New Guidelines for Data Result Suppression of Small Cell Numbers in IBIS-Query Output

Public Health Informatics Brown Bag

July 22, 2009

Kathryn Marti, Director

Office of Public Health Assessment

Outline

- Introduction
- Background
- Work Group
- Important Concepts
- Methodology
- The Guidelines
- Next steps

Introduction

- Why not report data?
  - Protect privacy
  - Ensure reliable data
- Why make data publicly available?
- What reports?
  - Published reports
  - Online static tables
  - Web-based data query system output

Background

- Current IBIS-Q suppression rules
- Enhancements to IBIS-Q BRFSS
- Increased likelihood of small cell sizes
- Grew to department-wide effort
- Develop UDOH-wide policy
- HIPAA and more
- Sensitivity for vulnerable populations

Data Suppression Decision Rules Work Group

- Charter
- Provide guidelines to UDOH programs
- Program into IBIS-Q
- Work Group Products:
  - Decision Rules
  - Decision Tree
  - Justification of the rules

Important Concepts

- Confidence Interval
- Coefficient of Variation
- Power to detect a difference
- Context/Use of the estimate

Coefficient of Variation

CV=[100 x (SE(R)/R)]%

- Where:
  - SE = the standard error of the estimate
  - R = the estimated rate
- Unitless measure
- Usually expressed as a percentage

Coefficient of Variation (cont')

- If the estimated percentage is greater than 50%, the CV is calculated by dividing the standard error of the estimate (SE(R)) by one minus the estimate (1-R).

CV=[100 x (SE(R)/(1-R))]%

Coefficient of Variation for Percentages

Example #1

The crude estimated percentage of Utah adults aged 85 and older who were current cigarette smokers in 2007 was 7.65% with a standard error of 3.81%.

CV = 100 x (0.0381/0.0765) = 49.8%

Coefficient of Variation for Percentages (cont')

Example #2

The crude estimated percentage of Utah adults aged 85 and older who were non-smokers in 2007 was 92.35% with a standard error of 3.81%.

$$CV = 100 \times (0.0381/(1-0.9235))$$

$$= 100 \times (0.0381/0.0765) = 49.8\%$$

Coefficient of Variation for Count Data

D = # of deaths ~ Poisson

P = # in population where deaths occurred

R = rate per 100,000 = (D/P) *100,000

Var(R) = Var [(D/P) x 100,000]

$= (100{,}000/P)^2 \text{ Var}(D)$

$= (100{,}000/P) \times ((100{,}000*D)/P)$

$= (100{,}000/P) \times R$

CV = 100 x [SQRT(Var(R))]/R

= 100 x [SQRT((100,000/P)*R)]/R

= 100 x SQRT[100,000/(P*R)]

Coefficient of Variation for Count Data (cont')

Example #1

D = 388 Alzheimer's disease deaths in Utah in 2006

P = UT Population in 2006 was 2,615,129

R = (388/2,615,129) x 100,000 = 14.84 per 100,000

CV = 100 x SQRT[100,000/(P*R)]

CV = 100 x SQRT[100,000/(2,615,129 x 14.84)]

= 5.08%

Coefficient of Variation for Count Data (cont')

Example #2

D = 5 Alzheimer's disease deaths in Utah adults between the ages 55-64 in 2006

P = 55-64 UT Population in 2006 was 201,340

R = (5/201,340) x 100,000 = 2.48 per 100,000

CV = 100 x SQRT[100,000/(P*R)]

CV = 100 x SQRT[100,000/(201,340 x 2.48)] = 44.75%

Methodology

- Literature review
- Compared suppression rules using BRFSS small area, low prevalence data.
  - < 5 observations in numerator or 30 in denominator
  - CV greater than 30%
  - CV greater than 50%
  - the CI length larger than the estimated percent
  - 1/2 the CI length >10% and denominator less than 50

Guidelines

Minimum Criteria

For Reporting Both Survey Data and Population Event Data:

- ✓ CV $\leq$ 50%
- ✓ If 30% < CV $\leq$ 50% an asterisk should be included with a footnote that says: *Use caution in interpreting, the estimate has a relative standard error greater than 30% and does not meet UDOH standards for reliability.

Guidelines

Strict Criteria

For Reporting Survey Data:

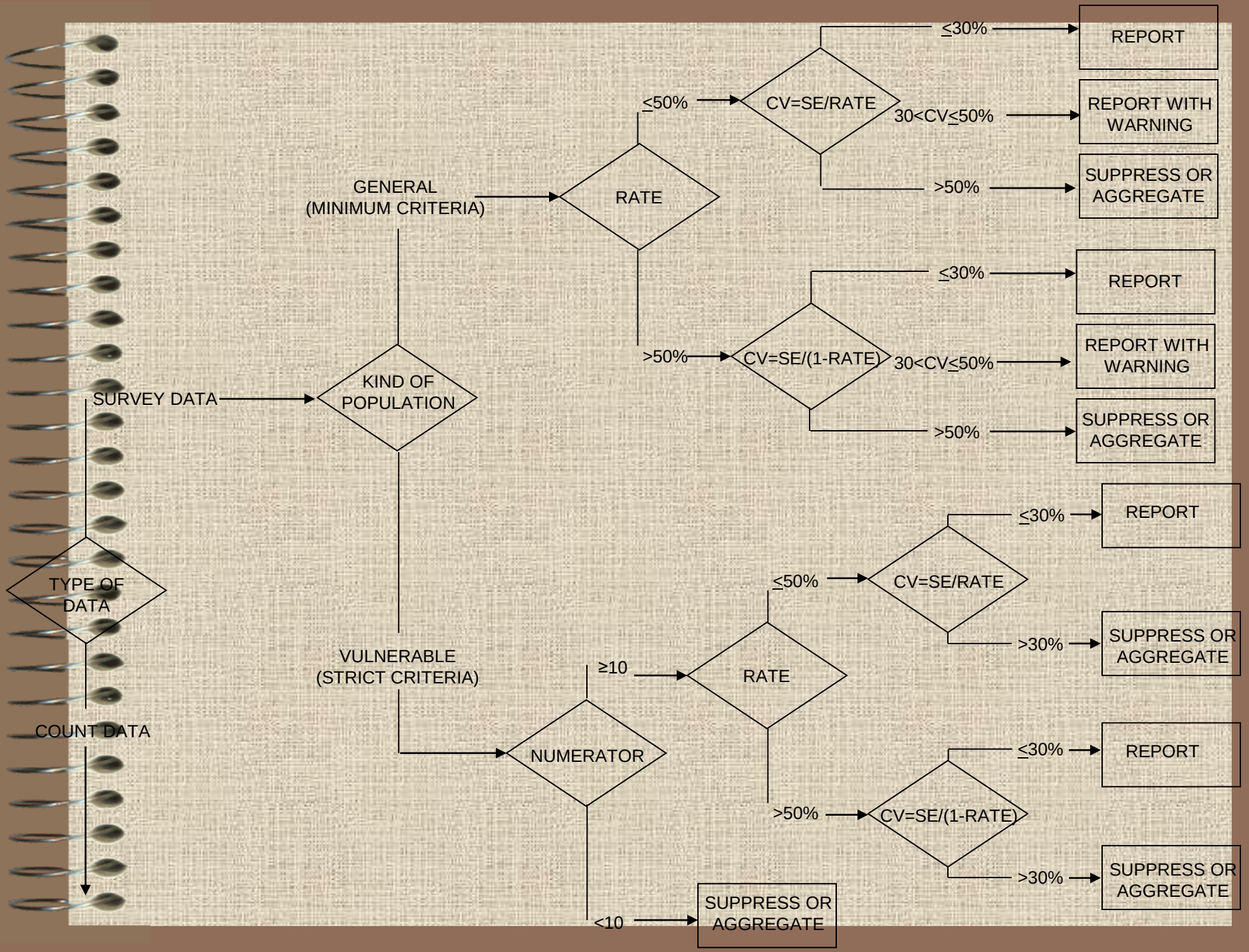
➓✓ ≥ 10 cases in the numerator

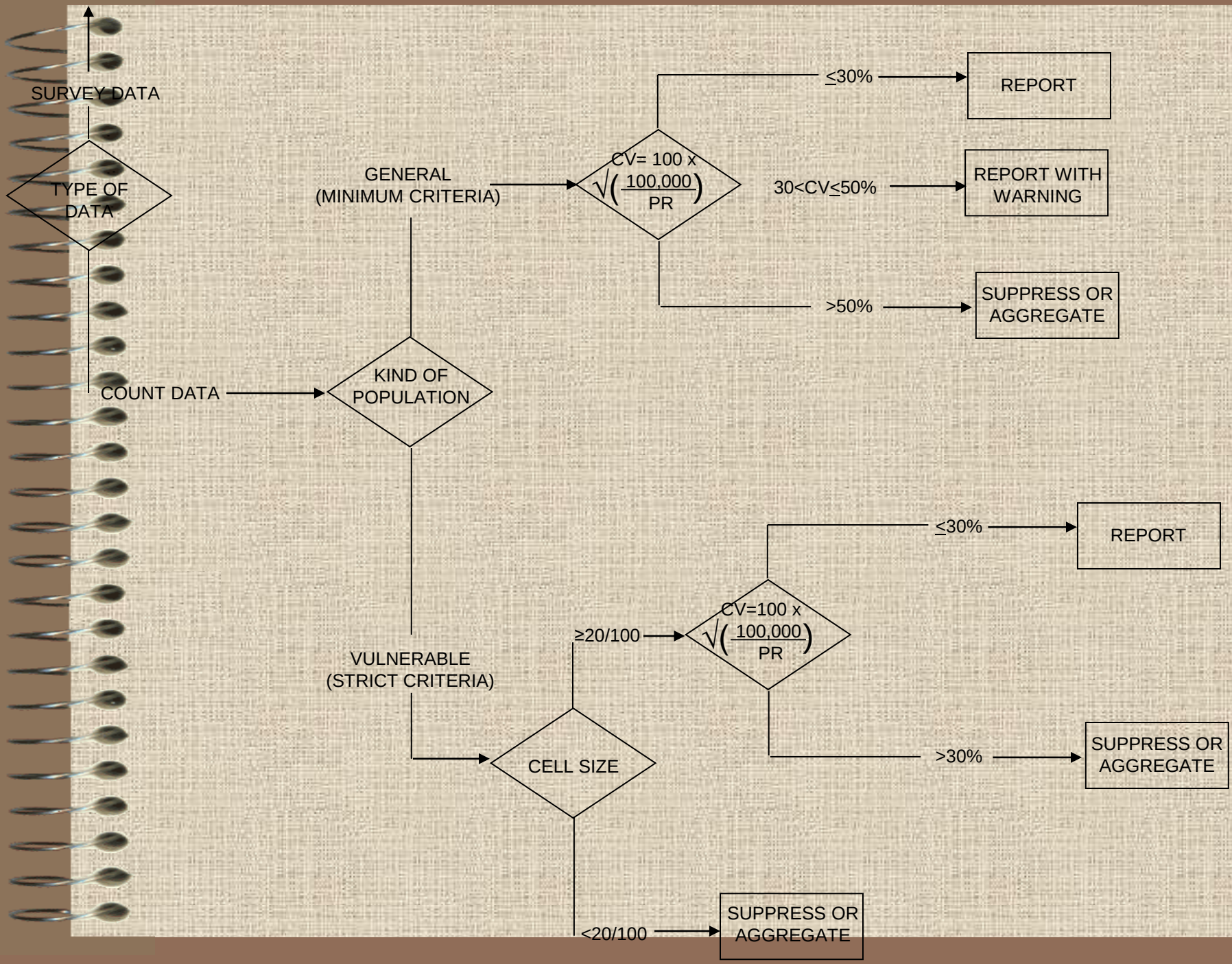
➓✓ AND a CV ≤ 30%

For Reporting Population Event Data:

➓✓ ≥ 20 cases in the numerator and ≥100 persons in the population

- AND a CV ≤ 30%


SURVEY DATA
TYPE OF DATA
COUNT DATA
KIND OF POPULATION
GENERAL (MINIMUM CRITERIA)
RATE
≤50%
CV=SE/RATE
≤30%
REPORT
30<CV≤50%
REPORT WITH WARNING
>50%
SUPPRESS OR AGGREGATE
>50%
CV=SE/(1-RATE)
≤30%
REPORT
30<CV≤50%
REPORT WITH WARNING
>50%
SUPPRESS OR AGGREGATE
VULNERABLE (STRICT CRITERIA)
NUMERATOR
≥10
RATE
≤50%
CV=SE/RATE
≤30%
REPORT
>30%
SUPPRESS OR AGGREGATE
>50%
CV=SE/(1-RATE)
≤30%
REPORT
>30%
SUPPRESS OR AGGREGATE
<10
SUPPRESS OR AGGREGATE


SURVEY DATA
TYPE OF DATA
GENERAL (MINIMUM CRITERIA)
CV= 100 x √(100,000/PR)
≤30%
REPORT
30<CV≤50%
REPORT WITH WARNING
>50%
SUPPRESS OR AGGREGATE
COUNT DATA
KIND OF POPULATION
VULNERABLE (STRICT CRITERIA)
CELL SIZE
≥20/100
CV=100 x √(100,000/PR)
≤30%
REPORT
>30%
SUPPRESS OR AGGREGATE
<20/100
SUPPRESS OR AGGREGATE

Next Steps

- Program into IBIS-Q and SAS CGI
- Determine how to display output and notes
  - New Mexico example
  - Utah Test server
- Get OK from Data Stewards
  - Minimum criteria vs. Strict criteria

Next Steps (con't)

Suggested text for 30% < CV ≤ 50% :

- *Use caution in interpreting, the estimate has a relative standard error greater than 30% and does not meet UDOH standards for reliability.

DIFY Query / Return
ueryBuilder
ct MEASURE / Go
election Page
DEFAULT Query
ctions
View
ort Results to
EL
ct GRAPH Display

Query Criteria for the Infant Sex Ratio Measure

Year Filter:	2004, 2005, 2006
Data Grouped By:	Mother's County of Residence
Data Chart:	None

Data List

Mother's County of Residence	Males	Females	Infant Sex Ratio (M/F)	95%CI -LL-	95%CI -UL-	Data Reliability
Total	44,491	42,601	1.04	1.03	1.06	
Bernalillo	14,289	13,695	1.04	1.02	1.07	
Catron	37	39	0.95	0.5	1.4	
Chaves	1,535	1,435	1.07	1	1.14	
Cibola	663	642	1.03	0.92	1.14	
Colfax	235	221	1.06	0.88	1.25	
Curry	1,457	1,307	1.11	1.04	1.19	
De Baca	35	20	1.75	1.2	2.3	
Dona Ana	5,144	4,887	1.05	1.01	1.09	
Eddy	1,101	1,046	1.05	0.97	1.14	
Grant	579	537	1.08	0.96	1.2	
Guadalupe	74	72	1.03	0.7	1.35	
Harding	7	6	1.17	0.08	2.26	† Warning
Hidalgo	98	90	1.09	0.8	1.38	
Lea	1,579	1,518	1.04	0.97	1.11	

IFY Query / Return
ueryBuilder
ct MEASURE / Go
election Page
DEFAULT Query
ctions
View
ort Results to
EL
ct GRAPH Display

Santa Fe	2,514	2,425	1.04	0.98	1.09	
Sierra	170	162	1.05	0.83	1.26	
Socorro	412	390	1.06	0.92	1.19	
Taos	536	559	0.96	0.84	1.08	
Torrance	239	302	0.79	0.62	0.96	
Union	75	54	1.39	1.04	1.74	
Valencia	1,459	1,326	1.1	1.03	1.17	

Record Count: 34

† This count or rate is not reliable (coefficient of variation >.30). Please interpret with caution.
‡ This count or rate is extremely unreliable (coefficient of variation >.50). You should combine years or otherwise increase your population size. More on data reliability.

Data Sources

New Mexico Birth Certificate Database, Bureau of Vital Records and Health Statistics, New Mexico Department of Health

Data Issues

Birth records are filed electronically by hospitals. Medical records staff use standard mother and facility worksheets and medical charts to complete the birth registration. Hospital training is provided by the Bureau of Vital Records and Health Statistics (BVRHS) and is based on the Vital Statistics Act and Regulations, BVRHS documentation, and handbooks produced by CDC's National Center for Health Statistics (NCHS). The electronic birth registration system has online edits and records are reviewed by BVRHS. Additionally, NCHS provides feedback to BVRHS on data quality. The state also provides feedback to the hospitals to improve data quality and contacts the hospital staff for clarification of missing, inconsistent or incorrect entries.

Time of Query

These data were queried on: Tue, 21 Jul 2009 13:44:52 MDT

number in a cell that has been suppressed.

y Definition
ult Query Result
y Dataset
iguration Selection
hical Charts
r Measures
ut to Excel
View

Data List

Local Health District	General Health Status	Number of Responses	General Health Status	Confidence Bounds - Lower Limit	Confidence Bounds - Upper Limit	Coefficient of Variation (%)
Bear River LHD	Excellent / Very Good / Good	26	100.0%	.	.	.
Bear River LHD	Fair / Poor	**	**	.	.	.
Bear River LHD	Total	**	**	.	.	.
Central Utah LHD	Excellent / Very Good / Good	24	98.31%	88.67%	99.77%	100.48
Central Utah LHD	Fair / Poor	**	**	0.23%	11.33%	100.48
Central Utah LHD	Total	**	**	.	.	.
Davis County LHD	Excellent / Very Good / Good	44	94.03%	82.13%	98.18%	59.06
Davis County LHD	Fair / Poor	**	**	1.82%	17.87%	59.06
Davis County LHD	Total	**	**	.	.	.
Salt Lake Valley LHD	Excellent / Very Good /	153	93.64%	87.76%	96.8%	34.40

y Definition
ult Query Result
y Dataset iguration Selection
hical Charts
r Measures
ut to Excel
View

Wasatch County LHD	Fair / Poor	**	**	.	.	.
Wasatch County LHD	Total	**	**	.	.	.
Weber-Morgan LHD	Excellent / Very Good / Good	41	97.18%	82.48%	99.6%	98.60
Weber-Morgan LHD	Fair / Poor	**	**	0.4%	17.52%	98.60
Weber-Morgan LHD	Total	**	**	.	.	.

97.1753

Record Count: 36

***The count or rate in certain cells of the table has been suppressed either because 1) the observed number of events is very small and not appropriate for publication, or 2) it could be used to calculate the number in a cell that has been suppressed.*

Data Notes
Would you say that in general your health is Excellent, Very Good, Good, Fair or Poor?

Data Sources
Utah Behavioral Risk Factor Surveillance System, Office of Public Health Assessment, Utah Department of Health

Data Issues
Denominator includes all survey respondents ages 18 years and older except those with missing, don't know and refused answers. If the query was limited to a particular sub-population-group, only those respondents are included in the denominator.

The confidence bounds are asymmetric.

Time of Query
These data were queried on: Tue, 21 Jul 2009 14:20:08 MDT

Top