CYP-C Data Analysis Using SAS I

CYP-C Research Champion Webinar November 3, 2017 Jason D. Pole, PhD



Overview

- SAS overview revisited
- Data Analysis
 - Bi-Variate Tables and Stratification
 - Correlation
 - Chi-square Test
 - Odds Ratios / Relative Risk
 - Introduction to Logistic Regression

SAS Overview

- For our purposes only two major things you can do in SAS
 - DATA step Manipulate the data in some way
 - Reading in Data
 - Creating and Redefining Variables
 - Sub-Setting Data
 - Working with Dates
 - · Working with Formats
 - Procedure step
 - · Analyze the data
 - Produce frequency tables
 - Estimate a regression model

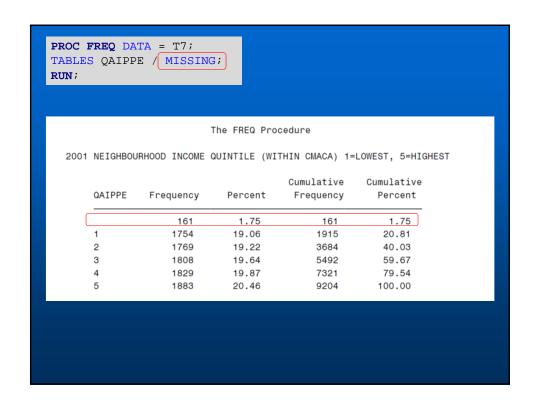
Bi-Variate Tables and Stratification

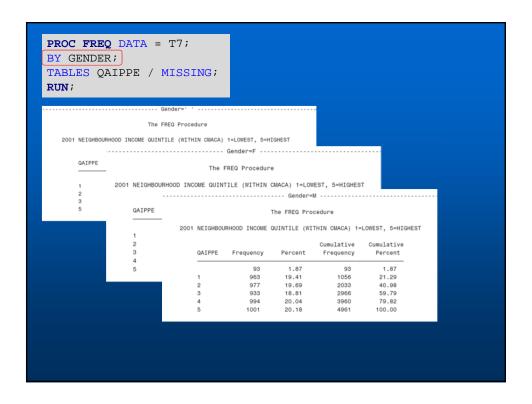
SAS PROC FREQ

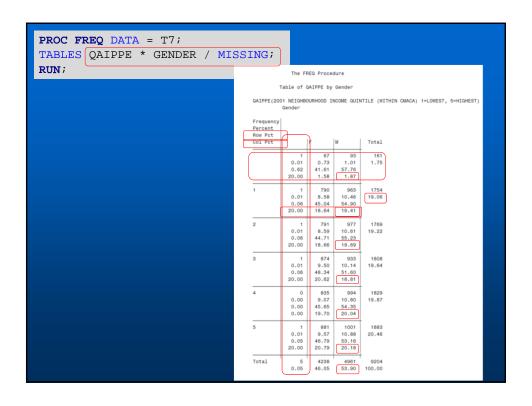
- · Allows you to get a n-way cross-tabulation of data
- Basic statistical tests are available

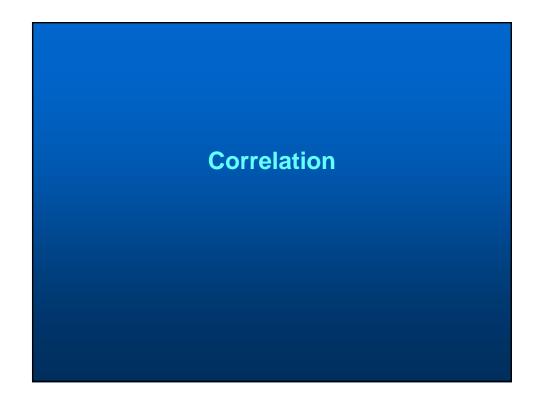
PROC FREQ <options>;
BY <variable list>;
TABLES <requests> / <options>;
RUN;

PROC FREQ DATA = T7; TABLES QAIPPE GENDER; RUN; The FREQ Procedure 2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST Cumulative Cumulative QAIPPE Frequency Percent Frequency Percent 19.40 19.40 2 1769 19.56 3523 38.96 3 1808 19.99 5331 58.95 20.23 7160 79.18 20.82 9043 100.00 1883 Frequency Missing = 161 Cumulative Cumulative Gender Frequency Percent Frequency Percent 4238 46.07 4238 46.07 4961 9199 100.00 53.93 Frequency Missing = 5





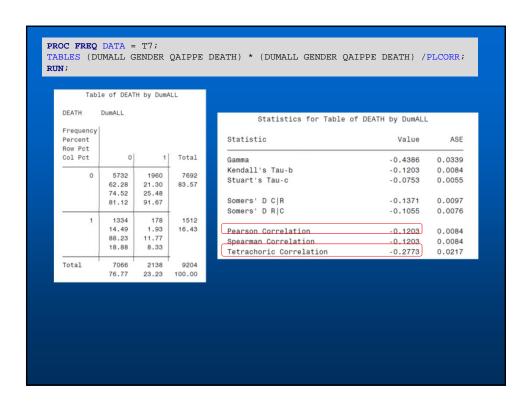


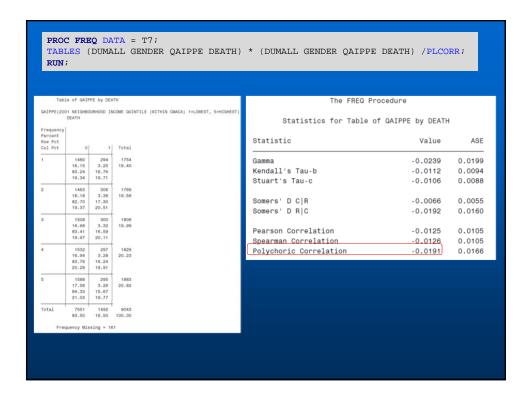


```
PROC CORR DATA = T7;
VAR DUMALL /*GENDER QAIPPE*/ DEATH;
                                      The CORR Procedure
                                 2 Variables: DumALL DEATH
                                      Simple Statistics
                                          Std Dev
                                                                Minimum
         DumALL 9204
DEATH 9204
                                          0.42232 2138
0.37055 1512
                            0.23229
                                                                            1.00000
                                                                            1.00000
                             Pearson Correlation Coefficients, N = 9204
                                   Prob > |r| under HO: Rho=0
                                           DumALL
                                                       DEATH
                                                    <.0001
                                DEATH
                                         -0.12028
                                                      1.00000
                                           <.0001
     p-value indicates probability of observing this or larger correlation coefficient
     under the null hypothesis that the correlation equals 0
```

How to get correlations for categorical data?

- Need to calculate polychoric or tetrachoric correlations
 - Techniques estimates correlation between theorized continuous variables, using observed ordinal variables
 - Tetrachoric for 2 x 2 tables
 - Polychoric for n x n tables

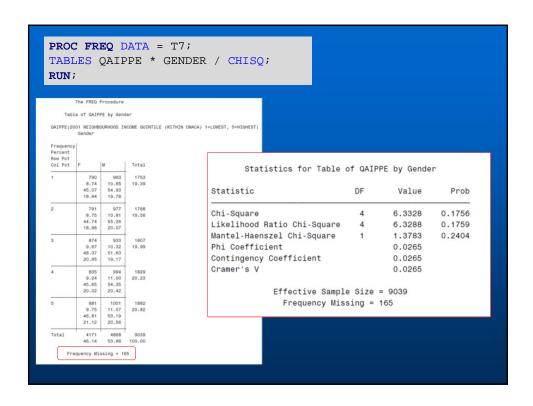




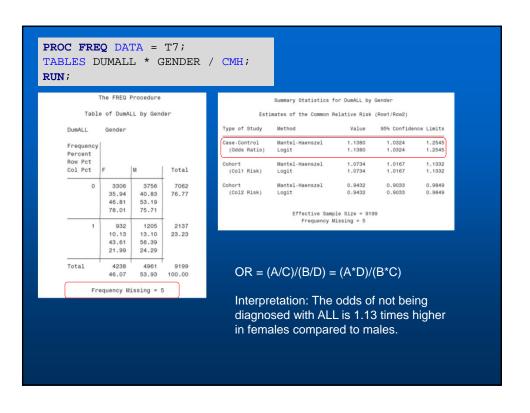
Pearson Chi-Square Test

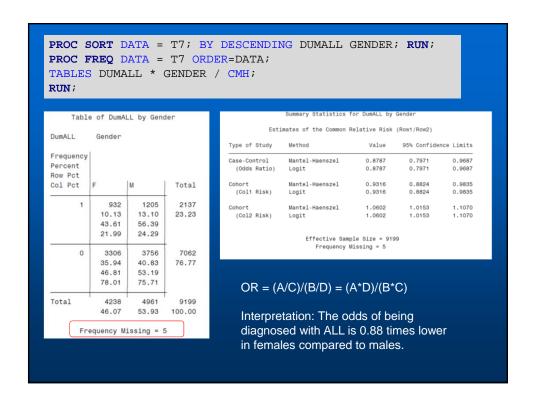
Pearson Chi-Square Test

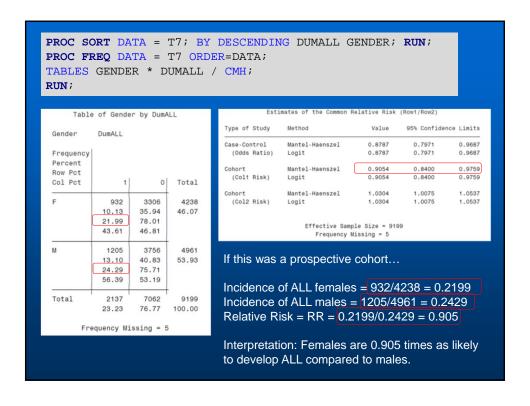
- Hypothesis test that uses the Chi-Square distribution under the null hypothesis
- Tests if the two variables are independent (related or associated)
- Tests difference between expected frequency and observed frequency in one or more categories











Other ways to generate...

- Odds ratios (OR) and relative risks (RR) are often called measures of association
- Can be generated using modelling procedures
 - Logistic regression (OR)
 - Log-binomial regression (RR)
- Models allow for further assessment
 - control of confounding
 - Estimation of effect modification

Logistic Regression

Logistic Regression

- Form of Generalized Linear Model (GLM)
- Uses the logit function to link dependent and independent variables
 - Other models use other link functions
 - Each link function comes with set of assumptions
 - LR assumptions are reasonable in most situations hence the models are robust
- Generally used for dichotomous outcomes (but not always)

```
PROC LOGISTIC DATA = T7 (DESCENDING;

CLASS GENDER) (REF='M') / PARAM = REF;

MODEL DEATH = GENDER;

RUN;

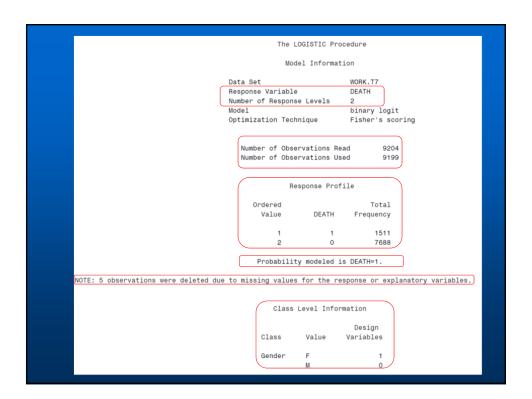
Descending: orders the outcome (death) so highest level event

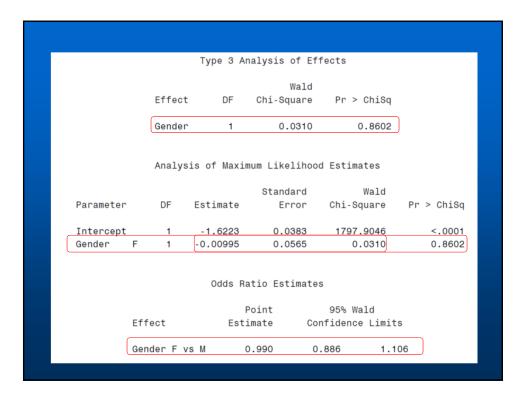
Class: tells SAS that these variables are categorical in nature

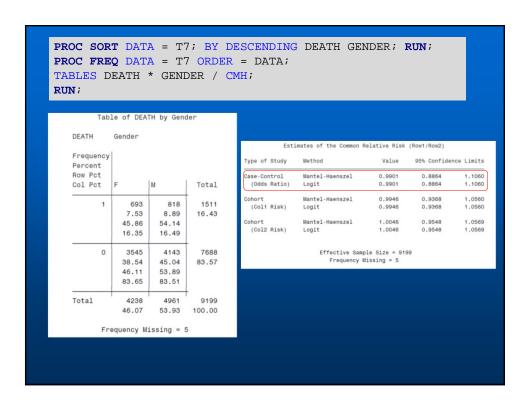
Ref: tells SAS you would like to use the 'M' (male) category as the reference group

Param = ref: tells SAS how you would like to parametrize categorical variables

Model: tells SAS what the dependent and independent variables are
```







```
PROC LOGISTIC DATA = T7 DESCENDING;
CLASS GENDER (REF='M') DumTumorType (REF='Leukemia') QAIPPE (REF='5')
        /PARAM = REF;
MODEL DEATH = GENDER DUMTUMORTYPE QAIPPE;
RUN;
                                             Model Information
                                 Data Set
                                                           WORK.T7
                                 Response Variable
                                                           DEATH
                                 Number of Response Levels
                                                           binary logit
                                 Optimization Technique
                                                           Fisher's scoring
                                   Number of Observations Read
                                                                  9204
                                   Number of Observations Used
                                                                  9039
                                              Response Profile
                                      Ordered
                                                               Total
                                                   DEATH Frequency
                                       Value
                                                                1491
                                                      0
                                                                7548
                                       Probability modeled is DEATH=1.
NOTE: 165 observations were deleted due to missing values for the response or explanatory variables
```

| | Class Level In | Tormati | .on | | |
|--------------|----------------|---------|---------|----------|---|
| Class | Value | D | esign V | ariables | |
| Gender | F | 1 | | | |
| | М | 0 | | | |
| DumTumorType | Brain Tumo | 1 | 0 | 0 | 0 |
| | Leukemia | 0 | 0 | 0 | 0 |
| | Lymphoma | 0 | 1 | 0 | 0 |
| | Missing | 0 | 0 | 1 | 0 |
| | Solid Tumo | 0 | 0 | 0 | 1 |
| QAIPPE | 1 | 1 | 0 | 0 | 0 |
| | 2 | 0 | 1 | 0 | 0 |
| | 3 | 0 | 0 | 1 | 0 |
| | 4 | 0 | 0 | 0 | 1 |
| | 5 | 0 | 0 | 0 | 0 |

| | | Тур | e 3 Analysi | s of Effects | | |
|---|-----------------------------------|-------------|--|--|--|--|
| | | | | Wald | | |
| | Effect | | DF C | hi-Square | Pr > ChiSq | |
| | Gender | | 1 | 0.7568 | 0.3843 | |
| | DumTumor | Туре | 4 | 168.9501 | <.0001 | |
| | QAIPPE | | 4 | 2.2675 | 0.6867 | |
| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiS |
| | | | | Error | Chi-Square | |
| Intercept | - | DF | -1.9867 | Error 0.0864 | Chi-Square 529.0750 | <.000 |
| Intercept Gender | F Brain Tumo | 1 | -1.9867 -0.0501 | 0.0864 0.0576 | Chi-Square 529.0750 0.7568 | <.000 0.384 |
| Intercept Gender DumTumorType | | 1 1 1 | -1.9867 -0.0501 0.7971 | 0.0864 0.0576 0.0767 | Chi-Square 529.0750 0.7568 107.9152 | Pr > Chis |
| Intercept Gender DumTumorType DumTumorType | Lymphoma | 1 1 1 1 | -1.9867 -0.0501 0.7971 -0.4975 | 0.0864 0.0576 0.0767 0.1334 | Chi-Square 529.0750 0.7568 107.9152 13.9102 | <.000 0.384 <.000 0.000 |
| Intercept Gender DumTumorType DumTumorType DumTumorType | Lymphoma Missing | 1 1 1 | -1.9867 -0.0501 0.7971 | 0.0864 0.0576 0.0767 | Chi-Square 529.0750 0.7568 107.9152 | <.000 0.384 <.000 0.000 0.482 |
| Intercept Gender DumTumorType DumTumorType | Lymphoma Missing | 1 1 1 1 1 | -1.9867 -0.0501 0.7971 -0.4975 -0.3701 | 0.0864 0.0576 0.0767 0.1334 0.5272 | 529.0750 0.7568 107.9152 13.9102 0.4927 | <.000 0.384 <.000 0.000 0.482 <.000 |
| Intercept Gender DumTumorType DumTumorType DumTumorType DumTumorType | Lymphoma Missing Solid Tumo | 1 1 1 1 1 1 | -1.9867 -0.0501 0.7971 -0.4975 -0.3701 0.3782 | 0.0864 0.0576 0.0767 0.1334 0.5272 0.0755 | 529.0750 0.7568 107.9152 13.9102 0.4927 25.0785 | <.000 0.384 <.000 |
| Intercept Gender DumTumorType DumTumorType DumTumorType DumTumorType QAIPPE | Lymphoma Missing Solid Tumo | 1 1 1 1 1 1 | -1.9867 -0.0501 0.7971 -0.4975 -0.3701 0.3782 0.0721 | 0.0864 0.0576 0.0767 0.1334 0.5272 0.0755 | 529.0750 0.7568 107.9152 13.9102 0.4927 25.0785 0.6274 | <.000 0.384 <.000 0.000 0.482 <.000 |

| Odds Ratio Estimates Point 95% Wald Estimate Confidence Limits Gender F vs M 0.951 0.850 1.065 DumTumorType Brain Tumo vs Leukemia 2.219 1.909 2.579 DumTumorType Lymphoma vs Leukemia 0.608 0.468 0.790 DumTumorType Missing vs Leukemia 0.691 0.246 1.941 DumTumorType Solid Tumo vs Leukemia 1.460 1.259 1.693 QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 QAIPPE 3 vs 5 1.067 0.894 1.274 |
|--|
| Effect Estimate Confidence Limits Gender F vs M 0.951 0.850 1.065 DumTumorType Brain Tumo vs Leukemia 2.219 1.909 2.579 DumTumorType Lymphoma vs Leukemia 0.608 0.468 0.790 DumTumorType Missing vs Leukemia 0.691 0.246 1.941 DumTumorType Solid Tumo vs Leukemia 1.460 1.259 1.693 QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 |
| DumTumorType Brain Tumo vs Leukemia 2.219 1.909 2.579 DumTumorType Lymphoma vs Leukemia 0.608 0.468 0.790 DumTumorType Missing vs Leukemia 0.691 0.246 1.941 DumTumorType Solid Tumo vs Leukemia 1.460 1.259 1.693 QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 |
| DumTumorType Lymphoma vs Leukemia 0.608 0.468 0.790 DumTumorType Missing vs Leukemia 0.691 0.246 1.941 DumTumorType Solid Tumo vs Leukemia 1.460 1.259 1.693 QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 |
| DumTumorType Missing vs Leukemia 0.691 0.246 1.941 DumTumorType Solid Tumo vs Leukemia 1.460 1.259 1.693 QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 |
| DumTumorType Solid Tumo vs Leukemia 1.460 1.259 1.693 QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 |
| QAIPPE 1 vs 5 1.075 0.899 1.285 QAIPPE 2 vs 5 1.139 0.954 1.359 |
| QAIPPE 2 vs 5 1.139 0.954 1.359 |
| |
| QAIPPE 3 vs 5 1.067 0.894 1.274 |
| |
| OAIPPE 4 vs 5 1.037 0.869 1.239 |

Topics Covered

- SAS overview revisited
- Data Analysis
 - Bi-Variate Tables and Stratification
 - Correlation
 - Chi-square Test
 - Odds Ratios / Relative Risk
 - Introduction to Logistic Regression