name (required) _____

You must turn in this hard copy (with your name on it) and your scantron to receive credit for this exam.

One answer and only one answer per question. Leaving a question blank or filling in 2+ answers will be incorrect no matter what.

Where relevant, the goal is underlined. *Italicized phrases are true.* Do not assume more than is given in a question.

A = True, B = False   unless indicated otherwise. If any part of an answer is incorrect, treat all of it as incorrect.

Data Quality: Errors and fixes
(RPA = rounding, precision, accuracy; H&T = human & technical)

1-5. (2.5pts each) Which type of error is indicated in each of the following paragraphs? (One answer per question)

1. You want to know the average weight of a grain of rice. Your scale only measures to the nearest 0.1 gram, but you suspect that a grain of rice weighs closer to 0.01 gram. So you count 500 grains of rice and weigh them together, dividing by 500. What type of error is reduced by weighing all 500 at once?
        (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) None

2. In clinical trials, a placebo is used to avoid which type of error?
        (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) None

3. A firm assesses customer satisfaction of its products in two different years. The two years yield different levels of customer satisfaction, but a statistical test indicates that the difference is 'not significant.' To what type of error is the difference between samples attributed?
        (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) None

4. Gardens in East Austin yield better produce than those in West Austin. A scientist compares soil samples between the two locations to isolate the cause. Seven different soil components are tested. Each of 5 samples taken from East Austin shows higher levels of all nutrients than the 3 samples taken from West Austin, and the difference between East and West Austin samples is statistically significant. What type of error in data accounts for the consistent difference between the two locations?
        (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) None

 5. A clinical trial of a new drug is careful to ensure that neither its subjects nor its observers know which group a subject is assigned to. Unknown to them, the first subjects to enroll were assigned to the treatment group until that group filled, then the last subjects to enroll were all assigned to the control group. The study results indicated that the drug had a statistically significant effect when compared to the control group, even though in reality there is no effect of the drug. What type of error accounts here for the finding of an effect of the drug when there really is none?
        (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) None

6-9. (**6 pts**) Why do researchers use statistics to draw conclusions about their data?   **(A)** = TRUE

   6. (A)(B) Researchers usually collect data (information) about everyone/everything in the population.
   7. (A)(B) The public is easily persuaded by numbers and statistics, so researchers use statistics to impress the public.
   8. (A)(B) Correct models can only be proved through statistical analyses.
   9. (A)(B) Researchers make inferences about a population using estimates from a smaller sample.

10-13. (6 pts) Following reports that caffeine content of different batches of the same energy drink are variable, two studies estimate the average caffeine content of the same brand of energy drink. Each study uses the same type of statistical test on a random sample of the energy drink. Study 1 uses 25 bottles, and study 2 uses 100 bottles. Which statements are true? (A) = TRUE, (B) = False

   10. (A)(B) Sampling error of the mean caffeine content will be smaller in study 2 than in study 1.
   11. (A)(B) There is expected to be more bias in the caffeine average in study 2 than in study 1
   12. (A)(B) Although randomization was used in both studies, randomization was unnecessary because all samples were
          from the same brand of energy drink.
   13. (A)(B) Human and technical error is not a possibility in a study of this type because the data are not subjective.

14-17. (2 pts each) You will send pairs of tubes to a lab for analysis. For each pair of tubes, you are to decide whether replication is present, absent or unknown to you (for the characteristic indicated) and also whether it would be known to the lab receiving the samples. You know everything given in the table. The lab only knows what is written on the tube: if a tube has a person's name on it, the lab can assume that the tube contents belong to the name of the person on the label and can infer gender but nothing else. If a tube is labeled with a number, the contents are completely unknown to the lab but known to you to the extent given in the table. A question mark (?) indicates that the analysis for that sample was ambiguous. You may be able to use other information in the table to decide its property. (Gender, marker type and blood type do not change from sample to sample of the same individual, even if the assays are sometimes ambiguous.) Your options for tube contents and tube labels are:

tube	tube label	Contents are from	Gender	Blood type	Marker type
(1)	Laura Baker	Laura Baker	Female	A	negative
(2)	Oz Wichman	Oz Wichman	Male	AB	+
(3)	Rachael Springman	Rachael Springman	Female	AB	negative
(4)	#101	Harold Zakon	Male	O	?
(5)	#218	Judy Timmins	Female	?	+
(6)	#10	Pat Hines	Male	AB	+
(7)	Jerry Allison	Jerry Allison	Male	B	negative
(8)	Brenda Iverson	Brenda Iverson	Female	B	negative
(9)	Judy Timmins	Judy Timmins	Female	A	+

In the following questions, indicate which pairs of tubes (if any) satisfy the specified criteria.

(A) Absence of replication is known to you, and the lab cannot infer the absence

(B) Absence of replication is known to you and the lab <u>can</u> infer the absence

(C) Presence of replication is known to you, and the lab cannot infer the replication

(D) Presence of replication is known to you, and the lab <u>can</u> infer the replication

(E) Replication is unknown to you and unknown to the lab

.

14. (A)(B)(C)(D) (E) tubes 2 & 6 analyzed for gender
15. (A)(B)(C)(D) (E) tubes 2 & 4 analyzed for marker type
16. (A)(B)(C)(D) (E) tubes 5 and 9 analyzed for blood type
17. (A)(B)(C)(D) (E) tubes 3 & 7 analyzed for gender

20 -27. Which data features are explicitly present?

18-22. (6 pts) A politician uses the scientific method to <u>decide on the best way to get votes</u>. He has 3 styles of giving speeches and, according to his gut instincts, tests his favorite style on an audience on one day, his second favorite style on a different audience the next day, and his least favorite style on yet another audience a day later. He has his campaign manager assess the audience response by informally chatting with several members of the audience after each speech, but not telling the members why he is asking for their responses. (A) = Present   (B) = absent or not described

18. (A)(B) Replication

19. (A)(B) Standards

20. (A)(B) Randomization

21. (A)(B) Blind (at least one way)

22. (A)(B) Blind (2 ways)

23-26. (6 pts) Before subjecting your employees to drug tests, you decide to assess the accuracy of the testing lab. Following the recommendations you receive from a consulting firm, you do the following test on two separate occasions. You take a sample from yourself, split it into 3 tubes, each with completely different identifying information, and send all three tubes for testing to the same lab. (**A**) = Present   (B**) = absent or not described

23 (A) (B) Replication
24 (A) (B) Standards
25 (A) (B) Random
26 (A) (B) Blind

27-29. (6pts) You have a class of 55 students in which you want to measure averages. You are interested only in the average for your class, not a larger sample. Which types of error could be present in the average for each of the following cases? The problem will tell you what is being measured and for what fraction of the class. If the problem does not specify how the measure or sample is obtained, then you must consider all feasible possibilities for how it might be done.

(**A**) Neither sampling nor bias   (**B**) Sampling only   (**C**) Bias only   (**D**) Both Sampling and Bias could be present

    27. (A)(B)(C)(D) Body mass index (measured by machines) of the entire class.

    28. (A)(B)(C)(D) Body mass index (measured by machines) of half the class

    29. (A)(B)(C)(D) The general appearance of all students in the class, measured casually by the president of the UT business student association.

Criminal Justice (RMP is random match probability)

We mentioned 4 features of an 'ideal' forensic method for matching a suspect with a forensic sample: (i) reference database, (ii) discrete characteristics, (iii) independent verification possible, (iv) labs/experts pass blind proficiency tests (95% or better).

30. (3 pts). Which feature of an ideal method gives you the most assurance that the lab is capable of getting the right answer, no matter what the methodology may be? (One answer only)

        (A) Reference database   **(C)** Independent verification   **(E)** None

        (B) Discrete characteristics   **(D)** Pass blind proficiency tests

31-36. (9 pts) For which types of matching method was it claimed (in the book and/or lecture) that the procedure had a functional reference database?

    31. (A)(B) Fingerprinting (before 1990)   **(A) = present**   (B) = Absent

    32. (A)(B) DNA typing

    33. (A)(B) Dog sniffing

    34. (A)(B) Hair matching (not DNA based)

    35. (A)(B) Bite mark identification

    36. (A)(B) Eye witness identification

37-40 (6 pts) A match between a suspect and a crime scene sample has been declared in trial. The defense has exposed the fact that the lab responsible for the analysis erroneously claims a match 5% of the time, and this number is not in dispute. Witnesses for the defense and prosecution are arguing about the RMP, and you only know that it could range from 1 in a billion up to 1 in 100. What can you legitimately conclude about the chance of an erroneous match -- that the suspect is not the source of the crimescene sample? **(A) = TRUE** (B) = False

    37. (A)(B) The chance of an erroneous match is 6% or more (0.06 and above).

    38. (A)(B) The chance of an erroneous match is 1 in a billion or less.

    39. (A)(B) You cannot set bounds on the chance of an erroneous match until you know the RMP exactly.

    40. (A)(B) The chance of an erroneous match is 1/billion + 1/100.

41-44 (6pts). In each of the following, a property of a forensic method, underlined, is described as absent or what it might consist of. The text following the underline gives a possible example of what the underlined text describes. Which examples correctly describe the underlined text?

(A) The non-underlined text correctly describes the underlined.   **(B)** The underlined and non-underlined do not match

    41. (A)(B) <u>A proficiency test is absent</u>: a dog sniffing expert whose ability to track a person is never checked by comparing a person's known path (directly observed) to the expert's claims of what the path was.

    42. (A)(B) <u>The possibility of 'Independent verification' is absent</u>: when only one expert in a case testifies about the evidence.

    43. (A)(B) <u>A possible reference database for soil sample matching would be</u>: a collection of soil samples from different parts of the geographical region relevant to the trial.

    44. (A)(B) <u>A possible reference database for matching hair samples would be</u>: a description of the characteristics and methods used to score hair properties.

Data Presentation and Numbers

45-49. (8 pts). Based on the Data Presentation lecture and chapter, which of the following points are true?

(A) = TRUE (B) = FALSE

45 (A) (B) A drug test applied to 10000 patients uses a method that gives the correct answer 99% of the time. In the absence of any information about the population being tested, it is possible that most people who get a positive result from the test could actually be negative.

46 (A) (B) The following statement gives an <u>absolute</u> risk: A drug reduces heart attacks from 3 in 30,000 patients to 1 in 30,000 patients. The corresponding <u>relative</u> risk would be that the drug causes a 67% reduction in heart attacks.

47 (A) (B) Data presented as natural frequencies were easier for our class to understand than data presented as conditional probabilities

48 (A) (B) Our class was more impressed by data presented as relative risks than data presented as absolute risks.

49 (A) (B) Graphs changed the class perception of information merely by changing the scale on the axes.
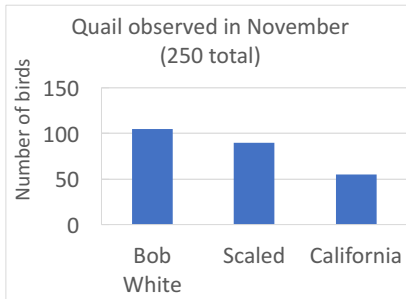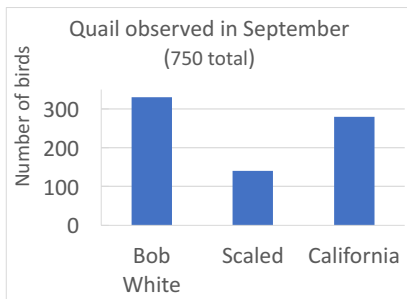
50-52 (6pts). In the figure shown, three types of quail were counted on a ranch in both September and November. All quail on the ranch were counted each month, and the monthly totals are given at the top of the figure. Which statements about these data are true? (A) = TRUE

50. (A)(B) The numbers of Scaled quail were higher in September than in November.

51. (A)(B) The proportion of Scaled quail (among all quail in the graph) was higher in November than in September.

52. (A)(B) The number of each type of quail declined from September to November.

53. (3pts) A gene test shows promising results in providing early detection for colon cancer. However, 5% of all test results on a cancer-free patient are falsely positive; that is, results indicate that cancer is present when the patient is, in fact, cancer-free. Given this false positive rate, how many people out of 10,000 cancer-free patients would have a false positive result? One answer only

    (A) 5
    (B) 50
    (C) 500
    (D) You cannot answer this question without knowing something else about the sample.

54-57. (2 pts each) We described 4 types of data/numbers regarding their similarity to what they represent. Which type of number is indicated by each question below?

    A) total counts   **B)** an extrapolation   **C)** data translations (conversions)   **D)** fabrications

54. The estimate of a population average from a sample of that population: **A B C D**

55. The basis of the claim that the lyrics to Louie Louie were obscene : **A B C D**

56. The odometer reading on your car (if not altered): **A B C D**

57. The creation of a number in the absence of data   **A B C D**

58. (4 pts.) Exam Key Code **A**: **Fill in** bubble **(A)** on question **58** to indicate your exam code; leave the other bubbles blank for this question. Also, fill in the correct <u>bubbles</u> for your name and EID on the scantron form.