The Super Six

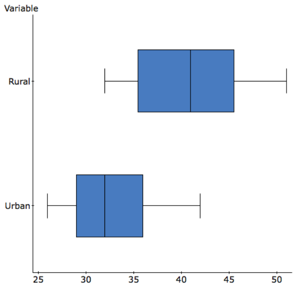
2005 #1: Urban and Rural Calories

a) *The Urban center is lower, with fewer calories typically than Rural. The spreads of the distributions are about the same. The Urban distribution is skewed right, the Rural distribution is roughly uniform.*

b) *No. A random sample of US schools was not selected. Only 1 school from each region.*

c) *Plan II would be better. One given day might have more or fewer calories than normal. A 7-day average would average out the day to day variability and more accurately estimate the true average.*

d) Construct parallel boxplots to compare the calorie distribution of the rural vs. the urban students.



Labeling the x-axis with "Calories" would also be great!

e) Verify whether or not there are outliers in either data set.
Rural:
IQR: 45.5 – 35.5 = 10
lower fence = 35.5 – 15 = 20.5
upper fence = 45.5 + 15 = 60.5
Urban:
IQR: 36 – 29 = 7
lower fence = 29 – 10.5 = 18.5
upper fence = 36 + 10.5 = 46.5
No outliers in either data set because all data is within the fences.

f) Describe how a researcher might use schools as clusters to gather data in a given county.
A researcher could take a list of all the schools in the county of interest. Randomly select some of the schools. Then survey all the students in those schools.

g) One researcher observed that rural students ate more home cooked meals than urban students. A journalist wrote an article stating that home cooked meals caused an increase in calorie intake. Describe a confounding variable that may be the cause of the higher calorie intake in rural students.
It is possible that rural students eat more calories because they are more active and thus eat larger portions. Thus we would be unable to determine in the increase in calories in calories for rural students was due to the home cooked meals or if it was due to the larger portions.

h) A researcher notes that rural students have less access to organic/vegan supermarkets than urban students and that this could explain the increase in

calories in rural areas. However, another researcher noted that rural students eat more meals at home (that are meat and carb heavy, thus high in calories) than urban students. Explain how these variables (lack organic stores and home-cooked meals) are confounded.
Rural students who do not have access to organic food will also eat less healthy meals at home more often. Eating less organic food and eating home more often will both increase calorie consumption. The researchers cannot tell how much an increase in calories is due to eating less organic food or eating more meals at home.

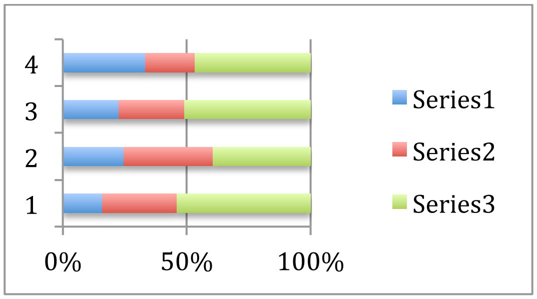
i) Describe how you would use your calculator and a list of 9th grade students from your school to conduct a simple random sample. Include a description of how you would implement your procedure.
Take a list of all 9th graders. Number the list, 1 to n. Randomly generate a random number and survey that student. Repeat the process, skipping repeats, until you have the desired sample size.

j) Describe one variable that might be important to create strata and why you chose that variable.
Because males and females tend to consume different amounts of calories, it would be helpful to stratify by gender. This would reduce variability created by gender differences in calorie consumption (ensuring that the correct proportion of males and females are in the sample.)

k) What inference procedure would you use to compare the two groups?
2 sample t test. Ho: mu(rural) = mu(urban)

2003B #2 Income & Age

a) *89/207 = 43%*
b) *35/96 = 36.5%*
c) *They are not independent. Because 43% the sample is 31-45, but only 36.5% of this age group makes over $50,000, this shows they're not independent.*
d) Make a graphical display to examine the relationship between Age and Income. Describe this graph.



Center, shape, and spread? NO!
Correct answer: We observe that the 31-45 group has the lowest percentage in the highest income category. The 21-30 year olds have the lowest percentage in the lowest income category. As you get the older, the percent in the lowest

income increases. (You could say a lot more. On the AP-test I recommend you make 3 observations.)

e) Name an inference procedure that could be carried out to answer the independence question on part (c).
Chi-square test for independence. Ho: Age and income are independent.

f) If Age and Income were completely independent, find the number of 46-60 year olds you would expect to have an income of over $50,000.
*(53*96)/207 = 24.6*

1999 #1 Aircraft in the 90's

a) *Yes, because the residual plot has no pattern.*

b) *233.5: There are approximately 233.5 more aircraft per year.*

c) *2939.9: In 1990, the model predicts 2939.9 aircraft.*

d) *2939.9 + 233.5(2) = 3406.9 aircraft predicted*

e) *40 = y – 3406.9, so y = 3446.9, so 3447 actual aircraft.*

f) Interpret *s* in the context of this problem.
The regression line misses the data by an average of 33.43 aircraft.

g) Create and interpret a 95% confidence interval for the slope.
233.5 ± 2.365(4.316) = (223.3, 243.7)
I am 95% that the true slope of the aircraft/year is between 233.3 and 243.7.

h) R^2= 87.4%. Interpret this value in context.
87.4% of the variation in aircraft has been successfully explained by regression on years.

i) Find and interpret the correlation coefficient.
$\sqrt{0.874}$ = .934 : There is a strong, positive, linear relationship between the # of aircraft and years.

j) If each new aircraft costs the FAA an additional $1000 in regulatory costs, how much are the costs increasing each year, on average?
*$1000*233.5 = $233,500 more per year.*

k) Is there statistically convincing evidence that the number of aircraft is related to year? Explain.
The p-value for the slope is zero which is less than any reasonable alpha. So we reject Ho (B = 0) and conclude there is a significant relationship between aircraft and year.

2005 #2 Telephone Lines

a) *0*0.35 + … = 1.6 telephone lines*

b) *I would expect the new average to be closer to 1.6 than the smaller sample. As the sample size increases the statistic will get closer to the parameter.*

c) *The median is 1. $P(X \leq 1) = 55\%$ and $P(X \geq 1) = 65\%$*

d) *The mean is greater than the median. This is expected, as the distribution is skewed right.*

e) What is the probability that 3 or more lines are in use at noon?

_15% + 10% + 5% = 30%_

f) What is the probability that at least 1 line is in use at noon?
_1 – 35% = 65%_

g) Given that 3 or more lines are in use at noon, what is the probability that all 5 are in use?
_5/30 = 16.7%_

h) Assuming that each day is independent of the next, what is the probability that on exactly 2 out of the next 5 days there are no lines in use at noon?
Binomial, n = 5, p = 35%. P(x = 2) = 33.6% (binomialpdf)

i) Suppose you come by every day at noon to see how many lines are in use. What are the chances that you don't find all 5 in use until your 7th visit?
_(0.95)^6*(0.05) = 3.675%_

j) Find the standard deviation of the number of lines in use this support center expects to have at noon.
√(0 – 1.6)^2*(0.35) + … = 1.56 telephone lines [1-var stats-probabilities in freq]

k) Each call lasts an average of 3.75 minutes. What is the mean and standard deviation of the number of minutes at noon?
1.6*3.75 = 6 minutes, 1.56*3.75 = 5.85 minutes

l) What are the mean and the standard deviation of the total number of lines in use this support center expects to have at noon over a 7-day week?
1.6*7 = 11.2 lines. √(1.56^2 + 1.56^2 + …) (7 times) = 4.13 lines

m) If another support center has a mean of 2.1 calls and a standard deviation of 1.8 calls, what is the mean and standard deviation of the total of number of calls of both centers at noon?
1.6 + 2.1 = 3.7 lines. √(1.56^2 + 1.8^2) = 2.38 lines.

2004B #3 Bauxite Ore Cars

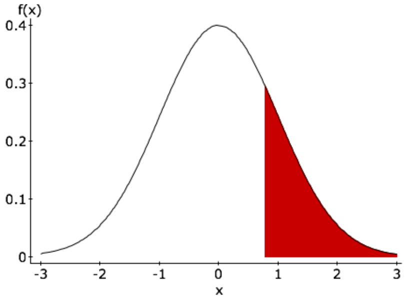
a) _z = 0.7778; normalcdf = 21.8%_

b) _No. Because 70.7 could happen by chance about 22% of the time._

c) _Now sd = 0.9/√10 = 0.285; So z = 2.46; normalcdf = 0.0069_

d) YES! A result like this would happen less than 1% of the time by chance.

e) Draw a careful sketch to show your answer to part (a)



Normal Distribution
Mean:0 Std. Dev.:1
P(X≥0.778) = 0.21828451

f) Given the initial mean and standard deviation, how full are the most heavy 10% of the cars?
71.152 tons: invnormal(0.90); [note z = 1.28]

g) Describe the sampling distribution of the sample mean, if a sample of size 10 were chosen.
mean = 70 tons; sd = 0.285 tons; shape = approx. normal

h) If we took a random sample of 40 cars instead of 10, how would that change your answer to part (g)?
The new sd = $0.9/\sqrt{40}$ = 0.142 [fun fact: exactly half as big as (g). See why!?!]

2000 #5 Cholesterol and Exercise

a) I would randomly sort the volunteers into 2 groups. 1 group would take the new drug and the other the current drug. Compare cholesterol levels at the end.

b) *Since exercise effects cholesterol level, I would block by the volunteers' exercise level. Divide the volunteers into high, medium, low exercise level. Randomly place half from each block in the treatment groups.*

c) *Yes. An assistant can setup the medications so neither the evaluators nor the subjects know which treatment they are receiving.*

d) Describe a method for implementing your design in part (a).
Put all the volunteers' names in a hat. Stir. Randomly select half the names and those subjects receive the new drug. The rest receive the current drug.

e) What inference procedure would you use to compare the results obtained by method (a)?
2 mean t test.

f) After the method in part (a) was carried out, researchers found a difference with a p-value of 0.003 (in favor of the new medication over the old drug). Does this mean that the researchers can conclude that drug caused a reduction in cholesterol?
Yes. They can conclude that for people like the volunteers the new drug is better.

g) Identify the subjects, the treatment(s), the factor(s), the level(s), and the response variable in this experiment.
The subjects are the volunteers.
The factor is the drug.
Two levels: current and new.
Two treatments: new and current drug.
Response variable: cholesterol level

h) After an increasing in funding, the company is able to run an experiment where they test both the drug treatment and the effects of exercise. They decide to ask volunteers to exercise at a high level, a low level, or not all. Describe the factors, their levels, and the treatments of this expanded experiment.
One factor is the drug, the other is exercise.
Two levels for drug: current and new. Three levels for exercise: high, low, none.

The Super Six

Six treatments: current & high, current & low, current & none, new & high, new & low, new & none.