**The Monday blues. Or are you more likely to have a heart attack on some days of the week than others?**

**Introduction**

A study done in Augsburg, Germany (Willich et al., 1993) looked at which days of the week people had heart attacks. They wanted to know whether or not heart attacks are distributed equally across the days of the week in the population. One subgroup of heart attack victims that they looked at were those who were employed. The researchers found that 884 heart attacks from this subgroup were distributed across the seven days of the week as shown in Table 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Day** | **Sun** | **Mon** | **Tue** | **Wed** | **Thur** | **Fri** | **Sat** |
| **Number of heart attacks** | 106 | 160 | 123 | 115 | 141 | 107 | 132 |

**Table 1:** A distribution of the frequency of heart attacks for 884 people that were employed at the time of their heart attack.

These data represent days of the week for 884 heart attacks in some larger population. We think of these 884 heart attacks as a sample. Although this isn’t a random sample, this group probably is representative of some larger population. (We will discuss this in more detail later.)

As a start, let’s first compare to one that is distributed evenly. What would this type of distribution look like? Well, because there were 884 total heart attacks we would expect about 884/7 126.3 heart attacks each day. In the distribution in Table 1, Tuesday’s frequency is closest to what would be expected, and Monday’s is the farthest from what would be expected.

Another way to rephrase the question that we want to answer is, if heart attacks were distributed evenly in the larger population, how likely would it be for us to get a distribution as “off” as this? Is that fairly likely to happen or is that very unlikely to happen?

**Typical Variation or Noise**

Let’s move away from our heart attack example for a moment to get an idea of what typical variation looks like in a simpler situation. Suppose I rolled a fair six-sided die 120 times. What do you expect to happen? You might say since 120/6 = 20, you would expect to get 20 ones, 20 twos, 20 threes and so on. But do you really expect to get exactly 20 of each of these? Not necessarily. You should expect to get close to 20 of each. And if you rolled the die 120 more times you should expect to get a distribution different than the first; but again all the frequencies will be close to 20. But now the question is, what does it mean to get close to 20?

To help get you to see what close to 20 might look like, we rolled a fair six-sided die 120 times and the number of times each face landed up is shown in Table 2 in the row labeled Trial #1. You can see that close to 20 means 1, 2 or 3 away for all of the outcomes except for the frequency of 5s. This frequency of 14 is 6 away from what would be expected. But again, these results come from 120 rolls a fair die, so these frequencies show the typical variation you could see or might expect. We repeated the process of rolling a die 120 times four more times, with the results shown in the last four rows of Table 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number showing** | **1** | **2** | **3** | **4** | **5** | **6** |
| **Trial #1**  | 22 | 23 | 23 | 21 | 14 | 17 |
| **Trial #2** | 13 | 27 | 19 | 23 | 21 | 17 |
| **Trial #3** | 16 | 21 | 23 | 22 | 17 | 21 |
| **Trial #4** | 25 | 18 | 25 | 21 | 18 | 13 |
| **Trial #5** | 26 | 22 | 18 | 17 | 16 | 21 |

**Table 2:** Five differentdistributions of the number of times each face occurred in 120 rolls of a fair die.

Again, these show the type of variation you might see under this fair process. Sometimes this type of variation is called *noise*. What we want to know, back in our heart attack data, are we just seeing this sort of noise or is the observed variation from what is expected more extreme than that. If the variation is more extreme, we might say there is something more going on than typical variation or noise. We might say that the extra variation from what is expected is not just noise but is a *signal*.

**Hypotheses**

Now let’s get back to our heart attack example. There are two possible reasons why the distribution of heart attacks deviates so much from one that would be equally distributed. One is that heart attacks in the population are equally distributed across the seven days and the variation we are seeing here is just “noise” in our data. The other reason is that heart attacks are not distributed evenly in the population so the amount of variation we see in the sample data is more than we would expect to see if heart attacks were evenly distributed in the population.

The two reasons we stated above are what we call hypotheses. In this example, the hypothesis that heart attacks are distributed equally in the population is called the null hypothesis and often the notation H0 is used to represent this hypothesis. The word null means zero, hence the subscript zero is used in the notation. You can think of this zero as no change or in this case no change or difference from what would be expected if heart attacks were equally distributed in the population. Formally, we could write out the null hypothesis as:

**H0: Heart attacks are distributed evenly across the days of the week for employed people in the population.**

The other hypothesis is basically the opposite, or that heart attacks are not distributed equally. We call this hypothesis the alternative hypothesis and often the notation Ha is used to represent this hypothesis. Some people will refer to this as the research hypothesis. Formally, we could write out the alternative hypothesis as:

**Ha: Heart attacks are not distributed evenly across the days of the week for employed people in the population.**

In testing these hypotheses, we assume that the null hypothesis is true, or that all the heart attacks from the population are distributed evenly. Then we determine the likelihood (or probability) of getting a distribution as “off” as the one we obtained. So we now need a way to calculate this probability.

**Statistic**

To determine whether the variation from what is expected is more than just noise, we first need a way to measure it. In statistical language, we need a *statistic*. A statistic is a measure of some attribute in a sample. Simple statistics are things like a sample mean or a sample proportion. In this case, we want to compare several categories, so we need a statistic that is a bit more complicated. We need something that will measure how far away the distribution of heart attacks is from what is expected if there was an equal distribution of heart attacks across the seven days of the week. Table 3 shows our original data for the 884 heart attacks, though we have renamed this row the **observed** number of heart attacks because this is the sample we observed. We have also added a row for the **expected** number of heart attacks. This means these numbers are expected if we have a perfectly equal distribution of heart attacks across the seven days. Because 884/3 ≈ 126.3, we have put that number in for each day. Don’t worry that this is not a whole number. Even though we can’t have a fraction of a heart attack, we can use this in developing our statistic.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Day** | **Sun** | **Mon** | **Tue** | **Wed** | **Thur** | **Fri** | **Sat** |
| **OBSERVED number of heart attacks** | 106 | 160 | 123 | 115 | 141 | 107 | 132 |
| **EXPECTED number of heart attacks** | 126.3 | 126.3 | 126.3 | 126.3 | 126.3 | 126.3 | 126.3 |

**Table 2:** The observed and expected number of heart attacks for each day of the week.

Now, how will we measure how far apart these two distributions are from one another? Subtracting each expected frequency from its corresponding observed frequency might seem like a good place to start. Then what should we do with the seven differences we get? We need our statistic to be a single number not seven numbers. Perhaps we should add them up or maybe average them. Let’s give this a try by first just adding up the differences.

What happened here? Why did we get such a small sum of all these differences? What happened was the positive and negative differences cancelled each other out. In fact, if we didn’t round the expected frequencies to 126.3, our sum of the differences would be exactly 0. If we simply sum the differences, they will sum to zero every time, no matter what data set we start with. What else can we do so that we don’t get these positive and negative differences that cancel each other out?

Hopefully you are thinking either take the absolute values of the differences or square the differences. Doing either of these will result in all positive values. Taking absolute values might seem the most logical thing to do because it is simpler, however squaring will work better. In a little bit we will go through a process to determine how unlikely our statistic would be if in fact the null hypothesis was true. (Remember that the null hypothesis was that the population distribution of heart attacks is equally distributed.) In fact, we are going to show you two different ways to do this. For one of those ways, using absolute values does not work well, but squaring will. We’ll point this out to you later when it arises.

We could just square the differences before adding them up and use that sum of squared differences as our statistic. Just doing that will result in a fine statistic, but we can make it better. We will also divide each squared difference by the expected frequency. We do this because the sample size matters. For example, doesn’t it seem like 106 and 126.3 are closer together than 6 and 26.3 even though their differences (or squared differences) are the same? After all, 26.3 is more than four times as large as 6 while 126.3 is nowhere near four times 106. By dividing by the expected frequencies we are standardizing the statistic so that different distributions with different sample sizes give statistics that can be compared on the same scale. We make these squared differences into relative squared differences.

Okay, let’s put this all together. The statistic we have been describing is called the **chi-square statistic** which has the symbol **χ2**. (Chi is a Greek letter and is pronounced ki like hi!) The formula for the χ2 statistic is the following.

Note that the Σ in the formula is commonly called a summation symbol. It just means to add up all the terms. (This symbol is also another Greek letter, in this case it is an upper-case sigma.) Also note in the formula that the χ2-statistic can never be negative since the numerator (because it is squared) and the denominator can never be negative. If all the observed frequencies were exactly the same as the expected (like our observed distribution was equally distributed) we would end up with a χ2-statistic of 0. Can you see why? The smallest a χ2-statistic can be is 0 and the further you move the observed frequencies away from the expected frequencies, the larger the χ2-statistic becomes.

Let’s calculate the χ2-statistic for our data.

This gives us a χ2-statistic of 18.27. Does this indicate that our distribution of heart attacks is far away from one that is equally distributed? To answer this, we need to know the values of typical χ2-statistics if there is just noise in the data and not a signal. This way we can determine whether our observed chi-square value is fairly small (just noise) or fairly large (evidence of a signal).

**A simulated distribution of typical χ2-statistics (with just some noise)**

We now need to determine typical values of the χ2-statistics that would occur if our null hypothesis was true. This will let us see the noise (or variability) that these statistics have and help us determine whether or our statistic of 18.27 fits in with this noise. It is probably a good time to remind you of the null hypothesis, so we write it again below.

**H0: Heart attacks are distributed evenly across the days of the week for employed people in the population.**

We want to take the 884 heart attacks and randomly distribute them across the 7 days of the week. To do this we can think of what we did earlier with rolling a die 120 times. This is the same process. Imagine a 7-sided die where each side had a day of the week on it instead of a number. For each of the 884 heart attacks, we would roll the die to randomly determine which day it occurred. The complete process to develop a χ2-statistic under the assumption that the null is true would be to:

1. Roll the die, note day of the week that is facing up.
2. Repeat this for a total of 884 times.
3. Develop a distribution of these 884 heart attacks (similar to our observed data back in Table 1).
4. Calculate the χ2-statistic from the randomized data.
5. Repeat this many times (like 1,000) so we can see typical values of the statistic under a true null.

We don’t have a 7-sided die and if we did, you can see that this process would be quite time consuming. However, technology can come to our aid. We have a computer applet that can simulate this process very quickly. We did this simulation in our applet five times and came up with the five distributions and their χ2-statistics shown in Table 3.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Day** | **Sun** | **Mon** | **Tue** | **Wed** | **Thur** | **Fri** | **Sat** | χ2 |
| **Trial #1**  | 127 | 124 | 133 | 125 | 125 | 122 | 128 | 0.60 |
| **Trial #2**  | 132 | 142 | 124 | 107 | 128 | 130 | 121 | 5.55 |
| **Trail #3** | 119 | 127 | 118 | 142 | 108 | 124 | 146 | 8.69 |
| **Trial #4** | 126 | 131 | 129 | 118 | 121 | 131 | 128 | 1.20 |
| **Trial #5** | 123 | 128 | 130 | 117 | 126 | 135 | 125 | 1.52 |

**Table 3:** Five simulated distributions giving how many heart attacks occurred on each day of the week for the 884 heart attacks under the assumption that they are distributed evenly in the population. The accompanying χ2-statistics are also given.

With our five simulated distributions and the resulting χ2-statistics we haven’t seen anything as extreme as our observed χ2-statistic of 18.27. We should do some more repetitions, so we can better see the noise in the resulting statistics. We had our applet do 95 more simulations for a total of 100. Each resulting χ2-statistic is plotted in a graph as shown in Figure 1.



**Figure 1:** A distribution of 100 simulated χ2-statistics under the assumption that the null hypothesis is true.

It doesn’t appear from the graph of the 100 simulated χ2-statistics shown in Figure 1 that any of the simulated statistics are as large as our observed value of 18.27. At this point it looks like a statistic as extreme as 18.27 would rarely happen by chance if the null hypothesis is true. We should still probably do more simulations. We had our applet do 900 more simulations for a total of 1,000. The resulting χ2-statistics are collected and shown in the distribution in Figure 2.

**Figure 2:** A distribution of 1,000 simulated χ2-statistics under the assumption that the null hypothesis is true.

From the distribution shown in Figure 2 we can see that a χ2-statistic as extreme as 18.27 (our observed statistic) is quite unlikely to occur. In only 6 out of the 1,000 repetitions did we get something at least as extreme. So what does this tell you? Is 18.27 part of the noise or is it sending us a signal that something different is going on? In other words, does 18.27 seem to be far enough out in the tail of this distribution that you would say it is unlikely to occur by chance?

The standard that is typically used to say something is unlikely to occur by chance is a probability of less than or equal to 0.05 or 5%. From our applet output, we can see that we estimate that a χ2-statistic at least as extreme as 18.27 happened about 0.6% of the time. Because this is less than 5%, we would conclude that there is strong evidence this observed statistic arose from something other than just noise. In other words we would say we have strong evidence against the null hypothesis that the population distribution is distributed evenly and say we have strong evidence for the alternative hypothesis that the population distribution is not distributed evenly.

Could a statistic like 18.27 happen just by chance if the population was distributed evenly? Yes, it happened 6 times in our 1,000 repetitions. However, a probability of 6/1,000 means it would be very unlikely to get a χ2-statistic at least as large as 18.27 if the population is equally distributed. Therefore it is more plausible that the population is not equally distributed.

**Theory-based p-values**

The probability of 0.006 that was given to us in the applet is called a p-value. More generally, a **p-value** is the probability of obtaining a value of the statistic at least as extreme as the observed statistic when the null hypothesis is true. We obtained our p-value of 0.006 through a simulation. If we repeated the simulation again we might get a slightly different p-value. (In fact we just repeated the simulation and obtained a p-value of 0.008). It shouldn’t be too concerning that we might get slightly different p-values, because they all would be fairly close together and should all be telling us the same thing in terms of strength of evidence.

Before computers could quickly do simulations like we just did, the way to get p-values had to rely on theory-based methods. In theory-based methods the simulated distribution that we obtained in the applet is predicted using mathematical formulas. The theory-based distribution that we would need to use in this example is appropriately called a χ2-distribution. The applet we were using earlier can both show a picture of the theory-based χ2-distribuiton and compute the theory-based p-value using this distribution. All this is shown in Figure 3.

**Figure 3:** A theory-based χ2-distribution is overlaid on our simulated χ2-distribution and a theory-based p-value is calculated in the applet.

Notice that we get a very similar p-value using theory-based methods. As long as the sample size is large enough this will happen. A theory-based χ2-goodness-of-fit test (this is the name of the test we have just used) to give a valid p-value the expected frequencies should all be at least 5. Remember that our expected counts were all 126.3, so our sample size was definitely large enough to get a valid theory-based p-value.

Remember back when we were developing our statistic and we decided to square the differences between the observed and expected frequencies? If we only use simulation techniques to get a p-value, using absolute values instead of squaring would have been fine. Simulation-based techniques can use a wide variety of statistics to obtain a p-value. The theory-based techniques can be a bit more finicky, however. Using absolute values could result in a simulated distribution in which the heights of the bars go up, then down, then back up again. They don’t tend to be as smooth as when we square the differences. This makes it hard to create a mathematical model that will fit a simulated distribution based on absolute values. Squaring doesn’t lead to this kind of problem.

**Scope of conclusion**

With a p-value of 0.006 (or 0.0056) we have strong evidence against the null and hence strong evidence that heart attacks are not distributed equally across the seven days of the week in the population. But what exactly is this population we keep referring to?

If our data were obtained from a random sample of all heart attacks from some specific population (like all German citizens in 1990) our population would be the group from which our sample was drawn. However, in many studies, like this one, the sample is not a random sample. In this case, the researchers collected data from 13 hospitals in and around Augsburg, Germany from 1985 to 1990. In particular, they looked at all the heart attacks that occurred in that period and found which day of the week they occurred. Furthermore, for this specific data set, they focused on heart attacks from people that were employed during that time.

Since this is not a random sample can we generalize our results to a larger population? We can, though we might do so with some hesitation as to exactly what that population is. Even though this wasn’t a random sample it is probably representative of some larger populations. We could probably be comfortable generalizing these results to all of Germany (or what was commonly called West Germany at the time of this study) since heart attacks in Augsburg are probably very similar to those around the rest of the country. Can we generalize to all of Europe? All industrialized nations? The entire world? Just how far you go depends on what populations you think are quite similar (in terms of having heart attacks) to those in and around Augsburg, Germany in the late 1980s.

What we need to emphasize here is that these heart attacks were from people that were employed. We saw that the largest frequency of heart attacks took place on Mondays. Might this also be true for those that are not employed? Are Mondays “special” for that population? We will have you look at this question with another data set in just a bit.

**Review**

Let’s review what we just did.

* We (or the researchers) wanted to see whether or not heart attacks were distributed evenly across the seven days of the week.
* In setting up a test of significance we first wrote out the null and alternative hypotheses:
	+ H0: Heart attacks are distributed evenly across the days of the week for employed people in the population.
	+ Ha: Heart attacks are not distributed evenly across the days of the week for employed people in the population.
* The researchers collected data and from that we computed a χ2-statistic of 18.27. This statistic measures how much the observed distribution differs from a distribution of the same sample size that has the heart attacks distributed equally across the seven days of the week.
* We then used an applet to find a p-value of 0.006. What does this number mean? Remember that we simulated χ2-statistics for this scenario under the assumption that the heart attacks are distributed evenly in the population. Only 6 of the 1,000 simulations resulted in χ2-statistics that were at least as large as the 18.27 that was observed in the study. This means that if heart attacks were distributed evenly across the seven days of the week in the population, it is very unlikely that we would get a χ2-statistic as large or larger than we did. (P-values less than 0.05 are considered small and ours was much smaller than 0.05.)
* We also found a theory-based p-value of 0.0056 which guides us to the same conclusion as did the simulation-based p-value.
* We then concluded that based on a p-value of 0.006, we have strong evidence that heart attacks in the population are not distributed evenly across the seven days of the week.

**Instructions for using the Analyzing One-way Tables Applet**

The applet we used can be found at: <http://www.rossmanchance.com/applets/GOF.html>. To find a p-value with our heart attack data do the following.

1. Open up the applet and put the heart attack table of data like it is shown below. Then click on Use Table and you should then see a bar graph of the data like the one shown below.



1. Below the table of data, click on the arrow in the box next to Statistic and change it to the χ2-statistic as shown.



1. Click on the Show Sampling Options check box on the right side of the applet. If heart attacks were distributed evenly, there should be 1/7 of them on each day. So for the hypothesized probabilities of heart attacks you need to enter 1/7 written as a decimal 7 times separated by commas as shown below.



1. Put 1000 in the Number of Samples box and click on the Sample button. The applet has now created 1000 distributions, calculated its χ2-statistic of each of them, and then plotted each statistic in a graph.



1. Put the χ2-statistic (which you should see on the lower left side of the applet) into the box as shown below and click on Count. You should now see a simulation-based p-value in red. It may not match the one shown below exactly, but it should be similar.

1. To calculate the theory-based p-value simply click on the Overlay Chi-square distribution box and you will see the theory-based p-value (which should be exactly the same as shown below) as well as a smooth distribution over the top of the simulated distribution.

**What about heart attacks for those not employed?**

We’ve seen that there is strong evidence that heart attacks are not distributed evenly across the seven days of the week for those that are employed. In particular, our data showed that Monday had the largest number by quite a bit. Since this data set came from people that are employed, maybe Mondays are particularly difficult for this group. What about people that are not employed? Is there the same Monday effect? We have data on that as well.

The same researchers found that there were 1,191 heart attacks among those that were not employed. The heart attacks in this group were distributed as shown in Table 4. Again we ask the same question as earlier. Do we have strong evidence that heart attacks are not distributed evenly across the days of the week in this population? To answer this, we ask you to answer the following questions.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Day** | **Sun** | **Mon** | **Tue** | **Wed** | **Thur** | **Fri** | **Sat** |
| **Number of heart attacks** | 168 | 180 | 157 | 169 | 180 | 169 | 168 |

**Table 4:** A distribution of the frequency of heart attacks for 1,191 people that were not employed at the time of their heart attack.

1. Write the null and alternative hypotheses for this test.
2. Put the table of data into the applet. Based on the distribution of heart attacks, do you think there will be strong evidence that heart attacks are not distributed evenly across the seven days of the week? Why or why not?
3. Calculate the χ2-statistic using the applet. How does this statistic compare with the 18.27 from the employed data? What does this difference mean?
4. Calculate both a simulation-based p-value and theory-based p-value using the applet. Are these the types of numbers you would expect based on what the distribution looked like and the size of the χ2-statistic?
5. Write out a conclusion in the context of this situation. (Note: When we have a small p-value we say we have strong evidence against the null hypothesis or strong evidence for the alternative. When we get a large p-value we don’t have strong evidence of anything. Instead, we say we do not have strong evidence against the null or we do not have strong evidence for the alternative in general.)

|  |  |
| --- | --- |
| **Employed Data Table**Day CountSun 106Mon 160Tue 123Wed 115Thu 141Fri 107Sat 132 | **Unemployed Data Table**Day CountSun 168Mon 180Tue 157Wed 169Thu 180Fri 169Sat 168 |