Comparing apples and oranges: Weights of factors with different numbers of levels

Vsevolod Kapatsinski, Irina Shport & Susan Guion-Anderson

vkapatsi@uoregon.edu, ishport@uoregon.edu, guion@uoregon.edu

University of Oregon

The problem and the solution

- Two possible cues to a phonological contrast
  - More generally, two independent variables with a binary dependent variable (multiple logistic regression)

- We want to know the weights of cues relative to each other
  - E.g., Escudero et al. (2009), Morrison (2005a): spectrum vs. duration as cues to vowel contrast in Dutch
  - Shport (2011): F0 peak location vs. F0 fall as cues to pitch accent in Japanese
  - More generally, relative importance of predictors
    - E.g., Kapatsinski (2009): conjoined constituent type vs. register as predictors of conjunction choice in Russian

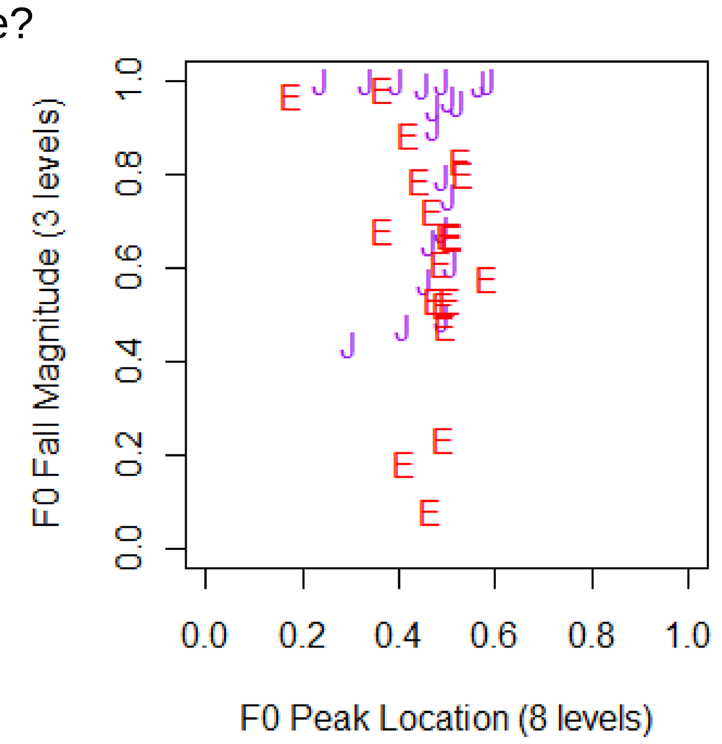
- The cues have different numbers of levels
  - Predictors have different numbers of levels

- What we show:
  - State-of-the-art methods (varbrule range, coefficient comparisons) are biased if the predictors are nominal and have different numbers of levels (cf. also Morrison 2005)
  - We can estimate *how* biased they are
  - And therefore how much of a difference in cue/factor weight is attributable to differences in numbers of levels
  - We can therefore estimate whether the observed difference could plausibly result from differences in numbers of levels (using Monte Carlo resampling techniques)

- Caution: Use only if there is no significant interaction between the cues (otherwise, the cues are not the real cues)

Example: Shport (2011)

- 1 = 100% syllable 1 prominence responses
- 0.5 = No reliance on the cue
- 0 = 100% syllable 2 prominence responses
- Which cue do the listeners rely on more? F0 peak location OR F0 fall magnitude?



General algorithm (with R code excerpts)

```
numSpeakers = Number of Speakers, same as actual study
nReps = Number of Replications #some large number
nVal1 = Number of values of first factor
nVal2 = Number of values of second factor
nPerCell = Number of stimuli per subject with the same values on factors 1 and 2

for (i in 1:nReps) #for each replication
{
    #the first predictor
   fakeIV1=as.factor(rep(0:(nVal-1),each=nStim/nVal1))

   #the second predictor
   fakeIV2=as.factor(rep(0:(nVal-1),each=nStim/nVal2)[sample(nStim,nStim)])

   for (j in 1:numSpeakers) #for each speaker in the simulated sample
   {
      #Get a sample of nStim random 1's and 0's: a simulated subject
      #Since zeros and ones are chosen at random, this is a subject
      #who is drawn from a population where the effects of both predictors
      #on the dependent variable are equal to zero
      p_yes_response=rnorm(n=1,mean=0.5,sd=0.1)
      fakeSample=rbinom(nVal1*nVal2, nPerCell, p_yes_response)

      #Run a logistic regression within the subject
      fake.glm<-glm(cbind(fakeSample, 10-fakeSample) ~ fakeIV1 + fakeIV2,
            family=binomial(link="logit"))

      #Extract factor weights from the results of the regression
      #Varbrul factor weight = exp(B)/(1+exp(B))  (Morrison 2005b)
      #The factor weights are handy because they scale between 0 and 1
      #As a result we get how important the predictors are for this individual subject
      #On average, all should be around .5 because the predictors have no effect
      #in the population from which subjects are drawn.
      factor.wts = exp(fake.glm$coefficients)/(1+exp(fake.glm$coefficients))

      #For each factor group, we get the difference between highest factor weight
      #and highest factor weight (range)
      #This is the standard ay of evaluating predictor importance in quantitative
      #sociolinguistics (e.g., Tagliamonte 2006)
      SubjCoeffs[j,1]<- abs(factor.wts[2]-.5)*2
      SubjCoeffs[j,2]<- max(factor.wts[3:11])-min(factor.wts[3:11])

   }

   #We compare the within-subject importance measures for predictor 1
   #to within-subject importance measures for predictor 2 using a t-test
   #Could use any other test.
   #We then extract the distribution of resulting t values.
   #If there is a bias in favor of a factor with more levels, the t's should tend to be <0
   tVec[i]<-t.test(SubjCoeffs[,1]-SubjCoeffs[,2])$statistic
}
#We can now see if the t value we obtained in actual study is within the range of the
simulated t's. If it is not, we can say that the factors differ in importance at the 1/nReps
level.
```

For full code, see pages.uoregon.edu/vkapatsi/CueWeightScript.R

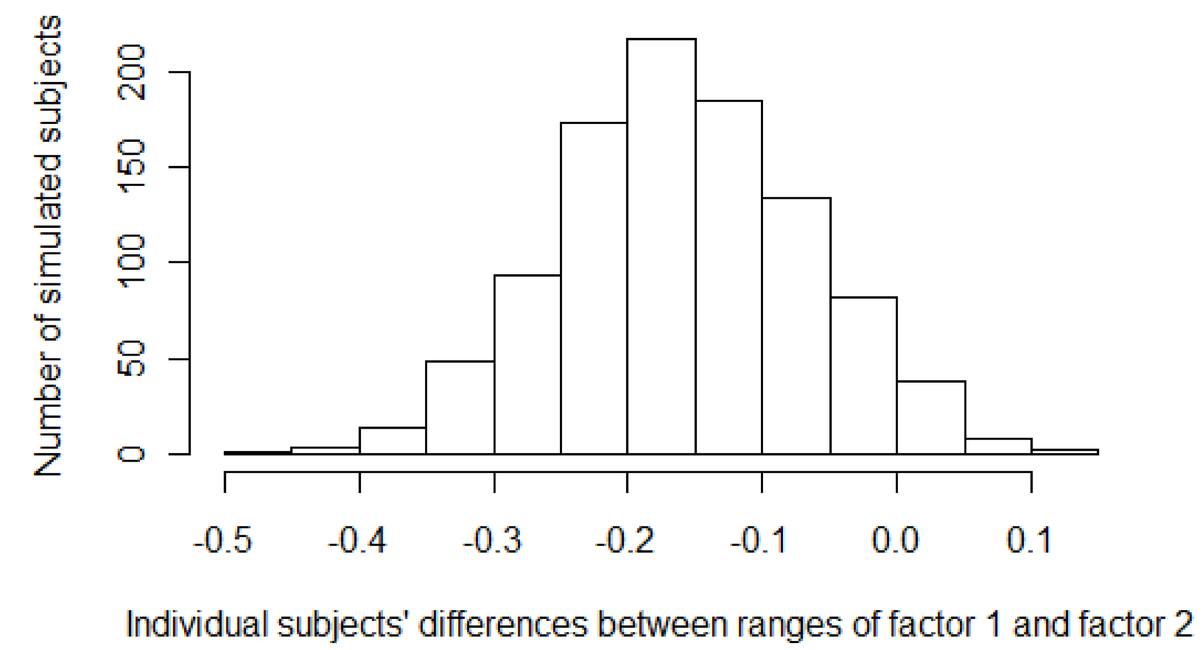
Two nominal predictors

Individual subjects:

- 1000 simulated subjects; nVal1=3, nVal2=8
- Result: Predictor with more levels tends to have a larger coefficient

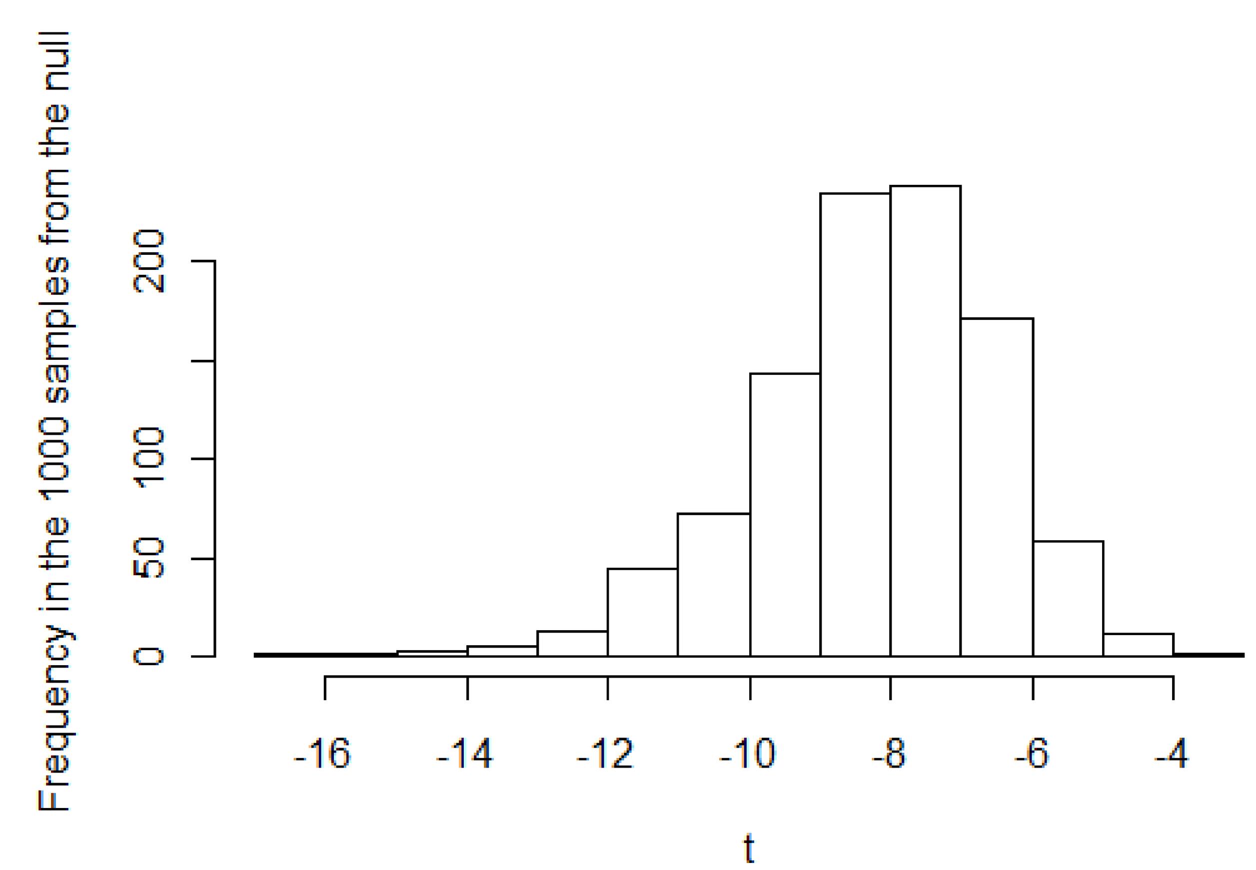
We know how much larger:

- Range of factor group with 3 levels must be greater than range of the factor group with 8 levels to conclude that the factor group with 3 levels is more important.
- Range of factor group with 8 levels must exceed the range of the factor group with 3 levels by 0.35 to conclude that the factor group with 8 levels is more important.
- Note: bias increases if subjects favor one value of the DV over another; The graph below is for subjects who on average do not have a bias in favor of one value of the DV.



Groups of subjects

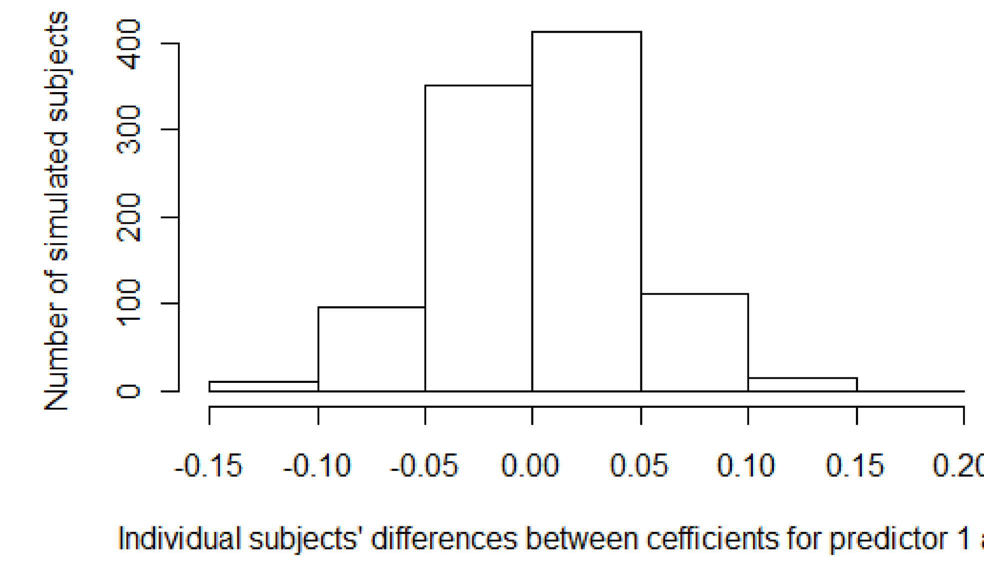
- nSubj=20, 1000 simulated groups
- Result: Predictor with more levels tends to have a larger coefficient
- We know how much larger:
  - Observed t's above -5 are evidence that the factor group with 3 levels is weighted more heavily than the group with 8 levels.
  - Observed t's below -12 are evidence that the factor group with 8 levels is weighted more heavily than the group with 3 levels.



Two numerical / ordinal predictors

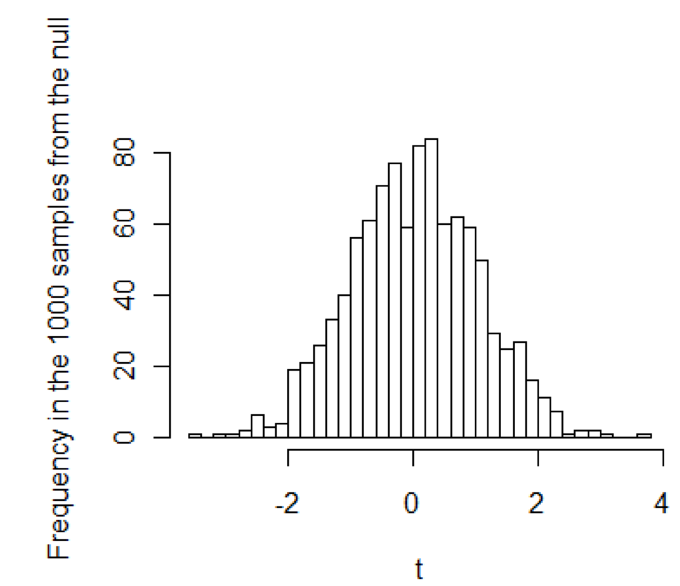
Individual subjects:

- 1000 simulated subjects, nVal1=3 (0,1, 2), nVal2=8 (0, 1, 2, 3, 4, 5, 6, 7)
- Result: Very little bias



Two nominal predictors with same number of levels but where there are equal numbers of stimuli having each value on predictor 1 but not on predictor 2:

- Predictor 1: 10 stimuli with each value
- Predictor 2: 5 stimuli with one value. 15 stimuli with other value
- Result: No observable bias



Conclusion

- Numerical predictors with different numbers of levels, scales → not to worry, R will take care of it.
- Categorical/nominal predictors → range biased in favor of predictor with more levels
- We can estimate how biased it is for a particular dataset (need to know whether subjects are biased in favor of a response)
- Therefore, we can test whether the observed difference in range between predictors could be due to differences in numbers of levels.
- **So, we can still compare them! Yay!**

References

- Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish learners. *Journal of Phonetics*, *37*, 452-465.
- Kapatsinski, V. 2009. Adversative conjunction choice in Russian: Semantic and syntactic influences on lexical selection. *Language Variation and Change*, 21, 157-73.
- Morrison, G. S. 2005a. An appropriate metric for cue weighting in L2 speech perception: Response to Escudero and Boersma (2004). *Studies in Second Language Acquisition*, *27*, 597-606.
- Morrison, G. S. 2005b. Dat is what the PM said: A quantitative analysis of Prime Minister Chrétien's pronunciation of English voiced dental fricatives. *Cahiers Linguistiques d'Ottawa*, 33, 1–21.
- Shport, I. A. 2011. *Cross-linguistic perception and learning of Japanese lexical prosody by English listeners.* Unpublished doctoral dissertation. University of Oregon, Eugene.
- Tagliamonte, S. 2006. *Analyzing sociolinguistic variation*. CUP.