Statistics and Linguistic Applications
Hale
March 15, 2010

Review of the t test

Previously, we saw that the t statistic — something you can compute from a sample — has many uses.

one sample t test	can judge whether to reject the null hypothesis that μ is some particular constant value
two sample t test	can judge whether to reject the null hypothesis that $\mu_1 = \mu_2$

Imagine now that you wish to do a more complicated experiment; one that systematically varies **levels** of a **factor**. Inquiring minds might wish to know whether or not the HARSHNESS of the advisor affects the amount of TEARS cried by the advisee. Say you decide to sample four different levels of advisor HARSHNESS: `ambivalent`, `annoying`, `mean`, `vicious`.

All possible two sample t tests on these data provide $\binom{4}{2} = 6$ opportunities to make a Type I error. Say we set a significance level of $\alpha = 0.05$ for each t test. The probability rejecting *at least one* null hypothesis is just the complement of the probability of not rejecting *any*. This is like a binomial where you have only a 5% chance of "succeeding." The probability of having zero "successes" of this type is

```
> pbinom(0, 6, 0.05)

[1] 0.7350919
```

Thus the probability of not rejecting any null hypothesis on the basis of the six separate t tests is $1 - 0.735 = 0.265$. The odds of mistakenly rejecting when in fact the null is true are inflated from 0.05 to 0.265 by these multiple comparisons. With more levels, the problem gets worse (Vasishth 5.1).

Stepping back for a second, it's not that we want to find just any difference between arbitrary levels; in this sort of experiment the real goal is typically to determine — overall — whether increasing advisor-HARSHNESS leads to more crying on the part of the graduate student.

The ANOVA

The ANOVA solves this problem by introducing a new test statistic with a known distribution whose wackiness we can quantify. Continue imagining that your sadistic experiment manipulates the role of the independent variable HARSHNESS while measuring the dependent variable TEARS. The resulting data could be organized as in Table 1.

		tears		
harshness	`ambivalent`	within	$= \bar{x}_{\texttt{ambivalent}}$	between
	`annoying`	within	$= \bar{x}_{\texttt{annoying}}$	
	`mean`	within	$= \bar{x}_{\texttt{mean}}$	
	`vicious`	within	$= \bar{x}_{\texttt{vicious}}$	

Table 1: One-way ANOVA scenario

The average amount of tears elicited by any treatment level, throughout the whole experiment is denoted $\bar{x}$ the **Grand Mean**. The big idea of ANOVA is that deviations from the Grand Mean can be "explained" in two different ways. A deviation could come about because individuals i just are randomly different in terms of their ability to not-cry when mistreated (the red variation). Or a deviation from the Grand Mean could be associated with the transition to ever-harsher treatments j. For instance, ratcheting up the cruelty from `mean` to `vicious` might globally increase the average number of tears in that condition for nearly all participants (the yellow variation). This second story is a fundamentally different explanation for deviations $x_{ij} - \bar{x}$ because it invokes the treatments, whereas the first story doesn't.

Vasishth shows (in 5.5) how rewriting deviations from the Grand Mean in a way that mentions the treatment means x_j

$$x_{ij} - \bar{x} = (\bar{x}_j - \bar{x}) + (x_{ij} - \bar{x}_j)$$

leads to an expression for total variability in terms of two addends, one for each potential source of variability, i.e. associated with either the red or yellow explanations. The test statistic, F is the ratio of these two variances. On $\mathcal{H}_0$ the F-ratio should be 1.00. That is, on the null hypothesis, variation (quantified as the **mean square**) between treatments is about the same as variation within treatments.

$$\begin{array}{ll} \mathcal{H}_0 & X_{ij} = \mu + \epsilon_{ij} \\ \mathcal{H}_1 & X_{ij} = \mu + \alpha_j + \epsilon_{ij} \end{array} \qquad \mathcal{H}_0 : \frac{MS_{\text{between}}}{MS_{\text{within}}} = 1$$

Under the alternative hypothesis, the observations are centered around a true mean, μ with some variation due to the being in group/level j, as well as a normally-distributed noise or "error" term ϵ. Crucially, the null hypothesis does not recognize any contribution made by HARSHNESS level j.

The ratio of Mean Squares follows the F distribution, the characteristic shape of the quotient of two variances. So a little wiggle room is permitted — but not much. Figure 1 (repeated from Vasishth 5.16) illustrates the case where the three treatment levels really do have the same mean. In ten thousand rounds of simulation, it is very rare that sampling error leads to ratios very far removed from 1.00.

```
> ss <- function(sample) {
+     m <- rep(mean(sample), length(sample))
+     m2 <- (sample - m)^2
+     result <- sum(m2)
+     result
+ }
> mswithin <- function(s1, s2, s3) {
+     N <- sum(length(s1), length(s2), length(s3))
+     DF <- N - 3
+     msw <- (ss(s1) + ss(s2) + ss(s3))/DF
+     msw
+ }
> msbetween <- function(s1, s2, s3) {
+     gm <- mean(c(s1, s2, s3))
+     m1 <- mean(s1)
+     m2 <- mean(s2)
+     m3 <- mean(s3)
+     msb <- (length(s1) * (m1 - gm)^2 + length(s2) * (m2 - gm)^2 +
+         length(s3) * (m3 - gm)^2)/2
+     msb
+ }
> Fratio <- function(msbetween, mswithin) {
+     Fvalue <- msbetween/mswithin
+     Fvalue
+ }
> pop1 <- rnorm(1000, mean = 60, sd = 4)
> pop2 <- rnorm(1000, mean = 60, sd = 4)
> pop3 <- rnorm(1000, mean = 60, sd = 4)
> Fs <- c()
> for (i in c(1:1000)) {
+     s1 <- sample(pop1, 11)
+     s2 <- sample(pop2, 15)
+     s3 <- sample(pop3, 20)
+     Fs <- append(Fs, msbetween(s1, s2, s3)/mswithin(s1, s2, s3))
+ }
```

```
> plot(density(Fs), xlim = range(0, 8))
```

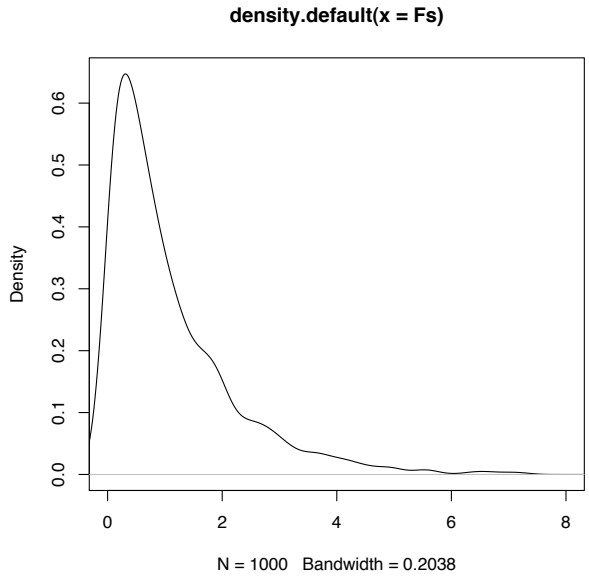


Figure 1: When $\mathcal{H}_0$ is true, the vast majority of samples exhibit an F-ratio close to 1.00.

Using R to do an ANOVA

To analyze variance with an ANOVA is simply to do a hypothesis test examining whether a model *without* the extra term corresponding to a treatment effect (α_j) fits as well has one that has such a term. If so, the treatment fails to account for enough variability to declare that there really is an effect. One compares the models:

$$\begin{array}{lll} \mathcal{H}_0 & X_{ij} = \mu + \epsilon_{ij} & \text{Model 0} \\ \mathcal{H}_1 & X_{ij} = \mu + \alpha_j + \epsilon_{ij} & \text{Model 1} \end{array}$$

where ϵ_{ij} are Normally distributed 'errors' with mean 0 and variance σ^2.

These models can (almost) literally be entered into R using **model formulas**, explained in much more detail in section 11 of Introduction to R. Briefly, if x stores all of the measured data, and f is a factor, then the model formula

```
x~f
```

corresponds to Model 1, namely

$$X_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

where the grand mean μ and the residual error ϵ are taken for granted. This model formula is read "the response `x` is modeled as the factor `f`."

Synset cardinality as a function of Dutch auxiliary verb

Harald Baayen provides some nice data on the choice of auxiliary verb that Dutch lexical verbs prefer.

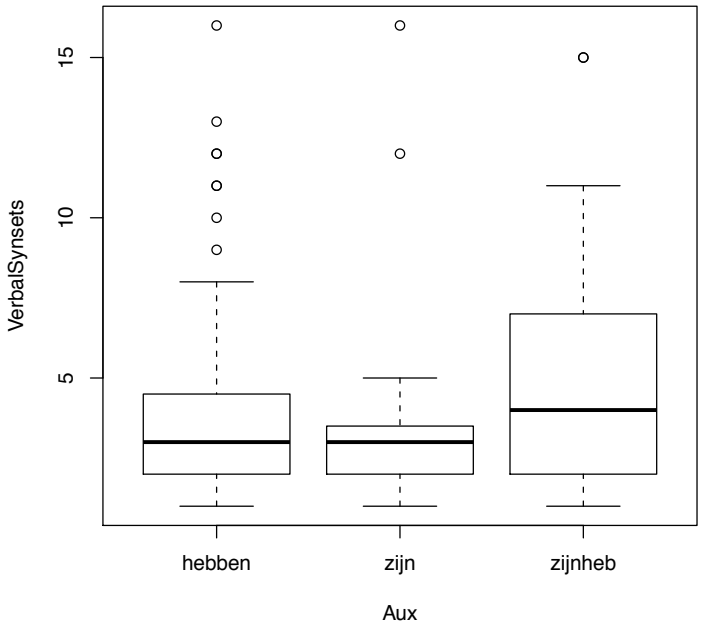
```
> library(languageR)
> head(auxiliaries)
```

```
     Verb     Aux VerbalSynsets Regularity
1 blijken    zijn             1  irregular
2 gloeien  hebben             3    regular
3 glimmen zijnheb             2  irregular
4  rijzen    zijn             4  irregular
5  werpen  hebben             3  irregular
6  delven  hebben             2  irregular
```

Main verbs select either for *zijn* ('be'), *hebben* ('have') or both. Does changing subcategorization frame affect the number of synonyms a verb has[1]? The null hypothesis is that verbs' preference for one auxiliary over another has nothing to do with a semantic property like synonym cardinality.

$$\begin{array}{ll} \mathcal{H}_0: & \mu_{hebben} = \mu_{zijn} = \mu_{zijheb} \\ \mathcal{H}_1: & \mu_j \neq \mu_j \text{ for at least one } i, j \in \{\text{zijn}, \text{hebben}, \text{zijnhebben}\} \end{array}$$

```
> plot(VerbalSynsets ~ Aux, data = auxiliaries)
```



One way to make R carry out an ANOVA is with the `aov` function. It takes a model formula as input and returns a fitted model which you can query using `summary`.

```
> result <- aov(VerbalSynsets ~ Aux, data = auxiliaries)
> summary(result)
```

```
             Df Sum Sq Mean Sq F value    Pr(>F)
Aux           2  117.8  58.901  7.6423 0.0005859 ***
Residuals   282 2173.4   7.707
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

[1] In the WordNet project, a synset is a collection words that are of truth-functionally equivalent in some context.

The first line of R's ANOVA summary table concerns the Aux factor. There are three levels of treatment, so $k - 1 = 2$ degrees of freedom with this factor; dividing by this *df* turns the sum of squares in a variance. The second line has to do with the remaining variation attributed to ϵ (rather than α). The ratio of the Mean Squares is the F value. The p-value, the probability of getting an F-ratio this extreme or more given $\mathcal{H}_0$ is also listed with an asterisk for 'significant' because the null hypothesis can be rejected as an event that would only occur by chance less than 1 in 20 times.

```
> attach(auxiliaries)
> wit <- mswithin(VerbalSynsets[Aux == "hebben"], VerbalSynsets[Aux ==
+     "zijn"], VerbalSynsets[Aux == "zijnheb"])
> bet <- msbetween(VerbalSynsets[Aux == "hebben"], VerbalSynsets[Aux ==
+     "zijn"], VerbalSynsets[Aux == "zijnheb"])
> 1 - pf((bet/wit), 2, 282)

[1] 0.00058587
```

Planned vs. post-hoc comparisons

Within a factorial design, an investigator might choose to examine particular levels judge whether their means are different enough to reject the null hypothesis that they vary only by luck. However it matters a lot whether this examination is clearly motivated by the hypothesis or is simply a fishing expedition trolling for a significant result. The former are "planned comparisons" — these could be significant even in the absence of a significant F ratio. The latter a "post hoc comparisons" which are only carried out in response to observing a significant F ratio, which justifies rejecting the null hypothesis that all the means are *the same*.

But as the first section of this handout argued, willy-nilly post hoc testing increases the probability of Type I error. As the number c of conditions goes up, the number of pairwise comparisons goes up as $\frac{c/(c-1)}{2}$. Baayen mentions the Bonferroni correction, which amounts to dividing all the significance levels α by the number of pairwise comparisons. This is often too stringent for practical use, for instance in a typical fMRI study where the question is "do any voxels differ from the others?"

Baayen pp115-116 demonstrate Tukey's "Honestly Significant Difference" on the warpbreaks dataset available in R. The significant F ratio reveals that, as the tension factor varies from "high" to "medium" to "low" the mean number of breaks doesn't stay the same. Tukey's HSD shows that the contrast between "high" tension and "low" tension is significant at the adjusted $p < 0.0014315$ level. Probably the switch from high to medium doesn't affect the number of breaks in a systematic way.

```
> attach(warpbreaks)
> warpbreaks.aov <- aov(breaks ~ tension, data = warpbreaks)
> summary(warpbreaks.aov)

            Df Sum Sq Mean Sq F value   Pr(>F)
tension      2 2034.3 1017.13  7.2061 0.001753 **
Residuals   51 7198.6  141.15
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

> TukeyHSD(warpbreaks.aov)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = breaks ~ tension, data = warpbreaks)

$tension
          diff       lwr        upr     p adj
M-L -10.000000 -19.55982 -0.4401756 0.0384598
H-L -14.722222 -24.28205 -5.1623978 0.0014315
H-M  -4.722222 -14.28205  4.8376022 0.4630831
```

Now You Try

1. I really want my result to come out. And my F-ratio looks pretty high. But the computer says the p-value is 0.35, whereas the standard for my sub-field is $p < 0.05$. What sort of skeptical argument am I vulnerable to if I go ahead and publish anyway?

2. Determine whether the following F ratios are significant at the 5% level: $F_{10,20} = 2.80$; $F_{2,8} = 3.10$; $F_{2,80} = 3.10$; $F_{2,800} = 3.10$; $F_{5,100} = 2.25$; $F_{10,100} = 2.25$

3. Carry out a one-way ANOVA on the result of a language test carried out at three different testing centers, A B and C. The data are posted as `wfr223.txt`. Comment on whatever result you obtain: what's driving it? What hypothesis have been rejected or upheld?