PLSC 502 – Autumn 2016
Two-Group Comparisons, I

October 20, 2016

"The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name).

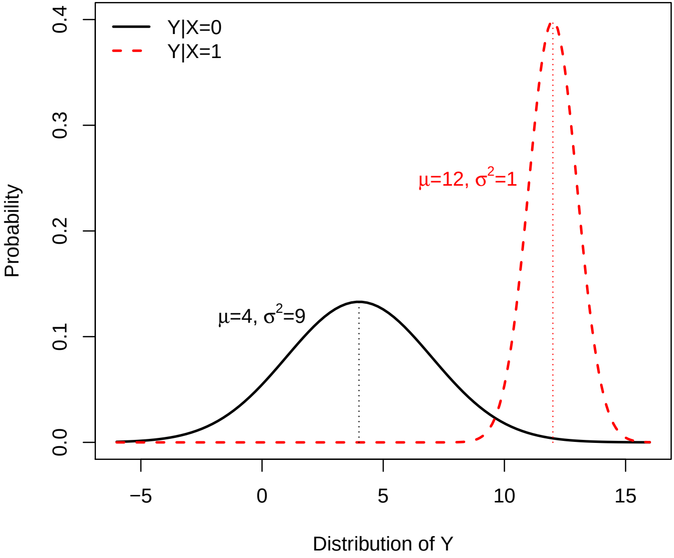
Gosset had been hired due to Claude Guinness's policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness's industrial processes. Gosset devised the *t*-test as an economical way to monitor the quality of stout. The Student's *t*-test work was submitted to and accepted in the journal *Biometrika* and published in 1908. Company policy at Guinness forbade its chemists from publishing their findings, so Gosset published his statistical work under the pseudonym "Student"."

- Student's *t*-test (Wikipedia)

The Setup

- N observations, $i \in \{1, 2, ...N\}$
- A dichotomous predictor X, so that $X_i \in \{0, 1\}$
- n_0 and n_1 are the number of observations in the data with $X = 0$ and $X = 1$, respectively (so $n_0 + n_1 = N$)
- An continuous (interval/ratio) outcome variable Y, with
  - $Y|X = 0 \sim N(\mu_0, \sigma_0^2)$ and
  - $Y|X = 1 \sim N(\mu_1, \sigma_1^2)$.
- Call
  - $\bar{Y}_0 = \bar{Y}|X = 0$, and
  - $\bar{Y}_1 = \bar{Y}|X = 1$

Example


Y|X=0
Y|X=1
μ=12, σ²=1
μ=4, σ²=9
Probability
0.0
0.1
0.2
0.3
0.4
−5
0
5
10
15
Distribution of Y

Difference of Means

Difference of (sample) means:

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

Has:

$$E(\bar{Y}_1 - \bar{Y}_0) = \mu_1 - \mu_0$$

and

$$Var(\bar{Y}_1 - \bar{Y}_0) = \sigma^2_{\mu_1 - \mu_0}.$$

Difference of Means (continued)

Can show that:

$$\sigma^2_{\mu_1 - \mu_0} = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

In practice we use:

$$s^2_{\bar{Y}_1 - \bar{Y}_0} = \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}$$

The t Statistic

$$
\begin{aligned}
t &= \frac{\bar{Y}_1 - \bar{Y}_0}{s_{\bar{Y}_1 - \bar{Y}_0}} \\
&= \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}}
\end{aligned}
$$

Can show that:

$$t \sim t(\nu)$$

where

$$\nu \approx \frac{\left(\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1} \right)^2}{\frac{s_0^4}{n_0^2(n_0-1)} + \frac{s_1^4}{n_1^2(n_1-1)}}$$

Remember...

- Assumes $Y \sim N(\mu, \sigma^2)$
- Note that if $s_0^2 = s_1^2$, then $\nu = n_0 + n_1 - 2$.
- $\nu = n_0 + n_1 - 2$ is also good if n_0 and $n_1 > 50$ or so

Test statistic for $H_0: \ \mu_1 - \mu_0 = k$:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_0) - k}{s_{\bar{Y}_1 - \bar{Y}_0}}$$

The $(1 - \alpha) \times 100$ c.i. for $\bar{Y}_1 - \bar{Y}_0$ is:

$$(\bar{Y}_1 - \bar{Y}_0) \pm t_{\alpha/2}(s_{\bar{Y}_1 - \bar{Y}_0}),$$

t Mnemonics

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. $= \infty$. * indicates that the directionality is "correct."

Differences of Proportions

For a proportion:

$$E(\mu) = \pi$$

and

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

So $\hat{\pi} = \bar{Y}$ and:

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

For two samples:

$$s_0 = \sqrt{\frac{\bar{Y}_0(1-\bar{Y}_0)}{n_0}} \quad \text{and} \quad s_1 = \sqrt{\frac{\bar{Y}_1(1-\bar{Y}_1)}{n_1}}$$

Example: Africa (2001) Data

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

By subsaharan

```
> 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
```

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

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:

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

t-test (via R)

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

Welch Two Sample t-test

data:  adrate by subsaharan
t = -5.4, df = 41, p-value = 0.000003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.942  -5.881
sample estimates:
mean in group Not Sub-Saharan     mean in group Sub-Saharan
                        1.267                        10.678
```

Another t-test: Literacy

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

Welch Two Sample t-test

data:  literacy by subsaharan
t = -0.77, df = 8.4, p-value = 0.5
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.35  10.12
sample estimates:
mean in group Not Sub-Saharan     mean in group Sub-Saharan
                        55.67                         60.78
```