Comparing two means

Math 530/630

Alison Presmanes Hill

2017-11-16

The *t*-test

The *t*-test is versatile procedure that we have seen can be used to test

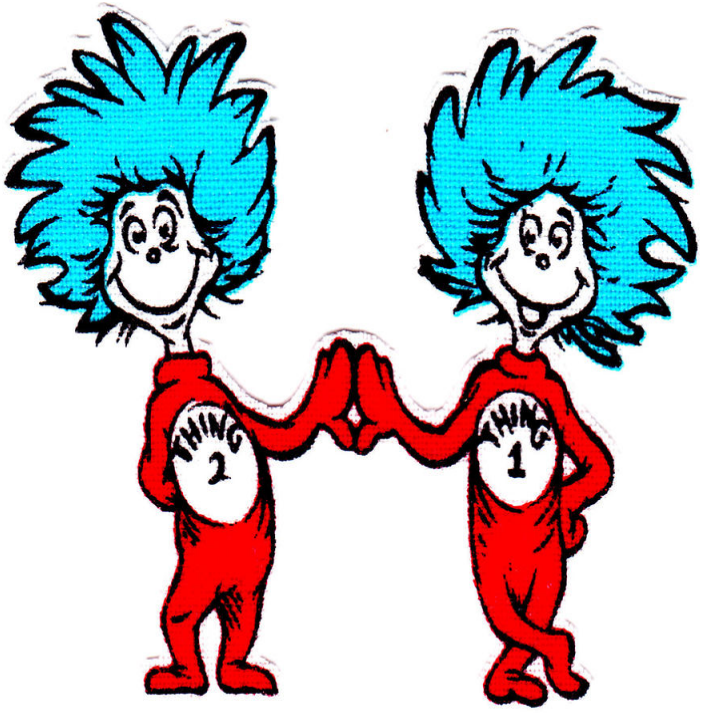
1. whether a regression coefficient is different from zero,
2. whether a sample mean is different from a population mean, and
3. whether two means are different from each other.

Family of *t*-tests

- One-sample *t*-test
- Independent samples *t*-test
- Dependent samples *t*-test (also known as paired)

What do they all have in common?

We want to compare two things- mean 1 and mean 2!

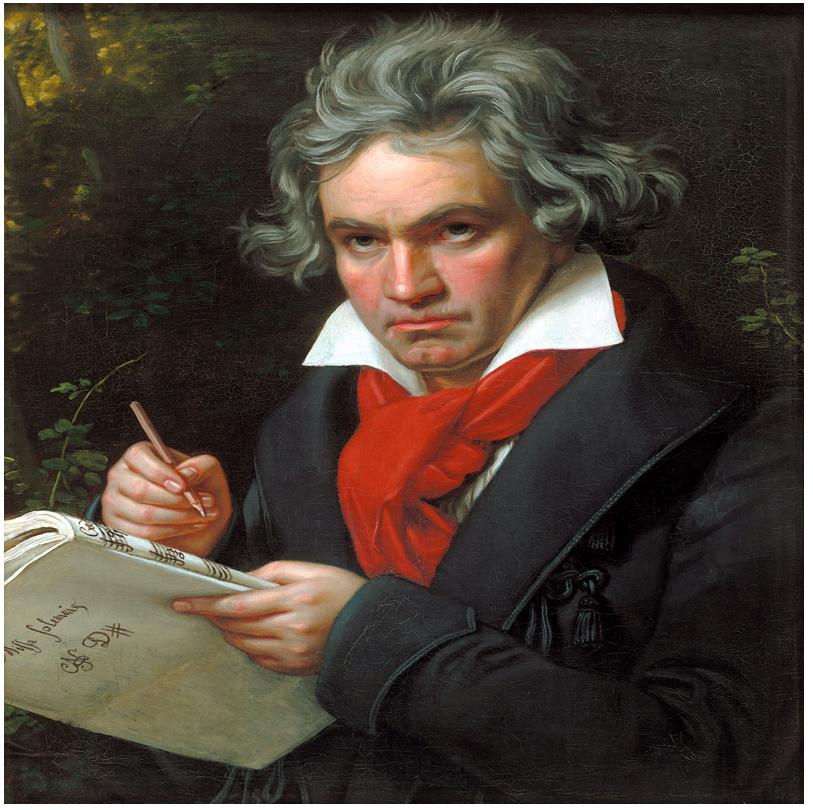


Research questions for a *t*-test

Was *Scream 2* scarier than the original *Scream*? You could measure heart rates (which indicate anxiety) during both films and compare them.

Research questions for a *t*-test

Does listening to classical music improve your writing? You could get some people to write an essay while listening to classical music versus silence and compare essay quality.



Ways to get independent samples

- Random assignment of participants to two different experimental conditions
  - Scream versus Scream 2
  - Classical music versus silence
- Naturally occurring assignment of participants to two different groups
  - Male versus female
  - Young versus old

Formula for the independent groups *t*-test

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE}$$

Generally,

> $H_0 : \mu_1 - \mu_2 = 0$ and $H_1 : \mu_1 - \mu_2 \neq 0$

So $\mu_1 - \mu_2$ is often excluded from the formula. Since we now have two sample means and therefore two sample variances, we need some way to combine these two variances in a logical way. The answer is to **pool** the variances to estimate the SE of the difference, $(\bar{y}_1 - \bar{y}_2)$:

$$s_{\bar{y}_1 - \bar{y}_2} = s_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and s_{pooled} is:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_{y_1}^2 + (n_2 - 1)s_{y_2}^2}{n_1 + n_2 - 2}}$$

Rationale for the *t*-test

- Two samples of data are collected and the sample means are calculated.
- If both samples come from the same population, then we would expect their means to be approximately equal to each other. Although they may differ by chance, we would expect large differences between sample means to happen infrequently. Under the null hypothesis, we assume that the two groups are not different from each other.
- We compare the two sample means to see if the difference is more than we would expect to get by chance under the null hypothesis. At α=.05, we therefore seek evidence that there is only a 5% chance that the magnitude of the difference we observe is consistent with what we would expect based on chance alone.
  Remember: The *p*-value does not tell you if the result was due to chance. It tells you whether the results are consistent with being due to chance. These two things are not the same.
- We use the standard error as the gauge of variability of the sample means. If it is small, we expect most samples to have very similar means. If it is large, then large differences in sample means are more likely.

Rationale for the *t*-test, continued

If the difference between the sample means is larger than we would expect based on the standard error, then we know that one of two things has happened:

- **We made a mistake (boo!).** There is no difference between the groups and our sample means fluctuate a lot. By chance, we have collected two samples that are atypical of the population we drew from. Here, the difference is a fluke and the null is true; we incorrectly reject the true null hypothesis and thus commit a Type I error.
- **We made a discovery (yay!).** The two samples come from different populations that are each typical of their respective populations. Here, the difference is genuine, and we correctly reject the null hypothesis.

The *t*-test as a general linear model (GLM)

All statistical procedures are basically the same thing:

$$outcome_i = (model) + error_i$$

In a simple linear regression my example will be:

$$prestige_i = (intercept + education_i) + error_i$$

Or more generally:

$$y_i = (b_0 + b_1 x_i) + error_i$$

Sidebar: The GLM in matrix notation

The GLM we have been dealing with thus far includes just one independent variable and thus just one b_1. However, the full GLM is better represented as a matrix

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where...

- $\boldsymbol{Y}$ is the *response vector* of length *N*;
- $\boldsymbol{\epsilon}$ is the *error vector* of length *N*;
- $\boldsymbol{\beta}$ is the vector of parameters of length *p*+1 where *p* is the number of IVs and the 1 accounts for the intercept;
- and $\boldsymbol{X}$ is called the *design matrix* consisting of a matrix of *N* rows and *p+1* columns

Design matrices with one independent variable

In both simple linear regression and the independent samples *t*-test, $\boldsymbol{X}$ is a matrix of *N* rows and *2* columns. Note that the number of columns in $\boldsymbol{X}$ must always equal the number of rows in β.

here

In simple linear regression, the vector *X* can take on any value. In the independent samples *t*-test, this vector simply contains 0's and 1's. Below, *n*=4 with 2 in each group to illustrate.

here

Dummy variables

Now, let's switch IVs in our example from education (a continuous variable) to a group variable:

$$prestige_i = (intercept + group_i) + error_i$$

where group is a dummy or indicator variable that can only take two values: 0 or 1.

This is easy to do in R. Remember in my Prestige dataset, I have a categorical variable called "type." Let's let the variable Group denote:

> 0 for blue collar (type="bc") and
>
> 1 for white collar (type="wc")

Creating a dummy variable in R

```
library(car)
table(Prestige$type)
```

```
##
##   bc prof   wc
##   44   31   23
```

```
Prestige$group <- ifelse(Prestige$type=="bc",0, ifelse(Prestige$type=="wc",1,NA))
table(Prestige$type, Prestige$group)
```

```
##
##          0  1
##   bc    44  0
##   prof   0  0
##   wc     0 23
```

Cleaning up the new dataset

```
Prestige.2 <- subset(Prestige, group %in% c(0,1))
Prestige.2$type <- droplevels(Prestige.2$type) #get rid of type=prof
```

OK, we are all set now with two groups

```
table(Prestige.2$type, Prestige.2$group)
```

```
##
##       0  1
##   bc 44  0
##   wc  0 23
```

```
head(Prestige.2)
```

```
##                     education income women prestige census type group
## nursing.aides            9.45   3485 76.14     34.9   3135   bc     0
## medical.technicians     12.79   5180 76.04     67.5   3156   wc     1
## radio.tv.announcers     12.71   7562 11.15     57.6   3337   wc     1
## secretaries             11.59   4036 97.51     46.0   4111   wc     1
## typists                 11.49   3148 95.97     41.9   4113   wc     1
## bookkeepers             11.32   4348 68.24     49.4   4131   wc     1
```

Plotting data for two groups

Not surprisingly, mean prestige ratings appear to be higher among white collar workers than blue collar workers.

But are the two groups different?

Let's do an independent samples *t*-test to find the answer:

```
t.test(prestige~group,data=Prestige.2, var.equal=T)
```

```
##
##     Two Sample t-test
##
## data:  prestige by group
## t = -2.6487, df = 65, p-value = 0.01013
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.780279  -1.652132
## sample estimates:
## mean in group 0 mean in group 1
##        35.52727        42.24348
```

What happens if I now run a linear regression?

Kidding. Here's is the linear regression summary...

```
fit <- lm(formula = prestige ~ group, data = Prestige.2)
summary(fit)
```

```
##
## Call:
## lm(formula = prestige ~ group, data = Prestige.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2273  -7.0273  -0.2273   6.8227  25.2565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.527      1.486  23.914   <2e-16 ***
## group          6.716      2.536   2.649   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.855 on 65 degrees of freedom
## Multiple R-squared:  0.09742,    Adjusted R-squared:  0.08353
## F-statistic: 7.016 on 1 and 65 DF,  p-value: 0.01013
```

But what does it MEAN?

Let's do a quick review of the linear regression formula:

$$y_i = (b_0 + b_1 x_i) + error_i$$

Solving for the intercept term, for example, we need to apply some summation algebra:

$$\frac{1}{n}\sum_i^n y_i = \frac{1}{n}\sum_i^n b_0 + \frac{1}{n}\sum_i^n b_1 x_i + \frac{1}{n}\sum_i^n error_i$$

A few reminders about summation algebra in this context:

- The mean is defined as: $\frac{1}{n}\sum_i^n y_i$
- The sum of a constant is just *n* times the constant: $\sum_i^n b_0 = n \times b_0$
- The sum of a constant times a random variable is the constant times the sum of the variable: $\sum_i^n b_1 x_i = b_1 \sum_i^n x_i$
- By definition, in GLM, we assume the mean of the error is 0: $\frac{1}{n}\sum_i^n error_i = 0$

Solving for the regression coefficients

Applying the summation algebra from the previous slide, we get:

$$\frac{1}{n}\sum_{i}^{n} y_i = \frac{1}{n} n b_0 + \frac{1}{n} b_1 \sum_{i}^{n} x_i$$

$$\bar{y} = b_0 + b_1 \bar{x}$$

This should look familiar! The formula for the regression intercept term, b_0, is:

$$b_0 = \bar{y} - b_1 \bar{x}$$

But look again: we also have the formula for $\bar{y}$. The intercept term in linear regression is the expected mean value of y when x_i=0.

Solving for $\bar{x}$

$$\bar{y} = b_0 + b_1\bar{x}$$

But, you protest, how do we calculate $\bar{x}$?
Is it the mean of the 0/1 values in *x*?
What is happening?

The intercept

For the *t*-test, we can solve for the intercept just as we can for the simple linear regression. Remember our formula for $\bar{y}$:

$$\bar{y} = b_0 + b_1\bar{x}$$

Let's re-write it two ways:

$$\bar{y}_{x=0} = b_0 + b_1\bar{x}_0$$

$$\bar{y}_{x=1} = b_0 + b_1\bar{x}_1$$

Start with the top formula: What is $\bar{x}_0$? This is simple to think about in matrix notation- the mean of a vector of 0's is 0. So, when the group variable is equal to zero (blue collar)...

$$\bar{y}_{bc} = (b_0 + b_1 \times 0) = b_0$$

Therefore, b_0 (the intercept term) is equal to the mean prestige score of the blue collar group (the one coded as 0).

The slope

Now, let's tackle the second formula:

$$\bar{y}_{x=1} = b_0 + b_1\bar{x}_1$$

When the group variable is equal to 1 (white collar), $\bar{x_1} = 1$ because the mean of a vector of 1's is 1.

$$\bar{y}_{wc} = (b_0 + b_1 \times 1)$$

$$\bar{y}_{wc} = b_0 + b_1$$

$$\bar{y}_{wc} = \bar{y}_{bc} + b_1$$

Solving for b_1:

$$b_1 = \bar{y}_{wc} - \bar{y}_{bc}$$

Therefore, b_1 (the slope) is equal to the difference between group means in prestige scores.

What does this mean?

We could represent a two-group experiment as a regression equation in which the regression coefficient b_1 is equal to the difference between group means and the intercept term b_0 is the mean of the group coded as 0.

Our independent samples *t*-test would take the form:

$$y_i = \bar{y}_{bc} + (\bar{y}_{wc} - \bar{y}_{bc})x_i + error_i$$

Think of it this way: the regression line must pass through these two points:

- (0, $\bar{y}_{bc}$)
- (1, $\bar{y}_{wc}$)

Trust but verify

```
y_bc <- mean(Prestige.2$prestige[Prestige.2$group==0])
y_wc <- mean(Prestige.2$prestige[Prestige.2$group==1])
diff <- y_wc-y_bc
cbind(y_bc,y_wc, diff)
```

```
##          y_bc     y_wc     diff
## [1,] 35.52727 42.24348 6.716206
```

So, y_bc is b_0 and diff is b_1...right?

```
coef(fit)
```

```
## (Intercept)       group
##   35.527273    6.716206
```

The General Linear Model

A number of different statistical models are extensions of this same idea of a GLM:

- Ordinary least squares (OLS) linear regression (simple and multiple): 1+ predictors may be continuous or factors
- *t*-test: a *t*-test is basically a regression model where the 1 predictor is a factor with exactly 2 levels
- Analysis of Variance/Covariance (ANOVA/ANCOVA): an ANOVA is basically a regression model where the 1+ predictors are factors
- Multivariate Analysis of Variance/Covariance (MANOVA/MANCOVA): a MANOVA is basically a regression model with 2+ DVs where the 1+ predictors are factors

Further food for thought...

The independent samples *t*-test is a special case of an ANOVA. Specifically, a one-way ANOVA (definition: 1 IV that is a factor and 1 DV that is continuous) in which the IV factor has *exactly* two levels.

Again, trust but verify!

```
anova(lm(prestige~group,data=Prestige.2)) #doing an ANOVA in R
```

```
## Analysis of Variance Table
##
## Response: prestige
##           Df Sum Sq Mean Sq F value  Pr(>F)
## group      1  681.3  681.32  7.0156 0.01013 *
## Residuals 65 6312.5   97.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(hint: square the *t*-statistic to get the *F*-statistic equivalent)

The non-centrality parameter of the *t* distribution

In an earlier class, I alluded to the fact that the *t*-distribution has an another parameter in addition to the degrees of freedom- a non-centrality parameter, δ.

> When H_0 is true, $\delta = 0$.
>
> When H_0 is false, $\delta \neq 0$.

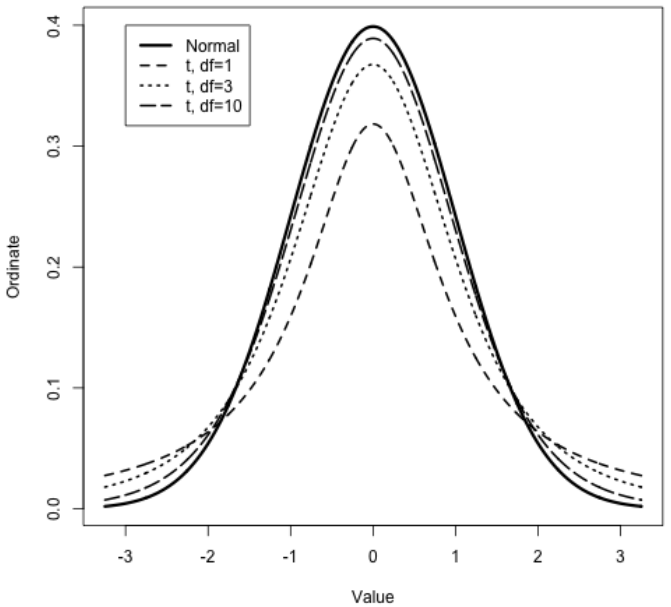
Why is this? Let's look at the formula:

$$\delta = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\frac{\mu_1 - \mu_2}{SE})$$

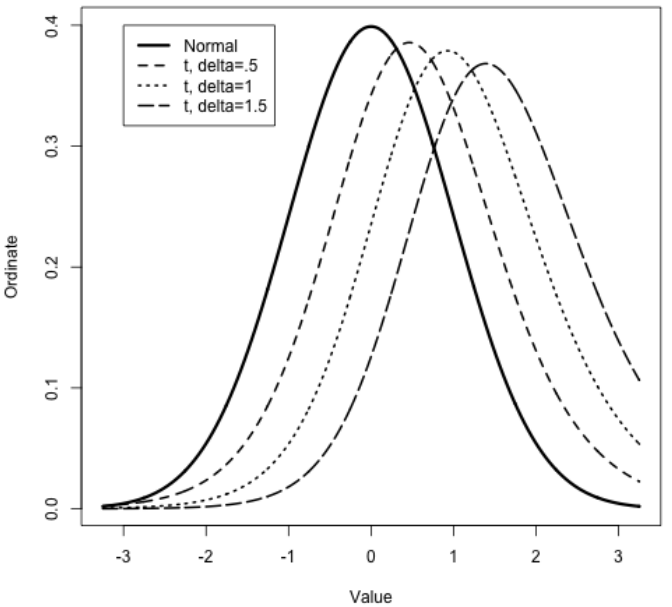
As you can see, if $\mu_1 - \mu_2$=0 then δ=0.

The *t*-distribution...

When the null is true

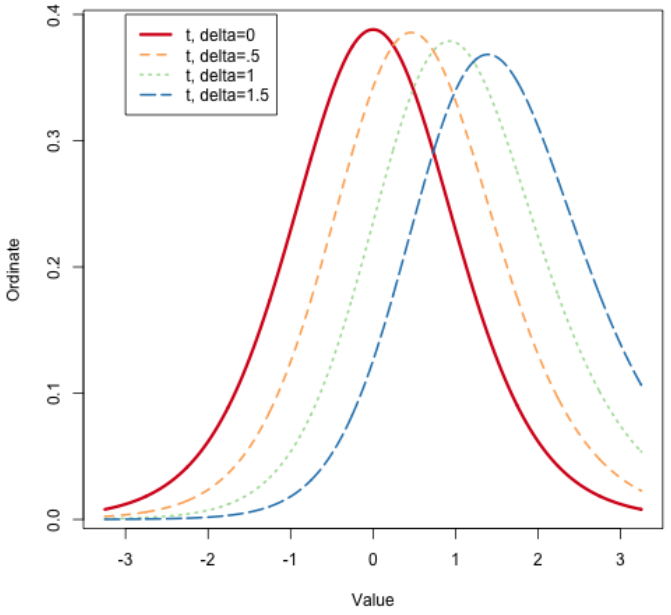


When the null is false

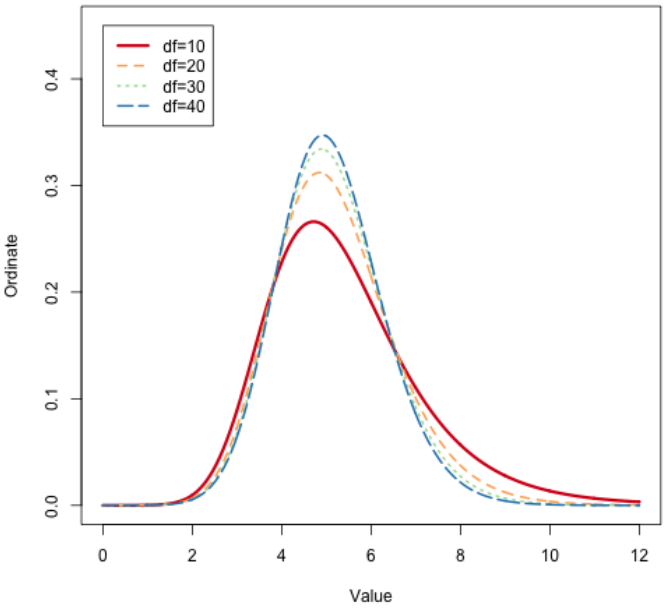


Non-central *t*-distributions...

Degrees of freedom=9



δ=5



Using non-central t-distributions...

```
help(TDist)
pt(-1, 20, 0)
```

```
## [1] 0.1646283
```

```
pt(-1, 20, 5)
```

```
## [1] 1.624699e-09
```

Recommended:

```
power.t.test()
```

Welch's *t*-test: Dealing with unequal variances

Recall the formula for the independent groups *t*-test:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$$

In Welch's formula, we calculate the SE differently:

$$SE' = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

So the formula for Welch's t':

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Welch's *t*-test: Modified degrees of freedom

Recall the degrees of freedom for the independent groups *t*-test:

$$\nu = n_1 + n_2 - 1$$

The degrees of freedom are modified for Welch's t':

$$\nu' = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

George E. P. Box

> "Equally, the statistician knows, for example, that in nature there never was a normal distribution. There never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world." - *Journal of the American Statistical Association, 71*(356), pp. 791-799.