3.1.5 68-95-99.7 rule

Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z table.

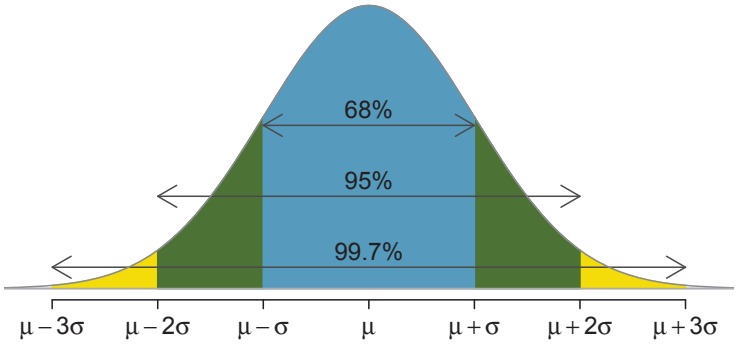

Figure 3.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

⊙ **Exercise 3.22** Use the Z table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3, standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = -1$ and $Z = 1$, which should have an area of about 0.68. Similarly there should be an area of about 0.95 between $Z = -2$ and $Z = 2$.[17]

It is possible for a normal random variable to fall 4, 5, or even more standard deviations from the mean. However, these occurrences are very rare if the data are nearly normal. The probability of being further than 4 standard deviations from the mean is about 1-in-30,000. For 5 and 6 standard deviations, it is about 1-in-3.5 million and 1-in-1 billion, respectively.

⊙ **Exercise 3.23** SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$. (a) About what percent of test takers score 900 to 2100? (b) What percent score between 1500 and 2100?[18]

3.2 Evaluating the normal approximation

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males. While using a normal model can be extremely convenient and helpful, it is important to remember normality is

[17] First draw the pictures. To find the area between $Z = -1$ and $Z = 1$, use the normal probability table to determine the areas below $Z = -1$ and above $Z = 1$. Next verify the area between $Z = -1$ and $Z = 1$ is about 0.68. Repeat this for $Z = -2$ to $Z = 2$ and also for $Z = -3$ to $Z = 3$.

[18] (a) 900 and 2100 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 900 and 2100. (b) Since the normal model is symmetric, then half of the test takers from part (a) ($\frac{95\%}{2} = 47.5\%$ of all test takers) will score 900 to 1500 while 47.5% score between 1500 and 2100.

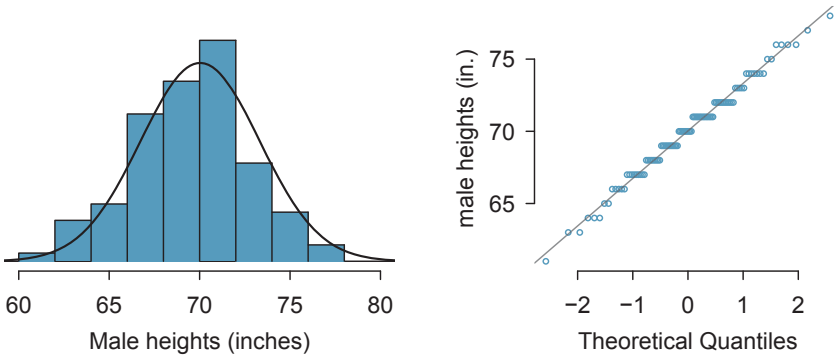

Figure 3.10: A sample of 100 male heights. The observations are rounded to the nearest whole inch, explaining why the points appear to jump in increments in the normal probability plot.

always an approximation. Testing the appropriateness of the normal assumption is a key step in many data analyses.

3.2.1 Normal probability plot

Example 3.15 suggests the distribution of heights of US males is well approximated by the normal model. We are interested in proceeding under the assumption that the data are normally distributed, but first we must check to see if this is reasonable.

There are two visual methods for checking the assumption of normality, which can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 3.10. The sample mean $\bar{x}$ and standard deviation s are used as the parameters of the best fitting normal curve. The closer this curve fits the histogram, the more reasonable the normal model assumption. Another more common method is examining a **normal probability plot**.[19], shown in the right panel of Figure 3.10. The closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model. We outline the construction of the normal probability plot in Section 3.2.2

● **Example 3.24** Three data sets of 40, 100, and 400 samples were simulated from a normal distribution, and the histograms and normal probability plots of the data sets are shown in Figure 3.11. These will provide a benchmark for what to look for in plots of real data.

The left panels show the histogram (top) and normal probability plot (bottom) for the simulated data set with 40 observations. The data set is too small to really see clear structure in the histogram. The normal probability plot also reflects this, where there are some deviations from the line. However, these deviations are not strong.

The middle panels show diagnostic plots for the data set with 100 simulated observations. The histogram shows more normality and the normal probability plot shows a better fit. While there is one observation that deviates noticeably from the line, it is not particularly extreme.

[19] Also commonly called a **quantile-quantile plot**.

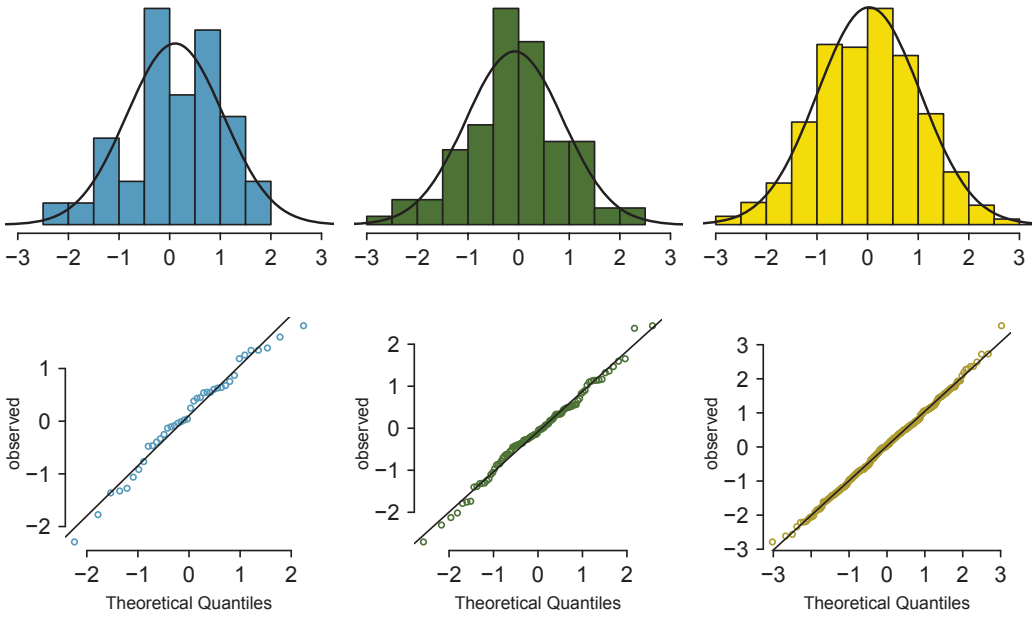

Figure 3.11: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

The data set with 400 observations has a histogram that greatly resembles the normal distribution, while the normal probability plot is nearly a perfect straight line. Again in the normal probability plot there is one observation (the largest) that deviates slightly from the line. If that observation had deviated 3 times further from the line, it would be of much greater concern in a real data set. Apparent outliers can occur in normally distributed data but they are rare.

Notice the histograms look more normal as the sample size increases, and the normal probability plot becomes straighter and more stable.

Example 3.25 Are NBA player heights normally distributed? Consider all 435 NBA players from the 2008-9 season presented in Figure 3.12.[20]

We first create a histogram and normal probability plot of the NBA player heights. The histogram in the left panel is slightly left skewed, which contrasts with the symmetric normal distribution. The points in the normal probability plot do not appear to closely follow a straight line but show what appears to be a "wave". We can compare these characteristics to the sample of 400 normally distributed observations in Example 3.24 and see that they represent much stronger deviations from the normal model. NBA player heights do not appear to come from a normal distribution.

[20] These data were collected from http://www.nba.com.

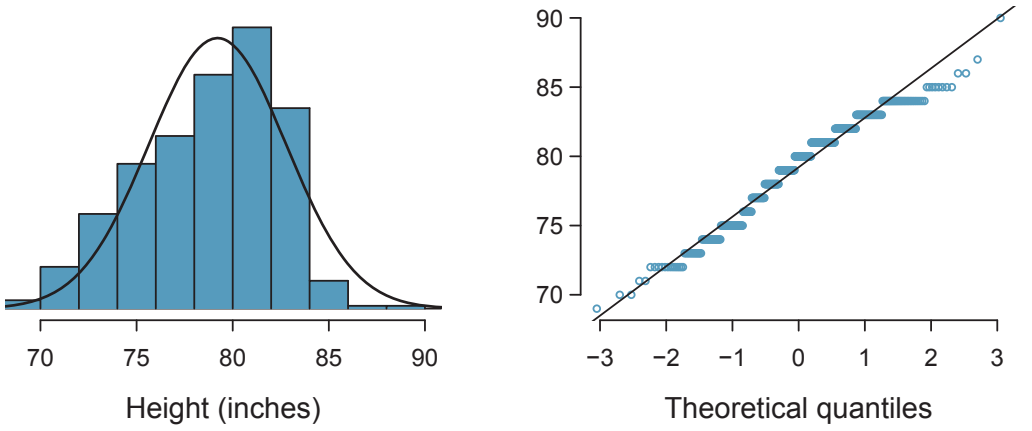

Figure 3.12: Histogram and normal probability plot for the NBA heights from the 2008-9 season.

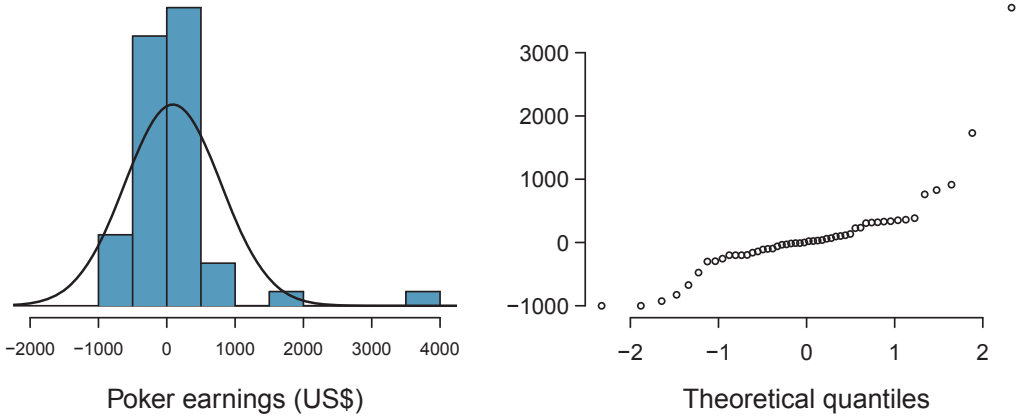

Figure 3.13: A histogram of poker data with the best fitting normal plot and a normal probability plot.

Example 3.26 Can we approximate poker winnings by a normal distribution? We consider the poker winnings of an individual over 50 days. A histogram and normal probability plot of these data are shown in Figure 3.13.

The data are very strongly right skewed in the histogram, which corresponds to the very strong deviations on the upper right component of the normal probability plot. If we compare these results to the sample of 40 normal observations in Example 3.24, it is apparent that these data show very strong deviations from the normal model.

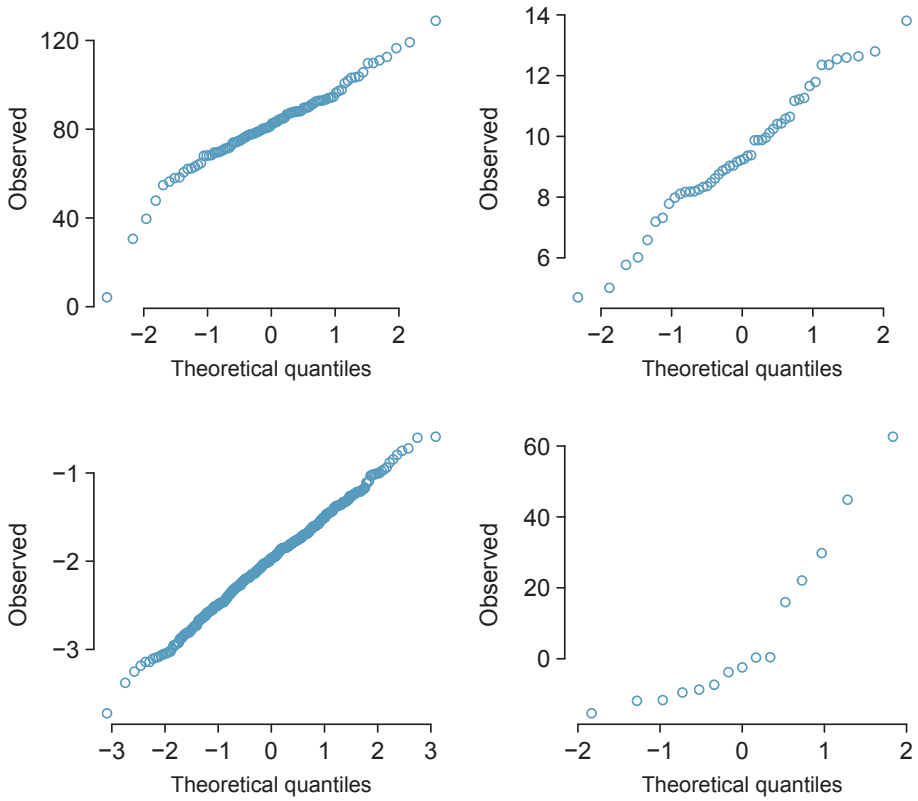

Figure 3.14: Four normal probability plots for Exercise 3.27.

⊙ **Exercise 3.27** Determine which data sets represented in Figure 3.14 plausibly come from a nearly normal distribution. Are you confident in all of your conclusions? There are 100 (top left), 50 (top right), 500 (bottom left), and 15 points (bottom right) in the four plots.[21]

⊙ **Exercise 3.28** Figure 3.15 shows normal probability plots for two distributions that are skewed. One distribution is skewed to the low end (left skewed) and the other to the high end (right skewed). Which is which?[22]

[21] Answers may vary a little. The top-left plot shows some deviations in the smallest values in the data set; specifically, the left tail of the data set has some outliers we should be wary of. The top-right and bottom-left plots do not show any obvious or extreme deviations from the lines for their respective sample sizes, so a normal model would be reasonable for these data sets. The bottom-right plot has a consistent curvature that suggests it is not from the normal distribution. If we examine just the vertical coordinates of these observations, we see that there is a lot of data between -20 and 0, and then about five observations scattered between 0 and 70. This describes a distribution that has a strong right skew.

[22] Examine where the points fall along the vertical axis. In the first plot, most points are near the low end with fewer observations scattered along the high end; this describes a distribution that is skewed to the high end. The second plot shows the opposite features, and this distribution is skewed to the low end.

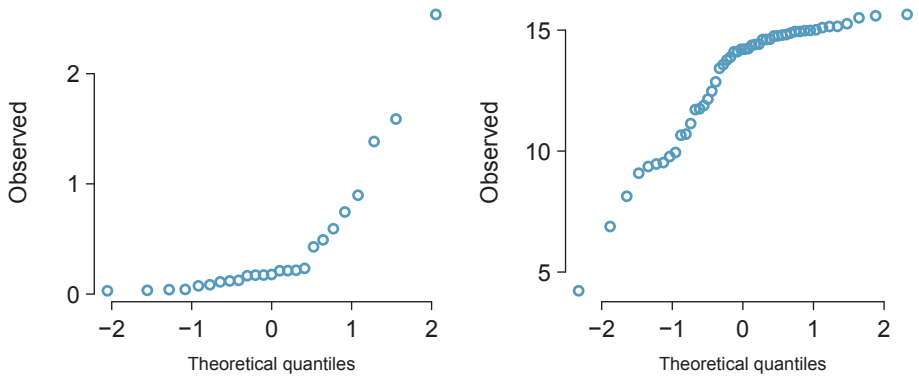

Figure 3.15: Normal probability plots for Exercise 3.28.

3.2.2 Constructing a normal probability plot (special topic)

We construct a normal probability plot for the heights of a sample of 100 men as follows:

(1) Order the observations.

(2) Determine the percentile of each observation in the ordered data set.

(3) Identify the Z score corresponding to each percentile.

(4) Create a scatterplot of the observations (vertical) against the Z scores (horizontal).

If the observations are normally distributed, then their Z scores will approximately correspond to their percentiles and thus to the z_i in Table 3.16.

Observation i	1	2	3	$\cdots$	100
x_i	61	63	63	$\cdots$	78
Percentile	0.99%	1.98%	2.97%	$\cdots$	99.01%
z_i	−2.33	−2.06	−1.89	$\cdots$	2.33

Table 3.16: Construction details for a normal probability plot of 100 men's heights. The first observation is assumed to be at the 0.99^{th} percentile, and the z_i corresponding to a lower tail of 0.0099 is −2.33. To create the plot based on this table, plot each pair of points, (z_i, x_i).

Caution: z_i correspond to percentiles
The z_i in Table 3.16 are *not* the Z scores of the observations but only correspond to the percentiles of the observations.

Because of the complexity of these calculations, normal probability plots are generally created using statistical software.

3.3 Geometric distribution (special topic)

How long should we expect to flip a coin until it turns up `heads`? Or how many times should we expect to roll a die until we get a `1`? These questions can be answered using the geometric distribution. We first formalize each trial – such as a single coin flip or die toss – using the Bernoulli distribution, and then we combine these with our tools from probability (Chapter 2) to construct the geometric distribution.

3.3.1 Bernoulli distribution

Stanley Milgram began a series of experiments in 1963 to estimate what proportion of people would willingly obey an authority and give severe shocks to a stranger. Milgram found that about 65% of people would obey the authority and give such shocks. Over the years, additional research suggested this number is approximately consistent across communities and time.[23]

Each person in Milgram's experiment can be thought of as a **trial**. We label a person a **success** if she refuses to administer the worst shock. A person is labeled a **failure** if she administers the worst shock. Because only 35% of individuals refused to administer the most severe shock, we denote the **probability of a success** with $p = 0.35$. The probability of a failure is sometimes denoted with $q = 1 - p$.

Thus, `success` or `failure` is recorded for each person in the study. When an individual trial only has two possible outcomes, it is called a **Bernoulli random variable**.

Bernoulli random variable, descriptive
A Bernoulli random variable has exactly two possible outcomes. We typically label one of these outcomes a "success" and the other outcome a "failure". We may also denote a success by `1` and a failure by `0`.

TIP: "success" need not be something positive
We chose to label a person who refuses to administer the worst shock a "success" and all others as "failures". However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

`0 1 1 1 1 0 1 1 0 0`

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{0+1+1+1+1+0+1+1+0+0}{10} = 0.6$$

[23] Find further information on Milgram's experiment at www.cnr.berkeley.edu/ucce50/ag-labor/7article/article35.htm.

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable.[24]

Bernoulli random variable, mathematical
If X is a random variable that takes value 1 with probability of success p and 0 with probability $1 - p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

3.3.2 Geometric distribution

Example 3.29 Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict the worst shock.[25] If the probability a person will *not* give the most severe shock is still 0.35 and the subjects are independent, what are the chances that she will stop the study after the first person? The second person? The third? What about if it takes her $n - 1$ individuals who will administer the worst shock before finding her first success, i.e. the first success is on the n^{th} person? (If the first success is the fifth person, then we say $n = 5$.)

The probability of stopping after the first person is just the chance the first person will not administer the worst shock: $1 - 0.65 = 0.35$. The probability it will be the second person is

$$\begin{aligned} &P(\text{second person is the first to not administer the worst shock}) \\ &\quad = P(\text{the first will, the second won't}) = (0.65)(0.35) = 0.228 \end{aligned}$$

Likewise, the probability it will be the third person is $(0.65)(0.65)(0.35) = 0.148$.

If the first success is on the n^{th} person, then there are $n - 1$ failures and finally 1 success, which corresponds to the probability $(0.65)^{n-1}(0.35)$. This is the same as $(1 - 0.35)^{n-1}(0.35)$.

---

[24] If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned} \mu = E[X] &= P(X = 0) \times 0 + P(X = 1) \times 1 \\ &= (1 - p) \times 0 + p \times 1 = 0 + p = p \end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned} \sigma^2 &= P(X = 0)(0 - p)^2 + P(X = 1)(1 - p)^2 \\ &= (1 - p)p^2 + p(1 - p)^2 = p(1 - p) \end{aligned}$$

The standard deviation is $\sigma = \sqrt{p(1-p)}$.

[25] This is hypothetical since, in reality, this sort of study probably would not be permitted any longer under current ethical standards.

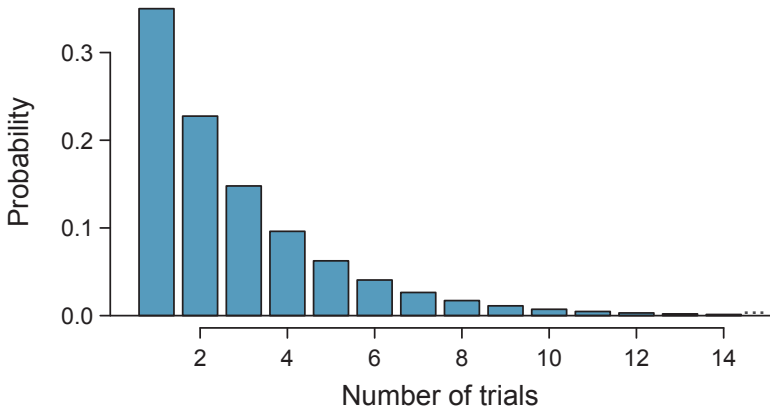

Figure 3.17: The geometric distribution when the probability of success is $p = 0.35$.

Example 3.29 illustrates what is called the geometric distribution, which describes the waiting time until a success for **independent and identically distributed (iid)** Bernoulli random variables. In this case, the *independence* aspect just means the individuals in the example don't affect each other, and *identical* means they each have the same probability of success.

The geometric distribution from Example 3.29 is shown in Figure 3.17. In general, the probabilities for a geometric distribution decrease **exponentially** fast.

While this text will not derive the formulas for the mean (expected) number of trials needed to find the first success or the standard deviation or variance of this distribution, we present general formulas for each.

Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p \tag{3.30}$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.31}$$

It is no accident that we use the symbol μ for both the mean and expected value. The mean and the expected value are one and the same.

The left side of Equation (3.31) says that, on average, it takes $1/p$ trials to get a success. This mathematical result is consistent with what we would expect intuitively. If the probability of a success is high (e.g. 0.8), then we don't usually wait very long for a success: $1/0.8 = 1.25$ trials on average. If the probability of a success is low (e.g. 0.1), then we would expect to view many trials before we see a success: $1/0.1 = 10$ trials.

⊙ **Exercise 3.32** The probability that an individual would refuse to administer the worst shock is said to be about 0.35. If we were to examine individuals until we found one that did not administer the shock, how many people should we expect to check? The first expression in Equation (3.31) may be useful.[26]

● **Example 3.33** What is the chance that Dr. Smith will find the first success within the first 4 people?

This is the chance it is the first ($n = 1$), second ($n = 2$), third ($n = 3$), or fourth ($n = 4$) person as the first success, which are four disjoint outcomes. Because the individuals in the sample are randomly sampled from a large population, they are independent. We compute the probability of each case and add the separate results:

$$\begin{aligned} &P(n = 1, 2, 3, \text{ or } 4) \\ &\quad = P(n = 1) + P(n = 2) + P(n = 3) + P(n = 4) \\ &\quad = (0.65)^{1-1}(0.35) + (0.65)^{2-1}(0.35) + (0.65)^{3-1}(0.35) + (0.65)^{4-1}(0.35) \\ &\quad = 0.82 \end{aligned}$$

There is an 82% chance that she will end the study within 4 people.

⊙ **Exercise 3.34** Determine a more clever way to solve Example 3.33. Show that you get the same result.[27]

● **Example 3.35** Suppose in one region it was found that the proportion of people who would administer the worst shock was "only" 55%. If people were randomly selected from this region, what is the expected number of people who must be checked before one was found that would be deemed a success? What is the standard deviation of this waiting time?

A success is when someone will **not** inflict the worst shock, which has probability $p = 1 - 0.55 = 0.45$ for this region. The expected number of people to be checked is $1/p = 1/0.45 = 2.22$ and the standard deviation is $\sqrt{(1-p)/p^2} = 1.65$.

⊙ **Exercise 3.36** Using the results from Example 3.35, $\mu = 2.22$ and $\sigma = 1.65$, would it be appropriate to use the normal model to find what proportion of experiments would end in 3 or fewer trials?[28]

The independence assumption is crucial to the geometric distribution's accurate description of a scenario. Mathematically, we can see that to construct the probability of the success on the n^{th} trial, we had to use the Multiplication Rule for Independent Processes. It is no simple task to generalize the geometric model for dependent trials.

---

[26] We would expect to see about $1/0.35 = 2.86$ individuals to find the first success.

[27] First find the probability of the complement: $P(\text{no success in first 4 trials}) = 0.65^4 = 0.18$. Next, compute one minus this probability: $1 - P(\text{no success in 4 trials}) = 1 - 0.18 = 0.82$.

[28] No. The geometric distribution is always right skewed and can never be well-approximated by the normal model.