Hypothesis tests for the difference between two population proportions using Stata

B V Girdler-Brown, FCPHM, FFPH, Hons BComm (Econ); **L N Dzikiti**, MSc

School of Health Systems and Public Health, Faculty of Health Sciences, University of Pretoria, South Africa

Corresponding author: *B V Girdler-Brown (brendangirdlerbrown@gmail.com)*

This educational article outlines the main methods available, using Stata statistical software, for testing hypotheses about the equality of population proportions, using sample-derived data. The article focuses on how to select the most appropriate test to use, the relevant Stata statistical software commands and the interpretation of the Stata output obtained following these commands. Both single-sample and two-sample hypothesis tests are covered.

South Afr J Pub Health 2018;2(3):63-68. DOI:10.7196/SHS.2018.v2.i3.71

In the previous edition of the journal we presented an overview of hypothesis testing for the difference between two population means, using Stata (StataCorp, USA) statistical software. In that article, we dealt with numerical data.[1] For those wishing to read further at this introductory level we recommend the text by Pagano and Gauvreau.[2]

In this article, we will give a similar overview, but for testing the difference between two population proportions. We will be dealing with binary variables where, at an individual level, a characteristic is either present (coded as 1) or absent (coded as 0). For each individual there is a characteristic of interest/outcome variable, such as lung cancer, with only two possible states, namely lung cancer present or lung cancer absent.

There is also, for hypothesis testing, a second classifying/exposure variable, which is also binary and coded 1 or 0. This second variable identifies the two groups that must be compared. An example might be smoker/non-smoker, etc.

The data would be laid out as in Table 1, for a sample of 100 study participants (lung cancer coded as 1 if present, 0 if absent; smokers coded as 1 if a smoker, 0 if a non-smoker).

Table 1. Long-format data entry for proportions

Participant_id	Lung_cancer	Smoker
1	1	1
2	1	1
3	0	0
4	0	1
…	…	…
79	0	1
80	1	0

We might summarise the information available from this table into counts of participants in a 2 × 2 contingency table (Table 2).

From Table 2 one can see that, in this sample of 100 people, the proportion of smokers with lung cancer is 15/45 (0.3), while the proportion of non-smokers with lung cancer is 5/55 (0.09). The proportion with lung cancer appears to be higher among smokers than non-smokers. However, this apparent difference may be due to a sampling error. If I were to draw a different sample of 100 people at random, then the same difference might not be observed in the second sample.

The hypothesis test would involve the relationship between the outcome state and the exposure state. For example, the hypothesis may be articulated as an answer to the question: 'Is there a difference in the proportions of smokers and non-smokers who develop lung cancer?' An appropriate null hypothesis would be that the proportions that develop lung cancer ($\pi_{smokers}$ and $\pi_{non\text{-}smokers}$) are equal, and may be written in three different ways:

H_0: $\pi_{smokers} = \pi_{non\text{-}smokers;}$ or;

H_0: $\pi_{smokers} - \pi_{non\text{-}smokers} = 0$; or

H_0: $\pi_{smokers} / \pi_{non\text{-}smokers} = 1$

(Note the use of the Greek π in these null hypothesis statements. This reminds us that the hypothesis test is testing a hypothesis about the study population parameters from which the samples have been drawn.)

Table 2. Contingency table (2 × 2) for lung cancer cases/non-cases by smoking status

		Lung cancer present?		
		Yes	No	
Smoking history?	Yes	15	30	
	No	5	50	
Total		20	80	100

Scope of the article

The foci of this article are on the selection of the most appropriate test; the Stata statistical software commands to use in order to perform the test; and the interpretation of Stata output for the test.

We have assumed that the reader possesses an understanding of the principles of statistical hypothesis testing.

A single-sample test, large sample size

A single-sample hypothesis test involving a proportion would involve, for example, the comparison of a sample-based population proportion estimate with a given gold standard or target.

For example, in 2018, there may be a target of 78% for voluntary HIV testing among pulmonary tuberculosis (TB) patients who make use of public sector health facilities, and who do not have a record of a previously positive HIV test result.

Official surveillance data in a rural district might show that this target has been met (or exceeded). However, a researcher might want to perform a study to determine the coverage that is based on carefully collected and verified information. The researcher could draw a simple random sample of 200 patients listed in the district's TB register, and then look for laboratory confirmation of the testing that has taken place. (S)he finds that 73% of the patients in her sample have in fact had an HIV test performed during the course of their anti-TB treatment. This result, 73%, is clearly below the target of 78%. However, could this difference be due to sampling error?

The null hypothesis (the null value would be the gold standard, 0.78, since this is a single-sample test) is:

H_0: $\pi_{tested} = 0.78$ (if one is interested in any difference from 0.78, either <0.78 or >0.78, a so-called 'two-tail' test); or

H_0: $\pi_{tested} \geq 0.78$ (if one is only concerned about the possibility that the target has not been met, a so-called 'single-tail' test).

If the sample size is large enough such that np and n(1-p) are both ≥5, then one may use a single sample *z*-test to test these null hypotheses. In Stata, this *z*-test is called a 'prtest', and the same command is used for both a single-sample test and a two-sample test. There is also an immediate command, 'prtesti', that may be used when the data are not already entered in the usual 1/0 format.

The Stata output provides a confidence interval (CI) that is wholly derived from the sample information, and a *z*-score *p*-value that is derived on the assumption that the null hypothesis is true.[3] This *p*-value should be used to decide on statistical significance. There may be discrepancies between this *p*-value and the CI.

For example, assume that one has a sample proportion of 0.73 (sample size = 200) for a null hypothesis that $\pi = 0.78$. Assuming a single-tail hypothesis test, the *p*-value is the probability that a random sample size of 200 will have a proportion of 0.73 or less if the null hypothesis is true. The value of *p* is found to be 0.044. Hence one rejects the null hypothesis (if $\alpha = 0.05$).

However, the 90% CI (90% since we are dealing with a single-tail test and want to know the upper and lower significance levels for rejection) comes to 0.678 - 0.782 (normal approximation method) or 0.674 - 0.781 (exact binomial method). The 90% CI comes to 0.674 - 0.779 if one uses the logit transformation method to estimate the 90% CI.

In this example, only the logit transformation method yields a 90% CI that does not include the null value of 0.78; the normal approximation and binomial exact methods produced 90% CIs that overlap with the null value of 0.78. Hence we would often fail to reject the null hypothesis using these methods, when we should, in fact, have rejected the null hypothesis.

Single-sample test, small sample size

When the conditions that $n\pi$ and $n(1-\pi)$ are both ≥5 have not both been met, then one may peform the binomial test (or 'bitest') in Stata. This test is based on the expected and observed number of successes for a given number of trials. The lower the number of trials, the more poorly the result of this test will compare with the prtest. Stata provides the calculated *p*-values for the observed number of successes under the null hypothesis. As the number of trials increases, the bitest and prtest will produce similar results for the *p*-values, even though the bitest results are estimated from whole numbers of successes, while the prtest results are obtained from the proportion treated as a continuous variable. There is no CI obtained from the bitest.

Two-sample tests

Suppose we have collected data from non-pregnant female patients with listeriosis, and from healthy non-pregnant female controls.

We find that 45/100 of the patients indicated that they had eaten uncooked polony during the 2 weeks prior to the onset of symptoms. Sixty of 200 controls also indicated that they had eaten uncooked polony during the 2 weeks before being interviewed.

Consider the set of results in Table 3 (fictitious data). It can be seen that the proportion of those with listeriosis who ate polony is 0.45 (45/100, 45%) The proportion who ate polony among the controls is 0.30 (60/200, 30%).

Is there a statistically significant difference between these two proportions? Might the difference that we observe be due to sampling error?

There are two variables in this situation, and both are binary. We are interested in the proportion of those classed as listeriosis patients who have a history of polony consumption, v. the proportion of those who do not have listeriosis and who have a history of polony consumption.

There are three main ways in which these questions may be addressed in Stata. The first is a two-sample prtest (which, as pointed out, is Stata's name for a *z*-test of two proportions), the second is a χ^2 test and the third is Fisher's exact test.

The prtest

The two-sample prtest is based on the assumption that both samples are 'large' (i.e. that *np* and *n*(1-*p*) are both ≥5 for each

Table 3. Fictitious data for listeriosis v. polony eaten (case-control study)

		Listeriosis		Totals
		Yes	No	
Polony eaten	Yes	45	60	105
	No	55	140	195
Totals		100	200	300

of the samples, where *n* is the sample size and *p* is the sample proportion with the outcome of interest). One should only perform a prtest if these large sample conditions are met. The reason for this is that the prtest is a normal approximation test that treats the mean of the 1 and 0 values as if it were a continuous variable with a normal probability sampling distribution.

This is only approximately acceptable if the sample sizes are sufficiently large. The null hypothesis is either:

H_0: $\pi_{listeriosis} = \pi_{controls}$; or

H_0: $\pi_{listeriosis} - \pi_{controls} = 0$

The two-sample prtest will yield a *p*-value as well as a CI for the difference between the two proportions.

The χ^2 test for independence, two binary variables

One of the other two alternative tests that are available in Stata is the χ^2 test. This test is performed on the count data in the 2 × 2 contingency table illustrated in Table 3. Table 3 contains the actual count data (whole numbers) in each cell. It also shows the row and column totals. In the χ^2 test, the null hypothesis is that the exposure variable (eating polony) observed counts are independent of the outcome variable (listeriosis v. control) observed counts.

To do this, Stata first estimates what the expected cell values would be if the null hypothesis is true. These expected cell values are estimated using the observed row and column totals as a given, and then expected cell values are assigned using simple probability theory. For the χ^2 test result to be valid for a 2 × 2 table, all these calculated expected cell values should be ≥5. In Stata, one is able to request that Stata show the expected cell values, so that one can then check and ensure that this important condition has been met.

If one or more of the expected cell values in a 2 × 2 table is/are <5, then one should not rely on the *p*-value that Stata has presented. Instead, one should perform Fisher's exact test on the data. The Stata commands for both these tests are presented below.

For large sample situations the prtest, χ^2 test and Fisher's exact test will all give very similar *p*-values. For smaller samples, the χ^2 and Fisher's exact tests will usually agree fairly well; for very small samples with an expected cell value <5, the Fisher's exact result will be quite different from that for the χ^2 test, and the Fisher's exact test *p*-value should be used. Where the conditions for a prtest are not met, it is recommended that the χ^2 (or Fisher's exact) test be used rather than the prtest.[4]

One of the drawbacks of using either the χ^2 test or the Fisher's exact test is that, while one obtains a *p*-value, one does not obtain a CI for the difference between the two proportions.

Performing the analyses using Stata statistical software

Data layout

Irrespective of whether one is performing a prtest, a χ^2 test or a Fisher's exact test, there will be an outcome variable and an exposure variable. Both should be coded as 1 (factor present) or 0 (factor absent).

Stata commands (given between < and >; when typing the command omit < and >)

1. For single-sample tests:

The following commands are presented for the single-sample prtest, where GS = the gold standard or target against which you are comparing actual performance. The output will give *p*-values for both single-tail and two-tail tests. Remember that np and n(1 – *p*) must both be ≥5.

<prtest variable = GS>

Where variable is the name of the 1/0 variable and GS is the gold standard or target proportion.

Should you require a 90% CI instead of the default 95% CI then use:

<prtest variable = GS, citype(90)>

Should you wish to use the immediate command:

<prtesti n p GS>

Where n is the sample size, p is the sample proportion (between 0 and 1) and GS is the gold standard proportion. Again, you may add in '… citype(90) after the GS if you want to obtain a 90% CI.

If nπ and/ or n(1 – π) is <5, then one may no longer use the prtest. Instead, one makes use of the binomial test. Only whole numbers are allowed.

<bitest variable==GS>

Where variable is the name of the binary variable coded as 1/0 and GS is the gold standard proportion (between 0 and 1). **Please note the use of the double equal sign for this command**.

For the immediate command:

<bitesti trials successes GS>

Where 'trials' is the sample size (a whole number), successes is the number of those who have the outcome of interest (also a whole number) and GS is the gold standard proportion (between 0 and 1).

2. The commands are presented for the two-sample prtest, with data stored in the long format (i.e. one binary variable indicating the presence or absence of the outcome of interest, and another indicating, for each participant, which comparison group that person belongs to).

The output will show *p*-values as well as 95% CIs for the difference between the population proportions of the two comparison groups.

Again, remember that n_1p_1; $n_1(1 - p_1)$; $n_2\ p_2$ and $n_2\ (1 - p_2)$ must all be ≥5 (n_1 and p_1 refer to the numbers and proportions in the first comparison group; n_2 and p_2 do the same for those in the second comparison group).

<prtest variable1, by(variable2)>

Where variable1 is the outcome variable and variable2 is the group identifier variable for the groups being compared. The immediate command is:

<prtesti n_a p_a n_b p_b>

Where n_a and p_a refer to the number of people in group A and the proportion with the outcome; and n_b and p_b refer to the number and proportion in group B.

3. Next, the commands are presented for the two-sample prtest with data stored in the wide format.

<prtest variable_a = (variable_b)>

Where variable_a is the binary (1/0) outcome measure for group

A, and variable_ is the binary (0/1) outcome measure for those in group B.

4. The commands for the χ^2 test in Stata are as follows (data must be in the long format for this command):

<tab variable1 variable2, chi2>

Where variable1 is the outcome variable and variable2 is the group identifier variable for the groups being compared.

Should you wish to also see the expected cell count values in order to decide whether to rather perform Fisher's exact test (if any one or more of the expected cell values is/are <5), then use the following command. This command will give you the expected value cell counts as well as the χ^2 test result:

<tab variable1 variable2, expected chi2>

The corresponding immediate commands are:

<tabi a b \ c d, chi2> and

<tabi a b \ c d, expected chi2>

Where a, b, c, and d represent the cell counts for the 2 × 2 table (Table 4).

5. Should you decide, after finding that an expected cell value is <5 that you would prefer to perform Fisher's exact test, then simply substitute 'exact' for 'chi2' in any of the above commands. In fact, Stata allows one to ask for all the results in a single step, and then one can just decide which test to rely on and which *p*-value to use, without having to repeat the commands. The following command, for example, would yield a great deal of information in a single step:

<tab variable1 variable2, expected chi2 cchi2 exact>

This would result in the following information being made available: the expected cell values, the χ^2 *p*-value, the individual cell χ^2 values, and the exact test *p*-values. The immediate command equivalent would be:

<tabi a b \ c d, expected chi2 cchi2 exact>

Stata version 14 outputs

(NB Stata version 15 outputs will be almost identical).

Example 1: Single-sample prtest (large samples)

Using the immediate command for the TB HIV testing example with 200 TB patients, 146 (73%) of whom had undergone HIV testing, the following output was obtained given a target of 78% (0.78) (Fig. 1).

If the individual level data had been entered into Stata as 1s (tested) and 0s (not tested), and if the variable has the name 'tested' then the Stata command would be:

<prtest tested=0.78>

The Ha: p<0.78 shows the *p*-value for a single-tail test of the alternative hypothesis that the proportion tested is <0.78. The Ha: p>0.78 shows the *p*-value for a single-tail test of the alternative hypothesis that the proportion tested is >0.78. The Ha: p!=0.78 shows the *p*-value for a two-tail test of the alternative hypothesis that the proportion tested is not equal to 0.78.

Since we are only interested in/concerned about the possibility that the proportion tested fails to meet the target of 0.78 we would concern ourselves with the single-tail test result, *p*=0.044. We would then reject the null hypothesis and conclude that we have probably failed to reach the target of 78%.

Table 4. A generic 2 x 2 contingency table

		Outcome?	
		Yes	No
Exposed?	Yes	a	b
	No	c	d

Example 2: Single-sample bitest (small samples)

Let us assume that, instead of a sample of 200 as we had in example 1, we only had a sample of 18 TB patients. The records show that 14 of these patients had undergone HIV testing. Have we met the target of 78% tested?

Recall that, for the prtest to be valid, we must meet the condition that np and n(1 – p) must both equal or exceed 5 (n is the sample size and p is the proportion that were tested). In this case, np = 18 × (14/18) = 14; n(1 – p) = 18 × (4/18) = 4. Clearly the conditions required for the prtest have not been met. We therefore resort to the binomial test.

The output in Fig. 2 has been obtained from Stata for the binomial test carried out using the immediate command option (numbers are small so this will be the most common situation).

If you had these data entered in Stata at the individual level as 1s and 0s, and if you called this variable 'tested', then the following Stata command may be used to obtain the same results:

<bitest tested=0.78>

Notice that there are no CIs presented.

In Fig. 2, 'k' is the number of successes. Stata has used the formula for calculating binomial probabilities for different numbers of successes from 18 trials if the null hypothesis of *p*=0.78 is true.

We see that if H_0 is true, then the probability of obtaining 13 or fewer success is 0.361. Clearly we have no grounds in this case to reject the null hypothesis. The deviation from 78% success could easily be due to a sampling error.

```
. prtesti 200 .73 .78

One-sample test of proportion                    x: Number of obs =      200

    Variable |      Mean   Std. Err.                     [95% Conf. Interval]
-------------+------------------------------------------------------------
           x |       .73   .0313927                      .6684715    .7915285

    p = proportion(x)                                             z =  -1.7070
Ho: p = 0.78

    Ha: p < 0.78                 Ha: p != 0.78                 Ha: p > 0.78
 Pr(Z < z) = 0.0439         Pr(|Z| > |z|) = 0.0878          Pr(Z > z) = 0.9561
```

Fig. 1. Stata output for a single-sample test of proportion.

```
. bitesti 18 13 .78

        N   Observed k   Expected k   Assumed p   Observed p
-------------------------------------------------------------
       18          13        14.04     0.78000      0.72222

  Pr(k >= 13)              = 0.813374   (one-sided test)
  Pr(k <= 13)              = 0.361298   (one-sided test)
  Pr(k <= 13 or k >= 16)   = 0.569714   (two-sided test)
```

Fig. 2. Stata output for a binomial test (immediate command).

Notice as well that for the binomial probabilities these are worked out for whole numbers of successes only. It is not possible to have, say, 13.3 people vaccinated.

As the expected value vaccinated under the null hypothesis is 14.04 (0.78 × 18) this should be rounded down to ≤13 (14 – 1) for those results falling below the expected value and rounded up to 16 or more (15 + 1) for those results exceeding the expected value.

Example 3: Two-sample prtest (large samples)

The Stata command and output in Fig. 3 was obtained using the case-control study data summarised in Table 3.

The 95% CI for the difference between the two proportions (–0.26 to –0.03) was calculated by Stata using the sample difference (–0.15), the standard error for the sample difference and the normal approximation.

The *p*-value (0.0102 for the two-tail test option) was calculated on the assumption that the null hypothesis is true (i.e. that the true difference in the proportions = 0).

Should you not have the data entered into Stata as an individual level 1/0 variable, then the immediate command in Fig. 4 would obtain the same results as those presented above (you would first need to work out the proportions with listeriosis among the polony-eating group and the non polony-eating group):

<prtesti 195 0.2821 105 0.4286>

```
. prtest listeriosis, by(polony)

Two-sample test of proportions                    0: Number of obs =      195
                                                  1: Number of obs =      105
------------------------------------------------------------------------------
    Variable |      Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |   .2820513   .032225                      .2188914    .3452112
           1 |   .4285714   .0482945                     .3339159     .523227
-------------+----------------------------------------------------------------
        diff |  -.1465201   .0580587                     -.2603131   -.0327272
             |   under Ho:  .0570614    -2.57   0.010
------------------------------------------------------------------------------
       diff = prop(0) - prop(1)                                  z =  -2.5678
   Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0051         Pr(|Z| < |z|) = 0.0102          Pr(Z > z) = 0.9949
```

Fig. 3. Stata output for a two-sample test of proportions.

```
. tab listeriosis polony, expected chi2

           Key
    frequency
expected frequency

           | Consumed polony in
           | previous two weeks
Listeriosi |         0          1 |     Total
s case     |-----------------------+----------
         0 |       140         60 |       200
           |     130.0       70.0 |     200.0
-----------+-----------------------+----------
         1 |        55         45 |       100
           |      65.0       35.0 |     100.0
-----------+-----------------------+----------
     Total |       195        105 |       300
           |     195.0      105.0 |     300.0

          Pearson chi2(1) =   6.5934   Pr = 0.010
```

Fig. 4. Stata output for a χ^2 test of independence (large sample).

```
. tab tb hiv, expected chi2 exact

           Key
    frequency
expected frequency

Tuberculos |    HIV sero-status
       is |         0          1 |     Total
----------+-----------------------+----------
        0 |         7          6 |        13
          |       4.2        8.8 |      13.0
----------+-----------------------+----------
        1 |         5         19 |        24
          |       7.8       16.2 |      24.0
----------+-----------------------+----------
    Total |        12         25 |        37
          |      12.0       25.0 |      37.0

          Pearson chi2(1) =   4.1937   Pr = 0.041
               Fisher's exact =                 0.067
  1-sided Fisher's exact =                      0.048
```

Fig. 5. Stata output for a small-sample Fisher's exact test.

Example 4: Chi square (χ^2) test

The Fig. 4 output was obtained from Stata for the listeriosis and polony data from Table 3. The output was obtained using the names of the data variables with the data entered into a Stata data set at the individual level.

Firstly, notice that the expected cell values (130, 70, 65 and 35) are all >5. We are therefore comfortable using a χ^2 square test. Secondly, there is no 95% CI presented for the difference between the two group proportions. This is because the null hypothesis is that polony consumption and listeriosis are independent of each other.

With this large sample situation, the *p*-value is almost identical to that obtained from the prtest.

Example 5: Fisher's exact test

Stata output is now presented (Fig. 5) for a sample where one of the expected cell values is <5. This means that we should be cautious about the χ^2 results, as they may be misleading. We would rather make use of the Fisher's exact test results in this case.

As an aside, the general approach is that if >10% of expected cell values are <5, then the χ^2 results may be misleading. However, not all people accept this guideline. In the case of a 2 × 2 table, such as that illustrated in the output displayed in Fig. 5, one cell makes up 25% of all the cells. In such a case, a low expected cell value affecting only one cell would be reason to prefer the Fisher's exact test result.

The χ^2 *p*-value is 0.041, suggesting statistical significance. That for the Fisher's exact test is 0.067, suggesting statistical non-significance.

In the past, many statisticians have made use of the Yates correction factor for discontinuity, especially in cases where numbers are small. This correction factor is not available with Stata. Nowadays, the trend is to use the Fisher's exact test, rather than invoking the Yates correction.[5] Prior

to the advent of desktop statistical software programmes, when many statistical analyses were done by hand, the Fisher's exact test proved burdensome to use, and alternative approximate methods were popular, but they are rarely used any more. Note here that the expected value for the upper left-hand cell is 4.2 (<5). Therefore it is preferable to rely on the Fisher's exact *p*-value.

1. Dzikiti LN, Girdler-Brown BV. Parametric hypothesis tests for the difference between two population means. Strengthen Health Syst 2017;2(2):40-46. https://doi.org/10.7196/SHS.2017.v2.21.60
2. Pagano M, Gauvreau K. Principles of Biostatistics, 2nd ed. Pacific Grove: Duxbury, 2000.
3. Gauvreaux K. Hypothesis testing proportions. Circulation 2006;114(14):1545-1548. https://doi.org/10.1161/circulationaha.105.586487
4. Rosner B. Fundamentals of Biostatistics, 6th ed. Belmont: Thomson Brooks/Cole, 2006.
5. Lydersen S. Statistical review: Frequently given comments. Ann Rheum Dis 2015;74(2): 323-325. https://doi.org/10.1136/annrheumdis-2014-206186

Accepted 12 March 2018.