AP Statistics
2000 Scoring Guidelines

These materials were produced by Educational Testing Service (ETS), which develops and administers the examinations of the Advanced Placement Program for the College Board. The College Board and Educational Testing Service (ETS) are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

The College Board is a national nonprofit membership association dedicated to preparing, inspiring, and connecting students to college and opportunity. Founded in 1900, the association is composed of more than 3,900 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three million students and their parents, 22,000 high schools, and 3,500 colleges, through major programs and services in college admission, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT™, the Advanced Placement Program® (AP®), and Pacesetter®. The College Board is committed to the principles of equity and excellence, and that commitment is embodied in all of its programs, services, activities, and concerns.

Question 1

Solution

a. Drug A produced average pain relief in the 55-70 range (or averaging approximately 65) for strengths between 210 and 400. Pain relief doesn't appear to depend on strength over the range 210 to 400.

b. Drug B didn't produce much (if any) pain relief for strengths less than about 270. For strengths between 270 and 400, pain relief increased steadily with dosage.

c. Drug A at strength 210: Choose drug A because the pain relief is about 65 (or in the 55 - 70 range) for all dosage levels whereas drug B needs to be given at 330 mg or higher to achieve pain relief of at least 50. Since the lowest dosage of drug A tested was 210 and all levels are about equally effective, prescribe 210 mg.

Scoring

Part (a) is

Essentially correct if	the answer includes **both** a statement that the pain relief is in the 55-70 range (or approximately 65) **and** that pain relief doesn't depend on strength.
Partially correct if	the answer includes only one of the two required statements.

Part (b) is

Essentially correct if	the answer includes **both** a statement that there is no pain relief for strengths below approximately 270 **and** that pain relief increases with strength above 270.
Partially correct if	the answer includes only one of the two required statements.

Question 1 — continued

part (c) is

Essentially correct if	response is drug A at strength 210, with justification of choice of drug **and** choice of dosage. Justification of drug A must involve explicit comparison to drug B.
Partially correct if	response is drug A at strength 210, with incomplete or no justification, OR response is drug A at a dosage other than 210 (or no mention of dosage) with justification of choice of drug A over drug B, OR response is drug B at strength 330 with justification of choice of dosage (e.g. because 330 is the lowest strength at which drug B gives at least 50% relief.)

4 Complete Response

Essentially correct on all three parts.

3 Substantial Response

Essentially correct on two parts and partially correct on the other part.

2 Developing Response

Essentially correct on two parts and incorrect on the other part,
OR
essentially correct on one part and partially correct on at least one other part,
OR
partially correct on all three parts.

1 Minimal Response

Essentially correct on one part,
OR
partially correct on two parts.

0 No credit

Question 2

Solution

a. The population of interest is adults who used the cave and the assumptions are:

1. the 20 measurements constitute a random sample **from the population of adults who used the cave**.

2. the adult foot length distribution is normal or approximately normal. (Some may state this assumption as "the population distribution is normal or the sample from the population is large (e.g., n >30)." This is acceptable.)

b. **Random sample is not reasonable**: This sample was not taken from the population of interest since the anthropologists took a random sample of footprints, not a random sample of adults who used the cave. There may be several different ways to explain that the sample was not taken from the population of interest. For example:

- the 20 observations may include several footprints from the same adult
- the footprints may be from children
- some of the original footprints may have eroded in time

Normality: Either of the responses below is acceptable.

Normality <u>is not</u> reasonable:

1. A boxplot or an analysis of the given summary statistics can be used to show that the distribution is skewed.

OR

2. The range of the data is 21.8, which is only 2.91 standard deviations, which is smaller than would be expected for a normal distribution.

OR

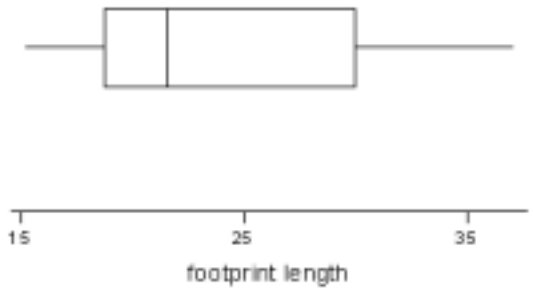
3. The minimum value is only 1.28 standard deviations below the mean, which is smaller than would be expected for a normal distribution.

OR

4. The maximum is only 1.63 standard deviations above the mean, which is smaller than would be expected for a normal distribution.

Normality <u>is</u> reasonable:

A boxplot shows that the distribution is not too skewed.



Question 2 continued

4 Complete Response

States both assumptions correctly in part (a). Justifies that the random sample assumption is not met and justifies whether or not the normality assumption is met.

3 Substantial Response

States both assumptions correctly and provides a correct discussion of only one of the assumptions.

2 Developing Response

States only one of the assumptions correctly in part (a) and provides a correct discussion regarding whether or not it is met in part (b).

1 Minimal Response

States one or both of the assumptions correctly in part (a) but does not provide an adequate discussion of either assumption in part (b)

0 No credit

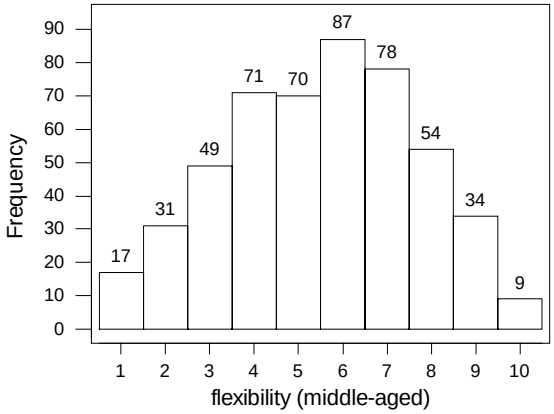
NOTES:

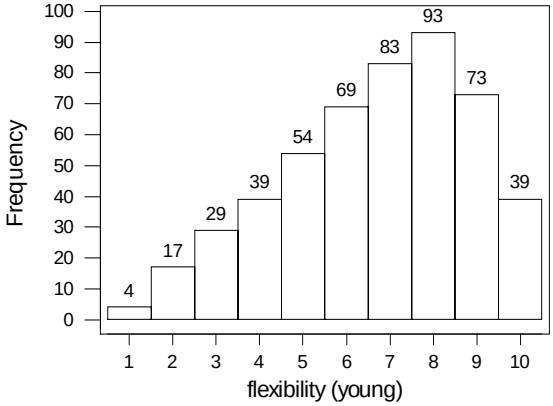
- Stating only "random sample" or "SRS" is insufficient.
- Assumption 2 is normality, not "no outliers." Stating only that "there are no outliers" is insufficient for establishing normality.
- Extraneous comments in either parts (a) or (b) should be ignored, as long as they are not contradictory to the given answer.
- If part (b) is addressed in the student's answer to (a) and not contradicted in what is written in (b), credit can be given for (b).
- If part (a) is addressed in the student's answer to (b) and not contradicted in what is written in (a), credit can be given for (a).
- A student can only be penalized once for failing to correctly identify the target population of interest: adults who use the cave.

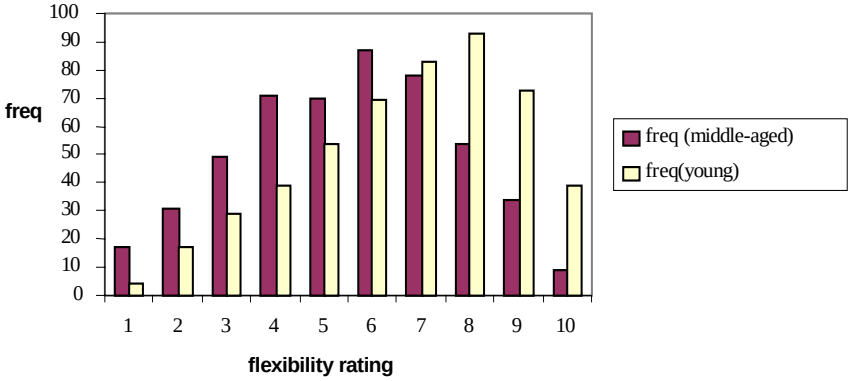
Question 3

Solution

a. Two histograms, drawn to the same scale or on the same axes. Can use either frequency or relative frequency since the sample sizes are equal (500).

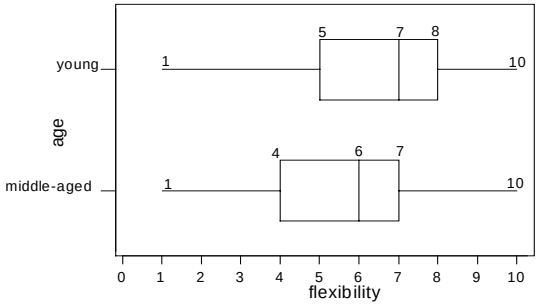




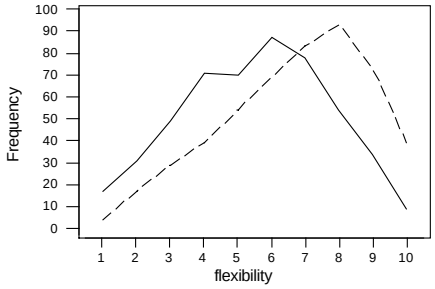


Other possibilities include boxplots or frequency polygons, if done correctly. For example,

Boxplots:



Frequency Polygons:

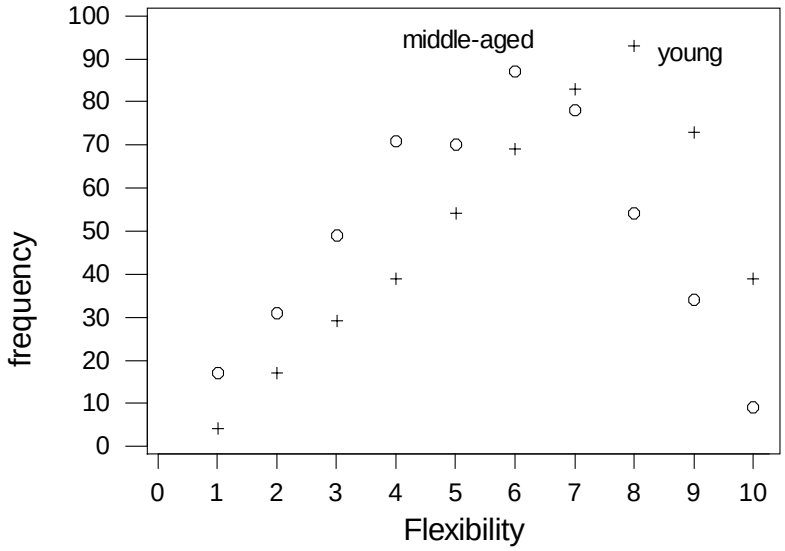


--- young
__ middle-aged

Papers that use cumulative frequency plots (ogives) may also be acceptable. These papers should be referred to a table leader.

Ideally graphs include scales, labels, title and legend.

If a student supplies the following graph, you should grade this problem holistically.



Judge from the answer to (b) whether the student is interpreting this as a frequency polygon or a scatterplot.

- With correct interpretation as a frequency polygon, the paper could be a 4.
- With weak or unclear interpretation as a frequency polygon, could be a 3.
- With mixed interpretations (e.g. association and shape), could be a 2.
- With interpretation as scatterplot (e.g. positive association/correlation), could be a 1.

b. The distribution of flexibility rating for middle-aged men is approximately symmetric, centered around 5.5, whereas the distribution for young adult men is skewed to the left (negatively skewed) centered around 6.5, higher than the middle aged men. There is quite a bit of variability in both distributions. In general there were more young men with flexibility ratings at the high end of the scale and fewer at the low end of the scale than for middle aged men.

Note: A clear description of the relative concentration of the two distributions (e.g. more flexibility ratings for young men at high values than middle aged men but more middle aged men at lower flexibility ratings than young men) is considered equivalent to a description of shape.

Scoring

Part a is

Essentially correct if	correct graph(s) are drawn, using the same scale so that comparisons can be made easily. Either frequency or relative frequency can be used. Missing labels and legends can be recovered if the description in (b) is clear and complete. Missing scaling cannot be recovered.
Partially correct if	there are errors in the construction of the graphs or the graphs are drawn using very different scales or the graphs are incomplete (but started correctly). Example: no scaling on graph
Incorrect if	any graphical displays that treat the frequencies as data are used (scatterplot, boxplots, dot plots or stem and leaf displays of the frequencies).

Part b is

Essentially correct if	the graphical displays from part (a) are interpreted in context, with comments on the differences and similarities in at least two of center, shape and spread, and the response shows good communication of ideas. (Discussion needs to be clearly linked to graphs and comparison between the two groups must be explicit.)

Note: It is not essentially correct to say the distribution of flexibility ratings for middle-aged men "**is** normal", some qualification must be given, i.e. “approximately Normal”.

Partially correct if	interpretation is correct but not in context **or** correct comparison of the two groups is made only on the basis of one of center, shape or spread **or** correct comparison of the two groups is made on at least two of center, shape and spread but communication is weak. **or** at least two of the same individual descriptions in both groups (e.g. center and shape) but no direct comparison between the two groups.

Incorrect if — response fails to compare the two groups on any of center, shape or spread (e.g. only compares the two groups for one value of the flexibility ratings.)

If a display that uses the frequencies as data is done in part (a), no credit is given in part (a) (e.g. frequency boxplots, or frequency vs. frequency/ frequency vs. flexibility scatterplots). However, if a student attempts an interpretation, and they do a credible job with the interpretation, part (b) can be scored as partially correct (resulting in a score of 1 for the problem). For example if a student says that there is a "positive linear relationship" or that center, shape and spread are similar because stem and leaf displays of the frequencies look similar, this is a credible interpretation.

4 Complete Response

Essentially correct on both parts.

3 Substantial Response

Essentially correct on one part and partially correct on the other.

2 Developing Response

Essentially correct on one part and incorrect on the other
OR
Partially correct on both parts

1 Minimal Response

Partially correct on one part

A paper using the frequencies as data can receive at most a 1.

0 No credit

Question 4

Solution

a. **part 1**: States a correct pair of hypotheses

H_o: $\mu_W = \mu_N$ $\qquad$ H_o: $\mu_W - \mu_N = 0$

OR

H_a: $\mu_W \neq \mu_N$ $\qquad$ H_a: $\mu_W - \mu_N \neq 0$

where μ_W is the mean mental skill score for babies who used walkers and μ_N is the mean for those who did not. Nonstandard notation must be explained. Hypotheses about statistics (e.g. $\bar{x}$ or $\hat{p}$) are unacceptable.

part 2: Identifies a correct test (by name or by formula), and **checks** appropriate assumptions.

Note: Problem states that samples are random samples, so this does not need to be addressed in the assumptions.

Independent samples t test. Assumptions: large sample or normal population distributions. Check: OK, because, for example, n_1&n_2>30.

OR

Pooled t test. Assumptions: large samples or normal population distributions, equal population standard deviations. Checks: OK because, for example, n_1&n_2>30 and $s_1 \approx s_2$.

OR

Independent samples z test. Assumptions: large samples. Check: OK because, for example, n_1&n_2>30.

part 3: Correct mechanics, including value of test statistic, df (if appropriate), and P-value or rejection region (except for minor arithmetic errors)

- For independent samples t test:

$$t = \frac{\bar{x}_W - \bar{x}_N}{\sqrt{\frac{s_W^2}{n_W} + \frac{s_N^2}{n_N}}} = \frac{113 - 123}{\sqrt{\frac{12^2}{54} + \frac{15^2}{55}}} = \frac{-10}{\sqrt{6.7576}} = -3.8468$$

(Calculator: t = -3.846843677)
df = 102.828 (OK to use 102), P-value = .0002

OR conservative df = 54 - 1 = 53, P-value = 2(.00016)=.00032
OR using tables (for either df) P-value < 2(.0005) = .001

- For pooled t test: s_p = 13.597, t = -3.839, df = 107, P-value = .0002 (or < .001 from tables)

- For independent samples z test, z = -3.8468, P-value = .0001 (or <2(.0002) = .0004 from tables)

Question 4 continued

part 4: Stating a correct conclusion in the context of the problem, using the result of the statistical test (i.e., **linking the conclusion to the result of the hypothesis test**).

Reject the null hypothesis because P-value is less than stated α (or because P-value is very small, or because test statistic falls in the rejection region). There is convincing evidence that the mean mental score of babies who used walkers is different from the mean score for babies who did not use walkers.

If both an α and a P-value are given, the linkage is implied. If no α is given, the solution must be explicit about the linkage by giving a correct interpretation of the P-value or explaining how the conclusion follows from the P-value.

If the P-value in part 3 is incorrect but the conclusion is consistent with the computed P-value, part 4 can be considered as correct.

NOTE: A confidence interval approach will earn full credit for

- correct hypotheses at outset or, implicitly, in conclusion,
- correct procedure (by name or formula) and assumptions checked,
- correct mechanics, including specification of a (reasonable) confidence level, degrees of freedom specified (if appropriate)
  - 2-sample t interval, unpooled, 95%, df=102 or 53: (−15.2, −4.8)
  - 2-sample t interval, pooled, 95%, df=107: (−15.2, −4.8)
  - 2-sample z interval, 95%: (−15.1, −4.9)
- correct conclusion in context: "Since 0 is not in the 95% confidence interval, there is a significant difference between the mean mental skill scores of babies with walkers and babies without at the α=.05 level of significance."

part (b): No. This was an observational study, and a causal relationship can not be inferred from an observational study.

- It ***is*** sufficient to say any of:
  - "no; observational study" (or "no; not controlled experiment").
  - "no; no randomization in grouping" or "no; parents choose which babies use walkers".
  - "no" and then cite a plausible confounding variable **and** indicate how it is confounded with the formation of the groups.
- It is ***not*** sufficient to either:
  - merely mention lurking and/or confounding variables without indicating how they are confounded with the formation of the groups.
  - mention a causal factor which is a treatment "side effect", e.g. that walkers may contain plastics which are toxic to children.

Question 4 continued

Scoring

Part (a) is evaluated based on the four parts of the test. Each part must be COMPLETELY correct (except for minor arithmetical errors in part 3) to consider the part correct.

Part (b) is either correct or incorrect. If a student just answers "no" without giving a correct explanation that relates to the design of the study, part (b) is incorrect.

Note: *A 1-sided test can earn, at most, a score of 3.*

4 Complete Response

All four parts of the hypothesis test in part (a) correct and part (b) correct. (4-E)
(4 parts correct in (a) -- Correct in (b))

3 Substantial Response

All four parts of the hypothesis test in part (a) correct and part (b) incorrect (4-I)
OR
Three parts of the hypothesis test in part (a) correct and part (b) correct. (3-E)

2 Developing Response

Two parts of the hypothesis test in part (a) correct and part (b) correct. (2-E)
OR
Three parts of the hypothesis test in part (a) correct and part (b) incorrect. (3-I)

Note: *For papers judged a 2 because a one-tailed test is done and assumptions are not checked, exceptionally strong answers to the rest of the problem can be used to score the paper a "holistic" 3.*

1 Minimal Response

Two parts of the hypothesis test in part (a) correct and part (b) incorrect (2-I)
OR
None or one part of the test in part (a) correct and part (b) correct.
(0-E or 1-E)

0 No credit
Note that a 1-I earns a score of 0.

Question 5

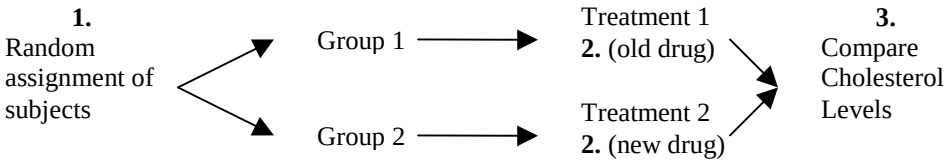
Solution

(a) Describes an experimental design that includes:

1. Random assignment of volunteers to the treatment groups
2. Identification of treatment groups as old drug and new drug
3. Indication that a comparison or measurement of cholesterol levels should be made

OR

The student may give a detailed diagram that addresses the three parts:



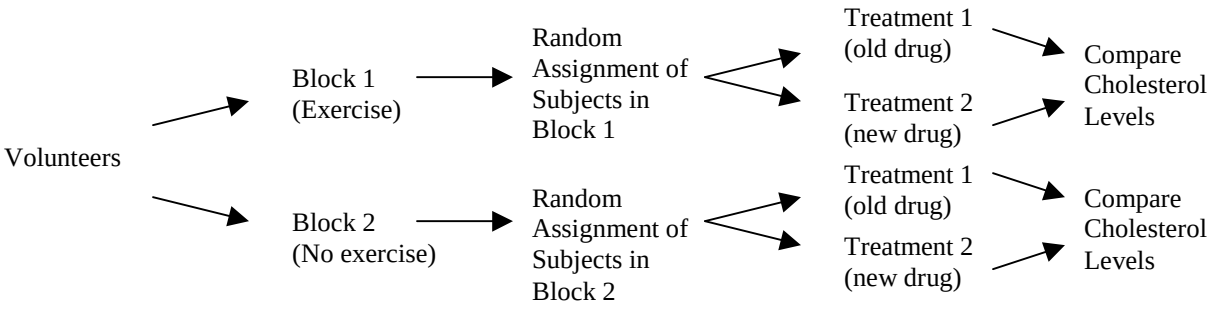
Note: In part (a), it is incorrect to use the terminology "treatment" and "placebo" for the treatment groups. It is considered correct to use "old drug" and "new drug", and "placebo," if a third group is used, for the treatment groups.

(b) Describes an experimental design that includes:

1. Creating blocks based on level of exercise or cholesterol level, or creating blocks using age, diet, gender, or any other factor plausibly related to cholesterol level **with explanation** (i.e., block on gender because males and females may respond differently)
2. Random assignment of subjects to treatments within blocks

OR

The student may give a detailed diagram that addresses the two parts as long as the blocking factor is described.



Note: No credit will be given in part (b) if a student does not use blocking in his/her design even though they randomize correctly.

Note: Crossover designs or matched-pairs designs that incorporate the idea of blocking are acceptable.

(c) Clearly explains a double blind experiment—neither the subjects nor those administering the drugs or monitoring results know which of the two drugs is being used.

An answer of yes without explanation receives no credit.

An answer of no could receive credit if the design described in part (b) does not allow for double-blinding.

Scoring:
-Parts (a) and (b) will be scored as either essentially correct (**E**), partially correct (**P**), or incorrect (**I**).
-Part (c) will be scored as either essentially correct (**E**) or incorrect (**I**).

Part (a) is:

Essentially correct if all three of the criteria are met
Partially correct if two of the three of the criteria are met
Incorrect if one or none of the three criteria are met

Part (b) is:
Essentially correct if the two criteria are met
Partially correct if only one of the two criteria is met, given that blocking has been indicated
Incorrect if none of the two criteria are met

Part (c) is:
Essentially correct only if the writer correctly communicates he/she knows what double blind means
Incorrect otherwise

Scoring

4 Complete Response

Essentially correct on all three parts.

3 Substantial Response

Part (c) is essentially correct, and parts (a) and (b) have exactly one essential and one partial.

OR

Part (c) is incorrect, and parts (a) and (b) are both essentially correct.

2 Developing Response

Part (c) is essentially correct, and parts (a) and (b) have at least one partial or exactly one essential.

OR

Part (c) is incorrect, and parts (a) and (b) have exactly one essential and one partial.

1 Minimal Response

Only part (c) is essentially correct.

OR

Part (c) is incorrect and parts (a) and (b) are both partially correct or have exactly one essential.

0 No credit

Note: Only one partial in parts (a) or (b) and an incorrect in part (c) will be a 0.

Exception: If part (a) includes an excellent explanation of a detailed randomization, a student can get a 1 even if parts (b) and (c) are incorrect.

Question 6

Solution

a. Large sample confidence interval for a population proportion. Assumptions: large sample. Here,

$n\hat{p} = 20 \geq 5$ (or 10) $\quad n(1-\hat{p}) = 380 \geq 5$ (or 10)

or

$\hat{p} \pm 3\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ is in the interval (0,1)

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.05 \pm 1.96\sqrt{\frac{(.05)(.95)}{400}}$$

$$.05 \pm .02146$$

(.02854, .07146)

Calculator solution: (.02864, .07136), but still need to name the interval used and check assumptions.

Interpretation: Based on this sample, we can be 95% confident that the proportion of married couples for which the wife is taller than her husband is between .028 and .071.

Question 6 continued

Part (a) is

Essentially correct if	(1) identifies the correct procedure either by name or by formula and checks to make sure sample size is large enough (2) has correct computations (3) gives a correct interpretation in context.
Partially correct if	correctly does two of the three things required for an essentially correct response.

Notes:
1. In checking assumptions, p, p-hat, and pi are all acceptable symbols for the sample proportion.
2. Stating the assumptions is NOT the same as checking the assumptions.
3. A common incorrect response refers to the proportion of times, in repeated sampling, that a future sample proportion would be contained in "this" interval. This should be read as an incorrect interpretation.

b. Let M be the height of a randomly selected married man and let W be the height of a randomly selected married woman. Then M - W has a distribution that is approximately normal with

$$\mu_{M-W} = 70 - 65 = 5$$

$$\sigma_{M-W} = \sqrt{\sigma_M^2 + \sigma_W^2} = \sqrt{3^2 + 2.5^2} = \sqrt{15.25} = 3.9051$$

Then P(M - W < 0) = P(Z < (0-5)/3.9051) = P(Z < -1.28) = .100

Part (b) is

Essentially correct if	calculates the mean and standard deviation of M-W (or W-M) correctly, and then correctly calculates the appropriate probability.
Partially correct if	calculates the mean and/or standard deviation incorrectly, but then uses these values and a correct process to compute an appropriate probability **OR** computes the mean and standard deviation correctly but is unable to compute the appropriate probability.

c. Based on the answer to part (b), if heights of husbands and heights of wives were independent, would expect approximately 10% of married couples to have the wife taller than the husband. Based on the interval in part (a), the estimate of the percent of married couples with the wife taller than the husband was between 3% and 7%. This is smaller than what we would have expected to see if the heights of husbands and wives were independent. So, the data suggests that heights of husbands and wives are not independent.

Part (c) is

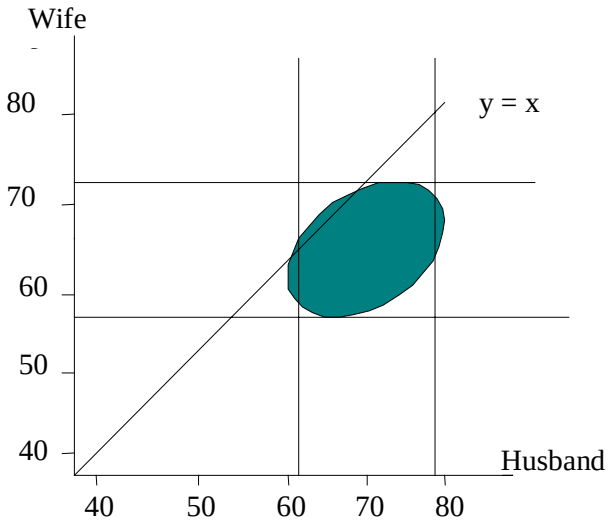
Essentially correct if	correctly judges independence (dependence) based on responses in parts (a) and (b) and gives a good explanation relating the probability in part (b) with the **interval** in part (a).
Partially correct if	judgment of independence is consistent with the responses in parts (a) and (b), but explanation is weak or poorly linked to parts (a) and (b).

Notes:

1. If the explanation compares the **point** estimate in (a) with the probability in (b), this is considered a weak argument.

2. If explanation is missing or shows no understanding of independence, then it should be regarded as incorrect.

d.



Key characteristics of a correct graph:

1. Center of ellipse is at about (70, 65)
2. Size of ellipse: approximately 3 standard deviations from center,
   i.e. +/- 7.5 from center on Wife axis and +/- 9 from center on Husband axis.
3. Orientation of ellipse--positive slope
4. Small overlap of the y = x line
5. Label the axes

Part (d) is

Essentially correct if	the ellipse drawn meets all 5 of the stated characteristics.
Partially correct if	the ellipse drawn meets 3 or 4 of the stated characteristics.

Scoring

PARTIALLY CORRECT RESPONSES COUNT AS ½ AN ESSENTIALLY CORRECT RESPONSE. THAT IS, TWO PARTIALLY CORRECT RESPONSES CAN COUNT AS ONE ESSENTIALLY CORRECT RESPONSE.

4 Complete Response

Essentially correct on four parts.

3 Substantial Response

Essentially correct on three parts.

2 Developing Response

Essentially correct on two parts.

1 Minimal Response

Essentially correct on one part.

0 No credit

IF A PAPER IS BETWEEN TWO SCORES (FOR EXAMPLE, 2 PARTS ESSENTIALLY CORRECT AND ONE PART PARTIALLY CORRECT, WHICH IS BETWEEN A 2 AND A 3) USE A HOLISTIC APPROACH TO DETERMINE WHETHER TO SCORE UP OR DOWN.