Transmission of Genetic Information

Introduction

If you are already familiar with the structure of these exercises, read the Introduction first.

> **Note**
>
> Reminder: Save your work regularly.

> **Important**
>
> If you are using a Mac, we recommend that you use either Chrome or Firefox to complete these exercises. Some of the default settings in Safari prevent these exercises from running.

Contact information

If you have questions about these exercises, please contact Dr. Kevin Middleton (middletonk@missouri.edu) or drop by Tucker 224.

Learning Objectives

The learning objectives for this exercise are:

- Calculate the probability of a particular gamete being produced from an individual, assuming independent assortment.
- Calculate the probability of a particular genotype in the offspring, given independent assortment and random fertilization between two individuals.
- Design genetic crosses to provide information about genes, alleles, and gene functions.
- Use a chi-squared test to determine how well data from a genetic cross fits theoretical predictions.
- Explain how sample size is related to the certainty of a chi-squared test.

Review of Inheritance

When gametes are produced, each gamete receives a single copy of each gene. Because each of the F_1 parents has two copies, there are four possible alleles that can be passed on (for example two copies of *D* and two copies of *d* in

Figure 1). A Punnett square can be used to enumerate the possible genotype combinations that result from the proposed cross. By joining the genes from each parent, the resulting possible offspring genotypes can be determined. These genotypes can be converted into phenotypes if the nature of dominance of the alleles in known (Figure 1). In this example the *D* allele for the "tall" phenotype is dominant relative to the *d* allele.

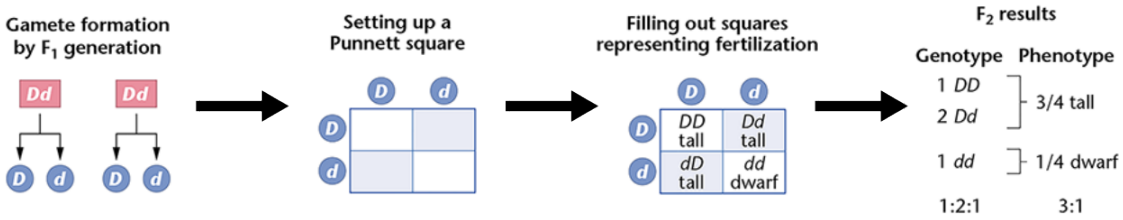


Figure 1: The process of gamete formation leads to predictable genotypic and phenotypic proportions in the F_2 generation. Image modified from Klug et al. (2019)

The 1:2:1 genotypic ratio and 3:1 phenotypic ratio in the heterozygote cross example above represent theoretical probabilities for the distributions of genotypes and phenotypes. These ratios can be tested statistically to determine if a sample deviates significantly from the expected proportions. The exercises that follow will allow you explore the statistics of genetics and give you some practice testing whether observations follow the predicted patterns.

Proportions and Probabilities

When thinking about genotypes and phenotypes, it can be useful to sometimes think of counts as *proportions* and sometimes as *probabilities.* The 1:2:1 genotypic ratio above can be thought of as

1. *Proportions*: 1/4 *DD*, 2/4 *Dd*, and 1/4 *dd*
2. *Percentages*: 25% *DD*, 50% *Dd*, and 25% *dd*
3. *Probabilities*: 0.25 *DD*, 0.5 *Dd*, and 0.25 *dd*

All are equal and mean the same thing. In some contexts, using one vs. another makes more sense. This set of proportions and probabilities all sum to one and represents the full range of possible combinations.

Probabilities are very predictable in the long run but are often unpredictable in the short run. Consider only one gamete being produced, say a sperm. Each sperm that is produced from a heterozygote *Dd* can carry either the *D* or *d* allele.

We can simulate the production of a single sperm with the following code. First we create an object that holds the possible *alleles*: “D” and “d”. The code *sample(allele, size = 1)* randomly samples 1 of the possible alleles. Run this code a few times to see the first few gametes produced.

```
allele <- c("D", "d")
sample(allele, size = 1)
```

```
[1] "d"
```

You probably noticed that, quite often, sequences of two or more *D* or *d* were produced randomly.

The code block below repeats this sampling process 10 times and counts up the numbers of *D* and *d* gametes produced. The function *replicate()* generates the samples and *table()* counts *D*s and *d*s.

How many D and d gametes do you predict from a total of 10? How confident are you in your answer?

There should be about 5 D and 5 d, but there will probably be variation. 6 and 4 (or 4 and 6) would be pretty common. 7 and 3 (3 and 7) less common, etc.

```
replicate(n = 10, sample(allele, size = 1)) |>
  table()
```

```
  D   d
244 256
```

Run the code a few times to generate new samples of 10 gametes. Revisit your prediction.

Now gradually increase the number of gametes produced by changing *n = 10* to *n = 50*, *n = 100*, etc. You can choose larger numbers, but avoid going over about 10,000 or it will run very slowly.

What happens to the relative counts of D and d as n *increases? What does this tell you about our ability to predict the exact counts either when sample sizes are small or when sample sizes are large? In general, how often do you see exactly 50% D and 50% d?*

As n increases, on average the difference between the count of D and of d decreases (as a percentage of the total sample). With small sample sizes. large deviations are possible (like 8 and 2 above), but when the sample size increases those deviations become relatively smaller. So, a difference of 6 in a sample size of 10 is large, but a difference of 6 in a sample size of 500 is not large at all,

relatively speaking. It will be much harder to predict counts when the sample size is small compared to when the sample size is large.

Blood types in humans

To explore the concepts of independent assortment and probabilities, we will use the blood-typing system that is used to classify human blood types based on the expression of different antigens on the surface of red blood cells (Figure 2). Blood types are important for transfusion medicine, where the mixing of incompatible blood types can have fatal results.

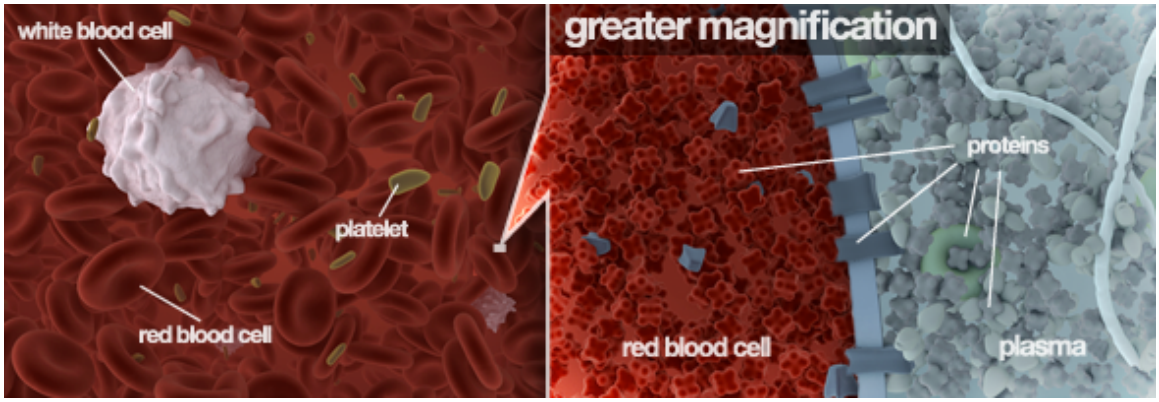


Figure 2: Three of the components of blood are white blood cells (involved in the immune response), platelets (involved in blood clotting), and red blood cells, which primarily carry oxygen. Image from learn.genetics

Blood typing in humans primarily uses the ABO system, although almost 30 other blood typing systems exist, which focus on other aspects of red blood cell structure and physiology.

A and/or B refer to the presence of A and/or B antigens of the surface of the red blood cells. O denotes the absence of both A and B antigens. The American Red Cross has a website with animations about what blood types are compatible with other blood types. In general, A must be matched with A and B with B or a cross-reaction will take place. Because type O blood lacks both A and B antigens, it is considered a “universal donor”. And because type AB blood has both A and B antigens, it is considered a “universal recipient”.

Figure 3 shows the Punnett square for human blood types. Each parent can provide alleles for A, B, or no antigens (O), leading to a complex 1:2:1:1:1:2:1 ratio.

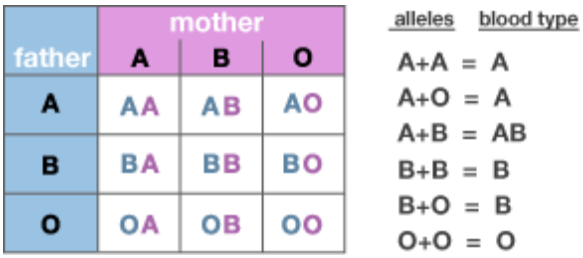


Figure 3: The possible gamete genotypes and combinations for human ABO alleles and associated blood types. Individuals don't have all of A, B and O – this table illustrates the possibilities for individual gametes. Image from learn.genetics

The right side of Figure 3 shows the blood types (i.e., phenotypes) associated with each of the possible genotypes. Although there are 7 possible genotypes, there are only 4 possible blood types: A, B, AB, and O.

Looking at the figure, you may have wondered (1) why there is no type AO or BO blood and (2) how type AB blood is produced.

1. Because the O allele has neither A nor B antigen, an individual with A+O alleles will have type A blood and an individual with B+O will have type B blood (the same is true for O+A and O+B). So we can say that A and B are **dominant** with respect to O.
2. Because the A and B alleles produce A and B antigens, respectively, an individual with A+B or B+A will have type AB blood. The A and B alleles are thus **co-dominant**.

Rh Factor

Another important feature of red blood cells that is important for transplant medicine is the Rh factor. The genetics of Rh factor is much more complicated than the ABO system, with approximately 50 different proteins involved (Dean 2005). Because a few proteins are most commonly involved, this system can be summarized as *Rh+* and *Rh-* (Rh-positive and Rh-negative) without going into all the details.

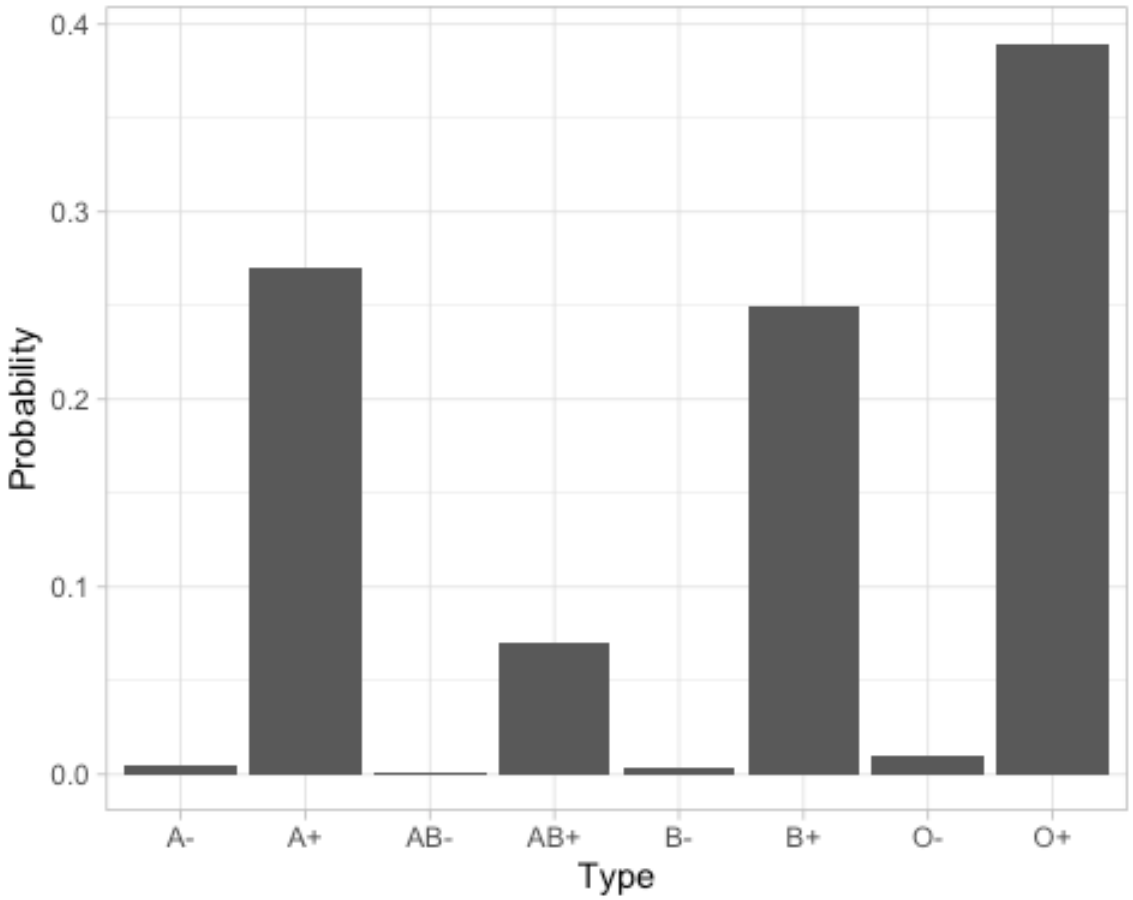
ABO type and Rh factor are determined by separate sets of genes, meaning that Rh factor is inherited separately from ABO type. Thus the rules of **Independent Assortment** apply. Each ABO blood type can be associated with Rh+ or Rh-.

Probabilities of blood types in a population

The distributions of A, B, AB, and O blood types as well as Rh factors differs among populations. In a hypothetical population, blood type probabilities for all the combinations of ABO types and Rh status are summarized in the following table.

ABO	Rh Factor	Probability
A	-	0.005
A	+	0.270
AB	-	0.001
AB	+	0.070
B	-	0.004
B	+	0.250
O	-	0.010
O	+	0.390

Looking at tables of numbers is difficult, so we will plot the data. The relative proportions of the eight different blood types can be shown on a bar chart, where the height of the bar is proportional to the probability of an individual having a specific ABO/Rh combination:



Visualizing the table above in this way really emphasizes the relative rarity of Rh-negative blood types in this population.

Note that sum of all the probabilities is 1. When studying the range of probabilities for an event (here a person's blood type), it is important that all the probabilities add up to 1, which means that we have accounted for all the possible outcomes.

```
sum(c(0.005, 0.270, 0.001, 0.070, 0.004, 0.250, 0.010, 0.390))
```

```
[1] 1
```

sum() adds up the values passed to it. *c()* makes a collection of numbers, which here is the set of probabilities.

If a large number of people have their blood typed, the proportions of people with each blood type should be approximately equal to the overall proportions in the table and figure above.

Combining probabilities

Because of independent assortment, each of the 8 possible ABO types + Rh factor is independent of each other, which is to say that no one person can have multiple blood types. We can use the rules of addition to determine the probability that a randomly selected individual has a certain blood type. The first few are provided for you. See if you can figure out the rest (here we are just using R as a calculator).

Type O blood

```
0.39 + 0.01
```

```
[1] 0.4
```

Type AB blood

```
0.07 + 0.001
```

```
[1] 0.071
```

Type A *or* type B blood with any Rh factor

```
0.005 + 0.27 + 0.004 + 0.25
```

```
[1] 0.529
```

Any ABO type with Rh+

```
0.27 + 0.07 + 0.25 + 0.39
```

```
[1] 0.98
```

Not Type AB blood

```
1 - (0.001 + 0.07)
```

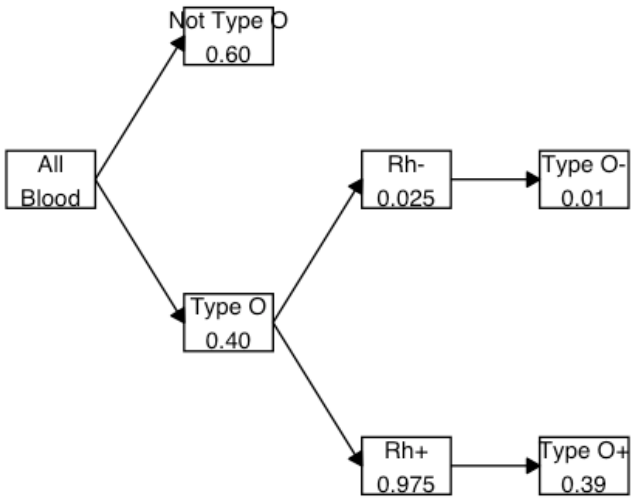
```
[1] 0.929
```

These are all examples of the *Addition Rule* for probabilities. For independent events, you can add the individual probabilities to get the overall probability for a set of events. Generally these take the form of A *or* B – like rolling a 1 or a 2 from a single die roll. Each has a probability of 1/6, so the probability of 1 *or* 2 is 1/6 + 1/6 = 2/6 = 1/3.

For conditional events, the kinds of which result from independent assortment, we have to multiply two probabilities (the *Multiplication Rule* for probabilities). The flowchart below shows the calculation of the probabilities for O+ and O- blood types.

Starting from the left, we have all blood types. In this population, 40% have type O blood (including both Rh+ and Rh-), and 60% have all the other types combined.

In this example, we will only follow the type O blood paths. Among those with type O blood, 97.5% have the Rh+ phenotype, and the remaining 2.5% have the Rh- phenotype. Because ABO type and Rh factor assort independently, the separate probabilities are multiplied to get the final probabilities. $0.4 \times 0.025 = 0.01$ and $0.4 \times 0.975 = 0.39$. You can confirm that these match the probabilities in the table.



We could repeat this process for types A, B, and AB, and make a giant chart with all the possibilities.

Sampling from a population

The probabilities in the blood type table above represent what are known as “long-run” probabilities, meaning that if we sample more and more people (up to the entire population size), the proportions that we observe will more closely match the expected probabilities.

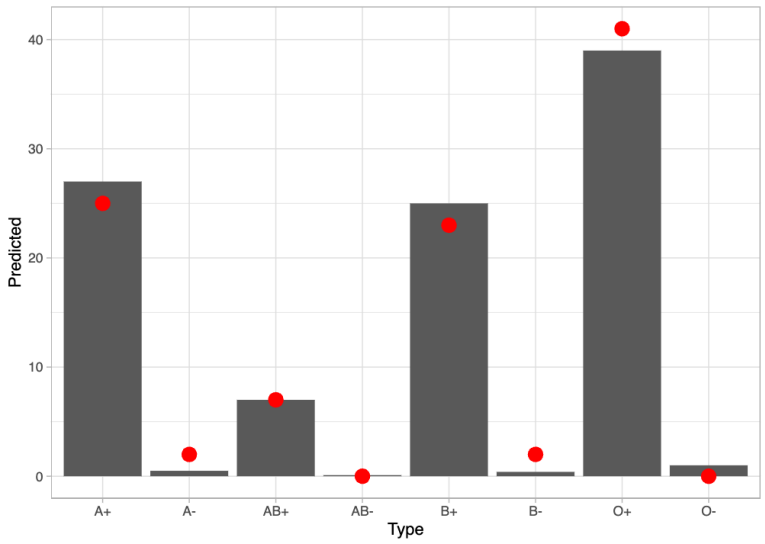
What happens if we sample from a small group, one that is much smaller than the entire population?

Questions like this become very important when thinking about blood donation programs. Not everyone donates blood, and the specific sample from which blood is sampled can have a dramatic effect on the relative supply of different blood types in a community.

The function below mimics the process of blood donations from a random sample of `people`. Random samples (i.e., donors) are drawn from a distribution where the probability of each of the eight blood types occurring is the same as in the table above.

The gray bars are the expected counts, and the red points are the observed counts in a particular sample. Run the code a few times and examine how the plot changes each time.

```
blood_sample(people = 100)
```



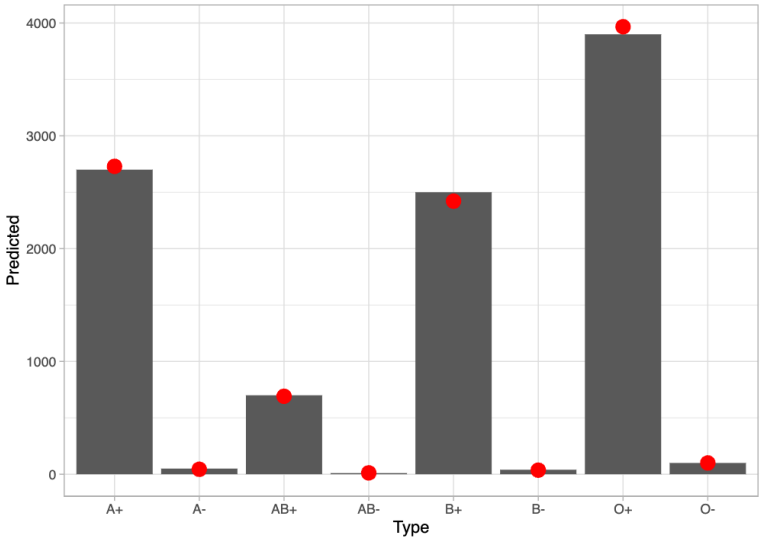
What do you observe about the distribution of counts in the observed sample (red points) compared to the predicted probabilities (gray bars)? What does this tell you about the need to specifically recruit donors with rare blood types?

The red dots (observed samples) are pretty close to the heights of the bars but can vary somewhat when n = 100. Because all the Rh-negative blood types are relatively rare, those show up rarely in sampling. Some samples have no Rh-negative individuals for some ABO types.

Go back to the code block above. Try increasing the number of people sampled and see how the distribution changes (once again, avoid very very large numbers)?

What looks different when 1000 or 10000 people are sampled?

As more people are sampled, the observed counts very closely match the predicted counts, even for the rare Rh-negative blood groups. Below is the image for n = 10000.



Goodness of fit

At this point you are gaining some intuition for how the random processes that are a normal part of sampling can lead to quite variable results each time a new sample is taken. Sometimes a sample will closely match the predicted probabilities:

- Occasionally 5 *D* alleles and 5 *d* alleles are present when 10 gametes are produced
- Occasionally the counts of blood types in a sample of 100 people agree well with the predictions

In both cases, as we increase the sample size (500, 1000, 10000, etc.), the observed counts more closely match the predicted counts.

It would be good to have a way to explicitly test whether the observed counts in a sample are a close match to what we expect. Essentially we want to know:

1. Do these observations represent a random sample from a given population?
2. Do these counts differ significantly from the expected counts for a given population?

These questions are basically asking the same thing. Both are answered by a statistical test for counts: a "goodness of fit" test. The name is fairly self-explanatory – we are testing how "good" the observations "fit" the expected.

There are many different goodness of fit tests in statistics, but the one that is used most often in genetics is the **chi-squared test**[1]. That is the test we will use here.

Although the chi-squared test is fairly straightforward as we will see, the equation is a little daunting at first:

$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

We can decode this "statistical sentence" piece by piece.

1. χ^2 is the value that we are going to calculate. It is a single number that represents the test statistic. Somewhat confusingly, this number is not, itself, squared. Think of "chi-squared" as just the number calculated on the right side of the equation.
2. Σ is the symbol for a summation, adding a set of numbers together
    - $i = 1$ tells us where to start adding
    - n tells us where to stop adding. In this case n represents the number of groups we are testing.
3. $(Observed_i - Expected_i)^2$ says to subtract the expected count from the observed count and then square that number, which is itself divided by the expect count ($Expected_i$). The subscript is are related to the $i = 1$ to n for the summation, just accounting for the fact that we'll do this calculation once for each group.

You can consider the equation as an efficient way to tell you what to do with all the counts you have. In plain language:

- For each group:

---

[1] Also called the chi-square test or the χ^2 test

    - Subtract the expected count from the observed count, square it, and divide by the expected count
- Add up all those numbers to get the χ^2 value.

Using a chi-squared test

Before we get to the case study, we will take slight detour and apply the chi-squared test to a new data set. The data below are the counts of births on each day of the week for 350 consecutive births in a hospital (not all the births on each day happened on the same day – that would be one very busy hospital).

```
DOB <- tribble(
  ~ Day, ~ Births,
  "Sunday", 33,
  "Monday", 41,
  "Tuesday", 63,
  "Wednesday", 63,
  "Thursday", 47,
  "Friday", 56,
  "Saturday", 47) |>
  mutate(Day = ordered(Day, levels = Day))

DOB
```

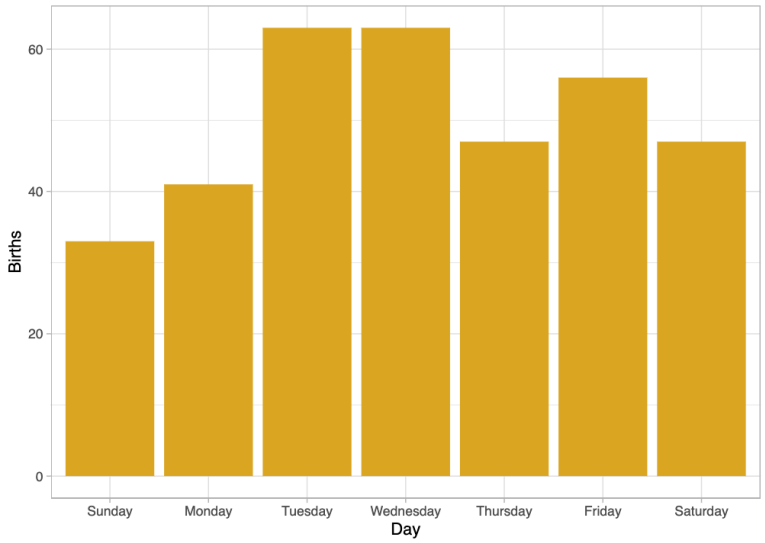
```
# A tibble: 7 × 2
  Day       Births
  <ord>      <dbl>
1 Sunday        33
2 Monday        41
3 Tuesday       63
4 Wednesday     63
5 Thursday      47
6 Friday        56
7 Saturday      47
```

Execute the code to save the data into R. If you are interested in what the code does:

- `tribble()` is a function that lets us enter data line by line. Here we are making columns for `Day` and `Births`. Each row represents one pair. Sunday has 33 births. Monday has 41, and so on.
- `mutate()` just puts the days in the correct order.

We can plot these data with the code block below.

```
ggplot(DOB, aes(x = Day, y = Births)) +
  geom_col(fill = "goldenrod") +
  theme_light()
```



For these data, the chi-squared test asks whether the number of births per day is equal across all the days. Since we expect that births happen randomly with respect to day of week (i.e., babies should have no concept of what day it is and should thus not prefer to be born on one day vs. another), this seems like a reasonable null hypothesis.

Looking at the plot above, do you predict that births are evenly distributed (null hypothesis) or unevenly distributed?

It appears that there are fewer births on Sunday and Monday compared to other days (but not really that different that Thursday and Saturday). Or Tuesday and Wednesday have more births than other days (again but not that different than Fridays).

If births are not evenly distributed throughout the week, what do you think might explain this pattern? What reason(s) might lead to more births on some days vs. others?

One hypothesis is that the apparent excess of births on some days are planned, for example Caesarean births are not scheduled on weekends, but instead are performed on Tuesdays and Wednesdays (in this sample). Or we simply have a small sample size here and there really isn't a difference (the null hypothesis).

We will find out.

Thinking back to the equation for the chi-squared test, let's figure out what we need to calculate. We have the observed counts (33, 41, 63, etc.). We first need to figure out the expect count of births per day. See if you can reason through what number that is.

What is the expected count of births per day?

There are 350 births total divided by 7 days. So, 350 / 7 = 50. We expect 50 births per day.

We can also use R to do the calculation:

```
Expected <- sum(DOB$Births) / 7
Expected
```

```
[1] 50
```

Here are using `sum()` to add up the `Births` column of the `DOB` data. With 350 births spread over 7 days, we expect 50 per day. Now that we have the expected value, we can calculate χ^2.

```
DOB <- DOB |>
  mutate(Obs_Exp = Births - Expected,
         Obs_Exp_Sq = Obs_Exp^2)
DOB
```

```
# A tibble: 7 × 4
  Day       Births Obs_Exp Obs_Exp_Sq
  <ord>      <dbl>   <dbl>      <dbl>
1 Sunday        33     -17        289
2 Monday        41      -9         81
3 Tuesday       63      13        169
4 Wednesday     63      13        169
5 Thursday      47      -3          9
6 Friday        56       6         36
7 Saturday      47      -3          9
```

We have done several things all at once in this code block. `mutate()` makes new columns. First we make a column `Obs_Exp` which is the observed count of `Births` minus `Expected` (50). We then use this new column to calculate the square of `Obs_Exp`.

Looking at the printout of the data, we can see that some days have more births than expected (Tuesday, Wednesday, Friday) and some have less (Sunday, Monday, Thursday, Saturday). All the squares are positive, as we expect (the squaring here is the derivation of the name χ^2).

All that remains is to divide the *Obs_Exp2* by *Expected*:

```
DOB <- DOB |>
  mutate(Obs_Exp_Sq_Std = Obs_Exp_Sq / Expected)
DOB
```

```
# A tibble: 7 × 4
  Day       Births Obs_Exp Obs_Exp_Sq Obs_Exp_Sq_Std
  <ord>      <dbl>   <dbl>      <dbl>          <dbl>
1 Sunday        33     -17        289           5.78
2 Monday        41      -9         81           1.62
3 Tuesday       63      13        169           3.38
4 Wednesday     63      13        169           3.38
5 Thursday      47      -3          9           0.18
6 Friday        56       6         36           0.72
7 Saturday      47      -3          9           0.18
```

This last step normalizes the squared deviation by the expected count. You can imagine if there are a lot of observations, then the squared deviations can get quite large.

Finally, we add up the values in the last column. These values represent each day's contribution to the overall χ^2 value.

```
X2 <- sum(DOB$Obs_Exp_Sq_Std)
X2
```

```
[1] 15.24
```

We have a value of 15.24. What do we do with it? Just on it's own, the χ^2 statistic doesn't mean anything. We need a value to compare it to.

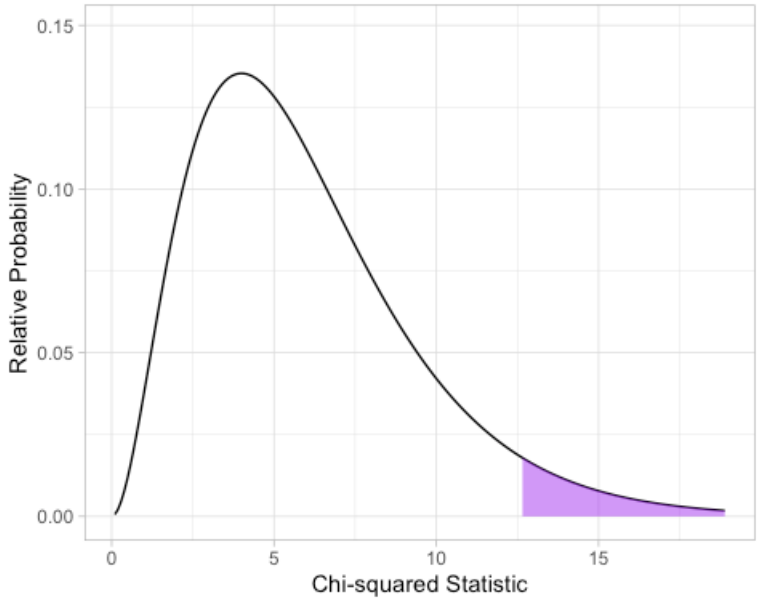
The way that the chi-squared test is set up, the test value (15.24) is compared to the value that marks the cutoff for some specified percentage of a chi-squared distribution.

The figure below shows the probability plot for a chi-squared distribution with 6 *degrees of freedom*. The concept of degrees of freedom in statistics is often confusing. One way to think about it is the number of values that can vary independently. For example, there are 10 total in two groups. If you know that one group has 6, then you also know that the other group has 4. So for two groups, there is just 1 degree of freedom (if you know one, then you can calculate the other).

There are 6 degrees of freedom in this example, because we have 7 groups (days). For this test degrees of freedom is $n - 1$. Like the sets of blood types

probabilities, this also sums to 1. But instead of adding up individual probabilities, here we calculate (integrate) the area under the curve. The area under the line is 1.

The area shaded in purple represents 5% of the area under the line. Most commonly in biological sciences, 5% is used as the cutoff when deciding that a particular test is “significant” or not.[2]



If the observed data were really drawn from a uniform distribution, meaning that counts for each day are equal, then 5% of the time, the χ^2 statistic will fall somewhere in the purple area.

For our test, all we need to know is the position on the x-axis of the left-hand edge of the purple area. This value is about 12.6 (you can confirm this because the left edge is just past the thin vertical line at 12.5). 12.6 is known as the *critical value* for the test.

If the χ^2 statistic is greater than the critical value, then we are able to reject the null hypothesis. Which says that these observations do not follow the expected distribution.

---

[2] For those interested, 5% is known as the alpha-level for the test. The choice of 5% is arbitrary but generally agreed on, for better or worse. There is a large literature on *P*-values, their use, and their misuse, but we don’t have time to dig into that today.

The test statistic we calculated was 15.24, which is greater than 12.6. So here we reject the null hypothesis. It appears that births do not happen randomly throughout the week.

Shortcut

R has a built-in function to carry out the calculations for us. Although it is very informative to work through the calculations manually, it's much easier to just let the function do all the work for us. That function is `chisq.test()`.

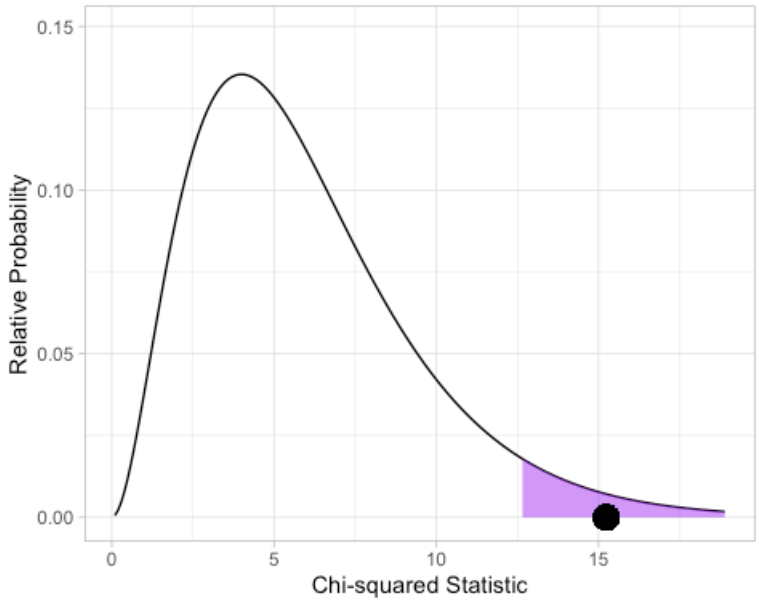
```
chisq.test(DOB$Births)
```

```
        Chi-squared test for given probabilities

data: DOB$Births
X-squared = 15.24, df = 6, p-value = 0.01847
```

The output shows much the same information that we already calculated: the χ^2 statistic of 15.24 and 6 degrees of freedom. One new value is the exact *P*-value. All that we knew above was that *P* was less than 0.05. Now we know that it is exactly 0.01847.

If we plot the chi-squared distribution again, we can add the observed statistic. *P* = 0.01847 means that 1.847% of the area under the line falls to the right of the black dot.



Take a few minutes to explore how the sample size impacts the χ^2 statistic and *P*-value. We can easily scale the data up and down:

```
chisq.test(DOB$Births * 10)
```

```
	Chi-squared test for given probabilities

data:  DOB$Births * 10
X-squared = 152.4, df = 6, p-value < 2.2e-16
```

```
chisq.test(DOB$Births / 10)
```

```
	Chi-squared test for given probabilities

data:  DOB$Births * 10
```

```
X-squared = 1.524, df = 6, p-value = 0.9579
```

> **Note**
>
> Note that if you scale down too far (e.g., `DOB$Births / 100`), you will receive a warning: `Chi-squared approximation may be incorrect`. This warning occurs because the chi-square test is not designed for very low sample sizes. This warning is just telling you to be careful with the results in this case. When sample sizes are small, there are alternative statistical tests that are used.

Do the results make sense to you based on what you know about how sample size impacts our certainty when making inferences about a sample?

When there are 3,500 births, the P-value is very small (less than 2×10^{-16}), so we can be very sure that the number of births is unequally distributed across days. In contrast, when there are only 35 births, the P-value is 0.96, which means that births are not unequally distributed. This pattern is similar to what we saw above -- when the sample size is low, small variations in sampling mean that it is difficult to draw strong conclusions. But when the sample size is high, it is easier to be certain.

Feedback

We would appreciate your anonymous feedback on this exercise. If you choose to, please fill out this optional 4-question survey to help us improve.

References

Dean, L. 2005. *Blood Groups and Red Cell Antigens*. National Center for Biotechnology Information (US).

Klug, W. S., M. R. Cummings, C. A. Spencer, M. A. Palladino, and D. Killian. 2019. *Concepts of Genetics*. 12th ed. Pearson.