Using ANOVA on percentages?

I have a table with four groups (4 BMI groups) as the independent variable (factor). I have a dependent variable that is "percent mother smoking in pregnancy".

Is it permissible to use ANOVA for this or do I have to use chi-square or some other test?

anova

edited May 27 '11 at 2:16
Jeromy Anglim
26.3k 11 87 189

asked May 27 '11 at 0:39
drew
36 1 1 2

4 Answers

There is a difference between having a binary variable as your dependent variable and having a proportion as your dependent variable.

- **Binary dependent variable**:
  - This sounds like what you have. (i.e., each mother either smoked or she did not smoke)
  - In this case I would not use ANOVA. Logistic regression with some form of coding (perhaps dummy coding) for the categorical predictor variable is the obvious choice if you are conceptualising the binary variable as the dependent variable (otherwise you could do chi-square).
- **Proportion as dependent variable**:
  - This does not sound like what you have. (i.e., you don't have data on the proportion of total waking time that a mother was smoking during pregnancy in a sample of smoking pregnant women).
  - In this case, ANOVA and standard linear model approaches in general may or may not be reasonable for your purposes. See @Ben Bolker's answer for a discussion of the issues.

answered May 27 '11 at 2:23
Jeromy Anglim
26.3k 11 87 189

For a binary dependent variable, in the case that I only have summary data for the binary proportions (ie # in the A, B, and C groups, and the # of successes in the A, B, and C group), and not the actual raw data, how can we go about using logistic regression? I am only familiar with using it with the raw data. – Bryan Jan 9 '15 at 5:39

It depends on how close the responses within different groups are to 0 or 100%. If there are a lot of extreme values (i.e. many values piled up on 0 or 100%) this will be difficult. (If you don't know the "denominators", i.e. the numbers of subjects from which the percentages are calculated, then you can't use contingency table approaches anyway.) If the values within groups are more reasonable, then you can transform the response variable (e.g. classical arcsine-square-root or perhaps logit transform). There are a variety of graphical (preferred) and null-hypothesis testing (less preferred) approaches for deciding whether your transformed data meet the assumptions of ANOVA adequately (homogeneity of variance and normality, the former more important than the latter). Graphical tests: boxplots (homogeneity of variance) and Q-Q plots (normality) [the latter should be done within groups, or on residuals]. Null-hypothesis tests: e.g. Bartlett or Fligner test (homogeneity of variance), Shapiro-Wilk, Jarque-Bera, etc.

answered May 27 '11 at 1:05
Ben Bolker
9,469 23 41

You need to have the raw data, so that the response variable is 0/1 (not smoke, smoke). Then you can use binary logistic regression. It is not correct to group BMI into intervals. The cutpoints are not correct, probably don't exist, and you are not officially testing whether BMI is associated with smoking. You are currently testing whether BMI with much of its information discarded is associated with smoking. You'll find that especially the outer BMI intervals are quite heterogeneous.

answered May 29 '11 at 13:18

Frank Harrell
34.4k 1 56 133

2 @Frank - why is it "not correct" to group BMI? this seems perfectly reasonable, so long as the results are appropriately interpreted. You could well be testing, for example, whether being "underweight" "healthy weight" "overweight" and "obese" are associated with smoking, where these terms are defined by the ranges of BMI. I see no "wrong" here. – probabilityislogic May 29 '11 at 13:42

I believe that the OP is working with a common instructional data set and may not have the raw BMI. While it's generally not ideal to discretize continuous regressors it's not "incorrect". It can even be helpful to resort to this when we suspect the measurements are noisy and there's no other recourse. Indeed, the real hypothesis we'd want to test is whether obesity is related to smoking; BMI is just one way to measure obesity (and has its problems from what I understand). – JMS May 29 '11 at 15:22

2 Even when measurements are noisy, analyzing variables as continuous is superior. Categorizing BMI creates more problems than different choices of analysis can fix. In fact the estimates upon categorization no longer have a scientific interpretation. A scientific quantity is one having meaning outside the current experiment. You'll find that group estimates (e.g., log odds that Y=1 for high vs low intervals of X) are functions of the entire set of observed BMIs. For example, if you were to add more extremely high or extremely low BMIs to the sample, the "effects" would get stronger. – Frank Harrell May 29 '11 at 19:20

For those who have installed R and RStudio, an interactive demonstration may be found at biostat.mc.vanderbilt.edu/BioMod - see the green NEW marking. You have to load the script into RStudio and also install the Hmisc package. – Frank Harrell May 29 '11 at 22:45

"Even when measurements are noisy, analyzing variables as continuous is superior" This is just incorrect (the generality of it, that is - usually it's true). Imagine you have a continuous covariate where the error in its measurement increases with its magnitude, for example. Of course the best thing to do is model the error, or get better measurements, etc. But to say that it's incorrect is simply too strong a statement to make. – JMS May 30 '11 at 15:47

If you choose to do an ordinary ANOVA on proportional data, it is crucial to verify the assumption of homogeneous error variances. If (as is common with percentage data), the error variances are not constant, a more realistic alternative is to try beta regression, which can account for this heteroscedasticity in the model. Here is a paper discussing various alternative ways of dealing with a response variable that is a percentage or proportion: http://www.ime.usp.br/~sferrari/beta.pdf

If you use R, the package **betareg** may be useful.

answered Nov 8 '12 at 22:33

Will Townes
173 6