TEACHING THE GENOME GENERATION

Establishing Genotype-Phenotype Associations: TAS2R38 SNP 1 Spreadsheets

The Jackson Laboratory
Leading the search for tomorrow's cures

NIH National Institutes of Health
Turning Discovery Into Health
SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

Establishing Genotype-Phenotype Associations

Introduction

What is a GWAS?

Did you know that most of our traits and many human diseases are not caused by a single genetic variant, but still are influenced by genetics?

For example, our heart health is influenced by both genetics and our environment. An individual with a variant in the *CDKN2A* gene might have an increased risk of developing heart disease, but whether that individual develops disease is also dependent on other genetic variants and on their environment. Even someone who does not have that variant in *CDKN2A* can develop heart disease, especially if factors in their behavior, such as smoking, or environment, such as exposure to air pollution, lead to increased risk.

For complex traits, like our heart health, how do scientists identify associations between genetic variants and a trait of interest when there are so many other variables? One tool that scientists use is called a Genome Wide Association study, or GWAS (pronounced *GEE-wahs*).

For a GWAS, scientists start by collecting genetic data from a large population with varied phenotypes for a particular trait or disease. Then the scientists look at millions of variants and determine if any of those variants are associated with a particular phenotype of that trait. Typically, scientists will look at single nucleotide polymorphisms, or SNPs (pronounced *snips*). SNPs are a type of variant where just one nucleotide varies from person to person.

For example, to try to find associations between genetic variants and heart disease, scientists would recruit a large group of people to participate in their study, some of whom have heart disease and others who do not. The scientists look at all of the participants' genetic variants and see if there are any variants that are more common in the group with heart disease than they are in the group without heart disease. Then, they perform statistical tests to determine if those variants are more common in the heart disease group than would be expected just by chance.

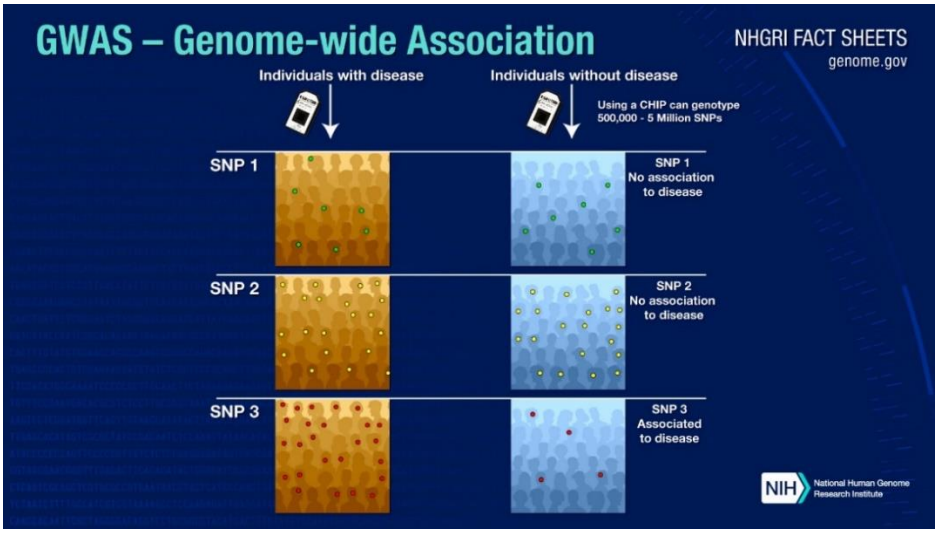

Figure 1. *Representation of potential outcomes of a GWAS study. In this example, the variants SNP 1 and SNP 2 are found in approximately equal percentages of the disease and non-disease populations. SNP 3, however, is more common in individuals with disease. (From "Genome-Wide Association Studies Fact Sheet" by the National Human Genome Research Institute).*

TAS2R38 SNP 1: SPREADSHEETS

The *TAS2R38* Gene

In this activity, you will look at an example of the types of calculations involved in a GWAS to demonstrate an association between bitter taste perception and variants in the *TAS2R38* gene.

The Taste 2 Receptor Member 38 (*TAS2R38*) gene produces the TAS2R38 protein, which functions as a receptor to perceive a wide range of bitter compounds. Bitter taste receptors, including TAS2R38, are proteins found on the taste cells of the tongue. Bitter taste perception is affected by multiple genes and by environmental factors, such as the foods we are exposed to and have regular access to.

There are three bitter taste perception phenotypes which you will investigate today: non-taster, taster, and super taster.

- Non-tasters do not perceive bitter taste.
- Tasters perceive bitter taste, although the taste may be mild.
- Super tasters perceive bitter taste, and the taste is typically strong.

While we group people into these three phenotypes, these phenotypes are subjective and correspond to a range of actual bitter taste sensing levels.

Part 1: Allele Counts

Background

For our investigation, you will start with one variant in *TAS2R38*. The *TAS2R38* variants are single nucleotide polymorphisms, or SNPs (pronounced *snips*). This means that just one nucleotide varies from person to person. There are three common SNPs in *TAS2R38*, which each occur at different locations in the gene.

- SNP 1: At this variant location, the nucleotide is usually either a guanine (G) or a cytosine (C).
- SNP 2: At this variant location, the nucleotide is usually either a thymine (T) or a cytosine (C).
- SNP 3: At this variant location, the nucleotide is usually either an adenine (A) or a guanine (G).

We have collected genetic data from a group of individuals, recording both their *TAS2R38* genotype and bitter taste perception phenotype. The samples were provided voluntarily, meaning that each person agreed to provide a sample. The samples were also anonymized, meaning they were labeled with random sample names so that they cannot be traced back to the person who provided the sample.

Recall that everyone has two copies, or alleles, of each gene. For example, for the first variant (SNP 1), individuals can have a C on both copies, a G on both copies, or one C and one G.

Activity

To start your analysis, you will count how many of each allele of your assigned SNP are associated with each tasting phenotype.

1. Create your own copy of the *TAS2R38* Google Sheet or *TAS2R38* Excel file.
    a. Note: It is good practice in data analysis to keep a copy of your raw data separate from the data that you analyze, so that you can restart from the raw data if you make a mistake in your analysis. If at any point you feel like you need to start over, you can always make or download a new copy of the *TAS2R38* spreadsheet.

2. Open your *TAS2R38* spreadsheet and navigate to the **SNP 1** tab.

3. Find **Table 1. Allele Data** on the top left of the spreadsheet. **Table 1** has 4 data columns and 60 data rows.

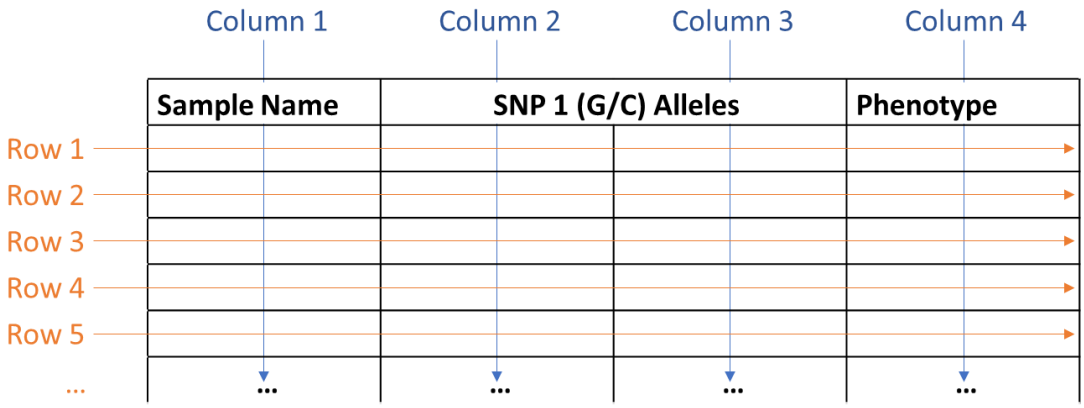

Figure 2. *Image of Table 1 with rows and columns numbered for reference. The full table has 60 rows—only 5 are shown here.*

a. What type of data is in columns 2 and 3? What does that data represent?

b. What type of data is in column 4? What does that data represent?

c. What does each row represent?

d. Some of the table entries are blank. What do you think this means about that data?

e. How many individuals are included in the dataset? How many alleles are included in the dataset?

As is, the data is organized alphabetically by sample name. However, to establish correlation between genotype and phenotype, you need to organize the data by phenotype. To do this, you will sort the data in the table.

4. Sort the data table by the **Phenotype** column using **Sort range** function in Google Sheets or the **Sort & Filter** function in Excel. Make sure that the range you select to sort contains only the data in **Table 1**. If you include the column names in your selected range, make sure that you indicate that your data has a header row.

Since the data is now sorted by phenotype, you can more easily count the number of G and C alleles that are associated with each of the bitter taste perception phenotypes. You will enter this information into one of the analysis tables, which is labeled **Table 2. Allele Counts**.

Table 2 has 3 data rows and 4 data columns, not including the row and column labels.

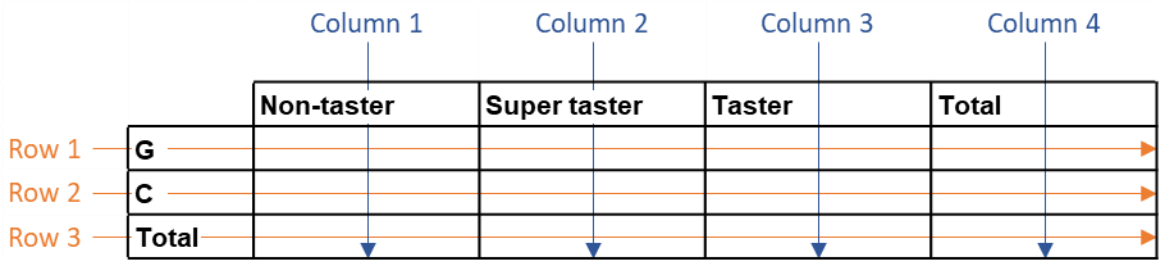

Figure 3. *Image of Table 2 with rows and columns numbered for reference.*

5. Find **Table 2. Allele Counts** and use the table to enter the following questions.
   a. What do columns 1, 2, and 3 represent?

   b. What does column 4 represent?

   c. What do rows 1 and 2 represent?

   d. What does row 3 represent?

e. Into which cell would you enter the data for the number of C alleles associated with the Taster phenotype? Put an X in the correct cell in the example table below.

Table 2. Allele Counts

	Non-taster	Super taster	Taster	Total
G				
C				
Total				

f. Into which cell would you enter the data for the number of G alleles associated with the Non-taster phenotype? Put an X in the correct cell in the example table below.

Table 2. Allele Counts

	Non-taster	Super taster	Taster	Total
G				
C				
Total				

Now that you are familiar with **Table 2**, you can count the alleles in each phenotype group and enter the counts data into **Table 2**. You can either count manually or count using a spreadsheet formula. As you enter the data, the **Total** column and **Total** row should update automatically.

Manual counting:

6. For each phenotype, count the number of G alleles and enter the data into the correct column of the **G** row of **Table 2**. Be sure to include data from both alleles for each individual!

7. For each phenotype, count the number of C alleles and enter the data into the correct column of the **C** row of **Table 2**. Be sure to include data from both alleles for each individual!

Using a formula:

8. Use the countif() formula in Excel or Google Sheets to calculate count data in **Table 2**. For each cell in **Table 2**, type:
   =countif(range, condition)
   where range is a subset of columns 2 and 3 from **Table 1** representing the allele data for the phenotype corresponding to your current column in **Table 2**, and condition represents the allele in the corresponding row of **Table 2** (either "G" or "C").

After completing all the calculations, use your completed **Table 2** to answer the following questions.

9. How many G alleles are associated with each phenotype?
    a. Non-taster:
    b. Super taster:
    c. Taster:
10. How many C alleles are associated with each phenotype?
    a. Non-taster:
    b. Super taster:
    c. Taster:
11. What patterns do you notice in the data?
12. Based on this data alone, without performing additional calculations, do you think this variant in *TAS2R38* is associated with bitter tasting phenotype? Why or why not?
13. How could you use additional data collection, analysis, or representation to better support your hypothesis?

Part 2: Allele Frequency

Background

Next, you can use the allele count data in **Table 2** to calculate allele frequencies. Allele frequency refers to the distribution of alleles across a population. For example, the allele frequency of the G allele in the total population is the fraction, or percentage, of the total alleles in the population that are G alleles. The equation we will use for allele frequency is:

$$Allele\ Frequency = \frac{v}{2N}$$

Where the variable v is the number of each allele in the dataset, and the variable N is the total number of diploid individuals in the dataset.

In Part 1, you counted the number of each allele in the dataset, so **Table 2** contains the values for v. Additionally, since N represents the number of diploid individuals and each individual has two alleles, $2N$ is equal to the total allele count. So, we can also write the allele frequency equation as:

$$Allele\ Frequency = \frac{Allele\ Count\ of\ a\ specific\ allele\ in\ the\ population}{Total\ Allele\ Count\ in\ the\ population}$$

For example, in our population of 60 individuals, we have 106 known alleles in total (**Table 2**, **Total** row, **Total** column). Of those 106 alleles, 65 are G alleles (**Table 2**, **G** row, **Total** column). We can calculate the allele frequency of the G allele in our total population as follows:

$$Total\ G\ Allele\ Frequency = \frac{Allele\ Count\ of\ G\ allele\ in\ the\ population}{Total\ Allele\ Count\ in\ the\ population} = \frac{65}{106} = 0.613$$

If we look at just the Taster population, there are 35 G alleles (**Table 2**, **G** row, **Taster** column). However, because we are only looking at the Taster population, our total number of alleles also changes. There are 30 Tasters, corresponding to 60 alleles (**Table 2**, **Total** row, **Taster** column). We can calculate the allele frequency of the G allele in the Taster population as follows:

$$Taster\ G\ Allele\ Frequency = \frac{Allele\ Count\ of\ G\ allele\ in\ Taster\ population}{Total\ Allele\ Count\ in\ the\ Taster\ population} = \frac{35}{60} = 0.583$$

Activity

First, practice calculating allele frequency using the data for the number of C alleles in the Taster population.

1. Before you begin, find the number of C alleles associated with the Taster phenotype in **Table 2. Allele Counts**. What cell in **Table 2** contains the number of C alleles associated with the Taster phenotype? Put an X in the correct cell in the example table below.

Table 2. Allele Counts

	Non-taster	Super taster	Taster	Total
G				
C				
Total				

2. What is the allele count of C alleles associated with the Taster phenotype?

3. Now, use the allele count to complete the calculation below. Round your answer to three decimal places.

$$Taster\ C\ Allele\ Frequency = \frac{Allele\ Count\ of\ C\ allele\ in\ Taster\ population}{Total\ Allele\ Count\ in\ the\ Taster\ population} = \frac{\_\_\_\_\_}{60} = \_\_\_\_\_$$

You will use **Table 3. Allele Frequencies** to record the results from your allele frequency calculations. Before completing the remaining calculations, get familiar with the format of **Table 3**.

Table 3 has 3 data rows and 4 data columns, not including the row and column labels.

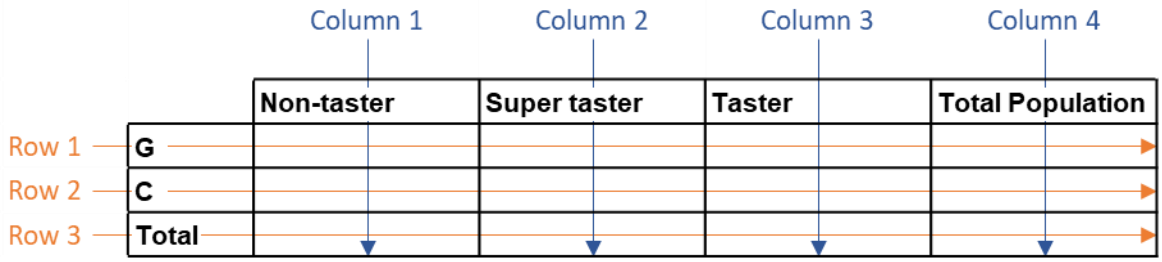

Figure 4. *Image of Table 3 with rows and columns numbered for reference.*

4. Find **Table 3. Allele Frequencies** and use the table to enter the following questions.
   a. What does column 4 represent?

b. What does row 3 represent? Predict the value of each cell in row 3.
Hint: In allele frequency, we are calculating the frequency of each allele within the population. What should be the frequency of both alleles combined in the population?

c. Into which cell would you enter the data for the frequency of G alleles in the total population? Put an X in the correct cell in the example table below.

Table 3. Allele Frequency

	Non-taster	Super taster	Taster	Total Population
G				
C				
Total				

d. Into which cell would you enter the data for the frequency of C alleles associated with the Super taster phenotype? Put an X in the correct cell in the example table below.

Table 3. Allele Frequency

	Non-taster	Super taster	Taster	Total Population
G				
C				
Total				

Now you can calculate the allele frequencies for each phenotype and enter the data into **Table 3**. You can either calculate manually using the equation, or you can use a spreadsheet formula to complete the calculations. As you enter the data, the numbers in the **Total** row should update automatically.

Manual calculation:

5. For each phenotype and for the total population, calculate the allele frequency of the G allele and enter the data into the correct column of the **G** row of **Table 3**.

6. For each phenotype and for the total population, calculate the allele frequency of the C allele and enter the data into the correct column of the **C** row of **Table 3**.

Using a formula:

7. There is no pre-existing formula for allele frequency, but you can set up your own. Create your own function in Excel or Google Sheets to calculate the allele frequencies for each phenotype and for the whole population.

   *Hint: For each cell in **Table 3**, divide the corresponding cell in **Table 2** by the cell in the **Total** row of the same column of **Table 2**.*

After completing all the calculations, use your completed **Table 3** to answer the following questions.

8. What is the G allele frequency for each of the following phenotype groups:
   a. Non-taster:
   b. Super taster:
   c. Taster:
   d. Total Population:
9. How is the C allele frequency for each of the following phenotype groups:
   a. Non-taster:
   b. Super taster:
   c. Taster:
   d. Total Population:
10. What patterns do you notice in the data?
11. The data in Table 2 and Table 3 come from the same raw data but show different ways of analyzing and representing that data. Which set of analyzed data, allele counts (**Table 2**) or allele frequency (**Table 3**), is easier for you to understand? Why?

Part 3: Expected Counts

Background

In **Table 2. Allele Counts** and **Table 3. Allele Frequency**, we have data about the distribution of the different *TAS2R38* SNP 1 alleles across the different bitter tasting phenotypes. Now we can ask the question: is there an association between variant and phenotype?

To answer this question, we can use a statistical test called a Pearson's chi-square test of independence. The Pearson's chi-square test of independence evaluates whether observations of measures on two variables (in our case, allele and phenotype) are independent of each other. The chi-square test requires us to compare our actual allele counts to **expected counts**.

Expected counts represent what the allele counts would be if there was no association between the variant and phenotype. In other words, **expected counts** represent a random distribution of alleles across the phenotypes, such that each phenotype has the same allele frequencies as the total population.

Example Calculation

For example, let's look at the expected count of G alleles associated with the Non-taster phenotype.

The frequency of G alleles in the total population is 0.613. Importantly, expected count calculations assume that the total number of alleles associated with each phenotype remains the same. In our population, there are 30 total alleles associated with the Non-taster phenotype.

1. The frequency of G alleles in the total population (0.613) is in **Table 3. Allele Frequency**. What cell in **Table 3** contains the frequency of G alleles in the total population? Put an X in the correct cell in the example table below.

Table 3. Allele Frequency

	Non-taster	Super taster	Taster	Total Population
G				
C				
Total				

2. The total number of alleles associated with the Non-taster phenotype (30) is in **Table 2. Allele Counts**. What cell in **Table 2** contains the total number of alleles associated with the Non-taster phenotype? Put an X in the correct cell in the example table below.

Table 2. Allele Counts

	Non-taster	Super taster	Taster	Total
G				
C				
Total				

The expected count of G alleles for the Non-taster population represents how many G alleles would be associated with the Non-taster phenotype if:

- the frequency of G alleles associated with the Non-taster phenotype was also 0.613
- the Non-taster phenotype was still associated with 30 alleles total

So, to calculate the expected count for a specific allele and phenotype, multiply:

a) the Allele Frequency of the specific allele in the total population BY
b) the total Allele Count for the specific phenotype.

The equation for the expected count of a specific allele associated with a specific phenotype is:

$$Expected\ Count =$$

$$Allele\ Frequency\ of\ specific\ allele\ in\ total\ population * total\ Allele\ Count\ for\ specific\ phenotype$$

Now, we can use this equation to calculate the expected count for G alleles associated with the Non-taster phenotype:

$$Expected\ Count =$$

$$Allele\ Frequency\ of\ G\ allele\ in\ total\ population * total\ Allele\ Count\ for\ Non\text{-}taster\ phenotype =$$

$$0.613 * 30 = 18.4$$

So, if this variant was not associated with bitter tasting perception, we would expect there to be 18.4 G alleles associated with the Non-taster phenotype.

It may seem strange to have an expected count of 18.4 alleles, since it is not possible to have just part of an allele in real life. However, since these are not real counts, it is okay to have decimals.

Activity

As another practice, calculate the expected count for C alleles associated with the Taster phenotype.

3. First, find the frequency of C alleles in the total population in **Table 3. Allele Frequency**.
   a. What cell in **Table 3** contains the frequency of C alleles in the total population? Put an X in the correct cell in the example table below.

Table 3. Allele Frequency

	Non-taster	Super taster	Taster	Total Population
G				
C				
Total				

   b. What is the allele frequency?

4. Next, find the total number of alleles associated with the Taster phenotype in **Table 2. Allele Counts.**
   a. What cell in **Table 2** contains the total number of alleles associated with the Taster phenotype? Put an X in the correct cell in the example table below.

Table 2. Allele Count

	Non-taster	Super taster	Taster	Total
G				
C				
Total				

   b. What is the total number of alleles associated with the Taster phenotype?

5. Now, use the numbers you just identified to calculate the expected count for C alleles associated with the Taster phenotype.

$$Expected\ Count =$$

$$Allele\ Frequency\ of\ C\ allele\ in\ total\ population * total\ Allele\ Count\ for\ Taster\ phenotype =$$

_____ * _____ = _____

Now that you've had some practice calculating expected counts, you can proceed with the rest of the analysis. In this step of the analysis, you will calculate the expected counts for each combination of allele and phenotype. You will enter this information into **Table 4. Expected Counts**. Before you begin, familiarize yourself with **Table 4**.

Table 4 has 3 data rows and 4 data columns, not including the row and column labels.

6. Find **Table 4. Expected Counts** and use the table to enter the following questions.
    a. What does column 4 represent? Predict the value of each cell in column 4.
    Hint: When calculating expected counts, we assume that the total number of each allele in the population remains the same as in the original dataset.

    b. What does row 3 represent? Predict the value of each cell in row 3.
    Hint: When calculating expected counts, we assume that the total number of individuals with each phenotype remains the same as in the original dataset.

Following the same process as the example calculations, calculate the expected count for each allele and phenotype combination and enter the data into **Table 4. Expected Counts**. You can either calculate manually using the equation, or you can use a spreadsheet formula to complete the calculations.

Manual calculation:

7. For each phenotype, calculate the expected count for the G allele using the expected count equation and enter the data into the correct column of the **G** row of **Table 4**.

8. For each phenotype, calculate the expected count for the C allele using the expected count equation and enter the data into the correct column of the **C** row of **Table 4**.

Using a formula:

9. There is no pre-existing formula for expected counts, but you can still set up your own. Create your own function in Excel or Google Sheets to calculate the expected count for each allele and phenotype combination.

    *Hint: For each cell in **Table 4**, multiply the cell that corresponds to the same phenotype in the **Table 2 Total row** by the cell that corresponds to the same allele in the **Table 3 Total Population column**.*

After completing all the calculations, use your completed **Table 4** to answer the following questions.

10. How do the values in the **Total** row and **Total** column compare to the values in the same row and column in **Table 2**? Why?

11. What patterns do you notice in the data?

12. Are the actual counts similar to or different from the expected counts? What conclusions could you draw from this comparison, if any?

Part 4: Chi-Square Test

Background

There is one final step to answering our question: is there an association between this *TAS2R38* variant and bitter taste perception phenotype?

Specifically, we will perform a Pearson's chi-square test of independence using our actual and expected allele counts. The Pearson's chi-square test of independence evaluates whether observations of measures on two variables (in our case, allele and phenotype) are independent of each other.

Remember that our expected counts represent what the allele counts would be if the counts were randomly distributed and independent of phenotype. So, by using the chi-square test to compare our expected and actual allele counts, we are testing how likely it is that the difference between our actual counts and the expected counts is due to chance.

Our null and alternative hypotheses for this test are:

- Null hypothesis (H_o): the count of each allele is independent of phenotype (the count of each allele is the same or very similar for each phenotype)
- Alternative hypothesis (H_a): the count of each allele is dependent on phenotype (the count of each allele is different for each phenotype)

For statistical tests, it is important that we establish a level of significance before we perform the testing. The level of significance, also called the alpha value, represents the probability of obtaining your results due to chance. Your chosen alpha value can change based on your research question and how much potential error you are willing to tolerate in your results.

In this case, we will choose an alpha value of .05, representing a 5% probability that our results are due to chance. So, if the outcome of our test suggests that the probability that our results are due to chance is *less than 5%,* we can reject our null hypothesis and conclude that there is an association between this *TAS2R38* variant and bitter taste perception phenotype.

Activity

The formula for the Pearson's chi-square test is complex. Luckily there is a formula built into Excel and Google Sheets that can do this calculation for us!

The formula is called CHISQ.TEST(). It takes the actual and expected counts as inputs and returns a probability value, or **p value**. The **p value** represents the probability that the distribution of alleles in our actual counts table is due to chance.

1. Calculate the Pearson's chi-square test of independence **p value** using the CHISQ.TEST() formula and your actual and expected counts tables.

   To use the CHISQ.TEST() function, choose a cell to store your results and type: =CHISQ.TEST(actual_range, expected_range)
   where actual_range represents the cells containing your actual count data from **Table 2** and expected_range represents the cells containing your expected count data from **Table 4**. Be sure to only include the count data and **not** the **Total** rows/columns or the row/column headers.

   Hint: If you have kept everything in your spreadsheet in its original location, your formula should read: =CHISQ.TEST(I12:K13, I28:K29)

2. Use the result from the Pearson's chi-square test of independence to answer the following questions:
   a. Does the **p value** meet our established criteria for statistical significance (alpha = 0.05)? Explain.

   b. What does this mean about association between this *TAS2R38* variant and bitter taste perception phenotype?

   c. Does this outcome reflect your earlier conclusions from Parts 1-3?

   d. How broadly can we generalize these results, given what we know (and don't know) about our sample population?

Part 5: Reflection

Use the following questions to draw conclusions and reflect on the dataset and calculations.

1. What can we conclude about allele/phenotype association from this dataset?

In this study, **Table 1** represents genotype and phenotype data from 60 individuals. The genotype data was determined from DNA sequencing. It is important to note that, while **Table 1** separates the genotype into two **allele** columns, we only know the genotype (i.e., what combination of alleles is present) for each SNP location. We do not know which allele is on which chromosome.

2. Can we draw any conclusions about genotype/phenotype association based on these calculations? Why or why not?

3. Each individual has two total *TAS2R38* alleles. Each of those two alleles has one specific nucleotide at each variant location. This combination of the three variants on a specific allele is called a haplotype. For example, one possible haplotype is GTA, representing a G at the location of SNP 1, a T at the location of SNP 2, and an A at the location of SNP 3.
    a. From this dataset, do we know which combination of variants were inherited together on a single chromosome in a single individual? Why or why not?

    b. Can we draw any conclusions about haplotype/phenotype association from this dataset? Why or why not?

4. How broadly can we generalize our results, given what we know (and don't know) about our sample population?

5. Based only on your results, can you conclude that these variants in *TAS2R38* cause an individual's bitter taste perception phenotype? Explain.

Remember that in a GWAS, scientists look at millions of variants across a large population with varied phenotypes for a trait, or disease, of interest. Typically, GWAS are used to study complex diseases or traits that are impacted by multiple genetic variants and environmental factors.

6. Generally, a GWAS can establish correlations between genetic variants and a disease or phenotype. However, a GWAS cannot be used to prove that a certain variant *causes* a phenotype. Why might it be important to establish a correlation between a small number of variants and a phenotype even if we cannot yet prove that those variants cause the phenotype?

7. Based on your experience completing calculations for one variant, why might computational tools and automation be important for GWAS?