

Lecture 6 : The Normal Distribution

Jonathan Marchini

November 12, 2004

1 Introduction

In previous lectures we have considered discrete datasets and discrete probability distributions. In practice many datasets that we collect from experiments consist of continuous measurements. For example, Figures 1, 2, 3 and 4 show histograms of real datasets consisting of continuous measurements. From such samples of continuous data we might want to test whether the data is consistent with a specific population mean value or whether there is a significant difference between 2 groups of data. To answer these question we need a probability model for the data. The Normal distribution is one such model and is used extensively throughout statistics.

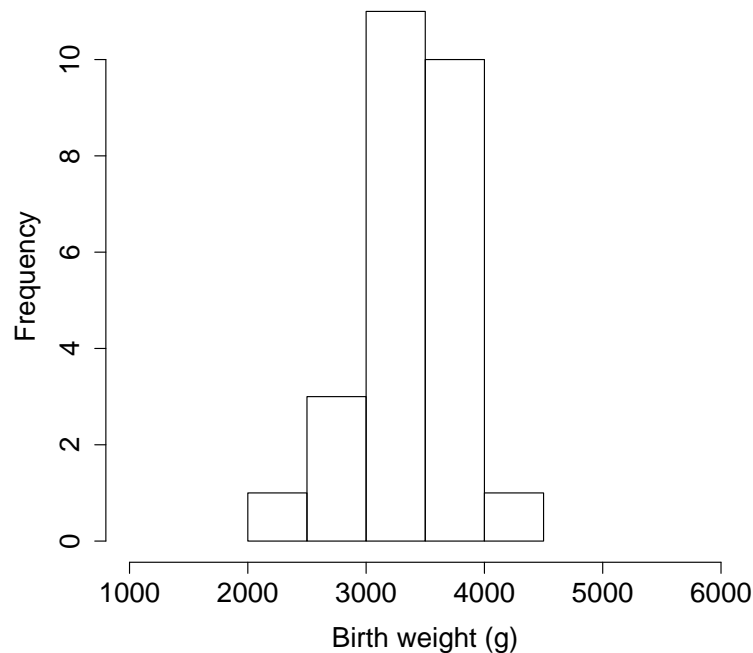


Figure 1: The birth weights of the babies in the Babyboom dataset

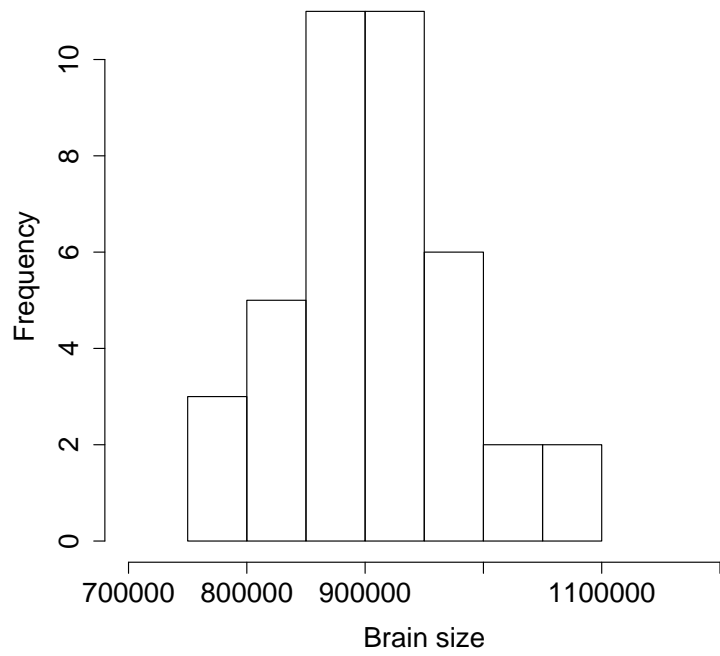


Figure 2: The brain sizes of 40 Psychology students

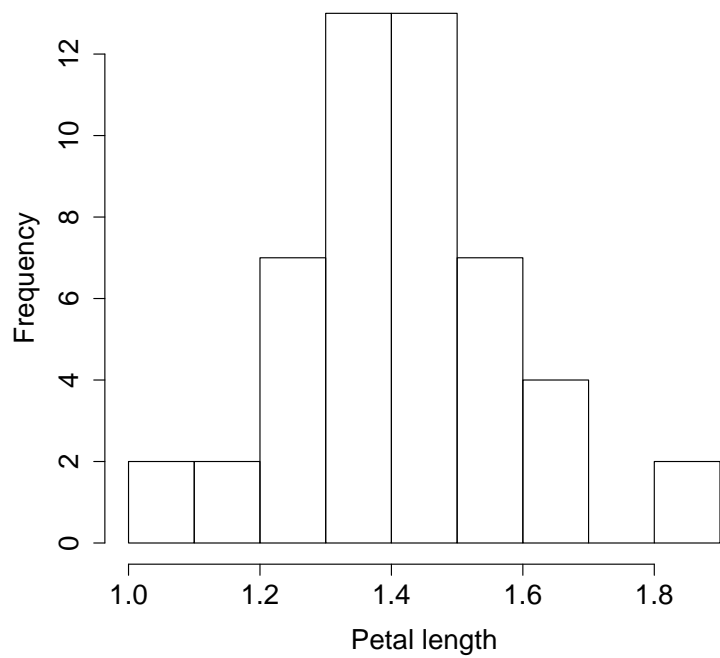


Figure 3: The petal length of a type of flower

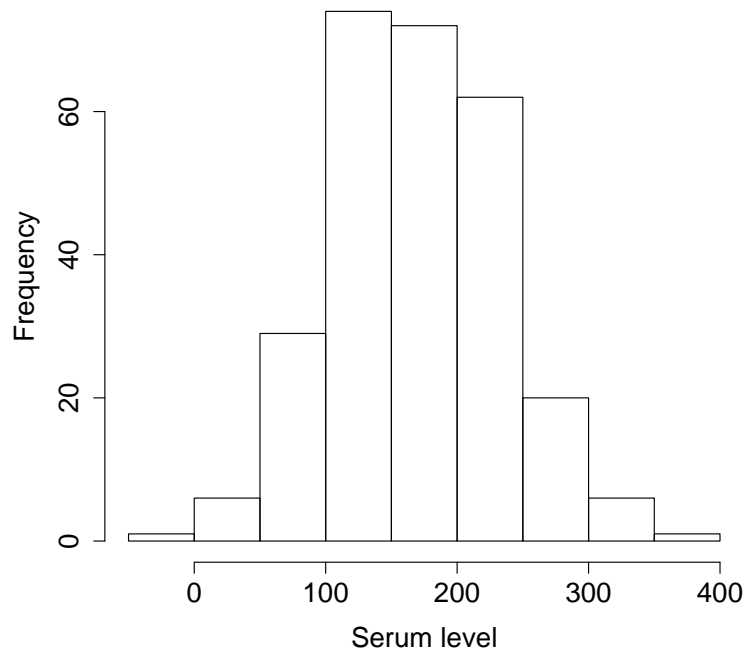


Figure 4: Serum level measurements from healthy volunteers

2 Continuous probability distributions

When we considered the Binomial and Poisson distributions we saw that the probability distributions were characterized by a formula for the probability of each possible discrete value. All of the probabilities together sum up to 1. We can visualize the density by plotting the probabilities against the discrete values (Figure 5). For continuous data we don't have equally spaced discrete values so instead we use a curve or function that describes the probability *density* over the range of the distribution (Figure 6). The curve is chosen so that the area under the curve is equal to 1. If we observe a sample of data from such a distribution we should see that the values occur in regions where the density is highest.

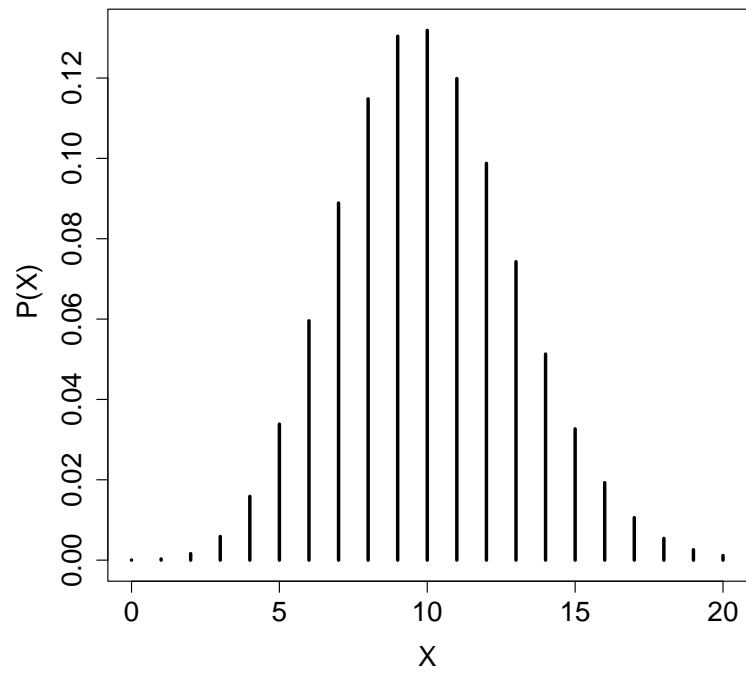


Figure 5: A discrete probability distribution

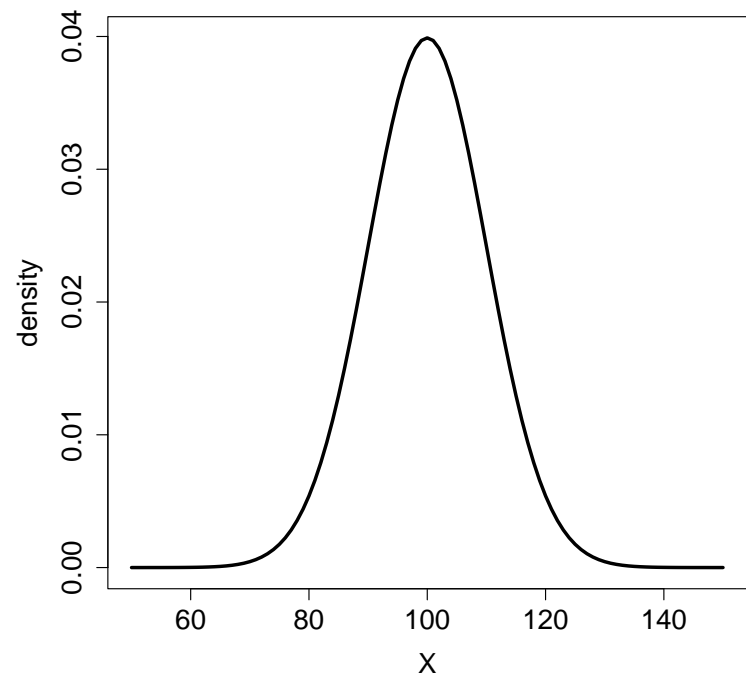


Figure 6: A continuous probability distribution

3 The Normal Distribution

There will be many, many possible probability density functions over a continuous range of values. The Normal distribution describes a special class of such distributions that are symmetric and can be described by the distribution mean μ and the standard deviation σ (or variance σ^2). 4 different Normal distributions are shown in Figure 7 together with the values of μ and σ . These plots illustrate how changing the values of μ and σ alter the positions and shapes of the distributions.

If X is Normally distributed with mean μ and standard deviation σ , we write

$$X \sim N(\mu, \sigma^2)$$

μ and σ are the **parameters** of the distribution.

The probability density of the Normal distribution is given by

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

For the purposes of this course we do not need to use this expression. It is included here for future reference.

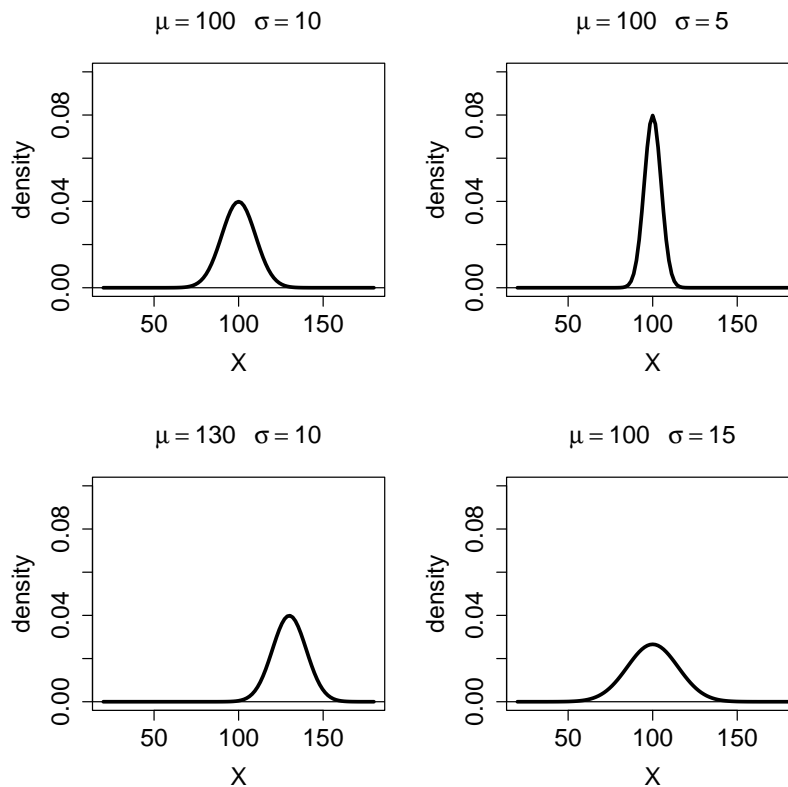


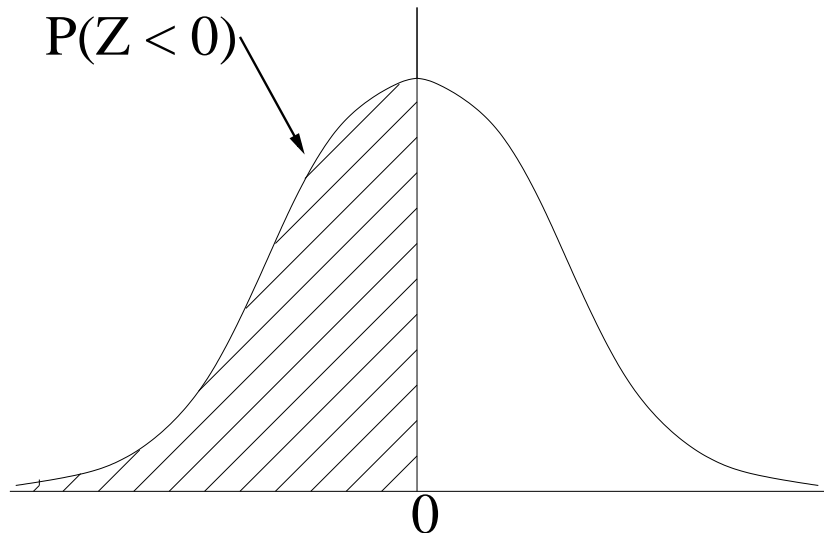
Figure 7: 4 different Normal distributions

3.1 Calculating probabilities from the Normal distribution

For a discrete probability distribution we calculate the probability of being less than some value x , i.e. $P(X < x)$, by simply summing up the probabilities of the values less than x .

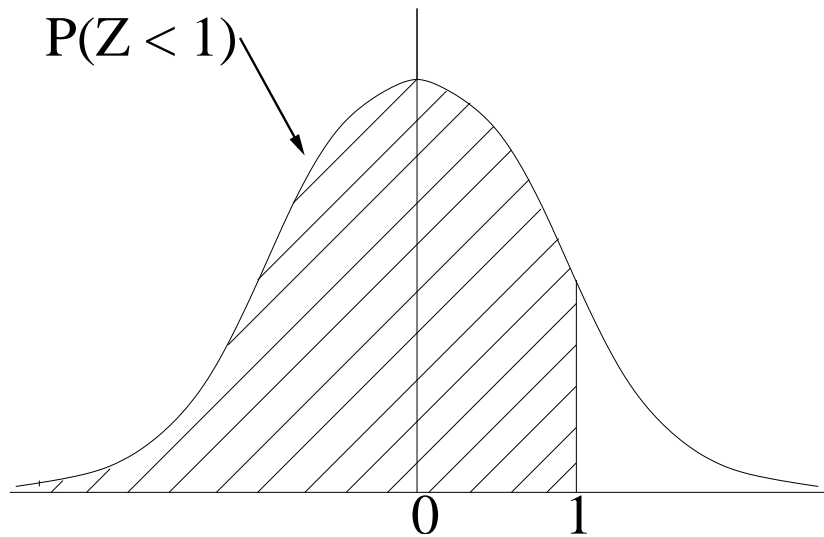
For a continuous probability distribution we calculate the probability of being less than some value x , i.e. $P(X < x)$, by calculating the area under the curve to the left of x .

For example, suppose $X \sim N(0, 1)$ and we want to calculate $P(X < 0)$?



For this example we can calculate the required area as we know the distribution is symmetric and the total area under the curve is equal to 1, i.e. $P(X < 0) = 0.5$.

What about $P(X < 1.0)$?

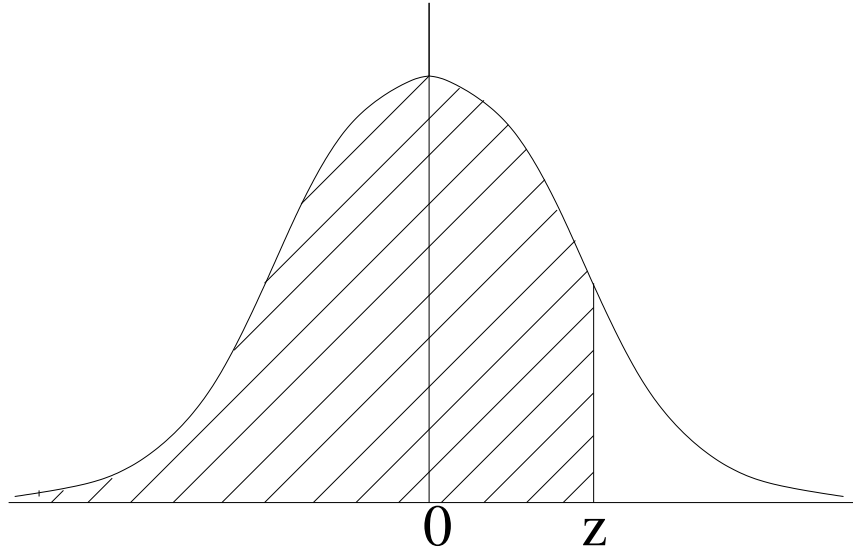


Calculating this area is not easy¹ and so we use probability tables. Probability tables are tables of probabilities that have been calculated on a computer. All we have to do is identify the right probability in the table and copy it down! Obviously it is impossible to tabulate all possible probabilities for all possible Normal distributions so only one special Normal distribution, $N(0, 1)$, has been tabulated.

The $N(0, 1)$ distribution is called the **standard Normal distribution**.

The tables allow us to read off probabilities of the form $P(Z < z)$. Most of the table in the formula book has been reproduced in Table 3.1. From this table we can identify that $P(X < 1.0) = 0.8413$ (this probability has been highlighted with a box)

¹For those Mathematicians who recognize this area as a definite integral and try to do the integral by hand please note that the integral *cannot* be evaluated analytically

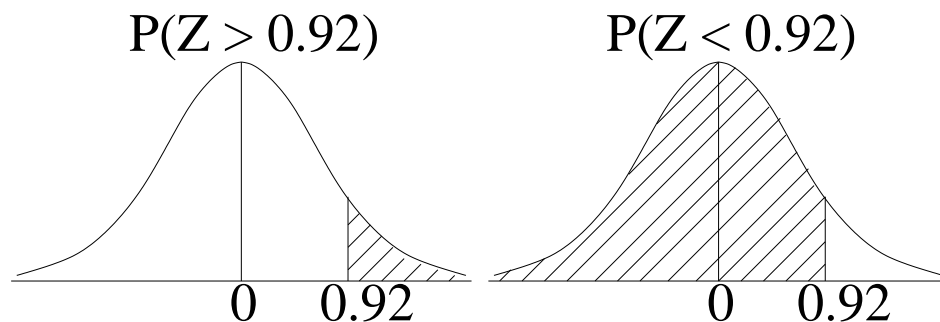


z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	0.5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	0.5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	0.6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	0.6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	0.7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	0.7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	0.7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	0.8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	0.8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	0.8643	8665	8686	8708	8729	8749	8770	8790	8810	8830

Table 1: $N(0, 1)$ probability table

Once we can know how to read tables we can calculate lots of other probabilities

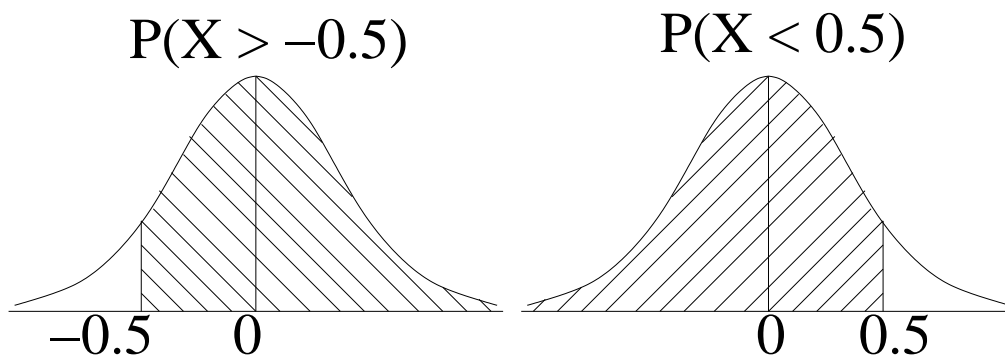
Example 1 $P(X > 0.92)$



We know that $P(X > 0.92) = 1 - P(X < 0.92)$ and we can calculate $P(X < 0.92)$ from the tables.

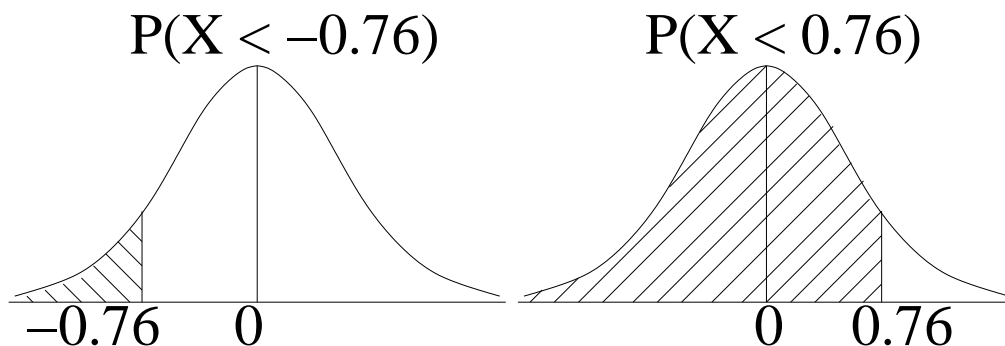
Thus, $P(X > 0.92) = 1 - 0.8212 = 0.1788$

Example 2 $P(X > -0.5)$?



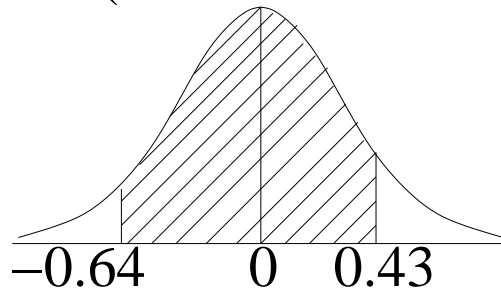
The Normal distribution is symmetric so we know that $P(X > -0.5) = P(X < 0.5) = 0.6915$

Example 3 We can use the symmetry of the Normal distribution to calculate $P(X < -0.76) = P(X > 0.76) = 1 - P(X < 0.76) = 1 - 0.7764 = 0.2236$

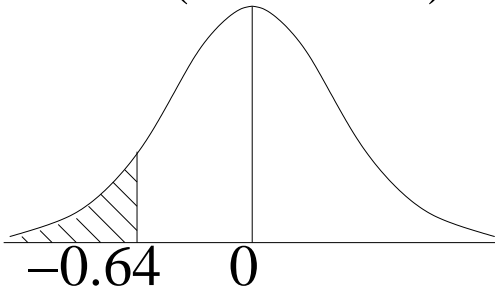


Example 4 $P(-0.64 < X < 0.43)$

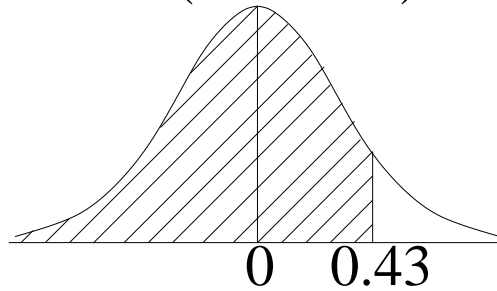
$$P(-0.64 < X < 0.43)$$



$$P(X < -0.64)$$



$$P(X < 0.43)$$



We can calculate this using

$$\begin{aligned} P(-0.64 < X < 0.43) &= P(X < 0.43) - P(X < -0.64) \\ &= 0.6664 - (1 - 0.7389) \\ &= 0.4053 \end{aligned}$$

Example 5 Consider $P(X < 0.567)$?

From tables we know that $P(X < 0.56) = 0.7123$ and $P(X < 0.57) = 0.7157$
To calculate $P(X < 0.567)$ we *interpolate* between these two values

$$P(X < 0.567) = 0.3 \times 0.7123 + 0.7 \times 0.7157 = 0.71468$$

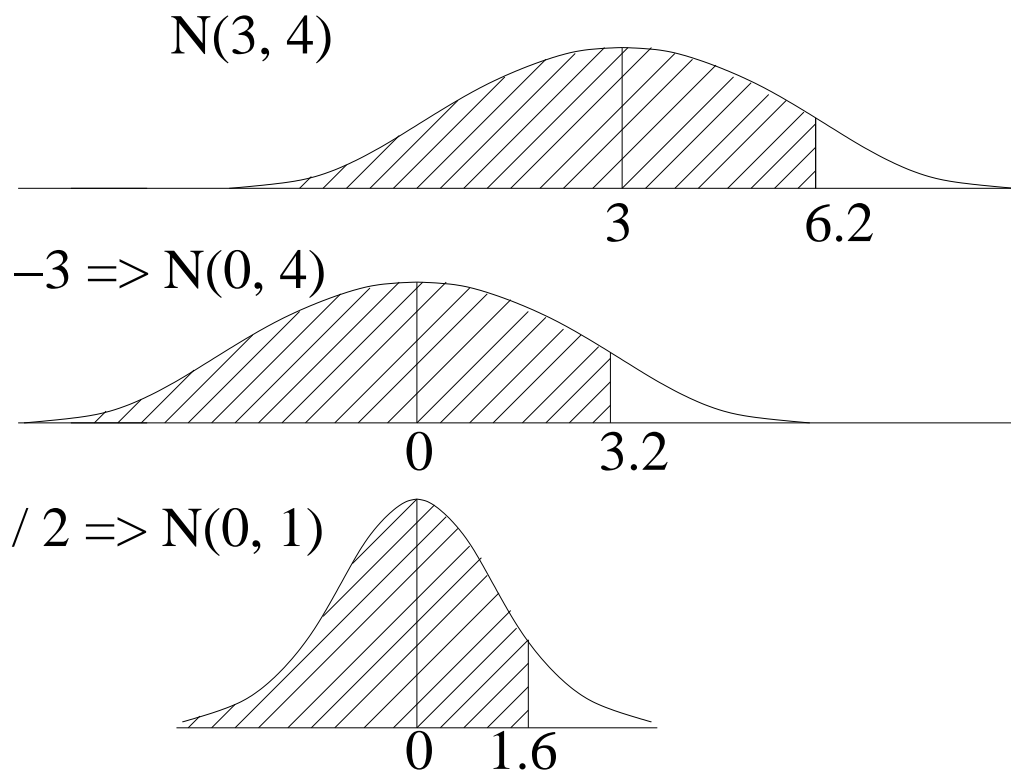
3.2 Standardization

All of the probabilities above were calculated for the standard Normal distribution $N(0, 1)$. If we want to calculate probabilities from different Normal distributions we convert the probability to one involving the standard Normal distribution. This process is called **standardization**.

Suppose $X \sim N(3, 4)$ and we want to calculate $P(X < 6.2)$. We convert this probability to one involving the $N(0, 1)$ distribution by

- (i) Subtracting the mean μ
- (ii) Dividing by the standard deviation σ

Subtracting the mean re-centers the distribution on zero. Dividing by the standard deviation re-scales the distribution so it has standard deviation 1. If we also transform the boundary point of the area we wish to calculate we obtain the equivalent boundary point for the $N(0, 1)$ distribution. This process is illustrated in the figure below. In this example, $P(X < 6.2) = P(Z < 1.6) = 0.9452$ where $Z \sim N(0,1)$



This process can be described by the following rule

$$\text{If } X \sim N(\mu, \sigma^2) \text{ and } Z = \frac{X - \mu}{\sigma} \text{ then } Z \sim N(0, 1)$$

Other rules that are often used are

If X and Y are two independent normal variable such that

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2)$$

then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$aX \sim N(a\mu_1, a^2\sigma_1^2)$$

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

Example 7 Suppose two rats A and B have been trained to navigate a large maze. The time it takes rat A is normally distributed with mean 80 seconds and standard deviation 10 seconds. The time it takes rat B is normally distributed with mean 78 seconds and standard deviation 13 seconds. On any given day what is the probability that the average time the rats take to run the maze is greater than 82 seconds?

$$X = \text{Time of run for rat A} \quad X \sim N(80, 10^2)$$

$$Y = \text{Time of run for rat B} \quad Y \sim N(78, 13^2)$$

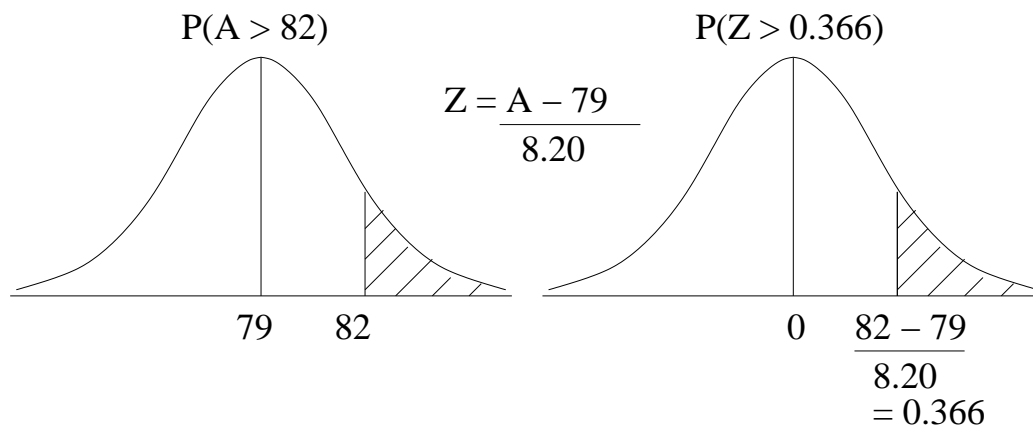
Let $A = \frac{X+Y}{2} = \frac{1}{2}X + \frac{1}{2}Y$ be the average time of rats A and B

$$\text{Then } A \sim N\left(\frac{1}{2}80 + \frac{1}{2}78, \left(\frac{1}{2}\right)^2 10^2 + \left(\frac{1}{2}\right)^2 13^2\right) = N(79, 67.25)$$

We want $P(A > 82)$

$$A \sim N(79, 67.25)$$

$$Z \sim N(0, 1)$$



$$\begin{aligned}
 P(A > 82) &= P\left(\frac{A - 79}{\sqrt{67.25}} < \frac{82 - 79}{\sqrt{67.25}}\right) = P(Z > 0.366) \quad \text{where } Z \sim N(0, 1) \\
 &= 1 - (0.4 \times 0.6406 + 0.6 \times 0.6443) \\
 &= 0.35718
 \end{aligned}$$

3.4 Using the Normal tables backwards

Example 8

The marks of 500 candidates in an examination are normally distributed with a mean of 45 marks and a standard deviation of 20 marks.

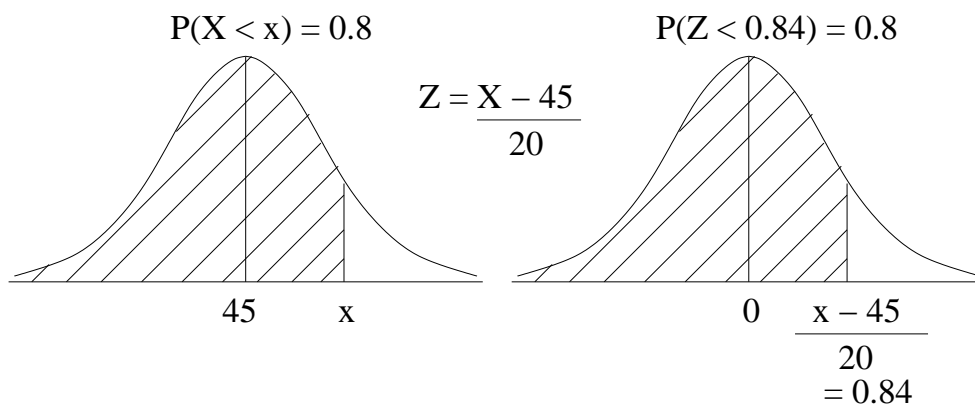
If 20% of candidates obtain a distinction by scoring x marks or more, estimate the value of x .

We have $X \sim N(45, 20^2)$ and we want x such that $P(X > x) = 0.2$

$$\Rightarrow P(X < x) = 0.8$$

$$X \sim N(45, 400)$$

$$Z \sim N(0, 1)$$



Standardizing this probability we get

$$P\left(\frac{X - 45}{20} < \frac{x - 45}{20}\right) = 0.8$$
$$\Rightarrow P\left(Z < \frac{x - 45}{20}\right) = 0.8$$

From the tables we know that $P(Z < 0.84) \approx 0.8$ so

$$\frac{x - 45}{20} \approx 0.84$$
$$\Rightarrow x \approx 45 + 20 \times 0.84 = 61.8$$

4 The Normal approximation to the Binomial

Under certain conditions we can use the Normal distribution to approximate the Binomial distribution. This can be very useful when we need to sum up a large number of Binomial probabilities to calculate the probability that we want.

For example, Figure 8 compares a $\text{Bin}(300, 0.5)$ and a $\text{N}(150, 75)$ which both have the same mean and variance. The figure shows that the distributions are very similar.

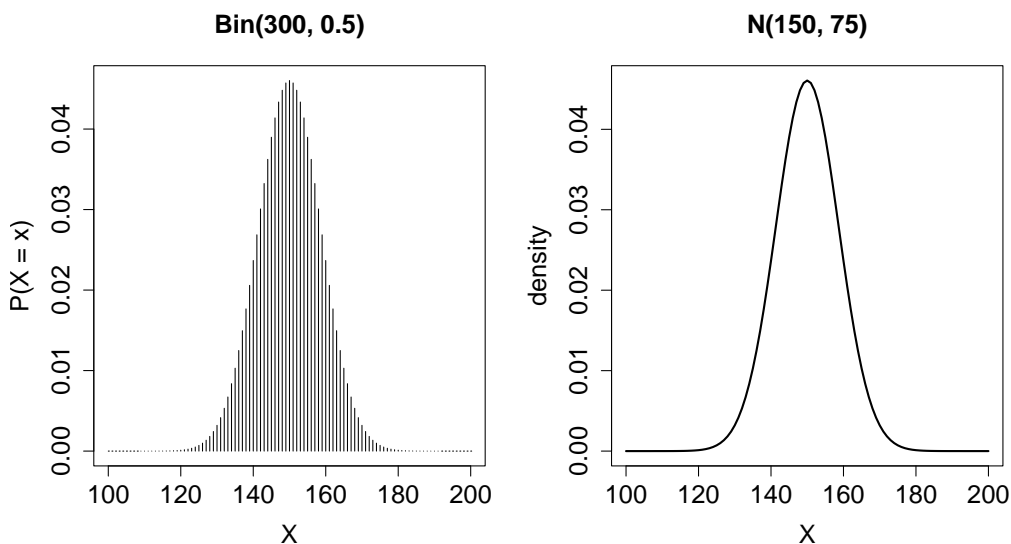


Figure 8: Comparison of a $\text{Bin}(300, 0.5)$ and a $\text{N}(150, 75)$ distribution

In general

If $X \sim \text{Bin}(n, p)$ then

$$\mu = np$$

$$\sigma^2 = npq \quad \text{where } q = 1 - p$$

For large n and p not too small or too large

$$X \sim \text{N}(np, npq)$$

$n > 10$ and $p \approx \frac{1}{2}$ OR $n > 30$ and p moving away from $\frac{1}{2}$

Example 8

Suppose $X \sim \text{Bin}(12, 0.5)$ what is $P(4 \leq X \leq 7)$?

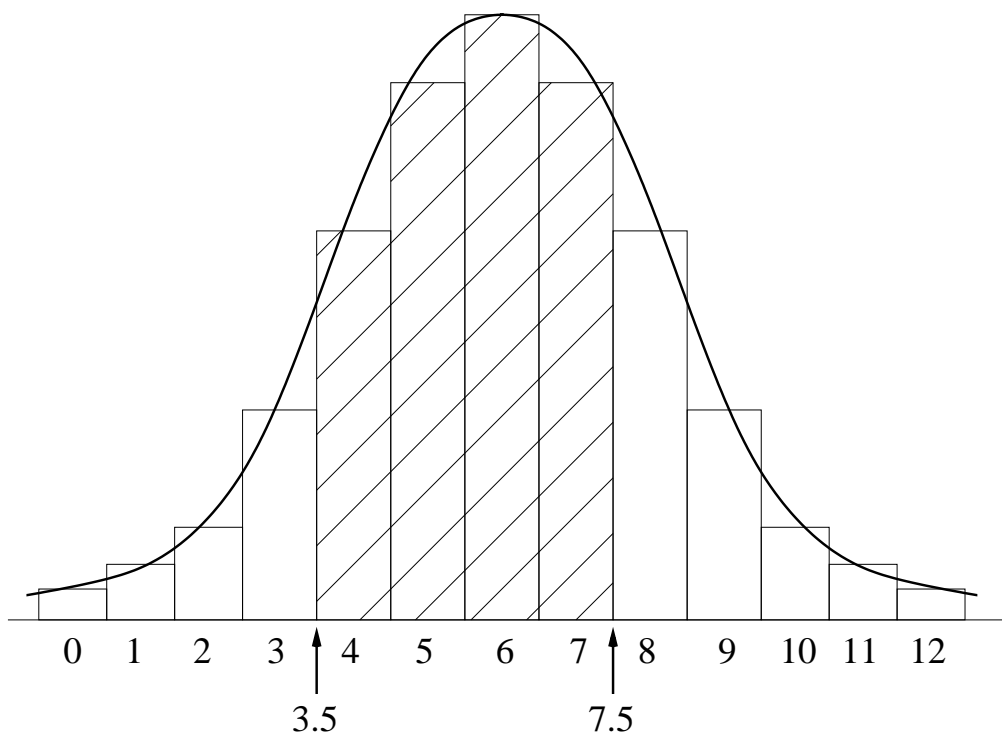
For this distribution we have

$$\begin{aligned}\mu &= np = 6 \\ \sigma^2 &= npq = 3\end{aligned}$$

So we can use a $N(6, 3)$ distribution as an approximation.

Unfortunately, it's not quite so simple. We have to take into account the fact that we are using a *continuous* distribution to approximate a *discrete* distribution. This is done using a **continuity correction**. The continuity correction appropriate for this example is illustrated in the figure below

In this example, $P(4 \leq X \leq 7)$ transforms to $P(3.5 < X < 7.5)$



$$\begin{aligned}P(3.5 < X < 7.5) &= P\left(\frac{3.5 - 6}{\sqrt{3}} < \frac{X - 6}{\sqrt{3}} < \frac{7.5 - 6}{\sqrt{3}}\right) \\ &= P(-1.443 < Z < 0.866) \quad \text{where } Z \sim N(0, 1) \\ &= 0.732\end{aligned}$$

The exact answer is 0.733 so in this case the approximation is very good.

5 The Normal approximation to the Poisson

We can also use the Normal distribution to approximate a Poisson distribution under certain conditions.

In general,

If $X \sim \text{Po}(\lambda)$ then

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

For large λ (say $\lambda > 20$)

$$X \sim N(\lambda, \lambda)$$

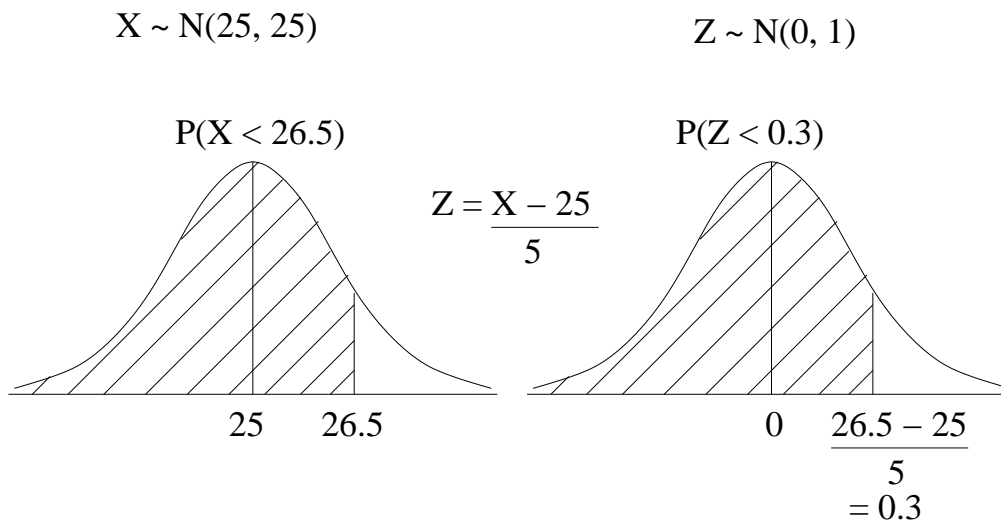
Example 9 A radioactive source emits particles at an average rate of 25 particles per second. What is the probability that in 1 second the count is less than 28 particles?

X = No. of particles emitted in 1 second $X \sim \text{Po}(25)$

So, we can use a $N(25, 25)$ as an approximate distribution.

Again, we need to make a continuity correction

So $P(X < 27)$ transforms to $P(X < 26.5)$



$$\begin{aligned}
 P(X < 26.5) &= P\left(\frac{X - 25}{5} < \frac{26.5 - 25}{5}\right) \\
 &= P(Z < 0.3) \quad \text{where } Z \sim N(0, 1) \\
 &= 0.6179
 \end{aligned}$$