BOOST YOUR CONFIDENCE (INTERVALS) WITH SAS

Brought to you by:

Peter Langlois, PhD
Birth Defects Epidemiology & Surveillance Branch,
Texas Dept State Health Services

Background

Confidence Interval Definition

- DEFINITION: An interval around a statistic that contains the true underlying value of the statistic (the population parameter) a certain amount of the time

Confidence Interval Definition

- Example: a survey of 50 SAS programmers finds that the average IQ is 130 $\pm$ 10

- If we did 100 surveys, the average IQ should be between 120 and 140 in 95 of them

Confidence Interval Definition

- 95% confidence interval bounded by the upper 95% confidence limit and the lower 95% confidence limit

- 95% just conventional. Can have for e.g.:
  - 90% CIs (narrower)
  - 99% CIs (wider)
- CI for any level (95% etc) is narrower if based on more observations

General Formula

- Can make CIs around almost any statistic you calculate, for example…

- Data summaries (1 var) such as:
  - Categorical variables: proportion
  - Continuous variables: mean

- Statistics resulting from hypothesis tests (2+ vars):
  - Correlation, regression slope
  - Relative risk, odds ratios

Using Confidence Intervals

Why Use CIs?

- Gives audience idea of impact of chance
- Gives reasonable bounds on your result(s)
- Can check if your data are compatible with a certain value
  - (From data summary): Does 95% confidence interval of IQ in SAS programmers include 100?
  - (From hypothesis testing): Is occurrence of schizophrenia higher in SAS programmers or SPSS users? (Does relative risk = 1.00?)

Why Not Use CIs?

- Some organizations consider their figures to be a census, not a sample
- Increases statistical work for staff
- Some data users may find the concept confusing

Overlapping CIs vs Hypothesis Tests

- Example: Want to compare prevalence of schizophrenia in SAS vs SPSS users
- You could:

(A) Calculate schizophrenia prevalence and 95% CI for each group and see if overlap OR

(B) Calculate prevalence ratio of one group vs another, and see if includes 1.00

Overlapping CIs vs Hypothesis Tests

- Answer: (B) usually better
- Why? More statistical power

- Why even consider the first approach?
  - Easier to do if already have data summaries published
  - Can’t anticipate all comparisons readers will want to make

Confidence Intervals
For Data Summaries

CIs for Means

- Based on normal distribution
- Where your study sample is large (# of subjects > 30), the sampling distribution → normal, and you can use:

CI = obsd mean ± 1.96 X standard error (of mean)

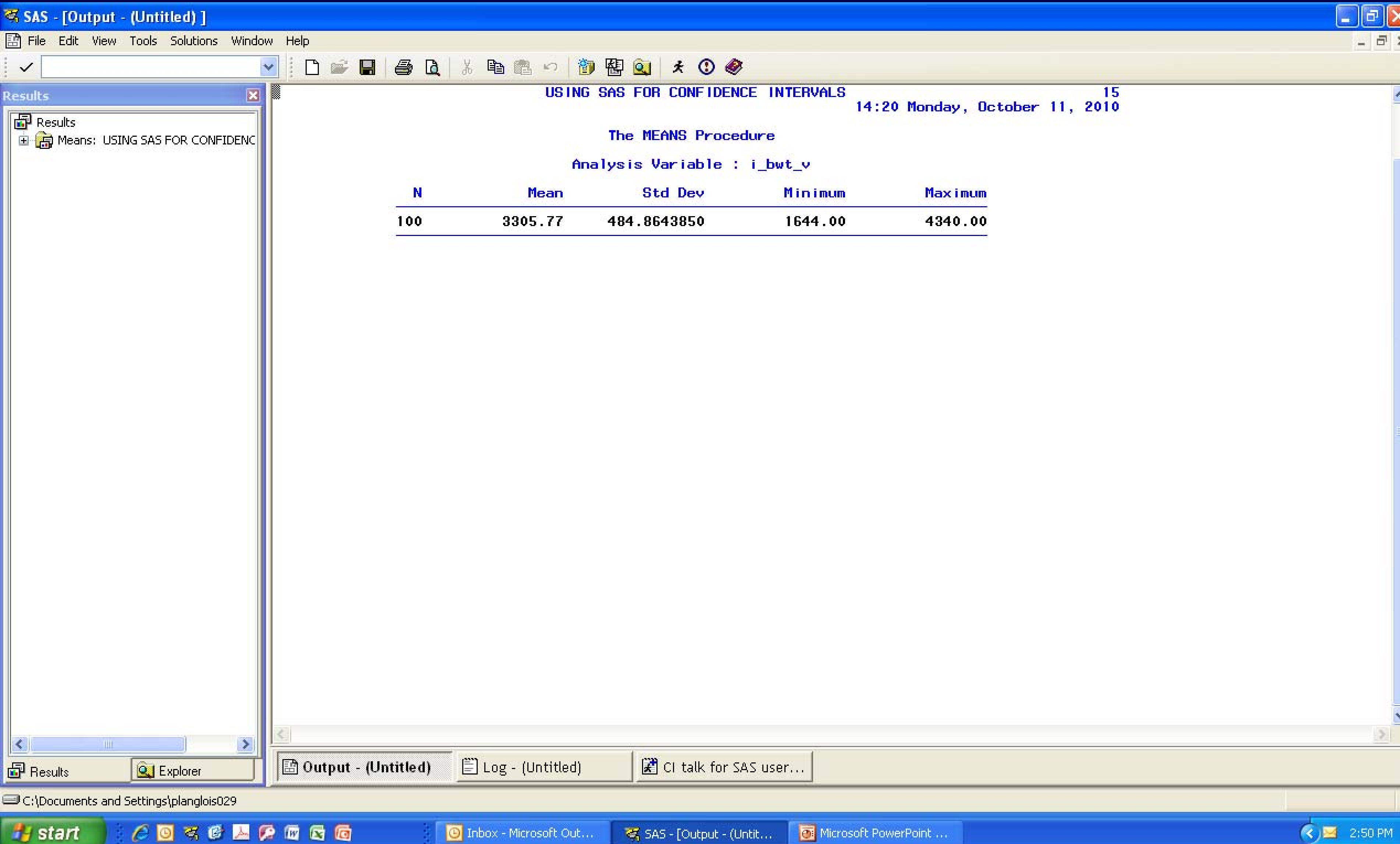
= obsd mean ± 1.96 X standard dev'n / sqrt(n)

$$= \bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

CIs for Means: Using SAS

- Let SAS do it!
- Recall: Using SAS for simple desc stats:

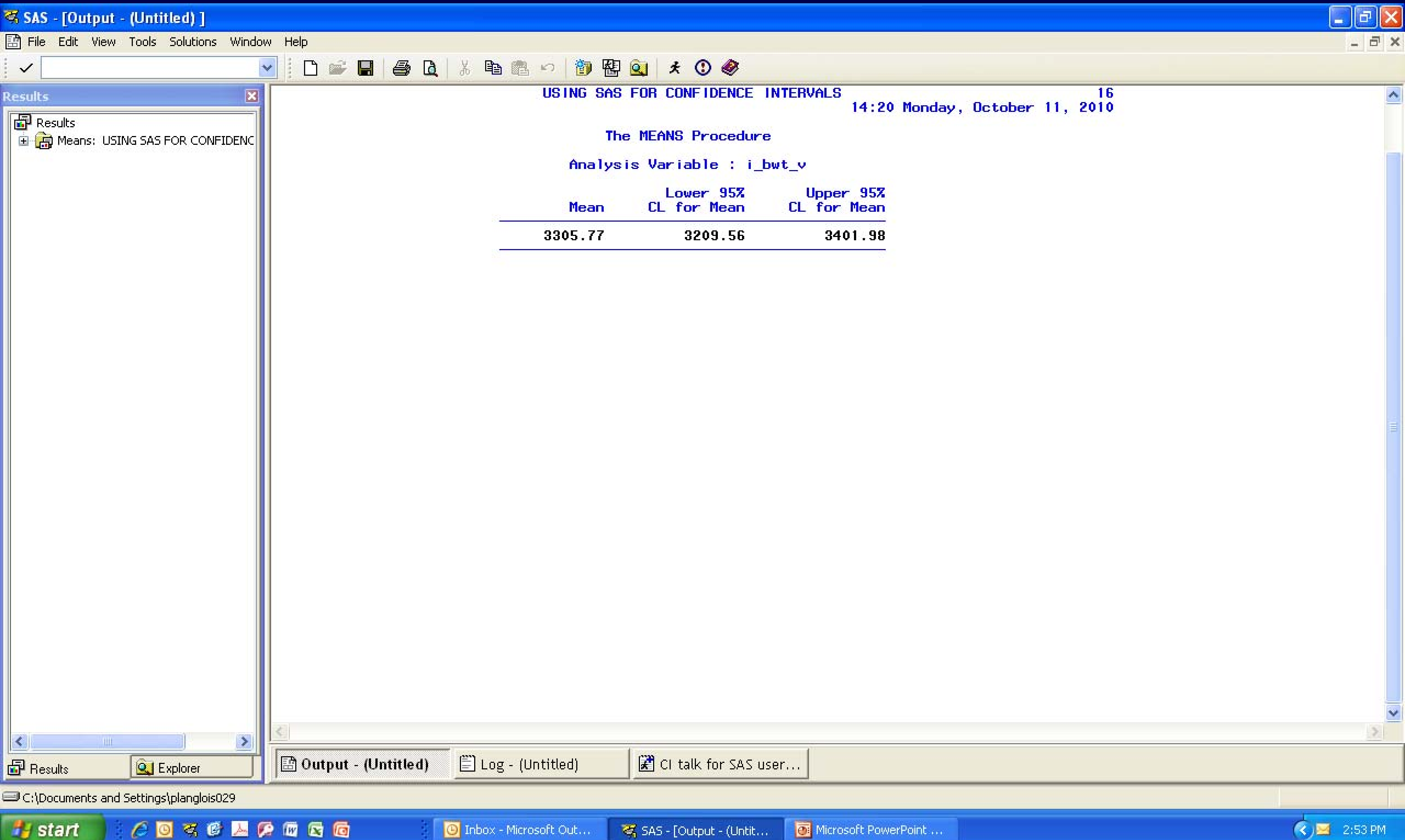
```
proc means;
   var <variable name>;
```


SAS - [Output - (Untitled)]
File Edit View Tools Solutions Window Help
Results
Means: USING SAS FOR CONFIDENC
USING SAS FOR CONFIDENCE INTERVALS 15
14:20 Monday, October 11, 2010
The MEANS Procedure
Analysis Variable : i_bwt_v
N Mean Std Dev Minimum Maximum
100 3305.77 484.8643850 1644.00 4340.00
Results Explorer
Output - (Untitled)
Log - (Untitled)
CI talk for SAS user...
C:\Documents and Settings\planglois029
start
Inbox - Microsoft Out...
SAS - [Output - (Untit...
Microsoft PowerPoint ...
2:50 PM

CIs for Means: Using SAS

- To get confidence limits, request “clm” as PROC MEANS option
- (Request “mean” to get the mean printed out too)

```
proc means mean clm;
   var <variable name>;
```


SAS - [Output - (Untitled)]
USING SAS FOR CONFIDENCE INTERVALS 16
14:20 Monday, October 11, 2010
The MEANS Procedure
Analysis Variable : i_bwt_v
Mean
Lower 95% CL for Mean
Upper 95% CL for Mean
3305.77
3209.56
3401.98
Results
Means: USING SAS FOR CONFIDENC
Output - (Untitled)
Log - (Untitled)
CI talk for SAS user...
C:\Documents and Settings\planglois029
2:53 PM

CIs for Proportions

- Based on binomial distribution
- Where your study sample is large (# with and without the characteristic > 5), the sampling distribution → normal, and you can use:
- = obsd proportion ± 1.96 X SE of prop'n

$$= p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

CIs for Proportions: Using SAS

- To get confidence limits, request "binomial" as TABLE option in PROC FREQ:

```
proc freq;
   table m_edu2g_v / binomial;
```

SAS - [Output - (Untitled)]

File Edit View Tools Solutions Window Help

```
                     USING SAS FOR CONFIDENCE INTERVALS                        27
                                                 14:20 Monday, October 11, 2010

                              The FREQ Procedure

                                                   Cumulative    Cumulative
   m_edu2g_v          Frequency     Percent     Frequency      Percent
   --------------------------------------------------------------------------
   < High School            35       35.00            35         35.00
   High School +            65       65.00           100        100.00

                           Binomial Proportion for
                          m_edu2g_v = < High School
                         ----------------------------------
                         Proportion                  0.3500
                         ASE                         0.0477
                         95% Lower Conf Limit        0.2565
                         95% Upper Conf Limit        0.4435

                         Exact Conf Limits
                         95% Lower Conf Limit        0.2573
                         95% Upper Conf Limit        0.4518

                          Test of H0: Proportion = 0.5

                         ASE under H0                0.0500
                         Z                          -3.0000
                         One-sided Pr <  Z           0.0013
                         Two-sided Pr > |Z|          0.0027

                              Sample Size = 100
```

Results
Freq: USING SAS FOR CONFIDENCE

Output - (Untitled) | Log - (Untitled) | CI talk for SAS user...

C:\Documents and Settings\planglois029

SAS - [Output - (Untitled)]

```
            USING SAS FOR CONFIDENCE INTERVALS                          27
                                      14:20 Monday, October 11, 2010

                     The FREQ Procedure

                                                Cumulative    Cumulative
m_edu2g_v        Frequency     Percent      Frequency       Percent
--------------------------------------------------------------------
< High School          35       35.00             35         35.00
High School +          65       65.00            100        100.00


                  Binomial Proportion for
                 m_edu2g_v = < High School
                 --------------------------------
                 Proportion               0.3500
                 ASE                      0.0477
                 95% Lower Conf Limit     0.2565
                 95% Upper Conf Limit     0.4435

                 Exact Conf Limits
                 95% Lower Conf Limit     0.2573
                 95% Upper Conf Limit     0.4518

                 Test of H0: Proportion = 0.5

                 ASE under H0             0.0500
                 Z                       -3.0000
                 One-sided Pr <  Z        0.0013
                 Two-sided Pr > |Z|       0.0027

                     Sample Size = 100
```

CIs for Rates of Rare Outcomes

- Follows Poisson distribution
- Standard error based on number of cases

CIs for Rates of Rare Outcomes

PREVALENCE BASED ON < 30 CASES

- Find upper and lower 95% conf limits for # cases:
  - In table (e.g. poisson.xls)
  - Using PEPI program POISSON
- Multiply both by (10000) and divide by # population or live births to express as birth prevalence

Microsoft Excel - poisson base data.xls

COUNT	LFACTOR	UFACTOR
0	0.0000	3.6889
1	0.0253	5.5716
2	0.2422	7.2247
3	0.6187	8.7673
4	1.0899	10.2416
5	1.6235	11.6683
6	2.2019	13.0595
7	2.8144	14.4227
8	3.4538	15.7632
9	4.1154	17.0848
10	4.7954	18.3904
11	5.4912	19.6820
12	6.2006	20.9616
13	6.9220	22.2304
14	7.6539	23.4896
15	8.3954	24.7402
16	9.1454	25.9830
17	9.9031	27.2186
18	10.6679	28.4478
19	11.4392	29.6709
20	12.2165	30.8884
21	12.9993	32.1007
22	13.7873	33.3083
23	14.5800	34.5113
24	15.3773	35.7101
25	16.1787	36.9049
26	16.9841	38.0960
27	17.7932	39.2836
28	18.6058	40.4678
29	19.4218	41.6488
30	20.2409	42.8269
31	21.0630	44.0020

POISSON

CIs for Rates of Rare Outcomes

EXAMPLE

- # cases of anophthalmia in CA 1983-1986 = 18
- Prevalence = (18 * 10,000 / 452,287) = 0.40
- Looking in table, lower 95% CL is 10.67 and upper 95% CL is 28.45 for # cases
- To express as prevalence CLs, multiply both by 10,000 and divide by 452,287 live births
- Lower 95% CL = 0.24, upper 95% CL = 0.63
- Thus we say prev = 0.40 cases per 10,000 live births, 95% CI = 0.24 - 0.63

CIs for Rates of Rare Outcomes

PREVALENCE BASED ON 30+ CASES

- Considered large # cases (more or less)

- Poisson → normal distribution
- Can use normal approximation in SAS code
- Several equations for doing this, yielding similar results

CIs for Rates of Rare Outcomes

- SAS code for obs with few cases: Combine with Poisson lower and upper limits for cases (get lfactor and ufactor for the observed # cases):

```
proc sort data=b1;  by count;

proc sort data=lib2.poisson out=poisson;
  by count;

data c1 prob2;
  merge b1(in=b) poisson(in=p);
  by count;
  if b;
```

CIs for Rates of Rare Outcomes

- SAS code: Calculate CIs for obs with many cases too

```
data c2;
  set c1;
  calcrate = count * 10000 / births;
  if count le 30 then do;
    calclcl = lfactor * 10000 / births;
    calcucl = ufactor * 10000 / births;
    end;
  else if count > 30 then do;
    calclcl = ((count / births) - (1.96 * sqrt(count) / births)) * 10000;
    calcucl = ((count / births) + (1.96 * sqrt(count) / births)) * 10000;
    end;
  rate = round(calcrate,.01);
  lcl = round(calclcl,.01);
  ucl = round(calcucl,.01);
  rename count=cases;
```

Reminder

- To compare groups (e.g. whether rates are statistically different), can calculate 95% confidence intervals and see if they overlap
- Better to do hypothesis testing

Confidence Intervals For Statistics From Hypothesis Tests / Measures of Association

Types of Hypothesis Tests / Measures of Association

Indep Var	Dep Var	Approach (SAS PROC)
Categorical	Categorical	Contingency tables (FREQ)
Cat. / Cont.	Categorical	Logistic reg. (LOGISTIC)
Categorical	Cont.	ANOVA (GLM); if 2 levels then T-test (TTEST or GLM)
Cont.	Cont.	Linear reg. (REG or GLM)
Cat. / Cont.	Rates	Poisson reg. (GENMOD)

Contingency Tables Using SAS

- Recall: Using SAS to produce the basic 2x2 table:

```
proc freq;
   tables <indep var> * <outcome var>;
```

- To get odds ratios and their CIs, request measures of association:

```
proc freq;
   tables <indep var> * <outcome var>
   / measures;
```

Contingency Tables Using SAS

QUESTION: Is mother's education associated with % low birth weight babies?

- Independent variable = m_edu2g_v
- Outcome variable = lbw
- If no statistically significant association, 95% CI will include 1.00

```
proc freq;
   tables m_edu2g_v * lbw / measures;
```

SAS - [Output - (Untitled)]

File Edit View Tools Solutions Window Help

Results

Results
Freq: USING SAS FOR CONFIDENCE

```
                     USING SAS FOR CONFIDENCE INTERVALS                      7
                                         09:40 Tuesday, October 12, 2010

                            The FREQ Procedure

                        Table of m_edu2g_v by lbw

                     m_edu2g_v       lbw

                     Frequency
                     Percent
                     Row Pct
                     Col Pct        |Low bwt |Normal b|  Total
                                    |        |wt      |
                     ---------------+--------+--------+
                     < High School  |  10769 | 113937 | 124706
                                    |   2.65 |  28.05 |  30.70
                                    |   8.64 |  91.36 |
                                    |  31.40 |  30.63 |
                     ---------------+--------+--------+
                     High School +  |  23525 | 258010 | 281535
                                    |   5.79 |  63.51 |  69.30
                                    |   8.36 |  91.64 |
                                    |  68.60 |  69.37 |
                     ---------------+--------+--------+
                     Total             34294   371947   406241
                                        8.44    91.56   100.00

                             Frequency Missing = 539


                     Statistics for Table of m_edu2g_v by lbw

              Statistic                              Value       ASE
              ------------------------------------------------------
              Gamma                                 0.0180    0.0061
              Kendall's Tau-b                       0.0046    0.0016
              Stuart's Tau-c                        0.0024    0.0008

              Somers' D C|R                         0.0028    0.0010
              Somers' D R|C                         0.0077    0.0026

              Pearson Correlation                   0.0046    0.0016
              Spearman Correlation                  0.0046    0.0016
```

Results | Explorer

Output - (Untitled) | Log - (Untitled) | CI talk for SAS user...

C:\Documents and Settings\planglois029

start | Inbox - Microsoft Out... | Microsoft PowerPoint ... | SAS - [Output - (Untit... | 10:12 AM

SAS - [Output - (Untitled)]

USING SAS FOR CONFIDENCE INTERVALS 8

09:40 Tuesday, October 12, 2010

The FREQ Procedure

Statistics for Table of m_edu2g_v by lbw

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.0366	1.0122	1.0616
Cohort (Col1 Risk)	1.0335	1.0112	1.0562
Cohort (Col2 Risk)	0.9969	0.9949	0.9990

Effective Sample Size = 406241
Frequency Missing = 539

Logistic Regression: Using SAS

- Can use PROC LOGISTIC in SAS; nice since it will exponentiate the slope (b) and its 95% confidence interval
- Basic syntax:

```
proc logistic;
   model <outcome var> = <independent var>;
```

- For low birthweight example:

```
proc logistic;
   model lbw = m_edu2g_v;
```

SAS - [Output - (Untitled)]

USING SAS FOR CONFIDENCE INTERVALS 14

09:40 Tuesday, October 12, 2010

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.7032	1	0.0032
Score	8.7372	1	0.0031
Wald	8.8380	1	0.0030

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3226	0.0213	11912.0530	<.0001
m_edu2g_v	1	-0.0362	0.0122	8.8380	0.0030

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
m_edu2g_v	0.964	0.942	0.988

Association of Predicted Probabilities and Observed Responses

Percent Concordant	21.8	Somers' D	0.008
Percent Discordant	21.0	Gamma	0.018
Percent Tied	57.2	Tau-a	0.001
Pairs	12755550418	c	0.504

Logistic Regression: Using SAS

- (One way) to get correct odds ratio: declare independent var to be a classification (categorical) var:

```
proc logistic;
   class m_edu2g_v;
   model lbw = m_edu2g_v;
```

SAS - [Output - (Untitled)]

File Edit View Tools Solutions Window Help

Results

Results
Logistic: USING SAS FOR CONFIDEN

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	235157.79	235151.08
SC	235168.70	235172.91
-2 Log L	235155.79	235147.08

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.7032	1	0.0032
Score	8.7372	1	0.0031
Wald	8.8380	1	0.0030

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
m_edu2g_v	1	8.8380	0.0030

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3768	0.00608	152684.523	<.0001
m_edu2g_v	< High School	1	0.0181	0.00608	8.8380	0.0030

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
m_edu2g_v < High School vs High School +	1.037	1.012	1.062

Results Explorer

Output - (Untitled) Log - (Untitled) CI talk for SAS user...

C:\Documents and Settings\planglois029

start Inbox - Microsoft Out... Microsoft PowerPoint ... SAS - [Output - (Untit... 10:34 AM

Logistic Regression: Using SAS

- Comparing results from PROC FREQ and PROC LOGISTIC:

SAS Proc	OR	95% CI
FREQ	1.0366	1.0122 – 1.0616
LOGISTIC	1.037	1.012 – 1.062

Logistic Regression With Multiple Predictor Variables

- Like other regression, the slope (b) is adjusted for all other independent variables in the model
- SAS takes both cont and categorical vars
  - SAS assumes ind vars are continuous
  - If categorical, list in CLASS statement and SAS creates dummy vars automatically

```
proc logistic;
   class <categorical independent vars>;
   model <dependent var> = <independent vars>;
```

peter.langlois@dshs.state.tx.us

Phone: 512-458-7111 x6183

Thanks