



MATH1. Part II

Probability and Statistics



Chapter 4

The Normal Distribution

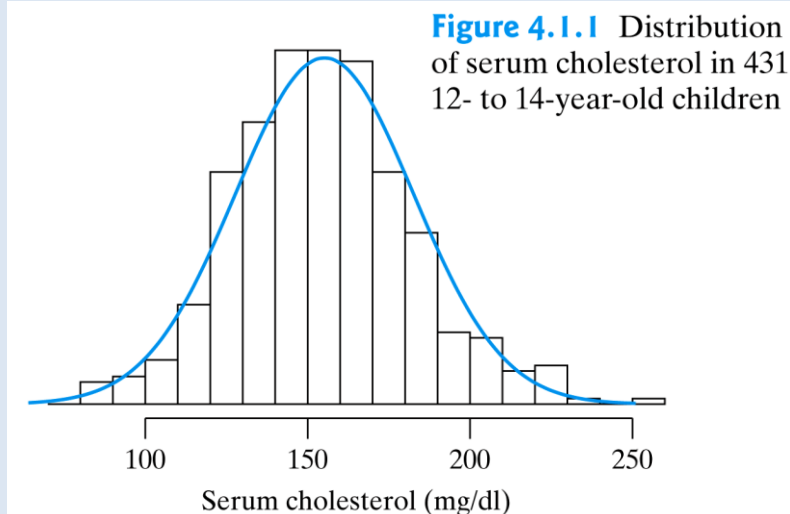
4.1 Introduction

Normal curve

- **Normal curve**: the most important type of **density curve**, which represents relative frequencies as areas under the curve.
- The normal curve is a symmetric “**bell-shaped**” curve.
 - Normal curve is not just any symmetric curve, but rather a specific kind of symmetric curve.

Example 4.1.1 Serum cholesterol

- The distribution of serum cholesterol levels for children between 12 and 14 years of age can be fairly well approximated by a normal curve with
 - mean $\mu = 155$ mg/dl , and
 - standard deviation $\sigma = 27$ mg/dl.



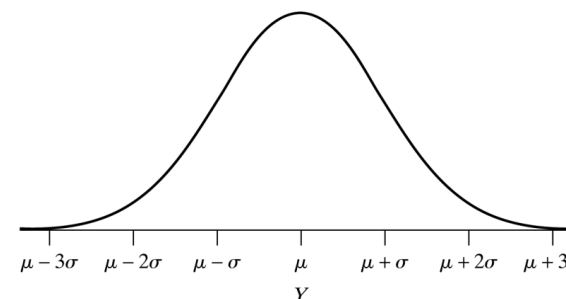
4.2 The Normal Curves

Normal distribution

Normal distribution: a distribution represented by a normal curve is called a normal distribution.

- If a variable Y follows a normal distribution with mean μ and standard deviation σ , $Y \sim N(\mu, \sigma)$, then the density curve of the distribution of Y is given by the following formula:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$



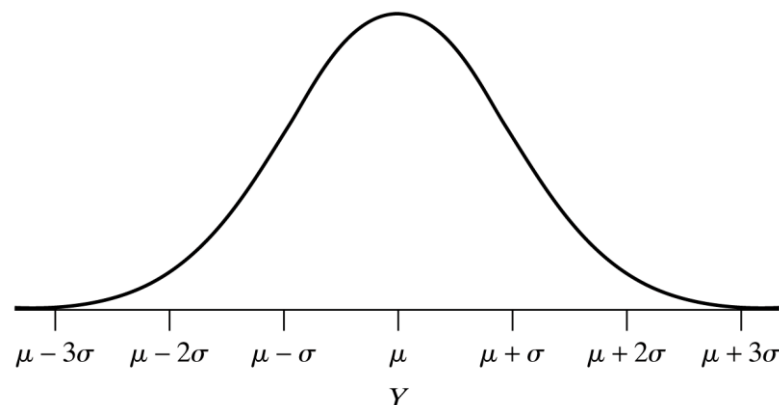
- This function, $f(y)$, is called the density function of the distribution.
- The quantities e and π that appear in the formula are constants, with e approximately equal to 2.71 and π approximately equal to 3.14.
- * We will not make any direct use of the formula in this book.

4.2 The Normal Curves

Shape of normal curve

- The shape of the curve is like a symmetric bell, centered at $y = \mu$.
- The direction of curvature is downward (like an inverted bowl) in the central portion of the curve, and upward in the tail portions.
- The points of inflection (i.e., where the curvature changes direction) are $y = \mu - \sigma$ and $y = \mu + \sigma$.
- In principle the curve extends to $+\infty$ and $-\infty$, never actually reaching the y-axis; however, the height of the curve is very small for y values more than three standard deviations from the mean.

Figure 4.2.1 A normal curve with mean μ and standard deviation σ

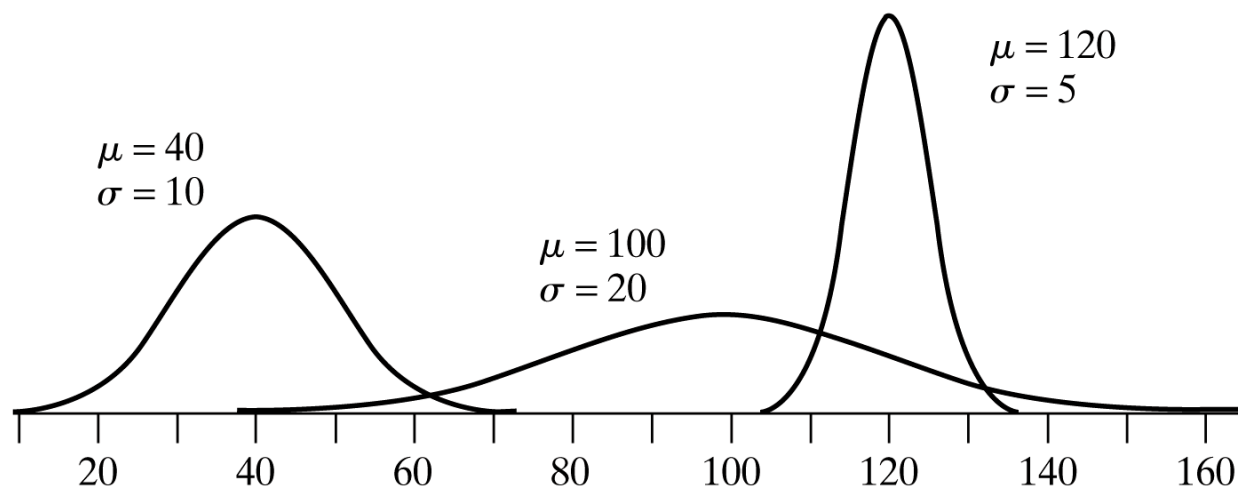


4.2 The Normal Curves

Shape of normal curve

- As the examples below, there are many normal curves; each particular normal curve is characterized by its mean and standard deviation.
- The location of the normal curve along the y-axis is governed by μ since the curve is centered at $y = \mu$;
- The width and the height of the curve (i.e., whether tall and thin or short and wide) are governed by σ .

Figure 4.2.2 Three normal curves with different means and standard deviations

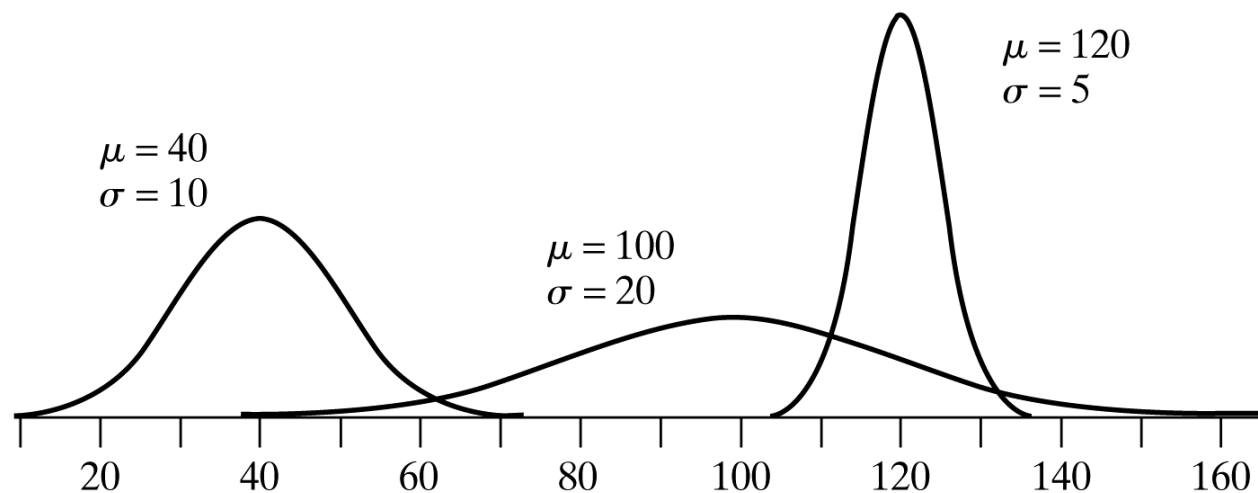


4.2 The Normal Curves

Shape of normal curve

- The area under the curve is exactly equal to 1.
- Since the area under each curve must be equal to 1, a curve with a smaller value of σ must be taller.
- This reflects the fact that the values of Y are more highly concentrated near the mean when the standard deviation is smaller.

Figure 4.2.2 Three normal curves with different means and standard deviations



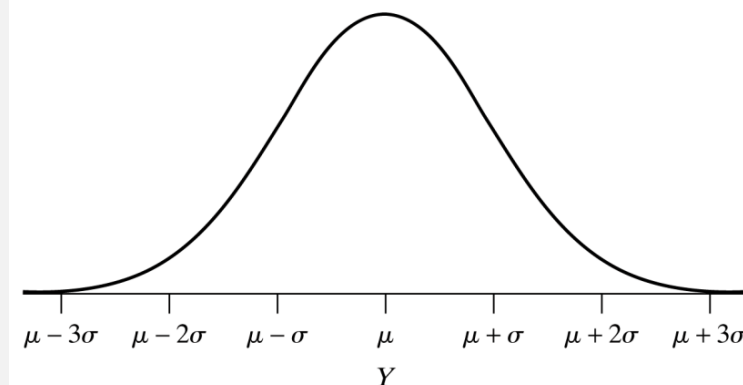
4.2 The Normal Curves

Areas under a normal curve

Why is the area under the curve exactly equal to 1?

- Hint: It is given that $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$

Figure 4.2.1 A normal curve with mean μ and standard deviation σ



4.2 The Normal Curves

Areas under a normal curve

Why is the area under the curve exactly equal to 1?

- Hint: We know that $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$

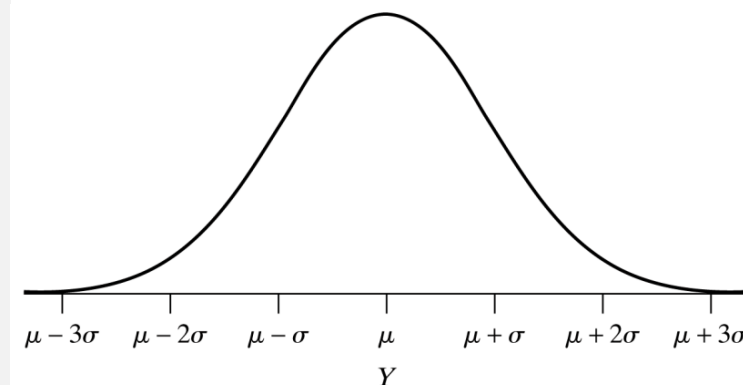
Proof:
$$\text{Area} = \int_{-\infty}^{\infty} f(y) dy = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{y-\mu}{\sqrt{2}\sigma}\right)^2} d\left(\frac{y-\mu}{\sqrt{2}\sigma}\right)$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$= 1$$

Figure 4.2.1 A normal curve with mean μ and standard deviation σ



4.3 Areas under a Normal Curve

The standardized scale

- Rescale the horizontal axis by **Standardization Formula**: $Z = (Y - \mu) / \sigma$
- The **rescaled variable** is denoted by Z . The variable Z is referred to as the standard normal.
- The **Z scale** is referred to as a **standardized scale**.
- The relationship between the two scales is shown in Figure 4.3.1.

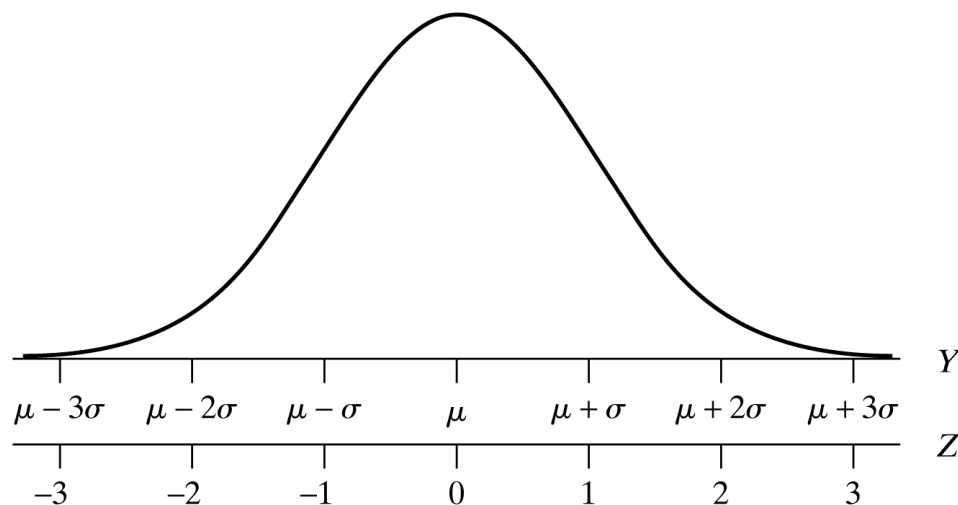


Figure 4.3.1 A normal curve, showing the relationship between the natural scale (Y) and the standardized scale (Z)

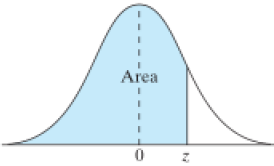
4.3 Areas under a Normal Curve

The standardized scale

- Table 3 gives areas under the **standard normal curve**, with distances along the horizontal axis measured in the Z scale.

616 Statistical Tables

TABLE 3 Areas Under the Normal Curve



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559

For example, for $z = -3.01$,
the tabled area is 0.0013

4.3 Areas under a Normal Curve

The standardized scale

Example: use Table 3 to find the area under a standard normal curve

- Area under $z = 1.53$

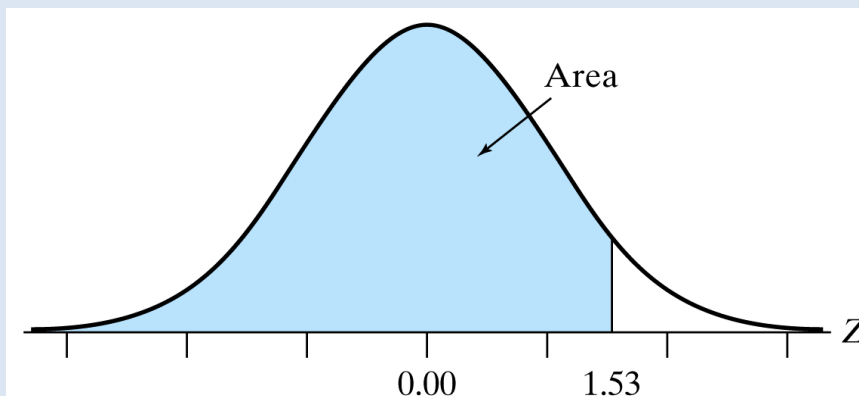


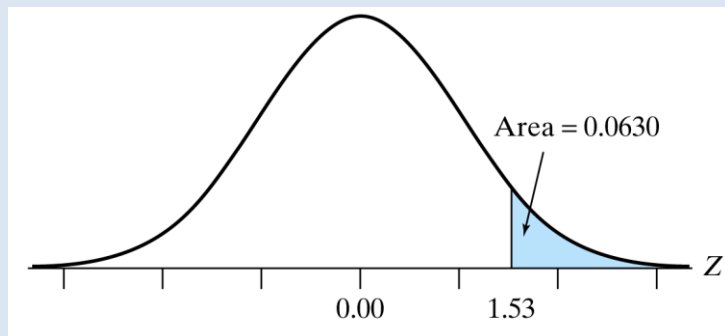
Figure 4.3.2 Illustration
of the use of Table 3

4.3 Areas under a Normal Curve

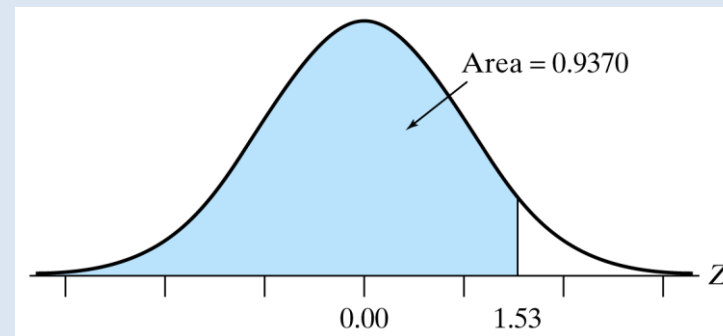
The standardized scale

Example: use Table 3 to find the area under a standard normal curve

- Area under $z = 1.53$



— Find the area above 1.53



— Area under $z = 1.53$ is
 $1.0000 - 0.0630 = 0.9370$

4.3 Areas under a Normal Curve

The standardized scale

Example: use Table 3 to find the area under a standard normal curve

- Area between $z = -1.2$ and $z = 0.8$

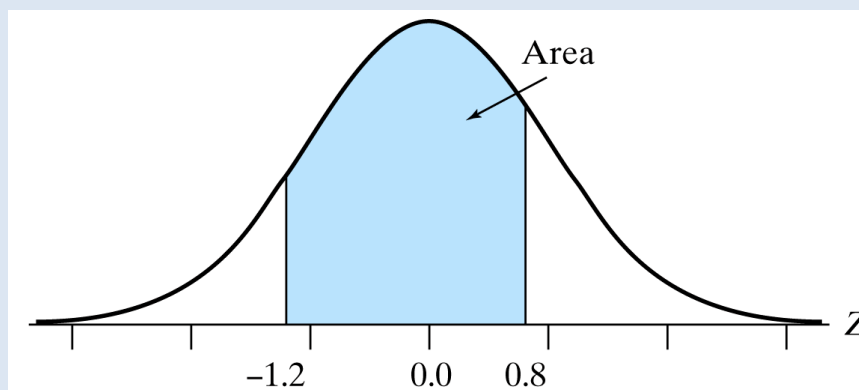


Figure 4.3.4 Area under a standard normal curve between -1.2 and 0.8

4.3 Areas under a Normal Curve

The standardized scale

Example: use Table 3 to find the area under a standard normal curve

- Area between $z = -1.2$ and $z = 0.8$
 - Area below 0.8 is 0.7881;
 - Area below -1.2 is 0.1151;
 - Area between two z values (also commonly called z scores), -1.2 and 0.8, is $0.7881 - 0.1151 = 0.6730$

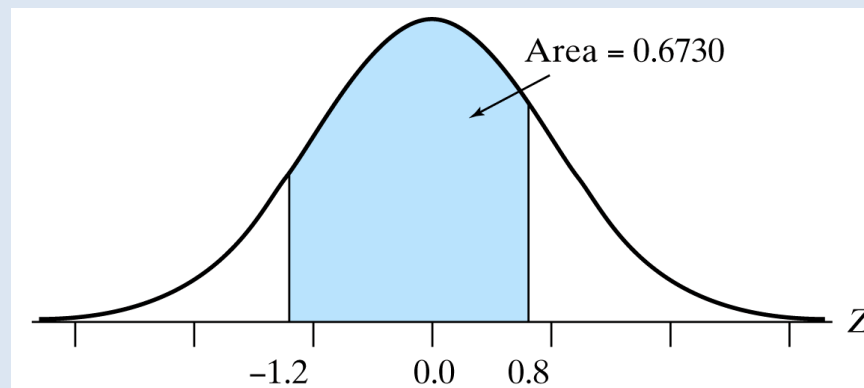


Figure 4.3.4 Area under a standard normal curve between -1.2 and 0.8

4.3 Areas under a Normal Curve

The standardized scale

- Using **Table 3**, we see that the area under the normal curve between $z = -1$ and $z = +1$ is $0.8413 - 0.1587 = 0.6826$. Thus, for any normal distribution, about **68%** of the observations are within **± 1 standard deviation** of the mean.
- Likewise, the area under the normal curve between $z = -2$ and $z = +2$ is $0.9772 - 0.0228 = 0.9544$. This means that for any normal distribution about **95%** of the observations are within **± 2 standard deviations** of the mean a
- The area under the normal curve between $z = -3$ and $z = +3$ is $0.9987 - 0.0013 = 0.9974$. This means that for any normal distribution about **99.7%** of the observations are within **± 3 standard deviations** of the mean.

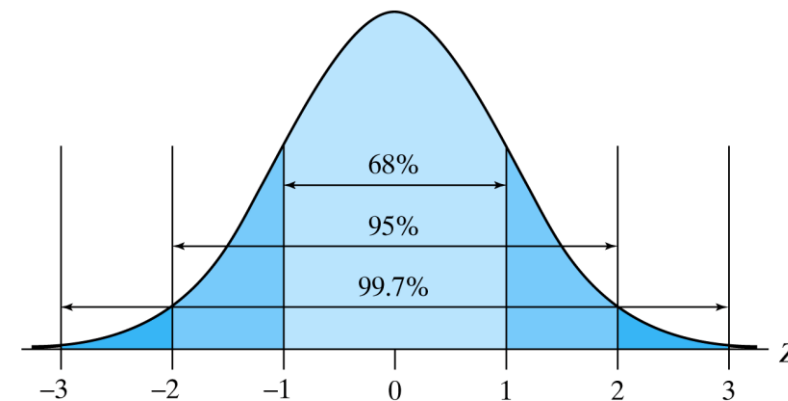


Figure 4.3.5 Areas under a standard normal curve between -1 and $+1$, between -2 and $+2$, and between -3 and $+3$

4.3 Areas under a Normal Curve

The standardized scale

If the variable Y follows a normal distribution, then

- about 68% of the y 's are within ± 1 SD of the mean.
 - about 95% of the y 's are within ± 2 SDs of the mean.
 - about 99.7% of the y 's are within ± 3 SDs of the mean.
- * These statements provide a very definite interpretation of the standard deviation in cases where a distribution is approximately normal.

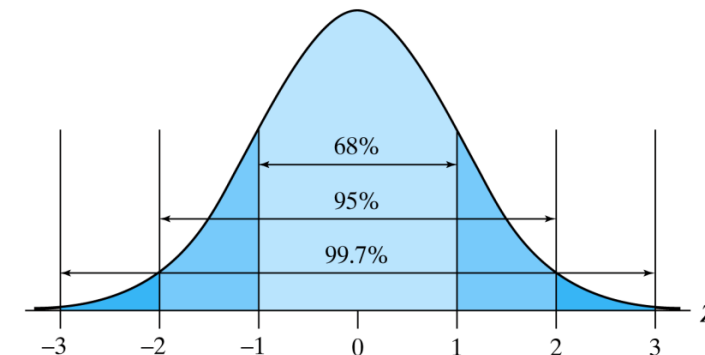


Figure 4.3.5 Areas under a standard normal curve between -1 and $+1$, between -2 and $+2$, and between -3 and $+3$

Example 4.1.1 Serum cholesterol (continued)

- The distribution of serum cholesterol levels for children between 12 and 14 years of age can be fairly well approximated by a normal curve with
 - mean $\mu = 155$ mg/dl, and
 - standard deviation $\sigma = 27$ mg/dl.
- Explain the meaning of Figure 4.3.6.

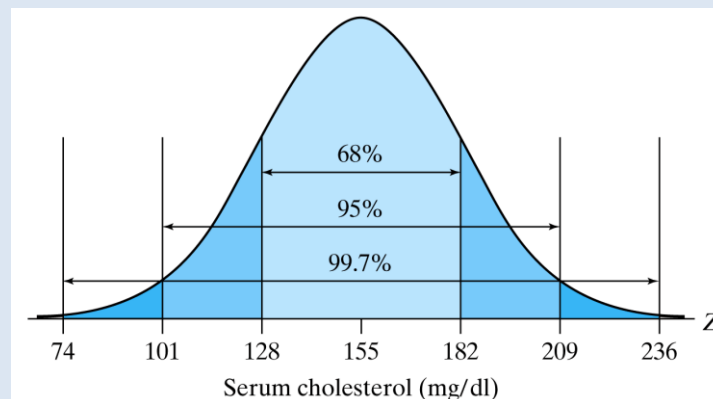


Figure 4.3.6 The 68/95/99.7 rule and the serum cholesterol distribution

4.3 Areas under a Normal Curve

The standardized scale

Example 4.1.1 Serum cholesterol (continued)

- The distribution of serum cholesterol levels for children between 12 and 14 years of age can be fairly well approximated by a normal curve with
 - mean $\mu = 155$ mg/dl , and
 - standard deviation $\sigma = 27$ mg/dl.
- Explain the meaning of Figure 4.3.6.
 - about 68% of the serum cholesterol values are between 128 mg/dl and 182 mg/dl,
 - about 95% are between 101 mg/dl and 209 mg/dl,
 - almost all are between 74 mg/dl and 236 mg/dl.

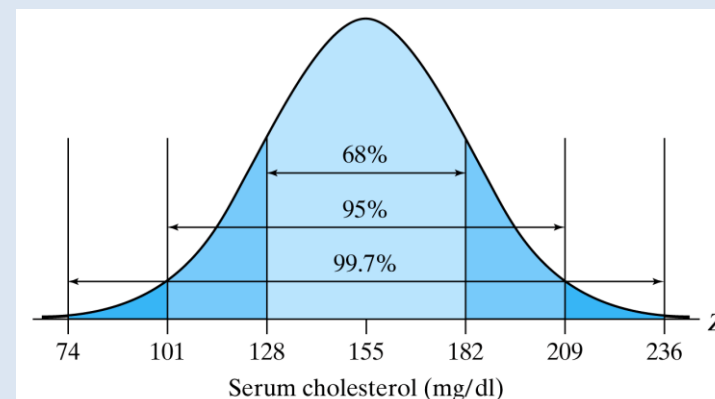


Figure 4.3.6 The 68/95/99.7 rule and the serum cholesterol distribution

4.3 Areas under a Normal Curve

Determining areas for a normal Curve

Example 4.3.1 Lengths of Fish

- The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. Use Table 3 to answer various questions about the population.
- (a) What percentage of the fish are less than 60 mm long?
- (b) What percentage of the fish are more than 51 mm long?
- (c) What percentage of the fish are between 51 and 60 mm long?
- (d) What percentage of the fish are between 58 and 60 mm long?

4.3 Areas under a Normal Curve

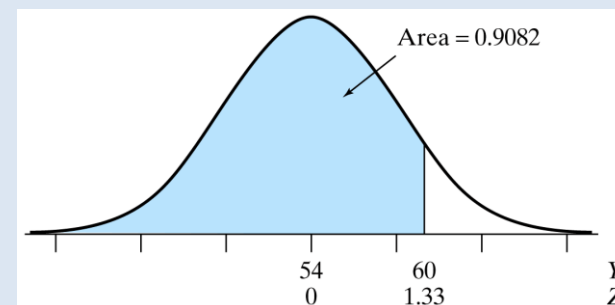
Determining areas for a normal Curve

Example 4.3.1 Lengths of Fish

- The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. Use Table 3 to answer various questions about the population.

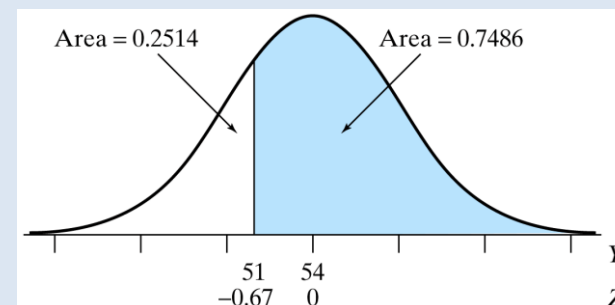
- (a) What percentage of the fish are less than 60 mm long?

- $Z = (Y - \mu) / \sigma = (60 - 54) / 4.5 = 1.33$
- 90.82% of the fish are less than 60 mm long



- (b) What percentage of the fish are more than 51 mm long?

- $Z = (Y - \mu) / \sigma = (51 - 54) / 4.5 = -0.67$
- 74.86% of the fish are more than 51 mm long



4.3 Areas under a Normal Curve

Determining areas for a normal Curve

Example 4.3.1 Lengths of Fish

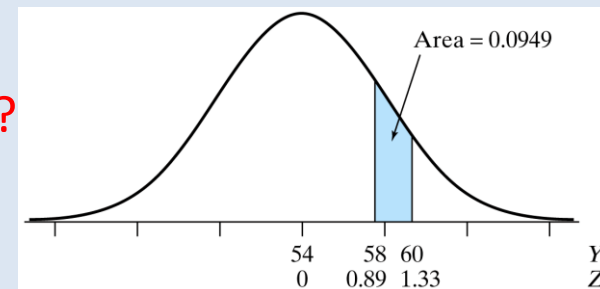
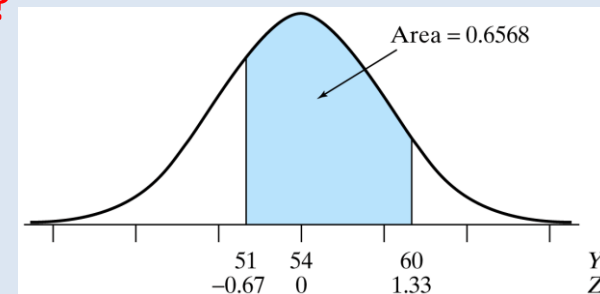
- The mean length of the fish is 54.0 mm, and the standard deviation is 4.5 mm. Use Table 3 to answer various questions about the population.

- (c) What percentage of the fish are between 51 and 60 mm long?

- $Z = (Y - \mu) / \sigma = (60 - 54) / 4.5 = 1.33$
- 90.82% of the fish are less than 60 mm long
- $Z = (Y - \mu) / \sigma = (51 - 54) / 4.5 = -0.67$
- 25.14 % of the fish are less than 51 mm long
- 65.68% of the fish are between 51 and 60 mm long

- (d) What percentage of the fish are between 58 and 60 mm long?

- $Z = (Y - \mu) / \sigma = (58 - 54) / 4.5 = 0.89$
- 9.49% of the fish are between 58 and 60 mm long



4.4 Assessing Normality

How to assess normality?

Example 4.1.1 Serum cholesterol (continued)

- The distribution of serum cholesterol levels for children between 12 and 14 years of age can be fairly well approximated by a normal curve with
 - mean $\mu = 155$ mg/dl , and
 - standard deviation $\sigma = 27$ mg/dl.
- How to check if the data follows the normal distribution?
 - From the data, we know
 - about 70.5% of the y's are within ± 1 SD of the mean.
 - about 94.4% of the y's are within ± 2 SDs of the mean.
 - about 99.8% of the y's are within ± 3 SDs of the mean.
 - agree very well with the theoretical percentages of 68%, 95%, 99.7%.
 - This agreement supports the claim that serum cholesterol levels for 12- to 14-year-olds have a normal distribution. This reinforces the visual evidence of Figure 4.1.1.

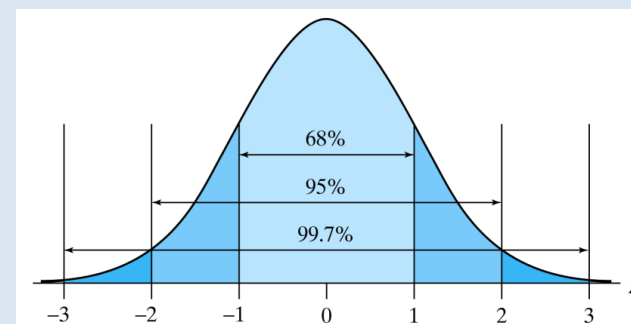
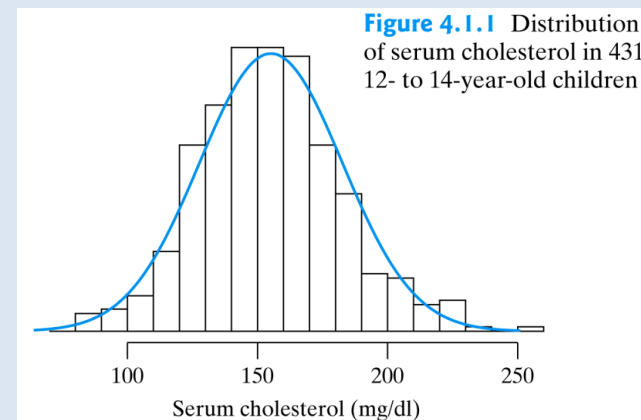


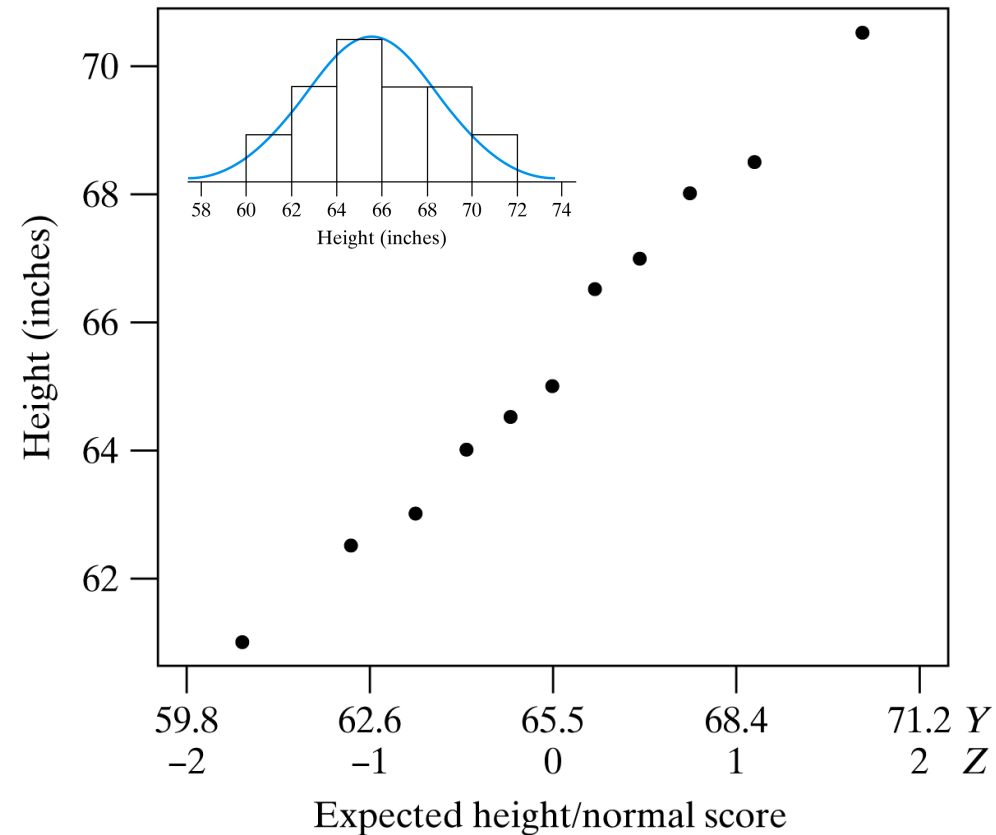
Figure 4.3.5 Areas under a standard normal curve between -1 and +1, between -2 and +2, and between -3 and +3

4.4 Assessing Normality

Normal quantile plot

(also called normal probability plot)

- a special statistical graph that is used to assess normality
- A **normal quantile plot** is a scatterplot that compares our observed data values to values we would expect to see if the population were normal.
- If the data come from a normal population, the points in this plot should follow a straight line, which is much easier to visually recognize than a bell shape of a jagged histogram.



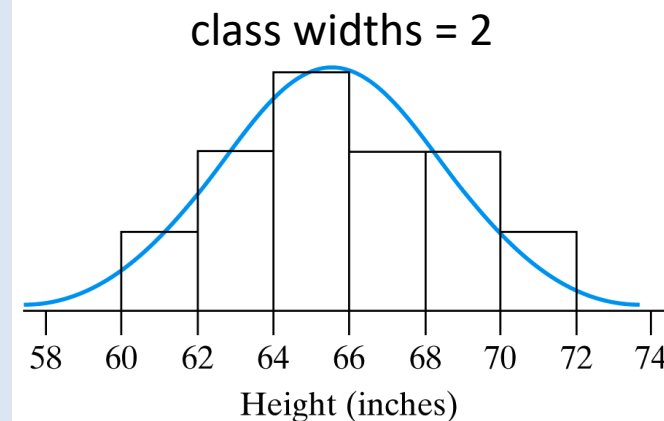
4.4 Assessing Normality

Normal quantile plot

Example 4.4.3 Height of eleven Women

- Heights (in inches) of a sample of 11 women, sorted from smallest to largest:
- 61, | 62.5, 63, | 64, 64.5, 65, | 66.5, 67, | 68, 68.5, | 70.5
- Determine if above data follows normal distribution.
- Normality can be assessed by **normal quantile plot**.
 - 1. calculate percentile
 - 2. calculate theoretical height for normal distribution
 - 3. compare data

Figure 4.4.2 Histogram of the heights of 11 women



4.4 Assessing Normality

Normal quantile plot

Example 4.4.3 Height of eleven Women

1. calculate percentile

Table 4.4.1 Computing indices and percentiles for the heights of 11 women

i	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Percentile $100(i/11)$	9.09	18.18	27.27	36.36	45.45	54.55	63.64	72.73	81.82	90.91	100.00
Adjusted percentile $100(i - \frac{1}{2})/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45

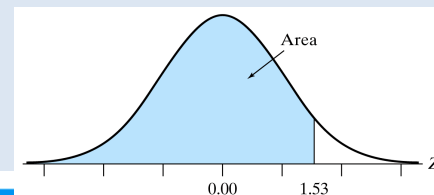


Figure 4.3.2 Illustration of the use of Table 3

Meaning of percentile

- Observed height: sample mean = 65.5, and standard deviation = 2.9
- Frequency distribution (LectureS2): 9.1% (1/11) of our sample is 61 inches or shorter, 18.2% (2/11) of our sample is 62.5 inches or shorter, ... 100% (11/11) of our sample is 70.5 inches or shorter.
- It seems unreasonable to believe that 100% of the population is 70.5 inches or shorter based on the small sample. Therefore, we make a correction: $100(i - 1/2) / 11$.

**Different software packages may compute these proportions differently and may also modify the formula based on sample size.*

4.4 Assessing Normality

Normal quantile plot

Example 4.4.3 Height of eleven Women

2. calculate theoretical height for normal distribution

Table 4.4.2 Computing theoretical z scores and heights for 11 women

i	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile $100(i - \frac{1}{2})/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45
z	-1.69	-1.10	-0.75	-0.47	-0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4

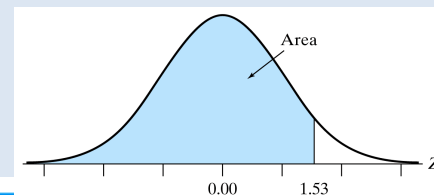


Figure 4.3.2 Illustration of the use of Table 3

Meaning of percentile

$$Y = Z \times \sigma + \mu$$

- Use adjusted percentiles and Table 3 to get z scores for standard normal curve
- Use z scores to get theoretical heights: $Y = Z \times \sigma + \mu$, for normal curve
 - **Eg:** In this example, the sample mean and standard deviation are 65.5 and 2.9, respectively, so the expected height of the shortest woman in a sample of 11 women from a normal population is $65.5 - 1.69 \times 2.9 = 60.6$ inches.

4.4 Assessing Normality

Normal quantile plot

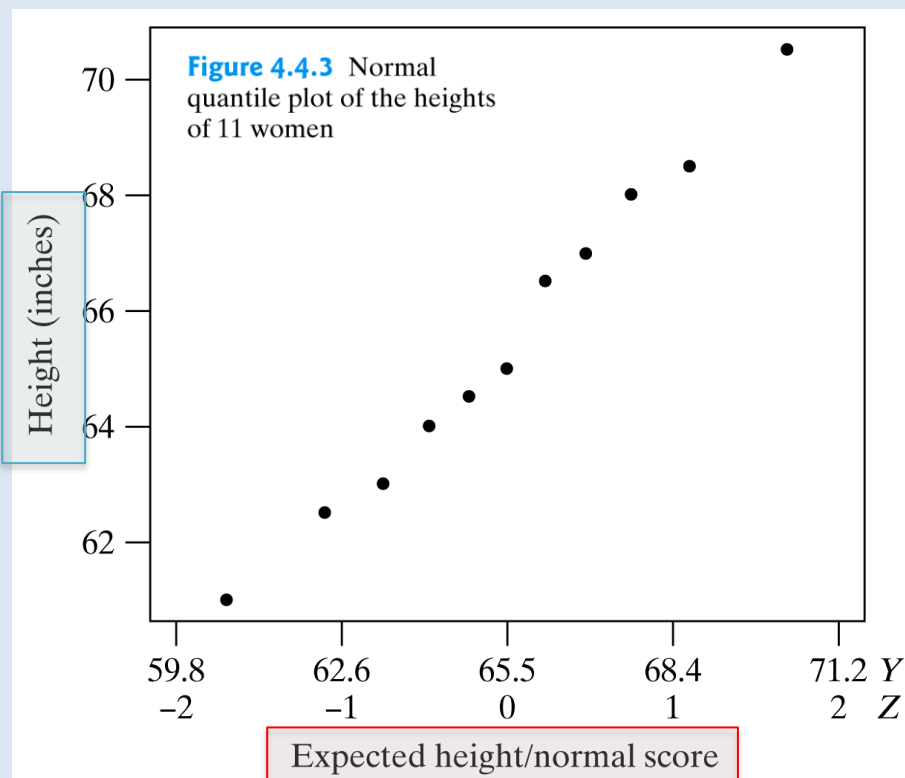
Example 4.4.3 Height of eleven Women

3. Compare data

Table 4.4.2 Computing theoretical z scores and heights for 11 women

i	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile $100(i - \frac{1}{2})/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45
z	-1.69	-1.10	-0.75	-0.47	-0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4

- Plot the observed heights against the theoretical heights in a scatterplot
- Linear - observed values generally agree with the theoretical values
- Non-linear - observed values do NOT agree with the theoretical values

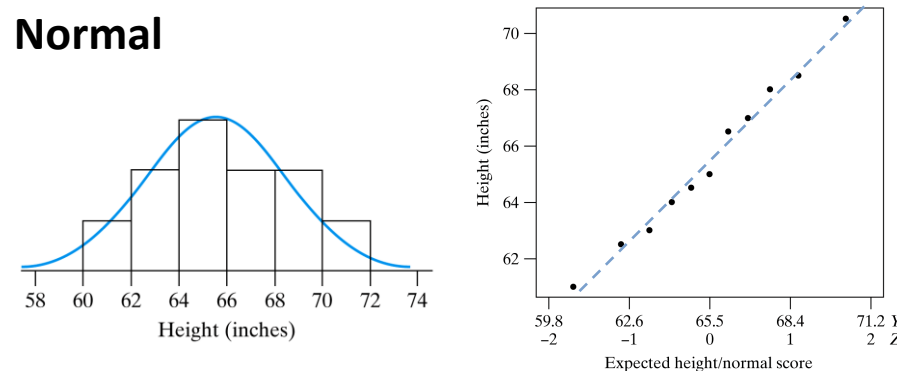


4.4 Assessing Normality

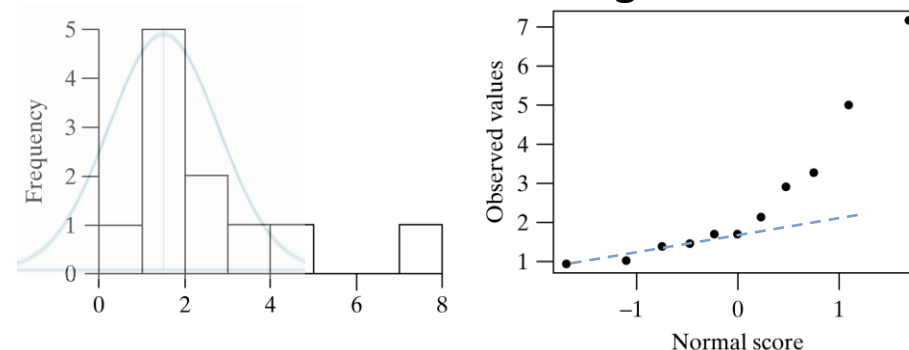
Making decisions about normality

- straight line in the normal quantile plot: normal population.
- top of the plot bends up: the y values at the upper end of the distribution are too large for the distribution to be bell-shaped; that is, the distribution is skewed to the right or has large outliers.
- bottom of the plot bends down: the y values at the lower end of the distribution are too small for the distribution to be bell-shaped; that is, the distribution is skewed to the left or has small outliers.

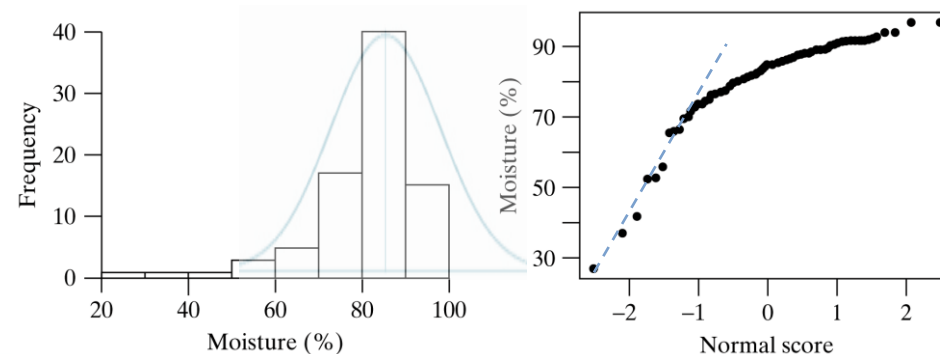
Normal



Nonnormal: skewed to the right



Nonnormal: skewed to the left

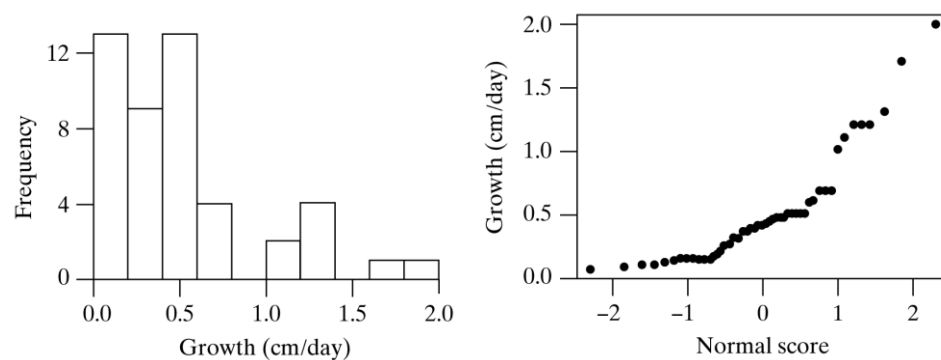


4.4 Assessing Normality

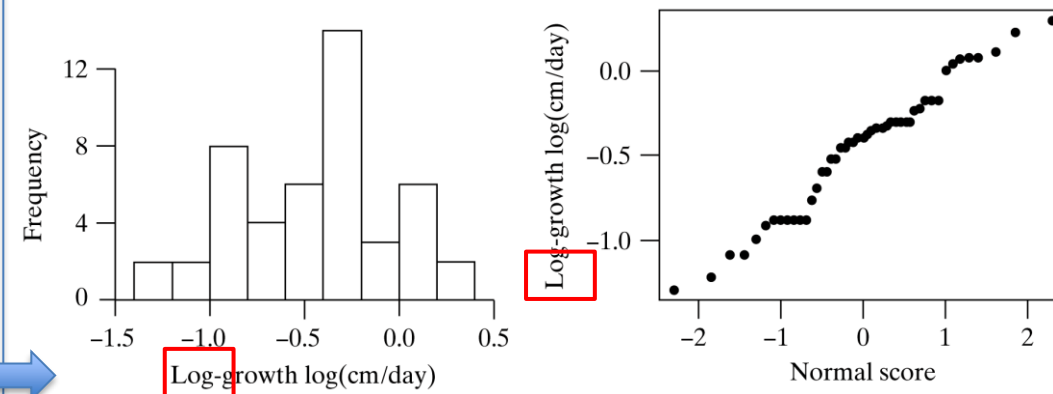
Transformations for non-normal data

- Sometimes a histogram or normal quantile plot shows that our data are nonnormal, but a transformation of the data gives us a symmetric, bell-shaped curve.
- In such a situation, we may wish to transform the data and continue our analysis in the new (transformed) scale.

Nonnormal: skewed to the right



Normal



4.4 Assessing Normality

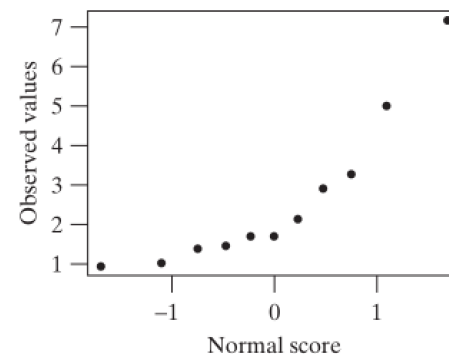
Transformations for non-normal data

- In general, if the distribution is skewed to the right then one of the following transformations should be considered:

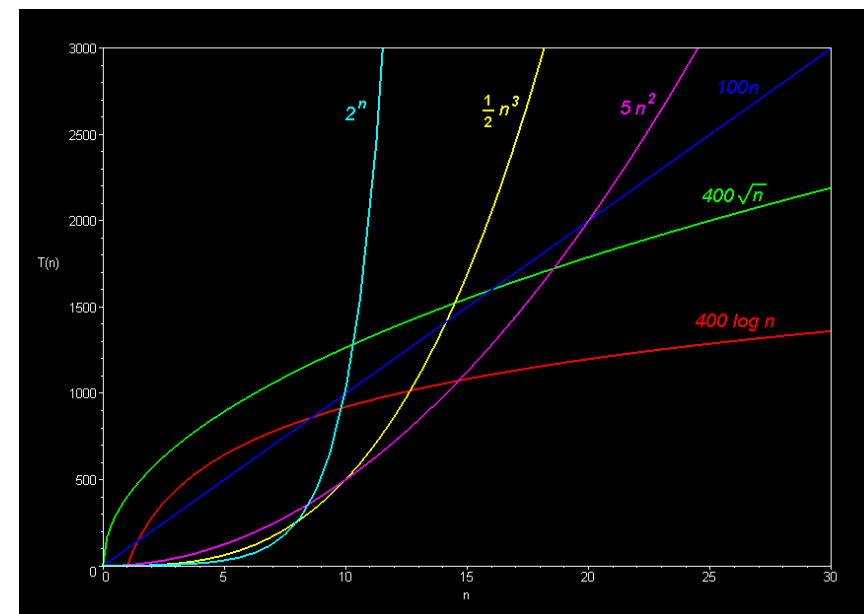
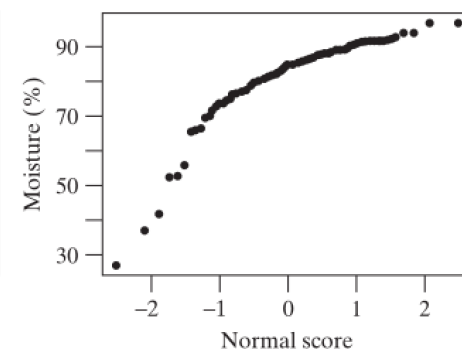
$$\sqrt{Y}, \log Y, 1/\sqrt{Y}, 1/Y.$$

- These transformations will pull in the long right-hand tail and push out the short left-hand tail, making the distribution more nearly symmetric.
- A square root transformation will change a mildly skewed distribution into a symmetric distribution
- A log transformation may be needed if the distribution is more heavily skewed.
- If the distribution of a variable Y is skewed to the left, then raising Y to a power greater than 1 can be helpful.

Nonnormal:
skewed to the right



Nonnormal:
skewed to the left



4.4 Assessing Normality

Normal quantile plot - exercise

Example 4.4.4 Wait time in the coffee shop

- The data in the table is a random sample of 16 individuals wait time in the coffee shop.
- Is there evidence to support the belief that the variable of waiting time follows a normal distribution?
- Normality can be assessed by **normal quantile plot**.
 - 1. calculate percentile
 - 2. calculate theoretical wait time for normal distribution
 - 3. compare data

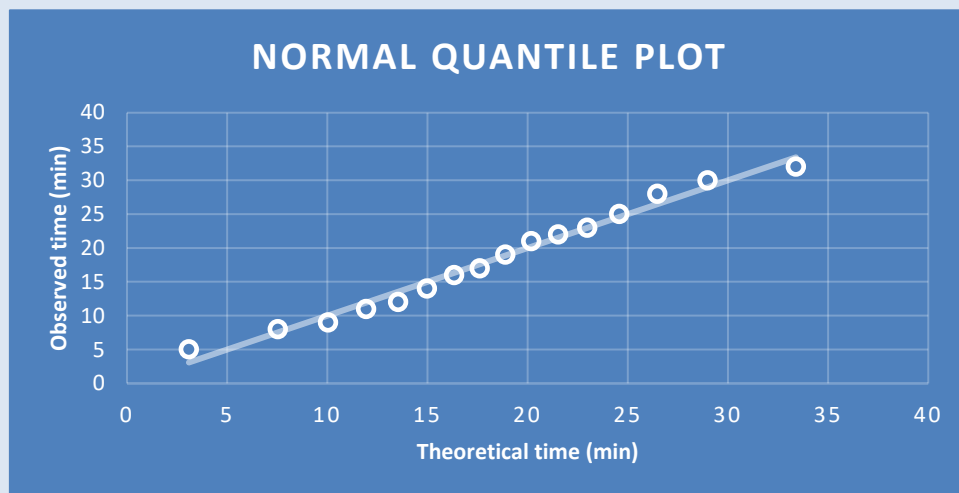
Observed value in (Minutes)
14
8
32
22
12
5
16
28
19
21
11
23
25
17
30
9

4.4 Assessing Normality

Normal quantile plot - exercise

Example 4.4.4 Wait time in the coffee shop

- The data in the table is a random sample of 16 individuals wait time in the coffee shop.
- Is there evidence to support the belief that the variable of waiting time follows a normal distribution?



i	Observed time (min)	1. Adjusted percentile	Z value	2. Theoretical time (min)
1	5	3.125	-1.86	3
2	8	9.375	-1.32	8
3	9	15.625	-1.01	10
4	11	21.875	-0.78	12
5	12	28.125	-0.58	14
6	14	34.375	-0.40	15
7	16	40.625	-0.24	16
8	17	46.875	-0.08	18
9	19	53.125	0.08	19
10	21	59.375	0.24	20
11	22	65.625	0.40	22
12	23	71.875	0.58	23
13	25	78.125	0.78	25
14	28	84.375	1.01	26
15	30	90.625	1.32	29
16	32	96.875	1.86	33
AVG	18.25			
SD	8.14			



4.5 Perspective

- The normal distribution is also called the Gaussian distribution, after the German mathematician K. F. Gauss.
- Naturally occurring biological distributions could be described better by a skewed curve than by a normal curve.
- A major use of the normal distribution is not to describe natural distributions, but rather to describe certain theoretical distributions, called sampling distributions, that are used in the statistical analysis of data





Summary

Chapter 4 -The Normal Distribution

- 4.1 Introduction
- 4.2 The Normal Curves
- 4.3 Areas Under a Normal Curve
- 4.4 Assessing Normality
- 4.5 Perspective





Homework

Chapter 4

- 4.3.3 ; 4.3.17 ;
- 4.4.6

