Final Exam Sample

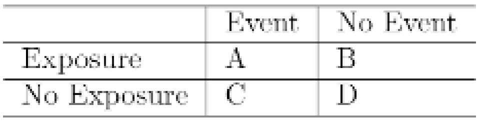
STA209-04: Applied Statistics

May 4, 2019

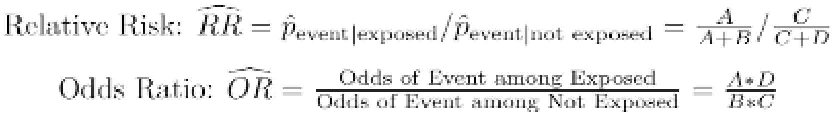
Statistic	**Standard Error**
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$\bar{x}_d$	$\frac{\sigma_d}{\sqrt{n_d}}$

Other Formula(s)

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$



	Event	No Event
Exposure	A	B
No Exposure	C	D



Relative Risk: $\widehat{RR} = \hat{p}_{\text{event}|\text{exposed}} / \hat{p}_{\text{event}|\text{not exposed}} = \frac{A}{A+B} / \frac{C}{C+D}$

Odds Ratio: $\widehat{OR} = \frac{\text{Odds of Event among Exposed}}{\text{Odds of Event among Not Exposed}} = \frac{A*D}{B*C}$

Confidence Level	**80%**	**90%**	**95%**	**99%**
z	1.282	1.645	1.960	2.576
$t_{df=5}$	1.476	2.015	2.571	4.030
$t_{df=10}$	1.372	1.812	2.228	2.764
$t_{df=15}$	1.341	1.753	2.131	2.602
$\chi^2_{df=1}$	1.640	2.710	3.840	6.630
$\chi^2_{df=2}$	3.220	4.610	5.990	9.210
$\chi^2_{df=3}$	4.640	6.250	7.810	11.340
$\chi^2_{df=4}$	5.990	7.780	9.490	13.280

1) [40 pts] The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) aims to conduct and support research on several common chronic conditions with hopes of improving the health and quality of life of those afflicted. One primary condition of interest to the NIDDK is diabetes.

One NIDDK study was interested in determining factors associated with diabetes among females at least 21 years old of Pima Indian heritage.

i) One factor commonly found to be associated with diabetes is BMI. The CDC defines an individual as obese if their BMI is greater than or equal to thirty. The following table cross-tabulates obese and non-obese patients by their diabetes diagnosis.

	Diabetic	**Non-Diabetic**
Obese	219	253
Non-Obese	49	247

Using the provided table, compute and interpret both the odds ratio and relative risk for having diabetes given a BMI ≥ 30.

ii) Suppose that these data were collected by recruiting individuals based on their diabetes diagnosis, as opposed to through a random sampling from the population of 21+ year old females of Pima Indian hertiage. Are both of the quantities computed in the previous question appropriate to use to quantify the strength of the association between obesity and diabetes? If not, explain why not and state which should be used.

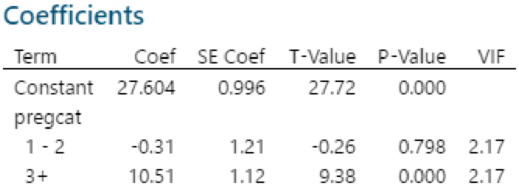
iii) Suppose that these data were collected by recruiting from among individuals without diabetes that were then followed forward in time for a number of years. Which of the quantities computed in (i) - odds ratio, relative risk, or both - could be used to quantify the strength of the association between obesity and diabetes? How would you characterize this study design?

iv) Suppose that investigators were interested in determining whether being obese caused diabetes. Which of the previously described study designs would you recommend the investigators use? Why?

v) In addition to BMI, investigators found that both the number of pregnancies (”None”, ”1-2”, ”3+”) and age of the subject were associated with a diabetes diagnosis. Given these two associations, researchers were interested in assessing whether age confounded the relationship between the number of pregnancies and a diabetes diagnosis. An ANOVA was then performed to determine whether number of pregnancies and age were associated. Complete the ANOVA table provided below to determine whether age is a confounding variable. Be sure to explain how the ANOVA results support or refute the idea that age is a confounder. (Note that the 95% critical value for the appropriate F-distribution is 3).

Source	DF	SS	MS	F-Value	P-Value
Group					
Error		84192			
Total	767	106078			

vi) An alternate approach to determining whether there is an association between age and number of pregnancies would be to fit a regression model with age as the outcome and number of pregnancies as a covariate. Doing so yields the following output:



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	27.604	0.996	27.72	0.000	
pregcat					
1 - 2	-0.31	1.21	-0.26	0.798	2.17
3+	10.51	1.12	9.38	0.000	2.17

Using this output, determine the equation for the fitted regression line.

vii) Interpret the coefficient corresponding to ”1-2”. Your answer should directly include or reference the idea of a ”reference category”.

viii) Based on this model, what is the predicted age for an individual who has never been pregnant?