PLSC 502: "Statistical Methods for Political Research"

Two-Group Comparisons, I
October 20, 2016

Introduction

We'll discuss bivariate statistics for the next few weeks. In essence, we'll discuss how to test for bivariate differences in a dependent/response variable (generically denoted Y) across different values of an independent variable/predictor/covariate (generically called X).

Today we'll discuss the case where Y is continuous (either an unbounded continuous variable or a proportion) and X is dichotomous. In general, we examine such differences through a *difference of means test*, also sometimes generically referred to as a *t-test*.

Differences of Means

For two groups in the data defined by a dichotomous variable X, call $\bar{Y}_0 = \bar{Y}|X = 0$ and $\bar{Y}_1 = \bar{Y}|X = 1$. The difference between these two values is:

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i} \tag{1}$$

where Y_{0i} and Y_{1i} denotes $Y_i|X = 0$ and $Y_i|X = 1$, respectively, and n_0 and n_1 are the number of observations in the data with $X = 0$ and $X = 1$, respectively. Think of this number $\bar{Y}_1 - \bar{Y}_0$ as a *sample statistic*; that is, there is some value $\bar{\mu}_1 - \bar{\mu}_0$ in the population which one wants to learn about through $\bar{Y}_1 - \bar{Y}_0$.

One can show (in a relatively straightforward way) that the statistic:

$$\bar{Y}_1 - \bar{Y}_0 \sim t\left(\sqrt{\sigma^2_{\bar{\mu}_1 - \bar{\mu}_0}}\right). \tag{2}$$

That is, $\bar{Y}_1 - \bar{Y}_0$ is distributed according to a t distribution with degrees of freedom equal to $\sqrt{\sigma^2_{\bar{\mu}_1 - \bar{\mu}_0}}$. The value of the latter term follows in a straightforward way from the sampling variability of $\bar{Y}_0$ and $\bar{Y}_1$:

$$\sigma^2_{\bar{\mu}_1 - \bar{\mu}_0} = \frac{\sigma^2_0}{n_0} + \frac{\sigma^2_1}{n_1} \tag{3}$$

where σ^2_0 and σ^2_1 are the variance of Y for $X = 0$ and $X = 1$, respectively. In practice, we do not know these values, and so we use estimates s^2 based upon the data. Thus,

$$s^2_{\bar{Y}_1 - \bar{Y}_0} = \frac{s^2_0}{n_0} + \frac{s^2_1}{n_1} \tag{4}$$

and so

$$s_{\bar{Y}_1 - \bar{Y}_0} = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}. \tag{5}$$

This is why difference of means tests are often referred to as "t-tests." Note that if both the variances of the two groups defined by X are the same – that is, if $s_0^2 = s_1^2$ – and if the sample sizes are the same, then the formula for the degrees of freedom reduces to $n_0 + n_1 - 2$.[1]

The t-test as defined in (2) is what is known as "Welch's" t-test; it is different than "Student's" original t-test in that it allows for the possibility that the variances of the two groups are different ("Student's" original t-test required that they be equal, which is a special case of what is outlined above).

Hypothesis Testing, Confidence Intervals, etc.

The importance of (2) is that it allows us to build confidence intervals, conduct hypothesis tests, and the like on the statistic $\bar{Y}_1 - \bar{Y}_0$. For example, the $(1 - \alpha) \times 100$ - percent confidence interval for $\bar{Y}_1 - \bar{Y}_0$ is:

$$\bar{Y}_1 - \bar{Y}_0 \pm t_\alpha(s_{\bar{Y}_1 - \bar{Y}_0}), \tag{6}$$

where $t_\alpha(\cdot)$ denotes the α significance level of the t distribution and $s_{\bar{Y}_1 - \bar{Y}_0}$ is defined as in (5) above. Remember that, particularly as the d.f. of a t distribution goes to infinity, the distribution takes on the shape of a standard normal (z).

Similarly, we can test a null hypothesis $H_0 : \bar{Y}_1 - \bar{Y}_0 = k_0$ by calculating the associated t-score:

$$t = \frac{(\bar{Y}_1 - Y_0) - k_0}{s_{\bar{Y}_1 - \bar{Y}_0}} \tag{7}$$

and seeing whether that "t-score" allows us to reject the appropriate null hypothesis that $\bar{Y}_1 - \bar{Y}_0 = k_0$. Alternatively, we can calculate the the P-value associated with the calculated t-score. We'll do an example of this below.

Tips for t

One thing that inevitably happens as one does more and more research is that one gets a better and better idea of what are "good" values for t. Assuming a relatively large number of degrees of freedom, Table 1 gives some values of t that are useful to keep in the back of your mind...

[1] This will be somewhat important later, and is also useful in designed experiments, when we can ensure that $n_0 = n_1$.

Table 1: Rough Values of t You'll Want To Get To Know

Absolute Value of t	One-Tailed P-Value*	Two-Tailed P-Value
≈ 1.3	0.10	0.20
≈ 1.65	0.05	0.10
≈ 2	0.025	0.05
≈ 2.4	0.01	0.02
≈ 2.6	0.005	0.01
> 3	< 0.001	< 0.002

Note: Assumes d.f. $= \infty$. Asterisk indicates that the directionality of the statistic is "correct" relative to expectations.

In other words, a t-score of less than 1.3 or so isn't even worth talking about in most cases, while one > 3 is "significant" at any level you'd care to mention. In between, $t = 2$ is a commonly-looked-for cutoff value.

Differences of Proportions

For instances where Y is a proportion, the same formulae hold; the key difference is that things are somewhat easier to compute. We know, for example, that if Y is a proportion, $E(\mu) = \pi$, the proportion of "1s" in the data (which can also be thought of as the unconditional probability that any one observation will have $Y = 1$) and:

$$\sigma_\mu^2 = \frac{\pi(1-\pi)}{\mathfrak{N}}$$

This means that, in the sample data, $\hat{\pi} = \bar{Y}$,

$$\begin{aligned} s^2 &= \frac{\hat{\pi}(1-\hat{\pi})}{N} \\ &= \frac{\bar{Y}(1-\bar{Y})}{N}, \end{aligned}$$

and s is the square root of this term. For our two samples defined by the values of X, we then have:

$$s_0 = \sqrt{\frac{\bar{Y}_0(1-\bar{Y}_0)}{n_0}}$$

and

$$s_1 = \sqrt{\frac{\bar{Y}_1(1-\bar{Y}_1)}{n_1}}.$$

That makes calculation of (5) straightforward. Moreover, in large samples, the difference of proportions is distributed asymptotically normally (which is equivalent to t with a very large number of degrees of freedom). This means that we can construct confidence intervals and conduct hypothesis tests in the usual way, using a z (standard normal) distribution.

Example: Africa (2001) Data

One worked-through example: `adrate`, by `subsaharan`.

```
> stat.desc(Africa$adrate)
     nbr.val     nbr.null       nbr.na          min          max
       43.00         0.00         0.00         0.10        38.80
       range          sum       median         mean      SE.mean
       38.70       402.70         6.00         9.37         1.52
CI.mean.0.95          var      std.dev     coef.var
        3.07        99.21         9.96         1.06
```

By values of X:

```
> with(Africa[Africa$subsaharan=="Not Sub-Saharan",],
+      stat.desc(adrate))
     nbr.val     nbr.null       nbr.na          min          max
       6.000        0.000        0.000        0.100        2.800
       range          sum       median         mean      SE.mean
       2.700        7.600        1.000        1.267        0.525
CI.mean.0.95          var      std.dev     coef.var
       1.350        1.655        1.286        1.016

> with(Africa[Africa$subsaharan=="Sub-Saharan",],
+      stat.desc(adrate))
     nbr.val     nbr.null       nbr.na          min          max
       37.00         0.00         0.00         0.10        38.80
       range          sum       median         mean      SE.mean
       38.70       395.10         7.20        10.68         1.67
CI.mean.0.95          var      std.dev     coef.var
        3.38       102.81        10.14         0.95
```

So:

$$\bar{Y}_1 - \bar{Y}_0 = 9.41$$

and

$$
\begin{aligned}
s^2_{\bar{Y}_1-\bar{Y}_0} &= \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1} \\
&= \frac{1.655}{6} + \frac{102.8}{37} \\
&= 0.28 + 2.78 \\
&= 3.06
\end{aligned}
$$

and:

$$
\begin{aligned}
s_{\bar{Y}_1-\bar{Y}_0} &= \sqrt{3.06} \\
&= 1.75.
\end{aligned}
$$

Then the t-statistic for $\bar{Y}_1 - \bar{Y}_0$ (assuming $k_0 = 0$; that is, that $H_0 : \bar{Y}_1 - \bar{Y}_0 = 0$) is

$$
\begin{aligned}
t &= \frac{9.41 - 0}{1.75} \\
&= \mathbf{5.38}
\end{aligned}
$$

Looking at a t-table yields a P-value that is far less than even 0.001; this is confirmed in the software, below.

R Results: t-tests

```
> with(Africa, t.test(adrate~subsaharan))

Welch Two Sample t-test

data:  adrate by subsaharan
t = -5, df = 40, p-value = 0.000003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.94  -5.88
sample estimates:
mean in group Not Sub-Saharan     mean in group Sub-Saharan
                         1.27                         10.68
```

We can do the same for the POLITY and `internalwar` variables as well:

```
> t.test(polity~subsaharan)

Welch Two Sample t-test
```

```
data:  polity by subsaharan
t = -9.644, df = 39.99, p-value = 5.452e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.856 -6.441
sample estimates:
mean in group Not Sub-Saharan     mean in group Sub-Saharan
                       -6.500                         1.649
```

```
> t.test(internalwar~subsaharan)

Welch Two Sample t-test

data:  internalwar by subsaharan
t = -0.8567, df = 7.382, p-value = 0.4185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5883  0.2730
sample estimates:
mean in group Not Sub-Saharan     mean in group Sub-Saharan
                       0.1667                        0.3243
```