Single-Sample Statistical Tests with a Binary Dependent Variable

Many surveys use a simple statistical test that is analogous to the single sample t test we used to investigate whether a company paid a higher than (state) average wage. In this survey example, the researcher is interested in whether one candidate (or side of an issue) would receive more votes than an alternative candidate. Survey participants are asked a single question which has two possible options, such as "yes" or "no." The statistical test investigates whether there are significantly more "yes" than "no" responses.

There are two tests designed for this circumstance. One of these tests is a z test that is very similar to the single-group t test, called the z test for the difference between two proportions. The formula looks like this:

$$z = \frac{p - \pi_0}{\sqrt{p(1-p)/n}}$$

In the formula, p is the proportion of the sample choosing one of the options in the survey (e.g., "yes"), π_0 is the null hypothesis value (i.e, the proportion expected if there is no difference between "yes" and "no"—usually .5 in head-to-head polls), and n is the sample size.

Standard Error and Significance Tests

The denominator of the z test equation above is the standard error estimate, $SE_p = \sqrt{p(1-p)/n}$. There are two things to notice. If you look carefully, you will see that this formula parallels the single-sample t test, $t = (\bar{Y} - \mu)/\sqrt{s^2/n}$, in which the denominator is the standard error estimate of the mean, $SE_{\bar{Y}} = \sqrt{s^2/n}$. The standard error of the proportion simply uses the short cut formula for the variance, s, because $p(1-p)$ is equivalent to the uncorrected standard deviation computation.[1] The top part of the equation is parallel as well, because it concerns the difference between the sample and population means ($\bar{X} - \mu$) and we know that p is equal to the sample mean if values are coded 0,1.

The second thing to notice is that the equation above uses the sample proportion p, for the calculation of the standard error. We use the sample standard deviation to estimate the standard error typically when the population value is unknown. Use of the sample proportion for the computation of the standard error gives the *Wald test* or Wald ratio (a term used whenever the sample statistic is divided by the estimate of the standard error), whereas use of the null population value in the standard error computation, with $SE_\pi = \sqrt{\pi_0(1-\pi_0)}$, is called the *score test* (Wilson or Lagrange Multiplier test). Use of the population null value is based on the "score" interval and is permissible in this case and in fact desirable, because its sampling distribution more closely approximates the normal distribution and works well even for small sample sizes (Agresti, 2013). The score test can be said to be an "approximate normal" test. The Wald test and Wald confidence intervals do not perform as well when the proportions are extreme unless n is very large. The score test is then

$$z = \frac{p - \pi_0}{SE_\pi} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

With a z-proportions test, one can also construct "confidence limits" or a "confidence interval." Generally speaking, the confidence limits describe the amount of sampling variability expected. The 95% confidence interval is an estimate of the range of these possible values (more precisely, 95% of this range). In the case of the z test, we use the normal distribution and our estimate of standard error to construct the interval using the following formula.

$$p \pm (z_{critical})(SE_\pi),$$

where the $z_{critical}$ is the critical value, which is 1.96 whenever the normal distribution is used. Wald confidence intervals also could be constructed if we replace the standard error computed from the null proportion, SE_π, with the standard error estimate computed from the observed proportion, SE_p. They tend to be too small (Type I errors) when the sample size is small and the observed proportion is very low or high, so the score is a

[1] Your text authors do not actually use a symbol for standard error in this case, but my use here is consistent with their use of SE elsewhere.

preferred test generally. Two other approaches you may hear about are the Agresti-Coull confidence interval (or test) or the Clopper-Pearson confidence interval. The Agresti-Coull (Agresti & Coull, 1998) is a correction to the Wald that improves estimates when the proportion is near 0 or 1 by adding two cases each to the failures and successes. The Clopper-Pearson (1934) interval approach is known as an "exact" test. It is based on the inverse of the binomial distribution but and is overly conservative. Agresti and Coull show that the score test outperforms the Wald and the Clopper-Pearson and that Agresti-Coull correction improves the Wald test for small n and proportion with similar performance to the score but outperforms it when the proportions are extreme.

Yet another alternative—one that also performs well—is the *likelihood ratio test* of proportions, which is a comparison between the log-likelihood fit of the data for the observed proportion, p, and the log-likelihood fit of the data for the null proportion, π_0. Using the binomial distribution equation for the null and observed proportions and then taking -2 times the natural log, ln.[2] The symbol G^2 is used for the likelihood ratio chi-square. Significance is determined by comparison to the critical value in the chi-square distribution with df = 1 (i.e., 3.84).

$$G^2 = -2\ln\left(\frac{L_0}{L_1}\right)$$

$$\binom{n}{k}\pi^k(1-\pi)^{n-k} = \frac{n!}{k!(n-k)!}\pi^k(1-\pi)^{n-k}$$, with p = π_0 and k = $n\pi_0$ for L_0, and p and np for L_1.

As with the other binomial proportion tests, a confidence interval is possible for the likelihood ratio test, but it is more complicated. Nonetheless this approach produces intervals that perform similarly to the confidence intervals from the score test. The value 3.84 is $z^2_{critical}$, the critical value for chi-square if df = 1.

$$\left[p\left(\frac{n}{n+3.84}\right)+\frac{1}{2}\left(\frac{3.84}{n+3.84}\right)\right] \pm \sqrt{\frac{1}{n+3.84}\left[p(1-p)\left(\frac{n}{n+3.84}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{3.84}{n+3.84}\right)\right]}$$

The center of the confidence interval is an adjustment which increases as the observed proportion departs from .5 rather than just the observed proportion p. The fact that the center is not p makes this approach inconvenient for presenting results.[3]

Computation Example

As an example, I use data extrapolated from reports of a 2020 presidential election poll in Georgia.[4] The results in a head-to-head question with just Biden and Trump, in which I excluded the undecideds (n = 988), had Biden up among registered voters by 53.1% to 46.9%. To determine whether this is a significant difference, we need only choose one proportion—the proportion for either Biden or Trump, it does not matter. The null hypothesis is that voters in the population are perfectly split 50/50 (i.e., the proportion is .50), so π = .5. If the proportion of the sample for one candidate differs from this value relative to what we expect due to sampling variability (chance), then one candidate has a significant lead over the other.
If we plug in our obtained values, we get the following result:

$$z = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$$
$$= \frac{.5314-.50}{\sqrt{.50(1-.50)/988}}$$
$$= \frac{.0314}{.0159}$$
$$= 1.97$$

[2] I'll review natural logarithms in the maximum likelihood handout.
[3] There are still other confidence interval approaches, with and without continuity corrections, including Jeffreys (Bayesian approach).
[4] These results are taken from a Quinnipiac University poll from Oct 14, 2020 in Georgia among likely voters, https://poll.qu.edu/georgia/release-detail?ReleaseID=3679. Methodological details are here https://poll.qu.edu/images/polling/ga/ga10142020_demos_bgwc96.pdf. These results are "extrapolated" here because the survey is weighted for demographics, because I excluded other categories ("other" "wouldn't vote" and "don't know/refused"), and because some rounding is necessary to construct the counts to match the percents given in the report.

This obtained value is compared to the critical value which is always 1.96 for two-tailed significance regardless of sample size (i.e., there is only one normal curve). Because our computed value of 1.97 exceeds this cutoff value there is a significant difference between the proportion that preferred Biden and the proportion that preferred Trump. The confidence limits for our example employ $p \pm (z_{critical})(SE_\pi)$, as given above, based on the score version of the test and we get the following values for the lower confidence limit (LCL) and the upper confidence limit (UCL):

$$LCL = .5314 - (1.96)(.0159) = .5314 - .031 = .5004$$
$$UCL = .5314 + (1.96)(.0159) = .5314 + .031 = .5624$$

Thus, the 95% confidence interval is .50-.56. This interval does not include (with sufficient decimals!) the null hypothesis value of .50, suggesting that the difference from an equal proportion is unlikely to be due to random sampling chance. Whenever the confidence limits include the null value, you will find that the significance test will have a non-significant result. Half of this confidence interval is what is commonly called the *margin of error*, and is typically expressed in terms of a percentage. We can just use the .031 subtracted to find the confidence interval multiplied by 100 to find a percent (i.e., $.031 \times 100 = 3.1\%$) or we can compute the margin of error by subtracting the LCL from the UCL and dividing by two [$(.56 - .50)/2 \times 100 = .06/2 \times 100 = 3.0\%$]. The two methods are equivalent but may differ slightly depending on whether or when rounding is used.

Goodness-of-fit Tests

A second, equivalent test for this problem is the *Pearson chi-square*. The chi-square is a "goodness-of-fit" test that compares frequencies obtained in the sample to those expected according to the null hypothesis (i.e., no difference in the population). The chi-square formula looks like this:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where $\sum$ is the summation sign, indicating addition across all the "cells," O_i is the observed frequency (obtained from the study), and E_i is the frequency expected if the two "cells" were equal. If we translate our presidential poll into observed frequencies, for Biden, O_i = 525, and for Trump, O_i = 463. The expected frequency, based on equal counts, is $E_i = (534 + 454)/2 = 494$.

$$\begin{aligned}\chi^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(525 - 494)^2}{494} + \frac{(463 - 494)^2}{494} \\ &= \frac{961}{494} + \frac{961}{494} \\ &= 1.95 + 1.95 \\ &= 3.90\end{aligned}$$

This computed value is compared to a critical value obtained from a chi-square table. It is a 1-*degree-of-freedom* (*df*) test, and chi-square for a two-tailed 1-*df* test is always 3.84. Our computed value does exceed this value, so voters were significantly more likely to prefer Biden over Trump. The *z test* and the chi-square test will always give identical results, in fact, $z^2 = (1.97)^2 = \chi^2 = 3.88$ (within rounding error), with an equivalent significance test assuming a two-tailed test for *z*.

An alternative goodness-of-fit test is the *likelihood ratio chi-square* of frequencies, which similarly quantifies the closeness of the observed and expected frequencies, and also is evaluated against a chi-square distribution with one *df*. The likelihood ratio equation is somewhat more complicated than the Pearson chi-square, because it involves a natural log, ln. As with the Pearson chi-square, larger values will be

$$G^2 = 2\sum O_i \ln\left(\frac{O_i}{E_i}\right)$$

As you might expect, z^2 from the likelihood ratio test of proportions is equal the likelihood ratio chi-square of frequencies. The nice feature of the goodness-of-fit tests is that their equations are very general, allowing for comparison of three or more frequencies or comparison of groups.

Software Examples
Below I illustrate the binomial tests and confidence intervals in SPSS, R, and SAS using the results of the Reuters poll. For the first several analyses below, I use only respondents who favored Biden or Trump and I omit cases who favored third party candidates or were undecided. The variable "response" is the dependent variable used for all of these analyses.

SPSS

Note that the Azen and Walker texts gives the menu steps for the binomial test. Please note that by default SPSS with give results in the "model viewer" form, which is a graphical depiction or the result. I turn this off by going to Edit->Options->output tab->select "Pivot tables and charts."

```
nptests   /onesample test (response) binomial (testvalue=.5 successcategorical=list(1) likelihood ).
*For binomial (z-proportion) test, successcategorical=list(1) chooses the value of 1 (Biden) as the comparison
proportion
*testvalue=.5 gives the null proportion (default and can be omitted)
*likelihood gives CIs based on the sample SE estimate (Wald) rather than the null value SE estimate.
*The z-value printed uses a continuity correction (and will not match other programs unless the continuity
correction is requested.
```

One-Sample Binomial Test Summary

Total N	988
Test Statistic	525.000
Standard Error	15.716
Standardized Test Statistic	1.941
Asymptotic Sig.(2-sided test)	.052

```
nptests    /onesample test (response) chisquare.
```

One-Sample Chi-Square Test Summary

Total N	988
Test Statistic	3.891[a]
Degree Of Freedom	1
Asymptotic Sig.(2-sided test)	.049

a. There are 0 cells (0%) with expected values less than 5. The minimum expected value is 494.

R

```
> #for binomial proportion test, first find the number of biden voters
> #SummaryStats(response)
> #then enter in the number of cases into prop.test(x,n,p,continuity correction option)
> #where x is the number of successes (Biden voters)
> prop.test(525, 988, p=0.5, correct=FALSE)

        1-sample proportions test without continuity correction

data:  525 out of 988, null probability 0.5
X-squared = 3.8907, df = 1, p-value = 0.04855
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5001991 0.5623109
sample estimates:
        p
0.5313765
```

```
>
> #this lessR BarChart function produces a chi-square test by default
> BarChart(response)

>>> Suggestions
BarChart(response, horiz=TRUE)  # horizontal bar chart
BarChart(response, fill="greens")  # sequential green bars
PieChart(response)  # doughnut (ring) chart
Plot(response)  # bubble plot
Plot(response, stat="count")  # lollipop plot


--- response ---


Missing Values of response: 0


                  0      1     Total
Frequencies:    463    525       988
Proportions:  0.469  0.531     1.000


Chi-squared test of null hypothesis of equal probabilities
  Chisq = 3.891, df = 1, p-value = 0.049
```

SAS

The binomial option produces the proportion test and the chisq option produces the Pearson chi-square.

```
proc freq data=one ;
      where (response=0) or (response=1);
      tables response / binomial(ac wald wilson level=2 p=.5) alpha=.05 chisq;
      *level=2 specifies the second category (Biden) as the referent. Note this is not
the code;
      *using order=freq just after data=one would do the same thing (because Biden is the
highest frequency);
      *p=.5 is the null proportion (this is the default and can be omitted);
      *three types of intervals are requested, ac=Agresti-Coull, wald=Wald, and
wilson=score;
run;
```

The FREQ Procedure

intended vote

response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Trump	463	46.86	463	46.86
Biden	525	53.14	988	100.00

Chi-Square Test for Equal Proportions

Chi-Square	3.8907
DF	1
Pr > ChiSq	0.0486

Binomial Proportion
response = Biden

Proportion	0.5314
ASE	0.0159

Confidence Limits for the Binomial Proportion

```
                Proportion = 0.5314

Type             95% Confidence Limits

Agresti-Coull    0.5002         0.5623
Wald             0.5003         0.5625
Wilson           0.5002         0.5623

     Test of H0: Proportion = 0.5

   ASE under H0              0.0159
   Z                         1.9725
   One-sided Pr >  Z         0.0243
   Two-sided Pr > |Z|        0.0486

          Sample Size = 988
```

Sample write-up

Note: In practice, you would never need to do both the binomial and chi-square test to compare two responses.

A binomal test was used to test whether significantly more likely voters preferred Joseph Biden over Donald Trump for president. Of the 988 voters surveyed, 525 (53%) preferred Biden and 463 (47%) preferred Trump. The difference was statistically significant, z = 1.97, p = .049, 95% score CIs=.50,.56, indicating that the preference for Biden was larger than what would be expected due to chance. The 95% score confidence intervals, did not include the null value of equal proportions. The margin of error for this survey was 3%.

A chi-square test was conducted to determine whether registered voters preferred Biden. Of the 988 voters surveyed, 525 (53%) preferred Biden and 463 (47%) preferred Trump. The difference was statistically significant, χ^2 (1) = 3.89, p = .049), indicating that the preference for Biden was larger than what would be expected due to chance.

References

Agresti, A. *Categorical data analysis, third edition*. New York: Wiley.
Agresti, A., & Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119-126.