Identifying and Addressing Student Difficulties in an Introductory Statistics Course

Bruce Dunham and Eugenia Yu

Department of Statistics

Seminar: 27th October, 2009

Introduction

Statistics Department work with CWSEI includes

1. Post–course concept retention study
2. Student attitude survey
3. Changes in pedagogy
4. Analysis of responses to test questions
5. Experiments involving lab activities

Various faculty involved, plus Rebekah Mohr and Ava Zhang. Focus is on STAT 200.

Some background

Students find Statistics difficult ... see for instance Chance *et al.* (2004), delMas *et al.* (2007), Garfield and Ben–Zvi (2007).

The concepts ARE difficult.

The language of the subject doesn't help (see for instance Rumsey 2009).

Post–course retention study

- Call for volunteers several months after completion of STAT 200

- 29 students interviewed, from several different cohorts

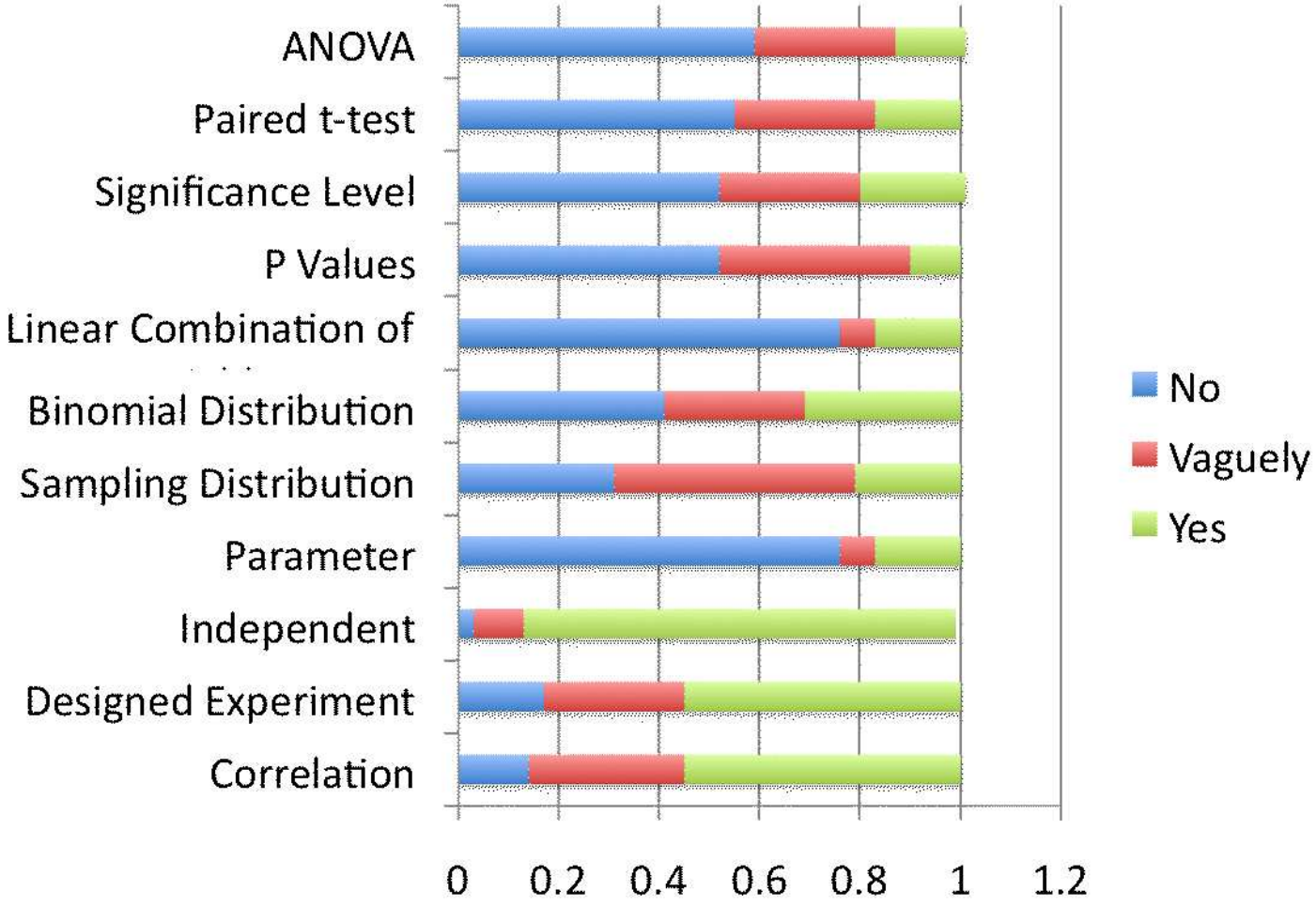
- Some selection criteria . . . and “better” students volunteered.

- Students asked about study habits, such as how often they attended lectures.

- Then asked to talk about certain concepts they had met in course – free response, with prompting if needed.

- Understanding level was classified as “yes”, “vaguely” or “no”.

Summary of findings



Parameter?

Only 17% could explain this term in the context of the course. Many were "contaminated" by use of the word elsewhere.

- "the limits that your experiment is set in.... like the extremes"

- "I use parameters in a lab and we set parameters for different things – we use it as a range and we're looking in this area."

- "a variable in an equation"

- "I always get it confused with Computer Science. In Computer Science it's programming. We have explicit or implicit parameters."

Sampling distribution?

About 21% could explain this. Many thought it was the way you take the sample or how the data in a sample vary.

- "Take small samples from a population rather than doing a survey on the whole population."
- ". . . different kinds of sampling methods . . . like stratified and random sampling."
- "The way in which you'd choose the subjects . . ."

ANOVA?

Only 14% recalled more than there was a table with some numbers. Some recalled procedural aspects, but didn't know the purpose of the procedure.

On pre-test in follow–on course, only 10% correctly identified ANOVA assumptions.

- "I remember it's a table and you fill in the values and there's formulas"

- "There were tables we had to fill out and every value in the table you calculated from the other numbers. I had no idea what it meant but I just memorized it and did it. "

- "it was at the end of the course, so I don't remember it"

Analysis of examination questions

(Fall 07, Q8) Common questions in the study of wildlife behaviour involve contests between "residents" of a particular area and "intruders". In each contest, the residents either win or lose the encounter. What is a possible parameter of interest? (Choose all that apply.)

(a) the probability that the residents win

(b) whether the residents win or lose

(c) the amount of funding the researchers receive for their study

(d) the number of researchers involved in the study

(e) the expected duration time of a contest

Responses:

- 34% gave correct response (a) & (e).
- 14% gave neither (a) nor (e).

(Winter 08, Q.14) A researcher for Agri-Giant Corporation wants to study how many pounds of tomatoes are typically picked by a manual labourer in an 8 hour day. The Corporation will consider using picking machines if the labourers cannot pick at least 240 pounds, on average. The parameter of interest to the researcher is

(a) whether or not a manual labourer can pick at least 240 pounds of tomatoes in an 8 hour shift

(b) 240 pounds

(c) how long the labourer works

(d) how many pounds of tomatoes each labourer picks in an 8 hour shift

(e) the average number of pounds of tomatoes typically picked by labourers in an 8 hour shift

Response: 77% correctly responded (e). 12% gave (a).

(Fall 08, Q.14) *(Description of experiment omitted.)* The memory test had a total of 25 items on it. The average number of items recalled was 15 for the caffeine-free group, 17 for the mixture group, and 16 for the regular cola group. Are the values 15, 16, and 17 statistics or parameters?

(a) Parameters

(b) Statistics

Responses: 95% correctly responded (b).

(Fall 07, Q.18) Recall that in the analysis of variance (ANOVA), the data are in groups. In performing the usual hypothesis test, which of the following is/are assumed? (Choose all that apply):

(a) The within–group sample means are all equal.

(b) The within–group sample standard deviations are all equal.

(c) The observations within a group are from a Normal distribution.

(d) The observations within each group are independent.

(e) Observations in different groups are independent.

Responses:

- 19% correctly gave (c), (d) and (e).
- 41% gave two of (c), (d) and (e).
- 7% gave none of (c), (d) or (e).

(Winter 09, Q.17) *(Full description omitted.)* In order to compare the mean number of hours spent studying between groups of students obtaining different letter grades in the course, which of the following tests should be used?

(a) The 2-sample z-test

(b) The 2-sample t-test

(c) The paired t-test

(d) The ANOVA F-test

(e) The Chi-square test of independence

Response: 74% correctly selected (d).

In similar question on Summer 09, 77% responded correctly.

(Fall 07, Q.15) In a hypothesis testing procedure, suppose the null hypothesis is rejected if a test statistic, W, is too large. Suppose the test with significance level 0.05 rejects the null hypothesis if $W > 4.2$. Which of the statements below is/are true? (Choose all that apply.)

(a) With this level 0.05 procedure, the probability of a Type I error is 0.05.

(b) With this level 0.05 procedure, if the null hypothesis is true, the probability of rejecting the null hypothesis is 0.05.

(c) A level 0.10 hypothesis test rejects the null hypothesis if $W > c$, where c is some number larger than 4.2.

(d) A level 0.10 hypothesis test rejects the null hypothesis if $W > c$, where c is some number smaller than 4.2.

(e) If the p-value is 0.07, then W must be bigger than 4.2.

Responses:

- 20% correctly gave (a), (b) and (d).
- 22% gave none of (a), (b) or (d).

(Winter 09, Q.14) The researcher studying turtle eggs ran a one-sample hypothesis test to determine if the eggs in regions with high levels of acid rain are thinner than the generally accepted thickness. The test was one sided, with the alternative hypothesis that these eggs have thinner shells. The test led to a p-value of 0.000839. An interpretation of the p-value in the context of this problem is:

(a) Given the Null Hypothesis is false, the probability of observing data as extreme or more extreme has probability 0.000839.

(b) The probability that the eggs from this region have mean shell thickness equal to the accepted thickness is 0.000839.

(c) If in fact the mean thickness of the egg shells from these regions is the accepted one, the probability of observing a sample mean as low or lower than the one observed is 0.000839.

(d) If in fact the mean thickness of the egg shells from these regions is not the accepted one, the probability of observing a sample mean as low or lower than the one observed is 0.000839.

(e) Both (a) and (d).

Responses:

- 59% correctly gave (c).
- 21% responded (e).

(Winter 08, Q.24) Three different labs tested two types of cream, A and B, recording the percentage of solubility in some liquid. Each lab repeated each experiment, and the data are given below:

	Cream type		
		A	B
Lab	1	6.8, 6.6	5.3, 6.1
	2	7.5, 7.4	7.2, 6.5
	3	7.8, 9.1	8.8, 9.1

Differences in the measurements may be due to differences in solubility in the cream types, differences between the labs or both of these possible sources of variation. To investigate this, you could use

(a) A Binomial distribution.

(b) A matched pairs t test.

(c) A Chi-squared test for association.

(d) A linear regression model.

(e) No method that has been encountered in STAT 200.

Responses:

- 20% correctly gave (e).
- 43% responded (b).

Lab experimentation on Teaching Sampling Distribution and the Central Limit Theorem

Outline

- Rationale behind the CLT experiment
- Description of the CLT lab
- Past exam questions and results
- Discussion and summary

Rationale behind the CLT experiment

- Students have difficulty visualizing:
  - repeated sampling from a population
  - the construction of a sampling distribution

- Students have difficulty distinguishing between:
  - the distribution of a random sample drawn from a population
  - the distribution of a statistic whose values come from repeated samples drawn from a population

- For the past few years, the CLT was taught using simulations via Excel in labs

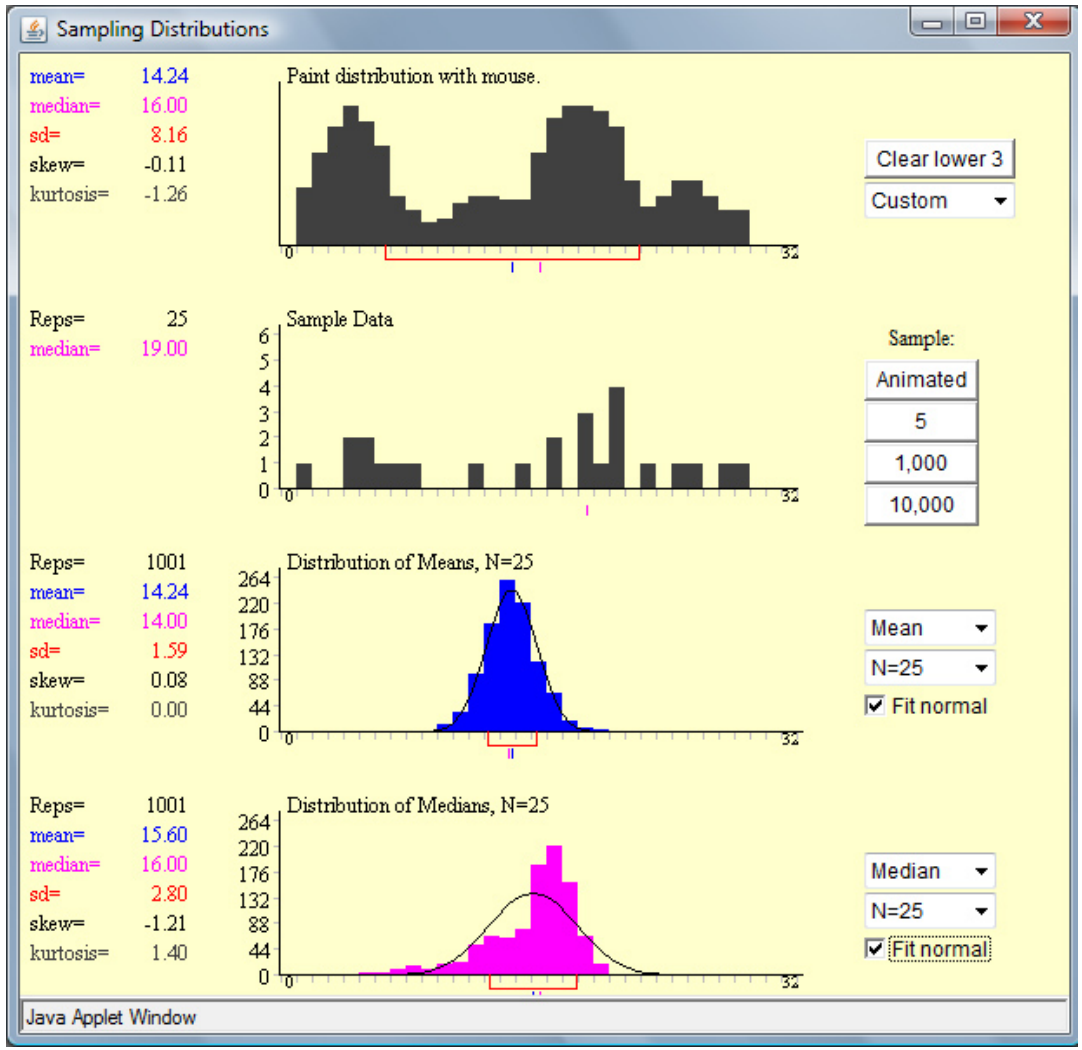
- We designed an experiment to compare the effectiveness of using simulation and applet in illustrating repeated sampling and the CLT

- We devised exam questions that identify and address misconceptions related to sampling distribution and the CLT

- Simulation:

	A	B	C	D	E	F
1	sample size = n					
2	number of samples = m					
3		sample 1	sample 2	sample 3	.....	sample m
4	obs 1					
5	obs 2		these cells contain data generated			
6	obs 3		from the normal / uniform distributions			
7	:					
8	:					
9	obs n					
10	sample mean	x-bar 1	x-bar 2	x-bar 3	.....	x-bar m
11						
12						

- Applet URL:
  http://onlinestatbook.com/stat_sim/sampling_dist/index.html



- We have had two runs of the experiment:

  1. Winter 2008 Term 2 (cohort size = 249 students)

  2. Summer 2009 (cohort size = 69 students)

- Lab sections were randomized to one of the two groups: simulation lab and applet lab

- Students were evaluated by a lab exercise, at midterm (given during the week after the CLT lab) and final exam

Description of the CLT lab

Students complete the following tasks:

- generate data from the normal and uniform distributions (the applet lab also involves generating data from a bi-modal distribution)
- construct a histogram of sample data
- construct a histogram of sample means
- compare the two histograms in the above
- matching descriptions to a set of histograms

Lab exercise

Match the following descriptions to the appropriate histogram (A-D). Justify your answers.

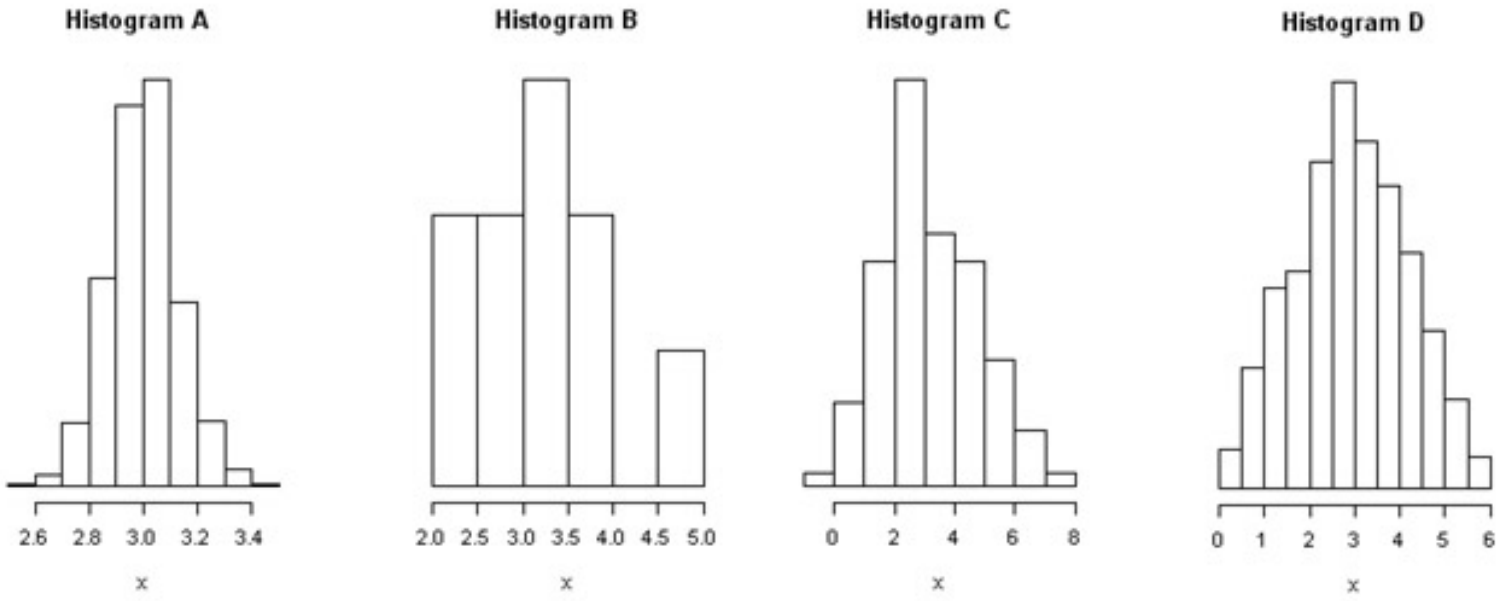
1. histogram of a sample of size 100 from a normal(3,1.8)

2. histogram of 1000 sample means based on 1000 random samples of size 200 from a normal(3,1.8)

3. histogram of a sample of size 10 from a uniform[2,5]

4. histogram of 1000 sample means based on 1000 random samples of size 2 from a uniform[0,6]

Question: For which (if any) of the above 4 descriptions did you need the Central Limit Theorem to deduce your selection?

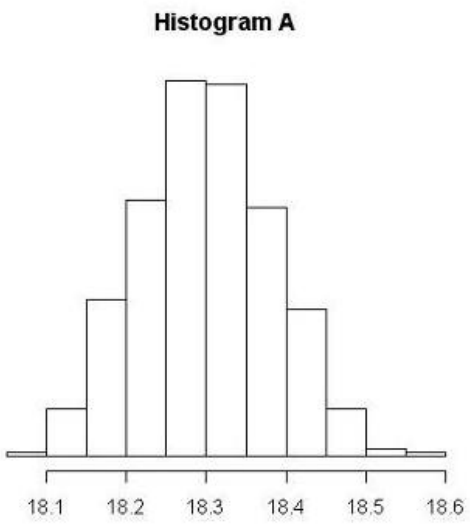


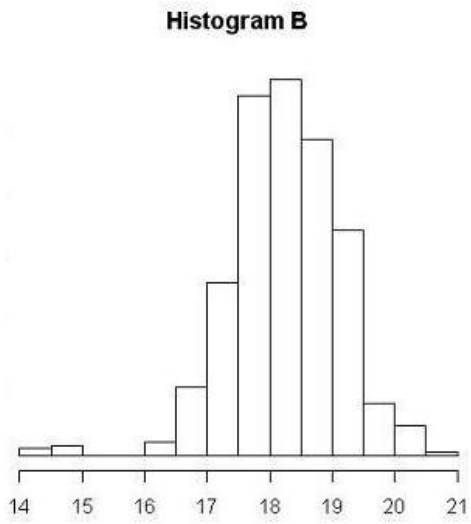
Past midterm questions

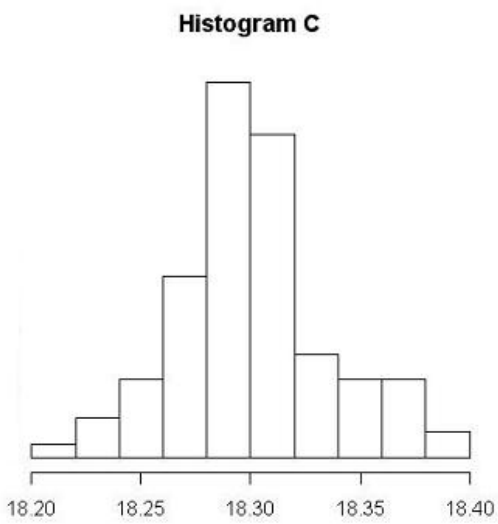
Winter 2008, Term 2, Midterm

In North America, students' ages at entry to university follow a distribution that is skewed to the left. The mean age is 18.3 years and the standard deviation of the ages is 0.8 years.

(a) Below are three descriptions related to ages of students when they enter university. Match the descriptions to the three histograms (A, B and C).







Description I: Histogram of ages of a random sample of 500 students when they enter university. This is Histogram ___.

Description II: Histogram of 100 sample mean ages at entry to university based on 100 random samples of size 500. This is Histogram ___.

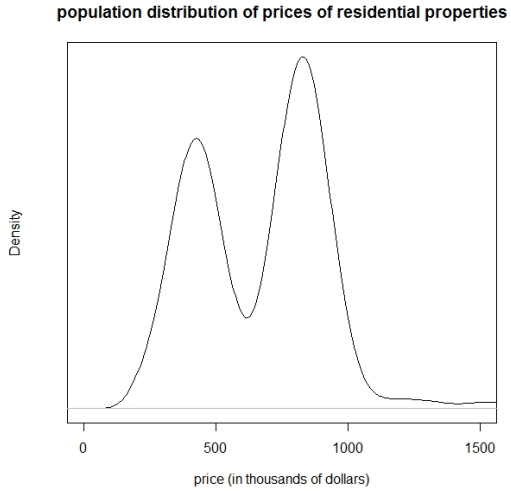
Description III: Histogram of 500 sample mean ages at entry to university based on 500 random samples of size 100. This is Histogram ___.

(b) In part (a), for which (if any) of the descriptions did you need the Central Limit Theorem to deduce your selection? Circle all that apply.

I II III None of the three

Summer 2009, Midterm

The prices of all residential properties (including houses, condos, apartments) in Vancouver follow a distribution that has a mean of 668.6 and a standard deviation of 216.4 thousands of dollars. The following graph shows the shape (a sketch of the density function) of the price distribution.



(a) Below are three descriptions related to the price of residential properties in Vancouver. Match the descriptions to the correct histograms (A, B, C, D and E) on the next page.

Description I:
Histogram of prices of a random sample of 500 residential properties. This is Histogram ___.

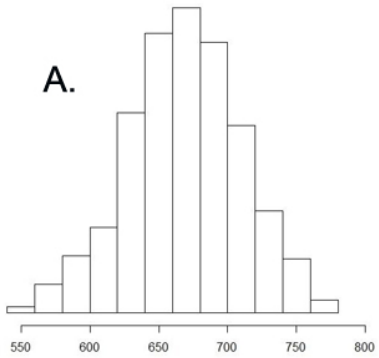
Description II:
Histogram of sample mean prices based on repeated random samples of 100 residential properties. This is Histogram ___.

Description III:
Histogram of sample mean prices based on repeated random samples of 30 residential properties. This is Histogram ___.

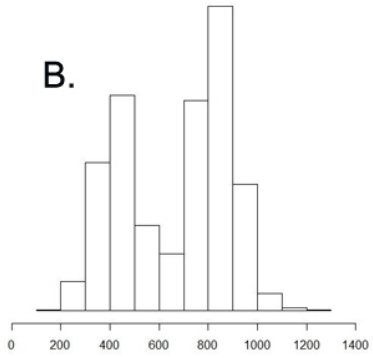
(b) In part (a), for which (if any) of the descriptions could one have applied the Central Limit Theorem to deduce the selection? Circle all that apply.

I II III None of the three

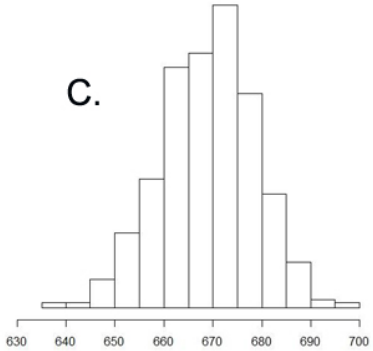
A.


550
600
650
700
750
800

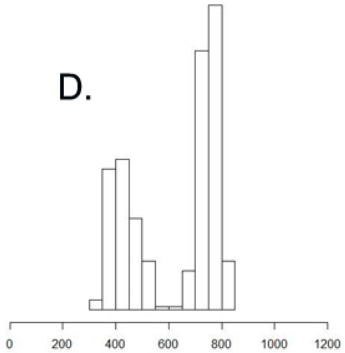
B.


0
200
400
600
800
1000
1200
1400

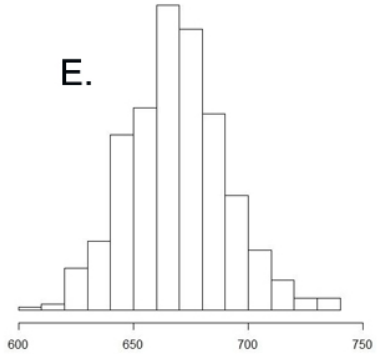
C.


630
640
650
660
670
680
690
700

D.


0
200
400
600
800
1000
1200

E.


600
650
700
750

Results for lab and midterm questions

- Analysis method: Ordered logistic regression
- Response: students' marks are treated as ordinal
- Main effect: applet versus simulation (1-applet, 0-simulation)
- Covariates:
  - TA
  - Lecture section (in W08,T2 analysis only)
  - Lab mark (in midterm analysis only)

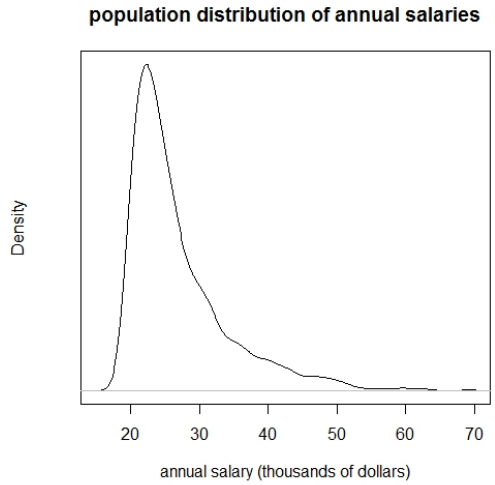
	Coefficient for the use of applet (1-applet, 0-simulation)	
Response	Winter08,T2	Summer09
Lab (matching histograms)	+	−
Lab (use of CLT)	−	+
Midterm (matching histograms)	+	+
Midterm (use of CLT)	+	+

None of the coefficients are significantly different from 0.

Past final exam question

Summer 2009, Final exam

The annual salaries of all high school graduates from BC follow a skewed distribution that has a mean of 27.1 and a standard deviation of 7.6 thousands of dollars. The following graph shows the shape (a sketch of the density function) of the annual salary distribution.



Below are two descriptions related to the annual salaries of high school graduates from BC. Match the descriptions to the correct histograms (A, B, C, D and E) on the next page. Also justify your choice.

Description I:
Histogram of sample mean annual salaries from 1000 random samples of 50 high school graduates. This is Histogram ___.

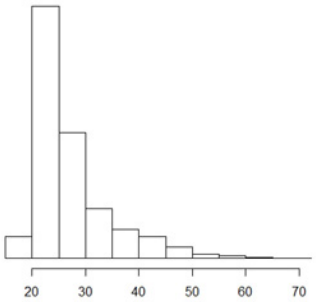
Explain your choice here:

Description II:
Histogram of annual salaries of a random sample of 200 high school graduates.

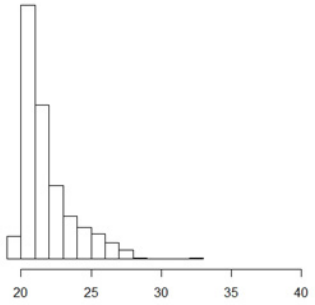
This is Histogram ___.

Explain your choice here:

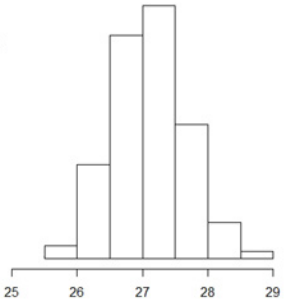
A.


20
30
40
50
60
70

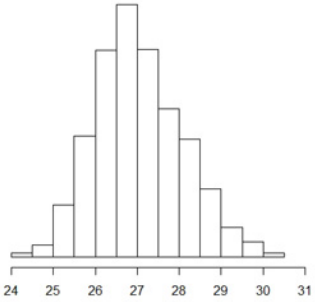
B.


20
25
30
35
40

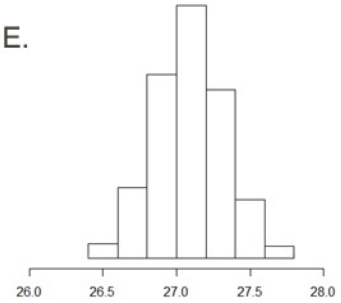
C.


25
26
27
28
29

D.


24
25
26
27
28
29
30
31

E.


26.0
26.5
27.0
27.5
28.0

Results for the summer09 final exam question

Description I: Histogram of sample mean annual salaries from 1000 random samples of 50 high school graduates.
(m = number of samples; n = sample size)

	Applet (N=41)	Simulation (N=28)
% correct histogram match	43.9%	48.2%
Justification:		
- correct explanation	**9.8%**	**33.3%**
- had m and n switched, would have been correct otherwise	9.8%	3.7%
- knew it is related to $\overline{X}$, but could not identify m and n	**34.2%**	**14.8%**
- other	46.3%	48.1%

For the justification part, the result is marginally significant (Fisher's Exact test: p-value = 0.054)

Description II: Histogram of annual salaries of a random sample of 200 high school graduates.

	Applet (N=41)	Simulation (N=28)
% correct histogram match	**63.4%**	**40.7%**
Justification:		
- correct explanation	**48.8%**	**37.0%**
- indicated that the sample data should have a smaller spread	24.4%	22.2%
- thought that it is a histogram of $\overline{X}$	17.1%	22.2%
- no or completely irrelevant explanation	**9.8%**	**18.5%**

The % of correct histogram match is not significantly different between the two methods (Chi-square test: p-value = 0.11). For the justification part, the result is not significant (Fisher's Exact test: p-value = 0.63)

Discussion and summary

- There is no evidence suggesting one of the two methods (applet and simulation) is more effective

- Despite insignifiant results, the applet lab seems to have a slightly more positive effect on the students' midterm performance

- Based on the summer09 final exam results,
  - students receiving the applet lab seem to better recognize the sample data and that it should resemble the shape of the population distribution

  - students receiving the simulation lab seem to be better able to identify the number of samples and the sample size in constructing a sampling distribution

- Common misconceptions:
  - as sample size increases, the spread decreases for both sample data and sampling distribution of $\overline{X}$
  - the sampling distribution of $\bar{X}$ resembles the population distribution when n is small
  - the normality of $\overline{X}$ for normally distributed $X_i's$ is the result of the CLT