Unit 5.
The Normal Distribution

"The means show a 'normal' distribution"

- Bradford Hill

The Amherst Regional High School measures the blood pressure of its very large number of students – over a thousand! A histogram or frequency polygon summary reveals the distribution to be ***bell shaped***. The ***normal*** distribution (also called the ***Gaussian*** distribution) is the familiar bell shaped probability distribution model. It is the most important probability distribution model for continuous data and for very good reasons!

(1) The distributions of many phenomena in nature are well described by normal probability distribution models;

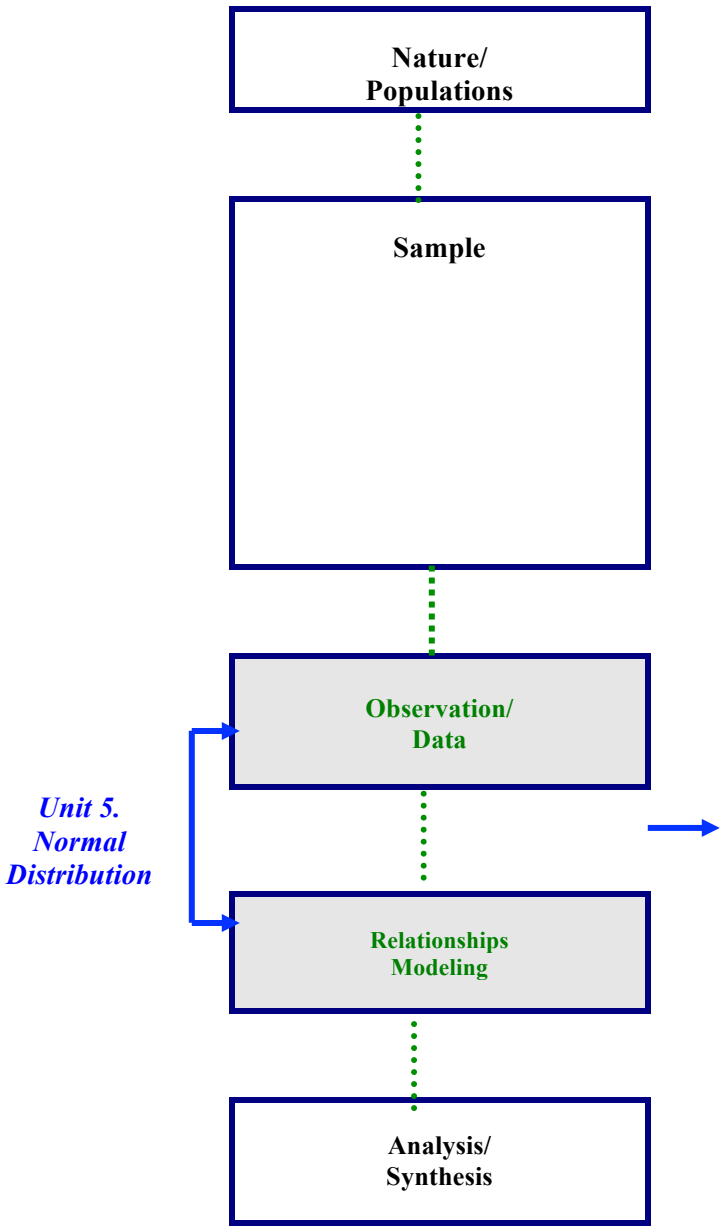
(2) Other distributions in nature are reasonably well described by normal probability distribution models;

(3) The distributions of some transformations of random variables are well described by a normal probability distribution model; and

(4) the sampling distribution of the sample mean tends to normality even when the population distribution in nature is non-normal.

Table of Contents

1. Unit Roadmap



This unit focuses on quantitative data that are continuous. Recall: A characteristic of a continuous random variable is that, between any two valid values, all of the intermediate values are also possible. Examples are weight, height, cholesterol, etc.

Recall, also, the "frequentist" definition of probability: the expected proportion of times that an event will occur as the number of trials tends to infinity. **The normal probability distribution model can be obtained by applying this reasoning to a Binomial distribution model of chance.** De Moivre did this for us, way back in 1720.

We will be using the normal distribution to model: (1) the population distribution of phenomena in nature; and (2) the sampling distributions of - the sample mean, sums and differences of sample means, and regression coefficients.

2. Learning Objectives

When you have finished this unit, you should be able to:

- Explain the distinction between a discrete versus a continuous probability distribution model.
- Define the Normal probability distribution model.
- Define the Standard Normal probability distribution model.
- Explain the relevence of the normal probability distribution .
- Calculate normal probability distribution probabilities.
- Define and explain the utility of Z-scores. Explain standardization.
- Obtain values of percentiles of the Standard Normal probability distribution..
- Obtain values of percentiles of normal probability distributions that do not have zero mean and unit variance.

3. Looking Ahead: The 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 $\overline{X}$, 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

4. Introduction

Much of statistical inference is based on the normal distribution.

- The patterns of occurrence of many phenomena in nature happen 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*). **More on this in Section 11, page ##.**

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 summarizing 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.

Recall. In Unit 4, we developed the definition of a discrete probability distribution. Much of that intuition carries over to the definition of a continuous probability distribution. However, some extensions of our thinking are required.

- $\Pr[X = x]$, the calculation of a point probability, is meaningless in the continuous variable setting. **This makes sense when you consider that a continuous random variable has infinitely many possibilities; how could you calculate infinitely many Pr[X=x], each having a positive value, and have their total be 1 or 100?** In its place, we calculate

  Pr [a < X < b], the probability of an interval of values of X.
  We may not be able to calculate point probabilities, but we can answer questions such as: "what is the probability that X has value between 'a' and 'b' "?

- 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 Probability Distribution (Unit 4)	**Continuous Probability Distribution (Unit 5)**
Example -	**DISCRETE** **X is the outcome of ONE spin of a roulette wheel.** **1st:** **The possible outcomes of X are finite and can be listed: 0, 1, 2, …9** **2nd:** **The probability of each outcome is 1/10 and the 10 of them add up to 1**	**CONTINUOUS** **X is the score on an IQ test.** **1st:** **The possible outcomes of X are infinitely many. We cannot list them, but we can give the range, roughly: 50 to 200** **2nd:** **We can't calculate the probability of an individual IQ score because there are infinitely many possibilities. But under the assumption that IQ scores are distributed normal, we can calculate probabilities such as: Pr [110 < X < 120] which answers the question: What is the probability that a randomly selected person has an IQ between 110 and 120?**
1st: "List" of all possible values that exhaust all possibilities	**E.g. –** **1, 2, 3, 4, …, N**	**"List" → range** **E.g. -∞ to +∞** **0 to +∞**
2nd: Accompanying probabilities of "each value"	**Pr [X = x]**	**"Point probability" → probability density** **Probability density of X, written $f_X(x)$**
Taken over all possibilities, the sum of the associated probabilities must be 100% or 1.	$\sum_{x=\min}^{\max} \Pr[X = x] = 1$	**"Unit total" → unit integral** $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

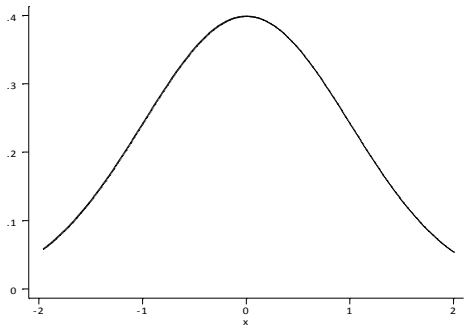
5. Definition of the Normal Distribution

Definition of the normal probability distribution density function.

- The concept "probability of X=x" (which we have been writing as Pr[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:

$f_X(x)$



X=x

Normal Distribution (μ, σ^2)

A random variable X that is distributed normal with mean=μ and variance=σ^2 has probability density function

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

x = Value of X

Range of possible values of X: -∞ to +∞

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

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, identified by the choice of values for μ and σ^2. It is an especially important tool.

- **The Standard Normal is the Normal Distribution for which $\mu=0$ and $\sigma^2=1$.**
- Tabulations of probabilities for this distribution are available. **Note – Actually, nowadays, the internet offers calculators for any normal distribution you like!**
- The Standard Normal random variable has some special names z-score, or normal deviate
- By convention, the Standard Normal random variable is usually written as Z, rather than X.

Standard Normal Distribution ($\mu=0$, $\sigma^2=1$)
Z-scores are distributed Normal(0,1)

A random variable Z that is distributed standard normal has probability density function

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

Introduction to the Z-Score: A tool to compute probabilities of intervals of values for X distributed Normal(μ,σ^2). **This is what the online calculators are doing for you, behind the scenes…**

- Suppose it is of interest to do a probability calculation for a random variable X that is distributed Normal(μ, σ^2)

Example – Suppose IQ scores are distributed Normal($\mu=100$, $\sigma^2 = 15^2$)
What is the probability of an IQ score between 110 and 120? That is,
Pr [110 < X < 120] = ??

- A bit of a glitch arises when normal probability calculations are available only for the Normal Distribution with $\mu = 0$ and $\sigma^2=1$. No worries. We can **translate** the desired calculation into an equivalent one for a Z-score that is distributed Normal with **$\mu=0$ and $\sigma^2=1$**. What we do is called **"standardization."**

- "Standardization" expresses the desired calculation for X distributed Normal(μ, σ^2) as an equivalent calculation for a Z-score distributed standard normal, Normal(0,1).

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

Example, continued –
Pr [110 < X < 120] = ?? for the IQ distribution here says:
a = 110, b = 120, $\mu=100$, and $\sigma = 15$. →

$$\Pr[110 < X < 120] = \Pr\left[\left(\frac{110-100}{15}\right) < \left(\frac{X-\mu}{\sigma}\right) < \left(\frac{120-100}{15}\right)\right] = \Pr[0.67 < \text{Z-score} < 1.33]$$

- More on calculations of Normal distribution probabilities in Section 8, page 18.

Thus,

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

- ***So you know ….*** 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/σ)

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.

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

Example – Consider again IQ scores that are distributed Normal(μ=100, $\sigma^2 = 15^2$)
What is the 80th percentile of this distribution?
Using the David Lane calculator introduced in Section 7, we obtain
$Z_{ptile} = Z_{.80} = 0.841$
Setting μ=100 and σ = 15 then yields
$X_{.80} = \sigma\,[Z_{.80}] + \mu = (15)\,[0.841] + 100 = 112.615$

- Calculations of Normal distribution percentiles are explained in Section 9, page ##

Putting it all together: Thus, we can go back and forth:

From: X distributed Normal(μ,σ^2) **To: Z-score distributed Normal(0,1)**	**From: Z-score distributed Normal(0,1)** **To: X distributed Normal(μ,σ^2)**
Use this when you want to calculate probabilities of event occurrence *(areas under the curve)*	**Use this when you want to obtain the values of percentiles of interest, eg; median, P25, etc.**
$\textbf{Z-score} = \frac{X - \mu}{\sigma}$	$X_{ptile} = \sigma\left[Z\text{-score}_{ptile}\right] + \mu$

Looking ahead to Units 6 and 7: 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.

6. A Feel for the Normal Distribution

What does the normal distribution look like?

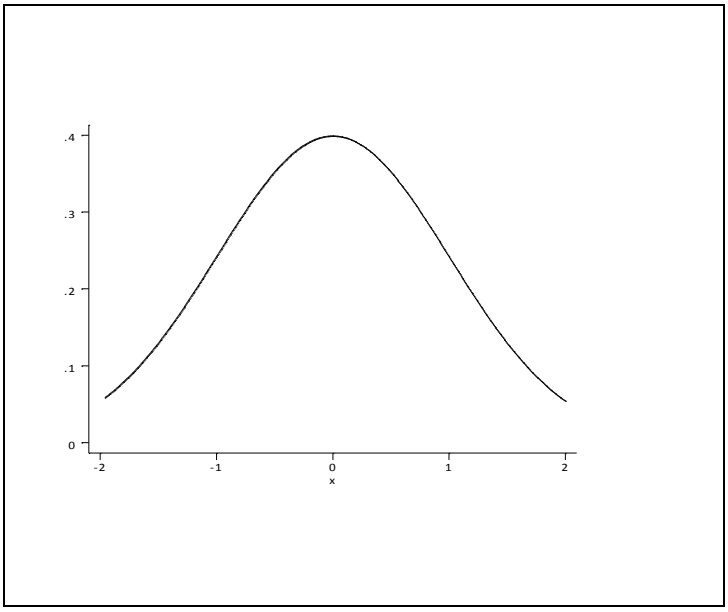
(1) a smooth curve defined everywhere on the real axis that is

(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:

Look again at the definition of the normal probability density function on page 8. Notice:

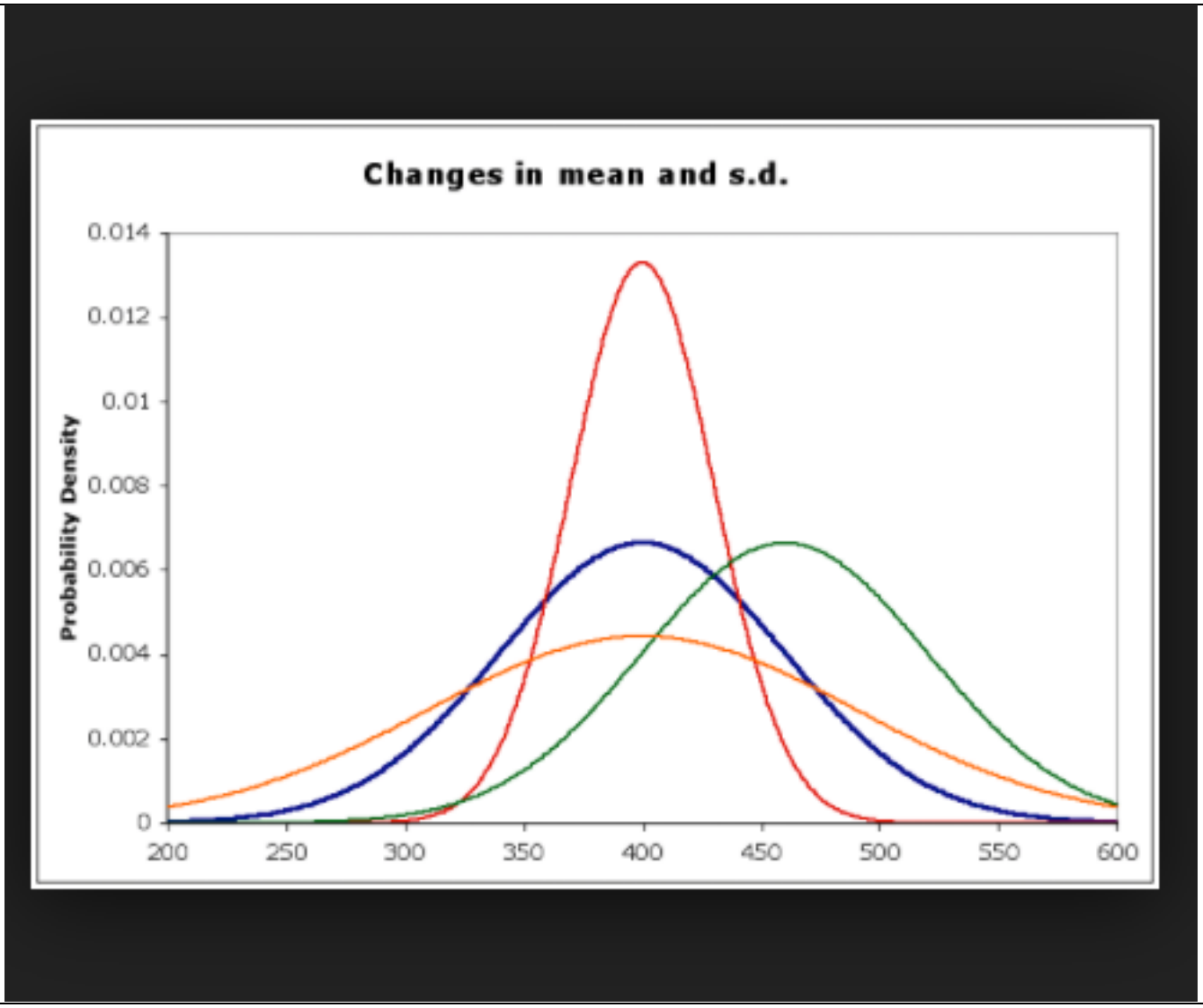
- The definition includes only two population parameters, the mean μ and variance σ^2
- There are no other population parameters present.
- This allows us to say that the normal probability density function is completely specified by the mean μ and variance σ^2

The mean μ tells you about location -

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

The variance σ^2 tells you about narrowness or flatness of the bell -

Increase σ^2 - Bell flattens. Values far away from the mean are more likely
Decrease σ^2 - Bell narrows. Values far away from the mean are less likely
Location is unchanged



Source: ww2.tnstate.edu/ganter

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.

Aside: A quick reminder about the sampling distribution of the average.

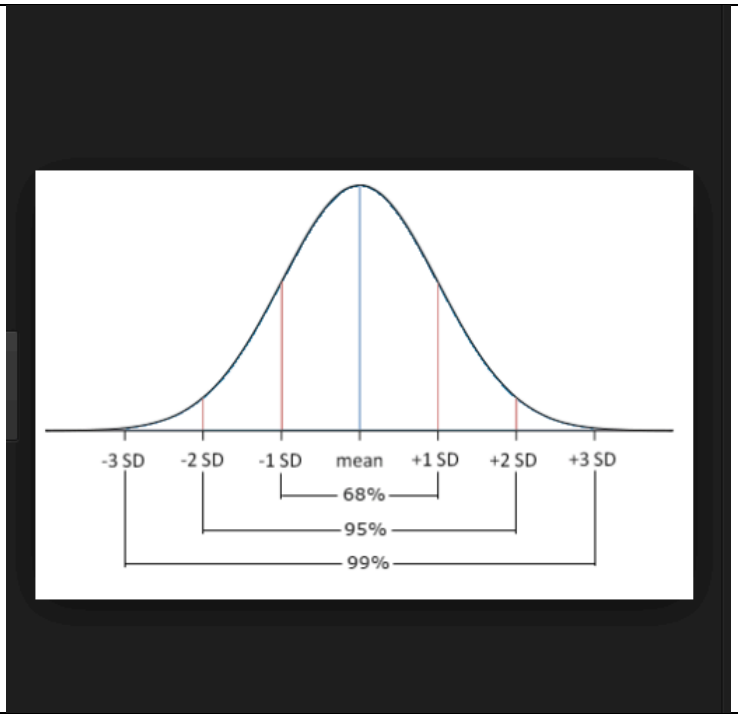
standard error $[\bar{X}] = \sqrt{\text{variance}[\bar{X}]} = \text{SE}[\bar{X}] = (s/\sqrt{n})$

Most often, this 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 $\pm (1)\text{SE} = \pm (1)(\sigma/\sqrt{n})$ about its mean value μ.

(ii) 95% of the distribution of $\bar{X}$ lies in an interval of $\pm (1.96)\text{SE} = \pm (1.96)(\sigma/\sqrt{n})$ about its mean value μ.

(iii) 99% of the distribution of $\bar{X}$ lies in an interval of $\pm (2.576)\text{SE} = \pm (2.576)(\sigma/\sqrt{n})$ about its mean value μ.



Source: www.totallab.com

7. 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) for a Z-score, 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:

Standard Normal Distribution ($\mu=0$, $\sigma^2=1$)

A random variable Z that is distributed standard normal has probability density function

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

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).

Notation

It is helpful to review and be comfortable with the ***notation*** involved in the calculation of probabilities. Notation for probability calculations for a random variable Z distributed standard normal, 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 8.
- Often, the notation Z-score, or the letter Z, is used to refer to a random variable distributed Normal(0,1)
- Prob[Normal(0,1) variable $\leq$ z] = Prob[$Z \leq z$]=$F_Z(z)$

Example 1 -
If Z is distributed standard normal, what is the probability that Z is at most 1.82?

Solution: Pr [$Z \leq 1.82$] = 0.9656

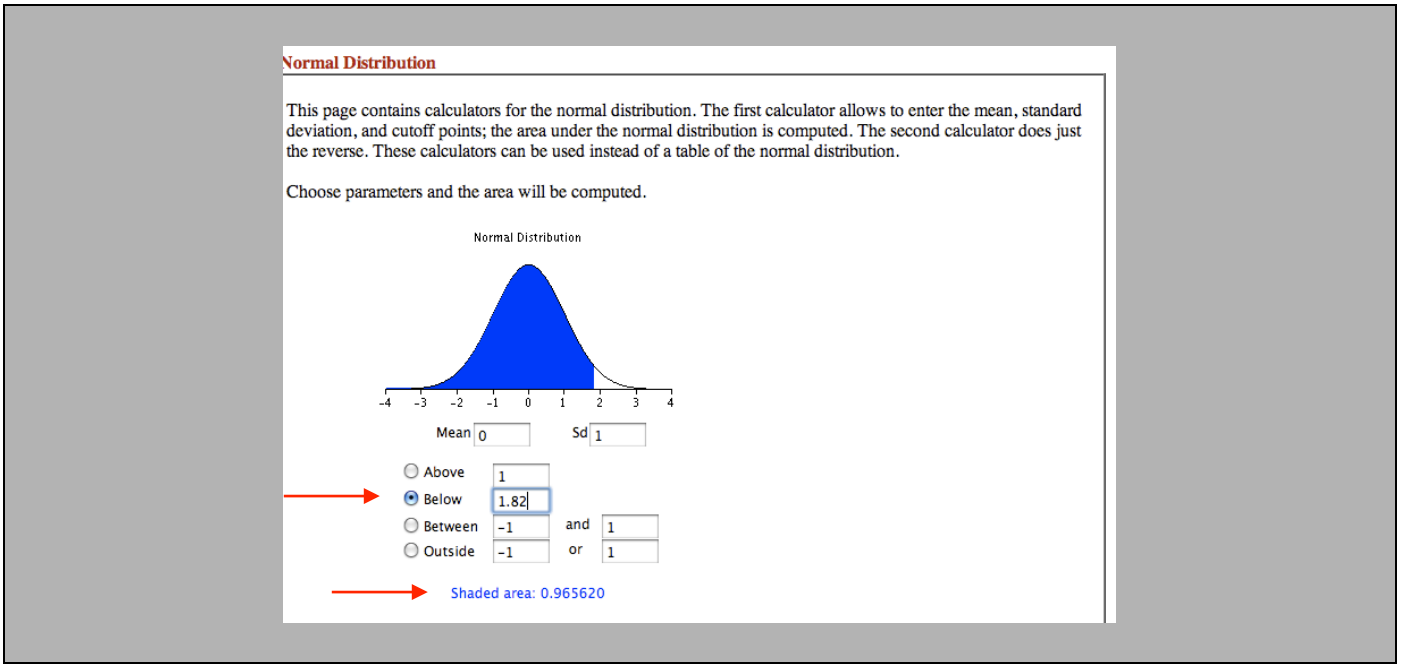
Calculator used: http://davidmlane.com/hyperstat/z_table.html

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 "David Lane" 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.**

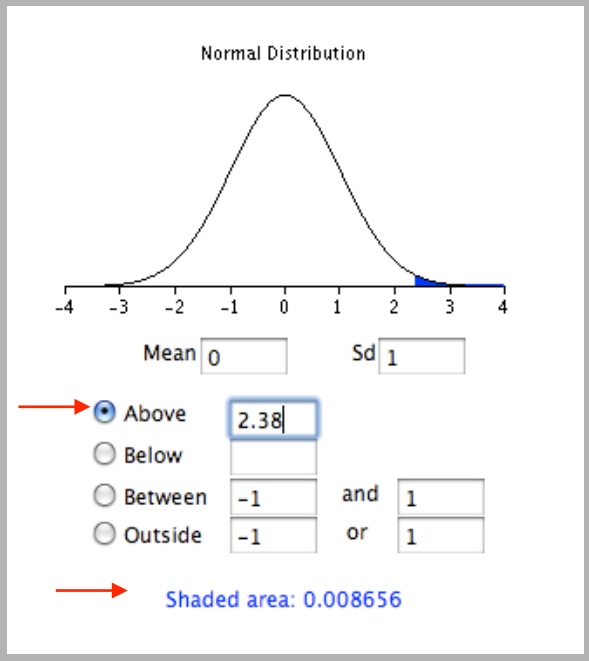


Example 2-
What is the probability that a standard normal random variable exceeds the value 2.38?

Solution = Pr [Z > 2.38] = 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.**

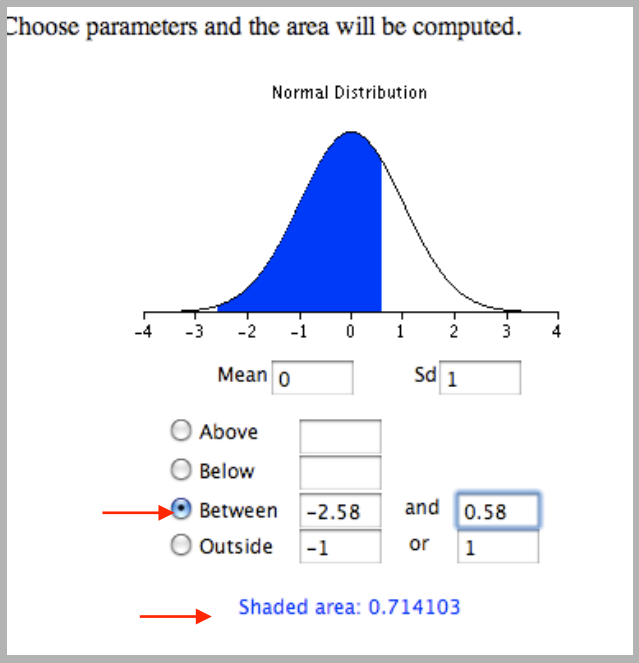


Example 3 -
How likely is it that a standard normal random variable will assume a value in the interval [-2.58, +0.58]?

Solution = Pr[-2.58 <= Z <= +0.58] = 0.714103
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 <= Z <= +0.58] = ?

Step 2:
The same calculator is used



8. Calculation of Probabilities for a Normal (μ, σ^2): From Normal (μ, σ^2) to Normal (0,1): Standardization and the Z Score

Three very useful "standardizing" transformations

In each of the transformations below, application of the formula is **standardization** to what is called a Z-score. Each Z-score is distributed Standard Normal; that is Normal(0,1).

1.	**If X is distributed Normal (μ, σ^2)**	$\text{z-score} = \frac{X-\mu}{\sigma}$ **is distributed Normal(0,1)**
2.	**If $\bar{X}$ is distributed Normal ($\mu, \sigma^2/n$)**	$\text{z-score} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ **is distributed Normal(0,1)**
3.	**If a "generic" random variable Y is distributed Normal with** $\mu_Y = E(Y)$ $\sigma_Y^2 = \text{Var}(Y)$	$\text{z-score} = \frac{Y-E(Y)}{\sqrt{\text{var}(Y)}}$ **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 (or the calculator that you have brilliantly found on the internet) to obtain interval probabilities for ANY Normal Distribution.

Example 4 -

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 profit is distributed normal, what are the chances that, in a given week, the profit will be between 8 and 10.5 million dollars?

Solution = Pr[8 < X < 10.5] = .367
Calculator used: http://davidmlane.com/hyperstat/z_table.html

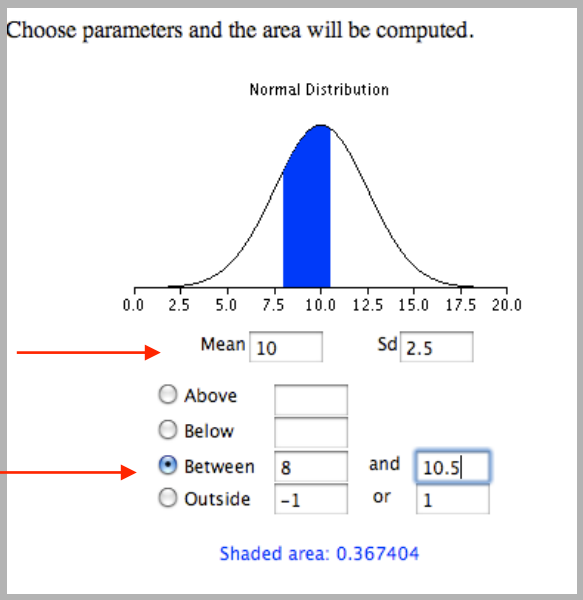
step 1:
Translate the "words" into an event.
Since we're no longer dealing with a standard normal random variable, it is convenient to use X to denote the random variable defined as the profit earned in a given week.
X assuming a value in the interval [8,10.5] is equivalent to Pr[8 <= X <= 10.5] = ?

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 18). 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. You supply the values of μ and σ:



0. Calculation of Percentile Values for a Normal (μ, σ^2): 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 (Confidence intervals are introduced in Unit 6 – Estimation)

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** $\bar{X}=\left(\sigma/\sqrt{n}\right)Z+\mu$ **is Normal (μ, σ^2/n)**
3.	**If Z is distributed Normal(0,1)**	generic $Y=\left(\sqrt{var(Y)}\right)Z+E(Y)$ **is Normal** $\mu_Y = E(Y)$ $\sigma^2_Y = Var(Y)$

Example 5 -

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

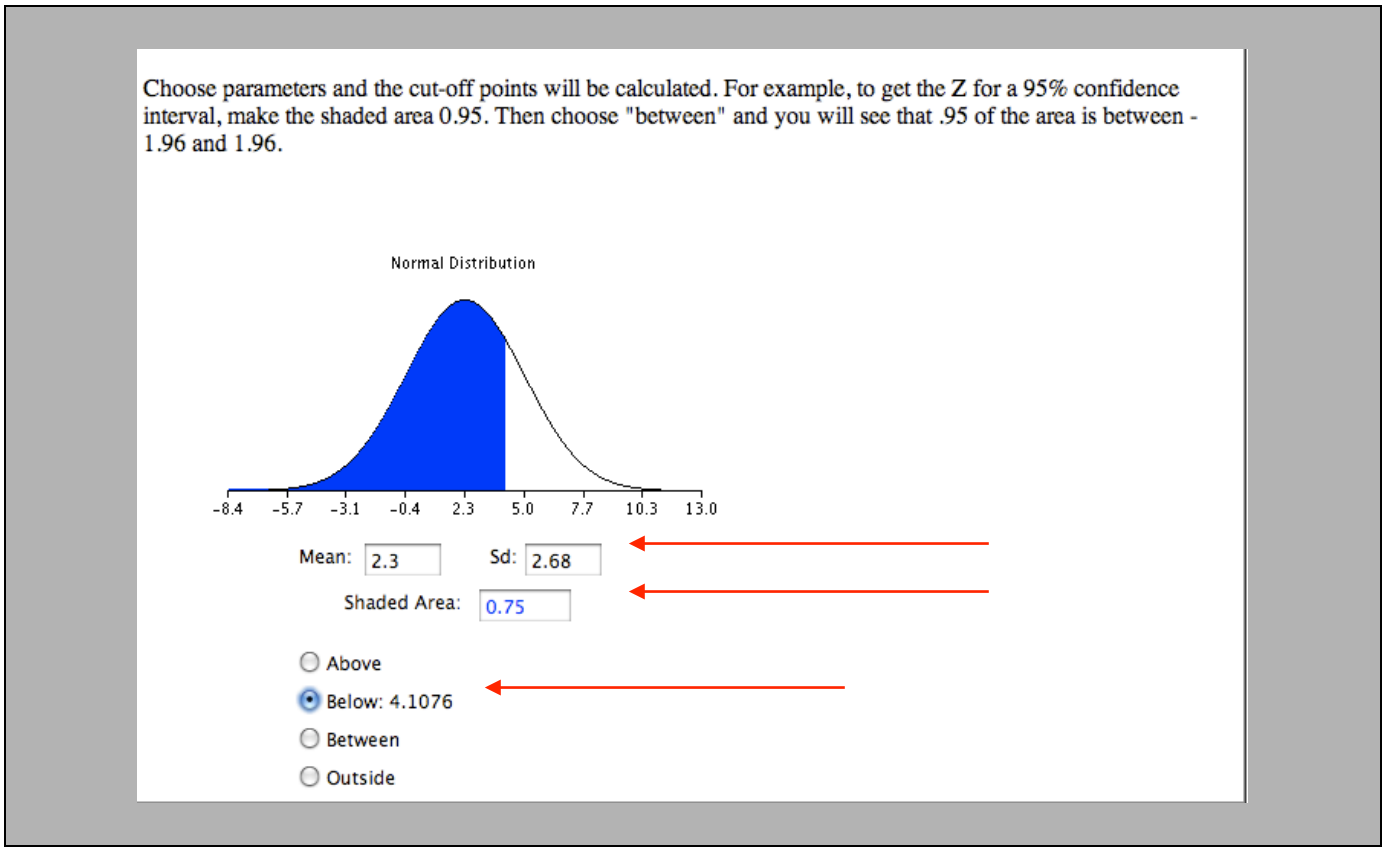
Solution = 4.11 years
Calculator used: http://davidmlane.com/hyperstat/z_table.html

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. 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.

10. Sums and Averages of Normal Random Variables are Normal

So What? Why is this of interest?

In actuality, there are lots of settings where we might be interested in the probability distribution of a sum or an average of normally distributed random variables. Here are just a few examples:

Example 1.

A certain town has two bodies of water. Suppose:
X_1 = Amount of pollutant in a body of water #1 is distributed Normal(μ_1, σ_1^2).
X_2 = Amount of pollutant in a body of water #2 is distributed Normal(μ_2, σ_2^2).
We are actually interested in the total for the whole town. Under independence,
$X_1 + X_2$ = Total amount of pollutant, over bodies of water #1 and #2
is distributed Normal $[\ (\mu_1 + \mu_1),\ (\sigma_1^2 + \sigma_2^2)\ ]$.

Example 2.

Often, in the doctor's office, a patient's blood pressure is taken 3 times, and the average is used. Suppose:
X_1 = 1st blood pressure reading is distributed Normal(μ, σ^2).
X_2 = 2nd blood pressure reading is distributed Normal(μ, σ^2).
X_3 = 3rd blood pressure reading is distributed Normal(μ, σ^2). Under independence, the average

$$\bar{X} = \frac{[X_1+X_2+X_3]}{3} \text{ is distributed Normal } \left(\left[\frac{\mu+\mu+\mu}{3}\right], \left[\frac{\sigma^2+\sigma^2+\sigma^2}{9}\right]\right) = \text{Normal}\left(\mu, \left[\frac{\sigma^2}{3}\right]\right)$$

Example 3 – The randomized controlled trial.
The randomized controlled trial with 2 independent groups, control and intervention.

Suppose we have for Group 1 = controls:

- $X_{11}, X_{12}, \ldots X_{1n1}$ is a simple random sample from a Normal (μ_1, σ_1^2). We then have
- $\bar{X}_{\text{Group 1}}$ is distributed Normal (μ_1, σ_1^2 / n_1)

And that we have for Group 2 = intervention:

- $X_{21}, X_{22}, \ldots X_{2n2}$ is a simple random sample from a Normal (μ_2, σ_2^2)
- $\bar{X}_{\text{Group 2}}$ is distributed Normal (μ_2, σ_2^2 / n_2)

We compare the average responses in the two groups through their difference. Under ***independence:***

$[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}]$ is also distributed Normal with

$$\text{Mean} = [\mu_{\text{Group1}} - \mu_{\text{Group 2}}]$$

$$\text{Variance} = \left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]$$

A General Result for Sums of Independent Variables
Even when they are NOT normal

If random variables X and Y are independent with

$$E[X] = \mu_X \text{ and } \text{Var}[X] = \sigma_X^2$$
$$E[Y] = \mu_Y \text{ and } \text{Var}[Y] = \sigma_Y^2$$

Then

$$E[aX + bY] = a\mu_X + b\mu_Y$$
$$\text{Var}[aX + bY] = a^2\sigma_X^2 + b^2\sigma_Y^2 \text{ and}$$
$$\text{Var}[aX - bY] = a^2\sigma_X^2 + b^2\sigma_Y^2$$

Tip! **This result ALSO says that, when X and Y are independent, the variance of their difference is equal to the variance of their sum. This makes sense if it is recalled that variance is defined using squared deviations which are always positive.**

11. Averages of *Non-Normal* Random Variables May be Normal Too: Introduction to the Central Limit Theorem

This is amazing!

Recall, our focus is on the behavior of the average, $\overline{X}_n$, 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 $X_1 \ldots X_n$; and**
2) **the $X_1 \ldots X_n$ 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 $\overline{X}_n = \left[\dfrac{\sum_{i=1}^{n} X_i}{n}\right]$ **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 $\overline{X}_n$, upon repeated sampling, is eventually Normal $\left(\mu, \dfrac{\sigma^2}{n}\right)$

Great! We can exploit the central limit theorem to compute probabilities of intervals of values for $\overline{X}_n$ distributed Normal(μ,σ^2/n) by using standardization to a z-score.

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

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

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 $\sigma/\sqrt{n}$ 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, ..., 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.