CYP-C Data Analysis Using SAS I

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

POGO
PEDIATRIC ONCOLOGY GROUP OF ONTARIO

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

QAIPPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1754	19.40	1754	19.40
2	1769	19.56	3523	38.96
3	1808	19.99	5331	58.95
4	1829	20.23	7160	79.18
5	1883	20.82	9043	100.00

Frequency Missing = 161

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	4238	46.07	4238	46.07
M	4961	53.93	9199	100.00

Frequency Missing = 5

```
PROC FREQ DATA = T7;
TABLES QAIPPE / MISSING;
RUN;
```

The FREQ Procedure

2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST

QAIPPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	161	1.75	161	1.75
1	1754	19.06	1915	20.81
2	1769	19.22	3684	40.03
3	1808	19.64	5492	59.67
4	1829	19.87	7321	79.54
5	1883	20.46	9204	100.00

```
PROC FREQ DATA = T7;
BY GENDER;
TABLES QAIPPE / MISSING;
RUN;
```

Gender=' '

The FREQ Procedure

2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST

Gender=F

The FREQ Procedure

2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST

QAIPPE
1
2
3
5

Gender=M

The FREQ Procedure

2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST

QAIPPE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	93	1.87	93	1.87
1	963	19.41	1056	21.29
2	977	19.69	2033	40.98
3	933	18.81	2966	59.79
4	994	20.04	3960	79.82
5	1001	20.18	4961	100.00

```
PROC FREQ DATA = T7;
TABLES QAIPPE * GENDER / MISSING;
RUN;
```

The FREQ Procedure

Table of QAIPPE by Gender

QAIPPE(2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST)
Gender

Frequency Percent Row Pct Col Pct	F	M	Total
	67	93	161
	0.73	1.01	1.75
	41.61	57.76	
	1.58	1.87	
1	790	963	1754
	8.58	10.46	19.06
	45.04	54.90	
	18.64	19.41	
2	791	977	1769
	8.59	10.61	19.22
	44.71	55.23	
	18.66	19.69	
3	874	933	1808
	9.50	10.14	19.64
	48.34	51.60	
	20.62	18.81	
4	835	994	1829
	9.07	10.80	19.87
	45.65	54.35	
	19.70	20.04	
5	881	1001	1883
	9.57	10.88	20.46
	46.79	53.16	
	20.79	20.18	
Total	4238	4961	9204
	46.05	53.90	100.00

Correlation

```
PROC CORR DATA = T7;
VAR DUMALL /*GENDER QAIPPE*/ DEATH;
RUN;
```

The CORR Procedure

2 Variables: DumALL DEATH

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
DumALL	9204	0.23229	0.42232	2138	0	1.00000
DEATH	9204	0.16428	0.37055	1512	0	1.00000

Pearson Correlation Coefficients, N = 9204
Prob > |r| under H0: Rho=0

	DumALL	DEATH
DumALL	1.00000	-0.12028 <.0001
DEATH	-0.12028 <.0001	1.00000

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**

```
PROC FREQ DATA = T7;
TABLES (DUMALL GENDER QAIPPE DEATH) * (DUMALL GENDER QAIPPE DEATH) /PLCORR;
RUN;
```

Table of DEATH by DumALL

DEATH	DumALL		
Frequency Percent Row Pct Col Pct	0	1	Total
0	5732 62.28 74.52 81.12	1960 21.30 25.48 91.67	7692 83.57
1	1334 14.49 88.23 18.88	178 1.93 11.77 8.33	1512 16.43
Total	7066 76.77	2138 23.23	9204 100.00

Statistics for Table of DEATH by DumALL

Statistic	Value	ASE
Gamma	-0.4386	0.0339
Kendall's Tau-b	-0.1203	0.0084
Stuart's Tau-c	-0.0753	0.0055
Somers' D C|R	-0.1371	0.0097
Somers' D R|C	-0.1055	0.0076
Pearson Correlation	-0.1203	0.0084
Spearman Correlation	-0.1203	0.0084
Tetrachoric Correlation	-0.2773	0.0217

```
PROC FREQ DATA = T7;
TABLES (DUMALL GENDER QAIPPE DEATH) * (DUMALL GENDER QAIPPE DEATH) /PLCORR;
RUN;
```

Table of QAIPPE by DEATH

QAIPPE(2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST)
DEATH

Frequency Percent Row Pct Col Pct	0	1	Total
1	1460 16.15 83.24 19.34	294 3.25 16.76 19.71	1754 19.40
2	1463 16.18 82.70 19.37	306 3.38 17.30 20.51	1769 19.56
3	1508 16.68 83.41 19.97	300 3.32 16.59 20.11	1808 19.99
4	1532 16.94 83.76 20.29	297 3.28 16.24 19.91	1829 20.23
5	1588 17.56 84.33 21.03	295 3.26 15.67 19.77	1883 20.82
Total	7551 83.50	1492 16.50	9043 100.00

Frequency Missing = 161

The FREQ Procedure

Statistics for Table of QAIPPE by DEATH

Statistic	Value	ASE
Gamma	-0.0239	0.0199
Kendall's Tau-b	-0.0112	0.0094
Stuart's Tau-c	-0.0106	0.0088
Somers' D C|R	-0.0066	0.0055
Somers' D R|C	-0.0192	0.0160
Pearson Correlation	-0.0125	0.0105
Spearman Correlation	-0.0126	0.0105
Polychoric Correlation	-0.0191	0.0166

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

```
PROC FREQ DATA = T7;
TABLES QAIPPE * GENDER / CHISQ;
RUN;
```

The FREQ Procedure

Table of QAIPPE by Gender

QAIPPE(2001 NEIGHBOURHOOD INCOME QUINTILE (WITHIN CMACA) 1=LOWEST, 5=HIGHEST)
Gender

Frequency Percent Row Pct Col Pct	F	M	Total
1	790 8.74 45.07 18.94	963 10.65 54.93 19.78	1753 19.39
2	791 8.75 44.74 18.96	977 10.81 55.26 20.07	1768 19.56
3	874 9.67 48.37 20.95	933 10.32 51.63 19.17	1807 19.99
4	835 9.24 45.65 20.02	994 11.00 54.35 20.42	1829 20.23
5	881 9.75 46.81 21.12	1001 11.07 53.19 20.56	1882 20.82
Total	4171 46.14	4868 53.86	9039 100.00

Frequency Missing = 165

Statistics for Table of QAIPPE by Gender

Statistic	DF	Value	Prob
Chi-Square	4	6.3328	0.1756
Likelihood Ratio Chi-Square	4	6.3288	0.1759
Mantel-Haenszel Chi-Square	1	1.3783	0.2404
Phi Coefficient		0.0265	
Contingency Coefficient		0.0265	
Cramer's V		0.0265	

Effective Sample Size = 9039
Frequency Missing = 165

Odds Ratio and Relative Risk

```
PROC FREQ DATA = T7;
TABLES DUMALL * GENDER / CMH;
RUN;
```

The FREQ Procedure

Table of DumALL by Gender

DumALL	Gender		
Frequency / Percent / Row Pct / Col Pct	F	M	Total
0	3306 35.94 46.81 78.01	3756 40.83 53.19 75.71	7062 76.77
1	932 10.13 43.61 21.99	1205 13.10 56.39 24.29	2137 23.23
Total	4238 46.07	4961 53.93	9199 100.00

Frequency Missing = 5

Summary Statistics for DumALL by Gender

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	1.1380	1.0324	1.2545
	Logit	1.1380	1.0324	1.2545
Cohort (Col1 Risk)	Mantel-Haenszel	1.0734	1.0167	1.1332
	Logit	1.0734	1.0167	1.1332
Cohort (Col2 Risk)	Mantel-Haenszel	0.9432	0.9033	0.9849
	Logit	0.9432	0.9033	0.9849

Effective Sample Size = 9199
Frequency Missing = 5

OR = (A/C)/(B/D) = (A*D)/(B*C)

Interpretation: The odds of not being diagnosed with ALL is 1.13 times higher in females compared to males.

```
PROC SORT DATA = T7; BY DESCENDING DUMALL GENDER; RUN;
PROC FREQ DATA = T7 ORDER=DATA;
TABLES DUMALL * GENDER / CMH;
RUN;
```

Table of DumALL by Gender

DumALL	Gender		
Frequency / Percent / Row Pct / Col Pct	F	M	Total
1	932 10.13 43.61 21.99	1205 13.10 56.39 24.29	2137 23.23
0	3306 35.94 46.81 78.01	3756 40.83 53.19 75.71	7062 76.77
Total	4238 46.07	4961 53.93	9199 100.00

Frequency Missing = 5

Summary Statistics for DumALL by Gender

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	0.8787	0.7971	0.9687
	Logit	0.8787	0.7971	0.9687
Cohort (Col1 Risk)	Mantel-Haenszel	0.9316	0.8824	0.9835
	Logit	0.9316	0.8824	0.9835
Cohort (Col2 Risk)	Mantel-Haenszel	1.0602	1.0153	1.1070
	Logit	1.0602	1.0153	1.1070

Effective Sample Size = 9199
Frequency Missing = 5

OR = (A/C)/(B/D) = (A*D)/(B*C)

Interpretation: The odds of being diagnosed with ALL is 0.88 times lower in females compared to males.

```
PROC SORT DATA = T7; BY DESCENDING DUMALL GENDER; RUN;
PROC FREQ DATA = T7 ORDER=DATA;
TABLES GENDER * DUMALL / CMH;
RUN;
```

Table of Gender by DumALL

Gender	DumALL		
Frequency / Percent / Row Pct / Col Pct	1	0	Total
F	932 10.13 21.99 43.61	3306 35.94 78.01 46.81	4238 46.07
M	1205 13.10 24.29 56.39	3756 40.83 75.71 53.19	4961 53.93
Total	2137 23.23	7062 76.77	9199 100.00

Frequency Missing = 5

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	0.8787	0.7971	0.9687
	Logit	0.8787	0.7971	0.9687
Cohort (Col1 Risk)	Mantel-Haenszel	0.9054	0.8400	0.9759
	Logit	0.9054	0.8400	0.9759
Cohort (Col2 Risk)	Mantel-Haenszel	1.0304	1.0075	1.0537
	Logit	1.0304	1.0075	1.0537

Effective Sample Size = 9199
Frequency Missing = 5

If this was a prospective cohort…

Incidence of ALL females = 932/4238 = 0.2199
Incidence of ALL males = 1205/4961 = 0.2429
Relative Risk = RR = 0.2199/0.2429 = 0.905

Interpretation: Females are 0.905 times as likely to develop ALL compared to males.

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

The LOGISTIC Procedure

Model Information

Data Set	WORK.T7
Response Variable	DEATH
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	9204
Number of Observations Used	9199

Response Profile

Ordered Value	DEATH	Total Frequency
1	1	1511
2	0	7688

Probability modeled is DEATH=1.

NOTE: 5 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information

Class	Value	Design Variables
Gender	F	1
	M	0

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	0.0310	0.8602

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.6223	0.0383	1797.9046	<.0001
Gender	F	1	-0.00995	0.0565	0.0310	0.8602

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Gender F vs M	0.990	0.886	1.106

```
PROC SORT DATA = T7; BY DESCENDING DEATH GENDER; RUN;
PROC FREQ DATA = T7 ORDER = DATA;
TABLES DEATH * GENDER / CMH;
RUN;
```

Table of DEATH by Gender

DEATH / Gender Frequency Percent Row Pct Col Pct	F	M	Total
1	693 7.53 45.86 16.35	818 8.89 54.14 16.49	1511 16.43
0	3545 38.54 46.11 83.65	4143 45.04 53.89 83.51	7688 83.57
Total	4238 46.07	4961 53.93	9199 100.00

Frequency Missing = 5

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	0.9901	0.8864	1.1060
	Logit	0.9901	0.8864	1.1060
Cohort (Col1 Risk)	Mantel-Haenszel	0.9946	0.9368	1.0560
	Logit	0.9946	0.9368	1.0560
Cohort (Col2 Risk)	Mantel-Haenszel	1.0046	0.9548	1.0569
	Logit	1.0046	0.9548	1.0569

Effective Sample Size = 9199
Frequency Missing = 5

```
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	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	9204
Number of Observations Used	9039

Response Profile

Ordered Value	DEATH	Total Frequency
1	1	1491
2	0	7548

Probability modeled is DEATH=1.

NOTE: 165 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information

Class	Value	Design Variables			
Gender	F	1			
	M	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

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	0.7568	0.3843
DumTumorType	4	168.9501	<.0001
QAIPPE	4	2.2675	0.6867

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.9867	0.0864	529.0750	<.0001
Gender	F	1	-0.0501	0.0576	0.7568	0.3843
DumTumorType	Brain Tumo	1	0.7971	0.0767	107.9152	<.0001
DumTumorType	Lymphoma	1	-0.4975	0.1334	13.9102	0.0002
DumTumorType	Missing	1	-0.3701	0.5272	0.4927	0.4827
DumTumorType	Solid Tumo	1	0.3782	0.0755	25.0785	<.0001
QAIPPE	1	1	0.0721	0.0910	0.6274	0.4283
QAIPPE	2	1	0.1301	0.0902	2.0809	0.1492
QAIPPE	3	1	0.0647	0.0904	0.5121	0.4742
QAIPPE	4	1	0.0367	0.0906	0.1646	0.6850

Odds Ratio Estimates

Effect		Point Estimate	95% Wald 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
QAIPPE	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