**Activity 7.6.2 – Goodness-of-Fit Test**

This activity introduces the *chi-square goodness-of-fit test*, a hypothesis test that assesses whether a frequency distribution fits a hypothesized distribution. Similar to hypothesis tests for other parameters, we will use randomization distributions to assess whether an observed sample statistic provides evidence against a claim about a population parameter.

**What is Your Favorite Sport to Watch on Television?**

A sports website posts a description of U.S. teenagers’ television viewing preferences for the four major professional sports. The website reports that U.S. teenagers’ preferences are distributed in the following way (left table). You randomly select 50 U.S. teenagers and ask them about their favorite sport to watch on television. The survey results are shown below (right table).

|  |  |  |
| --- | --- | --- |
| Favorite Sport to Watch on TV |  | Survey Results *n* = 50 |
| Football | 30% |  | Football | 24 |
| Basketball | 30% |  | Basketball | 13 |
| Baseball | 25% |  | Baseball | 8 |
| Hockey | 15% |  | Hockey | 5 |

Does the sample provide evidence that the student’s claim about the population’s distribution is false? To perform a goodness-of-fit test, we will assume that the hypothesized distribution is true and calculate the chi-square statistic $χ^{2}$ for the observed sample.

1. Complete the table below to calculate the *observed* chi-square statistic $χ^{2}$. Calculate the expected frequencies (*E*) by applying the hypothesized percentages to the sample size *n* = 50.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Favorite Sport to Watch on TV | $$O$$ | $$E$$ | $$O-E$$ | $$\left(O-E\right)^{2}$$ | $$\frac{\left(O-E\right)^{2}}{E}$$ |
| Football | 24 |  |  |  |  |
| Basketball | 13 |  |  |  |  |
| Baseball | 8 |  |  |  |  |
| Hockey | 5 |  |  |  |  |

**Constructing a** $χ^{2}$ **Randomization Distribution**

We can construct a randomization distribution of chi-square statistics by:

* Constructing a population with a distribution equal to the hypothesized distribution,
* Randomly sample (with replacement) from the hypothesized population to generate a randomized distribution of chi-square statistics $χ^{2}$.

We can model a population with a distribution that follows the hypothesized distribution as follows:

* The population consists of all two-digit numbers: 00 to 99
* Numbers from 00 to 29 correspond to students who prefer football (Football: 30%)
* Numbers from 30 to 59 correspond to “prefer basketball” (Basketball: 30%)
* Numbers from 60 to 84 correspond to “prefer baseball” (Baseball: 25%)
* Numbers from 85 to 99 correspond to “prefer hockey” (Hockey: 15%)

We can generate a randomized chi-square statistic $χ^{2}$ by: (a) sampling with replacement from the population to get a random sample, (b) determining the observed frequencies, and (c) calculating the chi-square statistic $χ^{2}$ for the random sample.

1. *Random Sample* – Randomly select 50 two-digit numbers. Count the number of two-digit numbers that correspond to each category to identify the observed frequencies. Complete the table below.

*Random Sample*

|  |  |
| --- | --- |
| Favorite Sport to Watch on TV | Observed Frequencies (*O*) |
| Football |  |
| Basketball |  |
| Baseball |  |
| Hockey |  |

1. Calculate the chi-square statistic $χ^{2}$ for your random sample. Calculate the expected frequencies (*E*) by applying the hypothesized percentages to the sample size *n* = 50.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Favorite Sport to Watch on TV | $$O$$ | $$E$$ | $$O-E$$ | $$\left(O-E\right)^{2}$$ | $$\frac{\left(O-E\right)^{2}}{E}$$ |
| Football |  |  |  |  |  |
| Basketball |  |  |  |  |  |
| Baseball |  |  |  |  |  |
| Hockey |  |  |  |  |  |

1. Create a dotplot of chi-square statistics using all the sample statistics from your class.



The randomization distribution above was formed under the assumption that the population’s distribution of television viewing preferences is 30% football, 30% basketball, 25% baseball, and 15% hockey. The variability in the chi-square sample statistics is due to sampling variability.

1. According to the randomization distribution, what is the probability of obtaining a random sample with a chi-square statistic greater than or equal to the *observed* chi-square statistic? This probability is the *P*-value for the chi-square goodness-of-fit test.

A *P*-value is the probability of obtaining a sample statistic as extreme or more extreme than the one observed assuming the population parameter is equal to a specific value.

* When a *P*-value is less than 5%, we say the sample statistic is *statistically significant*. This means that that it is likely that the statistic did not emerge due to chance alone. We reject the assumption about the population parameter.
* When a *P*-value is greater than or equal to 5%, we say the sample statistic is *not statistically significant*. This means that the sample statistic could have occurred solely due to chance. We do not reject the assumption about the population parameter.
1. Is the observed chi-square statistic statistically significant?
2. What can we conclude about the original claim regarding teenagers’ television viewing preferences for the four major professional sports?

**Age Distribution of Facebook Users**

In 2010 the Pew Research Center surveyed users of social networking sites to obtain information on the ages of users. They reported the following age distribution of Facebook users (left table). A researcher randomly surveyed150 Facebook users in 2015 and found the distribution shown below (right table). Does the 2015 age distribution provide evidence that the age distribution of Facebook users changed over time?

|  |  |  |
| --- | --- | --- |
| 2010 Survey ResultsAge Distribution of Facebook Users |  | 2015 Survey Results, *n* = 150Age Distribution of Facebook Users |
| 18 – 22 | 16% |  | 18 – 22 | 39 |
| 23 – 35 | 33% |  | 23 – 35 | 45 |
| 36 – 49 | 25% |  | 36 – 49 | 35 |
| 50 – 65 | 20% |  | 50 – 65 | 22 |
| 65+ | 6% |  | 65+ | 9 |

We can conduct a randomization test as follows:

* Assume the age distribution of Facebook users in 2015 is 16% age 18 – 22, 33% age 23 – 35, 25% age 36 – 49, 20% age 50 – 65, and 6% age 65+.
* Construct a randomization distribution of chi-square sample statistics from random samples of size *n* = 150.
* Find the probability of observing a chi-square statistic as extreme as the one found.
* Make a decision about the population.

The following graph shows a randomization distribution of 500 chi-square sample statistics from random samples of size *n* = 150.

Randomization Distribution of Chi-Square Sample Statistics $χ^{2}$, *n* = 150

****

1. Calculate the chi-square statistic $χ^{2}$ for the observed sample. Calculate the expected frequencies (*E*) by applying the hypothesized percentages to the sample size *n* = 150.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age Distribution of Facebook Users | $$O$$ | $$E$$ | $$O-E$$ | $$\left(O-E\right)^{2}$$ | $$\frac{\left(O-E\right)^{2}}{E}$$ |
| 18 – 22 | 39 |  |  |  |  |
| 23 – 35 | 45 |  |  |  |  |
| 36 – 49 | 35 |  |  |  |  |
| 50 – 65 | 22 |  |  |  |  |
| 65+ | 9 |  |  |  |  |

1. According to the randomization distribution, what is the probability of obtaining a random sample with a chi-square statistic greater than or equal to the *observed* chi-square statistic?
2. Is the observed chi-square statistic statistically significant?
3. What can we conclude about the original claim about the age distribution of Facebook users?