Name: KEY

Math 10 Midterm 2
May 12, 2010

Read all instructions carefully. Calculators *are* allowed, but you may also leave answers unevaluated provided you do not need the decimal for additional work; e.g., $6(\frac{1}{2})^5$. This is a closed book exam and no notes are allowed; the table of areas under the normal curve from the book is provided at the end of the exam. You are not to provide or receive help from any outside source during the exam except that you may ask the instructor for clarification of a problem. You have two hours and you should attempt all 10 problems. There will be partial credit, so show all work.

FERPA RELEASE: Because of privacy concerns, we are not allowed to return your graded exams in lecture without your permission. If you wish us to return your exam in lecture, please sign on the line indicated below. Otherwise, you will have to pick your exam up in your instructor's office after the exams have been returned in lecture.

SIGN HERE: _____

(1) (8 pts) Consider a long sequence of die rolls.

6 (a) As the number of rolls increases, what happens to the probability of each of the following events? [Clearly you will not be able to give specific numerical answers; just explain.]

(i) The number of 6s rolled is up to 5 rolls off from $\frac{1}{6}$ of the total number of rolls.

This probability decreases. The numerical range of common outcomes increases as the number of rolls increases, so eventually ± 5 is just a small portion of that.

(ii) The number of 6s rolled is up to 5% of the total number of rolls off from $\frac{1}{6}$ of the total number of rolls.

This probability increases. The percentage range of common outcomes decreases as the number of rolls increases, and eventually falls well within $\pm 5\%$.

2 (b) What statistical rule summarizes the results of part (a)?

The law of Averages.

(2) (12 pts) If the correlation between tadpole tail length and adult frog body length is 0.5 for a particular frog species and the data is homoscedastic, what percentage of the tadpoles with tail length one standard deviation above the mean do you expect to have adult body length at least one standard deviation above the mean (here meaning the SD and mean of *all* the frogs' body lengths)?

Let x be tadpole tail length and y adult frog body length.

We may keep everything in standard units (making up units is also okay; this does not depend on the specific values) so $\bar{x} = \bar{y} = 0$, $SD_x = SD_y = 1$; $r = 1/2$.

Average y associated with $x = 1$ is 0.5 $(= 1/2 \cdot 1)$

rms error is $1\sqrt{1-(.5)^2} = \sqrt{.75} = .866$

want the percentage for $y = 1$.

standard units within this strip:

$$\frac{1 - 0.5}{0.866} = .577 \approx .60$$

area $-.6$ to $.6$ by table: 45.15

area above $.6$: $\frac{100 - 45.15}{2} = 27.4$ 27.4%

or

round down to $z = .55$

area $\frac{100 - 41.77}{2} = 29.1$ 29.1%

(3) Letting x measure the height of Munchkins (in inches) and y measure the curliness of the tips of their shoes (in number of complete circles), we have the following homoscedastic data summary:

$$\bar{x} = 38 \quad \text{SD}_x = 4$$
$$\bar{y} = 1 \quad \text{SD}_y = 0.5 \quad r = -0.4$$

(a) (4 pts) What percentage of Munchkins have two or more complete circles on their shoes?

$2 = 1 + (0.5)(2)$, so 2 SDs above the mean

2.5 % (or using the table for $z = 2$, 2.275 %)

(b) (7 pts) What percentage of 30-inch-tall Munchkins have two or more complete circles on their shoes?

$30 = 38 - 2 \cdot 4$, so x is -2 SDs

average y paired to $x = 30$ is $1 + (-0.4)(-2)(0.5) = 1.4$

($\bar{y}$, r, z_x, SD_y)

SD inside strip = rms error $= 0.5\sqrt{1-(-0.4)^2} = 0.46$

$y = 2$ in new std. units: $\frac{2-1.4}{.46} = 1.3$

area -1.3 to 1.3 by table: 80.64

area above $1.3 = \frac{100 - 80.64}{2} = 9.68$ 9.7%

or

$y - 1 = \frac{(-.4)(0.5)}{4}(x-38)$ $x = 30$, $x - 38 = -8$

$y = 1 + (0.4)(0.5)(2)$

$= 1.4$

(c) (7 pts) What percentage of 40-inch-tall Munchkins have from one half to one complete circle on their shoes?

$40 = 38 + \frac{1}{2} \cdot 4$ x is ½ SD above the mean

avg y associated with $x = 40$: $1 + (-.4)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = 0.9$

SD = rms error = 0.46 again.

½ to 1 circle in std. units by new conversion!

$$\frac{\frac{1}{2} - .9}{.46} = -.87 \approx -.85$$

$$\frac{1 - .9}{.46} = .217 \approx .20$$

area $-.85$ to $.85$ by table: 60.47

$-.20$ to $.20$ 15.85

area $.85$ to $.2 = \frac{60.47}{2} + \frac{15.85}{2} = 38.16$

38.16 %

or again,

$$y - 1 = \frac{(-.4)(.5)}{4}(40 - 38)$$

$$y = 1 + \frac{(-.4)(.5)}{2} = 0.9$$

(4) A bin contains 100 balls. 20 of them have the number 9 on them, 30 have the number 4, and the remaining 50 have the number 2. Balls are selected at random, with replacement, and the numbers are recorded, summing successive draws.

(a) (4 pts) Find the mean and standard deviation of the bin. Round to one decimal place.

mean $\frac{20\cdot 9 + 30\cdot 4 + 50\cdot 2}{100} = \frac{400}{100} = 4$

SD $\sqrt{\frac{20(9-4)^2 + 30(4-4)^2 + 50(2-4)^2}{100}}$

$= \sqrt{\frac{500 + 200}{100}} = \sqrt{7} \approx 2.6$

(b) (4 pts) If 100 draws are made, estimate the percentage of sums between 387 and 413.

EV $= 100\cdot 4 = 400$

SE $= 10\cdot 2.6 = 26$

$400 - 387 = 13$ ½ SE below

$413 - 400 = 13$ ½ SE above

area −.5 to .5 under normal curve

$= 38.29$

38.29%

(c) (4 pts) If 400 draws are made, estimate the percentage of sums between 1574 and 1626.

$EV = 400 \cdot 4 = 1600$

$SE = 20 \cdot 2.6 = 52$

$1600 - 1574 = 26$

$1626 - 1600 = 26$ again, $\pm \frac{1}{2}$ SE

38.29%

(d) (3 pts) Which of your estimates, (b) or (c), do you expect to be more accurate, if either? Why?

c should be more accurate because as the number of draws increases the probability distribution is closer and closer to the actual normal curve (this is the central limit theorem).

(5) (11 pts) You are rolling a pair of dice and collecting tokens that you can later turn in for prizes. If the sum of the dice is 11 or more, you get 2 tokens. If the sum is 3 or less, you get 1 token. Otherwise you get no tokens.

(a) How many tokens do you expect to have after 24 rolls?

mean of one roll: $2(\text{prob } 11\text{-}12) + 1(\text{prob } 2\text{-}3) + 0$

prob 11-12 $= \frac{3}{36}$ (5/6, 6/5, 6/6)

prob 2-3 $= \frac{3}{36}$ (1/1, 1/2, 2/1)

mean $= \frac{6}{36} + \frac{3}{36} = \frac{9}{36} = \frac{1}{4}$ EV for 24 rolls: $24 \cdot \frac{1}{4} = \underline{\underline{6}}$

(b) In addition to token collection, there are the following three special prizes. You are allowed to decide whether you want to compete in a 12-round game or a 24-round game. For each prize, say whether your chances are better to win it after 12 rolls or after 24 rolls.

(i) A prize for having no tokens at all. : never getting roll 2, 3, 11, 12

12 rolls

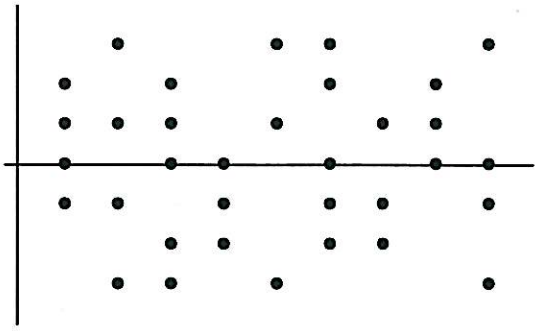
(ii) A prize for having the maximum possible number of tokens after the given number of rolls. : always getting 11, 12

12 rolls

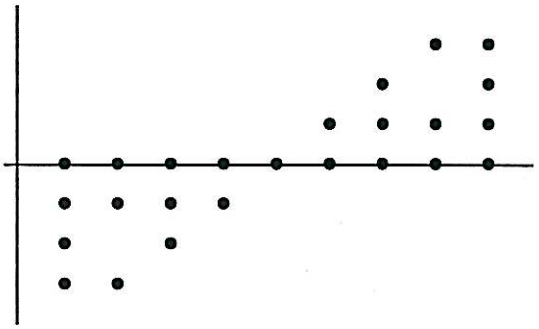
(iii) A prize for having exactly the average number of tokens after the given number of rolls. : error 0; exact # of tokens

12 rolls

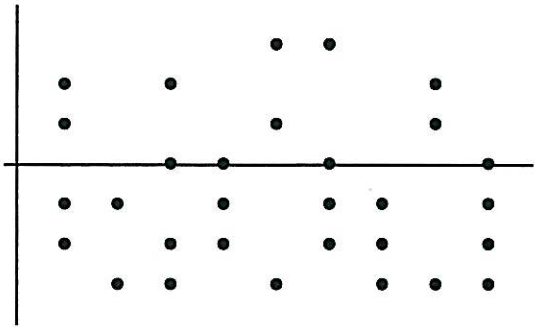
(6) (8 pts) For each of the following graphs, say whether it *could* be the graph of residuals for some scatter plot. For those that cannot be graphs of residuals, say why not, and for those that can, say whether or not they indicate linear regression was appropriate to apply.



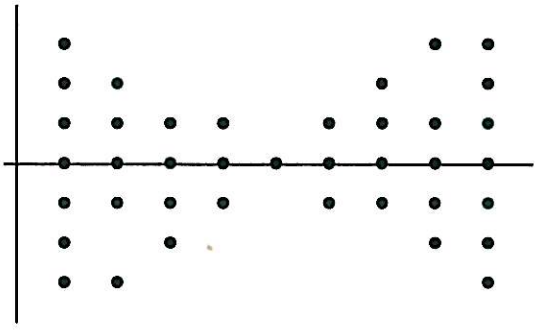
Yes & linear regression appears appropriate



No! the regression line for this plot is not the x-axis



No! the y-mean for this plot is negative rather than 0 (also leads to the regression line for this plot not being the x-axis)



Yes but the strong left to right pattern implies linear regression was not appropriate.

(7) (14 pts) Consider the following game: A wheel that is $\frac{1}{3}$ blue and $\frac{2}{3}$ red is spun. If the blue portion comes up, the player wins \$3. If the red portion comes up, the player loses \$2.

Three computer programs, labeled A, B, and C, are written to simulate 9 rounds of that game and print out the running total at that point. Each program is run ten times and the results are listed below. Do any of the output lists lead you to suspect the corresponding program is buggy? If so, what appears to be wrong with the output?

A	B	C
5	−8	−3
−11	2	−4
4	−3	−2
−9	−1	0
−5	3	−5
1	−4	−3
−2	−8	−1
−4	5	−6
0	−8	−5
−11	6	−1

box model: | 3 | -2 | -2 |

mean: $\frac{3-4}{3} = -\frac{1}{3}$

SD $= |3-(-2)|\sqrt{\frac{1}{3}\cdot\frac{2}{3}} = \frac{5\sqrt{2}}{3} = 2.36$

EV for 9 rounds: −3

SE for 9 rounds: 7.1

averages

A: −3.2 ok

B: −1.6 ← worth checking for bugs; not very close to −3

C: −3 ok

Spreads (for A & C)

SE predicts ~68% (~7 of 10) outcomes −10 to +4

A: outcomes −10 to 4? 7 ok

C: 10 hmm. In fact all C outcomes are −6 to 0, very tightly clustered. Check it for bugs.

(8) (4 pts) You have a data set of the brightness of fireflies in certain conditions. As it turns out, a given firefly is always 5% brighter when the environment is 80 degrees Fahrenheit than when it is 90 degrees. Given that information can you compute r for the data set whose variables are brightness measurements in 80- and 90-degree environments? If so, compute r, and if not, say what you are missing.

If x = brightness at 90°
y = brightness at 80°
then $y = 1.05x$ for all (x, y)

x & y are equal in standard units, so $r = 1$

(9) (5 pts) The correlation between variables A and B is 0.6. The correlation for variables A and C is -0.7. Which of B or C will give more accurate estimates for A via linear regression, and how can you quantify that accuracy (i.e., what computation are you using to justify your assertion)?

C will be more accurate because the rms error for the regression line of A on C is smaller than that for A on B: $\left(\sqrt{1-r^2}\right) SD_A$

A on C gives 0.714
A on B gives 0.8

Actual outcomes generally lie closer to the A on C regression line than the A on B line, in terms of SD_A.

(10) (5 pts) A lab assistant in charge of measuring how long it takes for rats to run a maze predicts that usually a rat will take less time on its second time through. A colleague of his, however, notes that the rats ought to regress to the mean maze running time, on average. Are these predictions compatible? Why or why not?

They are compatible. The lab assistant is predicting overall improvement. The colleague is predicting the regression effect, if somewhat loosely stated. The key to their agreement is that one expects the rats to (on average) have less extreme maze running times relative to the second run's mean and SD than their first run time was relative to the first run's mean and SD. If the second maze run has a lower mean we could see general improvement of times (which seems totally plausible) without losing the regression effect (which is a statistical fact, unless $|r| = 1$ which seems unlikely).

Note: there is no regression fallacy here. Neither person is proposing an explanation for anything, just a guess as to what the numbers will look like.