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.

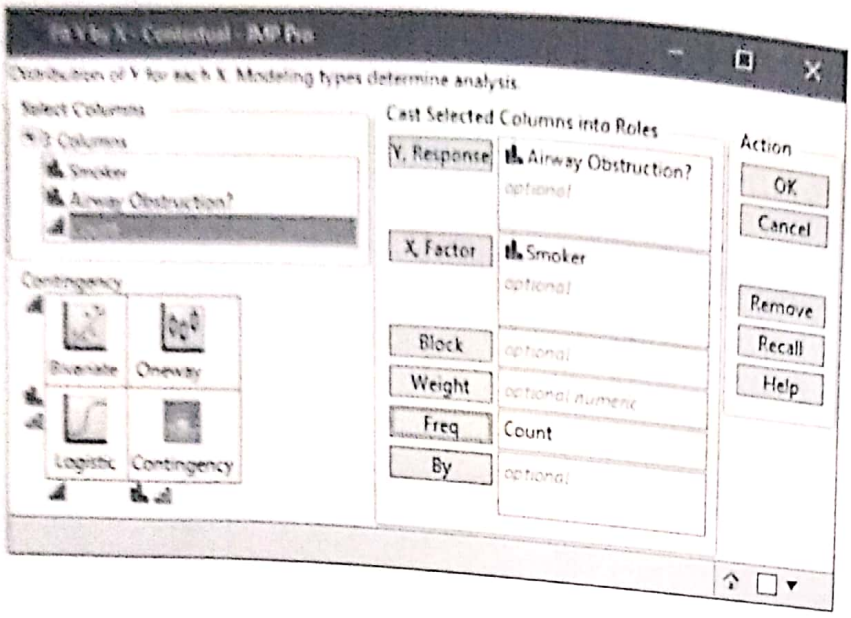
Smokers and Non-smokers tend to have different FVC measurements as smoking is known to reduce lung volume. Consider the following breakdown of smokers and non-smokers from this study.

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

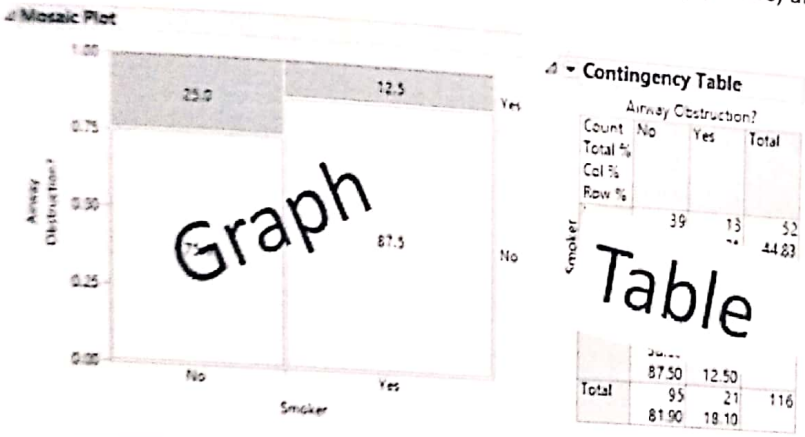
Getting this data in JMP

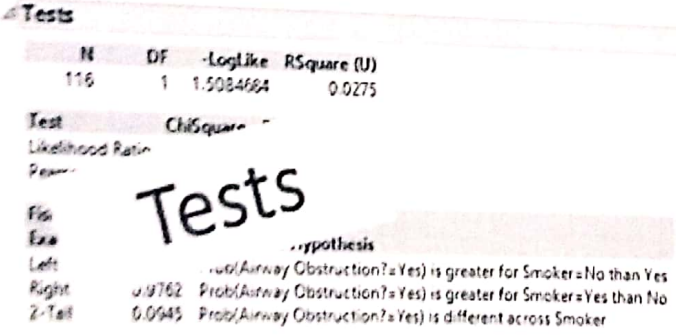
	Smoker	Airway Obstruction?	Count
1	Yes	Yes	8
2	Yes	No	56
3	No	Yes	13
4	No	No	39

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

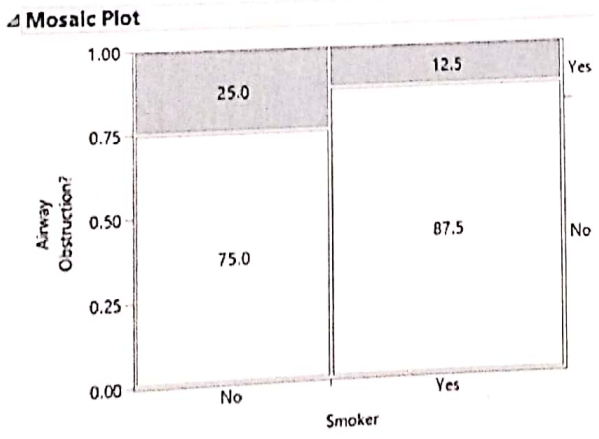
Contingency Table

Airway Obstruction?

Smoker	Count	No	Yes	Total
	No	39	13	52
	Yes	56	8	64
	Total	95	21	116

Conditioning on		Counts	Row Percentages	Graphs
	Smoker = No	Airway Obstruction? / Smoker: Count, No, Yes, Total; No: 39, 13, 52	Airway Obstruction? / Smoker: Count, Row %, No, Yes, Total; No: 39, 13, 52; 75.00, 25.00	Mosaic Plot
	Smoker = Yes	Airway Obstruction? / Smoker: Count, No, Yes, Total; Yes: 56, 8, 64	Airway Obstruction? / Smoker: Count, Row %, No, Yes, Total; Yes: 56, 8, 64; 87.50, 12.50	Mosaic Plot

Interpret the following output

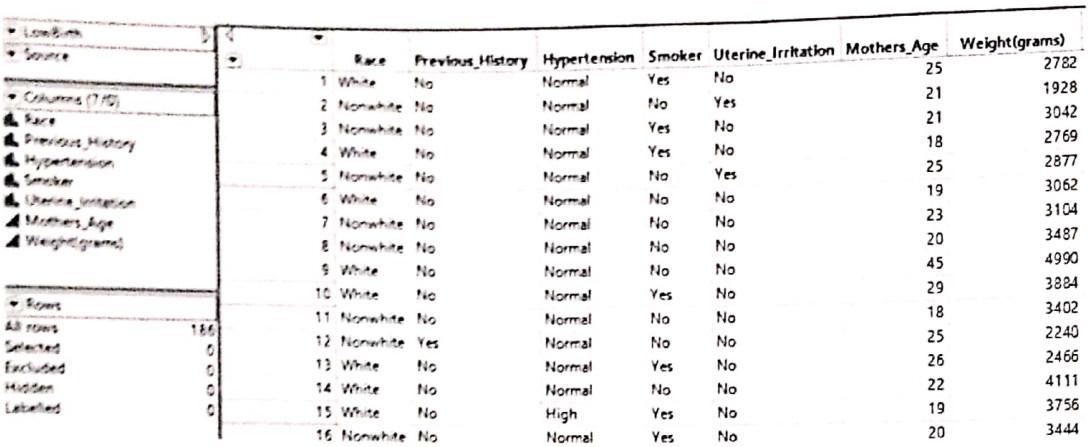


Contingency Table

Airway Obstruction?

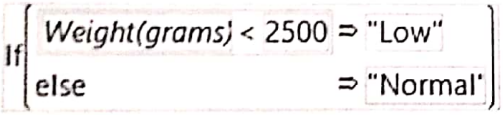
Smoker	Count / Row %	No	Yes	Total
	No	39 75.00	13 25.00	52
	Yes	56 87.50	8 12.50	64
	Total	95	21	116

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.



	Race	Previous_History	Hypertension	Smoker	Uterine_Irritation	Mothers_Age	Weight(grams)
1	White	No	Normal	Yes	No	25	2782
2	Nonwhite	No	Normal	No	Yes	21	1928
3	Nonwhite	No	Normal	Yes	No	21	3042
4	White	No	Normal	Yes	No	18	2769
5	Nonwhite	No	Normal	No	Yes	25	2877
6	White	No	Normal	No	No	19	3062
7	Nonwhite	No	Normal	No	No	23	3104
8	Nonwhite	No	Normal	No	No	20	3487
9	White	No	Normal	No	No	45	4990
10	White	No	Normal	Yes	No	29	3884
11	Nonwhite	No	Normal	No	No	18	3402
12	Nonwhite	Yes	Normal	No	No	25	2240
13	White	No	Normal	Yes	No	26	2466
14	White	No	Normal	No	No	22	4111
15	White	No	High	Yes	No	19	3756
16	Nonwhite	No	Normal	Yes	No	20	3444

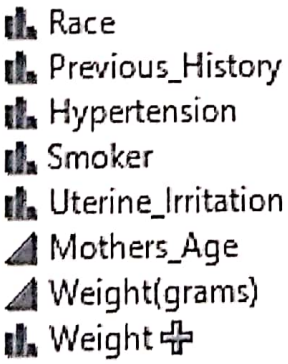
Making categories for weight



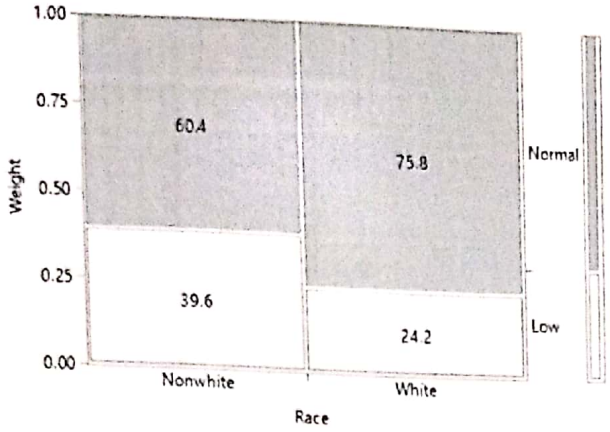
New Weight Column added to dataset

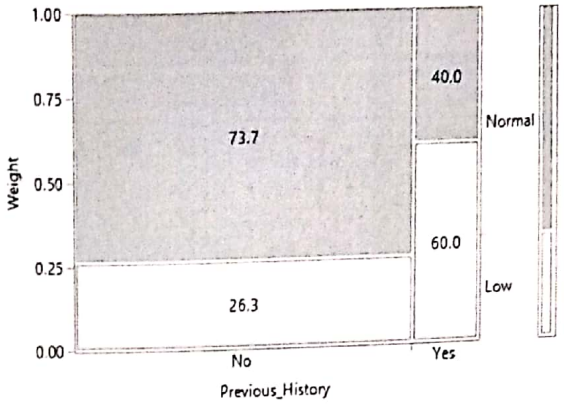
ıe	Weight(grams)	Weight
25	2782	Normal
21	1928	Low
21	3042	Normal
18	2769	Normal
25	2877	Normal
19	3062	Normal
23	3104	Normal
20	3487	Normal
15	4990	Normal
29	3884	Normal
18	3402	Normal

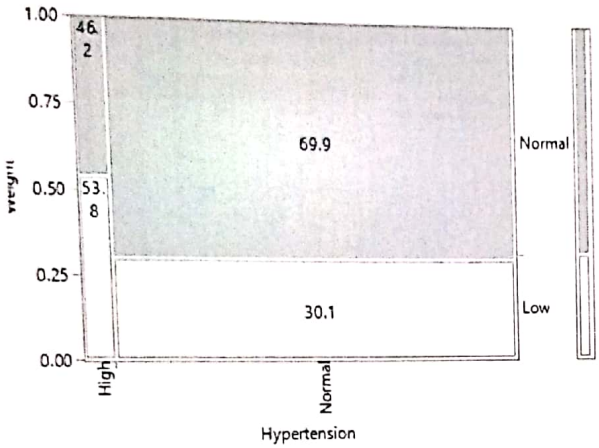
Consider the variable type in JMP. The designation of a red bar graph indicates a categorical variable.

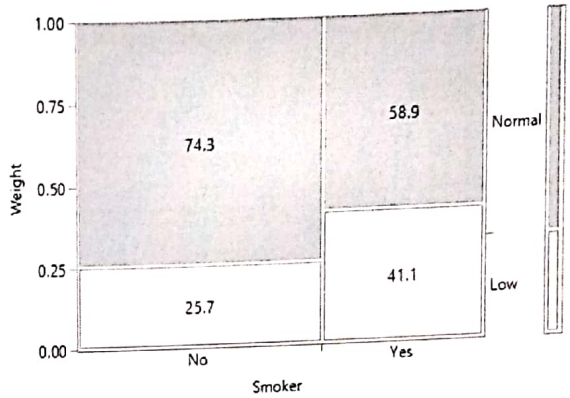


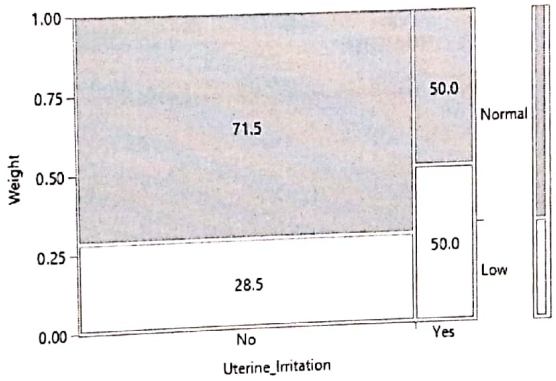
Looking at all risk factors











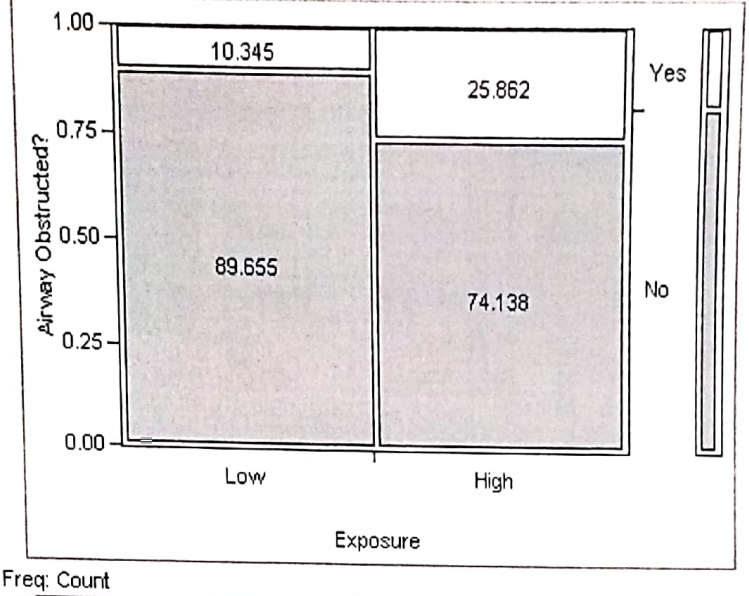
Question

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

The output from JMP

Contingency Analysis of Airway Obstructed? By Exposure

Mosaic Plot



Freq: Count

Contingency Table

Airway Obstructed?

Exposure	Count / Row %	No	Yes	
	Low	52	6	58
		89.66	10.34	
	High	43	15	58
		74.14	25.86	
		95	21	116

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) = .10345

and

P(Airway Obstruction = Yes | Exposure = High) = .25862

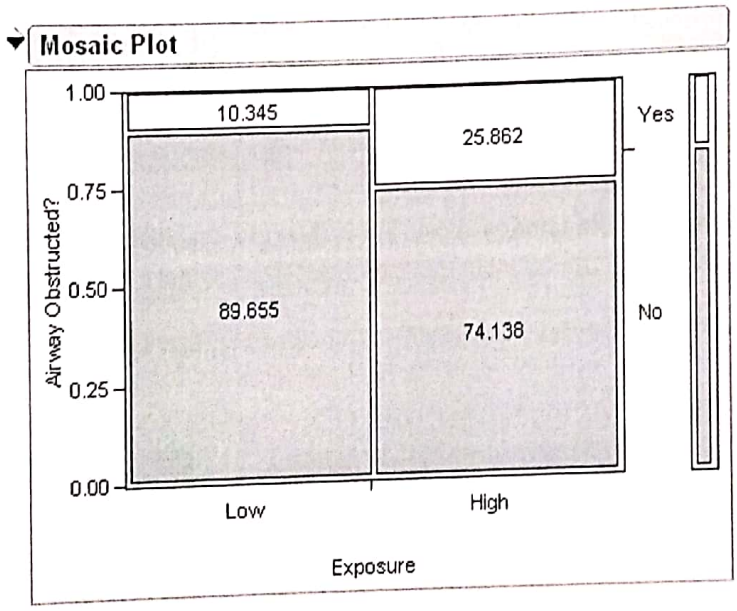
Relative Risk Ratio is method for making comparisons between Low and High exposure and is computed as follows.

$$\text{Relative Risk} = \frac{P(\text{Airway Obstruction} = \text{Yes} \mid \text{Exposure} = \text{Low})}{P(\text{Airway Obstruction} = \text{Yes} \mid \text{Exposure} = \text{High})} = \frac{.10345}{.25862}$$

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

$$\text{Relative Risk} = \frac{P(\text{Airway Obstruction} = \text{Yes} \mid \text{Exposure} = \text{High})}{P(\text{Airway Obstruction} = \text{Yes} \mid \text{Exposure} = \text{Low})} = \frac{.25862}{.10345}$$

Sketch the interpretation of relative risk on the plot below.



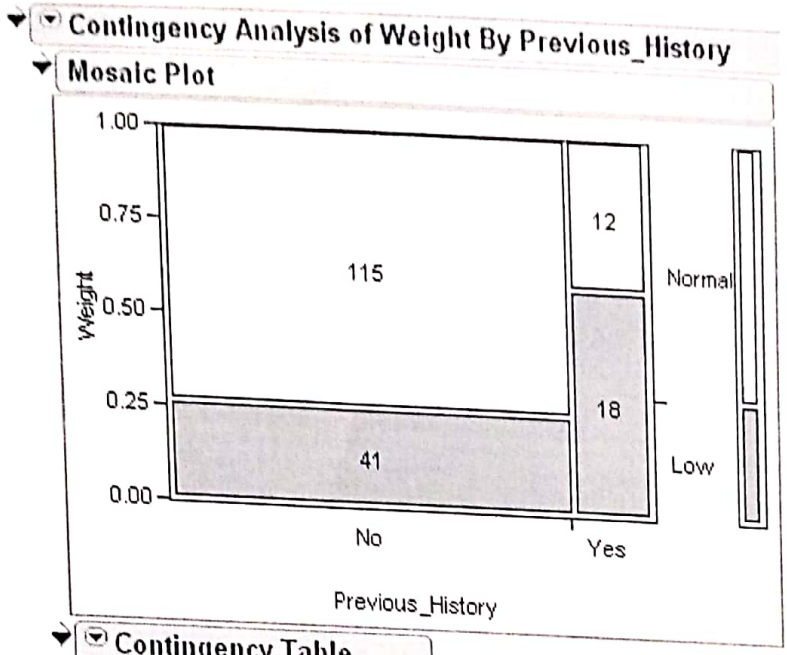
Interpret the relative risk computed above.

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.



Contingency Table

	Weight		
Count	Low	Normal	
No	41	115	156
Yes	18	12	30
	59	127	186

(Previous_History)

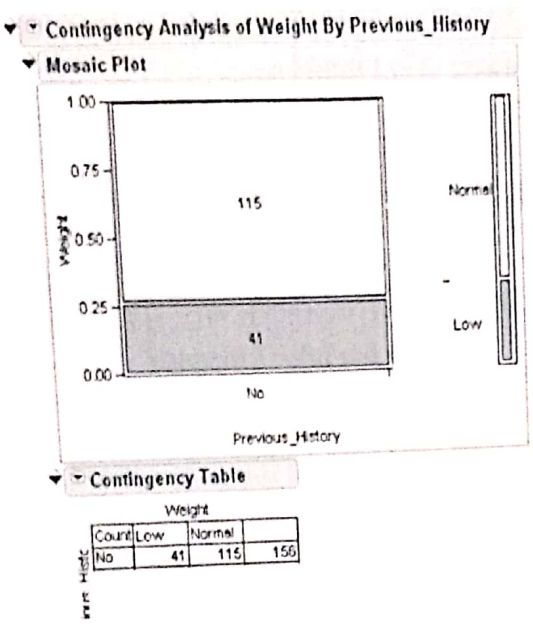
First, the odds for each group separately,

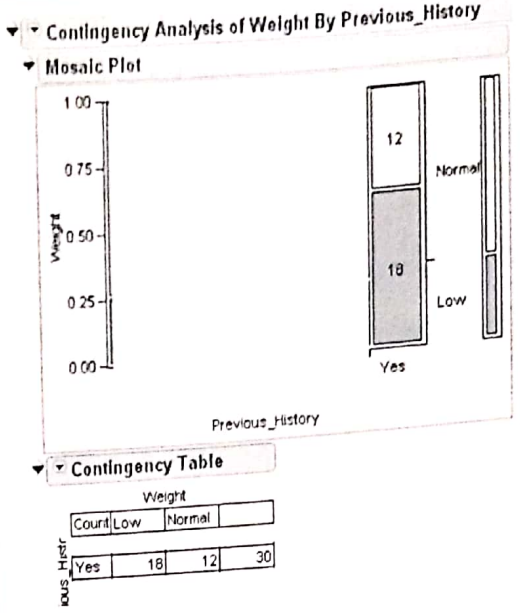
$$\text{Odds of Low Birth for (Prev Hist = No)} = \frac{\text{Number with (Weight = Low)}}{\text{Number with (Weight = Normal)}} = \frac{41}{115}$$

$$\text{Odds of Low Birth for (Prev Hist = Yes)} = \frac{\text{Number with (Weight = Low)}}{\text{Number with (Weight = Normal)}} = \frac{18}{12}$$

Visualization of each...

Odds for Previous History = No	Odds for Previous History = Yes





For comparisons, we compute the ratio

$$\text{Odds of Low Birth} = \frac{\text{Odds for (Prev Hist = No)}}{\text{Odds for (Prev Hist = Yes)}} = \frac{18/12}{41/115} \cong 4.21$$

odds of low birth are four times higher

Comments:

1. An Odds Ratio of 1.0 is again our reference value. What does an Odds Ratio of 1 mean?
   The odds are the same regardless of previous history

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.

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.

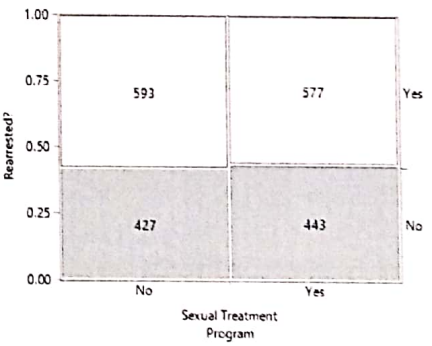
> "To evaluate the effectiveness of sex offender treatment programming, the DOC (Department of Corrections) examined the recidivism outcomes among 2,040 sex offenders released from prison between 1990 and 2003. Recidivism data were collected on 2,040 offenders through 2006. ..... Untreated and treated 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

	Sexual Treatment Program	Rearrested?	Count
1	Yes	Yes	577
2	Yes	No	443
3	No	Yes	593
4	No	No	427



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

Odds Ratio

Odds Ratio	Lower 95%	Upper 95%
1.066239	0.894606	1.270801

Questions

1. Compute the odds of a rearrest (for any reason) for a person not in a sexual treatment program.
   593 : 427
2. Compute the odds of a rearrest (for any reason) for a person who was in a sexual treatment program.
   577 : 443
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.

$$\frac{593:427}{577:443} \cong 1.066$$

4. 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.

$$\frac{145:432}{199:394} \approx \frac{.5051}{.3356} \approx 1.5$$

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

Table 2. Crude Odds Ratios for the Relationship between Individual Risk Factors Included in the RRASOR and the Static-99 and Criminal Recidivism Among Sex Offenders Released from Prison in Sweden

Items	Base rate[a]	Sexual recidivism		Any violent recidivism[b]	
		Odds ratio (95% CI)	Pearson's *r*	Odds ratio (95% CI)	Pearson's *r*
1. Prior sex offenses[c]	12		.29**		.17**
Score of 1[d]	7	4.13 (1.82–9.36)	.07*	2.23 (1.30–3.81)	.07*
Score of 2[d]	4	14.18 (6.68–30.09)	.20**	4.19 (2.25–7.81)	.12**
Score of 3[d]	1	26.93 (9.13–79.44)	.19**	5.35 (1.91–14.96)	.09**
2. Prior sentencing dates	31	4.16 (2.35–7.39)	.14**	4.00 (2.86–5.59)	.23**
3. Any noncontact sex offenses	23	2.93 (1.68–5.10)	.11**	1.92 (1.36–2.72)	.10**
4. Index nonsexual violence	28	1.10 (0.61–2.01)	.01	2.22 (1.60–3.10)	.13**
5. Prior nonsexual violence	29	2.49 (1.43–4.32)	.09**	3.93 (2.81–5.48)	.23**
6. Any unrelated victims[c]	51	3.90 (1.99–7.64)	.11**	2.03 (1.44–2.84)	.11**
7. Any stranger victims	23	3.67 (2.10–6.43)	.13**	3.27 (2.33–4.59)	.19**
8. Any male victims[c]	8	1.59 (0.67–3.82)	.03	0.58 (0.28–1.22)	−.04
9. Young age (18–24.99 years)[c]	8	0.92 (0.33–2.59)	.00	2.42 (1.51–3.88)	.10**
10. Single	27	1.97 (1.13–3.45)	.07*	2.15 (1.54–3.00)	.12**

Note. The average postrelease follow-up time was 3.69 years.
[a]Percent prevalence of each item in cohort.
[b]Including sexual offenses.
[c]Item included in the RRASOR.
[d]A score of 0 is used as reference category.
*$p < .05$. **$p < .01$.

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

5. What are the most important factors that influence the sexual recidivism in this study? What are the least? Prior Sex offenses

   Age seems to have the least influence (.92 ~ 1)

6. 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?