quad \text{-- standardizing the variable}$$

$$= p(z \leq 1.67) = 0.9525 \quad \text{-- value in the row labelled 1.6 and the column labelled 0.07 : 9525}$$

$$p(x \leq 165) = p\left(z \leq \frac{165 - 170}{12}\right) \quad \text{-- standardizing the variable}$$

$$= p(z \leq -1.25)$$

$$= p(z \geq 1.25) \quad \text{-- using symmetry of the distribution}$$

$$= 1 - p(z \leq 1.25) \quad \text{-- using } p(\text{not } A) = 1 - p(A)$$

$$= 1 - 0.8944 = 0.1056 \quad \text{-- value in the row labelled 1.2 and the column labelled 0.05 : 8944}$$

find c : $p(x \leq c) = 90\%$ -- so c is the 90th percentile height

$$p\left(z \leq \frac{c - 170}{12}\right) = 0.9000 \quad \text{-- standardizing the variable}$$

$$\frac{c - 170}{12} = 1.28 \quad \text{-- the value closest to 9000 in the table is 8997, in row 1.2 and column 0.08}$$

$$c = 185.4 \text{ cm}$$

e.g. The output of a machine producing bolts is checked, and a bolt is rejected if it is shorter than 99 mm or longer than 103 mm. At a certain setting of the machine, 3% of the bolts are rejected for being too short, and 9% for being too long. What are the mean and the standard deviation of the lengths of the bolts?

If the standard deviation cannot be adjusted, for what mean would the proportion of bolts that are rejected be a minimum? What is that proportion?

$$X \sim N(\mu, \sigma) : p(x \leq 99) = 3\%$$

$$p\left(z \leq \frac{99 - \mu}{\sigma}\right) = 0.03$$

$$\frac{99 - \mu}{\sigma} = \text{invNorm}(0.03) = -1.88$$

$$\text{similarly: } p(x \geq 103) = 9\%$$

$$p(x \leq 103) = 91\%$$

$$p\left(z \leq \frac{103 - \mu}{\sigma}\right) = 0.91$$

$$\frac{103 - \mu}{\sigma} = \text{invNorm}(0.91) = 1.34$$

$$\text{hence two equations: } 99 - \mu = -1.88 \sigma$$

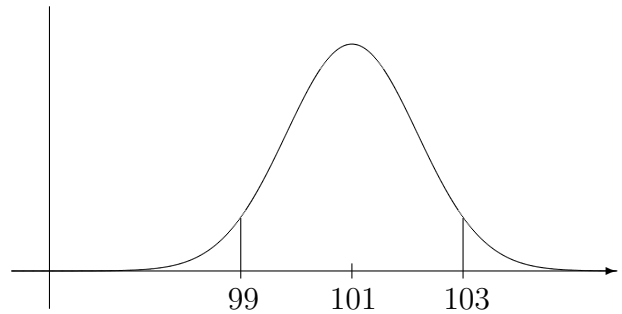
$$103 - \mu = 1.34 \sigma$$

$$\text{subtract and divide: } \sigma = 1.24$$

$$\mu = 101.33$$

the lowest proportion of bolts is rejected when the mean is set to be halfway between the critical values, so $\mu = 101$, as shown in the diagram;

if the curve was further to the left, i.e. $\mu < 101$ (while σ remains constant,) the area of the left tail would increase more than the area of the right tail would decrease, so the proportion of rejected bolts would be higher; similarly if the curve was further to the right, i.e. $\mu > 101$;



for $\mu = 101$, the proportion of acceptable bolts is $p(99 \leq x \leq 103) =$

$$. \quad \text{normcdf}(99, 103, 101, 1.24) = 0.893$$

so the proportion of rejected bolts is 10.7% (– instead of the previous 12%.)