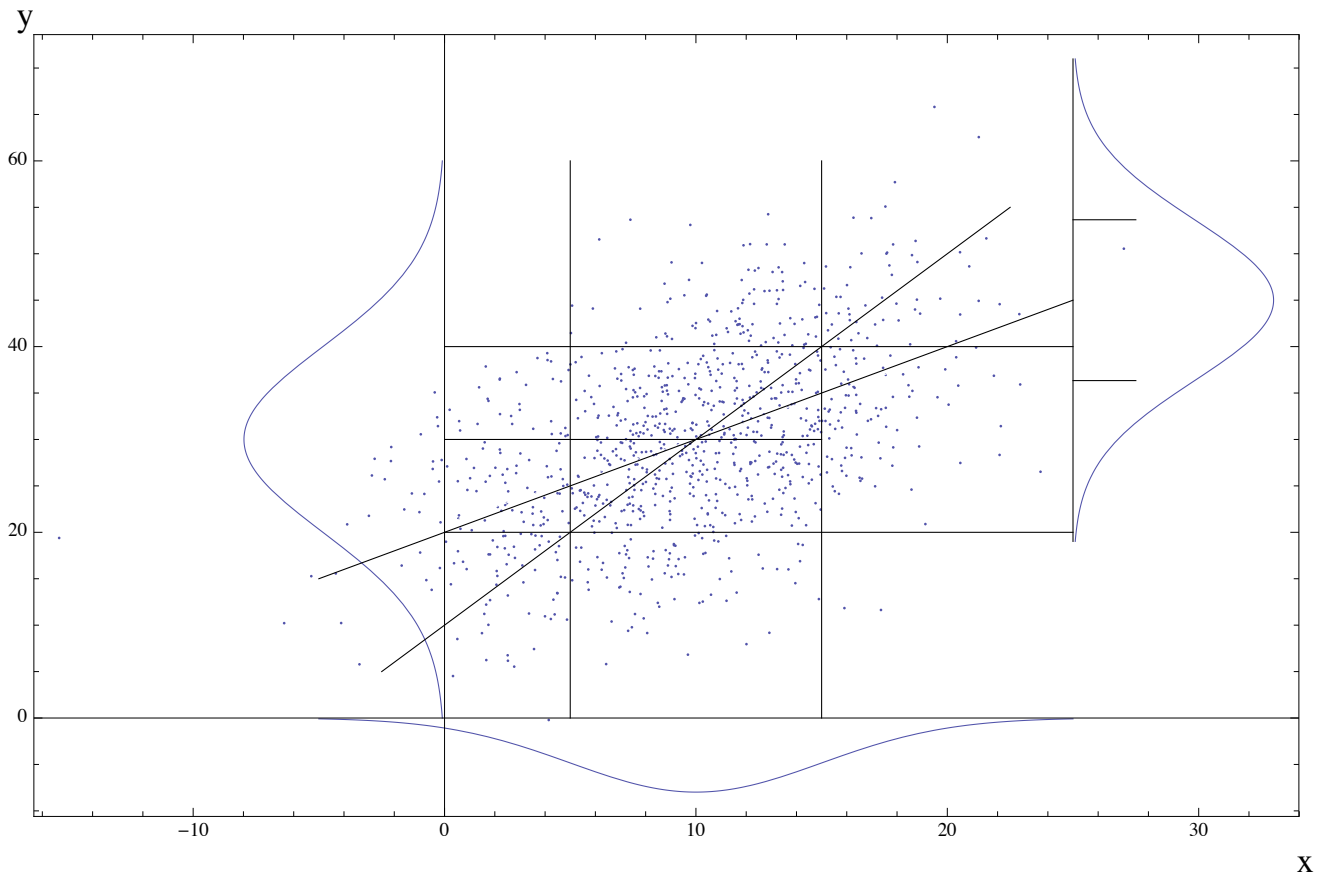
Questions 1 through 9 refer to the figure below. The distribution of (x, y) is 2-dimensional normal (bi-variate normal) vaguely represented by a large sample of points (x, y). All curves plotted are for the population not for the sample.



1. Determine the population mean of x denoted μ_x.
Ans. 10 from bottom bell curve.

2. Determine the population standard deviation of x denoted σ_x.
Ans. 15 (see 68% lines) in bottom bell curve.

3. Determine the *marginal probability* that y lies in the range [30, 40] denoted P(30 < Y < 40) .
Ans. .68 / 2 = .34 from 68% lines in left side bell curve (marginal density for Y).

4. Determine the *correlation* R between x and y.
Ans. R = 0.5 since the regression line has half the slope as the naive line as seen in their relative positions where they intercept the vertical line at x = 15.

5. **Using the naive line**, determine a *predicted value* of y for x = 25. Does it look as though this prediction is a good one?
Ans. The height y of the *naive line* at x = 25 is y = 60. This prediction is far to the left tail of the normal density for y conditional on x = 25 (see upper right bell curve).

6. **Using the regression line**, determine a *predicted value* of y for x = 25. Does it look as though this prediction is a good one?
Ans. The *regression line* at x = 25 is at height y = 45. This is right at the mean of the conditional distribution of y when x = 25. It seems y = 45 is the most reasonable predictor of y when x = 25.

7. Determine standard deviation of the bell curve in the upper right. It is the standard deviation of the *error of prediction* of y when **using the regression line** at x = 25. Do it by eye then confirm that your answer agrees with $\sqrt{1 - R^2}\,\sigma_y$.
Ans. By eye the standard deviation of the upper right bell curve is the gap between its mean and one of its 68% lines, which is around 8.5. The calculation using R = 0.5 and standard deviation of y = 10 gives $\sqrt{1 - R^2}\,\sigma_y = \sqrt{1 - 0.5^2}\,10 = 0.866$.

8. Determine standard deviation of the *error of prediction* (**using the regression line**) at x = 12.
Ans. The standard deviation of the error of prediction is the sd calculated in #7 above and iss the same for every value of x. So it is 0.866 for x = 12 as well as for x = 25. The upper right bell curve is just slid left & down along the regression line to the position at x = 12.

9. Determine a numerical *interval* for which the conditional probability of y being in the *interval* is around 0.68 if it is known that x = 25.
Ans. The obvious choice is 45 +/- 8.66. 45 is the predicted value and 8.66 is the sd of the distribution of y given x = 25.

Questions 10 through 12 assume that IQ in a particular population is distributed as normal with mean 100 and standard deviation 15.

10. Determine the standard score z of a person having an IQ of 124.
Ans. (124-100) / 15 = 24 / 15 = 1.60.

11. Use the table of areas under the standard normal curve to determine the fraction of the population having IQ < 124. This fraction is the fraction of the standard normal to the left of the z-score of 124.

Ans. Enter z = 1.60 to the margins of the table of P(std normal < z) getting

z	**.00**
1.6	**0.9452**

So P(std normal < 1.60) = 0.9452 to table accuracy.

12. What is the IQ you need to have in order to surpass 83% of the population? This would place you at the 83rd percentile of IQ.

a. To find out, first find the 83rd percentile of z by entering 0.83 (or the closest value near it) in the body of the table of left-tail z areas, then read off the z-value.
Ans. Enter 0.83 (or nearest entry) to the BODY of the table of P(std normal < z). Then read off z at the margins.

z	**.05**
0.9	**0.8289 (closest to 0.83 in the table body)**

Don't be confused by the answer z = 0.95 as it is a fluke that it reminds us of 95% as used in the rule of thumb. What we found is that Pstd normal < 0.95) = 0.8289 ~ 0.83. We wanted the z that would do that.

b. After you find the 83rd percentile of z convert this to the 83rd percentile of IQ according to IQ = 100 + 15 z. So what is the 83rd percentile of IQ?
Ans. IQ = 100 + 15 z ~ 100 + 15 0.95 = 114.25.

Questions 13 through 15 deal with algebraic properties of sample mean, sample standard deviation and correlation.

13. If sample mean of a list x is 15.2 what is the sample mean of list 1.6x - 16?
Ans. 1.6 15.2 - 16 = 8.32.

14. If sample standard deviation s_x of a list x is 3.4 what is sample standard deviation of list 1.6x - 16?
Ans. 1.6 3.4 = 1.6 3.4 = 5.44. Had it been the list -1.6x - 16 the answer would be the same. Remember, sd is never negative.

15. If correlation between (x, y) is 0.6 what is the correlation between (2x-4, 3y+2)?
Ans. 0.6 since 2 and 3 have the same sign (otherwise the answer would be -0.6). The location changes 4 and 2 have no effect on correlation, nor do non-zero scale changes both of the same sign.

Questions 16 through 23 are about calculations of means, standard deviations, correlation in rerlation to the following data of (x, y) pairs (column means are recorded at the bottom):

x	y	x^2	y^2	xy
0	0	0	0	0
0	9	0	81	0
9	3	81	9	27
—	—	—	—	—
3.	4.	27.	30.	9.

16. Mean of x
Ans. 3.

17. Sample standard deviation of x =
Note: Use a calculator program to get it then confirm in a separate calculation that it is equal to

$$s_x = \sqrt{\frac{n}{n-1}}\ \sqrt{\text{mean of } x \text{ squares} - \text{square of } x \text{ mean}}$$

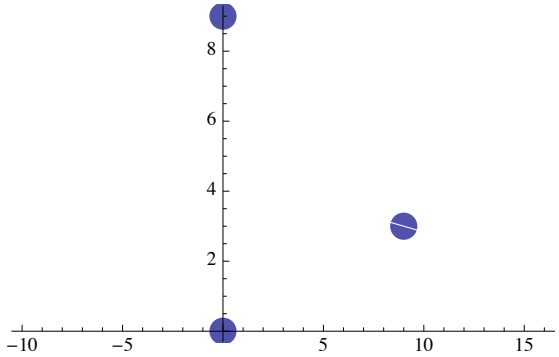
Ans. $\sqrt{\frac{3}{2}}\ \sqrt{27 - 3^2} = 5.196.$

18. Correlation R =

Ans. $\mathbf{R = \dfrac{9 - 3\ 4}{\sqrt{27 - 3^2}\ \sqrt{30 - 4^2}} = -0.19}$

Note: For the remaining problems we work with the naive and least squares line. Were the data *normal* then the least squares line would be (very close to) the regression line.

19. Plot the three points (x, y) and locate the point of means (x mean, y mean) then mark another point (x mean + s_x, y mean + s_y) and join these into a line. What is this line?



Ans. It is the *least squares* line.

20. On the same plot as #19 locate the point of means (x mean, y mean) then mark another point (x mean + s_x, y mean + R s_y) and join these into a line. What is this line?

Ans. Draw in the regression line joining (mean x, mean y) = (3, 4) to the point

$\mathbf{(3 + \sqrt{27 - 3^2}\ , 4 - 0.19\ \sqrt{30 - 4^2}\ ) \sim (7.24, 3.29)}$

In this case it will be the same as the *regression line* joining the vertical strip means, i.e. the line joining points (0, 4.5) and (9, 3).

21. What is the value predicted for y if x = 0?

Ans. 4.5 = the point on the least squares line at x = 0.

22. Does this plot appear normal?

Ans. No. It gives no evidence characteristic of data from a continuous probability model having elliptical contours. However the vertical strip means do plot as a line for this data. So there is a regression line and it must be the same as the least squares line (defined as the line passing through (mean x, mean y) and through (mean s + sd x, mean y + r sd y)).

23. Is the least squares line for this data the same as the regression line?
Ans. Yes. It is always the case when the plot of vertical strip means is a straight line, even if the data is not normal.

Questions 24 through 28 concern T and normal based confidence intervals and the margin of error. Throughout, we suppose a random sample is selected with-replacment and with equal probability from a population (the distribution of the population x scores is not necessarily normal unless it is specifically assumed in a problem). Suppose the sample mean is 23.2, the sample standard deviation is 4.2, the sample size is 50, and the population size is 4800.

23a. Give the *point estimate* of the population mean based on this data.
Ans. The sample mean is often used as a point estimator of the population mean. So the answer could be 23.2 (if the sample mean is used). If we are sampling incomes of the population of michigan people and one of our sample persons has an income of 3.3 million dollars we would likely drop that person from the sample. After all the Michigan per-person mean income is little changed if we exclude those with a very large income but the sample mean is very much impacted by including such persons.

24. Give the estimated margin of error for your estimate of #23.

Ans. $1.96\ s / \sqrt{n} = 1.96\ 4.2 / \sqrt{50}$ is commonly used.

25. Give the 95% confidence interval for the population mean based on this data.

Ans. sample mean +/- $1.96\ s / \sqrt{n} = 23.2 +/- 1.96\ 4.2 / \sqrt{50} = (22.0358, 24.3642)$.

26. Your answer to #25 is an interval calculated from data without knowlege of the population mean. What is the *approximate probability* that such an interval covers the population mean? That means out of all possible intervals constructed from random samples.
Ans. 0.95. In general, the approximate probability with which a confidence interval covers is close to the intended confidence level if everthing is done properly.

NOTE: It is not correct to say that the probabiliity the interval (3.03582, 5.36418) covers the population mean is around 0.95. Rather, the METHOD of producing such intervals has a round 95% hit rate. We don't know whether our confidence interval is one of the 5% misses or not.

27. If instead of sampling with replacement we sample *without replacement and with equal probability* what would be your answer to #24?

Ans. $1.96\ (s / \sqrt{n})\ \sqrt{(N-1)/(N-n)}$

$= 1.96\ (4.2 / \sqrt{50})\ \sqrt{(4800-50)/(4800-1)}$

It is virtually unchanged from #24 because n is far smaller than N so the *finite population correction* $\sqrt{(4800-50)/(4800-1)} = 0.994882 \sim 1$.

28. If instead of sampling with replacement we sampled *without replacement* what would be your answer to #25?

Ans. $\bar{X} +/- 1.96\ (s / \sqrt{n})\ \sqrt{(N-n)/(N-1)}$

$= 23.2 +/- 1.96\ (4.2 / \sqrt{50})\ \sqrt{(4800-50)/(4800-1)}$

29. If instead we had this same data but from a sample of only n = 6 *and if the population distribution is known to be close to normal* what number would we use in place of the z-score 1.96 in #25? What is the applicable degrees of freedom?

Ans. Look to the T table in the column with "confidence 95%" at the bottom and with degrees of freedom 6-1 = 5.

df	
5	2.571
∞	1.96 ⟸ NOTE: This line = fragment of z-table info.
Confidence Levels	95%

30. For a sample of n = 6 from a *normal population* give the 95% confidence interval for the population mean. Notice it is widened as compared with the n = 50 case since n is smaller and therefore $s / \sqrt{6}$ is larger than $s / \sqrt{6}$ but also the applicable T score exceeds 1.96. THE T CONFIDENCE INTERVAL IS EXACT, MEANING IT ACHIEVES EXACTLY THE NOMI-

NAL 0.95 COVERAGE PROBABILITY IF THE POPULATION DISTRIBUTION IS INDEED NORMAL, THE T-TABLE IS PERFECTLY ACCURATE AND ALL CALCULATIONS ARE INFINITE PRECISION.

Ans. 23.2 +/- 2.571 (4.2 / $\sqrt{6}$) = (18.7917, 27.6083). Notice this is far wider (less precise) than #25, reflecting the less conclusinve evidence provided by a sample of only n = 6.