University of Guelph
Department of Mathematics and Statistics

STAT*2040
Statistics I

Test 1 (White version)

Solutions

February 11 2017

Examiner: Jeremy Balka

This exam is 70 minutes in duration

Name:

ID:

Signature:

Please read the instructions:

1. Fill out your name and ID number above.
2. When the examination starts, make sure your question paper is complete. You should have 19 multiple choice questions, along with a formula sheet. The first question is just a bookkeeping question, and does not count for marks, but please fill it in to ensure your exam is properly graded.
3. Do all rough work on this paper.
4. You are allowed to bring in a calculator, and pens and pencils.
5. There is only **one** correct answer for each question. Fill in only one bubble for each question.
6. Fill out the computer answer sheet in pencil as you go. *There will be no extra time given at the end of the exam to fill in the sheet.*
7. The answers given in the exam are often rounded versions of the correct answer. Choose the closest value.

1. The colour of the first page of this examination booklet (the cover sheet) is:

   (a) White **** (Not worth marks)
   (b) Yellow

2. Which one of the following statements is FALSE?

   (a) A statistic is a numerical characteristic of a sample.
   (b) A parameter is a numerical characteristic of a population.
   (c) In practical problems, the value of a parameter is usually unknown.
   (d) We often use statistics to estimate parameters.
   (e) *The value of the population mean depends on the sample size.* ***

   The value of the population mean (a parameter) is a characteristic of the entire population, and does not depend on the sample size.

3. Which one of the following statements is TRUE?

   (a) Well designed observational studies often give strong evidence of a causal link between the explanatory and response variables.
   (b) Lurking variables are more of a problem in experiments than observational studies.
   (c) *Confounding can occur in both observational studies and experiments.* ***
   (d) Randomization plays an important role in observational studies, but not in experiments.
   (e) Observational studies are outdated, and are not used in modern scientific studies.

   Confounding (where it is impossible to separate the effect of effects of explanatory variables on the response), can occur in both observational studies and experiments.

4. Suppose $P(A) = 0.4$, $P(B) = 0.6$, and $P(A|B) = 0.5$. What is $P(A \cap B^c)$?

   (a) 0.1 ***
   (b) 0.2
   (c) 0.3
   (d) 0.4
   (e) 0.5

   $P(A|B) = \frac{P(A \cap B)}{P(B)} = 0.5 \implies P(A \cap B) = 0.5 \cdot 0.6 = 0.3.$
   Draw a Venn diagram to illustrate that:
   $P(A \cap B^c) = P(A) - P(A \cap B) = 0.4 - 0.3 = 0.1.$

5. Santori et al. (2014) investigated various characteristics of swimming in different species of semi-aquatic water rat. The study compared various swimming characteristics of 4 species of rat, but here we will look only at *Nectomys rattus* and *Nectomys squamipes*. In one part of the study, the swimming speed of the rats was recorded. Swimming speeds for 14 *N. rattus* and 15 *N. squamipes* rats are illustrated in Figure 1.

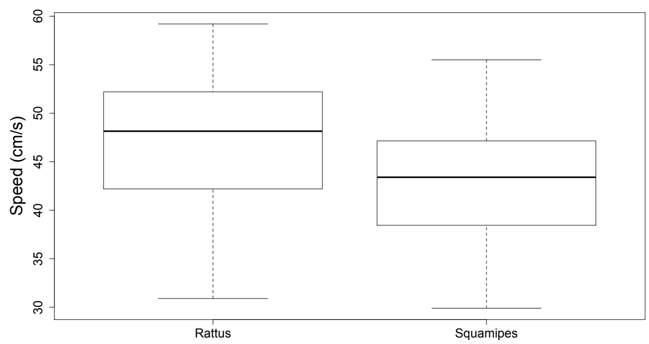

Figure 1: Boxplots of swimming speed (cm/s).

For the sample data used to create these boxplots, which one of the following statement is TRUE?

(a) The median speed for *N. rattus* is less than the median swimming speed for *N. squamipes*.
(b) Both distributions show right skewness.
(c) The IQR for *N. squamipes* is greater than the range for *N. rattus*.
(d) Q_1 for *N. rattus* is less than Q_3 for *N. squamipes*. ****
(e) None of the above.

6. Professors Jefferson and Akintola team teach a course that has an enrolment of 20 students. They want to draw a sample of 10 students from this class, but disagree about how to go about it. Professor Jefferson thinks they should put the class list in alphabetical order, then randomly determine whether to pick the first 10 on the list or the last 10 on the list by tossing a coin. Professor Akintola thinks they should put the student names on pieces of paper in a hat, mix them up, then draw 10 names randomly without replacement.

Which one of the following statements is FALSE? (If A-D are all true, choose option E.)

(a) In either sampling design, each student in the class has a 0.5 probability of being selected in the sample.
(b) Professor Akintola's method would result in a simple random sample.
(c) Professor Jefferson's method would result in a simple random sample. ***
(d) If professor Akintola's method is used, there are more than 100,000 possible samples.
(e) None of the above.

Professor Akintola's method results in one of $\binom{20}{10}$ *possible samples (where each of these is equally likely). Any possible sample of size* $n = 10$ *has the same chance of occurring under his plan, and that is in keeping with simple random sampling. Professor Jefferson's method results in one of only 2 possible samples, and not every sample of size* $n = 10$ *is even possible.*

7. Consider the following sample of 5 observations: 2.0, 3.0, 4.0, 5.0, 7.0.

For this sample data, order the following statistics from the one with the smallest value to the largest:

mean, median, variance, standard deviation.

Which one of the following is the correct ordering?

(a) standard deviation < variance < median < mean ****
(b) standard deviation < variance < mean < median
(c) mean < median < standard deviation < variance
(d) mean < median < variance < standard deviation.
(e) mean < standard deviation < variance < median.

$\bar{x} = 4.2$, $Median = 4$, $s^2 = 3.7$, $s = \sqrt{3.7} = 1.923538$

8. Suppose that in a shipment of 50 cartons of milk, 15 of the cartons contain spoiled milk, and the remaining 35 are unspoiled. If 6 containers are randomly selected without replacement, what is the probability that no more than 1 contains spoiled milk? (Choose the closest value.)

(a) 0.16
(b) 0.31
(c) 0.32
(d) 0.41 ****
(e) 0.42

Let X represent the number of cartons that contain spoiled milk. Then X has the hypergeometric distribution, and:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{\binom{15}{0}\binom{35}{6}}{\binom{50}{6}} + \frac{\binom{15}{1}\binom{35}{5}}{\binom{50}{6}} = 0.4085811$$

9. Suppose an ordinary six-sided die is rolled 10 times, and the number that comes up on the top face is recorded. What is the probability that the first roll is a six, the second roll is a six, and there is a total of 3 sixes rolled in the 10 rolls? (Choose the closest value.)

(a) 0.01 ***
(b) 0.03
(c) 0.05
(d) 0.10
(e) 0.16

Under almost all practical circumstances, it's reasonable to think that the die rolls are independent. We need the first two roles to come up with a six: $P(\text{First two rolls come up six}) = (\frac{1}{6})(\frac{1}{6})$, and then we need one six in the following 8 rolls: $P(X = 1) = \binom{8}{1}(\frac{1}{6})^1(1 - \frac{1}{6})^7$, so the probability the first two rolls are sixes and there are three sixes in total is:

$$(\frac{1}{6})(\frac{1}{6}) \times \binom{8}{1}(\frac{1}{6})^1(1 - \frac{1}{6})^7 = 0.01033636$$

10. Which one of the following statements is TRUE?

    (a) For right-skewed distributions, the median is greater than the mean.
    (b) For symmetric distributions, the standard deviation and mean are equal.
    (c) For mound-shaped distributions, approximately 50% of observations lie within 5 standard deviations of the mean.
    (d) For any distribution, the interquartile range is equal to 50.
    (e) If the range of a data set is equal to 0, then the variance and IQR will also equal 0. ****

    If Range $= Max - Min = 0$, this implies that $Max = Min$, which implies all the observations have the same value. If all the observations are equal, then $s^2 = 0$, and $IQR = Q_3 - Q_1 = 0$.

11. Suppose a sample has the following characteristics: $n = 12$, $\bar{x} = 10$, $s = 5$.

    Which one of the following statements is TRUE?

    (a) The median must be less than 10.
    (b) The z-score of the smallest observation would be positive.
    (c) The 90th percentile would be negative.
    (d) If every value in the sample was multiplied by 2, and the sample variance was calculated for these new values, the sample variance would be 100. **
    (e) None of the above.

    If all observations are multiplied by a constant, the variance gets multiplied by the square of that constant:

    $$s^2_{new} = 2^2 s^2 = 4 \cdot 5^2 = 100.$$

12. Which one of the following statements is TRUE?

    (a) A Poisson random variable can take on negative values.
    (b) For any binomial random variable X, $P(X = 0) = 0$.
    (c) The mean and variance of a Poisson random variable are always equal. **
    (d) If X has a geometric distribution with $p > 0$, then $P(X = 1) > P(X = 2)$. **
    (e) The mean of a binomial random variable can be negative.

    I inadvertently put two true answer options. If you answered either of the correct answer options, you got the mark.

    For a Poisson random variable, $\mu = \sigma^2 = \lambda$.

    For a geometric random variable, $P(X = x) = p(1-p)^{x-1}$, and so $P(X = 1) = p$ and $P(X = 2) = p(1-p)$.

13. The following table is loosely based on a recent poll in the U.S., in which 1000 randomly selected adults were asked if they approved or disapproved of the way Donald Trump is handling his job as president.

	Republicans	Democrats	Independents
Approve	300	40	100
Disapprove	30	360	170

Suppose one of these 1000 people is randomly selected. What is the conditional probability they approve, given they are a Democrat? (Choose the closest value.)

(a) 0.10 ***
(b) 0.20
(c) 0.30
(d) 0.40
(e) 0.50

$P(\text{Approve} \mid \text{Democrat} = \frac{40}{360+40}) = 0.1.$

14. Which one of the following statements is FALSE? (If A-D are all true, answer option E.)

(a) If $P(A) = 0$, and B is any event, then A and B are independent.
(b) If $P(A) = 0.5$, and B is any event such that $P(B) = 0.5$, then A and B are independent. ***
(c) If $P(A) > 0$, $P(B) > 0$, and A and B are mutually exclusive, then A and B are dependent.
(d) If $P(A|B) = P(A)$, then $P(A \cap B) = P(A)P(B)$.
(e) None of the above.

15. Approximately 20% of the adult Canadian population are smokers. Suppose adult Canadians are randomly sampled. What is the probability that the first smoker occurs on the 5th or 6th person sampled? (Choose the closest value.)

(a) 0.10
(b) 0.15 **
(c) 0.20
(d) 0.25
(e) 0.30

If we let X represent the number of people sampled required to get first smoker, then X has a geometric distribution with $p = 0.2$, and:

$$P(X = 5) + P(X = 6) = p(1-p)^4 + p(1-p)^5 = 0.147456$$

16. Of the following situations, which one would best be investigated with an experiment, as opposed to a survey or other type of observational study?

(a) A study of a possible relationship between heroine use and marital infidelity.
(b) A study of a possible relationship between maternal cocaine use during pregnancy and birth weight of the baby.

(c) A study of a possible relationship between non-prescription adderall use and cocaine use in undergraduate university students.
(d) A study investigating a possible relationship between binge-drinking and eating disorders.
(e) A study investigating a possible effect of a new high-protein diet on weight loss. ***

17. A dentist has 3 procedures to perform before she can leave for the day. The first procedure has a mean time to completion of 35 minutes, with a standard deviation of 4 minutes. The second procedure has a mean time to completion of 25 minutes, with a standard deviation of 5 minutes. The third procedure has a mean time to completion of 75 minutes, with a standard deviation of 10 minutes. The procedures must be completed sequentially, and they can be considered independent. What is the standard deviation of the total time to complete the 3 procedures? (Choose the closest value.)

    (a) 10 minutes
    (b) 12 minutes **
    (c) 14 minutes
    (d) 17 minutes
    (e) 19 minutes

    When random variables are independent, the variance of their sum is the sum of their variances, so the variance of the total time is:

    $Var(\text{Total time}) = 4^2 + 5^2 + 10^2 = 141$, and the standard deviation is $\sqrt{141} = 11.87434$ minutes.

18. In which one of the following situations would the Poisson distribution provide the most reasonable approximation to the binomial distribution?

    (a) X has a binomial distribution with $n = 5$ and $p = 0.5$.
    (b) X has a binomial distribution with $n = 10,000$ and $p = 0.9$.
    (c) X has a binomial distribution with $n = 10$ and $p = 0.2$.
    (d) X has a binomial distribution with $n = 500$ and $p = 0.01$. ***
    (e) X has a binomial distribution with $n = 2$ and $p = 0.1$.

    The Poisson approximation to the binomial works best when p is very small.

19. Consider the following probability distribution of a random variable X. (The question mark represents a missing probability, but you should be able to figure out what it is.)

x	20	30	40	50
$p(x)$	0.3	0.3	0.3	?

    What is the standard deviation of the random variable X?

    (a) 9.8 ***
    (b) 10.2
    (c) 12.9
    (d) 14.1
    (e) 15.7

Since the probabilities must sum to 1, the missing probability must be 0.1.

$\sigma^2 = E(X^2) - [E(X)]^2$, where

$E(X) = 20 \cdot 0.3 + 30 \cdot 0.3 + 40 \cdot 0.3 + 50 \cdot 0.1 = 32$

$E(X^2) = 20^2 \cdot 0.3 + 30^2 \cdot 0.3 + 40^2 \cdot 0.3 + 50^2 \cdot 0.1 = 1120$

and so $\sigma^2 = E(X^2) - [E(X)]^2 = 1120 - 32^2 = 96$ and $\sigma = \sqrt{96}$.

Sample variance: $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1}$. Equivalent alternative formula: $s^2 = \dfrac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$

Sample z-score for the ith observation: $z_i = \dfrac{x_i - \bar{x}}{s}$

If we transform the data using the linear transformation $x^* = a + bx$, then:

$$\bar{x}^* = a + b\bar{x}, s_{x^*} = |b|s_x, s_{x^*}^2 = b^2 s_x^2$$

Probability

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$.

$P(A|B) = \dfrac{P(A \cap B)}{P(B)}$.

Two events A and B are independent if and only if:

$P(A \cap B) = P(A) \cdot P(B), P(A|B) = P(A), P(B|A) = P(B)$.

The Expected Value and Variance of Discrete Random Variables

$E(X) = \mu = \sum xp(x)$.

$\sigma^2 = E[(X-\mu)^2] = \sum (x-\mu)^2 p(x)$.

A handy relationship: $E[(X-\mu)^2] = E(X^2) - [E(X)]^2$.

Properties of Expectation and Variance

$E(a + bX) = a + bE(X)$, $\sigma_{a+bX}^2 = b^2\sigma_X^2$, $\sigma_{a+bX} = |b|\sigma_X$

If X and Y are both random variables then $E(X+Y) = E(X) + E(Y)$ and $E(X-Y) = E(X) - E(Y)$.

If X and Y are independent: $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$ and $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$

Discrete Probablity Distributions

Binomial distribution: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. $\mu = np, \sigma^2 = np(1-p)$.

Hypergeometric distribution: $P(X = x) = \dfrac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$. $\mu = n\frac{a}{N}$.

Poisson distribution: $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}, \lambda = \mu = \sigma^2$.

Geometric distribution: $P(X = x) = (1-p)^{x-1}p$. $\mu = \frac{1}{p}$, $\sigma^2 = \frac{1-p}{p^2}$.