Why do we have another type of significance tests for Categorical Data?

- One-sample Z significance tests only analyze the proportion for _______ category
- Two-sample Z significance tests only analyze the ____________ in proportions for ______ category
- What if we want to analyze the proportions for ______ categories?

### Chi (kye) squared tests for categorical data

1. **χ² goodness of fit test** – This test takes one variable and analyzes its ____________ among all of the categories. It compares the ____________ distribution to the ____________ sample distribution to determine if there is a difference in the distributions. For this test we compare __________ not proportions.

2. **χ² Test for homogeneity** – We continue analyzing distributions but now we are checking if the distribution is the ______________ for _____________ for several populations or treatments.

3. **χ² Test for association/independence** - For this test we sample or experiment from a single population and break them up among ____________ variables. We then check to see if there is an association (not independent) between those variables.

### Chi squared distribution

- This distribution is a measure of the distance of the ____________ count from the ____________ count.
- Unlike a Normal distribution or a t-distribution, the chi-squared distribution is not __________.
- The chi-squared distribution does change based on the number of categories creating a ____________.

![Chi-squared distribution graph](image)

- As you can see the distribution is ____________ with only ____________ values.
- As degrees of freedom become larger, it becomes ______ skewed.
- As degrees of freedom become larger, larger values become ______ probable (higher p-value)
- The mode (peak) will be at ____________
- The mean will be the ______ as the degrees of freedom
- The variance will be ____________

Chi squared statistic can be calculated by

\[ \chi^2 = \]

### Conditions for any chi squared test

- **Random**: Simple random samples or randomized experiments
- **Large sample**: All ____________ values are greater than or equal to _____ (expected counts must be listed), it is okay if expected counts are _________.
- **Independent**: Individual observations are independent. Population is at least 10 times larger than the sample.
Chi squared goodness of fit test

State: We want to perform a chi-squared goodness of fit test at the ___ significance level

- Ho: what are the expected proportions based on the population (must list ALL proportions)
- Ha: At least one of the p’s is incorrect
- Where $p_{\text{category}}$ = the true proportion of . . . .

Plan: Check if conditions are met

- Random = Random sample OR randomized experiment,
- Large Sample = Expected COUNTs are all greater than or equal to five (okay if decimals)
- Independent = Observations are independent OR 10% rule

Do: Calculate $\chi^2 = \sum \left( \frac{\text{observed} - \text{expected}}{\text{expected}} \right)^2$ (list out at least the first two terms)

- Calculate the p-value = ______ with lower bound, upper bound, and df (categories – 1)
- Remember: P-value is the probability of getting the chi-squared statistic if Ho is true

Conclude: Compare the p-value to the ________________________.

- If P-value is less than $\alpha$, we
- If P-value is not less than $\alpha$, we
- Interpret those results in context
  - We reject $H_0$, so we _________________ convincing evidence of $H_a$
  - We Fail to reject $H_0$, so we _________________ convincing evidence of $H_a$

Sample: According to the 2000 Census, of all US residents age 20 and older, 19.1% are in their 20’s, 21.5% are in their 30’s, 21.1% are in their 40’s, 15.5% are in their 50’s, and 22.8% are 60 and older. The table below shows the age distribution for a sample of US residents age 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>141</td>
</tr>
<tr>
<td>30-39</td>
<td>186</td>
</tr>
<tr>
<td>40-49</td>
<td>224</td>
</tr>
<tr>
<td>50-59</td>
<td>211</td>
</tr>
<tr>
<td>60+</td>
<td>286</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1048</strong></td>
</tr>
</tbody>
</table>

Do these data provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all US residents?

State: We will perform a chi squared goodness of fit test at the ___ significance level.

Ho: $p_2 = 0.191$, $p_3 =$ , $p_4 =$ , $p_5 =$ , $p_6 =$

Ha: At least one of the p’s is incorrect

where $p_i$ = true proportion of people in 20’s, 30’s, 40’s, 50’s, and 60+ age groups

Plan: Check conditions

Random:

Large Sample:

Independent:

Do:

Conclude: Since our p-value is _______ than ________, we ____________ the null hypothesis.

We ____________ have convincing evidence that at least one of the proportions for people in their 20’s, 30’s, 40’s, 50’s, and 60+ answering landline telephone survey’s is not the same as the age distribution of US residents.
Follow-up: If we ____________, we can perform some follow-up to determine where the differences are occurring. Which category shows the ____________ deviations from the expected values?

- **Components** - individual term or calculation that contributes to the chi-squared statistic
- When using the GOF test in your calculator, these values are stored in CNTRB
- When calculating using lists, they will be the individual values in List 3
- When completing a follow-up, talk about the ____________ and ____________ of the component from expected values.

Sample: In his book *Outliers*, Malcolm Gladwell suggests that a hockey player’s birth month has a big influence on his chance to make it to the highest levels of the game. Specifically, since January 1 is the cut-off date for youth leagues in Canada (where many NHL players come from), players born in January will be competing against players up to 12 months younger. The older players tend to be bigger, stronger, and more coordinated and hence get more playing time, more coaching, and have a better chance of being successful. To see if this is true, a random sample of 80 National Hockey League players from the 2009-2010 season was selected and their birthdays were recorded. **Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter.**

A) Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the year?

**State:** We will perform a chi squared goodness of fit test at a _____ significance level.

- \( H_0: p_1 = _____, p_2 = _____, p_3 = _____, p_4 = _____ \)
- \( H_a: \) At least one of the \( p_i \)'s is incorrect

Where \( p_i = \) the proportion of birthdays of NHL players in each quarter of the year (1=Jan-Mar, . . .)

**Plan:** Check the conditions

- **Random:**

- **Large Sample:**

- **Independent:**

**Do:**

**Conclude:** Since our p-value is __________ than ________, we __________ the null hypothesis.

We __________ have convincing evidence that at least one of the proportions for people born in each quarter of the year for the NHL is equal throughout the year.

B) We concluded that the birthdays of NHL players were not uniformly distributed throughout the year. However, Gladwell’s claim wasn’t just that the distribution wasn’t uniform—he specifically claimed that NHL players are more likely to be born early in the year. Use the values for each component to discuss this claim.
### Chi-squared test for homogeneity

- Looking at what influences what (___________________ vs ___________________ variables)
- Compare the conditional distributions of __________ variable for each __________ variable.
- Calculate conditional distributions, _________ them, describe where the differences are
- We will test for any __________ differences
- IF there are differences, analyze __________ differs and by __________ much
- **Expected counts** – take marginal distribution for each __________ variable and multiply by total in each __________ variable.
- For chi-squared statistic, degrees of freedom will be (row – 1)(column – 1)

**State:** We want to perform a chi-squared test of homogeneity at the ___ significance level
- Ho: There is no difference in the distribution of __________ for ________________
- Ha: There is a difference in the distribution of __________ for ________________

**Plan:** Check if conditions are met
- Random = Random sample OR randomized experiment,
- Large Sample = Expected Value (as if the distributions are the same) = Marginal Distribution*total for one category, so on . . .
- Independent = Observations are independent OR 10% rule

**Do:** Calculate $\chi^2 = \sum \left( \frac{\text{observed} - \text{expected}}{\text{expected}} \right)^2$ (list out at least the first two terms)
- Calculate the p-value = _______ with lower bound, upper bound, and df
- df = (categories – 1) (categories – 1)
- Remember: P-value is the probability of getting the chi-squared statistic if Ho is true

**Conclude:** Compare the p-value to the ____________________________.
- If P-value is less than α, we
- If P-value is not less than α, we
- Interpret those results in context
  - We reject H₀, so we __________________ convincing evidence of Hₐ
  - We Fail to reject H₀, so we __________________ convincing evidence of Hₐ

**Follow-up:** Which cells show large deviations? Which direction are the deviations?
Sample: In Chapter 1, an example examined the distribution of superpower preference for a random sample of 200 children (ages 9-17) from the United Kingdom. Do American children have the same preferences? To find out, a random sample of 215 U.S. children (ages 9-17) was selected. Here are the results from both samples:

<table>
<thead>
<tr>
<th></th>
<th>UK</th>
<th>US</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fly</td>
<td>54</td>
<td>45</td>
<td>99</td>
</tr>
<tr>
<td>Freeze</td>
<td>52</td>
<td>44</td>
<td>96</td>
</tr>
<tr>
<td>Invisibility</td>
<td>30</td>
<td>37</td>
<td>67</td>
</tr>
<tr>
<td>Super Strength</td>
<td>20</td>
<td>23</td>
<td>43</td>
</tr>
<tr>
<td>Telepathy</td>
<td>44</td>
<td>66</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>215</td>
<td>415</td>
</tr>
</tbody>
</table>

(a) Construct an appropriate graph to compare the distribution of superpower preference for U.K. and U.S. children.

(b) Do these data provide convincing evidence that the distribution of superpower preference differs among U.S. and U.K. children?

State: We want to perform a chi-squared test for homogeneity at a _____ significance level.

Ho: There is no difference in the distribution of superpowers for kids from the UK and kids from the US.

Ha: There is a difference in the distribution of superpowers for kids from the UK and kids from the US.

Plan: Check the conditions

Random:

Large Sample:

Independent:

Do:
Conclude: Since our p-value is ________ than ________, we ___________ the null hypothesis. We __________ have convincing evidence that there is a difference in the distribution of superpowers for kids from the UK and kids from the US.

Follow-up: The biggest differences occurred in the categories

Chi-squared or 2-sample z-test
- When dealing with two categorical variables that each have only two categories (_________), we can use a chi-square test or a 2-sample z-test.
- P-values will be approximately the same (differences come from rounding errors)

In a study reported by the Annals of Emergency Medicine (March 2009), researchers conducted a randomized, double-blind clinical trial to compare the effects of ibuprofen and acetaminophen plus codeine as a pain reliever for children recovering from arm fractures. There were many response variables recorded, including the presence of any adverse effect, such as nausea, dizziness, and drowsiness. Here are the results:

<table>
<thead>
<tr>
<th></th>
<th>Ibuprofen</th>
<th>Acetaminophen plus Codeine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverse effects</td>
<td>36</td>
<td>57</td>
<td>93</td>
</tr>
<tr>
<td>No adverse effects</td>
<td>86</td>
<td>55</td>
<td>141</td>
</tr>
<tr>
<td>Total</td>
<td>122</td>
<td>112</td>
<td>234</td>
</tr>
</tbody>
</table>

(a) Is the difference between the two groups statistically significant? Conduct an appropriate chi-square test to find out.

State: We want to perform a chi-squared test for homogeneity at a ____ significance level.

Ho: There is no difference in the proportion of patients that had adverse effects when taking ibuprofen or acetaminophen plus codeine.

Ha: There is a difference in the proportion of patients that had adverse effects when taking ibuprofen or acetaminophen plus codeine.

Plan: Check the conditions

Random:

Large Sample:

Independent:

Do:

Conclude: Since our p-value is ________ than ________, we ___________ the null hypothesis. We __________ have convincing evidence that there is a difference in the proportion of patients that had adverse effects when taking ibuprofen or acetaminophen plus codeine.

(b) Calculate the 2-sample z statistic and the p-value. What do you notice?

Ho: $p_i = p_a$  \quad $p_i = \quad p_a = \quad p_c =$

Ha: $p_i \neq p_a$
### Chi-squared test for association or independence

When sampling from a ________________, we can analyze the relationship between the variables to see if there is an association (dependent) or no association (independent).

**State:** We want to perform a chi-squared test of **association** at the ___ significance level
- **Ho:** There is no association between __________ and __________
- **Ha:** There is an association between __________ and __________

**Plan:** Check if conditions are met
- **Random** = Random sample OR randomized experiment,
- **Large Sample** = Expected Value (as if the distributions are the same) = Marginal Distribution*total for one category, so on . . .
- **Independent** = Observations are independent OR 10% rule

**Do:** Calculate \( \chi^2 = \sum \left( \frac{\text{Observed} - \text{Expected}}{\text{Expected}} \right)^2 \) (list out at least the first two terms)
- Calculate the p-value = ______ with lower bound, upper bound, and df
- df = (categories – 1) (categories – 1)
- Remember: P-value is the probability of getting the chi-squared statistic if Ho is true

**Conclude:** Compare the p-value to the ________________________.
- If P-value is less than \( \alpha \), we
- If P-value is not less than \( \alpha \), we
- Interpret those results in context
  - We reject \( H_0 \), so we ________________ convincing evidence of \( H_a \)
  - We Fail to reject \( H_0 \), so we ________________ convincing evidence of \( H_a \)

**Follow-up:** Which cells show large deviations? Which direction are the deviations?

****Remember association does not mean _________________!****

---

**State:** We want to perform a chi-squared test of **independence** at the ___ significance level
- **Ho:** __________ and __________ are independent
- **Ha:** __________ and __________ are not independent

**Plan:** Check if conditions are met
- **Random** = Random sample OR randomized experiment,
- **Large Sample** = Expected Value (as if the distributions are the same) = Marginal Distribution*total for one category, so on . . .
- **Independent** = Observations are independent OR 10% rule

**Do:** Calculate \( \chi^2 = \sum \left( \frac{\text{Observed} - \text{Expected}}{\text{Expected}} \right)^2 \) (list out at least the first two terms)
- Calculate the p-value = ______ with lower bound, upper bound, and df
- df = (categories – 1) (categories – 1)
- Remember: P-value is the probability of getting the chi-squared statistic if Ho is true

**Conclude:** Compare the p-value to the ________________________.
- If P-value is less than \( \alpha \), we
- If P-value is not less than \( \alpha \), we
- Interpret those results in context
  - We reject \( H_0 \), so we ________________ convincing evidence of \( H_a \)
  - We Fail to reject \( H_0 \), so we ________________ convincing evidence of \( H_a \)

**Follow-up:** Which cells show large deviations? Which direction are the deviations?
**Sample:** In an example from chapter 5, we investigated the relationship between gender and having allergies for a random sample of 40 students who completed a CensusAtSchool survey. Here is a two-way table that summarizes the data:

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergies</td>
<td>10</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>No Allergies</td>
<td>13</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>17</td>
<td>40</td>
</tr>
</tbody>
</table>

Do the data provide convincing evidence of an association between gender and having allergies for U.S. high school students who filled out the survey?

**State:** We want to perform a chi squared test for association at a ____ significance level.
- Ho: There is no association between gender and allergies
- Ha: There is an association between gender and allergies

**Plan:** Check the conditions

*Random:*
- Large Sample:
- Independent:

**Do:**

**Conclude:** Since our p-value is ________ than ________, we __________ the null hypothesis.
We __________ have convincing evidence that there is an association between gender and allergies.

**Follow-up:**

### Which test should you use?

1. **Goodness of fit:** ________ variable, ________ population
   - Uses a ________ way table to analyze the distribution of ________ variable

2. **Homogeneity:** ________ variable, ________ populations
   - Uses a ________ way table with the populations as one side, the measured variable on the other side

3. **Association:** ________ variables, ________ population
   - Uses a ________ way table with one variable on one side, the other variable on the other side.

***The difference between #2 and #3 is how the data is collected
- **Homogeneity** uses ________ samples then measures ________ variable (i.e. a sample of males then another sample of females then asks about their age)
- **Association/Independence** uses ________ sample then measures the ________ variables (i.e. samples people then asks about their gender – variable 1 – then asks about their age – variable 2)