Math 2200 Spring 2016, Exam 2

You may use *any* calculator. You may use a 4×6 inch notecard as a cheat sheet.

1. New Mexico, Texas, California, Arizona, and Florida are the states with the five highest percentages of Hispanics among the population. The table below gives these percentages as well as the percentages of Hispanics among eligible voters. In parentheses beside each number is the deviation of the observation from the mean of the five observations in its column. (Source: Pew Research Center 2014 American Community Survey)

	Percentage among population (x)	Percentage among voters (y)
New Mexico	47.7 (11.8)	40.4 (13.18)
Texas	38.6 (2.7)	28.1 (0.88)
California	38.6 (2.7)	28.0 (0.78)
Arizona	30.5 (-5.4)	21.5 (-5.72)
Florida	24.1 (-11.8)	18.1 (-9.12)

What is the Pearson correlation coefficient r of the two types of percentages?

A) 0.8909 B) 0.9029 C) 0.9149 D) 0.9269 E) 0.9389
F) 0.9509 G) 0.9629 H) 0.9749 I) 0.9869 J) 0.9989

Solution. *Answer:* ***H***

The standard deviation of X is $\sqrt{\left((11.8)^2 + (2.7)^2 + (2.7)^2 + (-5.4)^2 + (-11.8)^2\right)/4}$, or 8.975244. The standard deviation of Y is $\sqrt{\left((13.18)^2 + (0.88)^2 + (0.78)^2 + (-5.72)^2 + (-9.12)^2\right)/4}$, or 8.529185. The sum of the five numbers obtained, one for each row, by multiplying the parenthesized figures in the row is 298.51. Therefore, Pearson's correlation coefficient r is given by

$$r = \frac{298.51}{(5-1)\,(8.975244)\,(8.529185)} = 0.9748664.$$

2. Scott Turner had tabulated the z-scores of a small ($N = 3$) bivariate sample, unwisely using a fountain pen filled with water-soluble ink. Alas, only two of his entries are legible in the table below. The other entries are not legible because the ink with which they were recorded ran, thanks to drool from Scott's dog Hooch, a canine with many fine attributes but a dog that was an undeniable slobberer nonetheless.

	Observation 1	Observation 2	Observation 3
z-score of X (z_x)	■	■	■
z-score of Y (z_y)	■	■	■
product $z_x \cdot z_y$	0.9991	-0.1537	■

Scott did have the Pearson correlation coefficient between X and Y recorded as $r = 0.662849$ on a saliva-free page. Scott was resourceful and he was able to recover the missing entry of the marginal row of z-score products. You can too. What was the final entry in the marginal row?

A) 0.4179 B) 0.4491 C) 0.4803 D) 0.5115 E) 0.5427
F) 0.5739 G) 0.6051 H) 0.6363 I) 0.6675 J) 0.6987

Solution. *Answer:* ***C***

By definition, Pearson's correlation coefficient r is the sum of the three z-score products in the marginal row divided by $N-1$, where $N=3$ is the size of this data set. Thus,

$$r = \frac{1}{3-1}\Big((z_x)_1 \cdot (z_y)_1 + (z_x)_2 \cdot (z_y)_2 + (z_x)_3 \cdot (z_y)_3\Big).$$

If we use the given values of the first two summands on the right side of the preceding equation and if we let w be missing product in the marginal row, then we have

$$r = \frac{1}{2}\,(0.9991 + (-0.1537) + w)\,,$$

or

$$w = 2\,r - 0.9991 + 0.1537 = 2\,(0.662849) - 0.9991 + 0.1537 = 0.480298.$$

Those of you who are who are very observant, who are very efficient in resource conservation, and who are gluttons for punishment, can answer this problem without using the given value of r. If z_1, z_2, z_3 are the three missing z-scores of the explanatory variable X, then $z_1 + z_2 + z_3 = 0$ and $z_1^2 + z_2^2 + z_3^2 = 2$ (because the mean is 0 and the variance is 1). Similarly, if w_1, w_2, w_3 are the three missing z-scores of the response Y, then $w_1 + w_2 + w_3 = 0$ and $w_1^2 + w_2^2 + w_3^2 = 2$. That's already four equations in six unknowns. But we are given $z_1\,w_1 = 0.9991$ and $z_2\,w_2 = -0.1537$. Now we have 6 equations for the 6 unknowns. A piece of cake! The mean = 0 and variance = 1 equations for the X z-scores lead to $z_1^2 + z_2^2 + z_1\,z_2 = 1$. The graph of this equation is an ellipse. Similarly, $w_1^2 + w_2^2 + w_1\,w_2 = 1$, or $(0.9991/z_1)^2 + (-0.1537/z_2)^2 + (0.9991/z_1)\,(-0.1537/z_2) = 1$. The graph of this equation consists of two lines through the origin. Using the Quadratic Formula to solve for the points of intersection, we find $z_1 = -1.144582270$ and $z_2 = 0.7043841705$. All that's left to do is mop up. We find $z_3 = 0.4401980995$, $w_1 = -0.8728948772$, $w_2 = -0.2182047900$, $w_3 = 1.091099667$, and $z_3 \cdot w_3 = (0.4401980995)\,(1.091099667) = 0.4802999998$.

3. Oh no! Hooch licked his chops again, and even more data became undecipherable. Seen below are the remains of another table of z-scores of a small ($N = 3$) bivariate sample, different from the one in the preceding problem.

	Observation 1	Observation 2	Observation 3
z-scores of X (z_x)	-0.756	-0.378	1.134
z-scores of Y (z_y)	-0.898	■	■

Scott did have the Pearson correlation coefficient between X and Y recorded as $r = 0.9844$ on a page Hooch missed. Scott was resourceful and recovered the two missing cell entries. In fact, he found two different ways to recover the missing values. You need only to find one of the ways. What was the entry in the last cell of the second row?

A) 0.518 B) 0.630 C) 0.742 D) 0.854 E) 0.966
F) 1.078 G) 1.190 H) 1.302 I) 1.414 J) 1.526

Solution. *Answer:* ***F***

Let u be the missing z-score of y_2 and let v be the missing z-score of y_3. We will need two equations to solve for these two unknowns. One equation is that the z-scores sum to 0, because the mean of the z-scores is 0. Thus, $u + v = 0.898$, our first equation. For the second equation, we note that the sum divided by $3-1$ of the products of corresponding z-scores is, by definition, the value of r. Thus,

$$(-0.756)\,(-0.898) - 0.378\,u + 1.134\,v = (3-1)\,(0.9844),$$

or $-0.378\,u + 1.134\,v = 1.289912$. *Let us obtain a third equation, actually a rewrite of the first equation, by multiplying each term by 0.378. Because* $(0.378)\,(0.898) = 0.339444$, *he third equation is* $0.378\,u + 0.378\,v = 0.339444$. *Now add corresponding sides of the second and third equations:*

$$\begin{array}{rcl} -0.378\,u + 1.134\,v & = & 1.289912 \\ 0.378\,u + 0.378\,v & = & 0.339444 \\ \hline 1.512\,v & = & 1.629356, \end{array}$$

or $v = 1.629356/1.512 = 1.077616$.

Let's find v *another way. We will use the first equation of the first solution, but, for the second equation, we will use the variance of the z-scores of* Y*: it is 1. Therefore*

$$\frac{1}{3-1}\left((-0.898)^2 + u^2 + v^2\right) = 1,$$

or $u^2 + v^2 = 1.193596$. *The first equation of our first solution is* $u + v = 0.898$, *or* $u = 0.898 - v$. *If we substitute this value for* u *in our quadratic equation, then we obtain* $(0.898 - v)^2 + v^2 = 1.193596$, *or* $2\,v^2 - 1.796\,v - 0.387192 = 0$. *The quadratic formula gives two solutions: 1.077646960 and -0.1796469597. If you notice that the two equations we have used,* $u^2 + v^2 = 1.193596$ *and* $u + v = 0.898$, *then it is apparent that the equations cannot distinguish* v *from* u*: they are symmetric in the two unknowns. That means that the two solutions of the quadratic equation are either* u *and* v, *or* v *and* u. *For* $u = 1.077646960$ *and* $v = -0.1796469597$, *we have*

$$r = \frac{1}{3-1}\left((-0.756)\,(-0.898) + (-0.378)\,(1.077646960) + (1.134)\,(-0.1796469597)\right) = 0.0339088984,$$

which differs from the given value. On the other hand, for $u = -0.1796469597$ *and* $v = 1.077646960$ *we get the correct value of* r*:*

$$r = \frac{1}{3-1}\left((-0.756)\,(-0.898) + (-0.378)\,(-0.1796469597) + (1.134)\,(1.077646960)\right) = 0.9844231020.$$

4. The *maternal mortality rate* (MMR) is the number of female deaths per 100,000 live births from any cause related to or aggravated by pregnancy or its management. The *infant mortality rate* (IMR) is the number of deaths of infants under one year old in a given year per 1,000 live births in the same year. The table below gives the MMR and IMR for six selected countries. (Source: The Central Intelligence Agency)

	Estonia	Singapore	Austria	Israel	South Korea	United States
MMR (2010)	2	3	4	7	16	21
IMR (2015)	3.85	2.48	3.45	3.55	3.86	5.87

What is the Kendall correlation coefficient τ for MMR and IMR for these six countries?

A) 1/3 B) 2/5 C) 7/15 D) 8/15 E) 3/5
F) 2/3 G) 11/15 H) 4/5 I) 13/15 J) 14/15

Solution. *Answer:* ***E***

Notice that MMR has already been sorted in increasing order from left to right. Because 3.85 exceeds 2.48, 3.45, and 3.55, we see that (2,3.85) & (3,2.48), (2,3.85) & (4,3.45), and (2,3.85) & (7,3.55) are discordant pairs. Because every second row entry from 2.48 to the right is smaller than every entry to its right, we see that there are no additional discordant pairs. Altogether, there are $\binom{6}{2}$, *or 15 pairs. It follows that 15 - 3, or 12 pairs are concordant. Therefore*

$$\tau = \frac{12-3}{15} = \frac{9}{15} = \frac{3}{5}.$$

5. World rankings for six countries are tabulated below for MMR and IMR. (The lower the ranking, the lower the mortality rate.) (Source: The Central Intelligence Agency)

	Estonia	Singapore	Austria	Israel	South Korea	United States
MMR (2010)	1	3	7	17	42	49
IMR (2015)	30	4	19	21	31	58

What is the Spearman correlation coefficient ρ for MMR and IMR for these six countries?

A) 0.3493 B) 0.4006 C) 0.4519 D) 0.5032 E) 0.5545
F) 0.6058 G) 0.6571 H) 0.7084 I) 0.7597 J) 0.8110

Solution. *Answer:* ***G***

First we rank these world rankings within the data sets of size 6:

	Estonia	*Singapore*	*Austria*	*Israel*	*South Korea*	*United States*
MMR (2010)	*1*	*2*	*3*	*4*	*5*	*6*
IMR (2015)	*4*	*1*	*2*	*3*	*5*	*6*

We'll use the strange formula. The differences in ranks are $d_1 = -3$, $d_2 = 1$, $d_3 = 1$, $d_4 = 1$, $d_5 = 0$, $d_6 = 0$. *Therefore,*

$$\rho = 1 - \frac{6}{6\left(6^2-1\right)}\left((-3)^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2\right) = \frac{23}{35} = 0.65714.$$

Alternatively, the z-scores of MMR and IMR are

	Estonia	*Singapore*	*Austria*	*Israel*	*South Korea*	*United States*
MMR (2010)	*-1.3363062*	*-0.8017837*	*-0.2672612*	*0.2672612*	*0.8017837*	*1.3363062*
IMR (2015)	*0.2672612*	*-1.3363062*	*-0.8017837*	*-0.2672612*	*0.8017837*	*1.3363062*
	-0.3571428	*1.071429*	*0.2142857*	*-0.07142855*	*0.6428571*	*1.785714*

Each entry in the marginal row is the product of the z-scores above it. The calculation of r, *which gives* ρ, *is*

$$r = \frac{1}{6-1}\left(-0.3571428 + 1.071429 + 0.2142857 - 0.07142855 + 0.6428571 + 1.785714\right) = 0.6571429.$$

6. When we fit a regression line of slope m and y-intercept b to the data in the table

	Observation 1	Observation 2	Observation 3
X	0	1	5
Y	-3	0	3

we minimize an algebraic expression involving the parameters m and b. The coefficient of b^2 in the expression that we minimize is 3. What is the coefficient of m^2 in the expression that we minimize?

A) 23 B) 24 C) 25 D) 26 E) 27
F) 28 G) 29 H) 30 I) 31 J) 32

Solution. *Answer:* ***D***

The regression line $y = m\,x + b$ *is also called the least squares line. That is because it minimizes the sum of the squares of the errors (residuals). For three bivariate observations* $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, *the predicted response values are* $\widehat{y}_1 = m\,x_1 + b, \widehat{y}_2 = m\,x_2 + b, \widehat{y}_3 = m\,x_3 + b$. *The errors (residuals) are* $y_1 - \widehat{y}_1, y_2 - \widehat{y}_2, y_3 - \widehat{y}_3$, *or* $y_1 - (m\,x_1 + b), y_2 - (m\,x_2 + b), y_3 - (m\,x_3 + b)$. *Substituting the given data* ($x_1 = 0$, $y_1 = -3$, $x_2 = 1$, $y_2 = 0$, $x_3 = 5$, $y_3 = 3$) *into these formulas for the residuals, we find that the*

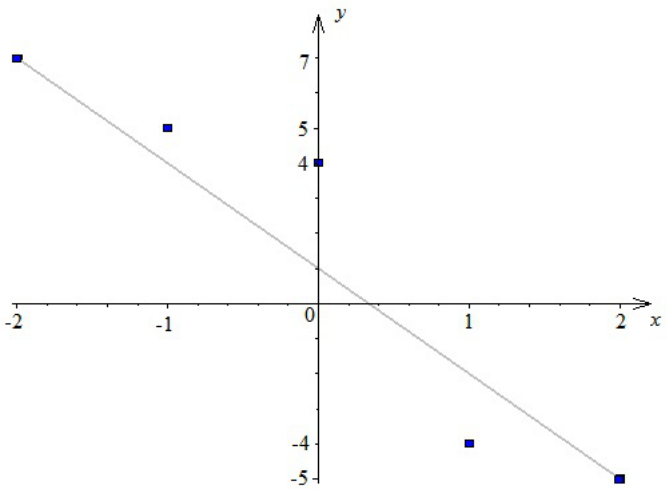
errors are $-3-(m\cdot 0+b)\,, 0-(m\cdot 1+b)\,, 3-(m\cdot 5+b)$. *We calculate the sum of the squares of these errors to obtain the expression that is minimized by the least squares line:*

$$\begin{aligned} SSE(m,b) &= \big(-3-(m\cdot 0+b)\big)^2+\big(0-(m\cdot 1+b)\big)^2+\big(3-(m\cdot 5+b)\big)^2 \\ &= (3+b)^2+(m+b)^2+\big(5\,m-(3-b)\big)^2 \\ &= (3+b)^2+m^2+2\,m\,b+b^2+25\,m^2-10\,m\,(3-b)+(3-b)^2 \\ &= 26\,m^2+12\,m\,b+3\,b^2-30\,m+18. \end{aligned}$$

7. Scott Turner was analyzing a bivariate data set with $N = 5$ data points.

	Observation 1	Observation 2	Observation 3	Observation 4	Observation 5
X	-2	-1	0	1	2
Y	7	5	4	-4	-5

Scott computed most of the good stuff (for example, $s_X = 1.581139, s_Y = 5.504544$, and $r_{XY} = -0.9479002$) and was about to calculate the equation of the regression line when he heard Hooch bark. That set him thinking: What would be the point of such a calculation if the equation for the regression line would just be lost in another inundation of spittle. So, instead, he simply connected the points for observations 1 and 5. Good enough, he thought.



What is SSE for the line Scott used to fit the data?

A) 5 B) 6 C) 7 D) 8 E) 9
F) 10 G) 11 H) 12 I) 13 J) 14

Solution. *Answer:* ***J***

The line Scott used to fit the data passes through the points $(-2,7)$ *and* $(2,-5)$. *We will obtain the equation for the line by finding its slope* m *and then using the "point-slope" equation for a line. Using both points* $(-2,7)$ *and* $(2,-5)$ *to calculate the slope* m, *we have* $m = (-5-7)/\big(2-(-2)\big)$, *or* $m = -3$. *With this value* m, *the point-slope equation of the line is* $y = m\,(x-x_0)+y_0$ *where* (x_0, y_0) *is* any *point on the line. We have a choice: we can use either* $(-2,7)$ *or* $(2,-5)$. *We will use* $(2,-5)$, *but the other point would give us the same equation. That equation we obtain is* $y = m\,(x-2)+(-5)$, *or* $y = -3\,(x-2)+(-5)$, *or* $y = 1-3\,x$. *For the observations* $x = -2,-1,0,1,2$ *of the explanatory variable, the response values* $\widehat{y}$ *predicted by the line*

are $1-3\cdot(-2)$, $1-3\cdot(-1)$, $1-3\cdot(0)$, $1-3\cdot(1)$, $1-3\cdot(2)$, *or 7, 4, 1, -2, -5 respectively. The errors are therefore* $7-7, 5-4, 4-1, -4-(-2)$, *and* $-5-(-5)$, *or 0, 1, 3, -2, 0. The sum of the squared errors is* $0^2+1^2+3^2+(-2)^2+0^2$, *or 14.*

8. Suppose that X is an explanatory variable with observations $x_1, x_2, \ldots, x_N$, that Y is a response variable with observations $y_1, y_2, \ldots, y_N$, that the Pearson correlation coefficient between the X and Y observations is 0.8, that the variances of the X and Y observations are equal, that the mean of the X observations is 15, and that the y-intercept of the regression line is -10. What is the mean of the Y observations?

   A) 2 B) 4 C) 6 D) 8 E) 10
   F) 12 G) 14 H) 16 I) 18 J) 20

Solution. *Answer:* ***A***

The regression line is $y = m\,x + b$ *where* $m = r\,s_Y/s_X = r = 0.8$ *and* $b = \bar{y} - m\,\bar{x}$. *Using the given information and the value of* m *we have deduced, the last equation becomes* $-10 = \bar{y} - (0.8)\,(15)$, *or* $\bar{y} = -10 + 12$.

9. For four observed points $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$, the correlation coefficient r is given by $r = 0.8466$, the standard deviation of the explanatory variable observations is 3.8622, and the regression line is $y = 4.4972\,x + 5.6369$. What is the regression sum of squares SSR? (Round to the nearest integer.)

   A) 900 B) 905 C) 910 D) 915 E) 920
   F) 925 G) 930 H) 935 I) 940 J) 945

Solution. *Answer:* ***B***

The slope m *of the regression line is 4.4972. Thus*

$$4.4972 = r\,\frac{s_Y}{s_X} = (0.8466)\,\frac{s_Y}{3.8622},$$

or $s_Y = (4.4972)\,(3.8622)/(0.8466)$, *or* $s_Y = 20.51628$. *It follows that*

$$SSR = (4-1)\,r^2\,s_Y^2 = 3\,(4.4972)^2\,(20.51628)^2 = 905.0651.$$

By way of a verification of our calculations, we mention that the distributions that served as the models for this problem were X = 0, 2, 7, 8 and Y = 10, 11, 24, 54. The given regression line was calculated using this data. Using the given regression line, the predicted responses to the observed values of X are calculated to be 5.6369, 14.6313, 37.1173, 41.6145. The mean of the Y observations is 24.75. The deviations of the predicted responses from this mean are -19.1131, -10.1187, 12.3673, 16.8645. The sum of the squares of these deviations is 905.0602.

10. A regression line was calculated using three points in the xy-plane. The standard deviation of the three response observations was 0.7234178, and the residuals were -0.183, 0.367, -0.183. What was the regression sum of squares SSR? (The next problem also refers to this data.)

    A) 0.5935 B) 0.6438 C) 0.6941 D) 0.7444 E) 0.7947
    F) 0.8450 G) 0.8953 H) 0.9456 I) 0.9959 J) 1.0462

Solution. *Answer:* ***F***

We calculate

$$\begin{aligned} SST &= (3-1)\,s_Y^2 = 2\cdot(0.7234178)^2 = 1.046667, \\ SSE &= (-0.183)^2 + (0.367)^2 + (-0.183)^2 = 0.201667, \quad \textit{and} \\ SSR &= SST - SSE = 1.046667 - 0.201667 = 0.845. \end{aligned}$$

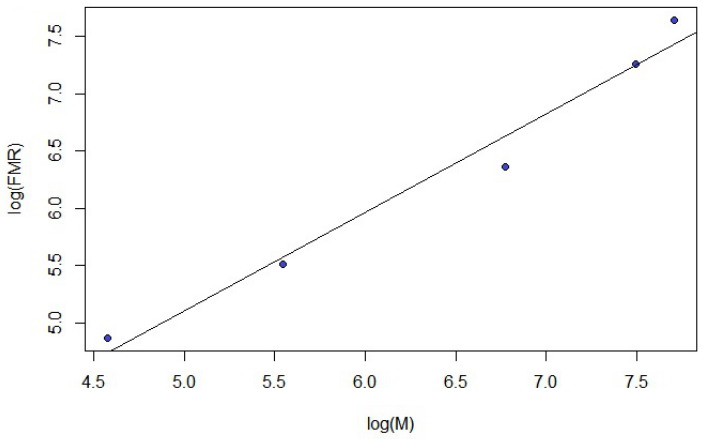
11. In the small data set of the preceding problem, what percentage of the response variance was explained by the regression line?

A) 80.7325 B) 81.8998 C) 83.0671 D) 84.2344 E) 85.4017
F) 86.5690 G) 87.7363 H) 88.9036 I) 90.0709 J) 91.2382

Solution. *Answer:* ***A***

The answer is $100\,SSR/SST$*, or 84.5/1.046667, or 80.7325.*

12. A scatterplot of observations (x_j, y_j) of (X,Y) where X is the mass (M) of a small mammal and Y is its field metabolic rate (FMR) suggests a relationship of the form $y = A\,x^p$ for some positive constants A and p. The suitability of this conjectured relationship can be investigated by scatterplotting $(\ln(x_j), \ln(y_j))$. In the Figure below, for each of five small mammals selected, the point $(\ln(x), \ln(y))$ was plotted. The regression line based on these five points was added to the log-log plot. Note: All logarithms in this discussion are natural logarithms with base e, not 10. The plot was created with R, in which language the natural logarithm is coded as `log`, not `ln`.



The five selected mammals in order of increasing mass are 1) the white-tailed antelope squirrel (Ammospermophilus leucurus), 2) the cascade golden-mantled ground squirrel (Spermophilus saturatus), 3) Blanford's fox (Vulpes cana), 4) the black-tailed jackrabbit (Lepus californicus), and 5) the swift fox (Vulpes velox). Quantities of interest for the five selected species are tabulated below. The masses are given in grams. To save horizontal space, $\overline{\ln(M)}$ is represented by μ_M and $\overline{\ln(FMR)}$ is represented by μ_R The metabolic rates are given in a unit that is appropriate for the rate of energy expenditure with respect to time.

M	FMR	ln(M)	$\ln(M) - \mu_M$	$(\ln(M) - \mu_M)^2$	ln(FMR)	$\ln(\text{FMR}) - \mu_R$	$(\ln(\text{FMR}) - \mu_R)^2$
97.5	130.6	4.579852	-1.8399305	3.3853444	4.872139	-1.45593586	2.119749
256	248	5.545177	-0.8746055	0.7649347	5.513429	-0.81464633	0.663649
874	578	6.773080	0.3532975	0.1248191	6.359574	0.03149879	0.000992
1800	1416	7.495542	1.0757590	1.1572575	7.255591	0.92751620	0.860286
2220	2079	7.705262	1.2854796	1.6524577	7.639642	1.31156721	1.720209

(Source for M and FMR: SPEAKMAN, J.R., The cost of living: Field metabolic rates of small mammals, Advances in Ecological Research **30** (2000), 177-297.) What is the slope of the regression line that fits the plotted $(\ln(M), \ln(FMR))$ points in the figure? (The next two questions are related.)

A) 0.3689 B) 0.4506 C) 0.5323 D) 0.6140 E) 0.6957
F) 0.7774 G) 0.8591 H) 0.9408 I) 1.0225 J) 1.1042

Solution. *Answer:* ***G***

The mean of ln(M), denoted by μ_M in the table, is 6.419783. The mean of ln(FMR), denoted by μ_R in the table, is 6.328075. The standard deviation $s_{\ln(M)}$ of ln(M) is obtained by adding the entries of the $(\ln(M) - \mu_M)^2$ column as follows:

$$s_{\ln(M)} = \sqrt{\frac{1}{5-1}\,(3.3853444 + 0.7649347 + 0.1248191 + 1.1572575 + 1.6524577)} = 1.330866.$$

Similarly, using the values in the $(\ln(FMR) - \mu_R)^2$ column, we have

$$s_{\ln(FMR)} = \sqrt{\frac{1}{5-1}\,(2.119749 + 0.663649 + 0.000992 + 0.860286 + 1.720209)} = 1.158111.$$

For the Pearson correlation between ln(M) and ln(FMR), we use the standard deviations just calculated together with the entries in the $\ln(M) - \mu_M$ and $\ln(FMR) - \mu_R$ columns:

$$r = \frac{1}{(5-1)}\,\frac{1}{s_{\ln(M)}}\,\frac{1}{s_{\ln(FMR)}} \qquad \Big((-1.8399305)(-1.45593586) + (-0.8746055)(-0.81464633) + (0.3532975)(0.03149879) + (1.0757590)(0.92751620) + (1.2854796)(1.31156721)\Big) = 0.9871957.$$

We calculate $m = r \cdot s_{\ln(FMR)}/s_{\ln(M)} = (0.9871957)(1.158111)/1.330866 = 0.8590516$ and $b = \mu_R - m \cdot \mu_M = 6.328075 - (0.8590516)(6.419783) = 0.8131501$. The regression line is $\ln(FMR) = m\ln(M) + b = 0.8590516\ln(M) + 0.8131501$. The two numbers on the right side of this equation answer this question and the next.

13. For the regression line that fits the plotted $(\ln(M), \ln(FMR))$ points in the figure of problem 12, what is the intercept on the vertical axis? (This would state "y-intercept" if we were not already using y for the response variable FMR; here we are using the vertical axis for ln(FMR).)

A) 0.5459 B) 0.6350 C) 0.7241 D) 0.8132 E) 0.9023
F) 0.9914 G) 1.0805 H) 1.1696 I) 1.2587 J) 1.3478

Solution. *Answer:* ***D*** *(See previous solution.)*

14. We know better than to extrapolate, but let's go ahead and do so just this once. The competition for the title of Earth's smallest known, surviving mammal is fierce, but the winner is the Etruscan shrew (Suncus etruscus), which has a mass of about 1.8g. (The bumblebee bat, aka Kitti's hog-nosed bat, aka Craseonycteris thonglongyai, is so close a runner-up that a well-nourished specimen might outweigh an Etruscan shrew that skipped a meal.) What FMR does our model predict for the Etruscan shrew?

A) 2.9120 B) 3.3243 C) 3.7366 D) 4.1489 E) 4.5612
F) 4.9735 G) 5.3858 H) 5.7981 I) 6.2104 J) 6.6227

Solution. *Answer:* ***C***

First, we apply the exponential function to each side of the regression line $\ln(FMR) = 0.8590516\ \ln(M) + 0.8131501$ *calculated in problem 12. We obtain*

$$\begin{aligned} FMR &= \exp\left(\ln(FMR)\right) \\ &= \exp\left(0.8590516\ \ln(M) + 0.8131501\right) \\ &= \exp\left(0.8590516\ \ln(M)\right) \cdot \exp\left(0.8131501\right) \\ &= 2.255\ \exp\left(\ln\left(M^{0.8590516}\right)\right) \\ &= 2.255\, M^{0.8590516}. \end{aligned}$$

Substitute $M = 1.8$ *to obtain the predicted FMR:* $2.255 \cdot 1.8^{0.8590516}$*, or 3.736275.*

15. Elbridge Gerry, Hannibal Hamlin, and Schuyler Colfax were the first three collectible cards chosen for Weetabix's *Vice Presidents of the United States You Never Heard Of* series. Each box of Weetabix cereal contained one card, but the cards were not inserted with equal likelihood: 50% of the cereal boxes had an Elbridge Gerry card, 30% of the cereal boxes had a Hannibal Hamlin card, and 20% of the cereal boxes had a Schuyler Colfax card. What is the probability p that the purchase of four cereal boxes results in the complete set of these illustrious civil servants? Let us estimate that probability using the following table of random digits grouped in 30 blocks of 4 digits. Choose three positive integers m, n, and k that sum to 10. Assign the first m digits in the list 0,1,2,3,4,5,6,7,8,9 to Elbridge, the next n digits to Hannibal, and the final k digits to Schuyler. Of course, you must choose m, n, and k in a way that is appropriate for the simulation. Use each group of four digits in the table to simulate the purchase of 4 cereal boxes.

7725	8275	3324	7640	5984	5015
5518	0726	7060	6925	9800	7408
5063	7335	6724	3366	3997	1587
3956	5688	3169	1910	6279	6134
8422	2407	0000	1802	4128	0773

What estimate of p results from the simulation? (Suggestion: Just to the right of the table, write each Vice Presidential first name followed by the digits by which his card is represented. It is easy to err without a visual reference.)

A) 0.2333 B) 0.2666 C) 0.3000 D) 0.3333 E) 0.3666
F) 0.4000 G) 0.4333 H) 0.4666 I) 0.5000 J) 0.5333

Solution. *Answer:* ***D***

With $m = 5, n = 3$*, and* $k = 2$*, we let 0, 1, 2, 3, and 4 represent a cereal box with an Elbridge Gerry card, 5, 6, and 7 represent a cereal box with a Hannibal Hamlin card, and 8, 9 represent a cereal box with a Schuyler Colfax card. Using S for success in acquiring the full set and F for failure, here are the results of our simulated cereal box purchases:*
F S F F S F
S F F S F S
F F F F S S
S F S F S F
F F F F F F

Thus, 10 successes in 30 trials.

The exact formula for p that is given in the course notes does not apply here. The formula in the course notes is for the case in which the collectible items have equal probabilities. For the current problem, the exact

formula for p is

$$\left(\frac{1}{2}+\frac{3}{10}+\frac{1}{5}\right)^4-\left(\left(\frac{1}{2}+\frac{3}{10}\right)^4+\left(\frac{3}{10}+\frac{1}{5}\right)^4+\left(\frac{1}{2}+\frac{1}{5}\right)^4\right)+\left(\left(\frac{1}{2}\right)^4+\left(\frac{3}{10}\right)^4+\left(\frac{1}{5}\right)^4\right),$$

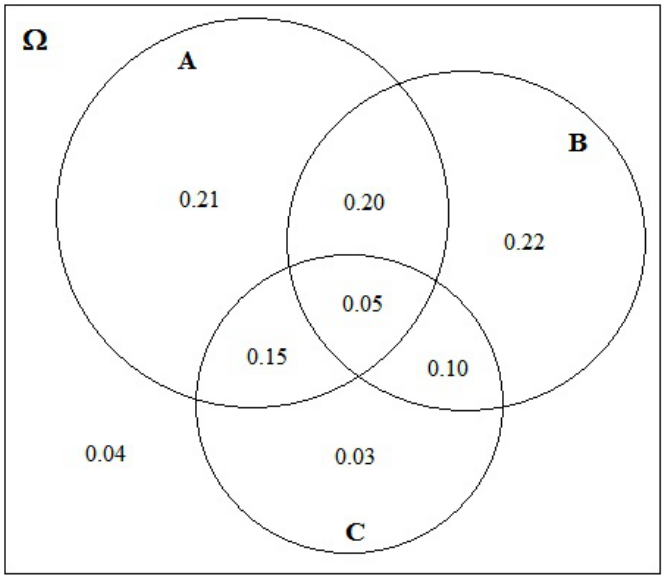
or $p = 9/25$*, or* $p = 0.36$*. A large simulation was run to confirm that there was no error in calculating this probability. Coded in Maple, a computer algebra system, a simulation with 1,000,000 trials resulted in 360,255 successes, or* $p \approx 0.360255$*.*

16. This problem and the two that follow it concern three events, E, F, and G, that are constructed from three more basic events, A, B, and C. Let E be the event that *at least* one of the three events A, B, and C occurs, let F be the event that *exactly* one of the three events A, B, and C occur, and let G be the event that *exactly* two of the three events A, B, and C occur. Suppose that P(A) = 0.61, that P(B) = 0.57, that P(C) = 0.33, that P(A and B) = 0.25, that P(A and C) = 0.20, that P(B and C) = 0.15, and that P(A and B and C) = 0.05. You are to calculate P(E), P(F), and P(G). In this problem, respond with P(E).

    A) 0.88 B) 0.89 C) 0.90 D) 0.91 E) 0.92
    F) 0.93 G) 0.94 H) 0.95 I) 0.96 J) 0.97

Solution. *Answer:* ***I***

Refer to the figure.



We have P(E) = 1 - 0.04 = 0.96.

17. Refer to problem 16. What is P(F)?

    A) 0.40 B) 0.41 C) 0.42 D) 0.43 E) 0.44
    F) 0.45 G) 0.46 H) 0.47 I) 0.48 J) 0.49

Solution. *Answer:* ***G***

We have P(F) = 0.21 + 0.22 + 0.03 = 0.46.

18. Refer to problem 16. What is P(G)?

A) 0.40 B) 0.41 C) 0.42 D) 0.43 E) 0.44
F) 0.45 G) 0.46 H) 0.47 I) 0.48 J) 0.49

Solution. *Answer:* ***F***

We have P(G) = 0.20 + 0.10 + 0.15 = 0.45.

19. Despereaux has before him 11 cheese wedges he can sample. Seven of the wedges are 100% cheese by anyone's definition, but four of the wedges contain cellulose (wood-pulp). Despereaux nibbles on two of the wedges. What is the probability that both cheeses sampled have no cellulose? (The next problem discusses Despereaux's snack further.)

A) 0.3727 B) 0.3818 C) 0.3909 D) 10.4000 E) 0.4091
F) 0.4182 G) 0.4273 H) 0.4364 I) 0.4455 J) 0.4545

Solution. *Answer:* ***B***

Let F_1 be the event that the first nibble is wood-free. Let F_2 be the event that the second nibble is wood-free. Then $P(F_1) = 7/11$ and $P(F_2 \,|\, F_1) = 6/10$. Therefore

$$P(F_1 \cap F_2) = P(F_2 \,|\, F_1) \cdot P(F_1) = \frac{6}{10} \cdot \frac{7}{11} = \frac{21}{55} = 0.3818182.$$

20. Refer to the preceding problem. What is the probability that exactly one of the cheeses sampled contains cellulose?

A) 0.4364 B) 0.4455 C) 0.4545 D) 0.4636 E) 0.4727
F) 0.4818 G) 0.4909 H) 0.5000 I) 0.5091 J) 0.5182

Solution. *Answer:* ***I***

Let C_1 and C_2 be the events that the first and, respectively, second, cheese sampled contains cellulose. Then

$$\begin{aligned} P(F_1 C_2 \cup C_1 F_2) &= P(C_2 \,|\, F_1) \cdot P(F_1) + P(F_2 \,|\, C_1) \cdot P(C_1) \\ &= \frac{4}{10} \cdot \frac{7}{11} + \frac{7}{10} \cdot \frac{4}{11} \\ &= \frac{56}{110} \\ &= 0.5090909. \end{aligned}$$

21. For Section 1 of First President University's Spring 2016 Elementary Statistics course, the table below shows the frequency count of all 40 drops categorized by the time interval during which the course was dropped.

Nov, 2015	Dec, 2015	Jan 1-18, 2016	1st Week of Classes	2nd Week Classes	3rd Week	4th Week
17	7	2	6	4	3	1

The instructor's class list shows all 105 students who ever registered for the course (including the 40 who dropped). If the instructor picks a name at random from the class list, what is the probability that the selected student is still enrolled in the course given that the student did not drop the course *before* the second week of classes?

A) 0.7671 B) 0.7808 C) 0.7945 D) 0.8082 E) 0.8219
F) 0.8356 G) 0.8493 H) 0.8630 I) 0.8767 J) 0.8904

Solution. *Answer:* ***J***

Let E be the event that the selected student is still enrolled. Then $P(E) = (105-40)/105 = 65/105$. Let B be the event that the selected student was still enrolled at the start of the 2nd week of classes. Then

$$P(B) = 1 - \left(\frac{17}{105} + \frac{7}{105} + \frac{2}{105} + \frac{6}{105}\right) = \frac{73}{105}.$$

Observe that $E \cap B = E$. Thus

$$P(E \,|\, B) = \frac{P(E \cap B)}{P(B)} = \frac{P(E)}{P(B)} = \frac{65/105}{73/105} = \frac{65}{73} = 0.890411.$$

22. A study conducted in San Diego (quite some time ago) compared trauma sustained by Hispanic children in vehicular accidents with trauma sustained by non-Hispanic White children. Of all the children considered in the study, 39.65% were Hispanic. Of all the children considered in the study, 77.40% were not wearing seat belts. Of the Hispanic children in the study, 90.13% were not wearing seat belts. What is the probability that a randomly chosen child in the study was non-Hispanic White and not wearing a seat belt?

    A) 0.3749 B) 0.4166 C) 0.4583 D) 0.5000 E) 0.5417
    F) 0.5834 G) 0.6251 H) 0.6668 I) 0.7085 J) 0.7502

Solution. *Answer:* ***B***

Let H be the event that a randomly selected child in the study is Hispanic. Let N be the event that a randomly selected child in the study was not seat-belted. Let Ω denote all children in the study. That is, $\Omega = H \cup H^{\mathrm{C}}$. We are given $P(H) = 0.3965$ and $P(N) = 0.7740$. We are also given $P(N \,|\, H) = 0.9013$. We are asked for the value of $P(\, N \cap H^{\mathrm{C}}\,)$. The key is to first find the value of $P(\, N \cap H\,)$. The given conditional probability permits this calculation:

$$P(N \cap H) = P(N|H)P(H) = (0.9013)(0.3965) = 0.3573655.$$

Next, observe that

$$N = N \cap \Omega = N \cap (H \cup H^{\mathrm{C}}) = (N \cap H) \cup (N \cap H^{\mathrm{C}}).$$

Therefore

$$P(N \cap H^{\mathrm{C}}) = P(N) - P(N \cap H) = 0.7740 - 0.3573655 = 0.4166345.$$

23. A patient exhibits symptoms of a disease. There is a screening test for the suspected disease that requires a blood sample. The sensitivity of the test is 0.87 and its specificity is 0.96. The prevalence of the disease is 0.4%. What is the probabilty that a patient who tests positive actually has the disease? Round to the nearest hundredth.

    A) 0.04 B) 0.05 C) 0.06 D) 0.07 E) 0.08
    F) 0.09 G) 0.10 H) 0.11 I) 0.12 J) 0.13

Solution. *Answer:* ***E***

Let S denote the event that a person has the disease in question. We have

$$\begin{aligned} P(S \,|\, POS) &= \frac{P(POS \,|\, S)P(S)}{P(POS \,|\, S)P(S) + P(POS \,|\, S^{\mathrm{C}})P(S^{\mathrm{C}})} \\ &= \frac{sensitivity \cdot prevalence}{sensitivity \cdot prevalence + (1 - specificity)(1 - prevalence)} \\ &= \frac{0.87 \cdot 0.004}{0.87 \cdot 0.004 + (1 - 0.96)(1 - 0.004)} \\ &= 0.080331. \end{aligned}$$

24. The probability that a randomly selected person suffering from a certain disease will survive 5 or more years is 0.3. If six persons with the disease are randomly selected, then what is the probability that at least 4 will live 5 or more years?

A) 0.0705 B) 0.0936 C) 0.1167 D) 0.1398 E) 0.1629
F) 0.1860 G) 0.2091 H) 0.2322 I) 0.2553 J) 0.2784

Solution. *Answer:* ***A***

Let X = 1 if a randomly selected person suffering from the disease survives 5 or more years, and let X = 0 otherwise. Then X is a Bernoulli distribution with probability of success equal to 0.3. The number of successes $S = X_1 + X_2 + \cdots + X_6$ *in 6 independent trials is a binomial distribution. We have*

$$P(S \geq 4) = P(S=4)+P(S=5)+P(S=6) = \binom{6}{4}(0.3)^4(0.7)^2+\binom{6}{5}(0.3)^5(0.7)^1+\binom{6}{6}(0.3)^6(0.7)^0 = 0.07047.$$

25. Rollin rolls a standard, fair die until he rolls a 6 (at which point he does not roll again). However, his fourth roll, if there is a fourth roll, is his last roll regardless of the result of the roll. Let X be the number of times Rollin rolls the die. What is E(X)?

A) 2.3848 B) 2.4879 C) 2.5910 D) 2.6941 E) 2.7972
F) 2.9003 G) 3.0034 H) 3.1065 I) 3.2096 J) 3.3127

Solution. *Answer:* ***H***

The possible values of X are 1, 2, 3, and 4. We calculate

$$\begin{aligned} E(X) &= 1 \times P(X=1) + 2 \times P(X=2) + 3 \times P(X=3) + 4 \times P(X=4) \\ &= 1 \times \frac{1}{6} + 2 \times \frac{5}{6} \times \frac{1}{6} + 3 \times \left(\frac{5}{6}\right)^2 \times \frac{1}{6} + 4\left(\left(\frac{5}{6}\right)^3 \times \frac{1}{6} + \left(\frac{5}{6}\right)^4\right) \\ &= \frac{671}{216} \\ &= 3.106481481. \end{aligned}$$

Note: P(X = 4) can also be calculated as 1 - (P(X = 1) + P(X = 2) + P(X = 3)).