Answer Key

Part I Study Design Practice Problems

Controlled Experiments—Researchers assigns subjects to treatment and control groups
Main Idea: Treatment and Control should be as much alike as possible

- Randomized, double-blind design is ideal because it eliminates systematic differences (bias). Random differences average out with enough subjects.
- Blocking reduces random differences that could be a problem for small studies by breaking subjects into similar sub-groups ***before*** randomization .
- Once subjects are randomized into treatment and control, NEVER rearrange them because it will introduce bias. That's why we compare the results of everyone in treatment to everyone in control whether or not they adhere.
- Non-randomized controls usually introduce systematic differences between treatment and control groups that could bias the result. These differences are called *confounders.*

Observational Studies—Subjects themselves or simple fate determines treatment and control groups. Researcher just observes.
Main Idea: Treatment and Control groups are likely to be systematically different, these differences can mix up or confound the results.

- Very difficult to conclude causation from association.
- With observational studies you must always think about what the likely confounders.
- Stratification adjusts for possible confounders by breaking subjects into sub groups where the confounding factor is the same.
- Simpson's Paradox is an example of extreme confounding. It's paradoxical because you get one result before stratification and the opposite afterwards!

Sample Problems:

1. A study was done to test whether Ginkgo biloba (GB) could alleviate symptoms of Alzheimer's and dementia. The 52-week study randomly assigned half of the patients take GB daily and half to take a placebo. Neither the subjects nor evaluators knew who was in each group. At the end of the study, there was significant evidence that GB improved the cognitive performance and the social functioning of the patients for 6 months to 1 year.

a) What type of bias could be present in this study ***Choose one:***
(i) No systematic bias ii) Subject Bias iii) Evaluator Bias iv) Selection Bias v) ii, iii, and iv
placebo double blind random

b) Which of the following could confound the results? ***Choose one:***

- **i)** Forgetfulness- Patients with dementia may forget to take the GB on a regular basis.
- **ii)** Increased Attention-- Participation in the study increased the attention these patients received. They felt less neglected and therefore more cognitively active.
- **iii)** More motivated-- Those who volunteered to be in the GB group were probably more conscientious and motivated to begin with since they actively sought a remedy for their condition.
- **iv)** All of the above
- **(v)** None of the above

* randomized controlled experiments do not have confounders

c) Not everyone in the treatment and control group adhered to the program and took their medicine/placebo. Which comparison is best when analyzing the final data? everyone in treatment to everyone in control

- **(i)** Compare everyone assigned to take the GB to everyone assigned to take the placebo.
- **ii)** Compare everyone who actually took GB to everyone who didn't actually take GB .
- **iii)** Compare only those who took the GB regularly to only those who took the placebo regularly.

2) Two experiments were done comparing the effects of listening to classical music versus pop music while studying. All the students in both experimental designs were given an identical 2-hour lesson and then allowed time to study for a short exam. obs. study

- **In Design A students themselves chose** to study either listening to classical or pop.
- In **Design B the students were randomly assigned** to study either listening to classical or pop. ← trust results

Design A found that the classical study group scored significantly higher on the exam than the pop group did. Design B found no significant difference in exam scores between the 2 groups. **The overall exam average in both designs was the same.**

a) Which design had randomized controls? A only **(B only)** Both Neither

b) Which design is more likely to have confounders? **(A)** B Both are equally likely

c) Which conclusion is best supported by the evidence? ***Circle one***

i) Students learn better when they are able to choose their own music while studying.
(ii) Students who choose classical are different in more ways than just their musical tastes than students who choose pop
iii) Classical music seems to enhance learning better than pop music.

3) A study published in the March 4, 2015 issue of the Journal of the American Medical Association evaluated whether peanut consumption might be more effective than peanut avoidance in preventing the development of peanut allergies in infants who are at high risk for the allergy. 640 infants aged 4 to 11 months with severe eczema and egg allergies (high risk indicators for peanut allergy) were ***randomly assigned*** to either consume (treatment) or avoid peanuts (control) until 5 years of age. The results were striking—17.2% of the children in the peanut-avoidance group tested positive for peanut allergy while only 3.2% of the group in the peanut-consumption group tested positive.

a) Which of the following best describes this study:

(i) A randomized controlled experiment
ii) An observational study with controls
iii) A non-randomized controlled experiment

Good study ⇒ trust results
* no confounders

b) Does the study show that eating peanuts helped prevent the children in the study from developing a peanut allergy?

i) No, it only shows that there is an association between peanut consumption and reduced rate of peanut allergy since many environmental, cultural, social and biological factors contribute to both diet and allergic responses.
ii) No, simply assigning children to 2 groups without considering the consequences of how peanut consumption or peanut avoidance may confer nutritional advantages limits any causal conclusions.
(iii) Yes, the study is strong evidence that peanut consumption helped prevent peanut allergy in these children although the causal mechanism can only be inferred.

c) Which of the following could confound the results? Circle Yes or No for each. none ⇒ randomized

i) Cultural/Ethnic differences- Peanuts and peanut oil are popular in West African and Southeast Asian cuisines, groups that have a relatively low incidence of peanut allergies. **a.** Yes (**b.** No)

ii) Health Benefits – Peanuts are a relatively healthy snack food. Children who eat peanuts may be healthier in general and less likely to develop allergies. **a.** Yes (**b.** No)

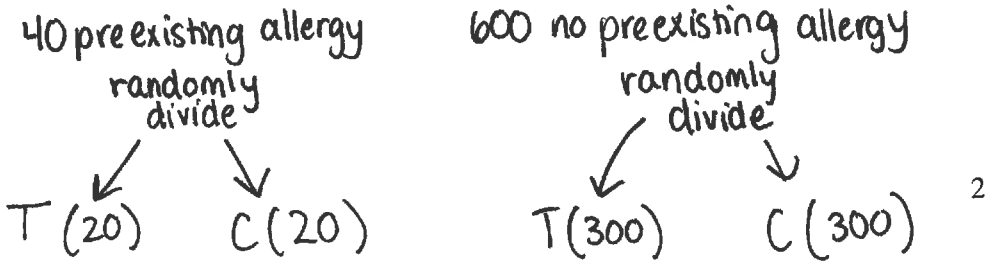
iii) Pre-existing Health Problems- The children all had severe health problems to begin with making it difficult to discern whether or not it was the peanuts or pre-existing conditions that led to the development of a peanut allergy. **a.** Yes (**b.** No)

iv) Overactive Immune System- Children with overactive immune systems are both more likely to have egg allergies (like the children in the study) and to develop a peanut allergy. **a.** Yes (**b.** No)

d) 40 of the 640 infants showed evidence (by a skin-prick test) of already having a peanut allergy before they were even assigned to treatment or control. The researchers want to make sure that the 40 children are exactly evenly divided between the treatment and control groups but they don't want to introduce bias. What should they do?

Blocking

(i) They should divide the infants into 2 groups (40 with pre-existing peanut allergy, and 600 without). Then randomly assign half of each group to treatment and half to control.
ii) Randomly assign half of the 640 infants to treatment and half to control. This will ensure the infants will be evenly divided on all characteristics relevant to the response including pre-existing peanut allergy.
iii) Randomly assign half of the 640 infants to treatment and half to control. In the unlikely event that the 2 groups are not balanced then, the researchers should balance the groups taking into account all variables to be as objective as possible.



4) A study published in the Feb 18, 2004 issue of the Journal of the American Medical Association compared pharmacy and medical records of 10,219 women and found that women who filled 25 or more prescriptions for antibiotics over a 17 year period received breast cancer diagnoses at twice the rate as those who took no antibiotics. The study concluded that high antibiotic usage increases one's risk of breast cancer.

a) Which of the following statements best describes this study? ***Circle one:***

i) This was a randomized controlled experiment without a placebo.
(ii) This was an observational study with controls. controls = comparison group
iii) This was a randomized controlled double-blind experiment.
iv) This was a non-randomized controlled experiment with a placebo.

b) Based on the results of this study alone, which of the following statements is best? ***Circle one.***

i) High antibiotic use causes an increased risk of breast cancer. can't conclude causation from obs. stud
(ii) High antibiotic use is *associated* with and *may* cause increased breast cancer risk.
iii) High antibiotic use is *associated* with but does *not* cause increased breast cancer risk.
iv) Having cancer is likely to cause increased use of antibiotics.

c) Below are either confounders, causal links or neither. Answer based only on given information.

i) Age of first pregnancy- women who have their first child after the age of 35 are more likely to get breast cancer.
a) Confounder **b)** Causal Link (c) Neither prescriptions for antibiotics ? [age] → breast cancer

ii) Destruction of Protective Bacteria- antibiotics kill healthy bacteria that may help prevent breast cancer.
a) Confounder (b) Causal Link c) Neither antibiotics → [destroy good bacteria] → breast cance

iii) Underlying Immune Problem- a weak immune system leads both to frequent infections necessitating antibiotics *and* also to a higher cancer risk.
(a) Confounder **b)** Causal Link c) Neither antibiotics ← [immune problem] → breast cancer

iv) Regular Check-ups- Women who regularly go to the doctor are both more likely to be prescribed antibiotics and more likely to receive a breast cancer diagnosis (especially for slow growing cancers that are unlikely to lead to serious health problems.)
(a) Confounder **b)** Causal Link **c)** Neither antibiotics ← [check ups] → breast cancer

d) Suppose the researchers thought that income was a possible confounder since high income women tend to take more antibiotics and tend to get more breast cancer. To separate out the effects of income from the effects of antibiotics researchers should ... **Circle one:** Strahify

i) split the data into high, middle and low income groups and compare the antibiotic usage between the 3 groups.
(ii) split the data into high, middle and low income groups and compare the cancer rate of those who took a lot of antibiotics to those who took no antibiotics within each group.
iii) split the data into high and low antibiotic users and compare the cancer rates between the groups.
iv) split the data into 2 groups—breast cancer and no breast cancer and compare antibiotic usage between the 2 groups.

Treatment Group
antibiotics
split by income

Control Group
no antibiotics
split by income

high ← compare → high
middle ← compare → middle
low ← compare → low

5) A study published in the August 15, 2017 issue of *Mayo Clinic Proceedings* tracked 44,000 people aged 20 to 87 for an average of about 16 years and found that those who drank 4 or more cups of coffee a day were 21% more likely to die than those who drank less than 4 cups a day. The risk was 50% higher for heavy coffee drinkers under 55 years of age.

b) Which of the following best describes this study? - can have confounders can't conclude causation

- (i) An observational study with controls
- ii) A randomized controlled experiment
- iii) A non-randomized experiment with historical controls

c) Does the study show that drinking 4 or more cups of coffee a day caused the higher death rate?

- **i)** No, the study was conducted over such a long time period that it's difficult to determine whether it was the original coffee drinking itself or something *else* about the coffee (for example, the way it was brewed) that caused the higher death rate.
- **ii)** Yes, particularly for young people, the study clearly shows that excessive coffee drinking caused an increased risk of death.
- (iii) No, it's possible that coffee drinkers share other traits (besides the coffee) that could put them at a higher risk of dying. confounders
- **iv)** No, you cannot conclude causation without a proven causal mechanism. The study does provide strong evidence that it's the coffee that's raising the death rate and not something else, but it fails to explain how or why.

c) The study reported that they controlled for cigarette smoking. This means they thought smoking might be a confounder so they eliminated its confounding effect. How did they do that? ***Choose one:*** Stratification

- **i)** At the beginning of the study, they divided the patients into smokers and non-smokers and then randomly divided the smokers and non-smokers equally between the coffee and no coffee groups.
- **ii)** Throughout the study they eliminated anyone who smoked from the study.
- (iii) At the end of the study, they stratified on smoking, and compared the death rate of coffee drinkers to non-coffee drinkers within each smoking level (non-smokers, light smokers, heavy smokers).

d) State whether the following are confounders, causal links, or neither:

- **i)** Increased popularity of coffee- The study was conducted over a 16-year time period that coincided with an enormous increase in coffee consumption. **a)** confounder **b)** causal link (c) neither

- **ii)** Caffeine—Excessive caffeine intake from 4 cups of coffee per day raises health risks because it increases a person's heart rate and blood pressure, which increase one's risk of death.
  a) confounder (b) causal link **c)** neither

what about the coffee causes death?

- **iii)** Unhealthy Diet – The study stated that people who drank 4 or more cups of coffee were also more likely to have an unhealthy diet that could increase one's risk of death.
  (a) confounder **b)** causal link **c)** neither

3rd factor
difference btwn people who drink coffee + people who don't

- **iv)** Pre-existing-conditions- Some members of the study may have had pre-existing conditions or illness that would cause them to die sooner. **a)** confounder **b)** causal link (c) neither.

6) A study was done to test the effectiveness of a new weight loss drug. The subjects were 2000 obese adults. Half were randomly assigned to take the drug every day and half were randomly assigned to take the placebo every day. Neither the subjects nor those who evaluated them knew who was in which group. The subjects were followed for 1 year and the percent of weight they lost or gained was recorded.

a) Based only on the information above which of the following best describes the study above?
Choose one:

- **i)** This was a non-randomized controlled experiment with a placebo.
- **ii)** This was a randomized controlled experiment without a placebo.
- **iii)** This was an observational study with controls.
- **(iv)** This was a randomized controlled double-blind experiment. good! ideal!

b) The table below gives the average percent weight change of "adherers" and "non-adherers" in both the drug and the placebo group. Adherers regularly took their pills while non-adherers took their pills less than 80% of the time.

← randomly divided →

	Drug		Placebo	
	Number	**%Weight change**	**Number**	**%Weight change**
Adherers	500	7% loss	502	7.1% lost
Non-Adherers	500	2% gain	498	2.1% gain
Total	1000	(2.5 loss)	1000	(2.52% lost)

* look @ overall results to keep the randomization!

Based on the results of the table would you conclude there is good evidence for the following statements?
Circle YES or NO after each statement:

- **i)** The drug worked better than the placebo for those who regularly took the medicine. YES (NO)
- **ii)** The drug works no better than a placebo (YES) NO
- **iii)** Adherers may be different than non-adherers in ways that help them lose weight. (YES) NO

(for example, more responsible about eating balanced meals, exercising regularly, etc.)

7) A country club gives a pass-fail golf test every year to professional and amateur golfers. Professionals have a much higher % passing than amateurs. The club members were happy that the overall % passing went up from 68% in 2007 to 70% in 2017 and wanted to know which group contributed to the improved rate.

	2007				2017			
	Number	**# Passes**	**# Failures**	**% Passing**	**Number**	**#Passes**	**# Failures**	**% Passing**
Professionals	100	92	8	92%	100	90	10	90%
Amateurs	300	180	120	60%	100	50	50	50%
Overall Total	400	272	128	(68%)	200	140	60	(70%)

Simpson's Paradox

misleading

a) Which group's % passing went up from 2007 to 2017? ***Choose one:*** a) Prof. b) Amat. (c) Neither d) Both

b) Is it possible for each group's % passing to go down if their overall % passing goes up? yes b/c of confounders

(i) Yes, it's possible because the overall makeup of the club has changed from 25% to 50% professionals which raises the overall % passing even though both groups % passing declined.

ii) No, it's not possible. If the overall passing rate goes up, then at least one group's passing rates must go up.

* with obs. studies, never look @ overall results

Part II Descriptive Statistics

Chapter 3 – Measures of Center and Spread

8) Look at this list of 5 numbers: 0, 1, -2, 2, 9

a) The average is 2 (Step 1) $\frac{0+1+(-2)+2+9}{5} = 2$

b) The median is 1. $-2, 0, 1, 2, 9$ * list in order first!

c) The **deviations** from the **average** are $-2, -1, -4, 0, 7$ Step 2

d) The sum of the deviations from the average should = 0. *Fill in the blank with a number.*

e) Compute the Standard Deviation. Round your answer to 2 decimal places.
Show your work. You may start with the deviations you found in part (c). ***Circle answer.***

Step 3: square deviations $4, 1, 16, 0, 49$ Step 4: avg of Step 3 $\frac{4+1+16+0+49}{5} = 14$ Step 5: $\sqrt{14} = 3.74$

9) A list of 10 numbers has an average = 5, median= 4, and SD = 3. Fill in the chart below with numbers.

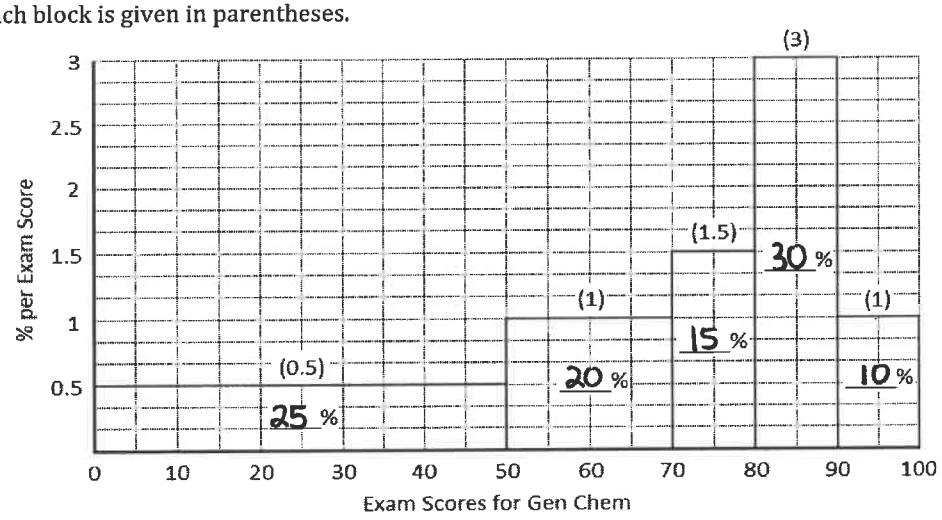
For (a-e) below, calculate the new average, median, and SD after the original list has been changed according to the given directions.	**New Average** (Write a number, not words, like "increase" or "decrease")	**New Median** (Write a number, not words, like "increase" or "decrease".)	**New SD** (Write a number, not words, like "increase" or "decrease" except for (e).
a) 5 is added to every number on the original list.	10	9	3
b) Every number on the original list is multiplied by **negative** 2.	-10	-8	6
c) Every number on the original list is divided by 2.	2.5	2	1.5
d) Subtract 5 from every number on the original list, and then divide every number by 3.	0	$-\frac{1}{3}$	1
e) Every number on the original list remains the same, EXCEPT that 10 is added to the largest number.	increase (6)	4	**Choose one:** (i) Increase ii) Decrease iii) Stays the same

$\text{avg} = 5 = \frac{\text{sum of numbers}}{10} \Rightarrow \text{sum of \#'s} = 50$

$\text{new avg} = \frac{\text{sum of \#'s} + 10}{10} = \frac{50+10}{10} = 6$

Chapter 4 Graphical Displays for Numerical Data

10) The figure below is a histogram for the first exam scores of 520 freshmen and sophomores in general chemistry. The height of each block is given in parentheses.



a) What percent of the students received an exam score between 0 and 50? Write your answer inside the blank provided in the 0-50 interval on the histogram. Do the same for the other 4 intervals. **Fill in ALL 5 blanks in each block of the histogram above with the correct areas.**

b) The area of the entire histogram is 100 % always!

c) The median exam score is **closest** to: ***Choose one:*** 50 70 (73) 80 90 — 50th percentile

d) Is the median >, <, or = to the average? > — long lefthand tail ⇒ avg < med

e) The percent of students who received exactly 75 on their first exam is closest to (Assume an equal distribution throughout the interval)
Choose one: 0.5% 1% (1.5%) 10% 15%

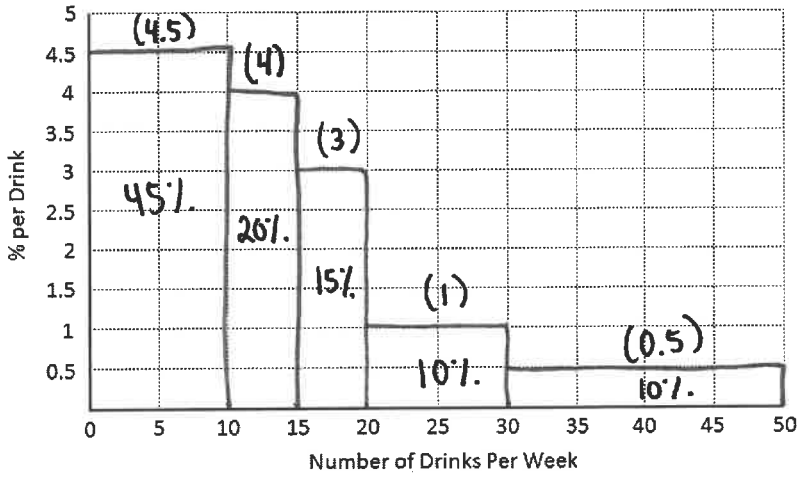
f) Suppose all the students in the 0-30 range were given extra credit that raised each of their scores 20 points? How would that affect the average, median and Standard Deviation?
(Check the appropriate boxes below, check only 1 box per row.)

	Increase	Decrease	Stay the same	Not enough information
Average would ...	X			
Median would ...			X	
Standard Deviation would ...		X		

11) A distribution table for the number of drinks a past semester of Stat 100 students said they typically consumed per week is shown below. The first row says that 45% of students said they had between 0 and 10 drinks per week. The table has 5 missing blanks. Fill them in with the correct widths, heights, and areas. Then draw the histogram. Write the area of each interval inside the block.

a) Fill in the 5 blanks in the table below and **then draw the histogram on the graph below.**

Interval	Width of Interval	Height (% per Drink)	Area (%)
0 to 10	10	4.5	45
10 to 15	5	4	20
15 to 20	5	3	15
20 to 30	10	1	10
30 to 50	20	0.5	10



b) The area column should sum to 100 %. **Fill in blank.**

c) If someone drinks more than 90% of the class, how much does he or she drink per week? 30 drinks **Fill in blank.**

d) Would it be appropriate to use a normal approximation for this data? long right hand tail

Choose one:

(i) No, the histogram is far from normal, so using a normal approximation would not be appropriate.

ii) Yes, because converting to z-scores will change the shape and make the histogram normal.

iii) Yes, because the normal approximation is suitable for all data sets.

iv) Yes, because we can determine the average and SD from the data.

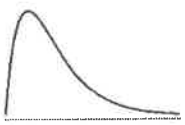
e) The Survey only allowed students to give answers up to 50 drinks. I gave everyone who answered 50 the opportunity to change their answers. A few of them changed their answer from 50 to 60 drinks. How would that affect the average, median and standard deviation?

(Check the appropriate boxes below, check only 1 box per row.)

	Increase	Decrease	Stay the same	Not enough information
Average would ...	X			
Median would ...			X	
Standard Deviation would ...	X			

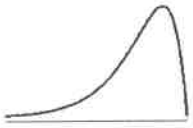
12) Below are rough sketches of 2 histograms. One depicts scores on an Easy exam were most students did well. One depicts scores on a hard exam where most students did poorly. The horizontal axis ranges from 0% to 100%.

Histogram A



0 100

Histogram B



0 100

a) Which histogram depicts the easy exam? (1 pt.)
Choose one:

i) Histogram A
ii) Histogram B

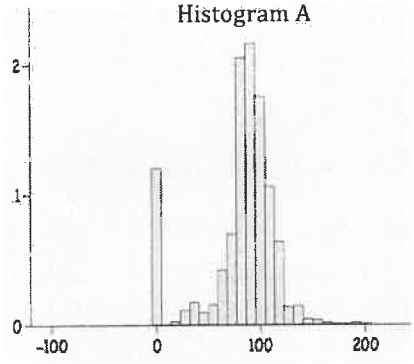
b) In Histogram A, is the average greater than, less than, or equal to the median? ***Circle one:*** (>) < = (1 pt)

c) In Histogram B, is the average greater than, less than, or equal to the median? ***Circle one:*** > (<) = (1 pt)

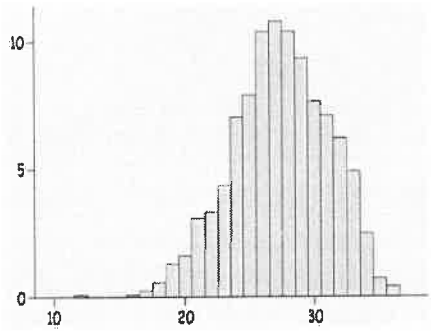
13) If a list of numbers has a SD of 0 then ....
a) All the numbers on the list must be the same. no spread
b) The average of the numbers must be 0.
c) All the numbers on the list must be 0.
d) There are 0 numbers on the list since the SD can never be 0.

14) Look at the 3 histograms below representing your survey responses to 3 questions:
What is your ACT score?
What's the fastest speed you've ever driven (in mph)?
What percent of your college costs are your parents paying for?

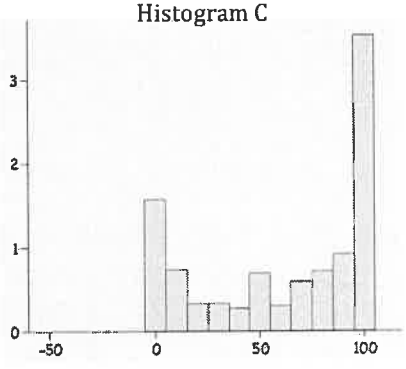
Histogram A



Histogram B



Histogram C



a) Which graph represents ACT scores? B Which graph represents fastest speed? A

b) I wrote the average and median of Histogram C down, but I forgot to label them.
Here are the 2 numbers: 62.25 and 80. Which is which? avge med

(i) 80 is the median ii) 80 is the average iii) Cannot be determined

↓
whole #
or .5

Chapter 5—Normal Approximation

15) According to our survey data, the histogram for the heights of females in our class is close to the normal curve with an **average = 65 inches and a SD =3 inches.**

a) If a female is below average in height, is her Z score positive or negative?
Choose One:

i) Positive
(ii) Negative
iii) Not enough information to tell

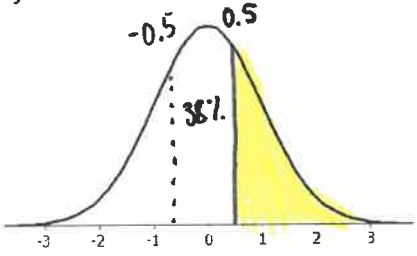
b) If a student is exactly at the 50th percentile, her Z score = 0 and she is 65 inches.
(Fill in the 2 blanks above with numbers.) → she is exactly average

c) What percent of the females are taller than 66.5 inches? (Use then normal curve , you may round percents on the table to the nearest whole number.)

i) First convert 66.5" to a Z score, show work.

Z = 0.5 $\quad z = \frac{66.5-65}{3} = 0.5$

ii) Then mark the Z score on the curve below and shade the area that represents everyone ***over*** 66.5".



Percent over 66.5" = 31 %

Write your answer in the blank above.

$\frac{100-38}{2} = 31\%$

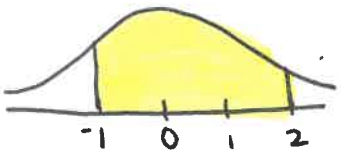
) Which of the following is closest to the percentage of females in the class who are between 62" and 68"?
Choose One:

(i) 68%
ii) 82%
iii) 91%
iv) 95%

$z = \frac{62-65}{3} = -1$ $\quad z = \frac{68-65}{3} = 1$ convert to z-scores + find the area in between.

) Which of the following is closest to the percentage of females in the class who are between 62" and 71"?
Choose One:

i) 68%
(ii) 82%
iii) 91%
iv) 95%



$z = -1$ $\quad z = \frac{71-65}{3} = 2$

Area = $\frac{1}{2}(68) + \frac{1}{2}(95) = 81.5\%$

) About 50% of the females are between 63" and 67". Are there more or less females between 65" and 69"?
Choose One:

$z = -0.67$ + $z = 0.67$ $\quad z = 0$ $\quad z = 1.33$

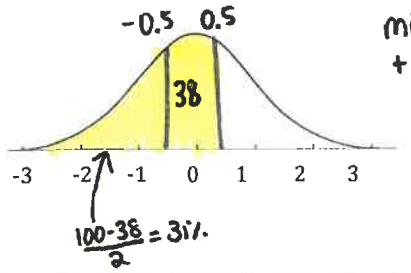
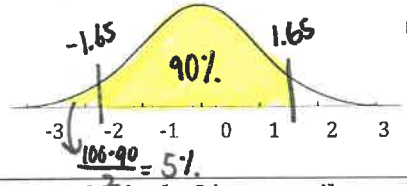
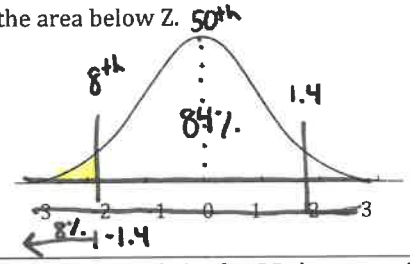
i) More females are between 65" and 69" than between 63" and 67".
(ii) Less females are between 65" and 69" than between 63" and 67".
iii) The 2 amounts are the same because the height difference is the same, 4" for both groups.
iv) There is not enough information to tell.

Area = $\frac{1}{2}(82\%) = 41\%$

) Suppose you found out that the heights were far from normally distributed but still had **average** = 65" and **SD** = 3", would your answers arts **d, e, f** above change or stay the same? * yes! we can't use the normal curve
Choose One:

i) The answers would be the same because the average and SD did not change.
(ii) The answers may change because the distribution is not normal and the table is therefore inaccurate.

16) Suppose IQ scores follow the normal curve with **an average=100 and a SD =16**. In the table below you're given either an IQ score, a Z score or percentile and you have to fill in the missing blanks. **For all these problems, please round the Areas given in the Normal Table to the nearest whole number.**

IQ	Z score	Percentile (% of people with lower IQ scores area to the left of Z.
a) Person A has IQ= 108	**Z=** 0.5 **Show work:** $Z = \frac{val - avg}{SD}$ $= \frac{108-100}{16} = 0.5$	***Person A is in the*** 69th **percentile** Mark Z score on curve and shade the area below Z 
IQ = 126.4 **Do NOT round answer.** **Show work:** val=avg+(z)(SD) = 100+(1.65)(16)=126.4	**Person B has Z= 1.65** person B is 1.65 SDs above average	***Person B is in the*** 95th ***percentile.*** Mark Z score on curve, and shade the area below Z 
IQ = 77.6 **Do NOT round answer.** **Show work:** val=avg+z(SD) = 100+(-1.4)(16)= 77.6	**Z=** -1.4	***Person C is in the*** **8th percentile** What middle area should you look up on the normal table to find the correct Z score? 84 % Mark the correct Z score on curve, and shade the area below Z. 50th  if 2 percentiles sum to 100, their z-scores are opposites
IQ= 122.4 **Do NOT round answer.** **Show work:** Z = 100+(1.4)(16) = 122.4	**Z=** 1.4	**Person D is in the 92nd percentile.** No work is necessary. Just use the Z score you got for the 8th percentile to get the Z score for the 92nd percentile. **Hint:** The 8th and 92nd percentiles are both the same distance from the 50th percentile.

Part III Probability

17) This question pertains to these 5 tickets. 0 2 3 3 7

a) Two tickets are drawn at random **with** replacement. What is the chance that both tickets shaded?
a) 3/5 x 2/4 **b)** 3/5 x 3/5 **c)** 3/5 **d)** 1/5 x 1/5 **e)** 2/5 x 1/4

b) Two tickets are drawn at random **without** replacement. What is the chance that both tickets are shaded?
a) 3/5 x 2/4 **b)** 3/5 x 3/5 **c)** 3/5 **d)** 1/5 x 1/5 **e)** 2/5 x 1/4

c) Five tickets are drawn at random **with** replacement. What is the chance of getting at least one shaded ticket?
a) $1 - (3/5)^5$ **b)** $(3/5)^5$ **c)** $1 - (4/5)^5$ **d)** $(4/5)^5$ **e)** $1 - (2/5)^5$

$P(\text{at least one}) = 1 - P(\text{none}) = 1 - (\frac{2}{5})^5$

d) One ticket is randomly drawn. What is the chance of getting either a shaded ticket or a ticket marked "3"?
a) 2/5 **b)** 4/5 **c)** 3/5 **d)** 1

$P(\text{shaded or 3}) = P(\text{shaded}) + P(3) - P(\text{both})$

$\frac{3}{5} + \frac{2}{5} - \frac{1}{5}$

18) This questions pertain to rolling fair dice.

2,3 1,4
3,2 4,1

a) Two dice are rolled. What is the chance that the sum of the spots is 5?
i) 2/36 ii) 3/36 iii) 4/36 iv) 5/36 v) 1/6*1/6 v) 1/6 + 1/6

b) One die is rolled 3 times. What is the chance of getting all 6's?
i) $(5/6)^3$ ii) $(1/6)^3$ iii) $1 - (5/6)^3$ iv) $1 - (1/6)^3$ v) 3/6

c) One die is rolled 3 times. What is the chance of not getting all 6's? $P(\text{not all 6's}) = 1 - P(\text{all 6's})$
i) $(5/6)^3$ ii) $(1/6)^3$ iii) $1 - (5/6)^3$ iv) $1 - (1/6)^3$ v) 3/6

d) One die is rolled 3 times. What is the chance of getting at least one 6? $= 1 - P(\text{no 6's})$
i) $(5/6)^3$ ii) $(1/6)^3$ iii) $1 - (5/6)^3$ iv) $1 - (1/6)^3$ v) 3/6

e) Two dice are rolled. What is the chance of getting a 3 on the first roll or a 4 on the second roll?
i) 1/6 ii) 2/6 iii) 11/36 iv) 13/36 v) 1/6*1/6 v) 1/6 * 1/6 – 1/36

$P(3 \text{ on } 1^{st} \text{ OR } 4 \text{ on } 2^{nd}) = \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$

19) A screening test for AIDs correctly gives positive results to about 99% of the people who have AIDs and incorrectly gives positive results to about 6% of the people who don't have AIDs. 1% of the population who take the test have AIDs.
Fill in the table below to give the results for 10,000 people.

	Tests Positive	**Tests Negative**	**Total**
Has AIDS	$.99(100) = 99$	1	$.01(10{,}000) = 100$
Does Not have AIDS	$.06(9900) = 594$	9306	9,900
Total	693	9307	**10,000**

a) What fraction of the people who test negative truly have AIDs?

$P(\text{AIDS} \mid \text{test } \ominus) = \frac{1}{9307}$

b) What fraction of the people who test positive truly have AIDs?

$P(\text{AIDS} \mid \text{test } \oplus) = \frac{99}{693}$

Part IV: Statistics for Random Variables

Chapters 8-9 Box Models, EV, SE and Histograms for Random Variables

Translating gambling games into Box models and computing the EV and SE for the sum, average and % of n draws from a box.

- **EV of the sum of n draws from a box = n times the average of the box**
- **Know the 3 SE formulas:**

  Remember SE = SD either multiplied or divided by $\sqrt{n}$ (multiply SD by $\sqrt{n}$ **only** for **SE of sum**)

  - **SE of the sum of n draws from a box =** $SD_{Box} * \sqrt{n}$
  - **SE of the average of n draws from a box =** $SD_{Box} / \sqrt{n}$
  - **SE of the % of 1's in n draws from a 0-1 box=** $SD_{Box} / \sqrt{n}$ $(* 100)\ \%$

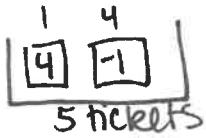
    (Multiply by 100 to change from a decimal to a percent, for example 0.1 x 100= 10%)

- **Know the short-cut formula for the SD of boxes that just have 2 types of tickets on page 50**
  If the box has only 1's and 0's this is the same as:

  $SD = \sqrt{p*(1-p)}$ where p is the proportion (fraction) of 1's in a 1-0 box.
- **Central Limit Theorem—The probability histogram for all possible sums (or averages, or percents) of draws from a box will get closer and closer to the normal curve.**
- **With enough draws we can use the normal curve to figure the chance that the sum (or average or percent) of the draws will fall within a given range by converting the endpoints of the interval into a Z score**
  Z = (Value – Expected Value)/ SE

* Draw box for one question

20) A 100 question multiple-choice test awards 4 points for each correct answer and subtracts 1 point for each incorrect answer. Each question has 5 choices.



i) Suppose a student guesses at random on each question, what is the corresponding box model?

a) It has two tickets: 1 and 0
b) It has 100 tickets: half 1's and half -1's
c) It has five tickets: 1, 0, 0, 0, 0
d) It has five tickets: 4, 0, 0, 0, 0
e) It has five tickets: 4, -1, -1, -1, -1

avg of box = 0

SD of box = $|4-(-1)|\sqrt{\frac{1}{5}\times\frac{4}{5}} = 2$

ii) The expected value for the student's score is $EV_{sum} = n \times \text{avg of box} = 100 \times 0$

a) 0 **b)** 10 **c)** 20 **d)** 40 **e)** 50

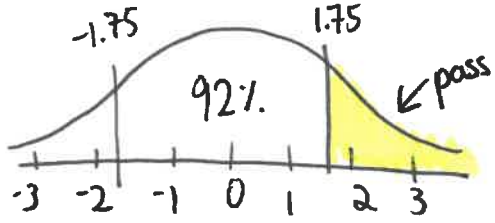
iii) The standard error of the student's score is $SE_{sum} = \sqrt{n} \times SD = \sqrt{100} \times 2 = 20$

a) 20 **b)** .4 **c)** 2 **d)** .2 **e)** not enough info

if it doesn't say avg or %, assume sum!

iv) Now suppose you're just interested in how many correct answers the student would get by guessing, not his score. Then the **EV = 20 and the SE = 4.** Suppose the student needs to get 27 answers correct in order to pass. What's the probability the student will pass? (Hint: convert to a Z score, and use the normal curve).

$Z = \frac{val - EV}{SE} = \frac{27-20}{4} = 1.75$



$\frac{100-92}{2} = 4\%$

21) Fill in the first blank with the number of draws, the second with either "with" or "without" and the third with the letter corresponding to the appropriate box model. Choose from the box models below. Use each box model exactly once.

Box A	Box B	Box C	Box D
1 -1/2 -1/2	0 1	1 2 3 4 5 6	0 0 0 1 0 0

a) A fair coin is tossed 25 times and the number or heads is counted. This corresponds to drawing 25 times with replacement from Box B

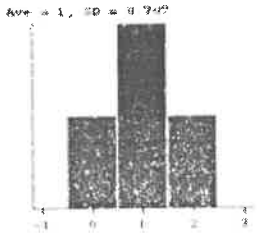
b) A pair of dice is rolled once and the total number of spots is counted. This corresponds to drawing 2 times with replacement from Box C

c) A die is rolled 50 times and the number of 4's is counted. This corresponds to drawing 50 times with replacement from Box D

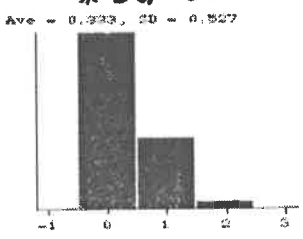
d) A multiple choice test has 100 questions. Each question had 3 options, only one of which is right. Suppose you randomly guess on all 100 questions, if you get a question right you get 1 point and if you get a question wrong you lose half of a point. This corresponds to drawing 100 times with replacement from Box A

e) Look at Box B and Box D above. The 4 histograms below are the probability histograms for the sum of 2 draws from Box B, 2 draws from Box D, 24 draws from Box B and 24 draws from Box D. Which is which?
Fill in the blanks below

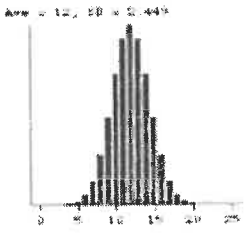
* more bars ⇒ more draws
* Box B is normal, Box D is lopsided ⇒ Box D will need more draws to look normal.



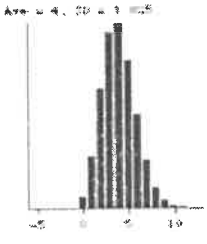
2 draws from Box B



2 draws from Box D



24 draws from Box B



24 draws from Box D

HINT—The Average and the SD given above each histogram is the EV and the SE of the sum of either 2 or 24 draws.

22) 400 draws are made at random with replacement from the box containing these 5 tickets: [2] [3] [4] [5] [6]

a) The smallest the sum of the 400 draws could possibly be is 800 (all 2's) and the largest is 2400 (all 6's)
(Fill in the 2 blanks above with the correct numbers)

b) What is the EV for the sum of the draws? **Show work below and circle your answer.**

$EV_{sum} = n \times \text{avg of box} = 4 \times 400 = (1600)$

c) What is the SE for the sum of the draws? (The SD of the box is 1.4)
Show work below and circle your answer.

$SE_{sum} = \sqrt{n} \times SD \text{ of box} = \sqrt{400} \times 1.4 = (28)$

d) Now suppose you draw at random with replacement from the same box above, but this time you're only interested in the percent of 3's that you get. What is the EV and the SE of the percent of 3's in 400 draws? *(Hint: draw a new box)*

i) What is the expected value of percent of 3's in 400 draws? 20 % 1/5 threes

[1] [0] 5 tickets

ii) *What is the SD of your new box?* 0.4 ***Show work.*** $SD = |1-0|\sqrt{\frac{1}{5} \times \frac{4}{5}} = 0.4$

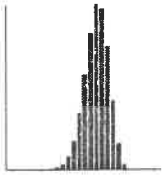
iii) What is the SE for the percent of 3's in 400 draws? ***Choose one.***

a) 1.33% (b) 2%) c) 4.67% d) 7% e) 20% f) 40%

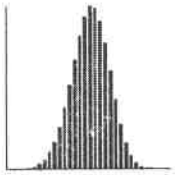
$SE_{\%} = \frac{0.4}{\sqrt{400}} \times 100 = 2\%$

23) The 3 histograms below (in scrambled order) are the probability histograms for the sum of **25, 50 and 150** random draws with replacement from a box that has 10 tickets 1 marked "0" and 9 marked "1". Which histogram depicts **25** draws, which **50** draws and which **150**? ***Fill in each blank below with the correct number of draws.***

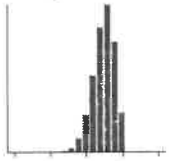
Histogram A



Histogram B



Histogram C



a) Histogram A is the sum of 50 draws **b)** Histogram B is the sum of 150 draws **c)** Histogram C is the sum of 25 draws

lopsided box

1 9

[0] [1]

10 tickets

* the more draws ⇒ the more normal the histogram will look

24) 64 draws are made at random with replacement from the box containing 5 tickets: 2 4 4 10 avg = 5

a) The ***smallest*** the sum of the 64 draws could possibly be is 128 (all 2's) and the ***largest*** is 640 (all 10's).
(Fill in the 2 blanks above with the correct numbers.)

b) What is the **EV** (expected value) of the **sum** of the **64** draws? (Show work, circle answer.)

EVsum = n × avg of box = 64 × 5 = (320)

c) What is the **SE** (Standard Error) **of the sum** of the **64** draws? Use the fact that the **SD of the box is 3.** (Show work, circle answer.)

SEsum = $\sqrt{n}$ × SD of box = $\sqrt{64}$ × 3 = (24)

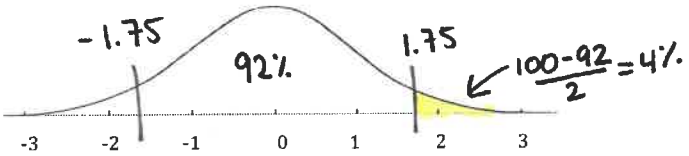
d) Use the normal approximation and ***your answers from (b) and (c) above*** to figure the **chance** that the sum of the 64 draws will be more than 362?

i) First calculate the Z score. ***Show work. Circle answer.***

$$Z = \frac{val - EV}{SE} = \frac{362 - 320}{24} = 1.75$$

ii) Now mark the Z score accurately and ***shade the correct area on the curve*** below. ***Round the middle area on the curve to the nearest whole number.***

Chance = 4 %



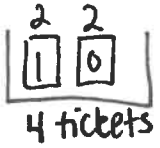
e) What is the **EV** of the **average** of the **64** draws? 5 (no work is necessary) EVavg = avg of box

f) What is the **SE** of the **average** of the **64** draws? 0.375 (Show work.)

$$SE_{avg} = \frac{SD}{\sqrt{n}} = \frac{3}{\sqrt{64}} = 0.375$$

g) Now suppose you draw at random with replacement from the same box above, but this time you're only interested in counting how many 4's you get. sums ⇒ doesn't say avg or %.

What is the **EV** and the **SE** of the ***number*** of 4's in **64** draws? ***(Hint: draw a new box)***



SD = 0.5

i. EV of the ***number*** of 4's in **64** draws = 32 ← EVsum = 64 × 0.5 = 32

ii. SE of the ***number*** of 4's in **64** draws = 4 ← SEsum = $\sqrt{64}$ × 0.5 = 4
(Show work by computing the SD of the new box, then use it to calculate the SE for the **sum** of the 100 draws.)

What is the **EV** and the **SE** of the ***percent*** of 4's in **64** draws?

iii. EV of the ***percent*** of 4's in **64** draws = 50%.
EV% = percent of 4's in box

iv. SE of the ***percent*** of 4's in **64** draws = 6.25%.

$$SE_{\%} = \frac{SD \text{ of box}}{\sqrt{n}} \times 100 = \frac{0.5}{\sqrt{64}} \times 100 = 6.25$$

Question 25 pertains to tossing a fair coin and counting the number of heads:

i) The appropriate box model has

a) Two tickets: 1 and 0
b) Two tickets: 1 and -1
c) Thousands of tickets marked with 1's and 0's. The exact percentage of each is unknown and estimated from the sample.
d) A box model is not appropriate for this situation.

sum for counting

ii) If you toss the coin 100 times you'd expect to 50 heads ± ____heads. Fill in the blank with the correct SE.

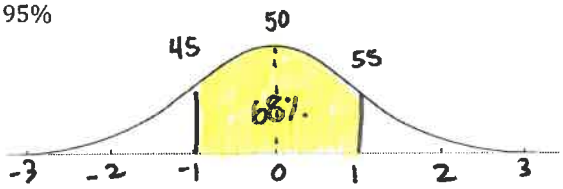
a) 2 b) 2.5 c) 5 d) 10 e) 20 $SE_{sum} = \sqrt{100} \times .5 = 5$

iii) What's the chance you'd get within 5 of 50 (between 45-55 heads)?

a) 34% b) 38% c) 68% d)95%

$Z = \frac{45-50}{5} = -1$

$Z = \frac{55-50}{5} = 1$



iv) If you toss a coin 400 times, you'd expect to get 200 heads ± ____heads? Fill in the blank with the correct SE.

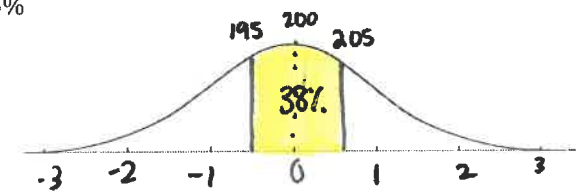
a) 2 b) 2.5 c) 5 d) 10 e) 20 $SE_{sum} = \sqrt{400} \times 0.5 = 10$

v) What's the chance you'd get within 5 heads of 200? (between 195-205 heads)

a) 34% b) 38% c) 68% d)95%

$Z = \frac{195-200}{10} = -0.5$

$Z = \frac{205-200}{10} = 0.5$



vi) If you toss the coin 100 times you'd expect 50% heads, give or take____%. Fill in the blank with the correct SE.

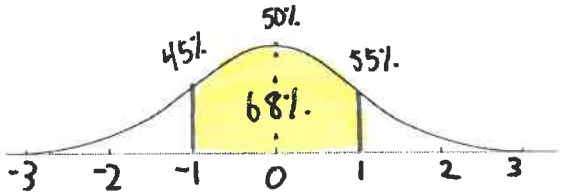
a) 2 b) 2.5 c) 5 d) 10 e) 20 $SE_{\%} = \frac{0.5}{\sqrt{100}} \times 100 = 5\%$

vii) What's the chance you'd get 45%-55% heads in 100 tosses?

a) 34% b) 38% c) 68% d)95%

$Z = \frac{45-50}{5} = -1$

$Z = \frac{55-50}{5} = 1$



viii) If you toss a coin 400 times, you'd expect to get 50%, give or take ____%. Fill in the blank with the correct SE.

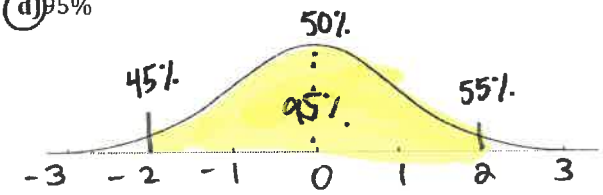
a) 2 b) 2.5 c) 5 d) 10 e) 20 $SE_{\%} = \frac{0.5}{\sqrt{400}} \times 100 = 2.5\%$

ix) What's the chance you'd get between 45%-55% heads in 400 tosses?

a) 34% b) 38% c) 68% d)95%

$Z = \frac{45-50}{2.5} = -2$

$Z = \frac{55-50}{2.5} = 2$



26) A slacker student has 4 Finals. Each Final consists of 100 multiple-choice questions. He knows nothing so he decides to randomly guess on every question so he can complete each Final in less than 5 minutes.

$EV_{sum} = n \times$ avg of box

i) To compute the Expected Value (EV) for the student's score for each Final, you may need additional information. Which of the following do you need to know? **Circle** "Yes" if needed or "No" if not.

a) How many students are taking each final. ***Circle one:*** Yes (No)

b) How many choices there are for each question. ***Circle one:*** (Yes) No

c) How many points are awarded or deducted for each choice. ***Circle one:*** (Yes) No

d) How much time is allotted for the exam. ***Circle one:*** Yes (No)

ii) Randomly guessing on all 100 questions corresponds to drawing 100 times with replacement from the appropriate box model. **(Fill in the first blank with a number and the second with either "with" or "without".)**

iii) For **a-d** match the Final exams to their corresponding box models ***Use each box model exactly once.***

Box I: [0] [1] [2] Box II: [-1] [0] [1] Box III: [0] [1] Box IV: [-1] [1]

a) Final A- Each question has 3 choices, one is a right answer, one is a wrong answer and one is an " I don't know" answer. Your score is computed as the number of right answers minus the number of wrong answers. The "I don't know" answers are scored as 0 points. This corresponds to Box... **i)** I (**ii)** II) **iii)** III **iv)** IV

b) Final B- Each question has 3 choices, one is the best answer and awarded 2 pts, one is a mediocre answer and awarded 1 pt. and one is a wrong answer and awarded no points. This corresponds to Box... (**i)** I) **ii)** II **iii)** III **iv)** IV

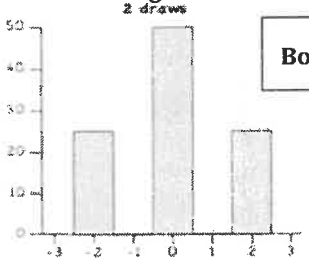
c) Final C--Each question is a true/false question. Your score is the number of answers you get right. This corresponds to Box... **i)** I **ii)** II (**iii)** III) **iv)** IV

d) Final D-Each question is a true/false question. Your score is the number of answers you get right minus the number of answers you get wrong. This corresponds to Box... **i)** I **ii)** II **iii)** III (**iv)** IV)

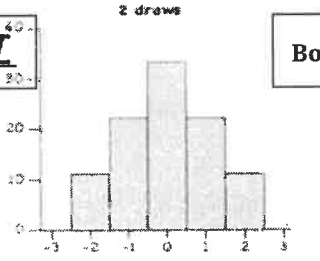
Parts iv and v are from Chapter 9 so you might want to wait until Monday's lecture to try them.

iv) The 4 histograms below represent the probability histogram for the **sum of 2** draws made at random with replacement from each of the boxes in part (iii) above. For each histogram identify the appropriate Box (I, II, III or IV). ***Use each box model exactly once.***

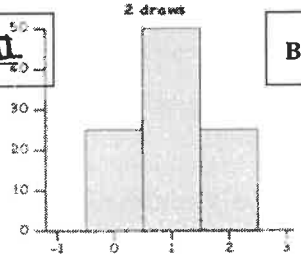
Histogram A



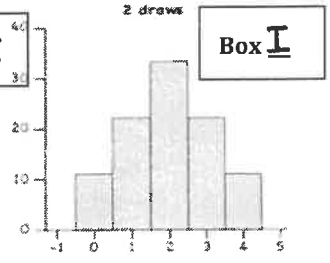
Histogram B



Histogram C



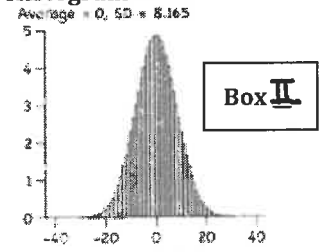
Histogram D



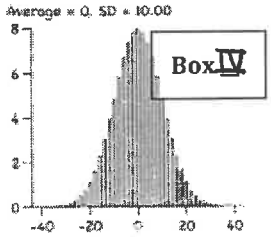
HINT— List all the possibilities for each box. For example, Box I has 9 possibilities: 0,1,1,2,2,2,3,3,4

v) The 4 histograms below represent the probability histogram for the **sum of 100** draws made at random with replacement from each of the boxes in part (iii) above. For each histogram identify the appropriate Box (I, II, III or IV). ***Use each box model exactly once.***

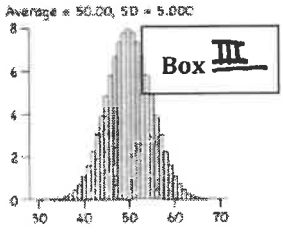
Histogram A



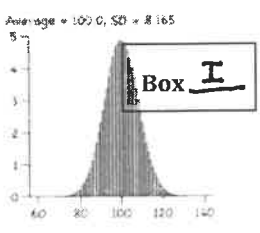
Histogram B



Histogram C

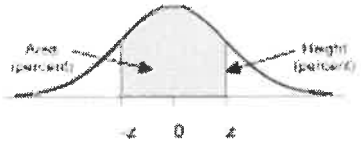


Histogram D



HINT—The Average and the SD given above each histogram is the EV and the SE of the sum of 100 draws.

STANDARD NORMAL TABLE



Standard Units

z	*Area*	z	*Area*	z	*Area*
0.00	0.00	1.50	86.64	3.00	99.730
0.05	3.99	1.55	87.89	3.05	99.771
0.10	7.97	1.60	89.04	3.10	99.806
0.15	11.92	1.65	90.11	3.15	99.837
0.20	15.85	1.70	91.09	3.20	99.863
0.25	19.74	1.75	91.99	3.25	99.885
0.30	23.58	1.80	92.81	3.30	99.903
0.35	27.37	1.85	93.57	3.35	99.919
0.40	31.08	1.90	94.26	3.40	99.933
0.45	34.73	1.95	94.88	3.45	99.944
0.50	38.29	2.00	95.45	3.50	99.953
0.55	41.77	2.05	95.96	3.55	99.961
0.60	45.15	2.10	96.43	3.60	99.968
0.65	48.43	2.15	96.84	3.65	99.974
0.70	51.61	2.20	97.22	3.70	99.978
0.75	54.67	2.25	97.56	3.75	99.982
0.80	57.63	2.30	97.86	3.80	99.986
0.85	60.47	2.35	98.12	3.85	99.988
0.90	63.19	2.40	98.36	3.90	99.990
0.95	65.79	2.45	98.57	3.95	99.992
1.00	68.27	2.50	98.76	4.00	99.9937
1.05	70.63	2.55	98.92	4.05	99.9949
1.10	72.87	2.60	99.07	4.10	99.9959
1.15	74.99	2.65	99.20	4.15	99.9967
1.20	76.99	2.70	99.31	4.20	99.9973
1.25	78.87	2.75	99.40	4.25	99.9979
1.30	80.64	2.80	99.49	4.30	99.9983
1.35	82.30	2.85	99.56	4.35	99.9986
1.40	83.85	2.90	99.63	4.40	99.9989
1.45	85.29	2.95	99.68	4.45	99.9991