Math 150 – Methods in Biostatistics

Spring 2019

Jo Hardin

iClicker Questions

to go with **Practicing Statistics**

Kuiper & Sklar

1. Hypothesis test

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

2. Confidence interval

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

3. Sample Mean

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

4. Central Limit Theorem

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

5. Standard Deviation

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

6. p-value

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

7. t-test

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

8. Interaction

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

9. Simple linear regression

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

10. Multiple linear regression

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

11. Logistic regression

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

12. Survival analysis

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

13. chi-square test

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

14. R

(a) Never heard of it

(b) Heard of it, but don’t know anything about it

(c) Know a little about it (or did once)

(d) Know something about it

(e) Confident about it

15. How did R go for you?

(a) Didn’t log in

(b) Logged in, but couldn’t run anything

(c) Able to do assignment in R, but didn’t do Markdown

(d) Able to do Markdown to turn in today

(e) Isn’t there something more complicated that we could use?

16. Why do we use a t distribution in the t-test?

(a) the technical assumptions don’t hold

(b) the means are extra variable

(c) we like the letter t

(d) we have two samples

(e) we don’t know the true standard deviation parameter

17. What happens if a t-test is used but isn’t appropriate (technical conditions don’t hold)?

(a) the p-value isn’t actually the probability of our data or more extreme if H0 is true.

(b) the software won’t give a p-value as output

(c) the rejection region needs to be calculated in the opposite direction

(d) the world blows up

18. We use linear regression to run a test of means

(xi = 0 for controls, group 1; xi = 1 for cases, group 2)

What is: ?

(a) n

(b) n1

(c) n2

(d) n1 1

(e) n2 2

19. We use linear regression to run a test of means

(xi = 0 for controls, group 1; xi = 1 for cases, group 2)

What is: ?

(a) n

(b) n1

(c) n2

(d) n1 1

(e) n2 2

20. With a strong correlation and very small p-value, what can we conclude about happiness and life expectancy?

(a) happiness causes longer lives

(b) longer lives cause happiness

(c) happiness and longer life are correlated

(d) happiness and longer life are perfectly predictive

(e) happiness and longer life are unrelated

21. Why do we check technical conditions?

(a) so that the inference is valid

(b) so that the estimates are valid

(c) so that the p-value is more likely to be small

(d) so that the confidence level is right

(e) for fun

22. When writing the regression equation, why do we put a hat ( ^) on the response variable?

(a) because the prediction is an estimate

(b) because the prediction is an average

(c) because the prediction may be due to extrapolation

(d) (a) & (b)

(e) all of the above

23. If there is no relationship in the population (true correlation = 0), then r=0

(a) TRUE

(b) FALSE

24. If there is no relationship in the population (true slope = 0), then b1=0

(a) TRUE

(b) FALSE

25. The regression technical assumptions include:

(a) The Y variable is normally distributed

(b) The X variable is normally distributed

(c) The residuals are normally distributed

(d) The slope coefficient is normally distributed

(e) The intercept coefficient is normally distributed

26. A smaller variability around the regression line (σ):

(a) increases the variability of b1.

(b) decreases the variability of b1.

(c) doesn’t necessarily change the variability of b1.

27. A smaller variability in the explanatory variable (SD(X) = sx):

(a) increases the variability of b1.

(b) decreases the variability of b1.

(c) doesn’t necessarily change the variability of b1.

28. A smaller sample size (n):

(a) increases the variability of b1.

(b) decreases the variability of b1.

(c) doesn’t necessarily change the variability of b1.

29. We transform our variables…

(a) … to find the highest r^2 value.

(b) … when the X variable is not normally distributed.

(c) … to make the model easier to interpret.

(d) … so that the technical conditions are met.

30. The variance of the individual observations at a given x

(a) is smaller than the variance of the mean (predicted) value

(b) is about the same as the variance of the mean (predicted) value

(c) is larger than the variance of the mean (predicted) value

(d) is not related to the variance of the mean (predicted) value

31. Prediction intervals are

(a) smaller than confidence intervals

(b) about the same width as confidence intervals

(c) larger than confidence intervals

(d) unrelated to confidence intervals

32. Prediction intervals have

(a) the same technical conditions as CIs

(b) stricter technical conditions than CIs

(c) more lenient technical conditions than CIs

(d) technical conditions which are unrelated to CIs

1/30/2017

33. What does “by random chance” (in calculating the p-value) mean here?

(a) random allocation

(b) random sample

34. “Our data or more extreme” is:

(a) fewer than 9

(b) 9 or fewer

(c) 9 or more

(d) more than 9

35. What is the mean value of the null sampling distribution for the number of Botox therapy who showed pain reduction?

(a) 0

(b) 9

(c) 5.3

(d) 11

(e) 15

36. The 2003 study by the Bureau of Labor Statistics reported the percentage of sample respondents in each age group who had performed volunteer work in the previous year:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age group** | **16–24 years** | **25–34 years** | **35–44 years** | **45–54 years** | **55–64 years** | **65 or more** |
| **% volun** | 21.9% | 24.8% | 34.1% | 31.3% | 27.5% | 22.7% |

Consider a chi-square test. Would it be a test of homogeneity of proportions or association between variables?

(a) homogeneity of proportions

(b) test of association (independence)

37. The Pew Internet and American Life Project sampled people from 12 regions of the US and measured internet use.

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **Internet users** | **Non-internet users** | **Sample size** |
| **New England** | 541 | 338 | 879 |
| **Mid-Atlantic** | 1354 | 901 | 2255 |
| **National Capital** | 522 | 317 | 839 |
| **Southeast** | 1346 | 955 | 2301 |
| **South** | 1026 | 1005 | 2031 |
| **Industrial Midwest** | 1543 | 1114 | 2657 |
| **Upper Midwest** | 560 | 422 | 982 |
| **Lower Midwest** | 728 | 562 | 1290 |
| **Border States** | 1054 | 660 | 1714 |
| **Mountain States** | 499 | 289 | 788 |
| **Pacific Northwest** | 468 | 223 | 691 |
| **California** | 1238 | 706 | 1944 |

Consider a chi-square test. Would it be a test of homogeneity of proportions or association between variables?

(a) homogeneity of proportions

(b) test of association (independence)

38. homogeneity of proportions means

(a) the proportion of obs units who have some response (e.g., believe = 4) is the same across all levels of the explanatory variable.

(b) each response is equally likely for any explanatory variable

(c) the variables are independent

(d) a & c

(e) b & c

38. homogeneity of proportions means

(a) the proportion of obs units who have some response (e.g., believe = 4) is the same across all levels of the explanatory variable.

(b) each response is equally likely for any explanatory variable

(c) the variables are independent (this is also true, however, if we have homogeneity of proportions)

(d) a & c

(e) b & c

39. Independence between two categorical variables means:

(a) one does not cause the other

(b) knowledge of one variable does not tell us anything about the probability associated with the other variable

(c) as one variable increases, the other variable increases

(d) as one variable increases, the other variable decreases

40. Relative Risk is

(a) the difference of two proportions

(b) the ratio of two proportions

(c) the log of the ratio of two proportions

(d) the log of the difference of two proportions

41. One reason we should be careful interpreting relative risks is if:

(a) we don’t know the difference in proportions

(b) we don’t know the SE of the relative risk

(c) we might be dividing by zero

(d) we don’t know the baseline risk

42. When we select individuals based on the explanatory variable, we cannot accurately measure

(a) the proportion of people in the population in each explanatory category

(b) the proportion of people in the population in each response group

(c) anything about the population

(d) confounding variables

43. The odds ratio is “invariant to which variable is explanatory and which is response” means:

(a) we always put the bigger odds in the numerator

(b) we must collect data so that we can estimate the response in the population

(c) which variable is called the explanatory changes the value of the OR

(d) which variable is called the explanatory does not change the value of the OR

44. In finding a CI for π1/π2, why is it okay to exponentiate the end points of the interval for ln(π1/π2)?

(a) Because if ln(π1/π2) is in the original interval, π1/π2 will be in the exponentiated interval.

(b) Because taking the natural log of the RR makes the distribution approximately normal.

(c) Because the natural log compresses values that are bigger than 1 and spreads values that are smaller than 1.

(d) Because we can get exact p-values using Fisher’s Exact Test.

45. In order to find a CI for the true OR, our steps are:

1. ln(OR)

2. add ± z\* sqrt( 1/A + 1/B + 1/C + 1/D)

3. take exp of the endpoints

(a) because the sampling distribution of is normal

(b) because OR is typically greater than 1

(c) because the ln transformation makes the sampling distribution almost normal

(d) because OR is invariant to the choice of explanatory or response variable

46. Given a (standard) logistic regression model, the odds of survival are:

(a)

(b)

(c) – (

(d)

47. In a logistic regression model, the variability is given by

(a) Normal Y given X

(b) Binomial Y given X

(c) Bernoulli Y given X

(d) Poisson Y given X

48. With a logistic regression model, the relative risk of survival (for a one unit increase in X) is:

(a)

(b)

(c)

(d) a non-linear function of X (dependent on the X )

49. If we want the relative risk of survival (for a one unit increase in X) to be independent of X, we should use which link:

(a) linear

(b) logistic

(c) complementary log-log

(d) log-linear

50. When trying to find estimates for and , we maximize the likelihood.

Take the derivative with respect to which variable(s):

(a) X

(b) Y

(c)

(d)

(e) and

51. Maximum likelihood estimation seeks to:

(a) Find the data which are most likely under the model.

(b) Find the parameters which are most likely under the model.

(c) Find the parameters which make the data most likely under the model.

(d) Find the data which make the parameters most likely under the model.

52. We use maximum likelihood estimation because:

(a) It gives an intuitive way for estimating the parameters.

(b) The estimates are asymptotically normally distributed.

(c) The estimates are always easy to compute.

(d) All of the above.

(e) Some of the above.

53. We know that for a given data set (a “hat” indicates an MLE):

(a) L(L( sometimes

(b) L(L( sometimes

(c) L(L( always

(d) L(L( always

54. How many parameters do we estimate when considering Length as a categorical variable? (the only variable)

(a) 0

(b) 1

(c) 2

(d) 33

(e) 34

55. How many df for the LRT addressing whether Length (as a categorical variable) belongs in the model?

(a) 0

(b) 1

(c) 2

(d) 33

(e) 34

56. How many df for the LRT addressing whether Incubate and Color belong in the model (given Length is determined to be in the model)?

(a) 0

(b) 1

(c) 2

(d) 3

(e) 4

57. An interaction term in a multiple regression model may be used when:

(a) the model fit is poor.

(b) there is a quadratic relationship between the response and explanatory variables.

(c) neither one of two explanatory variables contribute significantly to the regression model.

(d) the relationship between X1 and P(success) changes for differing values of X2.

58. The interpretations of the main effects (on their own) make sense only when the interaction component is not significant.

(a) True

(b) False

59. If the interaction is significant but the main effects aren’t:

(a) report on the significance of the main effects

(b) remove the main effects from the model

(c) avoid talking about main effects

(d) test whether the main effects are significant without interaction in the model

60. With two variables of interest, what should you always test first?

(a) The significance of variable 1.

(b) The significance of variable 2.

(c) The interaction between variables 1 and 2.

(d) None of the above.

61. Consider variable 1 is continuous and variable 2 has 4 levels. How many degrees of freedom are associated with the drop in deviance test (LRT) of their overall interaction?

(a) 1

(b) 2

(c) 3

(d) 4

(e) 5

62. When probability of being able to buy a candy bar is modeled as a function of the **number of coins**, the coefficient on number of coins is:

(a) positive

(b) negative

(c) zero

(d) no intuition exists for being able to answer this question

63. When probability of being able to buy a candy bar is modeled as a function of the **number of low coins**, the coefficient on number of low coins is:

(a) positive

(b) negative

(c) zero

(d) no intuition exists for being able to answer this question

64. When probability of being able to buy a candy bar is modeled as a function of the **number of coins and number of low coins**, the coefficient on number of coins is:

(a) positive

(b) negative

(c) zero

(d) no intuition exists for being able to answer this question

65. When probability of being able to buy a candy bar is modeled as a function of the **number of coins and number of low coins**, the coefficient on number of low coins is:

(a) positive

(b) negative

(c) zero

(d) no intuition exists for being able to answer this question

66. Consider the exam helper situation (100 questions total)

|  |  |  |  |
| --- | --- | --- | --- |
| Sage | 85 | Sage & Bruno | 85 |
| Bruno | 75 | Sage & Beta | 95 |
| Beta | 55 | Sage & Luna | 90 |
| Luna | 45 | Bruno & Beta | 85 |
|  |  | Bruno & Luna | 80 |
|  |  | Beta & Luna | 100 |

With forward selection which two are chosen (first)?

(a) Sage & Bruno

(b) Sage & Beta

(c) Sage & Luna

(d) Bruno & Luna

(e) Beta & Luna

67. Consider the exam helper situation (100 questions total)

|  |  |  |  |
| --- | --- | --- | --- |
| Sage | 85 | Sage & Bruno | 95 |
| Bruno | 75 | Sage & Beta | 85 |
| Beta | 55 | Sage & Luna | 90 |
| Luna | 45 | Bruno & Beta | 85 |
|  |  | Bruno & Luna | 80 |
|  |  | Beta & Luna | 100 |

With backward selection which two are chosen for the final model?

(a) Sage & Bruno

(b) Sage & Beta

(c) Sage & Luna

(d) Bruno & Luna

(e) Beta & Luna

68. P(surv if X = 0.4) = 0.3

P(surv if X = 0.2) = 0.9

The two points are concordant if

1. X = 0.4 survives and X = 0.2 dies
2. X = 0.4 dies and X= 0.2 survives
3. They both survive
4. They both die

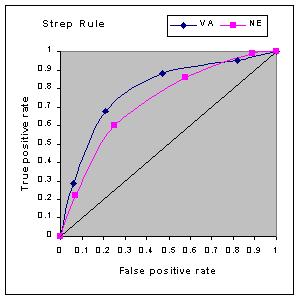
69. Which model is better (according to ROC)?

(a) pink because it goes closer to (1,1)

(b) pink because it is closer to y=x

(c) blue because it is farther from y=x

(d) blue because it is steeper

(e) neither

70. In ROC curve, the x-axis measures

(a) True Pos Rate which we want high

(b) False Pos Rate which we want low

(c) True Neg Rate which we want high

(d) False Neg Rate which we want low

71. With model building:

1. There are many ways to find a good model.
2. There is always one right answer.
3. There is no end to the fun.
4. Can we take a pure math class yet?

72. If we consider the censored times to be death times, the empirical survival curve will (on average)

(a) underestimate the parameter

(b) overestimate the parameter

(c) sometimes under and sometimes overestimate the parameter

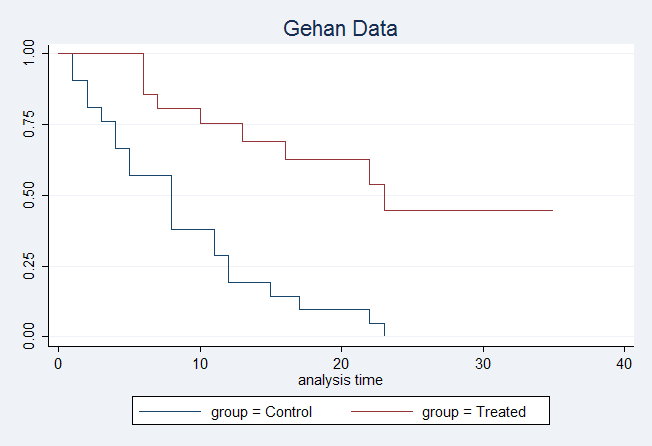
73. when:

1. there are no deaths at time
2. there is no censoring at time
3. there are no deaths at time
4. there is no censoring at time
5. there is no censoring at time

74. when:

1. there are no deaths at time
2. there is no censoring at time
3. there are no deaths at time
4. there is no censoring at time
5. there is no censoring at time

75. proportion who survive past 20 days for Treated (red)



(a) ~0.65

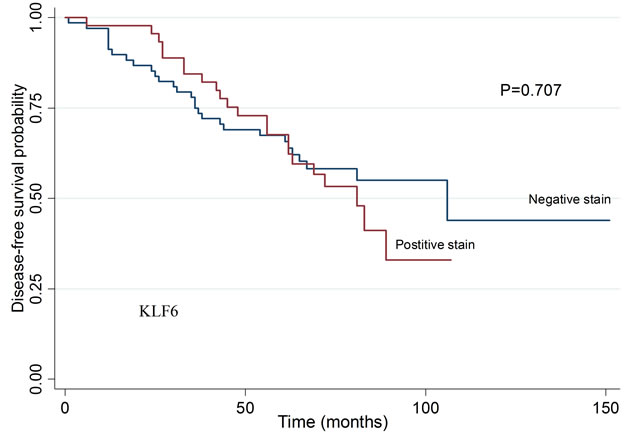
(b) ~0.35

(c) ~ 0.45

(d) we only know it’s bigger than blue

(e) we only know it’s smaller than blue

76.



(a) blue is clearly better

(b) red is clearly better

(c) can’t tell because they cross

(d) can’t tell because the p-value is big

(e) can’t tell because the p-value is small

77. The hazard at time t represents:

(a) the probability of the event

(b) the instantaneous rate of the event

(c) the relative risk of the event

(d) the odds ratio of the event

78. In A37, the last entry in the table, h(t), was NA because:

(a) the last observation was a death

(b) the last observation was censored

(c) the time interval is too big

(d) the time interval is too small

79. Censored observations are . . .?

(a) More important than non-censored ones in survival analysis

(b) Are assumed to be normally distributed over time

(c) Are assumed to have the same survival chances as uncensored observations

(d) Are essential to allow calculation of the Kaplan Meier plot

(e) Are allocated to the baseline survival curve

80. Survival Analysis: for a one unit change of an explanatory variable, the corresponding coefficient represents:

(a) baseline survival

(b) survival ratio

(c) baseline hazard

(d) hazard ratio

81. In survival analysis, the most appropriate way to interpret the value is:

(a) odds

(b) probability

(c) time to event

(d) relative risk

(e) odds ratio

82. Let the event be death. If larger values of the variable are associated with higher likelihood of *survival*, the coefficient (beta) should be

(a) bigger than 1

(b) smaller than 1

(c) positive

(d) negative

(e) zero

83. Let the event be death. If larger values of the variable are **NOT** associated with higher (or lower) likelihood of *survival*, the coefficient (beta) should be

(a) bigger than 1

(b) smaller than 1

(c) positive

(d) negative

(e) zero

84. A Cox regression analysis:

(a) Is used to analyze survival data when individuals in the study are followed for varying lengths of time.

(b) Can only be used when there are censored data

(c) Assumes that the relative hazard for a particular variable is always constant

(d) Uses the logrank statistic to compare two survival curves

(e) Relies on the assumption that the explanatory variables (covariates) in the model are normally distributed.

85. The effect of weight could violate PH if:

(a) people of different weights are in control vs treatment group

(b) people tend to weigh less over time

(c) the hazard function for weight is not monotonic

(d) the hazard function changes as a function of weight which is also changing over time

86. The effect of treatment could violate PH if:

(a) the treatment has no effect

(b) the treatment produces short term benefits only

(c) the treatment effect interacts with a different variable, like gender

(d) there is more than one treatment group

87. BP can violate the “**linear** HR” assumption if:

(a) the ln ratio of the hazard curves is not linear with respect to BP

(b) the ln ratio of the survival curves is not linear with respect to BP

(c) the effect of BP is to increase the hazard

(d) the effect of BP is to decrease the hazard

(e) there is no effect due to BP

88. AIC, BIC, model validation, and stepwise regression are methods for

(a) parameter estimation

(b) variable selection

85. If , I would expect 5% of all hypotheses to be rejected.

(a) TRUE

(b) FALSE

86. Power is:

(a) P(type I error)

(b) P(type II error)

(c) 1 – P(type I error)

(d) 1 – P(type II error)

87. The p-value is

(a) P(H0 is true | data)

(b) P(Ha is true | data)

(c) P(data | H0 is true)

(d) P(data | Ha is true)

(e) 1 – P(data | H0 is true)

RA Fisher (1929)

“… An observation is judged significant, if it would rarely have been produced, in the absence of a real cause of the kind we are seeking. It is a common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator, but it does not mean that he allows himself to be deceived once in every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation.”

88. For hypothesis testing, the problem of multiple comparisons (also known as the multiple testing problem) results from the increase in \_\_\_\_\_\_\_\_ that occurs when statistical tests are used repeatedly.

(a) Type I errors

(b) Type II errors

(c) Null hypothesis

(d) Statistical hypothesis testing

89. If H0 is true, the p-values should be distributed:

(a) Uniformly (equal prob) on 0 to 1

(b) Uniformly on -1 to 1

(c) Unimodal on 0 to 1

(d) Skewed left on 0 to 1

(e) Skewed right on 0 to 1

90. A good estimate of the number of null tests is:

(a) (# p-values > 0.5) / 2

(b) (# p-values > 0.5) \* 2

(c) (# p-values < 0.5) / 2

(d) (# p-values < 0.5) \* 2

91. What do I do if the adjusted p-value is bigger than 1?

(a) Leave it unadjusted

(b) Assign the value of the previous (“smaller”) p-value

(c) Round it to 1

(d) Divide by 2

92. With Holm’s method, what do I do if the (m+1)^th adjusted p-value is smaller than the m^th adjusted p-value?

(a) Leave it unadjusted

(b) Assign the value of the m^th adjusted p-value to the (m+1)^th adjusted p-value

(c) Round it to 1

(d) Divide by 2

93. The false discovery rate represents

(a) the proportion of true discoveries out of the total tests

(b) the proportion of true discoveries out of the total discoveries

(c) the ratio of the number of true discoveries divided by the number of null discoveries

(d) the number of null discoveries out of the total tests

(e) the number of null discoveries out of the total discoveries

94. FDR and FWER differ in that

(a) FDR is a rate and FWER is a probability

(b) FDR controls the rate of false positives

(c) FWER controls the probability of getting a false positive

(d) some of the above

(e) all of the above

­­­­­­95. Which multiple comparisons adjustment gives the highest power?

(a) Bonferonni

(b) Holm

(c) Benjamini-Hochberg

(d) Storey (q-values)

96. The linear regression technical assumptions include:

(a) The Y variable is normally distributed

(b) The X variable is normally distributed

(c) The residuals are normally distributed

(d) The slope coefficient is normally distributed

(e) The intercept coefficient is normally distributed

97. Which is bigger?

data = 1, 9, 25

1. Average(squares(data))
2. Squares(average(data))
3. Totally depends on the data
4. I have no idea

98. Which is bigger?

data = 1, 9, 25

1. Average(square root(data))
2. Square root(average(data))
3. Totally depends on the data
4. I have no idea

99. Consider the Poisson regression model:

How is interpreted?

1. A one unit change in X gives a difference change in E(Y)
2. A one unit change in X gives a difference change in E(Y)
3. A one unit change in X gives a factor change in E(Y)
4. A one unit change in X gives a factor change in E(Y)