Comparing Distributions Across Categories

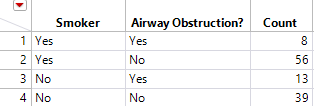
Example 7.1 Investigating Statistical Concepts, Applications, and Methods: Investigation 1.1.1). In this study researchers began conducting medical examinations and environmental surveys of workers employed at a microwave popcorn production plant. As part of this study, current employees at the plant underwent spirometric testing which measures FVC (forced vital capacity) which is the volume of air that can be maximally forcefully exhaled. There was a total of 116 employees who were underwent this testing. On this test, 31 employees had abnormal results, including 21 with airway obstruction.



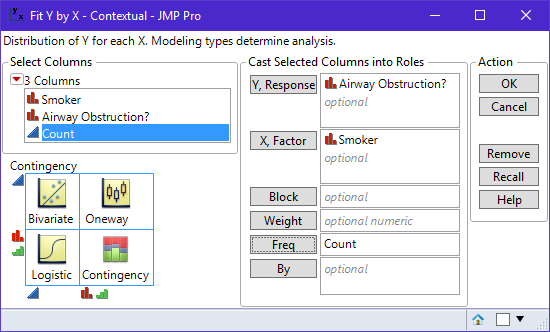
Smokers and Non-smokers tend to have different FVC measurements as smoking is known to reduce lung volume. Consider the following breakdown of smokers and non-smokers from this study.

|  |  |  |  |
| --- | --- | --- | --- |
| Smokers vs Nonsmokers | Number with Airway Obstruction | Number without Airway Obstruction | Total |
| Smokers | 8 | 56 | 64 |
| Non-Smokers | 13 | 39 | 52 |
| Total | 21 | 95 | 116 |

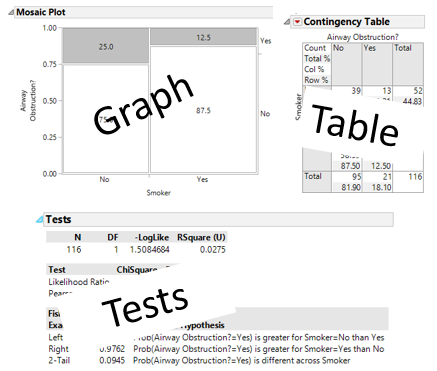
Getting this data in JMP



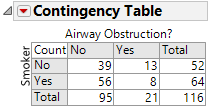
Getting the graphical and cross-tab summaries in JMP. Select Analyze > Fit Y by X.



The following output is returned and is divided into three pieces (1) Graph, (2) Table, and (3) Tests.



Making Comparisons Through Conditioning

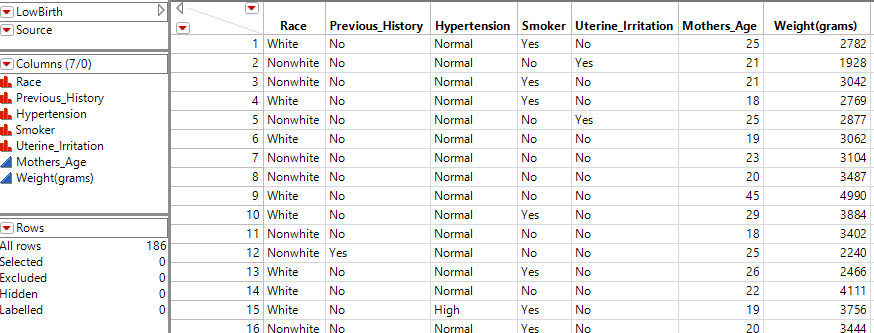


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Counts | Row Percentages | Graphs |
| Conditioning on | Smoker = No |  |  |  |
| Smoker = Yes |  |  |  |

Interpret the following output

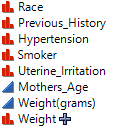
|  |  |
| --- | --- |
|  |  |

Example 7.2 Consider the following study of risk factors and their relationship to whether or not a mother is likely to have a low birth weight baby.



|  |  |
| --- | --- |
| Making categories for weight | New Weight Column added to dataset |

Consider the variable type in JMP. The designation of a red bar graph indicates a categorical variable.



Looking at all risk factors

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

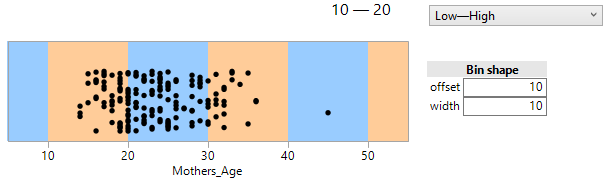
Question

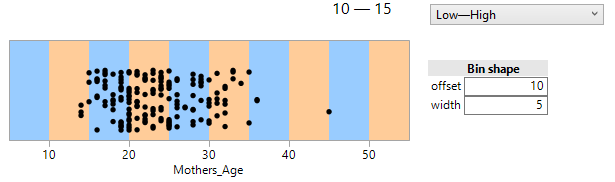
1. What is the most important risk factor? How did you make this determination?

Looking at the effect of Mothers Age

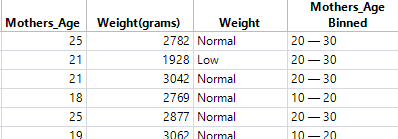
|  |  |
| --- | --- |
|  |  |

Creating Bins in JMP

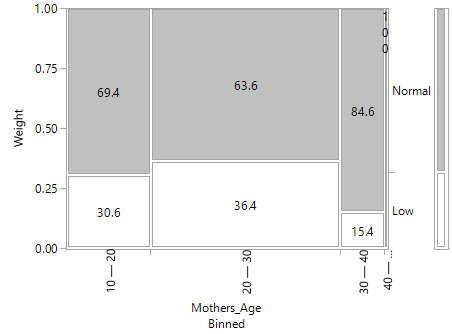




The desired output…



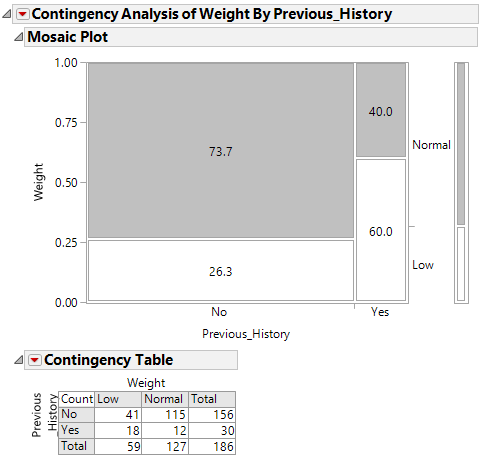
The mosaic plot for Mothers Age

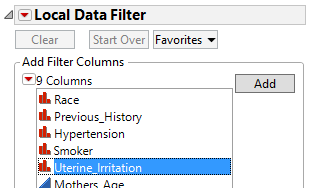
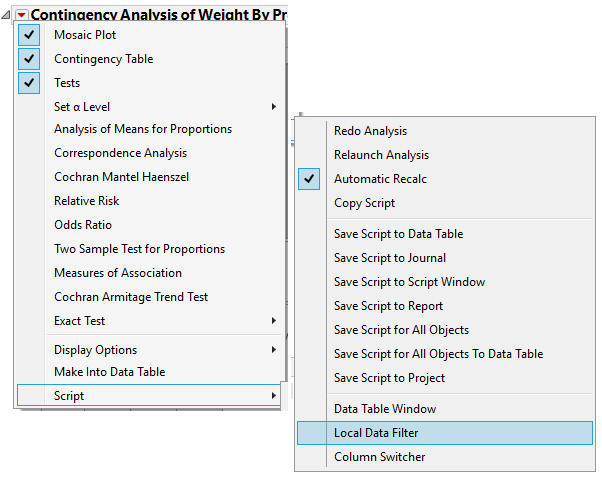


Question

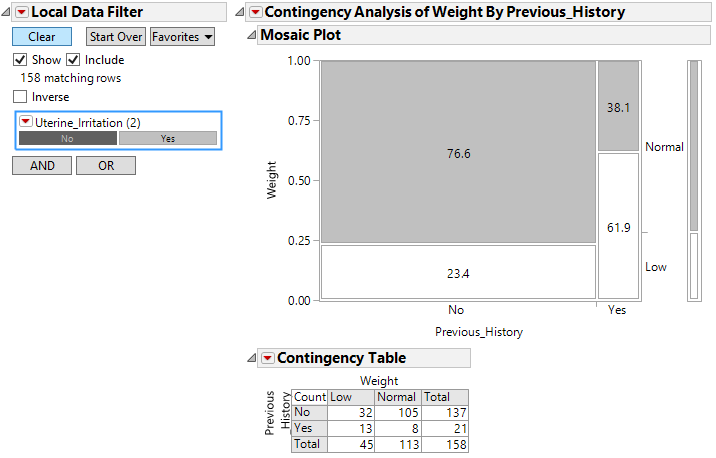
1. What is the most effect of Mothers Age on risk of low birth weight? Discuss.

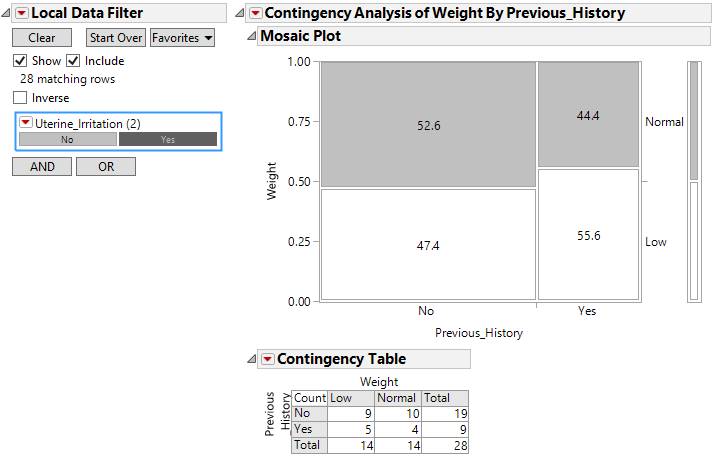
**Using Local Data Filter to dig deeper**





**Understanding Previous History vs Low Birth Weight (for each Uterine Irritation level)**





**Understanding Previous History vs Low Birth Weight (for each Smoking level)**

|  |  |
| --- | --- |
|  |  |

**Understanding Previous History vs Low Birth Weight (for each Mothers Age level)**

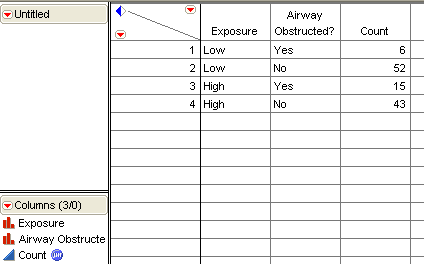
|  |  |
| --- | --- |
|  |  |
|  |  |

Summary Measures for Tables

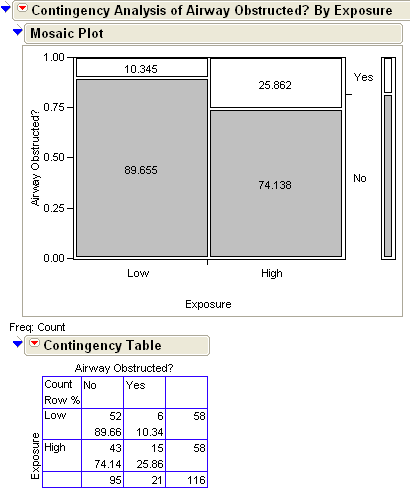
Example 7.4 In addition to other things done in this study, the popcorn plant was broken into several areas (the flavor-mixing room, packaging room, etc.). Air and dust samples in each area were measured to determine the exposure to diacetyl, a marker of organic-chemical exposure. Then, the average exposure for each study participant was determined by taking into account how long they spent at different jobs within the plant and the average exposure in that job area. Finally, they were classified as having either “low” or “high” exposure.

|  |  |  |  |
| --- | --- | --- | --- |
| Exposure Level | Number with Airway Obstruction | Number without Airway Obstruction | Total |
| Low | 6 | 52 | 58 |
| High | 15 | 43 | 58 |
| Total | 21 | 95 | 116 |

Putting this data into JMP



The output from JMP



Relative risk ratios requires the use of conditional probabilities which are simply probabilities or percentage that are computed based on a particular row or columns. Consider the following conditional probabilities.

P(Airway Obstruction = Yes | Exposure = Low) = 

and

P(Airway Obstruction = Yes | Exposure = High) = 

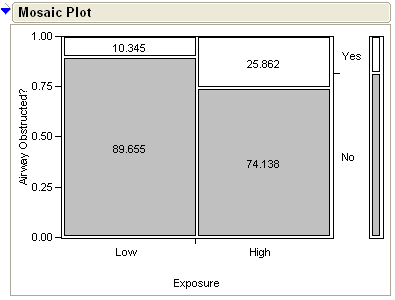
**Relative Risk Ratio**  is method for making comparisons between Low and High exposure and is computed as follows.



Comment: Relative Risk Ratios is usually computed so that they are bigger than one. Realize, we could have computed the relative risk ratio as



Sketch the interpretation of relative risk on the plot below.



Interpret the relative risk computed above.

Realize, a relative risk ratio of 1.0 is always your reference value. What does a relative risk of 1 imply? Sketch two examples with Relative Risk = 1.0.

|  |  |
| --- | --- |
|  |  |

**Comments**:

Sometimes relative risk can be a bit misleading. For example, suppose a relative risk ratio is computed as



In this case, the relative risk ratio is large, but the probabilities themselves are insignificant. Compare this to

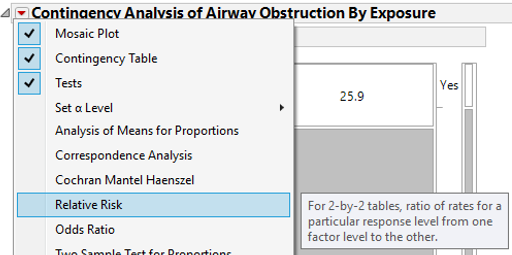


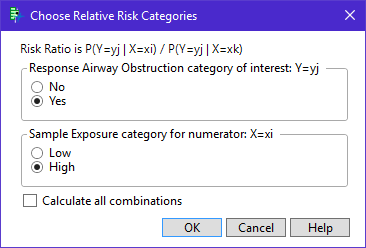
Note that the relative risk is the same as above, but in this case, the observable difference in the probabilities is much greater. A way around this problem is to compute the risk difference.

Risk Difference = P(Air Obs=Yes | Exp=High) – P(Air Obs=Yes | Exp=Low)

=

**Getting Relative Risk Ratio in JMP**





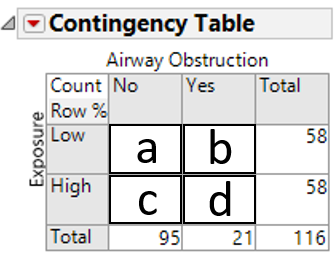


If you’re not sure which ones to select, ask for them all. However, only one of these relative risk ratios is the most appropriate for any given problem.

|  |  |
| --- | --- |
|  |  |

**Confidence Intervals for Relative Risk Ratio**

Suppose our table has the following general structure.



If the relative risk is computed as follows



Then the confidence interval is computed as follows and uses the natural log of the relative risk and is given here.





The final step is put this interval back on the original scale so interpretations can be done. Thus, a 95% confidence interval goes from  up to .

JMP computes this interval for us.



Sketch and interpret the 95% confidence interval for the relative risk difference here.

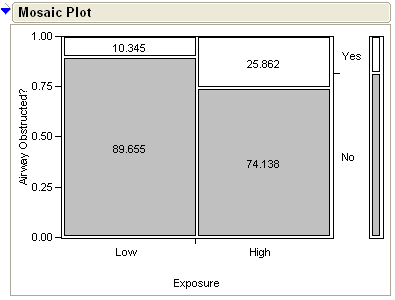
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Relative Risk Differences

As discussed above, the actual difference in risk may be more informative.



Sketch the interpretation of relative risk difference on the plot below.



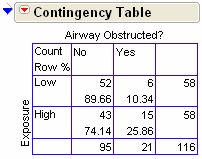
Interpret the relative risk computed above.

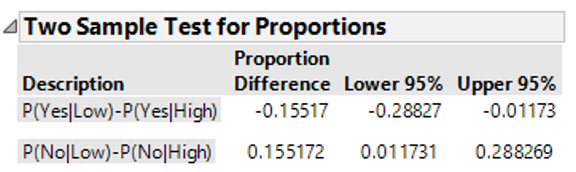
**Getting Relative Risk Difference in JMP**

|  |  |
| --- | --- |
|  |  |

**Confidence Intervals for Relative Risk Difference**

What are the other values given (i.e. Lower 95% and Upper 95%) for?





The formulas for the confidence intervals…



The missing interval multiplier comes from the normal distribution as is given here.

|  |  |  |  |
| --- | --- | --- | --- |
| Type of Interval | 90% | 95% | 99% |
| One-Tailed | 1.28 | 1.645 | 2.32 |
| Two-Tailed | 1.645 | 1.96 | 2.57 |

Plug in the numbers for our example to verify the confidence interval calculations…





Sketch the 95% confidence interval for the relative risk difference here.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Interpret this interval.

Comments:

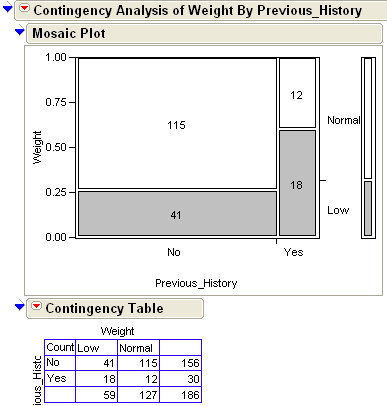
1. Zero is the reference value for this confidence interval. Why? What would it mean if the relative risk difference was zero?
2. This interval is completely positive, what does this mean?
3. What would an interval that is completely negative mean?

**Odds Ratios**

Another concept used to quantify the differences between two categorical variables is Odds Ratios. This are similar in concept to Relative Risk ratio, but can be applied more generally.

Example 7.5 Reconsider the study of risk factors and their relationship to whether or not a mother is likely to have a low birth weight baby.

The following displays the relationship between Previous History of Low Birth weight and current weight.

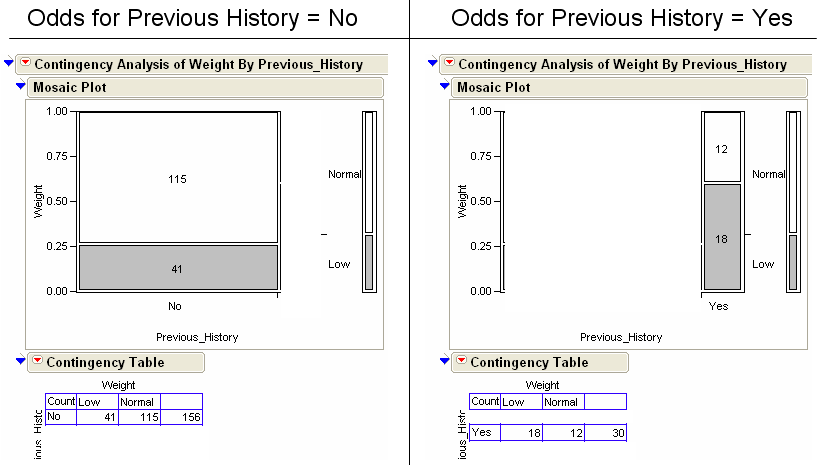


First, the odds for each group separately,





Visualization of each…



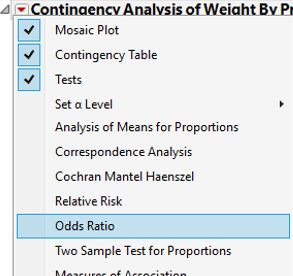
For comparisons, we compute the ratio



Comments:

1. An Odds Ratio of 1.0 is again our reference value. What does an Odds Ratio of 1 mean?
2. Again, often Odds Ratios are computed so that they are greater than 1.0. This is just for convenience and does not change our interpretation.

**Getting these in JMP**



The following output is obtained…

|  |  |
| --- | --- |
|  | Changing the Value Orderings |

**Confidence Intervals for Odds Ratios**



The confidence interval uses the natural log of the odds ratio and is given here.





The missing interval multiplier comes from the normal distribution as is given here.

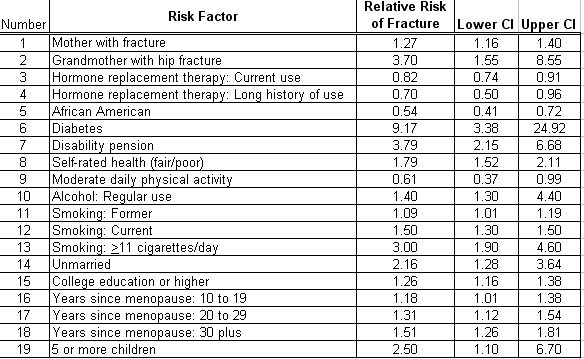
|  |  |  |  |
| --- | --- | --- | --- |
| Type of Interval | 90% | 95% | 99% |
| One-Tailed | 1.28 | 1.645 | 2.32 |
| Two-Tailed | 1.645 | 1.96 | 2.57 |

Once again, the final step is put this interval back on the original scale so interpretations can be done. The 95% confidence interval goes from  to .

Sketch the confidence interval for Example 3.8 below.

Interpret this confidence interval in the context of this problem.

Example 7.6 Consider the following study in which various risk factors were being considered as a method for screening for postmenopausal osteoporosis.  
  
Source: <http://www.ahrq.gov/clinic/3rduspstf/osteoporosis/osteosumm1.htm>  
  
The risk factors under consideration have been numbered from 1 – 19 in the following table.



As an example, consider the 1st Risk Fracture (Mother with Fracture). The reported relative risk was computed as follows.



Answer the following

1. Consider the relative risk for Diabetes at 9.17. Using everyday language, explain what this value means?
2. Consider the relative risk for being an African American at 0.54. Using everyday language, explain what this value means?
3. Look at Risk Factors #3 and #4. What can be said about the long term use of hormone replacement therapy in relation to bone fractures in females? Discuss.
4. Look at Risk Factors #11, #12, and #13. What can be said about the effects of smoking in relation to bone fractures in females? Discuss.
5. The risk factors listed above were those found to be statistically important. In particular, notice that none of the confidence intervals capture 1.0. Why is it the case that none of the confidence intervals contain 1.0? What would it mean if the confidence interval did contain 1.0? Explain.

Example 7.7 Consider the following data from the MN Department of Corrections web site. The investigation here is centered around whether or not sexual treatment programs work. Consider the following statement in their report.

“To evaluate the effectiveness of sex offender treatment programming, the DOC (Department of Corrections) examined the recidivism outcomes among 2,040 sex offenders released from prison between 1990 and 2003. Recidivism data were collected on 2,040 offenders through 2006. ….. Untreated and treated offenders were matched on commonly known risk factors, and multivariate statistical analyses were performed to control for other factors besides the treatment that may have an impact on recidivism. These measures were used to ensure that ‘apples were compared to apples’.”

Source: “The Impact of Prison-Based Treatment on Sex Offender Recidivism: Evidence from Minnesota”, *Research in Brief*, Minnesota Department of Corrections, March 2010.

The following is some of the data provided in their report.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sexual Treatment Program | Number Not  Rearrested | Number Rearrested  With Reason | | | Total |
| Sexual | Violent | Other |
| Yes | 443 | 145 | 314 | 118 | 1020 |
| No | 427 | 199 | 348 | 46 | 1020 |

|  |  |
| --- | --- |
| Note: Change value ordering on response to flip odds ratio so that it is bigger than 1 |  |

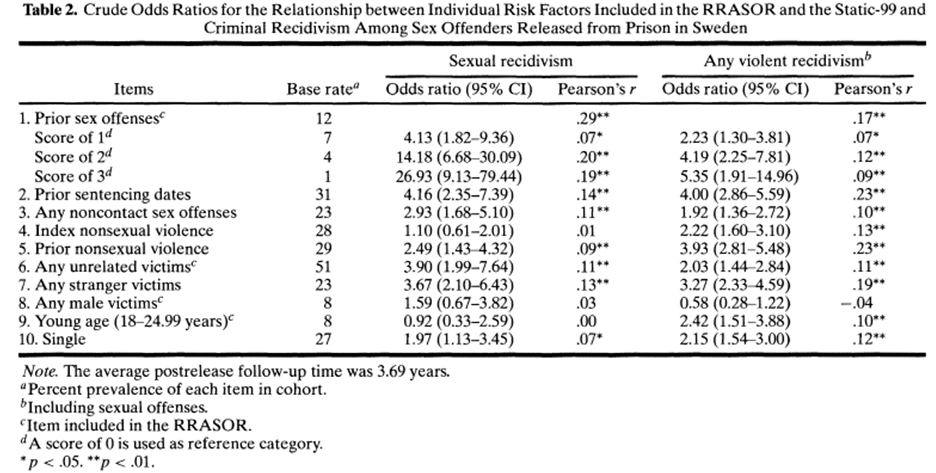
Questions

1. Compute the odds of a rearrest (for any reason) for a person not in a sexual treatment program.
2. Compute the odds of a rearrest (for any reason) for a person who was in a sexual treatment program.
3. Compute the appropriate odds ratio that would allow us to compare the odds of rearrest for

those that did not go through a sexual treatment program to those that did. What are the practical implications of this value? Discuss.

1. Recompute the odds of a rearrest for only sexual crimes for both those that completed the sexual treatment program and those that did not. Compute the appropriate odds ratio to measure the effect of the sexual treatment program. How is this odds ratio different than the one computed above. Discuss.

Example 7.8 Consider the following table of odd ratios for similar data from a study on sexual recidivism on individuals released from prison in Sweden



*Source*: Sjöstedt, G. and Långström, N. (2001) “Actuarial Assessment of Sex Offender Recidivism Risk: A Cross-Validation of the RRASOR and the Static-99 in Sweden.” *Law and Human Behavior*, Vol. 25, No. 6, pp. 629-645

Questions

1. What are the most important factors that influence the sexual recidivism in this study? What are the least?
2. The asterisk denotes the statistical significance of each item. Notice, that for each item that lacks statistical significance (i.e. does not have an asterisk), the 95% confidence interval contain 1. This should be the case. Why?