Chapter 70

The normal random variable and Z-scores

Questions answered in this chapter:

- What are the properties of the normal random variable?
- How do I use Excel to find the probabilities for the normal random variable?
- Can I use Excel to find the percentiles for normal random variables?
- Why is the normal random variable appropriate in many real-world situations?
- What are Z-scores?

Answers to this chapter's questions

What are the properties of the normal random variable?

In Chapter 67, "An introduction to random variables," you learned that continuous random variables can be used to model quantities such as the following:

- The price of Microsoft stock one year from now
- The market share for a new product
- The market size for a new product
- The cost of developing a new product
- A newborn baby's weight
- A person's IQ

Remember that if a discrete random variable (such as the sales of Big Macs during 2018) can assume many possible values, you can approximate the value by using a continuous random variable as well. As I described in Chapter 67, any continuous random variable X has a probability density function (PDF). The PDF for a continuous random variable is a nonnegative function with the following properties (a and b are arbitrary numbers):

- The area under the PDF is 1.
- The probability that $X<a$ equals the probability that $X\leq a$. This probability is represented by the

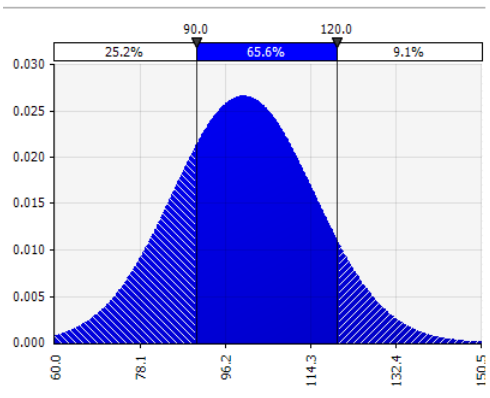
area under the PDF, to the left of *a*.

- The probability that $X>b$ equals the probability that $X\geq b$. This probability is given by the area under the PDF to the right of *b*.
- The probability that $a<X<b$ equals the probability that $a\leq X\leq b$. This probability is the area under the PDF between *a* and *b*.

Thus, the area under a continuous random variable's PDF represents probability. Also, the larger the value of the density function at *X*, the more likely the random variable will take on a value near *X*. For example, if the density function of a random variable at 20 is twice the density function of the random variable at 5, then the random variable is twice as likely to take on a value near 20 than a value near 5.

For a continuous random variable, the probability that *X* equals *a* will always equal 0. For example, some people are from 5.99999 feet through 6.00001 feet tall, but no person can be exactly 6 feet tall. This explains why you can replace the less-than sign (**<**) with the less-than-or-equal-to sign (**≤**) in the probability statements.

Figure 70-1 displays the PDF for *X*=IQ of a randomly chosen person. The area under this PDF is 1. If you want to find the probability that a person's IQ is less than or equal to 90 (0.252), you simply find the area to the left of 90. If you want to find the probability that a person's IQ is between 90 and 120 (0..656), you find the area under the PDF between 90 and 120. If you want to find the probability that a person's IQ is more than 120 (0.091), you find the area under the density function to the right of 120.



F70xx01

FIGURE 70-1 The probability density function for IQs.

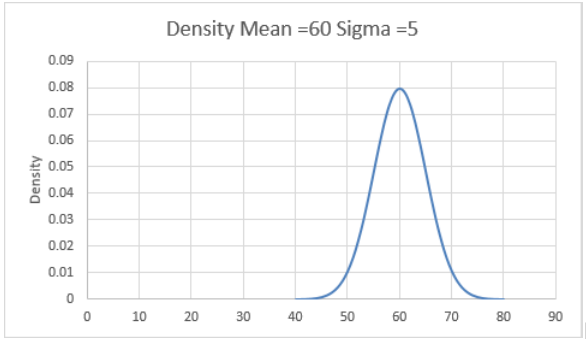
F70xx01: This figure shows the probability density function for normally distributed IQs. and various probabilities for IQs

Actually, the density plotted in Figure 70-1 is an example of the *normal random variable*. The normal random variable is specified by its mean and standard deviation. IQs follow a normal random

variable with μ=100 and σ=15. This is the PDF displayed in Figure 70-1. The normal random variable has the following properties:

- The most likely value of a normal random variable is μ (as indicated by the PDF peaking at 100 in Figure 70-1).
- As the value x of the random variable moves away from μ, the probability that the random variable is near x sharply decreases.
- The normal random variable is symmetric about its mean. For example, IQs near 80 are as likely as IQs near 120.
- A normal random variable has 68 percent of its probability within σ (sigma, representing the standard deviation) of its mean, 95 percent within 2σ of its mean, and 99.7 percent within 3σ of its mean. These measures should remind you of the rule of thumb I described in Chapter 42, "Summarizing data by using descriptive statistics." In fact, the rule of thumb is based on the assumption that data is "sampled" from a normal distribution, which explains why the rule of thumb does not work as well when the data fails to exhibit a symmetric histogram.

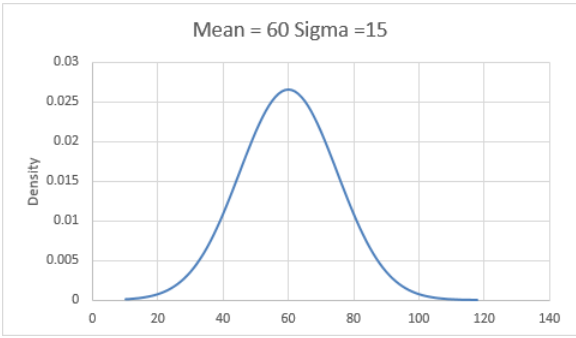
For a larger σ, a normal random variable is more spread out about its mean. This pattern is illustrated in Figures 70-2 and 70-3. (See the worksheet Sigma Of 5 And 15 in the file Normalexamples.xlsx.)



F70xx02

FIGURE 70-2 A normal random variable PDF, with a mean equal to 60 and a standard deviation equal to 5.

F70xx02: This figure shows the density for a normal random variable with a mean equal to 60 and a standard deviation equal to 5.



F70xx03

FIGURE 70-3 A normal random variable PDF, with a mean equal to 60 and a standard deviation equal to 15.

F70xx03: This figure shows the density for a normal random variable with a mean equal to 60 and a standard deviation equal to *15*.

How do I use Excel to find the probabilities for the normal random variable?

Consider a normal random variable X with a mean μ and standard deviation σ. Suppose for any number x, you want to find the probability that $X \leq x$, which is called the *normal cumulative function*. To use Microsoft Excel 2016 to find the probability that $X \leq x$, enter the formula =NORM.DIST(x,μ,σ,1). Of course, the fourth argument of *1* could be replaced by *True*.

The argument *1* tells Excel to compute the normal cumulative. If the last argument of the function is *0*, Excel returns the actual value of the normal random variable PDF. Beginning with Excel 2010, statistical functions (like NORM.DIST) have been modified or redesigned completely to provide better accuracy than their counterparts (like NORMDIST) in previous versions of Excel.

You can use the NORM.DIST function to answer many questions concerning normal probabilities. You can find examples in the Normal worksheet in the file Normalexamples.xlsx, which is shown in Figure 70-4, and in the following three scenarios.

	B	C	D
2		Prob	
3	IQ<90	0.25249254	=NORM.DIST(90,100,15,1)
4	90<IQ<120	0.65629624	=NORM.DIST(120,100,15,1)-NORMDIST(90,100,15,1)
5	IQ>120	0.09121122	=1-NORM.DIST(120,100,15,1)
6			
7	99 %ile Prozac	71.6317394	=NORM.INV(0.99,60,5)
8	10% Bloomington income	19747.5875	=NORM.INV(0.1,30000,8000)

F70xx04

FIGURE 70-4 Calculating the normal probability.

F70xx04: This figure shows examples of computations with the normal random variable.

- **What fraction of people have an IQ of less than 90?** Let X equal the IQ of a randomly chosen person. Then you seek the probability that $X<90$, which is equal to the probability that $X \leq 90$. Therefore, you can enter into cell C3 of the Normal worksheet the formula

=**NORM.DIST(90,100,15,1)**, and Excel returns 0.252. Thus, 25.2 percent of all people have an IQ less than 90.

- **What fraction of all people have IQs between 90 and 120?** When finding the probability that $a \le X \le b$, you use the form (area under the normal density function to the left of *b*)–(area under normal density function to the left of *a*). Thus, you can find the probability that $a \le X \le b$ by entering the formula =**NORM.DIST(b,μ,s,1)–NORM.DIST(a,μ,s,1)**. You can answer the question about IQs from 90 through 120 by entering into cell C4 of the worksheet Normal the formula =**NORM.DIST(120,100,15,1)–NORM.DIST(90,100,15,1)**. Excel returns the probability 0.656, so 65.6 percent of all people have an IQ from 90 through 120.

- **What fraction of all people have IQs of at least 120**? To find the probability that $X \ge b$, note that the probability that $X \ge b$ equals 1–probability $X < b$. You can compute the probability that $X \ge b$ by entering the formula 1–NORM.DIST(b,μ,σ,1). You seek the probability that $X \ge 120$. This equals 1–probability $X < 120$. You enter in cell C5 of the worksheet Normal the formula =1–NORM.DIST(120,100,15,1). Excel returns 0.091, so you know that 9.1 percent of people have an IQ of at least 120.

Can I use Excel to find the percentiles for normal random variables?

Consider a given normal random variable *X* with mean (μ) and standard deviation (σ). In many situations, you want to answer questions such as the following:

- A drug manufacturer believes that next year's demand for its popular antidepressant will be normally distributed, with a mean equal to 60 million days of therapy (DOT) and σ (standard deviation) equal to 5 million DOT. How many units of the drug should be produced this year if the company wants to have only a 1 percent chance of running out of the drug?

- Family income in Bloomington, Indiana, is normally distributed, with a mean equal to $30,000 and σ equal to $8,000. The poorest 10 percent of all families in Bloomington are eligible for federal aid. What should the aid cutoff be?

In the first example, you want to determine the ninety-ninth percentile of demand for the antidepressant. That is, you seek the number *x* so that there is only a 1 percent chance that demand will exceed *x* and a 99 percent chance that demand will be less than *x*. In the second example, you want the tenth percentile of family income in Bloomington. That is, you seek the number *x* so that there is only a 10 percent chance that the family income will be less than *x* and a 90 percent chance that the family income will exceed *x*.

Suppose you want to find the *p*th percentile (expressed as a decimal) of a normal random variable *X* with a mean (μ) and a standard deviation (σ). Simply enter the formula =**NORM.INV(p,μ,σ)**. This formula returns a number *x* so that the probability that $X \le x$ equals the specified percentile. You now can solve the examples. You'll find these exercises on the Normal worksheet in the file Normalexamples.xlsx.

For the drug manufacturing example, let *X* equal the annual demand for the drug. You want a value *x* so that the probability that $X \geq x$ equals 0.01 or the probability that $X < x$ equals 0.99. Again, you seek the ninety-ninth percentile of demand, which you find (in millions) by entering in cell C7 the formula =**NORM.INV(0.99,60,5)**. Excel returns 71.63, so the company must produce 71,630,000 DOT. This assumes, of course, that the company begins the year with no supply of the drug on hand. If, for example, they had a beginning inventory of 10 million DOT, they would need to produce 61,630,000 DOT during the current year.

To determine the cutoff for federal aid, if *X* equals the income of a Bloomington family, you seek a value of *x* so that the probability that $X \leq x$ equals 0.10, or the tenth percentile of Bloomington family income. You find this value with the formula NORM.INV(0.10,30000,8000). Excel returns $19,747.59, so aid should be given to all families with incomes less than $19,749.59.

Why is the normal random variable appropriate in many real-world situations?

A well-known mathematical result called the *central limit theorem (CLT)* indicates that if you add together many (usually at least 30 is sufficient) independent random variables, their sum is normally distributed. This result holds true even if the individual random variables are not normally distributed. Many quantities (such as measurement errors) are created by adding together many independent random variables, which explains why the normal random variable occurs often in the real world. The following are some other situations in which you can use the CLT.

- The total demand for pizzas during a month at a supermarket is normally distributed, even if the daily demand for pizzas is not.
- The amount of money you win if you play craps 1,000 times is normally distributed, even though the amount of money you win on each individual play is not.

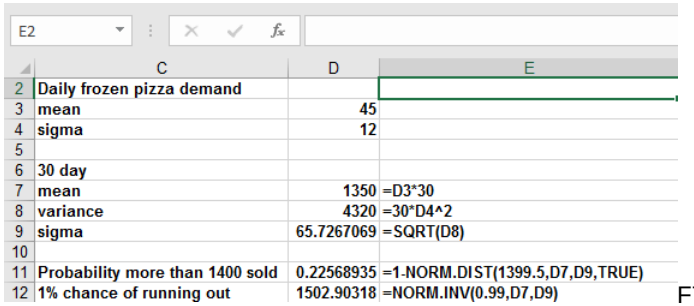
Another important mathematical result shows how to find the mean, variance, and standard deviations of sums of independent random variables. If you are adding together independent random variables $X_1, X_2, \ldots, Xn$ where the mean $X_i = \mu_i$, and the standard deviation $X_i = \sigma_i$, then the following are true:

1. Mean $(X_1+X_2+\ldots X_n)=\mu_1+\mu_2+\ldots\mu_n$
2. Variance $(X_1+X_2+\ldots X_n)$= G70xx01
3. Standard deviation $(X_1+X_2+\ldots X_n)$= G70xx02

Note that 1 is true even when the random variables are not independent. By combining 1 through 3 by means of the central limit theorem, you can solve many complex probability problems, such as modeling the demand for pizza over a 30-day period.

As an example of the central limit theorem, suppose daily demand for pizza at your local pizza parlor is not normally distributed but has a mean of 45 and standard deviation of 12. Suppose demand on different days is independent. What is the chance during a 30-day period that you will sell

at least 1,400 pizzas? If you order pizza dough at the beginning of the 30-day period, how much dough should you order to ensure you have a 1 percent chance of running out of pizza dough? See the example in Figure 70-5 and the Central Limit worksheet of the file Normalexamples.xlsx.



	C	D	E
2	Daily frozen pizza demand		
3	mean	45	
4	sigma	12	
5			
6	30 day		
7	mean	1350	=D3*30
8	variance	4320	=30*D4^2
9	sigma	65.7267069	=SQRT(D8)
10			
11	Probability more than 1400 sold	0.22568935	=1-NORM.DIST(1399.5,D7,D9,TRUE)
12	1% chance of running out	1502.90318	=NORM.INV(0.99,D7,D9)

F70xx05

FIGURE 70-5 Using the central limit theorem.

F70xx05: This figure shows an application of the central limit theorem.

Even though the daily demand for frozen pizzas is not normally distributed, you know from the CLT that the 30-day demand for frozen pizzas is normally distributed. Given this, 1 through 3 above imply the following:

- From 1, the mean of a 30-day demand equals 30(45)=1,350.
- From 2, the variance of a 30-day demand equals $30(12)^2$=4,320.
- From 3, the standard deviation of a 30-day demand equals G70xx03

Thus, the 30-day demand for pizzas can be modeled following a normal random variable with a mean of 1,350 and a standard deviation of 65.73. In cell D11, I compute the probability that at least 1,400 pizzas are sold as the probability that the normal approximation is at least 1,399.5, with the formula =1–NORM.DIST(1399.5,D7,D9,TRUE). (Note that a demand of 1,399.6, for example, would round up to 1,400.) I find the probability that demand in a 30-day period for at least 1,400 pizzas is 22.6 percent.

The number of pizzas that you must stock to have only a 1 percent chance of running out of pizzas is just the ninety-ninth percentile of the demand distribution. You determine the ninety-ninth percentile of the demand distribution (1,503) in cell D12 by using the formula =NORM.INV(0.99,D7,D9). Therefore, at the beginning of a month, you should bring your stock of pizzas up to 1,503 if you want only a 1 percent chance of running out of pizzas.

What are Z-scores?

When given a data set, you want to quickly determine how unusual a data point is. The most common approach is to standardize each data point by computing a *Z-score* the following way:

```
(Data value – Mean of data)/(Standard deviation of data)
```


Essentially a Z-score measures the number of standard deviations the data point differs from average. Since 95 percent of observations from a normal random variable are within two standard deviations of average, any data point with a Z-score exceeding 2 in absolute value is deemed an outlier.

As an example of the computation of Z-scores, the file Superbowlspreads.xlsx (see Figure 70-6) contains for the years 1967–2016 the Las Vegas prediction for the number of points by which the favorite would win the game and the number of points by which the favorite won. For example, in 2016 the Carolina Panthers were favored by 5 and lost by 14, so they performed -14 – 5 = -19 points relative to the point spread. We call the difference between the point spread and actual game outcome the *residual* or error for the game. You can compute Z-scores for the residuals as a measure for each game of how unexpected or unusual the game's outcome was. Proceed as follows:

- Compute the residual for each game by copying from cell C6 to C7:C55 the formula =E6-D6.
- In cell C4, compute the mean residual (-0.39) with the formula =AVERAGE(C6:C55).
- In cell I4, compute the standard deviation (16.10) of the errors with the formula =STDEV(C6:C55).

To compute the Z-score for each observation, copy from G6 to G7:G55 the formula =(C6-C4)/I4.

	C	D	E	F	G	H	I
2					Z Score		
3	Mean				Mean	2.9E-17	Stdev
4	-0.39				Sigma	1	16.1043
5	residual or error	Favorite	Result	Year	Z Score		
6	-19	5	-14	2016	-1.155588822		
7	4	0	4	2015	0.272597256		
8	-37.5	2.5	-35	2014	-2.30434719		
9	-7.5	4.5	-3	2013	-0.441495783		
10	-6.5	2.5	-4	2012	-0.379400736		
11	3	3	6	2011	0.210502209		
12	-19	5	-14	2010	-1.155588822		
13	-3	7	4	2009	-0.162068072		
14	-15	12	-3	2008	-0.907208635		
15	5	7	12	2007	0.334692303		
16	7	4	11	2006	0.458882396		
17	-4	7	3	2005	-0.224163119		

F70xx06

FIGURE 70-6 The Z-scores for the Super Bowl outcomes.

F70xx06: This figure computes the Z-scores for the Super Bowl outcomes.

You find, for example, that in 2014, the favored Denver Broncos performed 2.3 standard deviations below average (the largest outlier in Super Bowl history!), while in 1990 the favored San Francisco

49ers performed 2.07 standard deviations better than average. I used conditional formatting to highlight both these outliers.

For any data set, the mean Z-score will be 0, and the standard deviation of the Z-scores will equal 1. As shown in cells H3 and H4, the mean Z-score is 0, and the standard deviation of the Z-scores is 1.

Problems

1. Suppose you can set the mean number of ounces of soda that is put into a can. The actual number of ounces has a standard deviation of 0.05 ounces. Answer the following questions:
   - If you set the mean at 12.03 ounces, and a soda can is acceptable if it contains at least 12 ounces, what fraction of cans are acceptable?
   - What fraction of cans have less than 12.1 ounces?
   - To what should you set the mean if you want at most 1 percent of your cans to contain at most 12 ounces? Hint: Use the Goal Seek command.
2. The annual demand for a drug is normally distributed with a mean of 40,000 units and a standard deviation of 10,000 units. Answer the following questions:
   - What is the probability that annual demand is from 35,000 through 49,000 units?
   - If you want to have only a 5 percent chance of running out of the drug, at what level should you set annual production?
3. The probability of winning a game of craps is 0.493. If I play 10,000 games of craps and bet the same amount on each game, what is the probability that I'm ahead? Begin by determining the mean and standard deviation of the profit in one game of craps. Then use the central limit theorem.
4. The weekly sales of Volvo's Cross Country station wagons are normally distributed with a mean of 1,000 and standard deviation of 250. Answer the following questions:
   - What is the probability that from 400 through 1,100 station wagons are sold during one week?
   - There is a 1 percent chance that fewer than what number of station wagons is sold during a week?
5. My time for swimming 100 yards follows a normal random variable with a mean that equals 51 seconds and a standard deviation of 1.5 seconds. If I swim the race in under 49 seconds, I will win the race. What is the chance I win the race?

6. The federal government wants to charge a Medicare surcharge to Americans 65 or older whose income ranks in the top 2 percent of all American family incomes. If the American family incomes follow a normal random variable with a mean that equals $60,000 and a standard deviation that equals $20,000, what should be the cutoff for the surcharge?

7. The voltage held by a voltage regulator follows a normal random variable with a mean that equals 200 volts and a standard deviation that equals 5 volts. A regulator meets the specifications if the regulator can hold a voltage between 185 and 210 volts. What fraction of the regulators meet the specifications?

8. Texas wants to give welfare checks to the 5 percent poorest of all Texas families. If the family income in Texas follows a normal random variable with a mean that equals $50,000 and a standard deviation that equals $15,000, what income level should be the cutoff for welfare?

9. A rod meets its specifications if its diameter is between 0.98 and 1.02 inches. The mean diameter of a rod is 1 inch. What should the standard deviation of rods equal for 99 percent of rods to meet the specifications?

10. The file Problem10data.xlsx gives the number of touchdown passes thrown by each quarterback in 2014. Determine the Z-score for each quarterback, and use conditional formatting to highlight all the outliers.

11. The number of bottles of ice tea sold by a supermarket on a Monday follows a normal random variable with a mean that equals 100 and a standard deviation that equals 12. On a given Monday, what is the chance that the supermarket will sell at least 105 bottles of ice tea? Fill in the blank: There is a 10 percent chance that _____ or fewer bottles of ice tea will be sold on a Monday.

12. Assume the mean daily percentage change in the Dow Jones Index is 1 percent, with a standard deviation of 1.5 percent. Assume there are 252 trading days in a year. What is the chance the Dow Jones Index increases by 20 percent or more during a year?

13. I have 500 potential customers who come to my candy store each day. Seventy percent of the customers buy no pieces of candy, 15 percent buy one piece, 10 percent buy two pieces, and 5 percent buy three pieces. What is the chance that I sell at least 280 pieces of candy in a day?

