60 Questions, 4 pages   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. If different parts of an option are inconsistent with each other, consider it incorrect.

Data Quality: Errors and fixes

(RPA = rounding, precision, accuracy; H&T = human & technical; standards = knowns)

1-6. (2.5pts each) Which type of error is indicated in each of the following paragraphs? (One answer per question)

1. You want to randomly choose 100 names from among all Austin residents ages 18 and older. You get a file of the City of Austin phone book (listing all residents who own a landline) and use a random number generator to choose 100 names. After ensuring that all people chosen are at least 18 years old, what type of error may cause your sample to differ from your intent?
          (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) Faulty data analysis   (F) None

2. Bull wants to send a first class letter that weighs no more than 1.0 ounce. With a stamp, the envelop weighs 0.1 ounce (oz), but he does not know how much a sheet of paper weighs. His scale only reads to the nearest 4.0 oz, and one sheet of paper weighs far less than 4 oz. He uses the scale to weigh a stack of 500 sheets and finds a combined weight of approximately 80 oz. By dividing, he calculates that a sheet of paper weighs 0.16 oz, so he can safely add up to 5 sheets of paper to the envelope. What type of error is reduced by weighing 500 sheets at once?
          (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) Faulty data analysis   (F) None

3. Anna and Jean are employees subject to occasional random drug tests. Each is told there is a 10% random chance of being chosen for a test in the week ahead. On Monday, Anna but not Jean is chosen for a test.   What type of data error explains the difference between Anna being chosen but Jean not?
          (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) Faulty data analysis   (F) None

4. A DNA analysis is done on an adult woman and a fetus to decide if the woman might be the mother. A DNA profile from each is provided by one lab; the profiles are not identical, but they have similarities. Two experts do calculations from those profiles to decide whether the similarities are meaningful. One expert testifies that there is only a 1/300 chance that a non-mother would match so closely, but another expert testifies that chance is 1/15. Both cannot be right. What type of error accounts for the difference between the experts?
          (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) Faulty data analysis   (F) None

 5. What type of error is prone to arise if a machine is used to take measurements but the machine is not tested against standards/knowns?
          (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) Faulty data analysis   (F) None

6. What type of error may plausibly arise if a teacher grades essays while knowing the identities of each student?
          (A) Sampling   (B) Bias   (C) H&T   (D) RPA   (E) Faulty data analysis   (F) None

7-10. (6 pts) Following reports that average caffeine content of the energy drink *Black Bull* is not as advertised, two studies estimate the average caffeine content of all *Black Bull* bottles sold. Study S uses 25 bottles randomly chosen from a bottling plant in Mason, TX. Study L uses 100 bottles randomly chosen from a bottling plant in Amarillo, TX. (*Each plant produces thousands of bottles, and the caffeine content is not identical between bottles within the same plant.*) Each study measures caffeine of each bottle with a machine and uses the same statistical test on its respective sample; the two averages are compared. Which are true? **(A)** = TRUE, (B) = False

   7. (A)(B) Assuming no consistent differences between bottles from Mason and bottles from Amarillo, sampling error of the average caffeine content will be **smaller** in study S than in L.
   8. (A)(B) Randomization removes any concern that a consistent difference between S and L could affect the averages.
   9. (A)(B) RPA error would not exist in this study because caffeine is uniquely identifiable and cannot be confused with other ingredients in the drink
   10. (A)(B) Human and technical error is not a possibility because caffeine contents were measured on machines.

11-14. (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 for the characteristic indicated is present, absent or unknown to you 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 state of that particular sample is unknown to you. 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 -- what you and the lab each see	Contents are from – what only you see	Gender	Blood type	Marker type
(1)	Laura Baker	Laura Baker	Female	AB	negative
(2)	Oz Wichman	Oz Wichman	Male	A	+
(3)	Rachael Springman	Rachael Springman	Female	O	negative
(4)	#101	Harold Zakon	Male	O	?
(5)	#218	Pam Hines	Female	B	+
(6)	#10	Jules Timmins	Male	AB	+
(7)	Jerry Allison	Jerry Allison	Male	B	negative
(8)	Brenda Iverson	Brenda Iverson	Female	B	negative
(9)	Harold Zakon	Harold Zakon	Male	?	+
(10)	#719	Oz Wichman	Male	?	+

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 can 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 can infer the replication

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

11. (A)(B)(C)(D) (E) tubes 2 & 10 analyzed for blood type
12. (A)(B)(C)(D) (E) tubes 9 & 10 analyzed for blood type
13. (A)(B)(C)(D) (E) tubes 5 and 9 analyzed for gender
14. (A)(B)(C)(D) (E) tubes 3 & 7 analyzed for marker type

15 -23. Which data features are explicitly present?

15-19. (6 pts) Neal Caffrey is attempting to find the best ways to get donor support for crowdfunding projects. He invents 4 different project designs, each with different emotional appeals and web infrastructure. He has only one web site, so he chooses one design by drawing a card from a shuffled deck, runs the design for a week on the site and counts the donations. Each week, he chooses again the same way and counts donations, until a year is up. The donors are unaware of participating in his study.

. **(A) = Present** (B) = absent or not described

15. (A)(B) Replication	18. (A)(B) Blind (at least one way)
16. (A)(B) Standards	19. (A)(B) Blind (2 ways)
17. (A)(B) Randomization	

20-23. (6 pts) Jimmy is trying to find a treatment that will make unsightly mosquito bites vanish from his body. He tests 2 commercial remedies (cortisone cream, calamine lotion), each on different bites. He has 16 bites, so he applies a remedy to one bite, then applies the other remedy to the next bite, then goes back to the first remedy for the 3rd bite, alternating one after another. He takes pictures of each bite before treatment and a day after treatment and then does a comparison of the pictures. He marks the back of the pictures so he does not know whether a picture is 'before' or 'after' treatment. He does his comparison of before and after, then asks a friend to make the same comparison (without telling them his scores), so he can decide if his scoring is legitimate.

(**A**) = Present (B**)** = absent or not described

20. (A)(B) Replication	22. (A)(B) Randomization
21. (A)(B) Standards	23. (A)(B) Blind (at least one way)

24-26. (5 pts) A shopper wants to know the brand of popcorn that <u>gives the most kernels per jar on average</u>. She buys 3 seemingly equivalent jars of *Redenbacher* and 3 of *Hill Country Fare* brand. She and her two daughters separately count the numbers of kernels in each jar until everyone agrees. The numbers are 1213, 1179, 1205 for *HCF* and 1007, 983, 1013 for *Redenbacher*.   Which options are true about inferences you can draw?

(A) = true, (B) = false

24. (A) (B) The averages of each brand calculated from these data are affected by sampling error.
25. (A) (B) These counts are subject to RPA error.
26. (A) (B) The counts should have been done blindly to avoid bias.

27-30 (7 pts) A study on life expectancy used a random sample of 1,000 residents from the US. In this sample, the average life expectancy was 80.1 years for females and 74.9 years for males. Which are ways to <u>increase your understanding</u> of whether, for the total US population, women actually do live longer than men?

27. (A) (B) These averages indicate that women do, on average, live longer than men in the total US population. No further study is warranted.
28. (A) (B) Conduct a statistical analysis to determine if females live significantly longer than males.
29. (A) (B) Graph the mean (average) life expectancy values of females and males and visually analyze the data.
30. (A) (B) Analyze a second random sample and see if it also shows women living longer.

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.

31. (3 pts). Which feature of an ideal method is an assurance that the match can, in principle, be replicated by different labs? (One answer only)

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

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

32-37. (6 pts) For which types of matching method was it claimed (in the book and/or lecture) that independent verification was possible for the procedure?

32. (A)(B) Fingerprinting (before 1990)   **(A) = present**   (B) = Absent
33. (A)(B) DNA typing
34. (A)(B) Dog sniffing
35. (A)(B) Hair matching (not DNA based)
36. (A)(B) Bite mark identification
37. (A)(B) Eye witness identification

38-41 (6 pts) A match between a suspect and a crime scene sample has been declared in trial. The lab claims the RMP is <u>1/thousand</u>. However, the defense has exposed the fact that the officers handling the samples mix them up 2% of the time, and then, even when the lab receives the right samples, 1/10 of the time it falsely claims a match that doesn't exist. What can you correctly conclude about the chance that the suspect is not the source of the crimescene sample (is a false match)?

(A) = TRUE (B) = False

38. (A)(B) A change in the RMP to 1/million would greatly reduce the overall chance of a false match.
39. (A)(B) The chance of a false match cannot be less than 2%
40. (A)(B) The chance of false match cannot be determined by combining error rates from 3 completely different sources
41. (A)(B) The chance of a false match Is the product of the 3 error probabilities

42-47 (10 pts). Which of the following outcomes in a forensic method is possible when or because procedures are not blind?

42. (A)(B) A technician deliberately falsifyies results to incriminate a suspect
43. (A)(B) Unintentional mis-calculation of the RMP because of inadequate sampling
44. (A)(B) Unconscious interpretation of subjective data to 'fit' a suspect
45. (A)(B) Unconscious direction of an eyewitness toward a particular suspect
46. (A)(B) Subjective interpretation of discrete characteristics
47. (A)(B) Inaccurate estimation of a lab error rate becaue the lab knows when proficiency tests are conducted

48-51 (6pts). In each of the following, a property of a forensic method is said to be satisfied or not (as <u>underlined</u>). The text following the underline then gives an example possibly compatible with the underlined text . For which questions does the example correctly fit the underlined text?

(A) The non-underlined text correctly matches the <u>underlined</u>.   **(B)** The underlined and non-underlined do not match

48. **(A)(B)** <u>A proficiency test would be satisfied by</u>: a bite-mark expert has convinced each and every jury in over 100 trials to convict based on his testimonies.

49. **(A)(B)** <u>The possibility of 'Independent verification' would **Not** be satisfied when</u>: only one laboratory in the world is capable of the analysis.

50. **(A)(B)** <u>A reference database for bullet lead analysis would be satisfied by</u>: the data from analyses of lead from 10,000 different bullets from across the country.

51. **(A)(B)** <u>A reference database for matching hair samples between people would be satisfied by</u>: the data from characterizations of 1,000 different hairs from the suspect's head in the case being tried.

Data Presentation

52-56. 10 pts). Based on the Data Presentation lecture and chapter, which of the following points are true?

(A) = TRUE (B) = FALSE in at least one respect

52 (A) (B) A drug test applied to 10,050 patients gives the correct answer 99% of the time (without respect to whether the individual is truly positive or negative). If the sample has 50 patients who are truly positive and 10,000 who are truly negative, most individuals who test positive will actually be positive.

53 (A) (B) The following statement gives an <u>absolute</u> risk: A drug reduces heart attacks from 30 in 30,000 patients to 10 in 30,000 patients. The corresponding <u>relative</u> risk would be that the drug causes a 33% reduction in heart attacks.

54 (A) (B) Data on false positive rates were easier for our class to understand when presented as natural frequencies than when presented as conditional probabilities.

55 (A) (B) There are cases where the same data can be presented as proportions versus numbers and create *opposite* impressions.

56 (A) (B) For figures shown in lecture, merely changing the scale on the vertical axis changed the class perception of data.
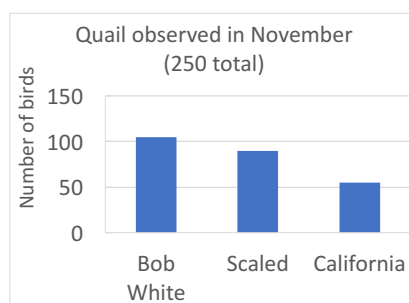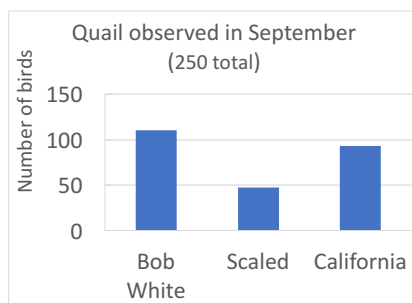
57-59 (6pts). In the figure shown, three types of quail were counted on a ranch in both September and November. All quail on the ranch were counted for both months, and the monthly totals are given at the top of the figure. Which statements about these data are true? (A) = TRUE (B) = false

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

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

59. **(A)(B)** The combined total number of the 3 types of quail on the ranch increased from September to November.

60. (4 pts.) Exam Key Code **A**: **Fill in** bubble **(A)** on question **60** 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.