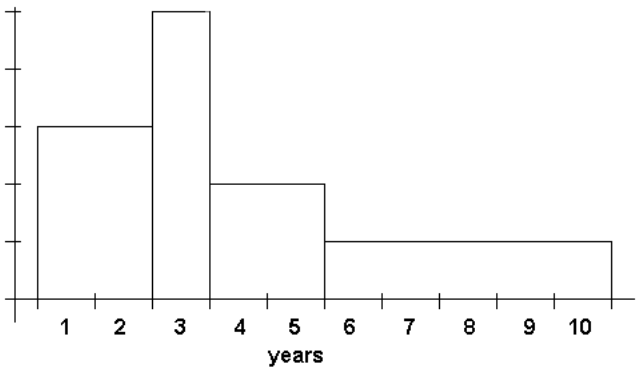
February 23, 2010

Math 102 / Core 143 BX — Exam I

An unsimplified answer like $12\sqrt{3.51} + 6/7$ is usually worth more than 23.3, because it is easier to understand where it came from.

1. *(20 points)* Consider the histogram of number of years of service of employees at a firm (in whole numbers of years):



   (a) What is the median of the data?

   (b) Approximately what is the average?

   (c) Which direction is the skew, if any?

   (d) Using the "density scale" in which the total area of the bars is 100%, what value should go on the lowest hash mark (above 0) on the vertical axis?

2. *(25 points)* A botanist studying a certain type of plant has many samples in pots in her greenhouse. She records the heights in centimeters, measuring from the level of soil in the pots. She finds that three are 8 cm tall, one is 9 cm, two are 11 cm, and one each 12 and 13 cm.

   (a) What is the average height?

   (b) What is the standard deviation of the heights?

   (c) What percent of the data is within one standard deviation of the average? (There are two equally acceptable answers, depending on how you interpret "within".)

   (d) In view of your answer to (c), would you describe the heights as normally distributed?

   (e) She decides to record the heights in decimeters (10 cm = 1 dm) measuring from the tabletop rather than the top of the soil. The soil level in each pot is 15 cm (i.e., 1.5 dm) above the tabletop. What are the new average and standard deviation?

3. *(7 points)* Heights of female college students are normally distributed, with an average of 67 inches and a standard deviation of 4 inches. Approximate the 80-th percentile for height.

4. *(20 points)* The botanist of an earlier problem also has plants of another species in her lab and is testing the effect of a certain fertilizer on the height of this species. She finds the following summary data:

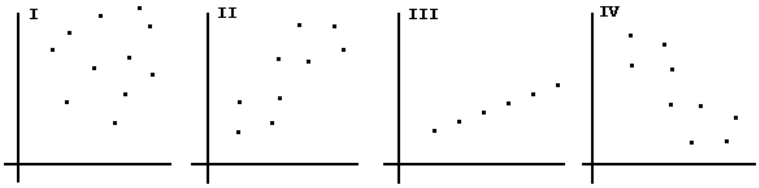
	fertilizer (cc)	height (cm)	
mean	50	30	$r = 0.8$
std dev	16	8	

   (a) Use linear regression to project the height of a plant that receives 58 cc of fertilizer.

   (b) Use linear regression to estimate the amount of fertilizer given to a plant 22 cm high.

   (c) How far is your estimate in (b) likely to be off?

   (d) Among the scatter diagrams below, which is most likely to be the graph of fertilizer amounts vs. heights?



5. *(8 points)* Classify the following as categorical or numerical variables.

   (a) Heights of plants

   (b) Sex, coded as 0 for males and 1 for females

   (c) Responses of "agree strongly", "agree", "neutral", "disagree", "disagree strongly" to a statement of opinion

   (d) Diastolic blood pressure

6. *(10 points)* [The point of this exercise is to explore the mantra "Correlation does not imply causation" — what causes what?] In a small community, an unusually large number of people seem to have colds, and many of the sufferers are found to frequent the community's only restaurant. Answer each of the following in a sentence or two:

   (a) How might going to the restaurant have caused its customers to have colds?

   (b) How might the fact that they have colds caused people to eat at the restaurant?

   (c) What other factor may have caused people both to get colds and to eat at the restaurant?

7. *(10 points)* Related to the articles "The Sins of the Fathers, Take 2" and "What Alters Our Genes" by Sharon Begley: Is epigenetics science? Is it a denial of Darwinism?

Some possibly useful formulas:

$$\sqrt{\text{average of } (x - AV_x)^2} \qquad SD_y \cdot \sqrt{1 - r^2}$$

$$r = \text{average of } ((x \text{ in std units}) \cdot (y \text{ in std units}))$$

$$z = \frac{x - AV_x}{SD_x} \qquad x = z \cdot SD_x + AV_x$$

$$y - AV_y = (\text{sign of } r)\frac{SD_y}{SD_x}(x - AV_x) \qquad y - AV_y = r\,\frac{SD_y}{SD_x}(x - AV_x)$$

Normal table (Area between $-z$ and z)									
z	Area(%)	z	Area(%)	z	Area(%)	z	Area(%)	z	Area(%)
0.0	0.0	0.9	63.19	1.8	92.81	2.7	99.31	3.6	99.968
0.05	3.99	0.95	65.79	1.85	93.57	2.75	99.4	3.65	99.974
0.1	7.97	1	68.27	1.9	94.26	2.8	99.49	3.7	99.978
0.15	11.92	1.05	70.63	1.95	94.88	2.85	99.56	3.75	99.982
0.2	15.85	1.1	72.87	2	95.45	2.9	99.63	3.8	99.986
0.25	19.74	1.15	74.99	2.05	95.96	2.95	99.68	3.85	99.988
0.3	23.58	1.2	76.99	2.1	96.43	3	99.73	3.9	99.99
0.35	27.37	1.25	78.87	2.15	96.84	3.05	99.771	3.95	99.992
0.4	31.08	1.3	80.64	2.2	97.22	3.1	99.806	4	99.9937
0.45	34.73	1.35	82.3	2.25	97.56	3.15	99.837	4.05	99.9949
0.5	38.29	1.4	83.85	2.3	97.86	3.2	99.863	4.1	99.9959
0.55	41.77	1.45	85.29	2.35	98.12	3.25	99.885	4.15	99.9967
0.6	45.15	1.5	86.64	2.4	98.36	3.3	99.903	4.2	99.9973
0.65	48.43	1.55	87.89	2.45	98.57	3.35	99.919	4.25	99.9979
0.7	51.61	1.6	89.04	2.5	98.76	3.4	99.933	4.3	99.9983
0.75	54.67	1.65	90.11	2.55	98.92	3.45	99.944	4.35	99.9986
0.8	57.63	1.7	91.09	2.6	99.07	3.5	99.953	4.4	99.9989
0.85	60.47	1.75	91.99	2.65	99.2	3.55	99.961	4.45	99.9991

Math 102 / Core 143 — Solutions to Exam I

1. (a) The area in the bars can be divided into 20 equal blocks (one horizontal unit by one vertical unit — so the right bar has 6 blocks, the next has 5, and so on). Over half are to the left of or on 3, so 3 is the median.

   (b) The point of balance seems to be about 4. (In fact, if we assume that the fractions of the list are as suggested by the bars, the average is 4.1.)

   (c) The skew (i.e., the tail) is to the right. (This is also shown by the fact that the average is larger than the median.)

   (d) Each of the 20 blocks represents 5% of the area, so the first hash mark represents 5% per year on the density scale, to make the total area (height times years) under the graph come out to 100%.

2. (a) $(3 \cdot 8 + 9 + 2 \cdot 11 + 12 + 13)/8 = 10$ cm

   (b) $\sqrt{[3(8-10)^2 + (9-10)^2 + 2(11-10)^2 + (12-10)^2 + (13-10)^2]/8} \approx 1.87$ cm

   (c) If "within" means <u>less than</u> 1.87 cm, 3 out of 8, or 37.5% of the 8 heights are within one standard deviation of the average. If "within" means 1.87 cm <u>or less</u>, rounded up to 2, 7 out of 8, or 87.5%, are within one standard deviation of the average.

   (d) Neither 37.5% nor 87.5% is really close to 68%, so the data does not seem to follow the Empirical Rule, and hence is not really normally distributed. (Most small data sets are not.)

   (e) The average is divided by 10 and increased by 1.5, becoming 2.5 dm. The standard deviation is divided by 10, becoming 0.187 dm.

3. We want an area of 60% (the given 80% minus the mirror image on the right of the 20% on the left), and the normal table says that comes at standard units value of about 0.85, or in terms of height, $67 + .85(4) = 70.4$ inches.

4. (a) 58 cc of fertilizer is 0.5 SDs above average, so we should project that the height is $0.8(0.5) = 0.4$ SDs above average, or $30 + 0.4(8) = 33.2$ cm.

   (b) 22 cm is 1 SD below average, so we should project that the fertilizer level is $0.8(1) = .8$ SD below average, or $50 - .8(16) = 37.2$ cc.

   (c) The RMS error for regression: $16\sqrt{1 - .8^2} = 9.6$ cc.

   (d) II. The correlation of I is too low to be 0.8, that of III is too high, and that of IV is negative.

5. (a) Numerical. (b) Categorical. (c) Categorical ("ordinal" would be possible, but "numerical" would be misleading). (d) Numerical.

6. (a) Maybe someone at the restaurant — an employee or a frequent patron — has a cold and is contagious.

   (b) Maybe cold sufferers feel too ill to cook for themselves.

   (c) Maybe those who go out and socialize get colds from those they socialize with, and they also go out to eat.

7. In the answers to these questions, I'm looking for evidence that you have read and thought about the articles, not for agreement with my opinions. In this case, my sense is that epigenetics ("on top of genetics") is science, the study of how, even though organisms may have the same genes, they may be expressed differently; it can be studied at the level of the whole organism or at the level of molecular biology. It isn't a denial of Darwinism; rather, it is a refinement of Darwin's theory — and Darwinism, like any scientific theory, is always subject to being refined in similar ways.