Yes, No and Maybe: Analyzing Categorical Data

Friday, September 9th, 2016

AnnMaria De Mars

Categorical data analysis: An overview of statistical techniques

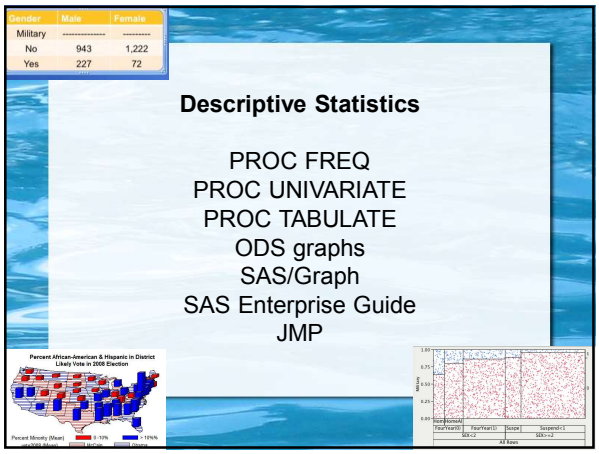
AnnMaria De Mars
The Julia Group
7 Generation Games

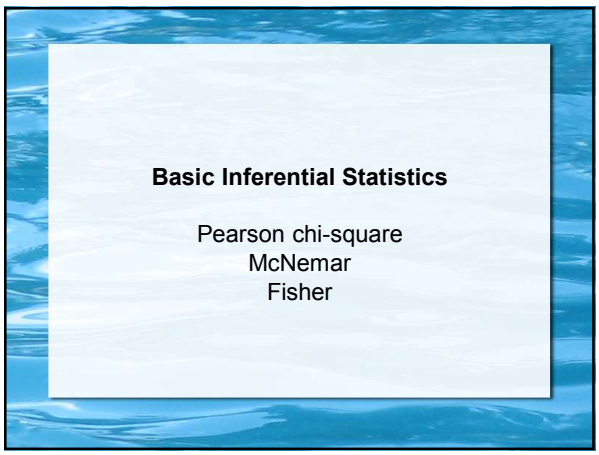
Anyone who thinks he knows all of SAS is clinically insane

Okay, Hemingway didn't really say that, but he should have

Three uses for descriptive statistics

- Describe a sample
- Check data quality
- Answer descriptive questions





Answers to deep questions

- What does a McNemar test test?
- Why would a Pearson chi-square and a McNemar test give different answers?

Pearson Chi-Square

- Tests for a relationship between two categorical variables, e.g. whether having participated in a program is related to having a correct answer on a test.
- Assumes randomly sampled data
- Assumes independent observations

Good for chi-square

Correct cause ---------- Group	YES	NO
Interactive	91	9
Handouts	55	45

Why is the previous example good?

- It includes two independent groups
- There are adequate numbers per cell

Bad for chi-square

Correct death ---------- Pre-Post	YES	NO
PRE	15	85
POST	91	9

Enter the McNemar

- This is a test of correlated proportions
- It is commonly used to test, for example, if the proportion showing mastery at time 1 = the proportion showing mastery at time 2

Bad for Pearson chi-square

Correct cause ---------- Group	YES	NO
Interactive	12	3
Handouts	8	4

Fisher's exact test

- Is used when the assumption of large sample sizes cannot be met
- There is no advantage to using it if you do have large sample sizes

A lot more …

- Cochran-Mantel- Haenszel test for repeated tests of independence
  - Do athletes in physical therapy report improvement in mobility more than those who do not receive PT and does this vary depending on if it is preseason or during the season ?

Other simple statistics

- Binomial tests
- Confidence intervals
- Odds ratios

Because, obviously, not everyone has the same tastes

What about logistic regression?

- Logistic is similar to linear regression in that a dependent variable is predicted from a combination of independent variables
- The dependent is the LOG of the ODDS ratio of being in one group versus another

Example: Death certificates

- The death certificate is an important medical document.
- Resident physician accuracy in completing death certificates is poor.
- Participants were in an interactive workshop or provided printed handouts.
- Pre-existing knowledge was measured

Example

- Dependent: Cause of death medical student is correct or incorrect
- Independent: Group
- Independent: Awareness of guidelines for death certificate completion

Surveylogistic

- Interpreted the same as the logistic output but allows inclusion of survey features such as strata and cluster

Other PROCs

- CATMOD
- CORRESP
- PRINQUAL

Hybrids

- T-test
- ANOVA
- NPAR1WAY
- FACTOR
- REG

It's all about questions

- Are your data any good?
- What is the distribution of X ?
- What is the distribution of X given Y?
- Is there a significant relationship between X and Y?
- Given X, what are the odds of Y?
- How well, and with what variables, can we predict which category of X a person falls into?
- Is this set of variables significantly better for predicting X than that other set of variables lying over there?

Our secret plan

- Bivariate descriptives
- Contingency, chi-square, probability
- Other descriptives
- Other simple statistics
- Logistic regression

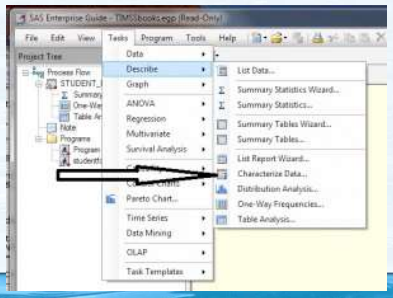
ANALYSIS OF CATEGORICAL DATA
Descriptive Statistics – SAS Enterprise Guide

ANNMARIA DE MARS, PH.D.
The Julia Group
7 Generation Games

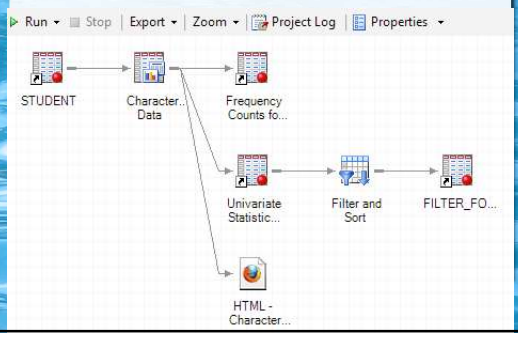
Three uses for descriptive statistics

- Describe a sample
- Check data quality
- Answer descriptive questions

Step 1



QUICK SAS ENTERPRISE WAY

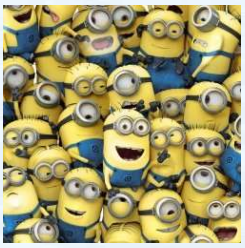


DATA QUALITY

It's a concern with categorical data, too

Why I don't have minions

The need to understand your own data



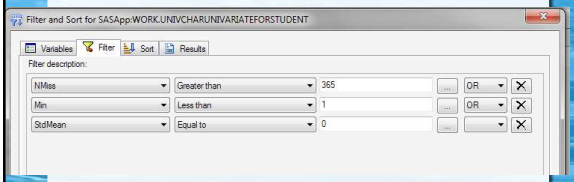
PROC UNIVARIATE

For categorical data?

That's strange ….

(not if you have lots of variables & the categories are coded numerically)

Step 2



Option 2: Write your own macro

http://www.thejuliagroup.com/blog/?p=1357

http://www.thejuliagroup.com/blog/?p=1364

```
LIBNAME MACLIB "C:\Users\MyDir\Documents\My SAS Files" ;
    %macro dataqual(dsn,idvar,startvar,endvar,obsnum) / store ;
    Title "Duplicate ID Numbers" ;
    Proc freq data =  lib.&dsn noprint ;
    tables &idvar / out = &dsn._freq (where = ( count > 1 )) ;
    format &idvar  ;
    proc print data = &dsn._freq (obs = 10 ) ;
    run ;
    proc summary data = lib.&dsn mean min n std ;
    output out = &dsn._stats ;
    var &startvar — &endvar ;

proc transpose data = &dsn._stats out = &dsn._stats_trans ;
    id _STAT_ ;
    data &dsn._chk ;
    set &dsn._stats_trans ;
    pctmiss = 1 – (n/&obsnum) ;
    if min < 0 then neg_min = 1 ;
    else neg_min = 0 ;
    if std = 0 then constant = 1 ;
    else constant = 0 ;
    if (pctmiss > .05 or neg_min = 1 or constant = 1) then output ;
    Title "Deviant variables to check " ;
    proc print data = &dsn._chk ;
    run;
```

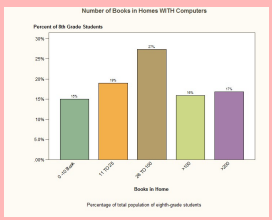
Enterprise Guide, Probability, Distributions , Contingency Tables & Chi-square

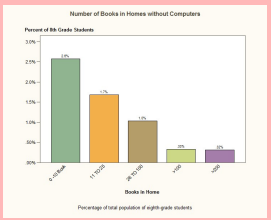
What did you expect?

Don't forget graphics!

These are easy to create with SAS Enterprise Guide and easy for a non-technical audience to interpret

Homes without computers have fewer books





Graphs with SAS Enterprise Guide

Definition

A listing of all the values the random variable can assume with their corresponding probabilities make a *probability distribution*

Table of momed3 by daded3

momed3	daded3			
Frequency Percent Row Pct Col Pct	0-11	12	Coll	Total
0-11	471 10.02 61.65 60.15	210 4.47 27.49 10.99	83 1.77 10.86 4.13	764 16.25
12	234 4.98 13.39 29.89	1145 24.35 65.50 59.95	369 7.85 21.11 18.37	1748 37.18
Coll	78 1.66 3.56 9.96	555 11.80 25.34 29.06	1557 33.11 71.10 77.50	2190 46.58
Total	783 16.65	1910 40.62	2009 42.73	4702 100.00

Two-way contingency table

Parts of a table

		FATHERS	EDUCATION	
MOTHERS EDUCATION	< HS	HS GRAD	COLLEGE	TOTAL
<HS	471	210	83	764
HS GRAD	234	1145	369	1748
College	78	555	1557	2190
TOTAL	783	1910	2009	4702

Marginal distributions are row or column totals divided by the grand total

Marginal Distributions

	FATHERS EDUCATION			
MOTHERS EDUCATION	< HS	HS GRAD	COLLEGE	TOTAL
<HS				**16.3%**
HS GRAD				**37.2%**
College				**46.6%**
TOTAL	**16.7%**	**40.6%**	**42.7%**	

CONDITIONAL DISTRIBUTION

Is the distribution of one variable on the condition of another variable

For example,
the distribution of mother's education for a given level of father's education

CONDITIONAL Distributions

	FATHERS EDUCATION		
MOTHERS EDUCATION	< HS	HS GRAD	COLLEGE
<HS	61.7	27.5	10.9
HS GRAD	13.4	65.5	21.1
College	3.6	25.3	71.1
TOTAL	**16.7%**	**40.6%**	**42.7%**

In words

The previous table shows that the marginal distribution of father's education is 17% less than high school, 41% high school graduates and 43% college graduates

Given the CONDITION that the mother had less than a high school education, the conditional distribution is 62% less than high school,28% high school grads and 11% college graduates

Just for closure ..

e	o	(e-o) ^2	((e-o)^2)/e
157	72	7225	46.01910828
142	227	7225	50.88028169
1028	943	7225	7.028210117
1137	1222	7225	6.354441513
			110.2820416

More on chi-square

http://www.thejuliagroup.com/blog/?p=661

Chi-square from SAS

```
PROC FREQ DATA = dsname ;
   TABLES var1 * var2 / chisq cellchi2 ;
```

SAS ENTERPRISE GUIDE

- Go to the TASKS menu
- Select DESCRIBE
- Select TABLE ANALYSIS
- Drag the variables you want on to row and column
- Under CELLS click the buttons next to EXPECTED and CELL FREQUENCIES

Cell chi-square (don't do anything stupid)

You have a significant chi-square value

One group is substantially larger than the other, e.g. 91% of students said "Yes"

When you look at the cell chi-square values you can see that most of the chi-square value comes from the smaller group.

Use Public ->	YES	NO	TOTAL
HOME			
YES	3,931 4,027 2.26 54.57	482 387 23.58 6.69	4,413 61.26
NO	2,642 2,547 3.58 36.67	149 244 37.28 3.07	2,791 38.74

Frequency
Expected Frequency
Cell chi-square
Percent

Use Public ->	YES	NO
HOME		
YES	3,931 4,027 2.26	482 387 23.58
NO	2,642 2,547 3.58	149 244 37.28 3.07

Total Chi-square = 66.7
Of that 60.7 – 90% - comes from two cells
Does that matter?

No. You are not a genius

Hypothetical example where cell chi-square is useful

Vote	Brown	Whitman
Hispanic	3,300	700
White	4200	4200
African-American	1000	1200
Asian-American	1200	800

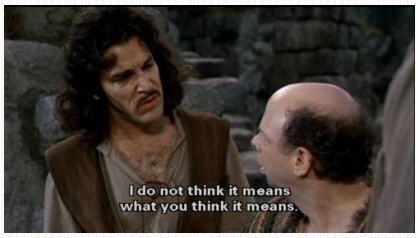
Categorical data analysis: For when your data DO fit in little boxes

AnnMaria De Mars, Ph.D.
The Julia Group
Santa Monica, CA

Our secret plan

- ~~Descriptives~~
- Chi-square
- Secrets of PROC FREQ
- Logistic regression

You keep saying that word



WE ALL KNEW FREQ DID THIS

PROC FREQ DATA = dsname ;
TABLES varname1 * varname2 / chisq ;

YOU GET

- Chi-square value (several)
- Phi coefficient
- Fisher Exact test (where applicable)

Table of momeduc by failgrade

momeduc	failgrade		
Frequency Percent Row Pct Col Pct	0	1	Total
0-11	714 12.91 73.31 15.44	260 4.70 26.69 28.63	974 17.61
12	1357 24.53 82.79 29.35	282 5.10 17.21 31.06	1639 29.63
13-15	356 6.44 82.60 7.70	75 1.36 17.40 8.26	431 7.79
16	1436 25.96 88.64 31.06	184 3.33 11.36 20.26	1620 29.28
17+	761 13.76 87.67 16.46	107 1.93 12.33 11.78	868 15.69
Total	4624 83.59	908 16.41	5532 100.00

Frequency Missing = 1845

Mothers Education & Failing a Grade

Statistic	DF	Value	Prob
Chi-Square	4	116.8321	<.0001
Likelihood Ratio Chi-Square	4	111.0668	<.0001
Mantel-Haenszel Chi-Square	1	91.9875	<.0001
Phi Coefficient		0.1453	
Contingency Coefficient		0.1438	
Cramer's V		0.1453	

Fisher's exact test

- Is used when the assumption of large sample sizes cannot be met
- There is no advantage to using it if you do have large sample sizes

Test for bias in sample

Frequency Percent Row Pct Col Pct	Table of respondent by gender		
	gender		
respondent	Female	Male	Total
0	1369 36.93 43.92 83.32	1748 47.15 56.08 84.69	3117 84.08
1	274 7.39 46.44 16.68	316 8.52 53.56 15.31	590 15.92
Total	1643 44.32	2064 55.68	3707 100.00
Frequency Missing = 3			

Fisher – magically happens

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1369
Left-sided Pr <= F	0.1390
Right-sided Pr >= F	0.8800
Table Probability (P)	0.0190
Two-sided Pr <= P	0.2590

A BUNCH OF THINGS YOU MAY NOT KNOW PROC FREQ DOES

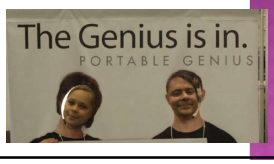
Other simple statistics

- Binomial tests
- Confidence intervals
- McNemar
- Odds ratios
- Cochran-Mantel-Haenszel test

Because, obviously, not everyone has the same tastes

WHAT ABOUT THIS ?

- PROC FREQ DATA = dsname ;
  TABLES varname /
  BINOMIAL (EXACT P = .333)
  ALPHA = .05 ;



WHAT'S IT DO

- The binomial (equiv p = .333) will produce a test that the population proportion is .333 for the first category. That is "No" for death. A Z-value will be produced and probabilities for one-tail and two-tailed tests.
- The exact keyword will produce confidence intervals and, since I have specified alpha = .05, these will be the 95% confidence intervals.

NOT NEW

death? y/n				
DTHFLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	5489	37.44	5489	37.44
Yes	9170	62.56	14659	100.00

HMMM…. THIS IS INTERESTING

Binomial Proportion for DTHFLAG = No	
Proportion	0.3744
ASE	0.0040
95% Lower Conf Limit	0.3666
95% Upper Conf Limit	0.3823
Exact Conf Limits	
95% Lower Conf Limit	0.3666
95% Upper Conf Limit	0.3823

NULL REJECTED !

Test of H0: Proportion = 0.333	
ASE under H0	0.0039
Z	10.6475
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

ODDS RATIOS !

```
PROC FREQ DATA = in.da4219p2 ;
  TABLES sex * dthflag / CHISQ CMH ;
```

SAME

Table of SEX by DTHFLAG			
SEX(SUM*sex)	DTHFLAG(death? y/n)		
Frequency Percent Row Pct Col Pct	No	Yes	Total
Female	3075 20.98 44.48 56.02	3838 26.18 55.52 41.85	6913 47.16
Male	2414 16.47 31.16 43.98	5332 36.37 68.84 58.15	7746 52.84
Total	5489 37.44	9170 62.56	14659 100.00
Frequency Missing = 71			

SAME

Statistic	DF	Value	Prob
Chi-Square	1	276.5634	<.0001
Likelihood Ratio Chi-Square	1	276.9268	<.0001
Continuity Adj. Chi-Square	1	275.9952	<.0001
Mantel-Haenszel Chi-Square	1	276.5445	<.0001
Phi Coefficient		0.1374	
Contingency Coefficient		0.1361	
Cramer's V		0.1374	

SAME

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3075
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	2.338E-62
Table Probability (P)	1.803E-62
Two-sided Pr <= P	4.180E-62

DIFFERENT

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	276.5445	<.0001
2	Row Mean Scores Differ	1	276.5445	<.0001
3	General Association	1	276.5445	<.0001

ODDS RATIO

Estimates of the Common Relative Risk (Row1/Row2)				
Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	1.7697	1.6541	1.8933
(Odds Ratio)	Logit	1.7697	1.6541	1.8933
Cohort	Mantel-Haenszel	1.4273	1.3682	1.4890
(Col1 Risk)	Logit	1.4273	1.3682	1.4890
Cohort	Mantel-Haenszel	0.8065	0.7859	0.8277
(Col2 Risk)	Logit	0.8065	0.7859	0.8277

Some More Coding

```
PROC FREQ DATA = dsname ;
      TABLES  varname1 * varname2 / AGREE ;
```

FOR CORRELATED DATA

Correlated Data

Table of prefail by postfail			
prefail	postfail		
Frequency Percent Row Pct Col Pct	0	1	Total
0	125 73.10 96.15 83.33	5 2.92 3.85 23.81	130 76.02
1	25 14.62 60.98 16.67	16 9.36 39.02 76.19	41 23.98
Total	150 87.72	21 12.28	171 100.00

McNemar's Test

McNemar's Test	
Statistic (S)	13.3333
DF	1
Pr > S	0.0003

Cohen's Kappa

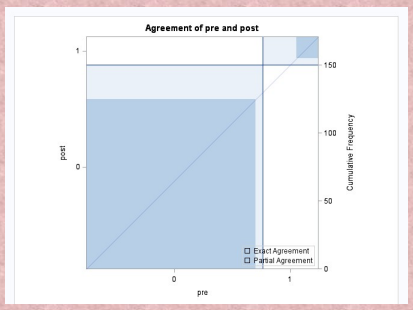
Simple Kappa Coefficient	
Kappa	0.4223
ASE	0.0837
95% Lower Conf Limit	0.2583
95% Upper Conf Limit	0.5863

$$= \frac{\text{Probability observed} - \text{Probability expected}}{1 - \text{Probability expected}}$$

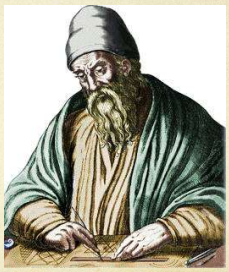
1.0 = perfect agreement

Negative Kappa is not an error, it means the two agree less than chance

Agreement plot



Logistic regression & Euclid



A very brief refresher

- Those of you who are statisticians, feel free to nap for two minutes

Assumptions of linear regression

linearity of the relationship between dependent and independent variables

independence of the errors (no serial correlation)

homoscedasticity (constant variance) of the errors across predictions (or versus any independent variable)

normality of the error distribution.

Residuals Bug Me

To a statistician, all of the variance in the world is divided into two groups, variance you can explain and variance you can't, called error variance.

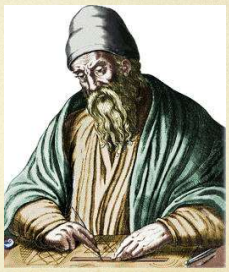
Residuals are the error in your prediction.

Residual error

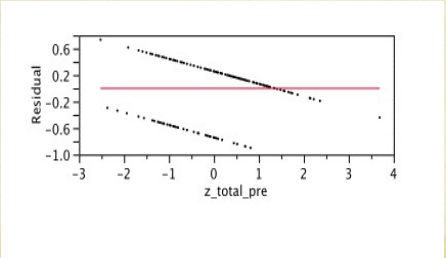
If your actual score on say, depression, is 25 points above average and, based on stressful events in your life I predict it to be 20 points above average, then the residual (error) is 5.

Euclid says …

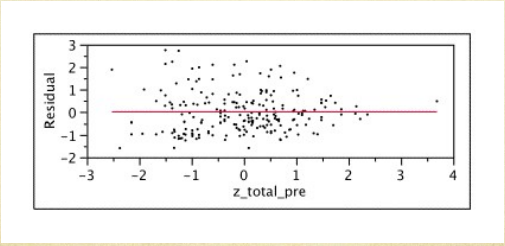
Let's look at those residuals when we do linear regression with a categorical and a continuous variable



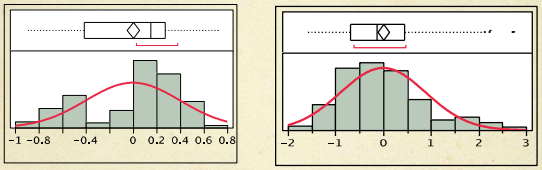
Residuals: Pass/ Fail



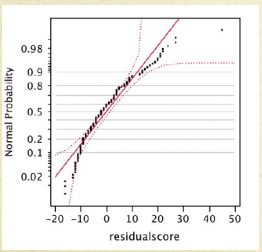
Residuals: Final Score

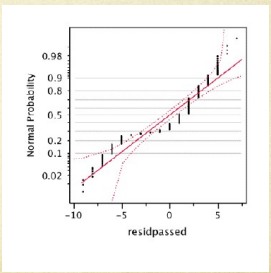


Which looks more normal?

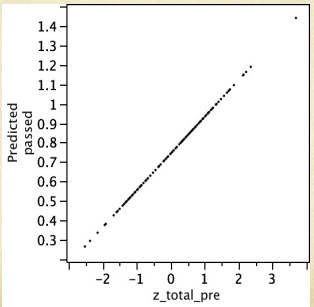


Which is a straight line?





Impossible events: Prediction of pass/fail



It's not always like this. Sometimes it's worse.

Notice that NO ONE was predicted to have failed the course.

Several people had predicted scores over 1.

Sometimes you get negative predictions, too

Logarithms, probability & odds ratios

In five minutes or less

Points justifying the use of logistic regression

Really, if you look at the relationship of a dichotomous dependent variable and a continuous predictor, often the best-fitting line isn't a straight line at all. It's a curve.

You could try predicting the probability of an event...

... say, passing a course. That would be better than nothing, but the problem with that is probability goes from 0 to 1, again, restricting your range.

Maybe use the odds ratio ?

which is the ratio of the odds of an event happening versus not happening given one condition compared to the odds given another condition. However, that only goes from 0 to infinity.

When to use logistic regression: Basic example #1

Your dependent variable (Y) :
There are two probabilities, married or not. We are modeling the probability that an individual is married, yes or no.

Your independent variable (X):
Degree in computer science field =1, degree in French literature = 0

Step #1

A. Find the PROBABILITY of the value of Y being a certain value divided by ONE MINUS THE PROBABILITY, for when X = 1

$p / (1- p)$

Step #2

B. Find the PROBABILITY of the value of Y being a certain value divided by ONE MINUS THE PROBABILITY, for when X = 0

Step #3

- B. Divide A by B

  That is, take the odds of Y given X = 1

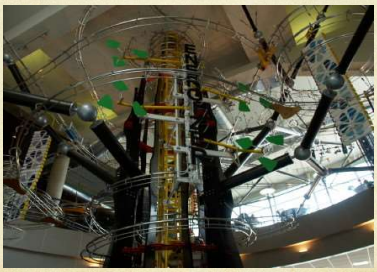
  And divide it by

  odds of Y given X = 2

Example!

- 100 people in computer science & 100 in French literature
- 90 computer scientists are married
  - Odds = 90/100 = 9
- 45 French literature majors are married
  - Odds = 45/55 = .818
- Divide 9 by .818 and you get your odds ratio of 11 because that is 9/.818

Just because that wasn't complicated enough …



Now that you understand what the odds ratio is …

The dependent variable in logistic regression is the LOG of the odds ratio (hence the name)

Which has the nice property of extending from negative infinity to positive infinity.

A table (try to contain your excitement)

	B	S.E.	Wald	Df	Sig	Exp(B)
CS	2.398	.389	37.949	1	.000	11.00
Constant	-.201	.201	.997	1	.318	.818

The natural logarithm (ln) of 11 is 2.398.
I don't think this is a coincidence

If the reference value for CS =1 , a positive coefficient means when cs =1, the outcome is more likely to occur

How much more likely? Look to your right

	B	S.E.	Wald	Df	Sig	Exp(B)
CS	2.398	.389	37.949	1	.000	11.00
Constant	-.201	.201	.997	1	.318	.818

The ODDS of getting married are 11 times GREATER
If you are a computer science major

Actual Syntax

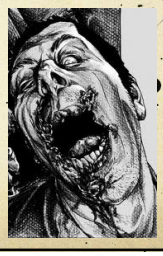
Thank God!

Picture of God not available

PROC LOGISTIC data = datasetname descending ;

By default the reference group is the first category.

What if data are scored
0 = not dead
1 = died



CLASS categorical variables ;

Any variables listed here will be treated as categorical variables, regardless of the format in which they are stored in SAS

MODEL dependent = independents ;

Dependent = Employed (0,1)

Independents

- County
- # Visits to program
- Gender
- Age

```
PROC LOGISTIC DATA = stats1
DESCENDING ;

CLASS gender county ;
 MODEL job =  gender county age visits ;
```

We will now enter real life

Table 1

Model Information	
Data Set	WORK.STATS1
Response Variable	job
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	135
Number of Observations Used	85

Response Profile		
Ordered Value	job	Total Frequency
1	1	28
2	0	57

Probability modeled is job=1.

Note: 50 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information						
Class	Value	Design Variables				
gender	Female	1				
	Male	-1				
county	.	1	0	0	0	0
	Anna	0	1	0	0	0
	Bob	0	0	1	0	0
	Clark	0	0	0	1	0
	Other	0	0	0	0	1
	Rufus	-1	-1	-1	-1	-1

This is bad

Model Convergence Status

Quasi-complete separation of data points detected.

Warning:

The maximum likelihood estimate may not exist.

Warning:

The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Complete separation

X	Group
0	0
1	0
2	0
3	0
4	1
5	1
6	1
7	1

If you don't go to church you will never die

Quasi-complete separation

Like complete separation BUT one or more points where the points have both values

1 1

2 1

3 1

4 1

4 0

5 0

6 0

there is not a unique maximum likelihood estimate

"For any dichotomous independent variable in a logistic regression, if there is a zero in the 2 x 2 table formed by that variable and the dependent variable, the ML estimate for the regression coefficient does not exist."

Depressing words from Paul Allison

Solution?

- Collect more data.
- Figure out why your data are missing and fix that.
- Delete the category that has the zero cell..
- Delete the variable that is causing the problem

Nothing was significant

& I was sad

Let's try something else!

Hey, there's still money in the budget!

Maybe it's the clients' fault

```
Proc logistic descending data = stats ;
    Class difficulty gender ;
    Model job = gender age difficulty
    ;
```

Oh, joy !

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

This sort of sucks

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	177.155	173.827
SC	180.038	185.358
-2 Log L	175.155	165.827

Yep. Sucks.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.7897	1.0768	0.5379	0.4633
gender	Female	1	0.1291	0.1886	0.4684	0.4937
Age		1	-0.0292	0.0210	1.9426	0.1634
difficulty	0	1	-0.3971	0.2279	3.0360	0.0814

Sucks. Totally.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gender Female vs Male	1.295	0.618	2.712
Age	0.971	0.932	1.012
difficulty 0 vs 1	0.452	0.185	1.104

Conclusion

Sometimes, even when you do the right statistical techniques the data don't predict well. My hypothesis would be that employment is determined by other variables, say having particular skills, like SAS programming.

Logistic regression is used when a few conditions are met:

1. There is a dependent variable.
2. There are two or more independent variables.
3. The dependent variable is binary, ordinal or categorical.

Medical applications

1. Symptoms are absent, mild or severe
2. Patient lives or dies
3. Cancer, in remission, no cancer history

Marketing Applications

1. Buys pickle / does not buy pickle
2. Which brand of pickle is purchased
3. Buys pickles never, monthly or daily

GLM and LOGISTIC are similar in syntax

```
PROC GLM DATA = dsname;
        CLASS class_variable ;
         model dependent = indep_var  class_variable ;

PROC LOGISTIC DATA = dsname;
   CLASS class_variable ;
    MODEL dependent = indep_var  class_variable ;
```

That was easy …

…. So, why aren't we done and going for coffee now?

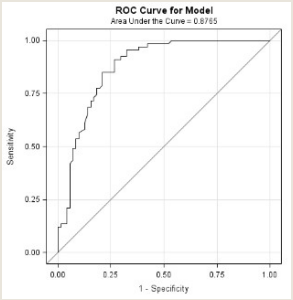
Why it's a little more complicated

1. The output from PROC LOGISTIC is quite different from PROC GLM
2. If you aren't familiar with PROC GLM, the similarities don't help you, now do they?

Important Logistic Output

- Model fit statistics
- Global Null Hypothesis tests
- Odds-ratios
- Parameter estimates

& a useful plot



A word from an unknown person on the Chronicle of Higher Ed Forum

Being able to find SPSS in the start menu does not qualify you to run a multinomial logistic regression

Take 2

Predicting passing grades

```
Proc Logistic data = nidrr  ;
        Class group ;
        Model passed = group education ;
```

Yay! Better than nothing!

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.6192	2	<.0001
Score	16.3023	2	0.0003
Wald	13.2622	2	0.0013

& we have a significant predictor

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Education	1	11.3050	0.0008
Group	1	2.5574	0.1098

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	3.5295	1.2682	7.7458	0.0054
Education		1	-0.3575	0.1063	11.3050	0.0008
Group	CONTROL	1	0.2972	0.1859	2.5574	0.1098

WHY is education negative?

Higher education, less failure

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Education	0.699	0.568	0.861
Group CONTROL vs EXP	1.812	0.874	3.755

Now it's later

Comparing model fit statistics

The Mathematical Way

Comparing models

Akaike Information Criterion

Used to compare models

The SMALLER the better when it comes to AIC.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	193.107	178.488
SC	196.131	187.560
-2 Log L	191.107	172.488

New variable improves model

Criterion	Intercept Only	Intercept and Covariates
AIC	193.107	178.488
SC	196.131	187.560
-2 Log L	191.107	172.488

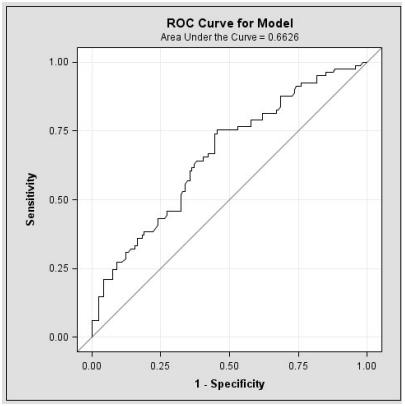
Criterion	Intercept Only	Intercept and Covariates
AIC	193.107	141.250
SC	196.131	153.346
-2 Log L	191.107	133.250

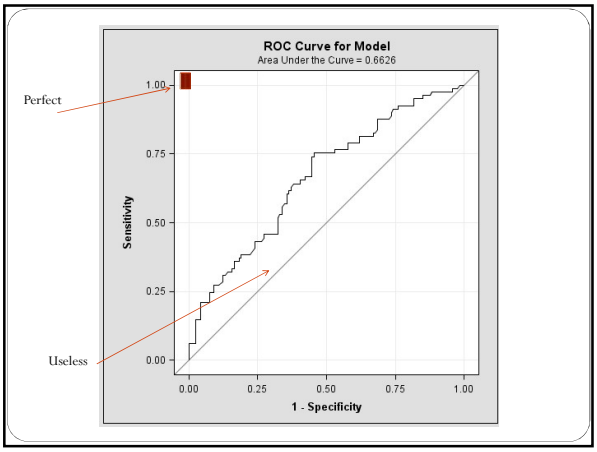
The Visual Way

Comparing models

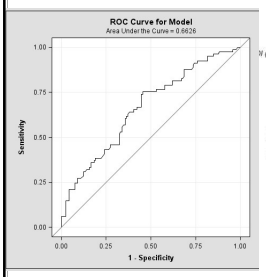
Reminder

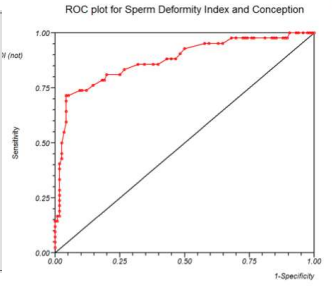
- **Sensitivity is the percent of true positives,** for example, the percentage of people you predicted would die who actually died.
- **Specificity is the percent of true negatives,** for example, the percentage of people you predicted would NOT die who survived.





WHICH MODEL IS BETTER ?





Special Topics: Data mining & stepwise logistic regression

With SAS

I'm unimpressed

Yeah, but can you do it again?

Data mining – sample & test

1. Select sample
2. Create estimates from sample
3. Apply to hold out group
4. Assess effectiveness

Create sample

```
proc surveyselect data = visual
  out = samp300 rep = 1
  method = SRS seed = 1958 sampsize = 315 ;
```

Create Test Dataset

```
proc sort data = samp300 ;
  by caseid ;
proc sort data = visual ;
  by caseid ;
 data testdata ;
  merge samp300 (in =a ) visual (in =b) ;
  if a and b then delete ;
```

Create Test Dataset

```
data testdata ;
  merge samp300 (in =a ) visual (in =b) ;
  if a and b then delete ;
```

*** Deletes record if it is in the sample ;

Create estimates

```
ods graphics on
proc logistic data = samp300 outmodel = test_estimates plots =
  all ;
  model vote = q6  totincome pctasian  / stb rsquare ;
  weight weight ;
```

Test estimates

```
proc logistic inmodel = test_estimates plots = all ;
  score data = testdata  ;
  weight weight ;
```

*** If no dataset is named, outputs to dataset named Data1, Data2 etc. ;

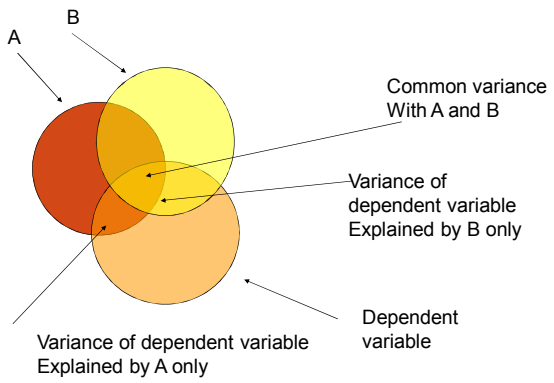
Validate estimates

```
proc freq data = data1;
  tables vote* i_vote  ;
proc sort data = data1 ;
  by i_vote ;
```

What is stepwise logistic regression ?

That's a good question. Usually, all the independent variables are entered in a model simultaneously.

In a stepwise model, the variable that has the largest zero-order correlation with the dependent is entered first. The variable that has the highest correlation with the remaining variance enters second.

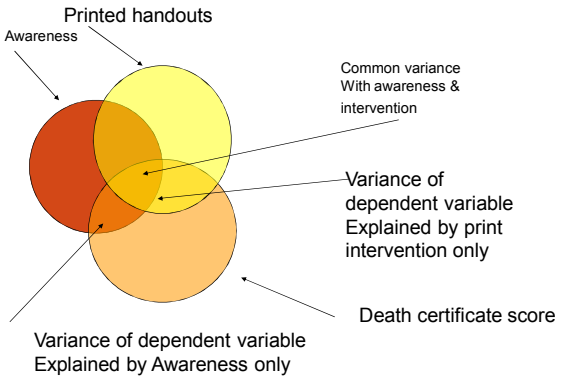


Example: Death certificates

- The death certificate is an important medical document.Resident physician accuracy in completing death certificates is poor. Participants were randomized into interactive workshop or provided with printed instruction material. A total of 200 residents completed the study, with 100 in each group.

Example

- Dependent: Cause of death medical student is correct or incorrect
- Independent: Group
- Independent: Awareness of guidelines for death certificate completion



Bottom line

Stepwise methods assign all of the shared variance to the first variable to enter the model

They take advantage of chance to maximize explained variance

Coefficients are not as stable as non-stepwise models

& this is all we'll have to say about stepwise today

Ordinal & Multinomial Logistic Regression

Featuring SAS

Default is ordinal

"When PROC LOGISTIC encounters a model with a dependent variable that has more than two categories, it automatically uses the cumulative logit to perform the analysis. Be careful: make sure that the dependent variable is ordinal and not nominal!

Ordinal logistic regression

Hosmer discusses various models.

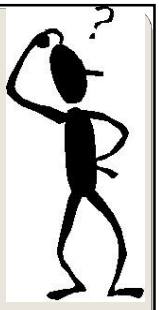
SAS default is the proportional odds model

What exactly is it?

- The probability of an equal or smaller response than j are compared to the probability of a larger response

$$c_k(\mathbf{x}) = \ln\left[\frac{\Pr(Y \le k|\mathbf{x})}{\Pr(Y > k|\mathbf{x})}\right]$$

Log of odds ratio



Probabilities modeled are cumulated over the lower ordered values

NOTE: Ordinal logistic regression in SAS

Odds ratios & parameter estimates

Logistic3.pdf

Probabilities modeled are cumulated over the lower ordered values

Because of this, the DESCENDING option has no effect. To get descending effect, you need to recode your dependent variable

Logistic5.pdf

Model as ordinal

```
proc logistic data = test descending ;
  class educ sedentary srsex income marit ;
  model health = educ sedentary srsex income srage_p mental marit
/ stb ;
```

Model as categorical

```
proc logistic data = test ;
  class educ sedentary srsex income marit ;
  model health = educ sedentary srsex income srage_p
mental marit / stb link=glogit ;
```

Ordinal, odds ratio

```
        proc logistic data = test ;
class educ srsex income marit ;
model health = educ  srsex income srage_p mental marit / stb ;
        oddsratio income ;
        oddsratio educ ;
```

Making data

```
data test ;
      set mydata.adult2009 ;
      if  srage_p > 39 ;
      if af24 = 1 and af83 = 1 then bloodst = 1;
            else bloodst = 0 ;
      if ab86 = 1 and ab85 = 1 then colonyr = 1 ;
            else colonyr = 0 ;
      ae_fruit = max(ae_fruit,21) ;
      if ac11 = 0 then soda = "0.None" ;
            else if ac11 < 5 then soda = "1.Few " ;
            else if ac11 < 31 then soda = "2.Much" ;
            else if ac11 >31 then soda = "3.Diabetes" ;
```