# MATH 4939 Questions

1. In order to assess factors related to reckless driving behaviors, investigators ran a study in which observers at different intersections with a stop sign recorded the number of cars that did not stop properly along with various information on each violator, including gender of the driver, type of car (sedan, sports utility, mini-van, wagon, truck, other), and approximate age of the driver (under 30, 30-40, 40-50, 50+). Pooling information from several intersections and for different observers, investigators recorded the number of violators in each category.
   Describe a generalized linear model for analyzing these data (specify a reasonable, possibly non-linear, model and its transformation to linear form as well as a conditional distribution and link function) and outline a specific approach for assessing the following questions of interest using frequentist methods: (1) is gender an important predictor? (2) is type of car an important predictor, and if so which types are predictive of a greater frequency of violations? (3) is there a trend with age in the frequency of violations?

2. In marketing research, it is widely believed that subliminal messages in advertisements can be effective in improving a consumer's impression of a product. To test whether the type and frequency of the subliminal message has an impact, investigators ran a study in which they enrolled male and female graduate students. Study subjects were all shown an outwardly-identical taped advertisement for a soft drink and then asked to rank their impression of the soft drink after watching the tape on a scale from 1-5 (1=strongly negative, 2=mildly negative, 3=no opinion, 4=mildly positive, 5=strongly positive). Tapes varied in the type of subliminal message (1=none, 2=attractive female face, 3=attractive male face) and (for tapes with a subliminal message) the frequency (1=low, 2=medium, 3=high).
   Describe a regression model for relating gender, type of subliminal message, and frequency to an individual's impression. Describe the specifics of a Bayesian approach for addressing the following questions of interest to the investigator: (1) do subliminal messages have an effect overall? (2) does this effect vary depending on the gender of the observer? (3) does the effect depend on the frequency? and (4) do males and females respond differently depending on the type of subliminal message? Detail the prior used, the form of the regression model, and the likelihood. Provide an outline of the method used for posterior computation and inferences on the above questions of interest.

3. Suppose that a random variable $Y$ has a Poisson distribution with mean $\lambda$ were $ln(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Assuming that $X_2$ is not constant and that $\beta_2 \neq 0$, show that $Y$ does not have a Poisson distribution if one fits a model using only $X_1$.

4. Uterine fibroids are a common reproductive tract tumor. To study factors related to fibroid incidence (that is, the rate of onset for women who do not have fibroids), women aged 20-30 who did not have fibroids at a baseline examination were enrolled in a prospective study. These women were then given a screening examination approximately every 5 years (though the specific ages at examination varied) to assess whether fibroids had yet developed. Each woman was followed for 15 years (3 examinations) or until she either developed fibroids or dropped out of the study. Various information was collected for each women, including race (white, black, other), age at menarche, age at entry into the study, age at each examination, and whether the woman was a smoker.
   Assuming that fibroids do not go away once they have developed, describe a regression model for these data that allows fibroid incidence to vary with age and other factors. Show the observed data likelihood under this regression model, and develop a Bayesian approach for estimating age-specific fibroid incidence for women in different groups. Detail the prior used and the form of the posterior. Outline an approach for posterior computation and describe specifically how you obtain point and interval estimates for the probability of developing fibroids by a given age for women in different groups.

5. For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4. What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males."

6. *(continued from previous question)* When would the quoted interpretation be approximately correct? Why?

7. *(continued from previous question)* The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.

8. Explain what is meant by overdispersion, and explain how it can occur for Poisson generalized linear models for count data.

9. Explain two ways in which the generalized linear model extends the ordinary regression model that is commonly used for quantitative response variables.

10. Each of 100 multiple-choice questions on an exam has five possible answers but one correct response. For each question, a student randomly selects one response as the answer. Specify the probability distribution of the student's number of correct answers on the exam, identifying the parameter(s) for that distribution. Would it be surprising if the student made at least 50 correct responses? Explain your reasoning.

11. Suppose `y`, `x`, and `z` are numerical variables in the R data frame `dd`.
Explain the difference between a linear model fitted with the formula `y ~ x*z` in comparison with the formula `y ~ I(x*z)`.

12. There is a lab test for a rare disease D that has a specificity of .98 and a sensitivity of .95. Suppose the prevalence of the disease is .01%. Explain how, if one thinks of the lab test as a hypothesis test for the null hypothesis of no disease, a positive result produces a p-value of 0.01.

13. *(continued from previous question)* If someone selected at random (for example if the test is used to screen for the disease) is given the test and gets a positive result, what is the probability that they have the disease?

14. An article states that the PSA blood test for detecting prostate cancer stated that, of men who had this disease, the test fails to detect prostate cancer in 1 in 4, and, of men who do not have it, approximately 2/3 received false positive results. Let $D$ ($D^c$) denote the event of having (not having) prostate cancer and let $Pos$ ($Neg$) denote a positive (negative) test result. What is the sensitivity and specificity of this test? (from Agresti, 2007)

15. *(continued from previous question)* Of men who take the PSA test, 1% have the disease. Find the cell probabilities in the $2 \times 2$ table for the joint distribution for having (not having) the disease versus a positive or negative test result.

16. *(continued from previous question)* Find the probability of having prostate cancer given a positive test result and find the probability of not having the disease given a negative test result.

17. A lab test for a rare disease D has a specificity of .95 and a sensitivity of .95. Suppose the prevalence of the disease is .01%. What proportion of the time is the test in error?

18. *(continued from previous question)* Given a positive test result, what is probability of error?

19. *(continued from previous question)* Given a negative test result, what is probability of error?

20. *(continued from previous question)* How do you reconcile the 3 preceding results?

21. A British study in 1998 report that, of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer". What kind of statistic is the figure '1.7' reported here? (Agresti, 2007)

22. A National Cancer Institute study about tamoxifen and breast cancer reported in 1998 that the women taking the drug were 45% less likely to experience invasive breast cancer compared with the women taking placebo. Find the relative risk for (i) those taking the drug compared to those taking placebo, (ii) those taking placebo compared to those taking the drug.

23. In the United States, data reported in 1993 indicates that the annual probability that a woman over the age of 35 dies of lung cancer equals 0.001304 for current smokers and 0.000121 for nonsmokers. Calculate and interpret the difference of proportions, the relative risk and the odds ratio. Which is more informative? Why? (Agresti, 2007)

24. *(continued from previous question)* Are the odds ratio and relative risk similar or dissimilar? Why?

25. With a $2 \times 2$ table what is the effect of interchanging two rows on the odds ratio? On the log-odds ratio?

26. AVP: Find an approximate 95% confidence intererval for $\beta_{XXXX}$. Show how you obtained it. You don't need to describe the process in detail but show evidence of the process you used.

27. *(continued from previous question)* Anova anova show the null model and the alternative model for each line. You may use R's linear model formula notation to define the models.

28. *(continued from previous question)* Linear combination question.

29. The eagle question with the gamma. Include information on the gamma

30. 23Q

31. 23Q

32. Beta space question

33. The following is output from an script in R using the Prestige data in the 'car' package:

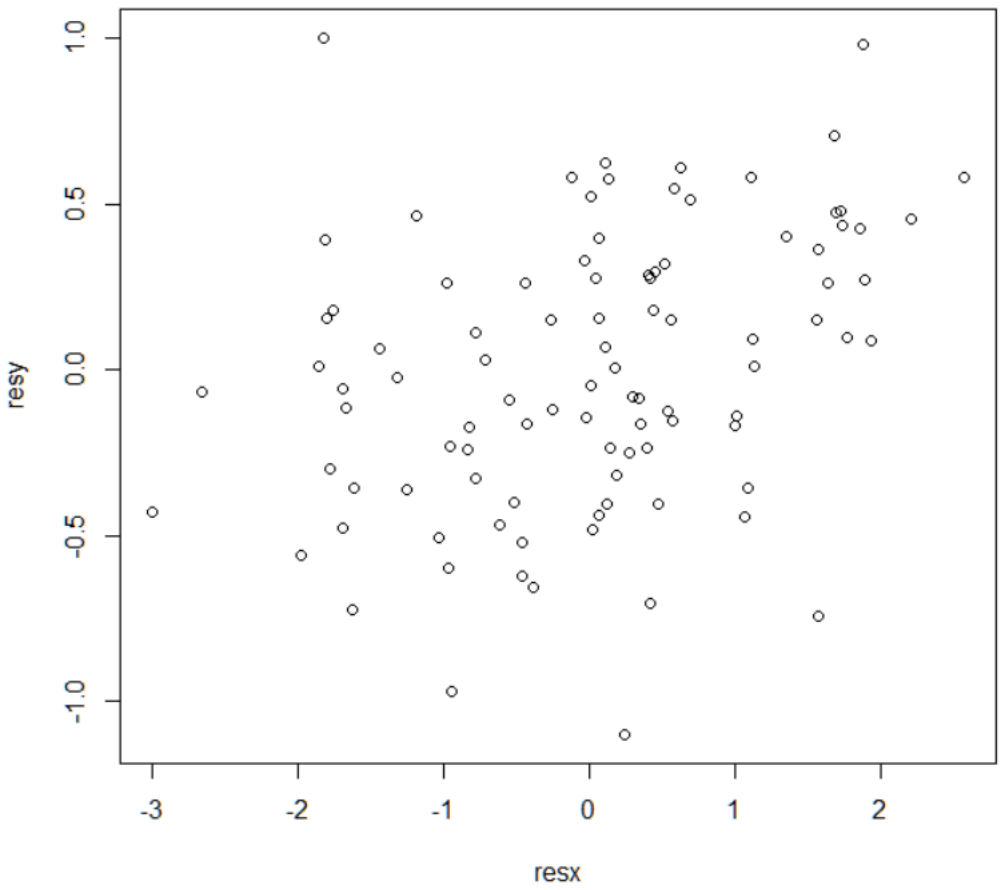
```
> library(car)
> dd <- Prestige
> head(dd)
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
physicists              15.64  11030  5.13     77.6   2113 prof
> tab(dd, ~ type)
type
   bc  prof    wc  <NA> Total
   44    31    23     4   102
> fit <- lm(log(income) ~ education + type, dd)
> fit

Call:
lm(formula = log(income) ~ education + type, data = dd)

Coefficients:
(Intercept)    education     typeprof       typewc
    7.48548      0.12284     -0.05802     -0.38323

> resy <- resid(lm(log(income) ~ type, dd))
> resx <- resid(lm(education ~ type, dd))
```

and this is the plot generated by the script:




If you have the information to construct an approximate 95% confidence interval for the coefficient of 'education' controlling for 'type', do so. Show how you constructed the interval. You do not need to explain how you constructed it in detail as long as it is obvious from your work on the exam. If you do not have the information, explain why not.

34. *(continued from previous question)* What is the difference in predicted income from this model for someone with 12 years of education in a professional occupation versus a white collar occupation?

35. Consider the problem of imputing a grade for a missed mid-term mark from a student's performance on a final exam. Suppose a professor is considering three possibilities: 1) Use the student's z-score on the final exam to impute the z-score on the mid-term, 2) perform a linear regression of mid-term grades on the final exam grade and impute the mid-term grade by using the predicted mid-term grade, and 3) use a linear regression of the final exam grade on the mid-term grade and impute the mid-term grade that would have predicted the student's final exam grade.
Which of these three methods would be advantageous for a student who has a very high grade on the final exam? Which would be advantageous for a student who has a very low grade on the final exam? Clearly show the reasoning behind your answer.

36. Explain what is meant by the Hauck-Donner phenomenon. How does it affect the practice of logistic regression?

37. In the early stages of an epidemic, the number of new cases grows exponentially. Suppose that the expected number of cases on day $t_i$ is modelled as $\gamma e^{\delta t_i}$. Specify a GLM that you could use to analyze data in which the response is the number of new cases on a number of specific days, for example: days 5, 6, 10, 12, 20 and 23. Describe in detail the transformation of the model to linear form and the interpretation of the original parameters above and of the linear parameters of the model.

38. Let $Y$ be number of fatal accidents on a day in Toronto. Suppose the expected number of fatal car accidents depends on a number of variables and that, if all of these variables were taken into account, the conditional distribution of $Y$ would be Poisson. Show that, assuming that the variables are not all constant, the unconditional distribution of $Y$ itself is overdispersed relative to that of a Poisson distribution.

39. Give a real world example of three variables, $X, Y$, and $Z$, for which we would expect $X$ and $Y$ to be marginally associated but conditionally independent controlling for $Z$.

40. The Cowles data frame has 1421 rows and 4 columns. These data come from a study of the personality determinants of volunteering for psychological research. Neuroticism (neuro) is classified in three levels: low, medium and high. Extraversion (extra) is measured on a scale that ranges from 1 to 25. The purpose of the study is to explore some personality predictors of the willingness to volunteer and how the prediction differs between men and women. This is some output from an R script:

```
> library(spida2)
> library(car)
> head(Cowles)
   neuro extra    sex volunteer
1 medium    13 female        no
2    low    14   male        no
3    low    16   male        no
4    low    20 female        no
5    low    19   male        no
6    low    15   male        no
> dim(Cowles)
[1] 1421    4
> Cowles %>% tab(~neuro+sex)
        sex
neuro    female male Total
  high       47   25    72
  low       234  280   514
  medium    499  336   835
  Total     780  641  1421
> Cowles %>% tab(~volunteer)
volunteer
   no   yes Total
  824   597  1421
> fit <- glm(volunteer ~ neuro*extra*sex, Cowles, family = binomial)
> print(fit)

Call:  glm(formula = volunteer ~ neuro * extra * sex, family = binomial,
    data = Cowles)

Coefficients:
              (Intercept)                   neurolow                neuromedium
                  -0.27178                   -0.61174                   -0.87311
                    extra                    sexmale              neurolow:extra
                  -0.00257                    5.45731                    0.06239
       neuromedium:extra           neurolow:sexmale         neuromedium:sexmale
                   0.07467                   -6.81494                   -5.24577
           extra:sexmale     neurolow:extra:sexmale  neuromedium:extra:sexmale
                  -0.43738                    0.52151                    0.39408
```

What is the predicted probability of volunteering for a male with high 'neuro' and extraversion equal to 20?

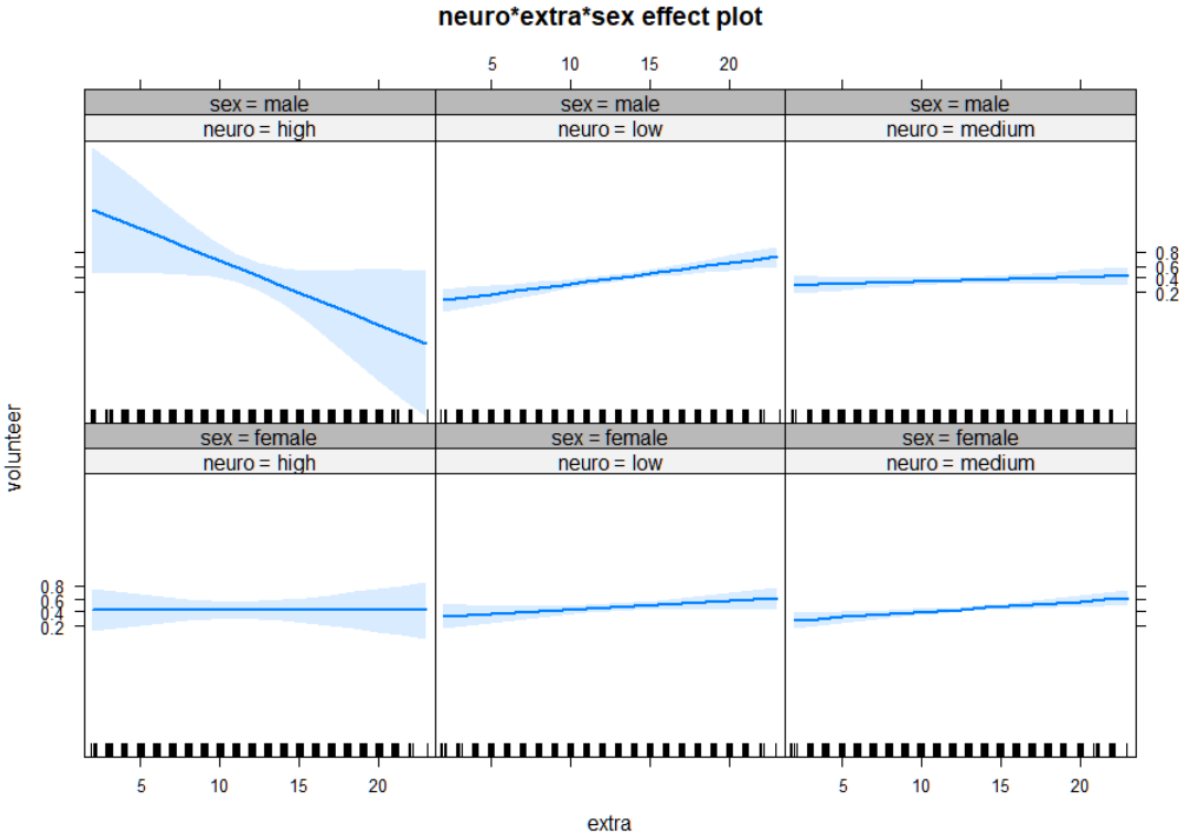
41. *(continued from previous question)* For each row of the following Anova table specify both the null hypothesis and the alternative hypothesis being tested. You may use R's linear model formulas to express the hypotheses or you may use the parameters of the model provided they are named in a way

that is clear.

```
> Anova(fit)
Analysis of Deviance Table (Type II tests)

Response: volunteer
                LR Chisq Df Pr(>Chisq)
neuro             2.1124  2   0.347769
extra            21.1480  1  4.251e-06 ***
sex               5.6187  1   0.017770 *
neuro:extra       8.9289  2   0.011511 *
neuro:sex         1.9909  2   0.369556
extra:sex         0.1109  1   0.739082
neuro:extra:sex   9.4006  2   0.009092 **
```

42. *(continued from previous question)* The following is an effect plot produced by the 'effects' package.




What are the main reasons that the confidence bands are wider for some combinations of neuro and sex than for others?

43. Sally Clark was convicted of murder of her two children who died with no signs of illness or trauma on the strength of statistical evidence that claimed that assuming the 'null hypothesis' that she is innocent, the probability of 2 deaths occuring due to the only known explanation, sudden infant death syndrome, was extremely small. In testimony, Sir Roy Meadow claimed the probability to be 1 in 73 million but better estimates would put it at a maximum of 1 in 100,000.
Explain whether 0.00001 can be considered a maximal p-value for assessing the hypothesis that Sally Clark is innocent. Discuss whether there are other ways of assessing the probability of her guilt? Work out a rough calculation based on reasonable guesses for the result of such a calculation.

44. Discuss how it could be possible for a regression with two linear predictors to produce two different final

models when using forward stepwise versus backward stepwise variable selection algorithms. Explain how a confidence region for the coefficients of the two linear predictors is related to forward and backward stepwise selection

45. A survey of students taking a large course showed (this is true) that students who viewed the recorded videos of the lectures many times performed less well on the final exam than students who viewed the videos fewer times. Upon discovering this study, your professor announces that they will discontinue recording the lectures because, the professor says, there is evidence that the videos cause students to do perform more poorly on courses. Comment on the professor's reasoning using concepts we have studied in this course.

46. Choose a possible confounding factor and use a Paik-Agresti diagram to show how controlling for this confounding factor could reverse the direction of association between the frequency of video viewing and performance on the course.

47. What output will the following R script produce? Explain briefly why.

48. What output will the following R script produce? Explain briefly why.

49. Let x be defined as:

    Write an R function that would turn `x` into a factor whose ordering corresponds to the numerical ordering of `x`.

50. In R, let `x <- 1:5`. What output would `x[NA]` produce? What output would `x[NA_real_]` produce? Describe the reason for the difference, if any.

51. In R, describe the result of subsetting a vector with positive integers, with negative integers, with a logical vector, or with a character vector?

52. In R, what's the difference between [, [[, and $ when applied to a list?

53. In R, when subsetting with [, when should you use drop = FALSE? Include arrays and factors in your discussion.

54. In R, If x is a matrix, what does x[] <- 0 do? How is it different from x <- 0?

55. In R, how can you use a named vector to relabel a categorical variable?

56. Fix each of the following common data frame subsetting errors in R:

57. In R, if mtcars is a data frame, why does mtcars[1:20] return an error? How does it differ from the similar mtcars[1:20, ]?

58. In R, if df is a data frame, what does df[is.na(df)] <- 0 do? How does it work?

59. Create the vector (20,19, ... ,2,1) in R.

60. Create the vector (1,2,3, ... ,19,20,19,18, ... ,2,1) in R.

61. Create the vector (4,4, ... ,4,6,6, ... ,6,3,3, ... ,3) in R, where there are 10 occurrences of 4, 20 of 6 and 30 of 3.

62. Write an expression in R to calculate the following $\sum_{i=10}^{100}(i^3 + 4i^2)$

63. Generate in R a vector of 30 labels: 'label 1', 'label 2', ... 'label 30'

64. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are divisible by 2.

65. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are divisible by 2.

66. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are within 200 of the maximum value.

67. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are less than the previous element.

68. Let `y <- sample(1000, 30, replace = TRUE)`. Write an expression in R to determine how many elements of `y` are an exact square.

69. Suppose data for a variable in R representing dollars has been entered in a variety of formats: '$1,000.00','1000.00','$1'. Write a function in R that transforms the variable to a numeric variable in dollars to the nearest cent.

70. Write a function in R that takes a character string and collapses multiple adjoining blanks to a single blank.

71. Use the site Gapminder.org to download at least three longitudinal variables into separate data sets. Merge the data sets into one for which each row represents one country and year and contains the values of each of the three variables you downloaded.

72. Write a function in R that removes every variable whose name starts with the letter 'X' and ends in a number from a data frame.

73. Create a 6 by 10 matrix of random integers in R as follows:

    Write a function to find the number of entries in each row that are greater than 4.

74. *(continued from the previous question)* Write a function to find how many row have exactly two instances of the number 7.

75. Describe the difference in R between `paste(x, y, sep = ':')` and `paste(x, y, collapse = ':')`

76. Using the `hs` data set in the spida2 package, create a plot with two panels showing histograms displaying the distribution of school sizes in the Public and in the Catholic sectors. Use the functions `capply` and `up` in the spida2 package. You may also use any other approach to compare with the use of `capply` and `up`.

77. Using the `hs` data set in the spida2 package, create a plot with two panels showing histograms displaying the distribution of sample sizes in each school in the Public and in the Catholic sectors. Use the functions `capply` and `up` in the spida2 package. You may also use any other approach to compare with the use of `capply` and `up`.

78. Using the `hs` data set in the spida2 package, create a plot with two panels showing scatterplots displaying the relationship between mean mathach and mean ses in each school in the Public and in the Catholic sectors. Explore reasonable transformations and regression lines: linear and non-parametric in the plots. Use the functions `capply` and `up` in the spida2 package. You may also use any other approach to compare with the use of `capply` and `up`.

79. Describe the difference in R between a 'generic function' and a method.

80. What is wrong with the following claim: "Data show that income and marriage have a high positive correlation. Therefore, your earnings will increase if you get married."

81. What is wrong with the following claim: "Data show that as the number of fires increase, so does the number of fire fighters. Therefore, to cut down on fires, you should reduce the number of fire fighters."

82. What is wrong with the following claim: "Data show that people who hurry tend to be late to their meetings. Don't hurry, or you'll be late."

83. A baseball batter Tim has a better batting average than his teammate Frank. However, someone notices that Frank has a better batting average than Tim against both right-handed and left-handed pitchers. How can this happen? Present your answer as a hypothetical table and in a Paik diagram.

84. Discuss whether you should use the aggregate (marginal) or the segregated (conditional) data to attempt to determine the true effect in the following situation: There are two treatments used on kidney stones: Treatment A and Treatment B. Doctors are more likely to use Treatment A on large (and therefore,

more severe) stones and more likely to use Treatment B on small stones. Should a patient who doesn't know the size of his or her stone examine the general population data, or the stone size-specific data when determining which treatment will be more effective?

85. Discuss whether you should use the aggregate (marginal) or the segregated (conditional) data to attempt to determine the true effect in the following situation: There are two doctors in a small town. Each has performed 100 surgeries in his career, which are of two types: one very difficult surgery and one very easy surgery. The first doctor performs the easy surgery much more often than the difficult surgery and the second doctor performs the difficult surgery more often than the easy surgery. You need surgery, but you do not know whether your case is easy or difficult. Should you consult the success rate of each doctor over all cases, or should you consult their success rates for the easy and difficult cases separately, to maximize the chance of a successful surgery?

86. Discuss whether you should use the aggregate (marginal) or the segregated (conditional) data to attempt to determine the true effect in the following situation: In a study of a group of male 50 to 55-year-old long-time smokers, researchers compared a group of heavy smokers with matched group (same age range, sex and similar socioeconomic and environmental backgrounds) of light smokers. It was found that lung function was worse in the group of heavy smokers than in the group of light smokers. The researchers also measured the amount of tar deposit in the lungs of the subjects and classified subjects as having heavy or light tar deposits. Would you get a better indication of the effect of smoking by comparing the aggregated data for the two groups or by comparing the tar level specific data?

87. When interpreting a study that purports to show a relationship between two variables, what do you think are the three most important questions that you should ask? Discuss as succinctly as you can the consequences of the answers to those questions.

88. R. A. Fisher insisted that causal inference was impossible in the absence of an experiment with random assignment to a 'treatment' variable. Discuss why Fisher's position could be considered correct but why it may be considered impractical?

89. Give an example of a situation in which we would be interested in predictive inference and an example in which we would be interested in causal inference.

90. Here are some fictitious data on the rate of complications for appendectomies performed at University Hospital, a large urban teaching and research hospital, and in County Hospital, a small-town hospital: at University Hospital there were 800 cases with 100 (12.5%) resulting in complication and at County Hospital there were 200 cases resulting in 5 (10%) complications. The p-value for a test of the hypothesis that there is no difference in the rate at the two hospitals is 0.0037. Suppose that appendectomies can be classified as high risk or low risk and that the high risk cases tend to be directed disproportionately to University Hospital instead of County Hospital. Construct two hypothetical tables, one for each level of risk, and draw a Paik diagram that shows how it is possible for both high- and low-risk patients to have a lower probability of complications at University Hospital than at County Hospital, although, overall the probability of complications is higher at University Hospital than at County Hospital.

91. Suppose a test for glaucoma has a sensitivity of .95 and a specificity of .90. You receive the test as a routine test on a regular visit to your optometrist. The prevalence of glaucoma in your age, ethnic and gender group among people who have not been previously diagnosed is 1 per 100. The test, alas, is positive. Use a natural frequency table to find the probability that you have glaucoma given the positive test result.

92. A group of major medical journals are now requiring that all authors who intend to publish in their journal must preregister their experimental designs and their intended analyses for all the clinical endpoints (responses) they intend to report before obtaining data if they intend to publish their results in their journal. Authors must also agree to publish their findings whether the results achieve statistical significance or not. In what ways does this policy contribute to mitigating lack of reproducibility?

93. A group of major medical journals are now requiring that all authors who intend to publish in their journal must preregister their experimental designs and their intended analyses for all the clinical

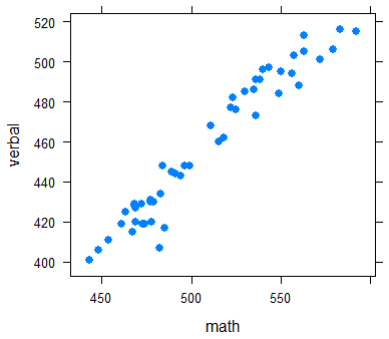
endpoints (responses) they intend to report before obtaining data if they intend to publish their results in their journal. Authors must also agree to publish their findings whether the results achieve statistical significance or not. In what ways does this policy contribute to mitigating lack of reproducibility?

94. A study investigated whether there was a higher risk of complications when women gave birth at home with the assistance of a midwife instead of giving birth in a maternity ward in a hospital. Of 400 women who chose to give birth at home 20 had complications and of 2,000 women who gave birth in a hospital 200 had complications. Do you think that this is an experimental study or an observational study? Why?

95. (continued from the previous question) The data suggest that it is safer (in the sense of a lower rate of complications) to give birth at home than to give birth in the hospital. Discuss whether this implies that a woman should consider giving birth at home in order to reduce her risk of complications. Identify at least one plausible confounding factor and one plausible mediating factor that could partly explain the results of the study

96. (continued from the previous question) Choose a possible confounding factor and use a Paik diagram to show how controlling for this confounding factor could reverse the direction of association between the rate of complications and the location of birth: home or hospital.

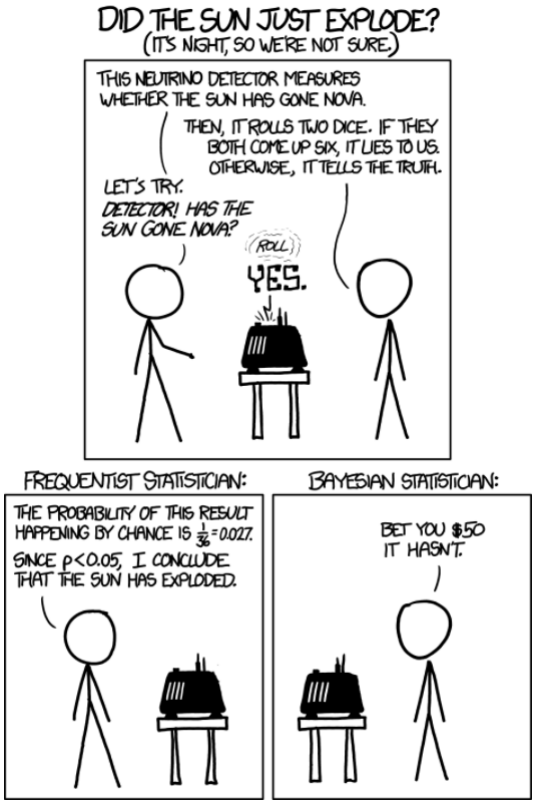
97. This graph shows the mean verbal and mean math scores on the SAT test in each of the 50 states of the US.




Make an intelligent guess of the correlation between these two variables. Show the basis for your guess – perhaps by drawing an appropriate geometric figure on the graph above.

98. Suppose you were to read about a study based on a random survey of Ontario medical records that shows that smokers have twice as high a risk of kidney disease as non-smokers. Is it reasonable to conclude that smoking causes a higher risk of kidney disease? Why or why not?

99. The following XKCD cartoon:




shows two statisticians interpreting the same data: one who uses a frequentist approach unquestioningly and one who uses a Bayesian approach. Make some reasonable assumptions, stating them explicitly, and calculate a reasonable value for the Bayesian statistician's posterior probability that the sun has exploded. Discuss why there is a difference between the 'p-value' of 0.027 and the Bayesian posterior probability.

100. Which of the following R expressions result in the following output?

```
[1]  8
```

(Write 'Y' for yes, 'N' for no, and 'D' or blank for 'do not know'. +1 for a correct answer, -1 for a wrong answer and 0 for 'D')

_____ `"+"(5,3)`

_____ `"/"(16,2)`

_____ `2^3`

_____ `4 + 4`

_____ `"^"(3,2)`

101. Suppose we run this command:

```
a <- matrix(1:8, rnow = 2)
```

Then which of the following R expressions result in the following output?

```
[1]  8
```

(Write 'Y' for yes, 'N' for no, and 'D' or blank for 'do not know'. +1 for a correct answer, -1 for a wrong answer and 0 for 'D')

_____ `a(8)`

_____ `a[8]`

_____ `a(2,4)`

_____ `a[4,2]`

_____ `a(2,4)`

102. Suppose you have data on two variables, $X$ and $Y$, in each of $J$ groups. Let $\hat{\beta}_j$ represent the vector of coefficients of least-squares regression ot $Y$ on $X$ within the the $j$th group, $j = 1, \ldots, J$. Prove that the inverse-variance weighted combination of the within-group estimated coefficients is the same as the vector of coefficients for the least-squares regression using the pooled data. Explain why (a sketch will suffice, no formal argument is needed) the pooled estimate of the slope estimates a combination of the within-group regression slope and the between-group regression slope.

103. Suppose you wish to estimate the relationship between income ($Y$) and education ($X$). Because of heteroscedasticity and curvature in the relationship you choose to fit a linear model using the log of $Y$:

```
fit <- lm( log(Y) ~ X, data)
```

Write the R code you would use to plot the estimated increase in income associated with an extra year of education as a function of years of education. It is not necessary to include error bars in the plot.