(1) Why is the mean an inappropriate measure of central tendency in a very skewed distribution?

The mean is not robust to outliers.

(2) a. What does $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)$ mean?

This notates the sum of the deviances from the mean.

b. What value does the above *always* equal?

Zero. This is the reason that we square the deviances when calculating the variance or sd.

(3) A researcher collected several sets of data. For each indicate which measure of central tendency would be best.

a. The following personality scores
0, 2, 3, 3, 8, 4, 9, 6, 7, 5, 6

Both the mean and median would be appropriate.

b. The following age scores
10,15,18,15,14,13,42,15,12,14,42

The median would be appropriate.

c. The following college years
senior, junior, junior, freshman, freshman, junior, sophomore, junior

Either the mode or median would be appropriate.

d. The following the political affiliations
Dem, Dem, Rep, Dem, Soc, Com, Com, Soc, Dem, Rep

The mode would be appropriate.

(4) You have subtracted the mean from each score in a sample that is approximately normally distributed. Which of the following deviation from the mean score is likely to have the highest frequency?

a. –12
b. –7
c. +2
d. +18

+2. The closer the number to the mean, the more frequent that number.

(5) a. Distinguish among the symbols σ and s.

σ = population standard deviation, s = sample sd

b. When are the above symbols accompanied by the square sign (^{2})?

Variance.

(6) Suppose you want to estimate population parameters from a sample of 30 scores. You decide to use the sample mean to estimate the population mean and then you estimate the population variance as

$$S_x^2 = \frac{1}{30}\left(597 - 30 \times (4.1)^2\right) = 3.09$$

a. What did you do wrong?

Should be N-1 in this denominator (population correction)

b. Should the estimate be larger or smaller?

Larger.

(7) For a distribution with a mean of 130 and a standard deviation of 15, approximately 68% of the scores will lie between which two scores? What about for approximately 95% of the scores?

68%: 115,145
95%: 100, 160

(8) The new statistician for the football team at MIT has calculated the following player statistics. Player A's average running yards per carry is 5 with a standard deviation of 3. Player B's average running yards per carry is 5 with a standard deviation of 10.

a. Which player is the more consistent yardage gainer? Why?

Player A. Smaller SD

b. Which player is more likely to break loose for a long yardage gain? Why? Player

B. Larger SD.

(9) Dr. Jones has administered a test to her students. She calculated an average of 86 and a standard deviation of 12.

a. What is the z-score of a student with a raw score of 80?
   -.5
b. What is the z-score of a student with a raw score of 98?
   +1
c. What is the raw score for a student for a student with a z-score of –1.5?
   68
d. What is the raw score for a student for a student with a z-score of +1?
   98

10) Two psychologists, Tversky and Kahneman, asked a group of subjects to carry out the following task. They were told that:

- Linda is 31, single, outspoken, and very bright. She majored in philosophy college. As a student she was deeply concerned with racial discrimination and other social issues, and participated in anti-nuclear demonstrations.

The subjects were then asked to rank the likelihood of various alternatives, such as:

(1) Linda is active in the feminist movement
(2) Linda is a bank teller
(3) Linda is a bank teller and active in the feminist movement

Tversky and Kahneman found that between 85% and 90% of subjects rated alternative (1) most likely, but alternative (3) more likely than alternative (2). Is it?

They call this phenomenon "conjunction fallacy", and note that it appears unaffected by prior training in probability and statistics. Explain why this is a fallacy. We are interested only in a mathematical explanation of the nature of the fallacy, not in your suppositions of psychological motivation.

Solution: Let the events A = ``Linda is a bank teller" and B = ``Linda is active in the feminist movement". Since alternative (3) corresponds to the event "A and B", it's less likely than alternative (2) which corresponds to the event "B" (event "A and B" is obviously a subset of event "B").

(11) What proportion of the area under the standard normal curve would you expect to be (round answers to the closest percent)

a. Between z=-1.2 and z=+0.6
   ~61%
b. Below z=1.4
   ~92%

c. Below z=-2.6
~.5%

d. Above z=-2.0
~97.5%

(12) What are the shape, mean, and standard deviation of a sampling distribution of means calculated from samples of size N taken from a normally distribution population with mean μ and variance σ^2?

Shape: Normal
Mean: Population Mean
SD: σ /sqrt(N) [N = # of samples, NOT # of trials]

(13) Given a standard deck of 52 playing cards determine the probability of the following events

a. of drawing a 7
4/52
b. of drawing a club
13/52
c. of drawing a 7 or a 10
4/52 + 4/52

d. of drawing a 7 or a club 4/52 +
13/52 – 1/52 (subtract for the 7 of
clubs)
e. of drawing a 7 then a 10 (without replacing the 7) 4/52 * 4/51

(14) In a sample with a mean of 46 and a standard deviation of 8, what is the probability of randomly selecting each of the following raw scores

a. a score above 62
2.5%

b. a score between 30 and 54
81.5%

c. a score below 3 8
16%

(15) What is the probability of getting 4 or 5 heads in 5 tosses of a

fair coin? ${}_5C_4 * .5^5 + {}_5C_5 * .5^5 = 5/32 + 1/32 = 6/32 = 3/16$

(16) Consider a population for which the mean is 53 and standard deviation is 15. Using a criterion of p=0.05 and both tails of the sampling distribution, which of the following samples (N=25) can be called unrepresentative of the population

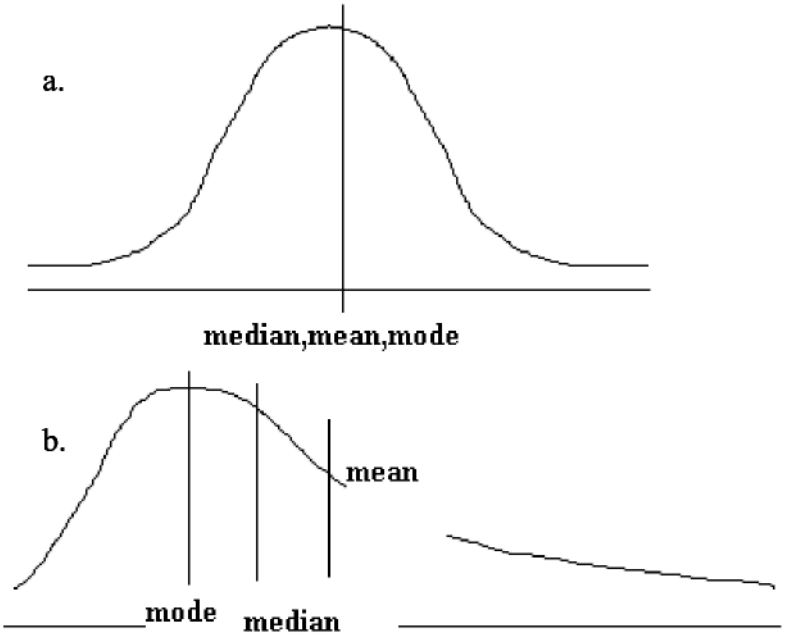
a. A sample with mean 56
15/5 = 3 (sigma of x) ; z = 56 – 53 / 3 = 1 (< 1.96)

b. A sample with mean 47: 47-53 / 3 = -2 (< -1.96)

(17) How many voters need to be polled to be 95% confident of having a 1% margin of error? Assume the actual percent of voters that favor each of the two candidates is 50%.

margin of error $e = \sigma_p * z = 1.96 * \sim(0.5(1-0.5)/n) = 0.01$; n = 9604

(18) For each of the following, note where the mean, median, and mode roughly fall.



19) Suppose 32,753 students nationwide take a standardized test for which the cutoff for passing is 1.2 standard deviations below the mean.
a) About how many students will pass?
b) What is the probability that all of the randomly selected 10 students will pass the test?

To solve: a) assume scores are normally distributes. p=P(pass) = P(score >mean-1.2sigma)=P((score-mean)/sigma>-1.2)=P(z>-1.2)=1-.1151=.8849
passing = 32,753*p=28,983
b) P(all 10 will pass) = p^10=0.2944028

20) There is a theory that the anticipation of a birthday can prolong a person's life. While examining this notion statistically, experimenters found that only 60 of 747 people whose obituaries were published in Salt Lake City newspapers died in the three month periods preceding their birthdays. Test the appropriate hypothesis at the 0.01 significance level.

Solution: Let p the proportion of the deaths that occur in the 3 month period preceding the birthday. Assuming that the deaths occur randomly uniform in the 12 month period of a year p is 1/4.We are testing the null hypothesys: H_0: p=1/4 versus H_1: p<1/4. The alternative is p<1/4 because we suspect that the proportion of the death in the 3 month period preceding the birthday is smaller (only 60 out of 747 deaths). Let Y be the number of deaths in the three month period preceding the birthday. use the normal approximation to the binomial and we obtain the test statistic: z_{obt}=(60-747 x 1/4)/ (sqrt(747* 1/4* 3/4)) = -10.71 Because z_{obt}=-10.71<-z_{crit}(0.01=-2.33 we reject the null hypothesis.

21) Is school performance getting worse? Each year the National Assesment of Educational Progress (NAEP) administers tests to nationwide sample of highschool graduates. The average score of students in 1973 was 55. In 1996 average score for a sample of 20 students was 52 with standard deviation of 10.

a) Is the difference of 3 real or due to chance? Use 5% significance level
b) Build a 95% confidence interval for true mean score in 1996.

Solve:
a) Test H_0: Mu=55 vs H_1: mu not equal 55. Since don't know true population variance of the scores, and sample size is small (20 <30) will use t-test.

t_{obt} = (m-55) / sqrt(SD^2/n) = (52-55) / sqrt(100/20) = -1.34 -- has t distribution with df=19. For 5% significance level (2-sided alternative) t_crit(0.025)=2.09 Since |t_obt|<t_crit accept H_0 and conclude that the difference of 3 in scores is due to chance.

b) Invert |t_obt| < t_crit:

-t_crit * sqrt(SD^2/n) < m-mu < t_crit * sqrt(SD^2/n)
-m - t_crit * sqrt(SD^2/n) < -mu < -m + t_crit * sqrt(SD^2/n)
m - t_crit * sqrt(SD^2/n) < mu < m + t_crit * sqrt(SD^2/n)
thus, 95% CI for mu is:
(m - t_crit * sqrt(SD^2/n); m + t_crit * sqrt(SD^2/n) = (52 - 2.09 *
sqrt(100/20); 52 + 2.09 * sqrt(100/20)) = (52-4.67; 52+4.67) = (47.33, 56.67)