

CEP 932 Quantitative Methods in Educational Research I
Summer 2010
Homework #2

43 points

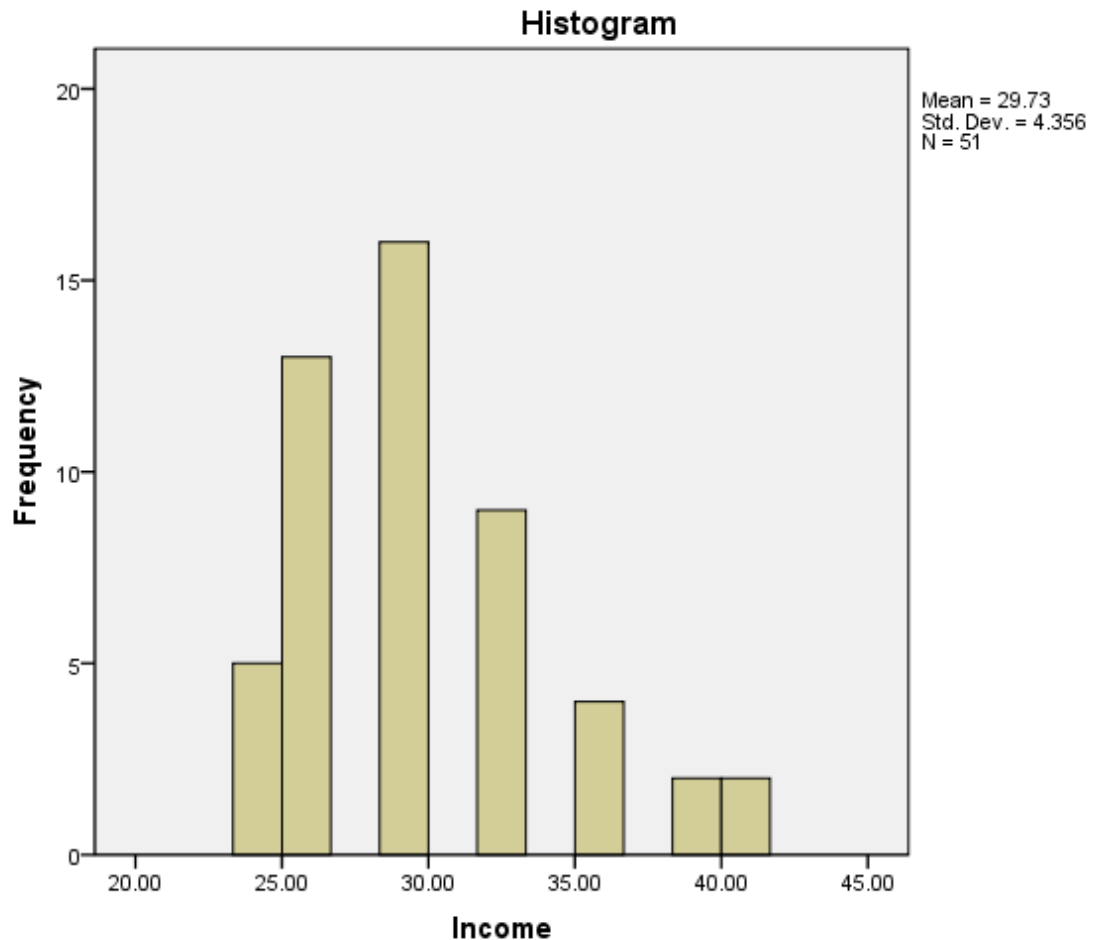
Due: July 5, 2010 (10:00 AM)

1. (3 points) What is a parameter? What is a sample statistic? Describe the differences and relationships between them.
 - a. The basic difference is that the parameter is the numerical descriptive measures for a population, while a sample statistic is the numerical descriptive measures for a sample. A parameter is usually unknown because populations are so large, while samples are defined because the data sets allows for arriving at a known measure.

2. (4 points) As the mobility of the population in the US has increased and with the increase in home-based employment, there is an inclination to assume that the personal income in the US would become fairly uniform across the country. The following table provides the per capita personal income for each of the 50 states and the District of Columbia.

Income (Thousands of dollars)	Number of States
23.5	5
26.5	13
29.0	16
32.0	9
35.5	4
38.5	2
41.5	2

(2 points) Construct a histogram for the income data using SPSS.



- a. (1 point) Describe the shape of the histogram. Is it symmetric? No, this histogram is not symmetric.
 - a. It has a cluster of values around the mean (around the central tendency) and a few outliers at the higher end.

- b. (1 point) Would you describe per capita income as being fairly homogeneous across the U.S.? Why or why not?
 - a. No, I would not describe per capita income as being fairly homogenous. There is a lot of difference within the range of values, with the mean being pulled up by the outliers at the high end of the income scale.

3. (8 points)

(5 points) Given the following data, calculate the following sample statistics: mean, median, mode, range, standard deviation, IQR using SPSS.

Statistics

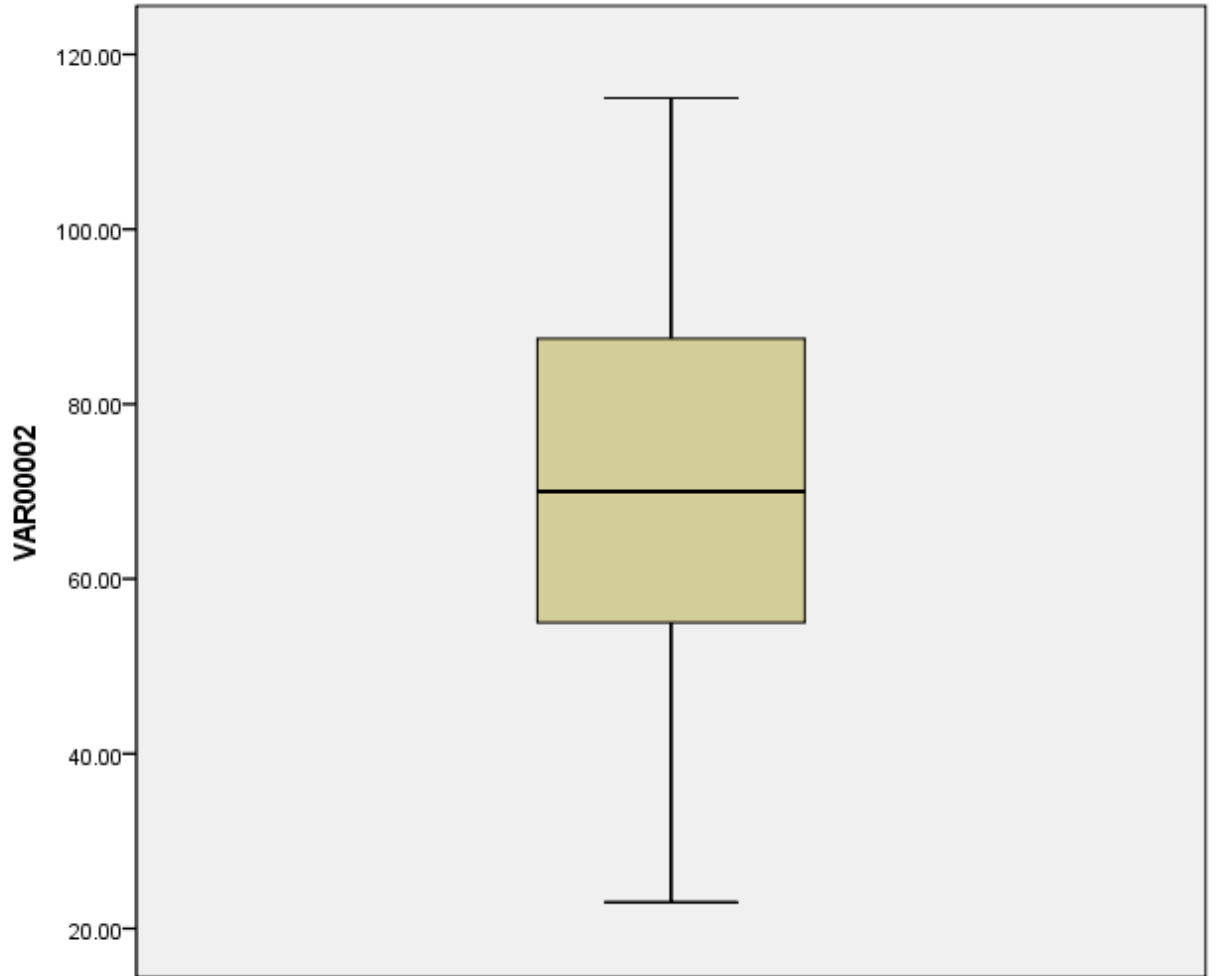
VAR00002

N	Valid	20
	Missing	31
Mean		70.5000
Median		70.0000
Mode		90.00
Std. Deviation		23.69210
Percentiles	25	55.0000
	50	70.0000
	75	88.7500

$$\text{IQR} = 88.75 - 55.00 = 33.75$$

55	85	90	50	110
115	75	85	82	23
70	65	50	60	90
90	55	70	59	31

(3 points) Create a box plot for the data using SPSS:



4. (9 points) Data are collected on the weekly expenditures of a sample of urban households on food (including restaurant expenditures). The data, obtained from diaries kept by each household are grouped by the number of members of the household. The expenditures are as follows:

1 member:	67	62	168	128	131	118	80
	53	99	68	76	55	84	77
	70	140	84	65	67	183	
2 members:	129	116	122	70	141	102	120
	75	114	81	106	95	94	98
	85	81	67	69	119	105	94
	94	92					
3 members:	79	99	171	145	86	100	116
	125	82	142	82	94	85	191
	100	116					
4 members:	139	251	93	155	158	114	108
	111	106	99	132	62	129	91

5+ members:	121	128	129	140	206	111	104
	109	135	136				

- a. (2.5 points) Calculate the mean expenditure separately for each group of members. Use excel or SPSS.
- 1 member =93.75
 - 2 members=98.65
 - 3 members=113.31
 - 4 member=118.42
 - 5+ members=131.9
- b. (2.5 points) Calculate the median expenditure separately for each group of members. Use excel or SPSS.
- 1 member=78.5
 - 2 members=95
 - 3 members=100
 - 4 members=112.5
 - 5 + members=128.5
- c. (1 point) Calculate the mean of all the data (for all members). Use excel or SPSS.
- The mean for all the data is 107.64.
- d. (1 point) Can the combined mean be calculated from the means of the groups of members? If so, how? Notice that the number of values differs by group of members.
- With samples of unequal sizes, the mean of the means is a weighted of average of the means using sample sizes. So we have five samples here, the mean can be calculated by multiplying the size of sample of 1 member households (20) by the mean (93.75) and repeating that for all the means. Add those together and divide by 5, and you have the mean of the means.
- e. (1 point) Calculate the median of all the data (for all members). Use excel or SPSS.
- The median for all members is 104.
- f. (1 point) Can the combined median be calculated from the means of the groups of members? If so, how?
- I am pretty sure there is not a way to move from the means of the groups to the combined median. Since mean and median are computed in different ways, that just seems strange to me. In the event that this question is really asking whether the combined median can be determined by the medians of the subsets, than the answer is 100% no.

5. (3 points) Use the following data:

69, 80, 77, 40, 59, 38, 99, 19, 27, 63, 70, 37, 62

- a. (2 points) What is the mean and median of the data? Use excel or SPSS.
 - i. Mean= 56.92
 - ii. Median=62.00
- b. (1 point) Comparing the mean and median, would you characterize the distribution as symmetric, negatively or positively skewed?
 - i. Because the mean is lower than the median, the distribution would be characterized as negatively skewed. However, it is not skewed that much and would be considered within the range of normal.

6. (6 points)

The table below represents Math and English scores obtained by 8 students in a 4th grade class (consider them a sample from a population). Use the table below to answer the questions.

Student Number	Math Score (x)	English Score (y)
1	22	33
2	26	35
3	27	35
4	25	30
5	21	29
6	26	34
7	23	31
8	24	29

1. What is the math mean score: Use excel or SPSS.
 - a. *Math mean score is 24.25*
2. What is the English mean score: Use excel or SPSS.
 - a. *The English mean score is 32*
3. What is the variance in math scores? Use excel or SPSS.
 - a. *The variance in the math scores is 4.5*
4. What is the variance in English scores? Use excel or SPSS.
 - a. *The variance in the English scores is 6.57*
5. Show that the following equality is true $\sum (x_i + y_i) = \sum x_i + \sum y_i$
 $(22 + 33) + (26 + 35) + (27 + 35) + (25 + 30) + (21 + 29) + (26 + 34) + (23 + 31) + (24 + 29) = 450$
And

$(22 + 26 + 27 + 25 + 21 + 26 + 23 + 24) + (33 + 35 + 35 + 30 + 29 + 34 + 31 + 29) = 450$
Therefore, the equality is true. Also, based on the associative property, it is true on its face.

Show that the following equality is true $\sum (x_i + y_i)^2 \neq \sum x_i^2 + \sum y_i^2$

I'm saving time, so I'll use the (x +y) values from above

$$55^2 + 61^2 + 62^2 + 55^2 + 50^2 + 60^2 + 54^2 + 53^2 = 25,440$$

$$194^2 + 256^2 = 37636 + 65536 = 103,172$$

Clearly, the two sides of the equation are not equal. This is also a basic rule in algebra as well when dealing exponents. This looks like when exponents are commutative, but it isn't when they are being added in this way.

7. (6 points)

The Tennessee class size experiment has been called one of the great experiments in the history of education. It involved randomly assigning both students and teachers to one of three experimental conditions (a small class of 13-17, a regular sized class of size 23-27, or a regular sized class with a full time instructional aide) within each participating school. Thus each school had classrooms corresponding to (at least) three conditions. The experiment started with a cohort of Kindergarten students and assignments were maintained until the end of third grade. When new students entered the participating schools, they were randomly assigned to one of the three conditions. At the end of the experiment the achievement test scores were compared to evaluate the effects of small classes. Long term follow up comparisons of the three groups have continued for nine years (until the participants graduated from high school).

1. (3 points) What are likely to be the most serious threats to the validity of causal inferences in this experiment when the outcome is measured at the end of third grade? Discuss each one of them?
 - a. The most serious threats to validity difference in an outcome and a control group
 - i. The most serious threats to validity is if the groups are not matched in terms of factors other than independent variable (for instance, if the smaller class size had a larger than chance group of high SES students, one could not prove causality was from the class size.) Additionally, I would worry about bias since it is obvious who was in each group. The switching among control and treatment groups would be a big issue if it occurred within this time period: there would be no way to show causality if students were moving among the different groups. Also, I wonder how student development might impact this. Finally, I would worry about the Hawthorne effect: that the treatment effect might be contaminated, as it were, by the fact that the teachers knew they were being observed.
2. (3 points) What are likely to be the most serious threats to the validity of this experiment when the outcome is measured at the end of high school? Explain each of them and why you think they may be more serious in long term follow up studies.
 - a. Cohort of students may be affected by student mobility (Differential attrition). I would say this would be the most serious threat to validity. Additionally, with so many additional interventions compounding over the course of a student's schooling (were they tutored more than another student in 7th grade? What about

9th grade?), I am not sure it would be true causality if a student were shown to have increased achievement.

8. (4 points)

Many non-experimental studies have compared the academic achievement of Catholic school students and public school students (e.g., the 1987 book by Coleman, Hoffer, and Kilgore). Because there is no random assignment, research designs must take into account that Catholic school students may be different in ways other than the treatment (they or the school they attend).

1. How would you design a study to estimate the causal effect of going to Catholic (as opposed to public) schools? Specify which variables you would use in the study and give a rationale for including them.
 - a. In this case, I would attempt to design my samples to control by matching: I would try to match students and teachers by race (because race has been known to have correlations to achievement), SES (because we know there is a correlation between increased income and increased achievement), mother's education (I read somewhere that mother's educational achievement has a high correlation to a child's achievement) or probably previous achievement (because we know that kids who have done well before will do well in the future). I might exclude students with learning disabilities. In order to best control by matching, I would focus on a geographical location (cluster sampling) where the Catholic schools matched well the local district schools in terms of student characteristics which correspond to the variables I've identified. If this were not possible, I would attempt to control by statistical adjustment based on the variables I've identified.