Math 2200 Spring 2017, Exam 2

You may use *any* calculator. You may use ONE "cheat sheet" in the form of a 4" x 6" note card (the medium size of the standard three sizes).

The exam is out of 100 points. The first 19 problems are worth 5 points each. The last 5 problems are worth 1 point each.

1. For what value of ζ are the z-scores of the five integer distribution 0, 1, 2, 3, 4 in the table below correct? (Some of your calculations in this problem might be of use in some subsequent problems.)

X	0	1	2	3	4
Z	$-2\,\zeta$	$-\zeta$	0	ζ	$2\,\zeta$

A) 0.4217 B) 0.5271 C) 0.6325 D) 0.7379 E) 0.8433
F) 0.9488 G) 1.054 H) 1.1596 I) 1.2650 J) 1.3704

Solution C) 0.6325

The mean m of X is $(0 + 1 + 2 + 3 + 4)/5$, or 2. The standard deviation s of X is given by $s = \sqrt{((0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2)/4}$, or $s = 1.581139$. Therefore, $\zeta = (3-m)/s = (3-2)/1.581139 = 0.6324555$. As a verification using R, we have

```
> scale(0:4)
           [,1]
[1,] -1.2649111
[2,] -0.6324555
[3,]  0.0000000
[4,]  0.6324555
[5,]  1.2649111
```

2. Risser's Sign, which is ascertained by pelvic X-ray, is used to help determine the potential for progression of scoliosis. It is a six stage rank, from 0 to 5. The first five stages are linear: for an integer k from 0 to 4, Risser Sign k signifies $25\,k\%$ iliac apophysis ossification. Notice that Risser Sign 4 indicates that ossification is complete. Risser Sign 5 signifies a further growth development, fusion to iliac crest, that is obviously something other than further ossification. Consequently, the 0 to 5 Risser scale, unlike the 0 to 4 Risser scale, is nonlinear. The table below gives the age and Risser Sign for five girls in the linear Risser Sign range. The distribution of their ages has mean 13.78 and standard deviation 2.146392. Find the Pearson r correlation coefficient. (In Problems 5–7, you might have use for the information given in this problem as well as its answer.)

Age	10.2	13.8	14.3	14.7	15.9
Risser Sign	0	1	2	3	4

A) 0.6464 B) 0.6835 C) 0.7206 D) 0.7577 E) 0.7948
F) 0.8319 G) 0.8690 H) 0.9061 I) 0.9432 J) 0.9803

Solution H) 0.9061

The Age z-scores are -1.667915023, 0.009317961, 0.242266987, 0.428626207, 0.987703869. The Risser Sign z-scores (from Problem 1) are -1.2649111, -0.6324555, 0.0000000, 0.6324555, 1.2649111. The products of corresponding z-scores are 2.109764167, -0.005893196, 0.000000000, 0.271087016, 1.249357551. The sum of these five products is 3.624316. Thus, $r = 3.624316/(5-1) = 0.906079$. Verification in R:

```
> Age = c(10.2,13.8,14.3,14.7,15.9)
> m = mean(Age)
> s = sd(Age)
```

```
> m
[1] 13.78
> s
[1] 2.146392
> z.Age = (Age - m)/s
> z.Age  # Age z-scores
[1] -1.667915023  0.009317961  0.242266987  0.428626207  0.987703869
> Risser = 0:4
> z.Risser = (Risser - mean(Risser))/sd(Risser) # Or get values from Problem 1
> z.Risser # Risser Sign z-scores
[1] -1.2649111 -0.6324555  0.0000000  0.6324555  1.2649111
> r = sum(z.Age*z.Risser)/4
> r  #Pearson r
[1] 0.9060789
> cor(Age,Risser) # Pearson r using R's builtin correlation function
[1] 0.9060789
```

3. The table below gives the age and Risser Sign for six boys. Find the Kendall tau correlation coefficient for Age and Risser Sign.

Age	14.1	15.3	15.2	16.5	16.4	18.0
Risser Sign	0	1	2	3	4	5

A) 0.3333 B) 0.4000 C) 0.4666 D) 0.5333 E) 0.6000
F) 0.6666 G) 0.7333 H) 0.8000 I) 0.8666 J) 0.9333

Solution G) 0.7333

Notice that the second row is in increasing order going left to right. Every one of the five values after 14.1 in the first row is greater than 14.1. That results in 5 concordant pairs and 0 discordant pairs. Exactly three of the four values after 15.3 in the first row is greater than 15.3. That results in 3 concordant pairs and 1 discordant pair. Every one of the three values after 15.2 in the first row is greater than 15.2. That results in 3 concordant pairs and 0 discordant pairs. Exactly one of the two values after 16.5 in the first row is greater than 16.5. That results in 1 concordant pair and 1 discordant pair. Finally, the single value after 16.4 in the first row is greater than 16.4. That results in 1 concordant pair and 0 discordant pairs. The total number of concordant pairs is $5+3+3+1+1$, or 13. The total number of discordant pairs is $0+1+0+1+0$, or 2. The numerator of Kendall tau is $13-2$, or 11. The denominator of Kendall tau is $6\times(6-1)/2$, or 15. Thus, $\tau = 11/15 = 0.7333333$.

In R:

```
> Age = c(14.1, 15.3, 15.2, 16.5, 16.4, 18.0)
> Risser = 0:5
> cor(Age, Risser, method = "kendall")
[1] 0.7333333
```

4. For the dataset in the preceding problem, find the Spearman rho correlation coefficient for Age and Risser Sign.

A) 0.5126 B) 0.5659 C) 0.6192 D) 0.6725 E) 0.7258
F) 0.7791 G) 0.8324 H) 0.8857 I) 0.9390 J) 0.9923

Solution H) 0.8857

We will use R to calculate Spearman rho in three different ways. First we use R's builtin `rank( )` function to calculate the ranks of Age and Risser Sign. Then we calculate Pearson r of the ranks by using `cor( )` without specifying a correlation method—the default is Pearson r. For the second way to calculate

Spearman rho, we use the builtin `cor( )` function applied to the raw data, but this time we include the optional parameter `method = ‘‘spearman’’`. Finally, for the third approach, we calculate Spearman rho using the strange formula that involves the squares of the differences of ranks.

```
> rk.Age = rank(Age) # ranks of Ages
> rk.Age
[1] 1 3 2 5 4 6
> rk.Risser = rank(Risser) # ranks of Risser Signs (= Risser + 1)
> rk.Risser
[1] 1 2 3 4 5 6
> cor(rk.Age,rk.Risser) # Spearman rho = Pearson r of ranks
[1] 0.8857143
> cor(Age,Risser, method = "spearman") # R's builtin for Spearman rho
[1] 0.8857143
> d = rk.Age - rk.Risser # The vector of rank differences
> d
[1]  0  1 -1  1 -1  0
> N = length(Age)
> N
[1] 6
> 1 - 6/(N*(N^2-1))*sum(d^2) #The strange formula for rho (no ties case)
[1] 0.8857143
```

5. For the 5-pair dataset of Problem 2, with X = Age and Y = Risser Sign (in the range from 0 to 4), calculate the regression line. In this problem respond with the slope of the regression line.

   A) 0.2973 B) 0.4207 C) 0.5441 D) 0.6675 E) 0.7909
   F) 0.9143 G) 1.0377 H) 1.1611 I) 1.2845 J) 1.4079

Solution D) 0.6675

If r is Pearson's linear correlation coefficient, which equals 0.9060789 (from Problem 2), if s_{Age} is the standard deviation of Age, which equals 2.146392 (as given in Problem 2), if s_{Risser} is the standard deviation of Risser Sign, which equals 1.581139 (as calculated in Problem 1), and if m is the slope of the regression line, then we have

$$m = r\,\frac{s_{\text{Risser}}}{s_{\text{Age}}} = (0.9060789)\,\frac{1.581139}{2.146392} = 0.6674627.$$

Or, using R,

```
> Age = c(10.2,13.8,14.3,14.7,15.9)
> Risser = 0:4
> lin.model = lm(Risser ~ Age)
> lin.model$coefficients
(Intercept)         Age
 -7.1976340   0.6674626
> m = cor(Age,Risser)*sd(Risser)/sd(Age)
> m
[1] 0.6674626
> b = mean(Risser) - m*mean(Age)
> b
[1] -7.197634
```

6. What is the vertical axis intercept of the regression line calculated in the preceding problem.

   A) −7.1976 B) −5.5981 C) −3.9987 D) −2.3992 E) −0.7997
   F) 0.7997 G) 2.3992 H) 3.9987 I) 5.5981 J) 7.1976

Solution **A)** -7.1976

If $\overline{Risser}$ is the mean of Risser, which equals 2, if $\overline{Age}$ is the mean of Age, which equals 13.78 (as given in Problem 2), and if m is the slope of the regression line, which equals 0.6674626 (as calculated in the preceding problem), then the vertical intercept b is given by

$$b = \overline{Risser} - m\,\overline{Age} = 2 - (0.6674626)(13.78) = -7.197635.$$

7. At what age does the linear model predict 100% ossification (Risser Sign 4) for girls?

   A) 15.74 B) 15.87 C) 16.00 D) 16.13 E) 16.26
   F) 16.39 G) 16.52 H) 16.65 I) 16.78 J) 16.91

Solution **I) 16.78**

The regression line is $\widehat{\text{Risser}} = 0.6674626\,\text{Age} - 7.1976340$. The value of Age that results in Risser $= 4$ satisfies $4 = 0.6674626\,\text{Age} - 7.1976340$, or $\text{Age} = 11.1976340/0.6674626$, or $\text{Age} = 16.77642$.

8. For the bivariate dataset

X	0	0	2	2.1	2.1
Y	0	1	2	2	2.2

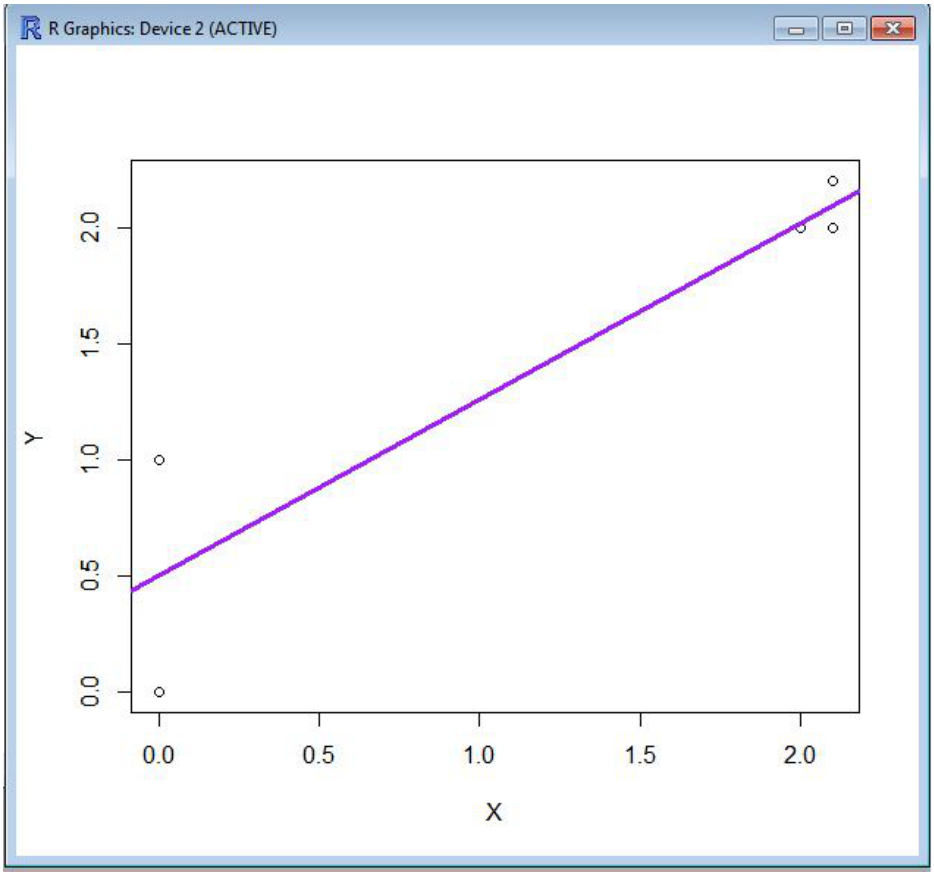
   the regression line has equation $\widehat{y} = 0.7583788\,x + 0.4996103$. Among the absolute values of the residuals, which is greatest?

   A) 0.4096 B) 0.4323 C) 0.4550 D) 0.4777 E) 0.5004
   F) 0.5231 G) 0.5458 H) 0.5685 I) 0.5912 J) 0.6139

Solution **E) 0.5004**

On inspecting the points, we realize that the regression line passes among the plotted points in a position similar to that shown in the figure below. We can be reasonably sure that the largest absolute residual arises from either the point (0,0) or the point (0,1). Because the equation is $\widehat{y} = 0.7583788\,x + 0.4996103$, the fitted point for each of our two candidate observed points is $\big(0, (0.7583788)(0) + 0.4996103\big)$, or $(0, 0.4996103)$. The absolute resduals are 0.4996103 and $1 - 0.4996103$, or 0.4996103 and 0.5003897. The latter is the larger.

```
> X = c(0,0,2,2.1,2.1)
> Y = c(0,1,2,2,2.2)
> plot(X,Y)
> lin = lm(Y~X)
> abline(lin, col = "purple", lwd = 3)
> lin$coefficients
(Intercept)           X
  0.4996103   0.7583788
> abs(lin$residuals)
         1          2          3          4          5
0.49961029 0.50038971 0.01636789 0.09220577 0.10779423
```



Scatterplot of (X,Y) and Regression Line

9. Let X be the year and Y the total outstanding credit card debt in billions of U.S. dollars. Let W = ln(Y). Values are tabulated below. Pearson's linear correlation coefficient for X and Y is 0.9691283. Pearson's linear correlation coefficient for X and W is 0.9771093.

							Mean	Sd
X	2002	2003	2004	2005	2006	2007	2004.5	1.870829
Y	756.7	774.6	807.9	839.5	905.4	981.8	844.3167	85.46263
W	6.628967	6.652347	6.694438	6.732806	6.808377	6.889388	6.734387	0.09893643

For the plotted ordered pairs (X,W), what is the W-intercept b of the regression line $\widehat{\mathrm{W}} = m\,\mathrm{X} + b$? (The next problem will use your calculation of this regression line.)

A) −98.0791 B) −96.8445 C) −95.6099 D) −94.3753 E) −93.1407
F) −91.9061 G) −90.6715 H) −89.4369 I) −88.2023 J) −86.9677

Solution B) −96.8445

The intercept is equal to

$$\text{mean(W)} - r_{\mathrm{X\,W}}\,\frac{sd(\mathrm{W})}{sd(\mathrm{X})}\,\text{mean(X)} = 6.734387 - 0.9771093\left(\frac{0.09893643}{1.870829}\right)2004.5 = -96.84453.$$

For use in the next problem, we note the value of the slope m:

$$m = 0.9771093\left(\frac{0.09893643}{1.870829}\right) = 0.05167319.$$

10. Transform the regression line $\widehat{\mathrm{W}} = m\,\mathrm{X} + b$ of the preceding problem to a "regression curve" $\widehat{\mathrm{Y}} = f(\mathrm{X})$ for an appropriate function f. In 2007, debt of all types soared and the observed value of Y for X = 2007 was significantly greater than the fitted value $f(2007)$. Then the year 2008 happened: the credit bubble burst and the financial system nearly collapsed. After 2008, the value of Y fell significantly, and, as of early 2017, it has not yet returned to the level of 2007. For example, the observed value of Y for 2009 was 898.0. To illustrate the danger of extrapolation, find the value of Y for 2009 that is predicted by the regression curve.

A) 1005 B) 1012 C) 1019 D) 1026 E) 1033
F) 1040 G) 1047 H) 1054 I) 1061 J) 1068

Solution I) 1061

The transformation is

$$\widehat{\mathrm{Y}} = \exp(m\,\mathrm{X} + b) = \exp(0.05167319\,\mathrm{X} - 96.84453).$$

We let X = 2009 in this equation to obtain

$$\widehat{\mathrm{Y}} = \exp(0.05167319 \times 2009 - 96.84453) = \exp(6.966909) = 1060.938.$$

11. For paired data $(\mathrm{X}_1, \mathrm{Y}_1), (\mathrm{X}_2, \mathrm{Y}_2), \ldots, (\mathrm{X}_{11}, \mathrm{Y}_{11})$, the variance of Y is 311.4909, the sum of the squares of the errors of the linear model for Y X is 88.28182, and the Pearson correlation coefficient r is positive. What is the value of r?

A) 0.83188 B) 0.8489 C) 0.8660 D) 0.8831 E) 0.9002
F) 0.9173 G) 0.9344 H) 0.9515 I) 0.9686 J) 0.9857

Solution J) 0.9857

Given that var(Y) = 311.4909 and N = 11, we have SST = $(11-1)\,311.4909 = 3114.909$. Also, SSE = 88.28182. Therefore SSR = SST − SSE = 3114.909 − 88.28182 = 3026.627. It follows that r^2 = SSR/SST = 3026.627/3114.909 = 0.9716582 and $r = \sqrt{0.9716582} = 0.9857272$.

The information that was given was obtained in the following R session:

```
> X = 1:11
> Y = c(3,8,10,22,28,31,40,38,46,53,52)
> N = length(Y)
> N
[1] 11
> var(Y)
[1] 311.4909
>  sse.ssr.sst = function(X,Y)
+   {
+    N = length(X)
+    m = cor(X,Y)*sd(Y)/sd(X)
+    b = mean(Y) - m*mean(X)
+    Yhat = m*X + b
+    sse = sum( (Y-Yhat)^2 )
+    ssr = sum( (Yhat-mean(Y))^2 )
+    sst = sum( (Y-mean(Y))^2 )
+    c(sse,ssr,sst)
+   }
> sse.ssr.sst(X,Y)
[1]   88.28182 3026.62727 3114.90909
> cor(X,Y)
[1] 0.9857273
```

12. Weetabix Cereal is at it again with a new series of collectible cards inserted in its boxes of delicious food product. Woodrow Wilson, Calvin Coolidge, Herbert Hoover, and Ronald Reagan are the cards in the company's *Alliterative Presidents of the United States* collection. Each box of Weetabix cereal contains one card, but the cards are not inserted with equal likelihood: 30% of the cereal boxes have a Woodrow Wilson card, 10% of the cereal boxes have a Calvin Coolidge card, 20% of the cereal boxes have a Herbert Hoover card, and 40% of the cereal boxes have a Ronald Reagan card. What is the probability p that the purchase of four cereal boxes will result in the complete subset that consists of the three Republican executive order signers (all but the Woodrow Wilson card)? Note: If the collection has CC, HH, and RR, then it is complete whether or not it also contains WW.

    Let us estimate p using the following table of random digits grouped in 30 blocks of 4 digits. Choose four positive integers w, c, h, and r that sum to 10. Assign the first w digits in the list 0,1,2,3,4,5,6,7,8,9 to W. Wilson, the next c digits to C. Coolidge, the next h digits to H. Hoover, and the final r digits to R. Reagan. Of course, you must choose w, c, h, and r 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 repetitiosly-initialed president followed by the digits by which his card is represented. It is easy to err without a visual reference.)

A) 0.0333 B) 0.0667 C) 0.1000 D) 0.1333 E) 0.1667
F) 0.2000 G) 0.2333 H) 0.2667 I) 0.3000 J) 0.3333

Solution **D) 0.1333**

The collections that have a CC (a 3 is present) are: 3324, 5063, 7335, 3366, 3997, 3956, 3169, 6134, 0773.
Of these, the collections that have a H.Hoover (a 4 or 5 is present): 3324, 5063, 7335, 3956, 6134.
Of these, the collections that have a RR (6, 7, 8, or 9 is present): 5063, 7335, 3956, 6134.

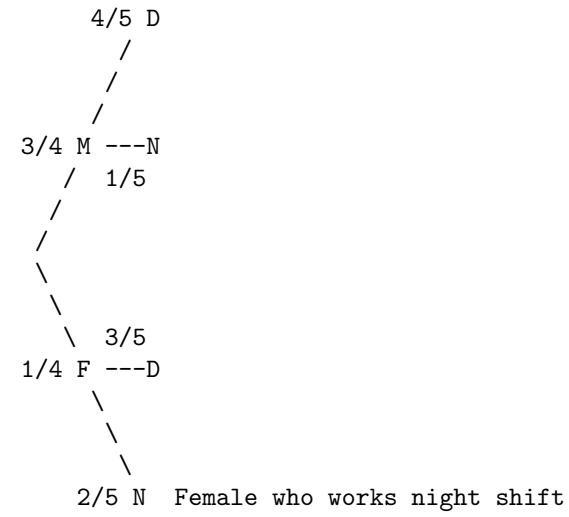
Our simulation has resulted in 4 successes in 30 trials. Our estimate of p is 4/30, or 0.1333.

13. At a plant there are 400 workers: 300 male and 100 female. Workers either work a day shift or a night shift: exactly 80% of the male workers work the day shift and exactly 60% of the female workers work the day shift. An employee is selected at random. What is the probability that the selected employee is a female who works the night shift? (The next two questions pertain to the same plant and the same set of workers.)

A) 0.05 B) 0.10 C) 0.15 D) 0.20 E) 0.25
F) 0.30 G) 0.35 H) 0.40 I) 0.45 J) 0.50

Solution **B) 0.10**

Using the following tree diagram, we see that the requested probability is $(1/4)(2/5) = 2/20 = 0.10$

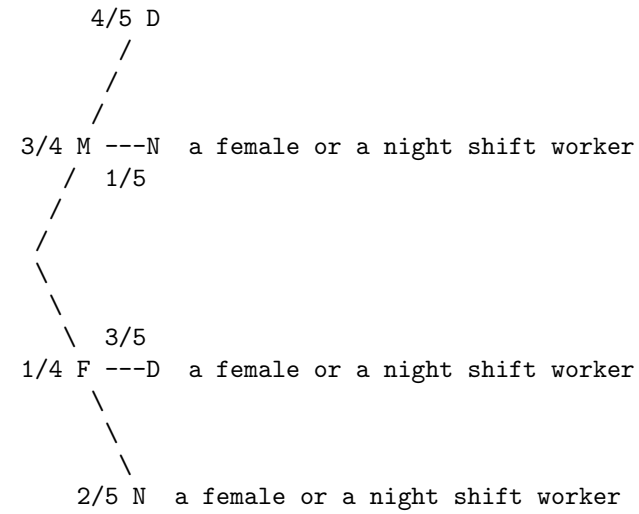


14. At the plant of the preceding problem, an employee is selected at random. What is the probability that the selected employee is either a female or a night shift worker?

A) 0.05 B) 0.10 C) 0.15 D) 0.20 E) 0.25
F) 0.30 G) 0.35 H) 0.40 I) 0.45 J) 0.50

Solution **H) 0.40** and **F) 0.30**

We can solve this with the tree diagram that follows; you will note that the interpretation of "... or ..." in the calculation that immediately follows the tree diagram is "and/or". The inclusion of the English word "either" in the statement of the problem, however, adds ambiguity. Moreover, the correct answer that results from the "exclusive or" (xor) interpretation is among the answer choices. Therefore, both interpretations of "or" have been accepted for this problem.



The requested probability is (3/4)(1/5) + 1/4 = 3/20 + 5/20 = 8/20 = 0.40 For the "exclusive or" interpretation, we use

P(female xor nightworker) = P(female and/or nightworker) - P(female nightworker) = 0.40 - (1/4)(2/5) = 0.30.

15. An urn contains three coins: two genuine, fair coins, and one gag, two-headed coin. You reach into the urn, pick one coin at random, and, without inspecting it, you toss it. What is the probability p_H that it lands head side up? Now suppose that somebody who witnessed the coin toss informs you of the outcome of the toss, which is that the coin did land head side up. Let $p_{G|H}$ be the probability that the coin you selected was the gag two-headed coin, given that the result of your coin toss was a head. What is the sum $p_H + p_{G|H}$?

A) 1/2 B) 7/12 C) 2/3 D) 3/4 E) 5/6
F) 11/12 G) 1 H) 13/12 I) 7/6 J) 5/4

Solution I) 7/6

Let H be the event "Heads". Let T be the event "Tails". Let F be the event "Fair coin selected". Let G be the event "Gag coinn selected". Then

$$\begin{aligned} p_{\mathrm{H}} &= \mathrm{P}(\mathrm{H} \cap \Omega) \\ &= \mathrm{P}\big(\mathrm{H} \cap (\mathrm{F} \cup \mathrm{G})\big) \\ &= \mathrm{P}\big((\mathrm{H} \cap \mathrm{F}) \cup (\mathrm{H} \cap \mathrm{G})\big) \\ &= \mathrm{P}\big(\mathrm{H} \cap \mathrm{F}\big) + \mathrm{P}\big(\mathrm{H} \cap \mathrm{G})\big) \\ &= \mathrm{P}(\mathrm{H} \,|\, \mathrm{F})\,\mathrm{P}(\mathrm{F}) + \mathrm{P}(\mathrm{H} \,|\, \mathrm{G})\,\mathrm{P}(\mathrm{G}) \\ &= \left(\frac{1}{2}\right)\left(\frac{2}{3}\right) + (1)\left(\frac{1}{3}\right) \\ &= \frac{1}{3} + \frac{1}{3} \\ &= \frac{2}{3}. \end{aligned}$$

Next, observe that the equation P(H ∩ G) = 1/3 was obtained as part of the preceding calculation. Therefore

$$p_{\mathrm{G|H}} = \mathrm{P}(\mathrm{G} \,|\, \mathrm{H}) = \frac{\mathrm{P}(\mathrm{G} \cap \mathrm{H})}{\mathrm{H}} = \frac{1/3}{2/3} = \frac{1}{2}.$$

Thus,

$$p_{\mathrm{H}} + p_{\mathrm{G|H}} = \frac{2}{3} + \frac{1}{2} = \frac{7}{6}.$$

16. In September 2016, the Centers for Disease Control and Prevention announced that the percentage of Americans without health insurance was 8.6%. If seven Americans were randomly selected at that time, what would have been the probablity that three or more were without health insurance?

A) 0.0002 B) 0.0171 C) 0.0340 D) 0.0509 E) 0.0678
F) 0.0847 G) 0.1016 H) 0.1185 I) 0.1354 J) 0.1523

Solution B) 0.0171

Let $p = 0.086$ and let $q = 1 - p = 0.914$. The probability of 0 uninsured among the seven is q^7. The probability of 1 uninsured among the seven is $7\,p\,q^6$. The probability of 2 uninsured among the seven is $(7 \times 6/2)\,p^2\,q^5$. The probability that three or more are without health insurance is therefore $1 - \big(q^7 + 7\,p\,q^6 + (7 \times 6/2)\,p^2\,q^5\big)$, or $1 - (0.53287258 + 0.35097297 + 0.09907114)$, or $1 - 0.9829167$, or 0.0170833.

There are three more-or-less natural ways to get the answer in R. Of these three ways, the middle one below seems most natural. Still, all three ways get you to the same place, and you can't argue with success.

```
> 1-pbinom(2, size = 7, prob = 0.086)
[1] 0.01708331
> pbinom(2, size = 7, prob = 0.086, lower.tail = FALSE)
[1] 0.01708331
> sum(dbinom(3:7, size = 7, prob = 0.086))
[1] 0.01708331
```

17. In a population consisting of persons with a history of cardiovascular disease, the prevalence of future cardiac events is 0.16. The sensitivity of exercise stress echocardiography is 85% and the specificity is 77%. What is the probability that a member of the population who stress tests positive will experience a cardiac event?

    A) 0.4131 B) 0.4757 C) 0.5383 D) 0.6009 E) 0.6635
    F) 0.7261 G) 0.7887 H) 0.8513 I) 0.9139 J) 0.9765

Solution A) 0.4131

Let CE denote a future cardiac event and let POS and NEG denote positive and negative stress tests respectively. We are given $\mathrm{P}(\mathrm{CE}) = 0.16$, $\mathrm{P}(\mathrm{POS} \,|\, \mathrm{CE}) = 0.85$, and $\mathrm{P}\left(\mathrm{NEG} \,|\, \mathrm{CE}^{\mathrm{c}}\right) = 0.77$. Then, acording to Bayes Theorem,

$$\begin{aligned}
\mathrm{P}(\mathrm{CE} \,|\, \mathrm{POS}) &= \frac{\mathrm{P}(\mathrm{POS} \,|\, \mathrm{CE})\,\mathrm{P}(\mathrm{CE})}{\mathrm{P}(\mathrm{POS} \,|\, \mathrm{CE})\;\; \mathrm{P}(\mathrm{CE}) + \mathrm{P}\left(\mathrm{POS} \,|\, \mathrm{CE}^{\mathrm{c}}\right)\,\mathrm{P}\left(\mathrm{CE}^{\mathrm{c}}\right)} \\
&= \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity})(1 - \text{prevalence})} \\
&= \frac{(0.85)(0.16)}{(0.85)(0.16) + (1 - 0.77)(1 - 0.16)} \\
&= 0.4131227.
\end{aligned}$$

18. For every live birth, let X be the number of babies born. In the United States, among live births, the probability of singleton births (X = 1) is 0.9649, the probability of twin births (X = 2) is 0.0339, the probability of triplet births (X = 3) is 0.0011, the probability of quadruplet births (X = 4) is 0.0001, and the probability of other multiple births (X > 4) is, within the accuracy of this problem, 0. What is the expectation of X? (The next problem is also about this random variable, X.)

    A) 1.0002 B) 1.0183 C) 1.0364 D) 1.0545 E) 1.0726
    F) 1.0907 G) 1.1088 H) 1.1269 I) 1.1450 J) 1.1631

Solution C) 1.0364

The set V of values that X assumes with positive probabilities is $V = \{1, 2, 3, 4\}$. We are given $f_X(1) = 0.9649$, $f_X(2) = 0.0339$, $f_X(3) = 0.0011$, $f_X(4) = 0.0001$. These four probabilities sum to 1.0000, which is consistent with the statement that $\mathrm{P}(\mathrm{X} > 4) = 0$ within the accuracy of the problem. We have

$$\mathrm{E}(\mathrm{X}) = (1)(0.9649) + (2)(0.0339) + (3)(0.0011) + (4)(0.000) = 1.0364.$$

19. For the random variable X of the preceding problem, what is the variance?

    A) 0.0065 B) 0.0222 C) 0.0379 D) 0.0536 E) 0.0693
    F) 0.0850 G) 0.1007 H) 0.1164 I) 0.1321 J) 0.1478

Solution C) 0.0379

```
> (1 - 1.0364)^2*0.9649  + (2 - 1.0364)^2*0.0339
+  + (3 - 1.0364)^2*0.0011 + (4 - 1.0364)^2*0.0001
[1] 0.03787504
```

20. In R, what number is the return of the following code?

```
> X = c( 0.6, 1.2, 0.8, 1.6, 1.0)
> sum(X[which(X > X^2)])
```

A) 0.6 B) 0.8 C) 1.0 D) 1.2 E) 1.4
F) 1.6 G) 1.8 H) 2.0 I) 2.2 J) 2.4

Solution E) 1.4

```
> X = c( 0.6, 1.2, 0.8, 1.6, 1.0)
> sum(X[which(X > X^2)])
[1] 1.4
```

21. In a poll, 1000 surveyees were asked to choose which one of three positions,
A) ...
B) ...
C) ... ,
was closest to their point of view. Responses were recorded as 0 for A, 1 for B, and 2 for C and stored in the fourth column of an R table named `Data`. There are several ways in R to count the number of responses for each of the three values, but suppose we set up a basic two-line procedure involving a for-loop to count the number of surveyees who responded with "C" and to assign that number to a variable named "count.C". Which one of the following strings could *not* have been used in the code (absent willfully deranged coding). A string is the sequence of characters, including spaces, between quotation marks.

A) "`count.C =`" B) "` = 0`" C) "`for(i in`" D) "`1:1000`" E) "`Data[i`"
F) "`,4]`" G) "` = 2`" H) "`{if(Data`" I) "` + 1`" J) "`}}`"

Solution G) " = 2"

Nothing should be assigned the value 2, which is what " = 2" does. The comparison to check whether a number equals 2 is " == 2".

```
> random.digits = sample(0:2, 120, replace = TRUE)
> Data = matrix(random.digits, ncol = 5, byrow = FALSE)
> count.C = 0
> for(i in 1:24){if(Data[i,4] == 2){count.C = count.C + 1}}
> count.C
[1] 5
> Data[1:24,4]
 [1] 1 1 2 1 0 0 1 1 0 2 0 1 1 1 1 1 2 1 0 0 2 2 0 0
```

22. A statistics professor is looking for an example to illustrate two variables that have near 0 linear correlation. A file of grades is at hand, so he decides to use the student IDs in the first column to explain the first exam scores found in the fifth column. He begins an R session with

```
> course.data = read.table(''C:/stats/grades.txt'')
> E1 = course.data[,5]
> ID = course.data[,1]
```

Which of the following commands would be the most *appropriate* continuation?

A) `lm(X,Y)` B) `lm(Y,X)`
C) `lm(X ~Y)` D) `lm(Y ~X)`
E) `lm(E1,ID)` F) `lm(ID,E1)`
G) `lm(E1 ~ID)` H) `lm(ID ~E1)`
I) `plot(ID,E1)` J) `abline(ID,E1)`

Solution I) `plot(ID,E1)`

Answer choices A, B, C, and D refer to variables that have not been defined. Answers E and F use incorrect syntax for `lm( )`. Answer choices G and H are possible continuations, but, given the expectation that E1 and ID have no linear relationship, it seems silly to find a linear relationship without any evidence to contradict a common sense expectation. Answer choice J uses the wrong syntax for `lm( )`, and, because there is no active graphics window, it is pointless to call on `abline( )`.

The answer choice `cor(ID,E1)` was omitted on purpose. That continuation might be proposed as an appropriate continuation. I would argue that such a proposal ignores the moral of one of the homework exercises. However, it might be counter-argued that an R session is not like flying a plane, an activity in which the sequential order of pressing buttons, flipping switches, and pulling levers might be crucial. No lives will be lost if the order of `plot(ID,E1)` and `cor(ID,E1)` are reversed. You can include the command `lm(E1 ID)` in that statement as well. However, the point of the homework exercise involving the Anscombe quartet of datasets is that the absence of a linear relationship does not prevent a high linear correlation coefficient, nor does a high linear correlation coefficient ensure a linear relationship. Obtaining visual evidence by scatterplotting bivariate data is the appropriate first step to take before calculating Pearson r or a regression line.

23. A statistics professor has ascertained that the first exam scores in his course can be used to explain the second exam scores in his course and that a regression line to illustrate the relationship is warranted. The first and second exam scores are in columns 5 and 6, repectively. Which command is the appropriate continuation of the indicated R session?

```
> course.data = read.table(''C:/stats/grades.txt'')
> E1 = course.data[,5]
> E2 = course.data[,6]
```

A) `abline(lm(X,Y))` B) `abline((lm(Y,X)`
C) `abline(lm(X ~Y)` D) `abline(lm(Y ~X)`
E) `abline(lm(E1,E2))` F) `abline(lm(E2,E1))`
G) `abline(lm(E1 ~E2))` H) `abline(lm(E2 ~E1))`
I) `points(E1,E2)` J) `plot(E1,E2)`

Solution J) `plot(E1,E2)`

The commands `abline( )` and `points( )` can be used only to superimpose a new plot in the existing active graphics window. Answers (A) through (I) are therefore ruled out. Additionally, several answers have incorrect syntax (A, B, E, and F), some have undefined variables (B, C, and D), one has the wrong explanatory and response variables (G). Although answer choices (H) and (I) will not result in error messages, they won't do anything because there is no active graphics window. Creating that window by entering `plot(E1,E2)` to obtain a scatterplot of the data is the appropriate continuation.

24. A student taking a statistics exam, on reaching the last question, reflected on what had come before. On looking back at the 23 previous problems, he realized that he had guessed on 12 of them. For each of his guesses, the probability of a correct guess was 0.10. Each guess was independent of the other guesses. He figured he needed 4 or more correct guesses to escape a grade of C. It was too depressing to calculate the probability P of guessing correctly on 4 or more of the questions, but at least he knew how to answer the last problem on the exam, which was, and is, Indentify the R code for the probability P of 4 or more successes in a sequence of 12 independent Bernoulli trials if the probability of success in each trial is 0.10. Among the choices, more than one is correct. If you answer with any correct choice, you get credt.

A) `dbinom(4, prob = 0.10, size = 12)`
B) `dbinom(4:12, prob = 0.10, size = 12)`
C) `sum(dbinom(4:12, prob = 0.10, size = 12))`
D) `pbinom(3, prob = 0.10, size = 12)`
E) `pbinom(4, prob = 0.10, size = 12)`
F) `pbinom(3, prob = 0.10, size = 12, lower.tail=FALSE)`
G) `pbinom(4, prob = 0.10, size = 12, lower.tail=FALSE)`

H) `1-pbinom(3, prob = 0.10, size = 12)`
I) `qbinom(4, prob = 0.10, size = 12)`
J) `1 - qbinom(3, prob = 0.10, size = 12)`

Solution **C or F or H**

```
> dbinom(4, prob = 0.10, size = 12) # A) Incorrect - this is the probability of exactly 4
[1] 0.02130813
> dbinom(4:12, prob = 0.10, size = 12) # B) Incorrect - these are the probs of exactly 4,5, ...,12
[1] 2.130813e-02 3.788111e-03 4.910515e-04 4.676681e-05 3.247695e-06
[6] 1.603800e-07 5.346000e-09 1.080000e-10 1.000000e-12
> sum(dbinom(4:12, prob = 0.10, size = 12)) #C) Correct
[1] 0.02563747
> pbinom(3, prob = 0.10, size = 12) # D) Incorrect - the probability of at most 3
[1] 0.9743625
> pbinom(4, prob = 0.10, size = 12) # E) Incorrect - the probability of at most 4
[1] 0.9956707
> pbinom(3, prob = 0.10, size = 12, lower.tail = FALSE) # F) Correct
[1] 0.02563747
> pbinom(4, prob = 0.10, size = 12, lower.tail = FALSE) # G) Incorrect - prob of at least 5
[1] 0.004329343
> 1-pbinom(3, prob = 0.10, size = 12) # H) Correct
[1] 0.02563747
> qbinom(4, prob = 0.10, size = 12) # I) Incorrect - qbinom is not a probability
[1] NaN
Warning message:
In qbinom(4, prob = 0.1, size = 12) : NaNs produced
> 1 - qbinom(3, prob = 0.10, size = 12) # J) Incorrect - qbinom is not a probability
[1] NaN
Warning message:
In qbinom(3, prob = 0.1, size = 12) : NaNs produced
```

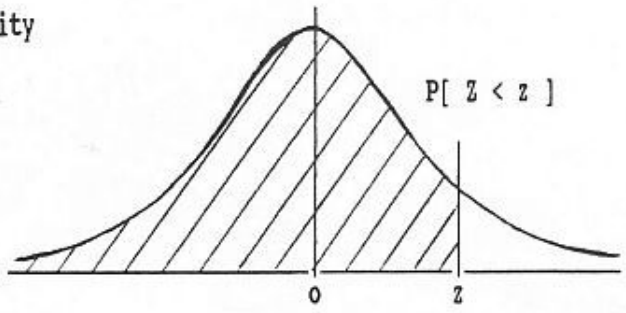
STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7020	0.7054	0.7089	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000