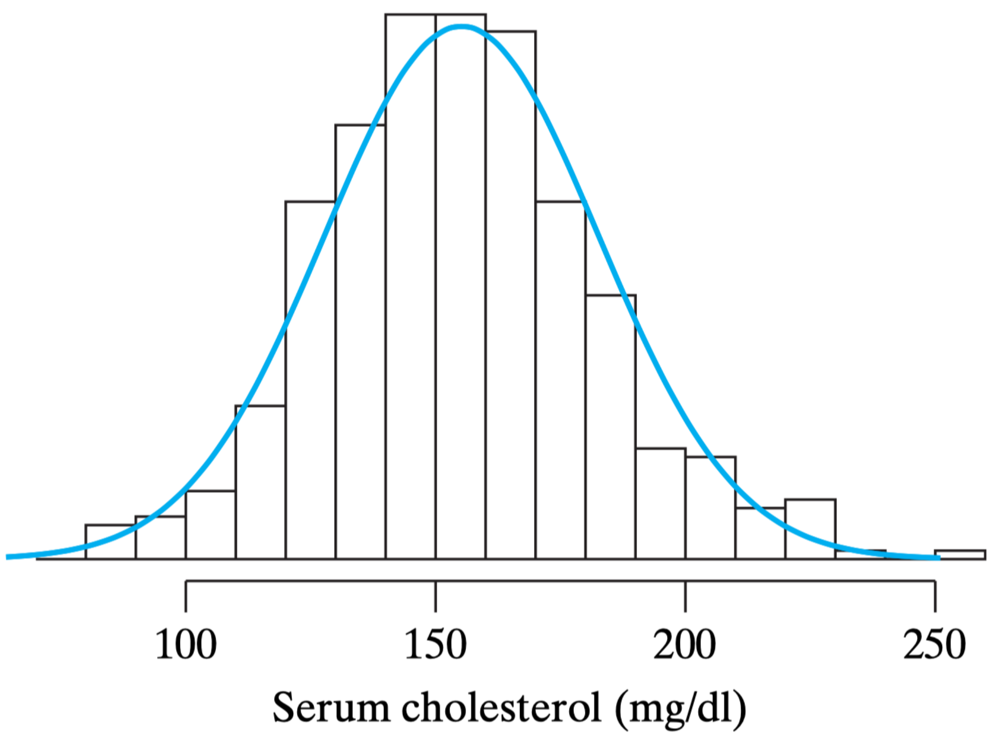
The Normal Distribution

Introduction

- In this chapter we study the most important type of density curve: the normal curve.
- The normal curve is a symmetric "**bell-shaped**" curve whose exact form we will describe next.
- A distribution represented by a normal curve is called a **normal distribution**.

Example: serum cholesterol

The relationship between the concentration of cholesterol in the blood and the occurrence of heart disease has been the subject of much research. As part of a government health survey, researchers measured serum cholesterol levels for a large sample of Americans, including children. The distribution 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. The following figure shows a histogram based on a sample of 431 children between 12 and 14 years old, with the normal curve superimposed.



The Normal Curves

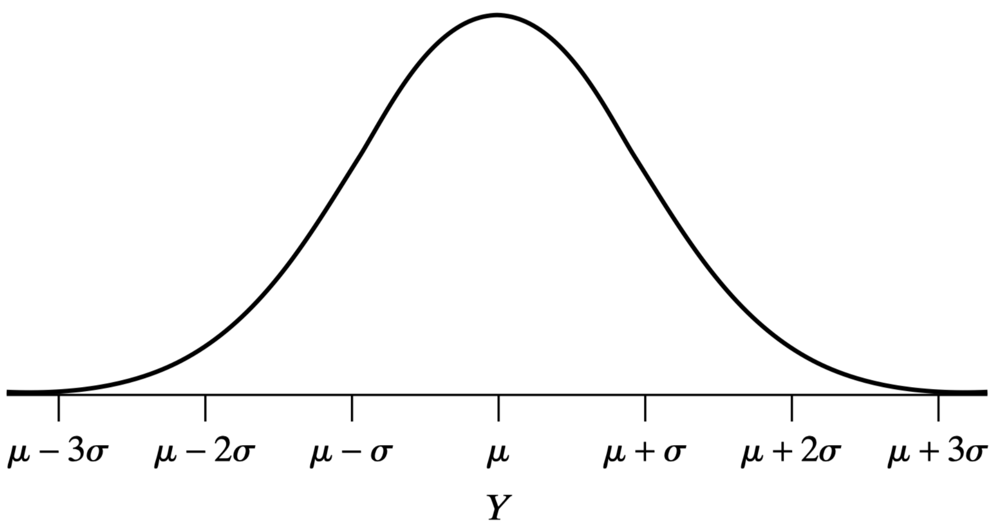
- There are many normal curves; each particular normal curve is characterized by its mean and standard deviation.

- If a random variable Y follows a normal distribution with mean μ and standard deviation σ, then it is common to write $Y \sim N(\mu, \sigma^2)$.
- The probability density function (pdf) of $Y \sim N(\mu, \sigma^2)$ is

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

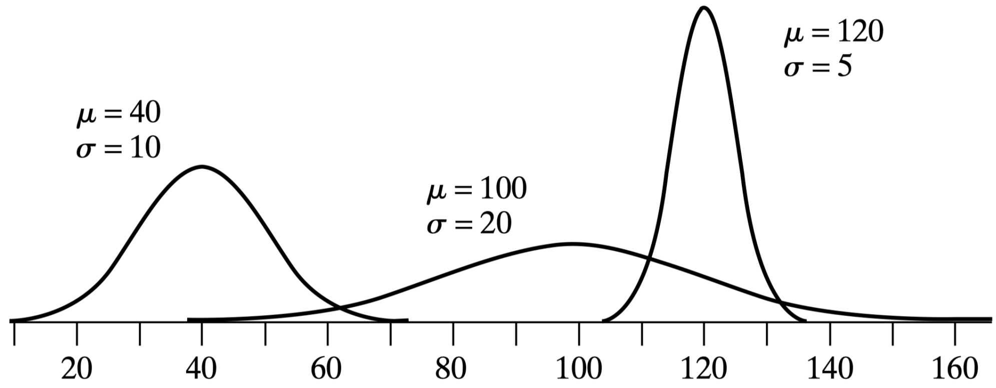
  which expresses the height of the normal curve as a function of the position along the horizontal axis. The quantities e and π that appear in the formula are constants, with e approximately equal to 2.71 and π approximately equal to 3.14.

- The figure below shows a graph of a 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.
- 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.
- The area under the curve is exactly equal to 1.



Normal curves with different means and SDs

- 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.

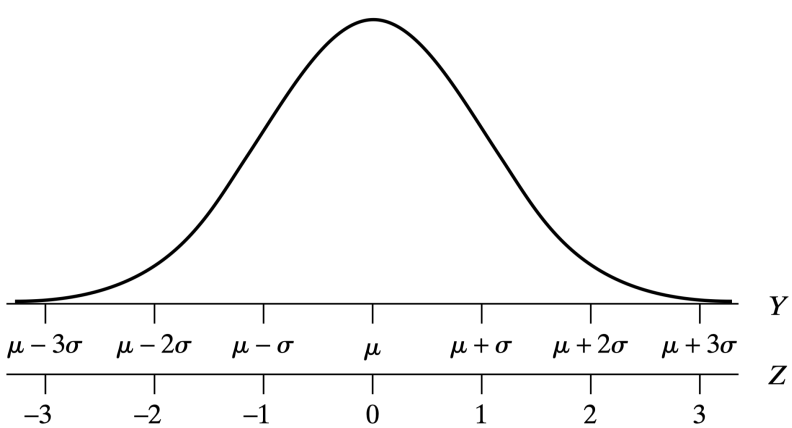


Areas under a Normal Curve

- The standard normal distribution, represented by Z, is the normal distribution having a mean of 0 and a standard deviation of 1. That is, $Z \sim N(0,1)$.
- If X is a random variable from a normal distribution with mean μ and standard deviation σ, its Z-score (standardization) may be calculated from X by subtracting μ and dividing by the standard deviation σ:

$$Z = \frac{Y-\mu}{\sigma}.$$

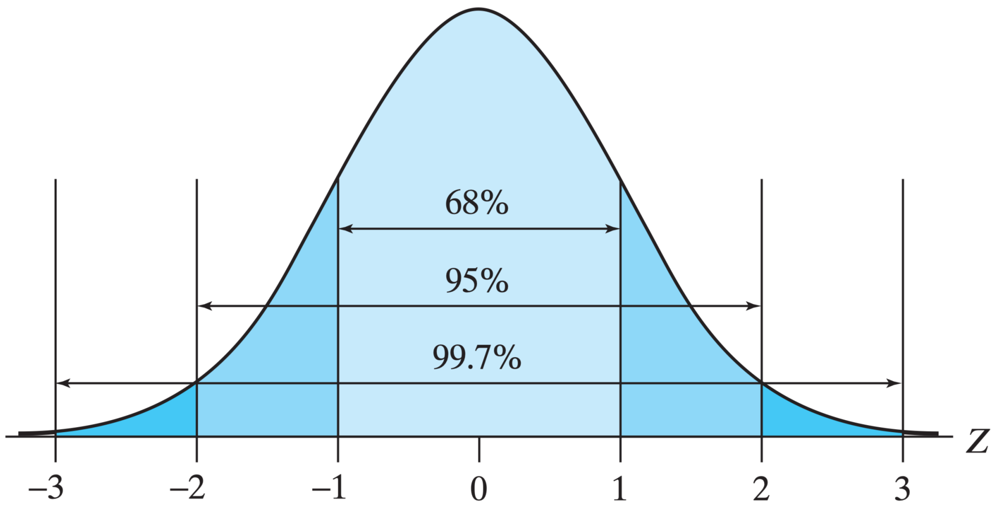
- Z table gives areas under the standard normal curve, with distances along the horizontal axis measured in the Z scale.
- Each area tabled in the body of Z table is the area under the standard normal curve **below** a specified value of z, tabled in the margins.
- If we want to find the area *above* a given value of z, we subtract the tabulated area from 1.
- How to find the area between two z values?



The empirical rule for normal distribution

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.



Determining areas for a normal curve

By taking advantage of the standardized scale, we can use Z table to answer detailed questions about any normal population when the population mean and standard deviation are specified.

A professor's exam scores are approximately distributed normally with mean 80 and standard deviation 5.

- What is the probability that a student scores an 82 or less? 0.65542
- What is the probability that a student scores a 90 or more? 0.02275
- What is the probability that a student scores between 74 and 82? 0.54035

Inverse reading of Z table

We often need to determine corresponding z-values when we want to determine a percentile of a normal distribution. For example, suppose we want to find the 70th percentile of a standard normal distribution. We need to look in Z table for an area of 0.7000. The closest value is an area of 0.6985, corresponding to a z value of 0.52.

- What is the first quartile of the exam score distribution? 76.65
- What is the 70th percentile of the exam score distribution? 82.6

Assessing Normality

Many statistical procedures are based on having data from a normal population. In this section we consider ways to assess whether it is reasonable to use a normal curve model for a set of data and, if not, how we might proceed.

Normal quantile plots

A **normal quantile plot** is a special statistical graph that is used to assess normality. We present this statistical tool with an example using the heights (in inches) of a sample of 11 women, sorted from smallest to largest:

$$61, 62.5, 63, 64, 64.5, 65, 66.5, 67, 68, 68.5, 70.5$$

Based on these data, does it make sense to use a normal curve to model the distribution of women's heights?

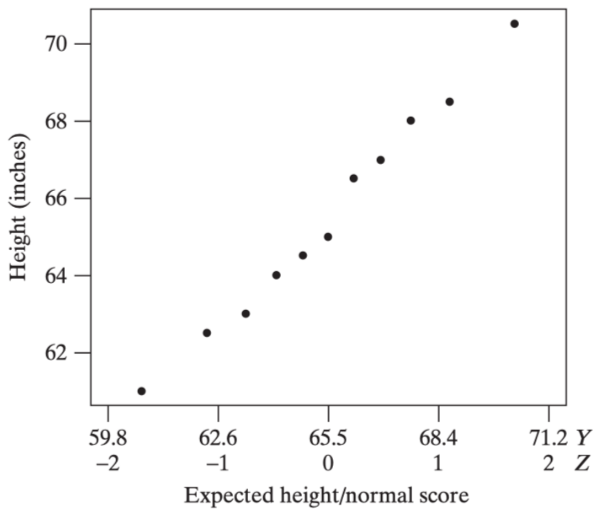
Table 4.4.1 Computing indices and percentiles for the heights of 11 women

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 $100\left(i-\frac{1}{2}\right)/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45

- sort the data from smallest to largest.
- calculate the adjusted percentiles $100(i-1/2)/n$.
- find the corresponding Z scores.
- calculate the theoretical quantiles $\mu + Z \times \sigma$.
- plot the sample quantiles against the theoretical quantiles in a scatterplot.

Table 4.4.2 Computing theoretical z scores and heights for 11 women

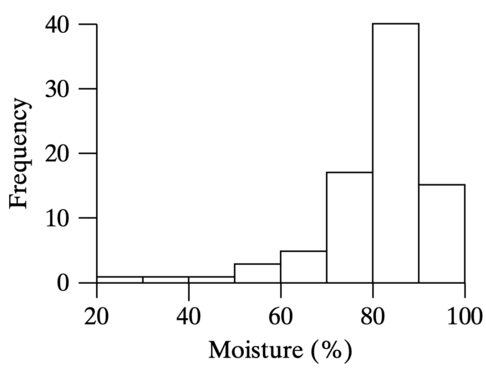
i	1	2	3	4	5	6	7	8	9	10	11
Observed height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile $100\left(i-\frac{1}{2}\right)/11$	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.36	95.45
z	−1.69	−1.10	−0.75	−0.47	−0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6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

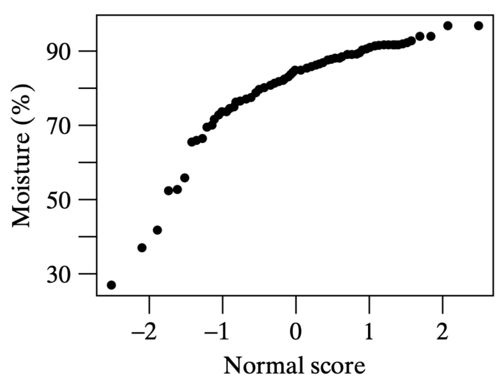


- In this case our plot appears fairly **linear**, suggesting that the observed values generally agree with the theoretical values and the normal model provides a reasonable approximation to the data.
- If the data do not agree with the normal model, then the plot will show strong **nonlinear** patterns such as curvature or S shapes.

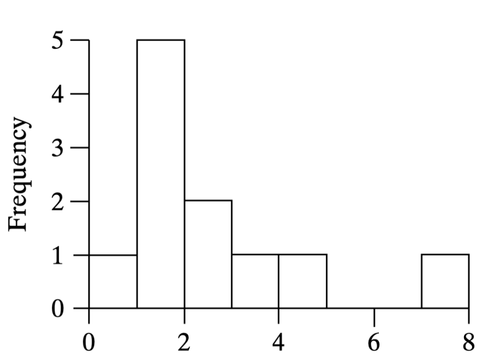
Skewness in normal quantile plots

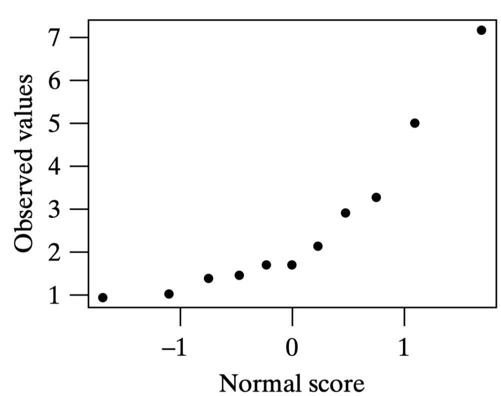
- Histogram and normal quantile plot of a distribution that is skewed to the left



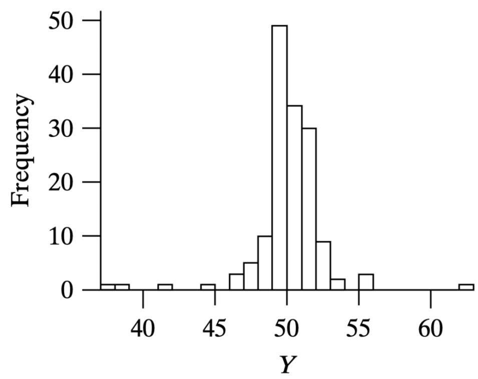


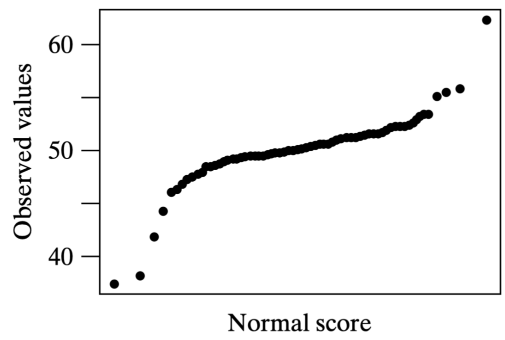
- Histogram and normal quantile plot of a distribution that is skewed to the right





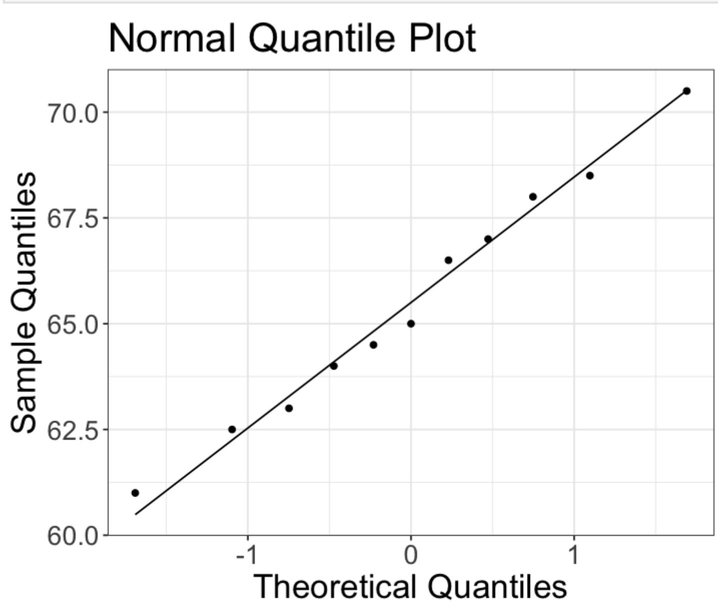
- Histogram and normal quantile plot of a distribution that has long tails





In [7]:
```
library(ggplot2)

# Create the normal quantile plot using ggplot2
g <- ggplot(data.frame(y = c(61, 62.5, 63, 64,
                             64.5, 65, 66.5, 67,
                             68, 68.5, 70.5)),
            aes(sample = y)) +
    stat_qq() +
    stat_qq_line() +
    labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
    ggtitle("Normal Quantile Plot") +
    theme_bw() +
    theme(text = element_text(size = 20))
options(repr.plot.width=6, repr.plot.height=5)
g
```



Transformations for nonnormal 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.

- 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. **Each of these is more drastic than the one before**. Thus, a square root transformation will change a mildly skewed distribution into a symmetric distribution, but a log transformation may be needed if the distribution is more heavily skewed, and so on.
- If the distribution of a variable Y is skewed to the *left*, then raising Y to a power greater than 1 can be helpful.