

Chapter 9: Normal Distributions

The Normal Curve

Areas under Normal Curves

The Central Limit Effect

Checking for Normality

Mound-Shaped Distributions in Nature (page 243)

Many distributions we see are somewhat mound-shaped.

Examples:

- Distribution of Percentage Changes in Daily Chrysler Stock Prices
- Distribution of Heights of 351 Elderly Women
- Distribution of Residuals from Regression Model for NBA Data

The Distribution of Means of Samples (page 245)

Consider the process of shooting free throws in basketball.

In particular, consider taking twenty shots in a row.

Try 20 shots and record the average number of made baskets. If you make 12, your average is $12/20 = 0.6$ for those 20 shots.

Now try 20 shots again. This time you might make 10 and your average is $10/20 = 0.5$.

Try 20 shots again and again and again each time recording the average.

Your average will vary from sample to sample!

What can we say about *the distribution of averages* over many repeats of 20 shots?

By repeated sampling and repeated calculation of the average, we are creating a *distribution of averages* that does not exist “in nature.”

Let's do a few simulations on Minitab to illustrate.

The Normal Curve

The formula:



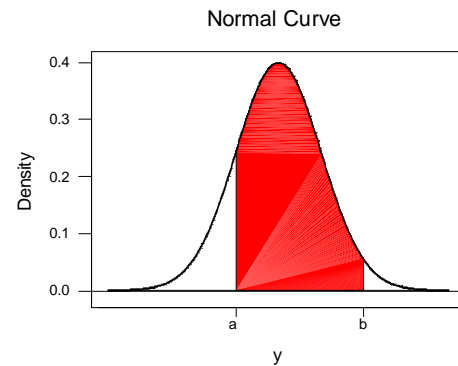
$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{(y-\mu)}{\sigma}\right]^2}$$

Here e is the base of the natural logarithms, π is pi.

μ is the mean of the theoretical distribution and $\sigma > 0$ is the standard deviation.

The normal curve serves as a model for the density “histogram” of many mound-shaped distributions.

As a density histogram the area under the normal curve between any two points a and b on the horizontal axis represents the proportion of measurements within the interval from a to b .



The Standard Normal Curve

When $\mu = 0$ and $\sigma = 1$ the curve is called the standard normal distribution.

If drawn on the same scale, ***all normal curves look alike!***

Another way to say it: all normal distributions are the same if variables are expressed in σ units up or down from the mean μ .

Facts about normal curves: (page 249)

- The curve is symmetric about its mean μ
- The total area between the curve and the $y = 0$ axis is 1

The basis of our earlier statements about mound-shaped distributions are the following facts about the standard normal curve: (page 250 and also page 96)

- About 68% of the distribution's area lies between -1 and $+1$
- About 95% of the distribution's area lies between -2 and $+2$
- Nearly all (99.7%) of the distribution's area lies between -3 and $+3$

Standardized Units (page 251)

If y has a normal distribution with mean μ and standard deviation σ , then the standardized variable

$$z = \frac{y - \mu}{\sigma}$$

has a standard normal distribution.

We say that z is measured in **standardized units**.

z measures the number of standard deviations that y is above (or below) the mean.

Example:

In a normal distribution with mean of 25 and standard deviation of 3, a y -value of 28 corresponds to a z -value of +1. A y -value of 19 corresponds to a z -value of -2.

The Normal Distribution Table

(pages 252 and 253.
Repeated on pages 796 and 797)

With this Table we can get areas under the normal curve for any z -values (or y -values) we want.

Examples:

- **In general, convert word problem to a statement about y -values**
- **Restate the problem in terms of standardized values**
- **Use the normal distribution table to obtain the required area**

Example:

A women must be at least 5' 10" tall to be a member of the *Boston Beanstalks*. What percent of women are eligible?

Let y represent the height of adult women. The collection of y -values has (approximately) a normal distribution with mean 65.5 inches and standard deviation 2.5 inches.

To be eligible for membership y must be 70 or greater: $y \geq 70$

In standardized terms this is:

$$\frac{y - 65.5}{2.5} \geq \frac{70 - 65.5}{2.5}$$

or $z \geq 4.5/2.5 = 1.8$

From the normal distribution table, the area **below** 1.8 is 0.9641.

Thus the area above 1.8 is

$$1 - 0.9641 = 0.0359.$$

So about 3.6% of women are eligible for the club.

Percentiles (page 256)

Sometimes we need to turn things around and go from an area to a standard normal variable z and back to the original units:

$$y = \mu + z\sigma$$

Example:

From the table the 75th percentile of the standard normal distribution is 0.67.

For the women's height distribution, this translates to a 75th percentile of

$$65.5 + (0.67)(2.5) = 65.5 + 1.675 = 67.175$$

or about 5 feet 7 inches.

75% of women are shorter than this.

The Central Limit Effect for Means (page 259)

If \bar{y} is the mean of n process values y_1, y_2, \dots, y_n drawn randomly from a distribution of individuals with mean μ and standard deviation σ , then the distribution of \bar{y} has mean μ and standard deviation σ/\sqrt{n} .

In equations:

$$\begin{aligned}\mu_{\bar{y}} &= \mu \\ \sigma_{\bar{y}} &= \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Recall the simulation of 20 basketball free throws and the successive calculation of averages of the 20 shots. We found that the created distribution of averages was quite mound-shaped. This was no fluke!

The distribution of \bar{y} is approximately a normal distribution.

The larger the sample size n , the better the approximation.

The shape of the distribution of individual values is (almost) irrelevant!

Central Limit Effect (page 259)

Under certain conditions:

The (sampling) distribution of \bar{y} is approximately a normal distribution with

$$\mu_{\bar{y}} = \mu$$
$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

This result is about the **shape** of the distribution—we already knew the facts about about the mean and standard deviation.

- The larger the sample size n , the better the approximation.
- The closer the distribution of individual values is to a normal distribution the better the approximation.
- The approximation is better when applied to the middle portion of the distribution of \bar{y} than to areas in the extremes.

Example:

A company produces potato chips in “10 ounce” bags. The actual fill weight varies according to a distribution with mean 10.2 ounces and standard deviation 0.5 ounces.

The company samples 20 bags, weighs them, and records their average weight. If this sampling procedure were repeated many times, what fraction of the average weights would be less than 10 ounces?

Let \bar{y} be the average in a sample of 20 bags.

By the Central Limit Effect the (sampling) distribution of averages is approximately a normal distribution with mean 10.2 and standard deviation $0.5/\sqrt{20} = 0.5/4.472 = 0.1118$

Standardizing, we need the area below $(10-10.2)/0.1118 = -1.79$

From the normal table this is 0.0367 or about 3.7% of the averages will be below 10 ounces.

The Central Limit Effect for Totals (page 262)

Let $n\bar{y}$ represent the total in a sample of n observations.

Under the same conditions as the Central Limit Effect for Means, the (sampling) distribution of the total, $n\bar{y}$, is approximately a normal distribution with

$$\mu_{n\bar{y}} = n\mu$$

$$\sigma_{n\bar{y}} = \sqrt{n}\sigma$$

Chap. 9 - page 21

We need the fraction of \bar{y} 's that exceed 25,000 tons.

For the total 25,000 tons, we have a standardized value of

$$\frac{25000 - 23000}{800} = \frac{2000}{800} = 2.5$$

By symmetry, the area above 2.5 is the same as the area below -2.5 which we can look-up directly.

This area is 0.0062 so about 6.2% of the loads will have to be lightened.

Chap. 9 - page 23

Examples:

A container ship carries 100 standard-size containers. The weight of the containers varies depending on what is being shipped but can be described by a distribution with mean 230 tons and standard deviation 80 tons.

If the total weight of the containers exceeds 25,000 tons the ship becomes unstable and must be lightened at considerable expense.

Over many loads, what fraction or percentage of loads will have to be lightened?

$n = 100$, $\mu = 230$, and $\sigma = 80$. Also

$$\mu_{n\bar{y}} = n\mu = 100 \times 230 = 23,000$$

$$\sigma_{n\bar{y}} = \sqrt{n}\sigma = \sqrt{100} \times 80 = 800$$

Chap. 9 - page 22

Checking for Normality (page 264)

Rough: Look at distribution display of the data (dotplot, stem-and-leaf, or histogram) Look for a mound-shaped distribution—hard to do!

A little easier if the data are standardized first.

Chap. 9 - page 24

Normal Counts (page 265)

Compare the percentage of data points within 1 or 2 or 3 standard deviations from the mean with the percentage expected for a normal distribution.

Should be about 68% within plus and minus 1 standard deviation of the mean.

About 95% within plus and minus 2 standard deviations of the mean.

Nearly all within plus and minus 3 standard deviations of the mean.

Chap. 9 - page 25

Second Example:

Consider the distribution of residuals from the regression model for the NBA Pts/Min data. Here the mean is zero and the residual standard deviation is 0.075 with 94 observations.

65.96% lie within 1 standard deviation of the mean.

94.68% lie within 2 standard deviations of the mean.

100% lie within 3 standard deviation of the mean.

Chap. 9 - page 27

Examples:

Consider the percentage changes in the daily Chrysler stock price data. These % changes have mean 0.082 and standard deviation 2.066.

Looking at the data and counting we find that 116 of the 166 percentage changes lie within 1 standard deviation of the mean. This is 69.88% compared to 68% for a "true" normal distribution.

Further counting finds 158 out of 166 within 2 standard deviations of the mean. This is 95.18% (compared with 95% for normal).

Finally, all but one observation is within 3 standard deviations of the mean. This is 99.4% (compared with 99.7% for normal).

Chap. 9 - page 26

Normal Scores (page 266)

For n observations the normal scores are ideal (standard) normal numbers.

They divide the data axis so that the normal areas are all equal (to $1/(n+1)$).

We match up the ordered observations with the ordered normal scores and plot the pairs to get a **normal probability plot**.

If the data come from (any) normal distribution, the normal probability plot should resemble a straight line.

The straighter the plot, the more evidence we have for a normal distribution.

Nonnormal data will plot with some curvature. The more curve, the more evidence against normality.

Chap. 9 - page 28