

Unit 5.
The Normal Distribution

Topics	1. Introduction	2
	2. Definition of the Normal Distribution	4
	3. The Sample Average is Often Normally Distributed	
	Introduction to the Central Limit Theorem	7
	4. A Feel for the Normal Distribution	10
	5. The Relevance of the Normal Distribution	12
	6. Calculation of Probabilities for the Normal(0,1)	13
	7. From Normal(μ , σ^2) to Normal(0,1) – The Z-Score	19
	8. From Normal(0,1) to Normal(μ , σ^2)	22

1. Introduction

Much of statistical inference is based on the **normal** distribution.

- The pattern of occurrence of many phenomena in nature happens to be described well using a normal distribution model.
- Even when the phenomena in a sample distribution are not described well by the normal distribution, the sampling distribution of sample averages obtained by repeated sampling from the parent distribution is often described well by the normal distribution (*Central limit theory*).

You may have noticed in your professional work (especially in reading the literature for your field) that, often, researchers choose to report the **average** when he/she wishes to summarize the information in a sample of data.

The normal distribution is appropriate for continuous random variables only.

- Recall that, in theory, a continuous random variable can assume any of an infinite number of values.

Therefore, we'll have to refine our definition of a probability model to accommodate the continuous variable setting.

- $\Pr[X = x]$, the calculation of a point probability, is meaningless in the continuous variable setting. In its place, we calculate $\Pr[a < X < b]$, the probability of an **interval** of values of X .

- For the above reason, $\sum_{-\infty}^{\infty} \Pr[X = x]$ is also without meaning.

We can extend the ideas of a probability distribution for a discrete random variable to gain an understanding of the ideas underlying the meaning of a probability distribution for a continuous random variable. A bit of calculus (sorry!) helps us out.

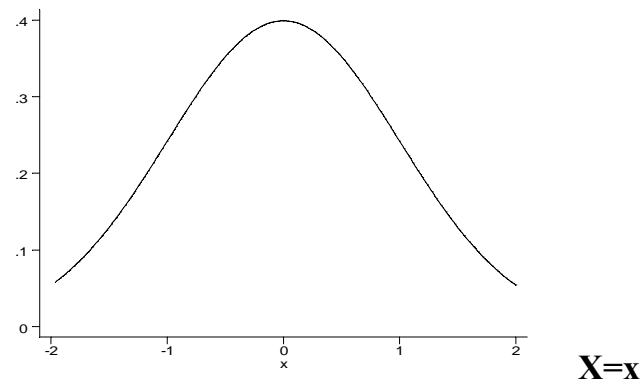
	Discrete Random Variable	Continuous Random Variable
1st: “List” of all possible values that exhaust all possibilities	E.g. – 1, 2, 3, 4, ..., N	“List” → range E.g. $-\infty$ to $+\infty$ 0 to $+\infty$
2nd: Accompanying probabilities of “each value”	$\Pr [X = x]$	“Point probability” → probability density Probability density of X , written $f_X(x)$
Total must be 1	$\sum_{x=\min}^{\max} \Pr[X = x] = 1$	“Unit total” → unit integral $\int_{-\infty}^{\infty} f_X(x) dx = 1$

2. Definition of the Normal Distribution

Definition of the normal probability distribution density function.

- The concept “probability of $X=x$ ” is replaced by the “probability density function $f_x()$ evaluated at $X=x$ ”
- A picture of this function with $X=x$ plotted on the horizontal and $f_x()$ evaluated at $X=x$ plotted on the vertical is the familiar bell shaped (“Gaussian”) curve

$f_x(x)$



$$f_x(X=x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \text{ where}$$

x = Value of X

Range of possible values of X : $-\infty$ to $+\infty$

Exp = e = Euler's constant = 2.71828 ... *note: $e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots$*

π = mathematical constant = 3.14 *note: π = (circumference/diameter) for any circle.*

μ = Expected value of X (“the long run average”)

σ^2 = Variance of X . *Recall – this is the expected value of $[X - \mu]^2$*

The Standard Normal Distribution is a particular normal distribution. It is the one for which $\mu=0$ and $\sigma^2=1$. It is an especially important tool.

- It is the one for which $\mu=0$ and $\sigma^2=1$.
- Tabulations of probabilities for this distribution are available.
- A random variable whose pattern of values is distributed standard normal has the special name: z-score, or normal deviate
- By convention, it is usually written as Z , rather than X .

$$f_Z(Z=z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]$$

Introduction to the Z-Score: A tool to compute probabilities of intervals of values for X distributed Normal(μ, σ^2).

- Suppose it is of interest to do a probability calculation for a random variable X that is distributed Normal(μ, σ^2)
- A bit of a glitch arises when tabulated normal probability calculations are available only for the Normal Distribution with $\mu = 0$ and $\sigma^2=1$. We solve our problem by exploiting an equivalence argument.
- “Standardization” expresses the desired calculation for X as an equivalent calculation for Z where Z is distributed standard normal, Normal(0,1).

$$\text{pr}\left[a \leq X \leq b\right] = \text{pr}\left[\left(\frac{a-\mu}{\sigma}\right) \leq Z \leq \left(\frac{b-\mu}{\sigma}\right)\right]. \text{ Thus,}$$

$$\text{Z-score} = \frac{X - \mu}{\sigma}$$

- **Note** - The technique of *standardization* of X involves *“centering”* (by subtraction of the mean of X which is μ) followed by *“rescaling”* (using the multiplier $1/\sigma$)

Sometimes, we might want to know the values of selected percentiles of a Normal(μ, σ^2) distribution. To do this, we work the standardization technique in the other direction.

For example, we might want to know the median of a normal distribution of gross income

- We have only percentile values tabulated for Z distributed Normal(0,1)
- The inverse of “Standardization” relates the percentile for X to that for Z.

$$X_{ptile} = \sigma [Z_{ptile}] + \mu$$

The z-score and its relatives the t-score, chi square and F statistics are central to the methods of confidence intervals and hypothesis testing.

3. The Sample Average is Often Normally Distributed Introduction to the Central Limit Theorem

Recall, our focus is on the behavior of the average, \bar{X}_n , of a sample. It is the Central Limit Theorem that gives us what we need.

The Central Limit Theorem

IF

- 1) We have an independent random sample of n observations $X_1 \dots X_n$; and
- 2) the $X_1 \dots X_n$ are all from the same distribution, *whatever that is.*; and
- 3) this distribution has mean = μ and variance = σ^2

THEN as $n \rightarrow \infty$

the sampling distribution of $\bar{X}_n = \left[\frac{\sum_{i=1}^n X_i}{n} \right]$ is eventually

Normal with mean = μ and variance = σ^2/n

In words:

“In the long run, averages have distributions that are well approximated by the Normal”

“The sampling distribution of \bar{X}_n , upon repeated sampling, is eventually Normal $\left(\mu, \frac{\sigma^2}{n} \right)$ ”

Later (Section 7) we'll learn how to compute probabilities of intervals of values for \bar{X}_n distributed Normal($\mu, \sigma^2/n$) by using the z-score technique.

$$\text{pr} \left[a \leq \bar{X} \leq b \right] = \text{pr} \left[\left(\frac{a - \mu}{\sigma / \sqrt{n}} \right) \leq Z \leq \left(\frac{b - \mu}{\sigma / \sqrt{n}} \right) \right]. \text{ Thus,}$$

$$\text{Z-score} = \frac{\bar{X} - E(\bar{X})}{\text{se}(\bar{X})} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

A variety of wordings of the central limit theorem give a feel for its significance!

1. " according to a certain theorem in mathematical statistics called the central limit theorem, the probability distribution of the sum of observations from any population corresponds more and more to that of a normal distribution as the number of observations increases; ie - if the sample size is large enough, the sum of observations from any distribution is approximately normally distributed. Since many of the test statistics and estimating functions which are used in advanced statistical methods can be represented as just such a sum, it follows that their approximate normal distributions can be used to calculate probabilities when nothing more exact is possible."

Matthews DE and Farewell VT. Using and Understanding Medical Statistics, 2nd, revised edition. New York: Karger, 1988. page 93.

2. "With measurement data, many investigations have as their purpose the estimation of averages - the average life of a battery, the average income of plumbers, and so on. Even if the distribution in the original population is far from normal, the distribution of sample averages tends to become normal, under a wide variety of conditions, as the size of the sample increases. This is perhaps the single most important reason for the use of the normal".

Snedecor GW and Cochran WG. Statistical Methods, sixth edition. Ames: The Iowa State University Press, 1967. page 35.

3. "If a random sample of n observations is drawn from some population of any shape, where the mean is a number μ and the standard deviation is a number σ , then the theoretical sampling distribution of \bar{X}_n , the mean of the random sample, is (nearly) a normal distribution with a mean of μ and a standard deviation of σ/\sqrt{n} if n , the sample size, is 'large'".

Moses LE. Think and Explain with Statistics. Reading: Addison-Wesley Publishing Company, 1986. page 91.

4. "It should be emphasized that the theorem applies almost regardless of the nature of the parent population, that is, almost regardless of the distribution from which X_1, \dots, X_n are a random sample. ... How large n must be to have a "good" approximation does depend, however, upon the shape of the parent population."

Anderson TW and Sclove SL. Introductory Statistical Analysis. Boston: Houghton Mifflin Company, 1974. page 295.

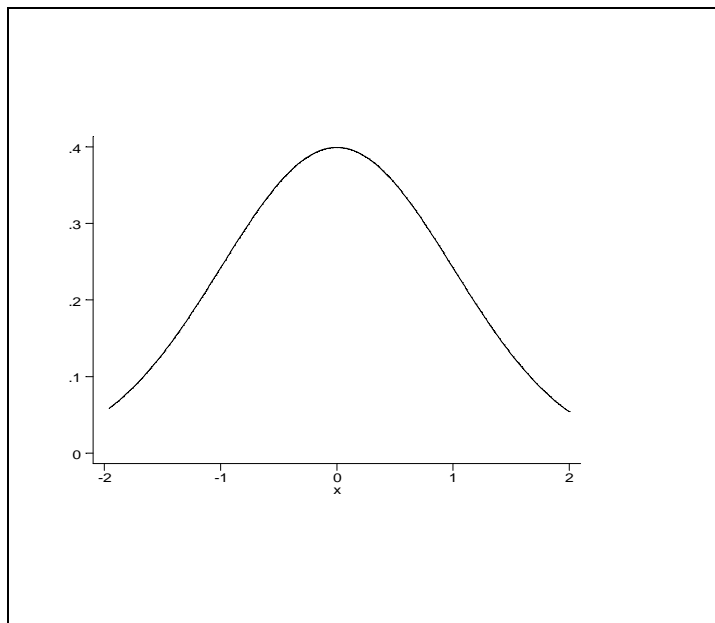
4. A Feel for the Normal Distribution

What does the normal distribution look like?

- (1) a smooth curve defined everywhere on the real axis that is
- (2) bell shaped and
- (3) symmetric about the mean value.

note – Because of symmetry, we know that the mean = median.

**General
shape of a
Normal
Distribution**



**This normal distribution
has mean and median =
0**

Some Features of the Normal Distribution:

- (1) Look again at the definition of the normal probability density function on page 4. Notice that it includes only two population parameters, the mean μ and variance σ^2 . Notice that there are no other population parameters present. This allows us to say that the normal probability density function is completely specified by the mean and variance. This feature is very useful in the calculation of event probabilities which will be described later.

(2) **The mean μ tells you about location -**

Increase μ - Location shifts right
 Decrease μ - Location shifts left
 Shape is unchanged

(3) **The variance σ^2 tells you about narrowness or flatness of the bell -**

Increase σ^2 - Bell flattens. Extreme values are more likely
 Decrease σ^2 - Bell narrows. Extreme values are less likely
 Location is unchanged

(4) **Very Useful Tool for Research Data:** If you are exploring some data, let's say it is a sample of data X that is distributed normal with mean μ and variance σ^2 , then *roughly*

- (i) 68% of the distribution of X lies in an interval of $\pm (1)(\sigma)$ about its mean value μ .
- (ii) 95% of the distribution of X lies in an interval of $\pm (1.96)(\sigma)$ about its mean value μ .
- (iii) 99% of the distribution of X lies in an interval of $\pm (2.576)(\sigma)$ about its mean value μ .

(5) **Most often, this “seat of the pants” rule is applied to the distribution of the sample mean \bar{X} . *Roughly ...***

- (i) 68% of the distribution of \bar{X} lies in an interval of $\pm (1)(\sigma/\sqrt{n})$ about its mean value μ .
- (ii) 95% of the distribution of \bar{X} lies in an interval of $\pm (1.96)(\sigma/\sqrt{n})$ about its mean value μ .
- (iii) 99% of the distribution of \bar{X} lies in an interval of $\pm (2.576)(\sigma/\sqrt{n})$ about its mean value μ .

Reminder: Look again at the quantity (σ/\sqrt{n}) . Recall that standard error $[\bar{X}] = SE[\bar{X}] = (\sigma/\sqrt{n})$

5. Relevance of the Normal Distribution

What Data Follow the Normal Distribution?

There are two kinds of data that follow a normal probability distribution.

First Type – Nature gives us this. Nature includes many continuous phenomena yielding sample data for which the normal probability model is a good description. For example,

- Heights of men
- Weights of women
- Systolic blood pressure of children
- Blood cholesterol in adults aged 20 to 100 years

Second Type – Repeated sampling and the Central Limit Theorem gives this. If we repeat our research study over and over again so as to produce the sampling distribution of the sample mean \bar{X} , this distribution is well described by a normal distribution model by virtue of the Central Limit Theorem.

This second class is particularly useful in research since, often, the focus of interest is in the behavior (reproducibility and variability) of sample means rather than individual values.

- Average response among persons randomized to treatment in a clinical trial

6. Calculation of Probabilities for the Normal (0,1)

With respect to studies of any normal distribution, it eventually boils down to knowing how to work with one particular distribution, the Normal(0,1), also called the **standard** normal or standard gaussian distribution.

A random variable Z is said to follow the standard normal distribution if it is distributed normal with mean=0 and variance=1. Recall again the probability density function for this distribution:

$$f_Z(Z=z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right]$$

Recall also

- For a continuous random variable, we cannot compute point probabilities such as Probability [$Z=z$]
- What we calculate, instead, are probabilities of intervals of values, such as Probability [$a \leq Z \leq b$] for some choice of “a” and “b”

A Probability of an Interval such as Probability [$Z \leq z$] is called a **cumulative probability**

- Tables for the Normal(0,1) distribution typically provide values of cumulative probabilities of this form. Sometimes, more is provided.

Fortunately, we don't actually have to do these calculations

- Such calculations are exercises in calculus and involve the integral of the probability density function.
- Values of Probability [$a \leq Z \leq b$] can be found by utilizing statistical tables for the Normal(0,1).
- They can also be gotten from the computer (either software or web).

Some Useful Tips for Using the Statistical Tables for the Normal(0,1)

Note – Many of the calculators for the normal distribution on the internet provide direct calculations of all of the probabilities described in these tips so that you do not have to do the translations yourself.

(Tip 1) Symmetry of the Normal(0,1) distribution about its mean value 0 has useful implications with respect to the calculation of probabilities. Recalling the notation "Z" to denote the random variable Z and "z" to denote a possible actual value, symmetry of the standard normal distribution about $\mu=0$ means:

$$\Pr [Z \leq -z] = \Pr [Z \geq +z]$$

(Tip 2) Because the integral of a continuous probability distribution (you can think of this as the sum of the probabilities associated with all possible outcomes when the distribution is discrete) must total exactly one,

$$\Pr [Z < z] + \Pr [Z \geq +z] = 1$$

$$\Pr [Z < z] = 1 - \Pr [Z \geq z]$$

$$\Pr [Z \geq z] = 1 - \Pr [Z < z]$$

(Tip 3) The facts of symmetry of the Normal(0,1) distribution about $\mu=0$ and its integral equaling one gives us

$$\Pr [Z \geq 0] = .5$$

$$\Pr [Z \leq 0] = .5$$

(Tip 4) Some tables for the Normal distribution provide values **ONLY** of the type $\Pr [Z < z]$.

If you want to calculate ...	Then use the table this way ...
$\Pr(Z > a)$	$\Pr(Z > a) = 1 - \Pr(Z \leq a)$
$\Pr(Z < -a)$	$\Pr(Z < -a) = 1 - \Pr(Z < +a)$
$\Pr(a < Z < b)$	$\Pr(a < Z < b) = \Pr(Z < b) - \Pr(Z < a)$

(Tip 5) Inequalities $\Pr(Z \leq z)$ versus $\Pr(Z < z)$ or $\Pr(Z \geq z)$ versus $\Pr(Z > z)$ can seem confusing at first. The key is to know that because Z is continuous when it is distributed $\text{Normal}(0,1)$, point probabilities are meaningless. That means

$$\Pr [Z \geq z] = \Pr [Z > z],$$

because Z is continuous

$$\Pr [Z \leq z] = \Pr [Z < z],$$

because Z is continuous

Beware – Be careful not to make this assumption when you are working with discrete variables!!

Notation

With apology, it is nevertheless useful to know the *notation* involved in the calculation of probabilities. Here it is for the setting of probability calculations for the standard normal, $\text{Normal}(0,1)$.

- $F_Z(z)$ is called the cumulative probability density function. It is the integral of the probability density function $f_Z(z)$ that was introduced on page 4.
- Often, the letter Z is used to refer to a random variable distributed $\text{Normal}(0,1)$
- $\text{Prob}[\text{Normal}(0,1) \text{ variable} < z] = \text{Prob}[Z \leq z] = F_Z(z)$

Example

If Z is distributed standard normal, what is the probability that Z is at most 1.82?

Solution = 0.9656

Calculator used: http://davidmlane.com/hyperstat/z_table.html

Step 1:

Translate the "words" into an event.

"Z is at most 1.82" is equivalent to the event ($Z \leq 1.82$).

The required probability is therefore $\Pr(Z \leq 1.82) = ?$

Step 2:

The calculator used (noted above) has two calculators, actually – one for the calculation of probabilities and one for the determination of percentiles.

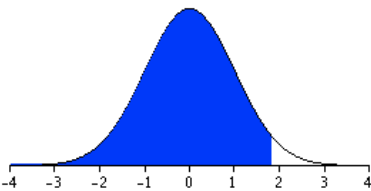
In this example, the first calculator is used: **I clicked on the button for "below", typed in the value and pressed the return key on my keyboard. The calculator returned the required probability as "shaded area" 0.965620 at the bottom of the screen.**

Normal Distribution

This page contains calculators for the normal distribution. The first calculator allows to enter the mean, standard deviation, and cutoff points; the area under the normal distribution is computed. The second calculator does just the reverse. These calculators can be used instead of a table of the normal distribution.

Choose parameters and the area will be computed.

Normal Distribution



Mean Sd

Above
 Below
 Between and
 Outside or

Shaded area: 0.965620

$$\Pr(Z \leq 1.82) = 0.9656$$

Example

What is the probability that a standard normal random variable exceeds the value 2.38?

Solution = 0.0087

Calculator used: http://davidmlane.com/hyperstat/z_table.html

step 1:

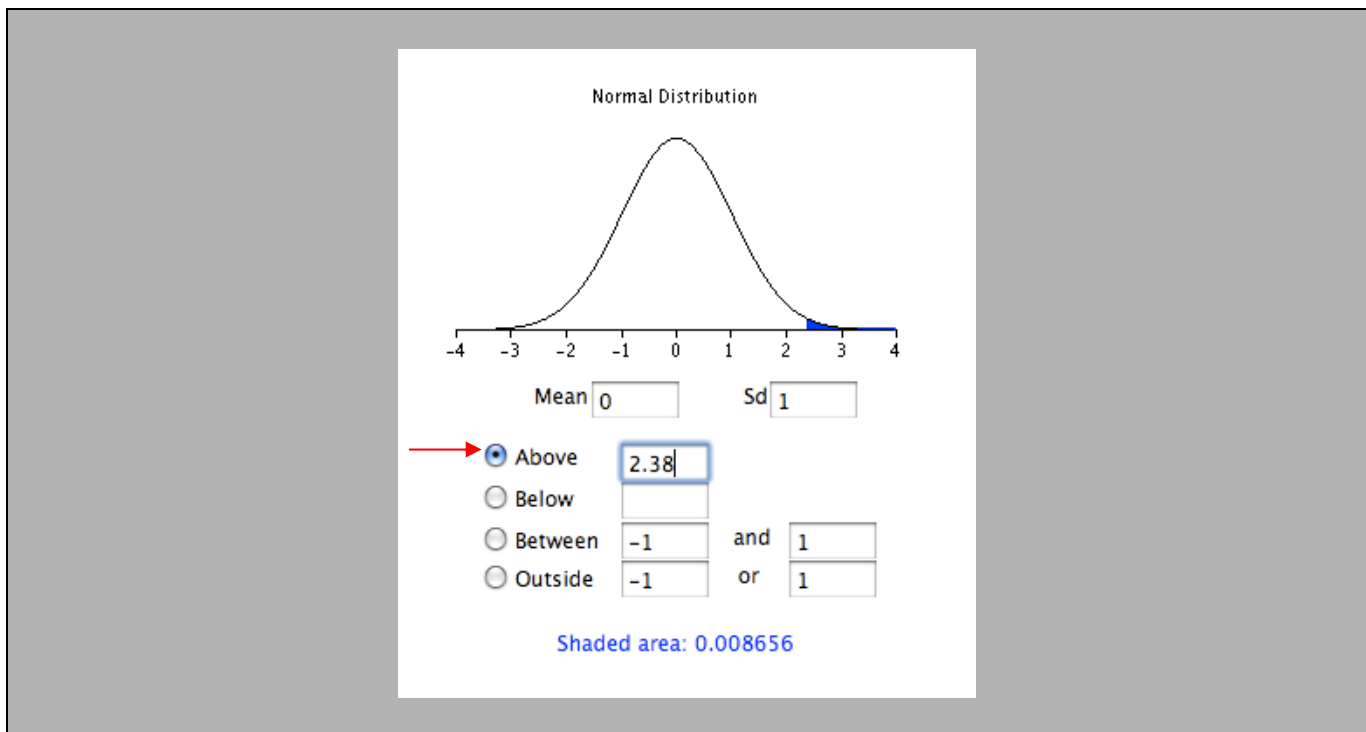
Translate the "words" into an event.

"Z exceeds 2.38" is equivalent to the event ($Z > 2.38$).

The required probability is therefore $\Pr(Z > 2.38) = ?$

Step 2:

The same calculator is used: This time, I clicked on the button for "above", and typed in the value and pressed the return key on my keyboard. The calculator returned the required probability as "shaded area" 0.008656 at the bottom of the screen.



$$\Pr(Z > 2.38) = 0.0087$$

Example

How likely is it that a standard normal random variable will assume a value in the interval $[-2.58, +0.58]$?

Solution = 0.714

Calculator used: http://davidmlane.com/hyperstat/z_table.html

step 1:

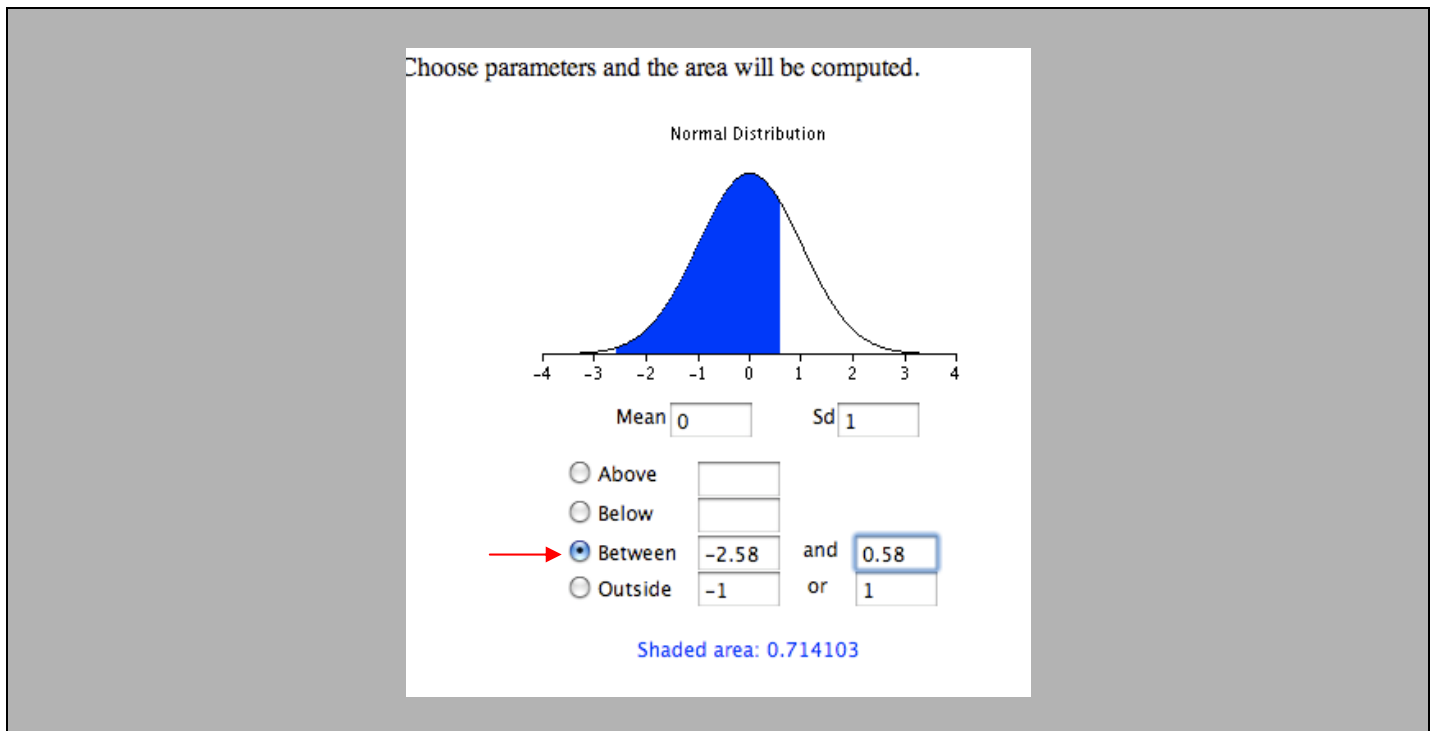
Translate the "words" into an event.

Z assuming a value in the interval $[-2.58, +0.58]$

is equivalent to $\Pr[-2.58 \leq Z \leq +0.58] = ?$

Step 2:

The same calculator is used



$$\Pr[-2.58 \leq Z \leq +0.58] = 0.714103$$

7. From Normal (μ, σ^2) to the Standardized Normal (0,1) – The Z Score

Seen already are two z-score transformations. From these, a useful generalization is also seen.

1.	If X is distributed Normal (μ, σ^2)	z-score = $\frac{X - \mu}{\sigma}$ is distributed Normal(0,1)
2.	If \bar{X} is distributed Normal ($\mu, \sigma^2/n$)	z-score = $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is distributed Normal(0,1)
3.	If a “generic” random variable Y is distributed Normal with $\mu_Y = E(Y)$ $\sigma_Y^2 = \text{Var}(Y)$	z-score = $\frac{Y - E(Y)}{\sqrt{\text{var}(Y)}}$ is distributed Normal(0,1)

Note – To appreciate the third row, notice that in #1, the choice is $Y=X$. In #2, the choice is $Y=\bar{X}$

It is the “z-score” transformation that allows us to obtain interval probabilities for ANY Normal Distribution.

Example

The Massachusetts State Lottery averages, on a weekly basis, a profit of 10.0 million dollars. The variability, as measured by the variance statistic is 6.25 million dollars squared. If it is known that the weekly profits is distributed normal, what are the chances that, in a given week, the profits will be between 8 and 10.5 million dollars?

Solution = .367

Calculator used: http://davidmlane.com/hyperstat/z_table.html

step 1:

Translate the "words" into an event.

Since we're no longer dealing with a standard normal random variable, it is convenient to use X to denote the random variable defined as the profit earned in a given week.

X assuming a value in the interval $[8,10.5]$ is equivalent to $\Pr[8 \leq X \leq 10.5] = ?$

step 2:

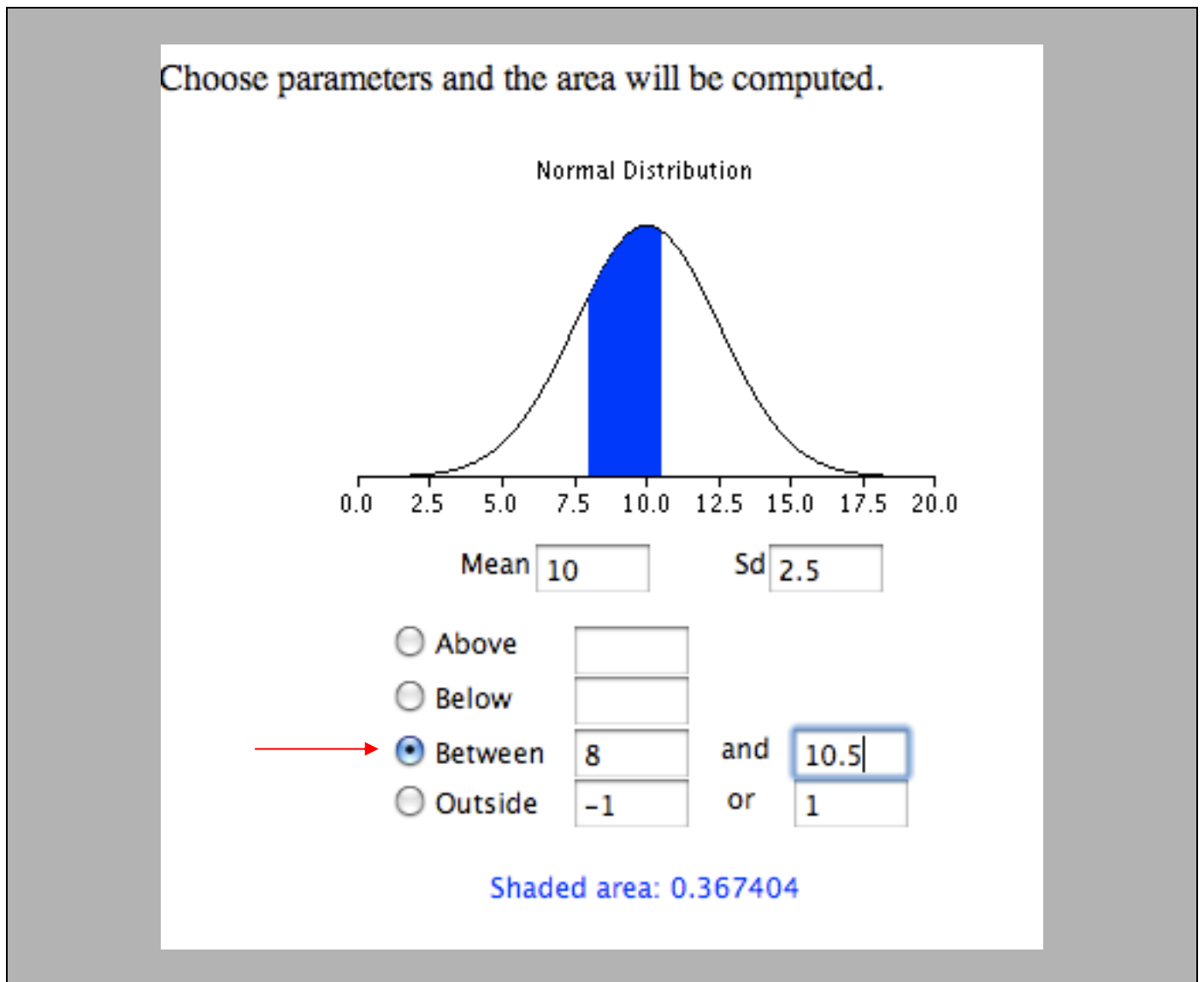
Before using the transformation formula, solve for the $\sqrt{\text{variance}}$ of the normal distribution of X . To see that this is correct, look at the 3rd row of the chart on the previous page (page 19). This is the standard deviation of the normal distribution of X .

If $\sigma^2 = 6.25$

Then $\sigma = \sqrt{6.25} = 2.5$

Step 3:

The same calculator is now used as follows:



$$\Pr[8 \leq X \leq 10.5] = 0.367404$$

8. From Normal (0,1) to Normal (μ, σ^2)

Sometimes we will want to work the z-score transformation backwards

- We might want to know values of selected **percentiles** of a Normal distribution (e.g. median cholesterol value)
- Knowledge of how to work this transformation backwards will be useful in **confidence interval** construction

1.	If Z is distributed Normal(0,1)	Then $X = \sigma Z + \mu$ distributed Normal(μ, σ^2)
2.	If Z is distributed Normal(0,1)	Then $\bar{X} = \left(\frac{\sigma}{\sqrt{n}}\right)Z + \mu$ is Normal ($\mu, \sigma^2/n$)
3.	If Z is distributed Normal(0,1)	<p>generic $Y = \left(\sqrt{\text{var}(Y)}\right)Z + E(Y)$ is Normal</p> <p style="text-align: center;"> $\mu_Y = E(Y)$ $\sigma_Y^2 = \text{Var}(Y)$ </p>

Example

Suppose it is known that survival time following a diagnosis of mesothelioma is normally distributed with $\mu=2.3$ years and variance $\sigma^2=7.2$ years squared. What is the 75th percentile, the elapsed time during which 75% of such cases are expected to die?

Solution = 4.11 years

Calculator used: http://davidmlane.com/hyperstat/z_table.html

step 1:

Translate the "words" into an event.

“What is the 75th percentile” tells us that we are seeking the value of X corresponding to the separation of the lower 75% of the probability distribution from the upper 25% of the distribution. This corresponds to a left tail area equal to 0.75.

step 2:

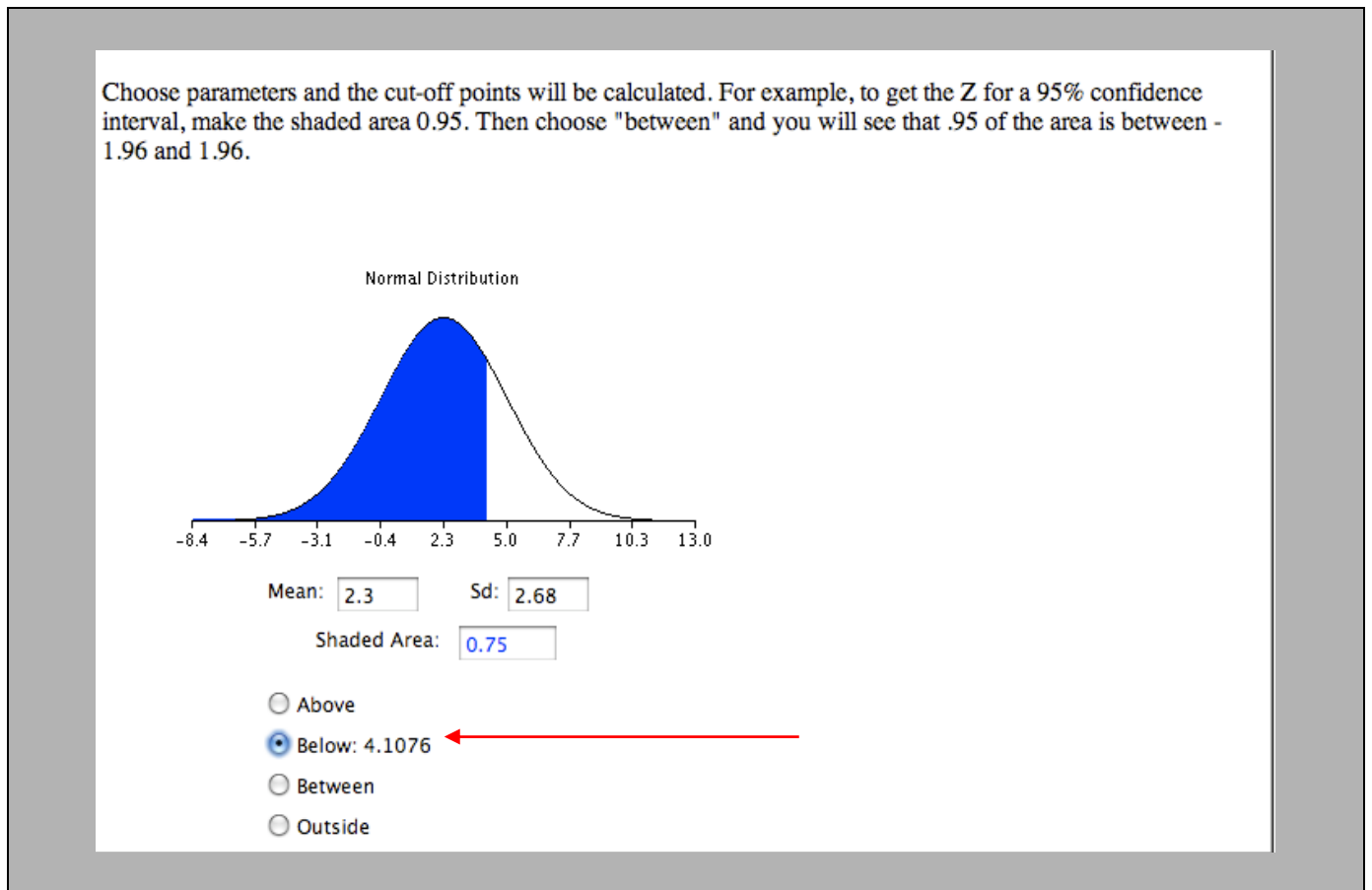
Because the calculator wants the standard deviation, , solve for the $\sqrt{\text{variance}}$ of the normal distribution of X.

If $\sigma^2 = 7.2$

Then $\sigma = \sqrt{7.2} = 2.68$

Step 3:

This time, unlike the previous examples, we will use the second calculator that is provided at this site. You might need to scroll down to get to it. In this example: I typed in the values for the mean and sd (note – this url could care less whether this box is a standard deviation or a standard error. The key is to type in the square root of the variance) I then typed in the value 0.75 in the box next to ‘shaded area’. Finally, I clicked on the button next to ‘below’ The calculator returned the required percentile, not at the bottom of the screen. Instead, you find the required percentile to the right of the radio button for ‘below’



Thus, it is expected that 75% of persons newly diagnosed with mesothelioma will have died within 4.11 years. This is the same as saying that there is an expected 25% chance of surviving beyond 4.11 years.