Course > Inference: Relationships C→C > Case C→C > Case C→C: Summary

Bookmark this page

Case C→C: Summary

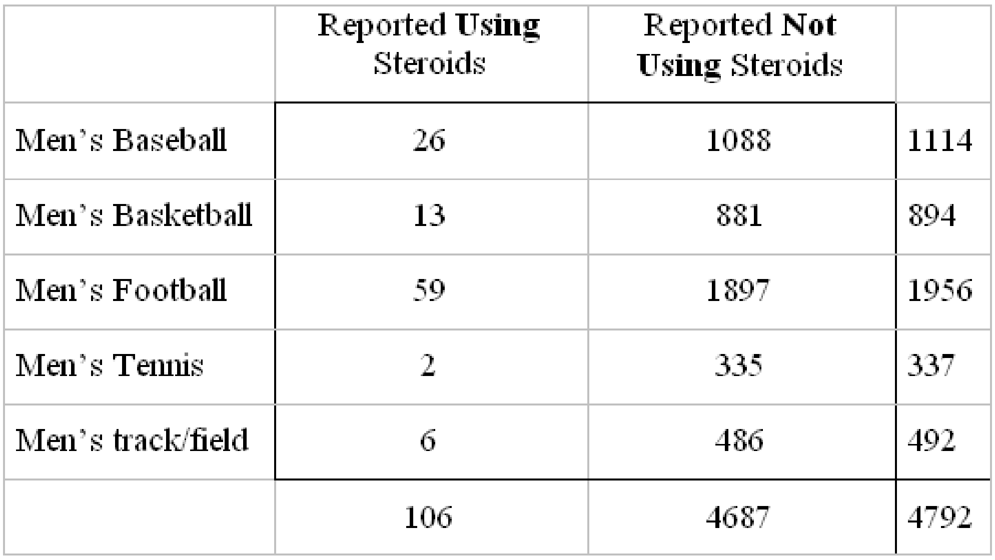
Learning Objective: In a given context, carry out the appropriate inferential method for comparing relationships and draw the appropriate conclusions.

Learning Objective: Specify the null and alternative hypotheses for comparing relationships.

Let's look at another example.

Example: Steroid Use in Sports

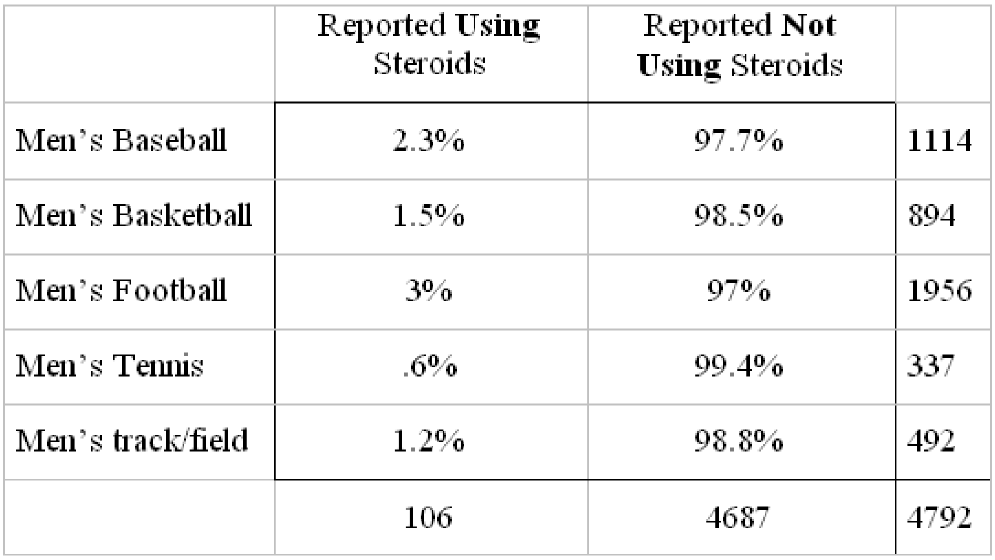
Major-league baseball star Barry Bonds admitted to using a steroid cream during the 2003 season. Is steroid use different in baseball than in other sports? According to the 2001 National Collegiate Athletic Association (NCAA) survey (http://www.ncaa.org/library/research/substance_use_habits/2001/substance_use_habits.pdf), which is self-reported and asked of a stratified random selection of teams from each of the three NCAA divisions, reported steroid use among the top 5 college sports was as follows:



	Reported Using Steroids	Reported Not Using Steroids	
Men's Baseball	26	1088	1114
Men's Basketball	13	881	894
Men's Football	59	1897	1956
Men's Tennis	2	335	337
Men's track/field	6	486	492
	106	4687	4792

Do the data provide evidence of a significant relationship between steroid use and the type of sport? In other words, are there significant differences in steroid use among the different sports?

Before we carry out the chi-square test for independence, let's get a sense of the data by calculating the conditional percents:



	Reported Using Steroids	Reported Not Using Steroids	
Men's Baseball	2.3%	97.7%	1114
Men's Basketball	1.5%	98.5%	894
Men's Football	3%	97%	1956
Men's Tennis	.6%	99.4%	337
Men's track/field	1.2%	98.8%	492
	106	4687	4792

It seems as if there are differences in steroid use among the different sports. Even though the differences do not seem to be overwhelming, since the sample size is so large, these differences might be significant. Let's carry out the test and see.

Step 1: Stating the hypotheses

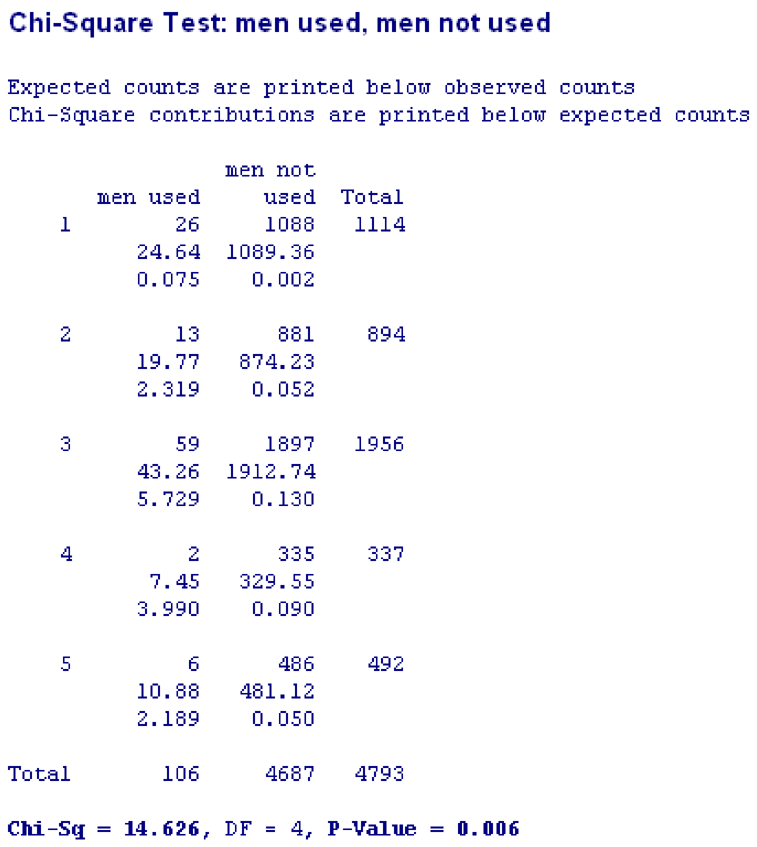
The hypotheses are:

H_0: steroid use is not related to the type of sport (or: type of sport and steroid use are independent)

H_a: Steroid use is related to the type of sport (or: type of sport and steroid use are not independent).

Step 2: Checking conditions and finding the test statistic

Here is the statistical software output of the chi-square test for this example:



Chi-Square Test: men used, men not used

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

                   men not
       men used       used  Total
    1        26       1088   1114
          24.64    1089.36
          0.075      0.002

    2        13        881    894
          19.77     874.23
          2.319      0.052

    3        59       1897   1956
          43.26    1912.74
          5.729      0.130

    4         2        335    337
           7.45     329.55
          3.990      0.090

    5         6        486    492
          10.88     481.12
          2.189      0.050

Total       106       4687   4793

Chi-Sq = 14.626, DF = 4, P-Value = 0.006
```

- Conditions:

  1. We are told that the sample was random.

  2. All the expected counts are above 5.

- Test statistic:

  The test statistic is 14.626. Note that the "largest contributors" to the test statistic are 5.729 and 3.990. The first cell corresponds to football players who used steroids, with an observed count larger than we would expect to see under independence. The second cell corresponds to tennis players who used steroids, and has an observed count lower than we would expect under independence.

Step 3: Finding the p-value

According to the output p-value it would be extremely unlikely (probability of 0.006) to get counts like those observed if the null hypothesis were true. In other words, it would be very surprising to get data like those observed if steroid use were not related to sport type.

Step 4: Conclusion

The small p-value indicates that the data provide strong evidence against the null hypothesis, so we reject it and conclude that the steroid use is related to the type of sport.

Let's Summarize

- The chi-square test for independence is used to test whether the relationship between two categorical variables is significant. In other words, the chi-square procedure assesses whether the data provide enough evidence that a true relationship between the two variables exists in the population.

- The hypotheses that are being tested in the chi-square test for independence are:

  - H_0: There is no relationship between ... and ....
  - H_a: There is a relationship between ... and ....
  - or equivalently,
  - H_0: The variables ... and ... are independent.
  - H_a: The variables ... and ... are not independent.

- The idea behind the test is measuring how far the observed data are from the null hypothesis by comparing the observed counts to the expected counts—the counts that we would expect to see (instead of the observed ones) had the null hypothesis been true. The expected count of each cell is calculated as follows:

$$Expected\ \ Count = \frac{Column\ \ Total*Row\ \ \ Total}{Table\ \ \ Total}$$

- The measure of the difference between the observed and expected counts is the chi-square test statistic, whose null distribution is called the chi-square distribution. The chi-square test statistic is calculated as follows:

$$\chi^2 = \sum_{all\ \ cells} \frac{(Observed\ \ Count - ExpectedCount)^2}{Expected\ \ \ Count}$$

- Once we verify that the conditions that allow us to safely use the chi-square test are met, we use software to carry it out and use the p-value to guide our conclusions.