Using Binary Variables to Estimate Proportions

R Tutorials for Applied Statistics

---

1 Introduction

When you compute the *mean* of a *binary* variable, it is mathematically equivalent to a proportion, Since all the x_i xi are equal to 0 or 1, summing them all the observations together is the same as counting the number x_i xi that equal to 1. The following formula for the sample mean reveals it is equal to the sample proportion

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{n(x_i = 1)}{n},$$
$$x¯=∑i=1nxin=n(xi=1)n,$$

where $n(x_i = 1)$n(xi=1) denotes the count for the observations in the sample that are equal to nn, and the ratio $n(x_i = 1)/n$n(xi=1)/n is exactly the proportion of a sample that is equal to 1.

Similarly, the *average value from the population for a binary variable* is the proportion of times the binary variable is equal to 1 or the *probability* that $x_i = 1$xi=1.

To estimate a probability of an event occurring, one can use standard statistical methods using a *binary outcome variable*.

2 Example

Suppose a national poll is used to measure whether or not someone is voting for the republican candidate for president. A binary variable is set equal to 1 if the respondent said ‘yes’, they do intend to vote for the republican candidate, and set to 0 if they intent to vote for someone else. If one wishes to estimate the probability that the Republican candidate will win with more than 50% of the vote, we can use a simple single sample t-test for a mean with the following hypotheses:

$$H_0 : \mu = 0.5$$
H0:μ=0.5

$$H_A : \mu > 0.5$$
HA:μ>0.5

3 Data

The data set, `loanapp.RData`, includes actual data from 1,989 mortgage loan applications, including whether or not a loan was approved, and a number of possible explanatory variables including variables related to the applicant's ability to pay the loan such as the applicant's income and employment information, value of the mortgaged property, and credit history. Also included in the data set are variables measuring the applicant's race and ethnicity.

The code below loads the `R` data set, which creates a data set called `data`, and a list of descriptions for the variables called `desc`.

```
load(url("https://murraylax.org/datasets/loanapp.RData"))
```

4 Inference on Single Proportion

Let us estimate the proportion of loans that were approved. The variable `approve` is a binary variable equal to 1 if the loan was approved and 0 otherwise. We simply compute the mean of approved:

```
mean(df$approve, na.rm=TRUE)
```

```
## [1] 0.8739449
```

We can see that 87% of loan applications are approved in our sample.

Suppose someone claims that more than 85% of all mortgage loan applications are approved. We can test this claim with a single-sample hypothesis test. The null and alternative hypotheses are given by:

$$H_0 : \mu = 0.85$$
H0:μ=0.85

$$H_A : \mu > 0.85$$
HA:μ>0.85

We conduct the test with the `t.test` function:

```
t.test(df$approve, mu=0.85, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  df$approve
## t = 3.0403, df = 1776, p-value = 0.001199
## alternative hypothesis: true mean is greater than 0.85
## 95 percent confidence interval:
##  0.8609834       Inf
## sample estimates:
## mean of x
## 0.8739449
```

The p-value is equal to `0.001199`0.001199 which is less than 0.05. We reject the null hypothesis and conclude that more than 85% of loan applications are approved.

The statistical results above include a 95% confidence interval, but it is only a one-sided "interval" because we asked for a one-sided test. Generally, we think of confidence intervals as two-sided. We can also compute a 95% confidence interval for the proportion of loans approved by looking at the two-sided result:

```
t.test(df$approve, conf.level=0.05, alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  df$approve
## t = 110.96, df = 1776, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 5 percent confidence interval:
##  0.8734509 0.8744388
## sample estimates:
## mean of x
## 0.8739449
```

With 95% confidence we can report an interval estimate for the average approval rate between 87.35% to 87.44%.

5 Estimating Difference Between Two Proportions

Let us continue using the loan approval data set and make comparisons across different groups.

Suppose we wish to estimate whether the average approval rate is lower for self-employed people versus people not self employed. The variable, `self` is a binary variable equal to 1 if the person in the observation is self employed and zero otherwise.

Our null and alternative hypotheses are given by,

$$H_0 : \mu_0 - \mu_1 = 0$$
H0:μ0−μ1=0

$$H_A : \mu_0 - \mu_1 > 0$$
HA:μ0−μ1>0

where μ_1μ1 is the probability a self-employed person (`self=1`) is approved a loan and μ_0μ0 is the probability someone who is not self-employed (`self=0`) is approved for a loan.

The alternative hypothesis has a *greater-than* sign, because our research question asks whether group 1 has a smaller proportion than group 0. Since we are subtracting the smaller proportion, the difference will be *greater than* 0.

The following call to `t.test` runs an independent-samples t-test for a difference in means depending on the value for `self`:

```
t.test(approve ~ self, data=df, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  approve by self
## t = 1.7165, df = 296.09, p-value = 0.04356
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.001691381         Inf
## sample estimates:
## mean in group 0 mean in group 1
##       0.8797921       0.8361345
```

Our sample evidence shows that 87.9% of people who are not self-employed are approved for a mortgage loan and 83.6% of self-employed people are approved.

The p-value for the hypothesis test is equal to 0.04356. This is less than 0.05, so we reject the null hypothesis. We conclude that self-employed people on average are approved for mortgages less than others.

We can compute a 95% confidence interval for the difference in approval rates:

```
t.test(approve ~ self, data=df, conf.level=0.95, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  approve by self
## t = 1.7165, df = 296.09, p-value = 0.08711
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.006396378  0.093711616
## sample estimates:
## mean in group 0 mean in group 1
##       0.8797921       0.8361345
```

With 95% confidence, the difference in approval rate for non-self-employed people versus self-employed people is between -0.6% and 9.4%