October 3, 2006

Exam I — Math 102 / Core 143 CX

Points are in parentheses. Show your work to receive partial credit; an answer like $[1 - (6/7)^3]/4$ is worth more than 0.09257, because it displays your reasoning more clearly.

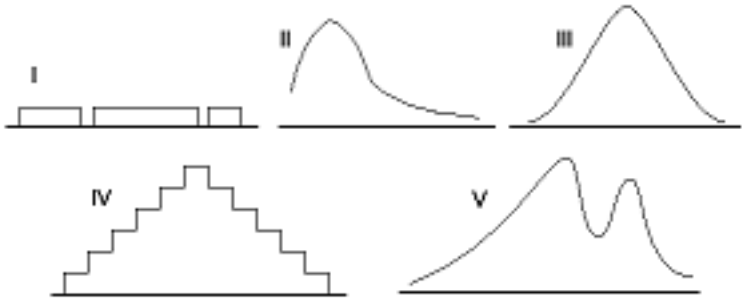
1. Identify the histogram below that best matches the data below. If none of the histograms match, sketch one that gives a better match than any given and explain.

   (a) Years of education from a large group of Americans

   (b) Numbers on the caller's markers from an old bingo game, missing several markers

   (c) Heights of a large group of nine-year-old boys

   (d) Numbers of ways to get a given number on a roll of a pair of dice



2, Find, for the list of 10 numbers

$$-4, -1, -1, -1, 0, 0, 0, 2, 2, 3 :$$

(a) the average (b) the standard deviation (c) the median

(d) the 35th percentile (e) the IQR

3. Over a two-year period, world oil production (in Mbbl/day) averaged 5, with a standard deviation of 2; and gasoline prices (in \$/gal) averaged 1.5 with a standard deviation of 0.8. The correlation was -0.5.

   (a) Is the sign of the correlation reasonable? Explain.

   (b) Estimate the gas price when the oil production is 6 Mbbl/day.

   (c) About how far should you expect your estimate in (b) to be off?

   (d) If you were to estimate the oil production when the gas price is 2.3, would your estimate be closest to 1, 4, 6 or 9?

4. (a) For a long list of numbers that has an average of 140 and a standard deviation of 20, use normal approximation to estimate the 80-th percentile.

   (b) For the same data as in (a) estimate the percentile rank of the datum 130.

(c) The list of 20 numbers

$$-14, -13, -12, -11, -10, -9, -8, -7, -6, -5, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14$$

has an average of 0 (that should be clear from the symmetry) and a standard deviation of not quite 10; so only 50% of the data is within one standard deviation of the average. How does this relate to the "Rule of Thumb" in the text?

(d) For the data in (c), without doing arithmetic, would you guess that the standard deviations of the shorter lists

$$-14, -13, -12, -11, -10, -9, -8, -7, -6, -5 \qquad \text{and} \qquad 5, 6, 7, 8, 9, 10, 11, 12, 13, 14$$

are equal to, less than, or greater than the standard deviation in (c)?

5. You want to know whether vitamin C alleviates the symptoms of a cold. Give a quick comment on each of the following methods of gathering data to answer the question.

   (a) Put a questionnaire on the web asking for responses from those who have taken vitamin C for cold symptoms.

   (b) Get a list of people who have purchased vitamin C, and interview as many as possible, including in your study the ones who say they took it for cold symptoms.

   (c) Get a list of cold sufferers, give them vitamin C and interview them for results.

   (d) Get a list of cold sufferers, give half of them (chosen at random) vitamin C, and interview all of them for results.

6. Related to the article, "The median isn't the message", by Stephen Jay Gould: Explain why Gould feels that his knowledge of statistics played a major role in saving his life.

Some possibly useful formulas:

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n}} \qquad z_x = \frac{x - \overline{x}}{s} \qquad r = \frac{\sum z_x z_y}{n}$$

$$\text{RMS error} = \sigma_y \sqrt{1 - r^2} \qquad y - \overline{y} = r \frac{s_y}{s_x}(x - \overline{x}) \qquad y - \overline{y} = (\text{sign of } r) \frac{s_y}{s_x}(x - \overline{x})$$

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

Solutions to Exam I

1. (a) V — the humps come from the many people who finish high school or college. (b) I — each number from 1 to 75 appears at most once, but some are missing. (c) III. (d) IV.

2. (a) $[-4 + 3(-1) + 3(0) + 2(2) + 3]/10 = 0$
   (b) $\sqrt{[(-4-0)^2 + 3(-1-0)^2 + 3(0-0)^2 + 2(2-0)^2 + (3-0)^2]/10} \approx 1.9$
   (c) 0, between the fifth and sixth numbers in the list
   (d) -1, between the third and fourth numbers in the list
   (e) The third quartile may be 0 or 2 or anything between, but the first quartile is surely -1, so their difference, the IQR, could be anything from 1 to 3.

3. (a) Yes: As oil production decreases, gas prices probably go up.
   (b) 6 Mbbl/day is $(6-5)/2 = 0.5$ standard deviations above average, so gas prices are projected to be $(-0.5)(0.5) = -0.25$ standard deviations above average, or $1.5+(-0.25)(0.8) = 1.3$ \$/gal.
   (c) We should expect it to be off by a "root mean square error for regression", $.8\sqrt{1-(.5)^2} \approx .69$
   (d) A gas price of 2.3 is one SD above the average, so we should expect oil production to be .5 SD's below average, or $5-(.5)(2) = 4$ (in other words, below average, but not <u>as much</u> below average as the gas price was above its average, both measured in standard units).

4. (a) The 80-th percentile has, between it and its corresponding value on the opposite side of the average, 60% of the data. From a normal table, that value, in standard units, is about .84, so in the original values it is about $140 + .84(20) = 156.8$.
   (b) 130 is half an SD below average, and from the normal table, the area below $-.5$ is $(100-38)/2 = 31$ percent. So 130 is about the thirty-first percentile.
   (c) The data aren't normally distributed and so need not (and clearly do not) follow the Rule of Thumb.
   (d) It should be smaller. The pieces each have the same SD, but their averages are different, so the joined group would have, in general, higher deviations from the new average of 0 than from their old averages (of -9.5 and 9.5 respectively).

5. (a) Volunteer samples are useless.
   (b) It's an observational study, and your subjects are probably biased in favor of vitamin C, but it's a start toward more reliable data.
   (c) Better still, but a control group would tell you whether symptoms improved because of the vitamin C or just because they were recovering from the cold.
   (d) Good! A controlled experiment with random assignment to groups (hopefully conducted double-blind, but we have no info about that) can establish that vitamin C actually causes an improvement in cold symptoms. But no scientific result is ever final.

6. Gould feels that his knowledge of statistics helped him to realize that, though the prognosis for most victims of his type of cancer is not good, he was among the people (younger than average, otherwise healthy and with access to good care) who had a better chance for long survival (in the right "tail" of survivors). The positive outlook this gave him helped to survive longer and finally to beat that cancer (though he died of another kind of cancer not long ago).