Warning
This practice final does not have box plot questions. There WILL be some on the final! Please study accordingly. You can practice box plots by doing the practice homework on box plots. There are also box plot questions on all exam 2 practice exams.

Stat 100 Final Exam **Form is either A B C D E F G H I** (you don't know) Spring 2012

PRINT Last Name FINAL EXAM KEY by Jackie Caperon **PRINT** First Name _____

Signature _____ Net ID _____

Circle section: F1 (9:30 am) R1 (12:30 pm)

Instructions- This is a closed book, closed notes exam. You have 3 hours to complete it.

- **Print and sign your name above, then fill in your Net ID, then circle your section.**
- **At the end of this exam, you must return this Exam Booklet complete with all pages, and you must put your Scantron inside the booklet. You don't need to show any work on the exam booklet.**
- **If you do not turn in a complete Exam Booklet, with your Answer Sheet inside you will receive the grade AB (Absent) for this exam.**
- Use a **#2 pencil. Each question has only *one* answer.** If you bubble in more than one answer it will automatically be marked wrong. Erase mistakes completely.
- This Exam Booklet is **either Form A, B, C, D, E, F, G, H or I.** You don't know which form you have so you MUST put your Scantron form inside the exam booklet so the TA's can correctly mark your Scantron form after the exam.
- Print and bubble in your NET ID in the NETWORK ID box. This is IMPORTANT, you may lose points if your netid is wrong (e.g. net id's should have no spaces; kim 87 is WRONG, kim87 is correct).
- Print and bubble in your Student ID number in the Student Number box.
- Print and bubble in 00001 for F1 section, 00002 for R1 in the Section Box.
- Print and bubble in the date in the Date box
- Print and bubble in your **LAST NAME with NO SPACES** starting in the left most column. Print your **FIRST INITIAL** in the right-most column.
- Write Stat 100 on the COURSE line.
- Write Fireman on the INSTRUCTOR line.
- Write either F1 (9:30am) or R1 (12:30pm) on the SECTION line.
- **Sign your name, and right underneath PRINT your name on the STUDENT signature line.**

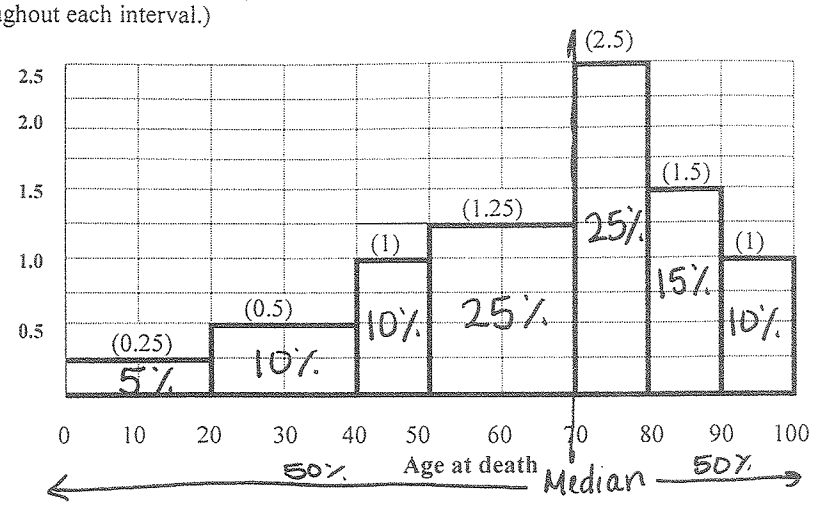
Final Exam Scores will be posted on Compass on Friday. Bonus Notebook points will be posted on Compass by the end of this Final. Check Compass to make sure your points were recorded.

CHECK NOW THAT YOU HAVE COMPLETED ALL OF THE STEPS. Before starting work, check to make sure that your test booklet is complete. You should have **15 pages (100 problems)**, including **3 tables**, the Normal, the *t*-Table and the chi-square table.

Questions 1-8 pertain to the histogram below.

The histogram below represents the age at death of a large population. The height of each block is given in parentheses. (Assume an even distribution throughout each interval.)

Height × Width = Area



1) What percent of the population died in the 70-80 interval ? a) 10% b) 15% c) 20% (d) 25% e) 30% $10 \times 2.5 = 25\%$

2) The median is closest to … a) 40 b) 50 c) 60 (d) 70 e) 80

3) The median is _____ the average. a) less than (b) greater than c) equal to d) cannot be determined

4) The 25th percentile is a) 20 b) 25 c) 30 d) 40 (e) 50 $5\% + 10\% + 10\% = 25\%$ (1st block, 2nd, 3rd) ↳ at 50

5) The percent of the population who died at 75 years is closest to….. a) 1% b) 1.25% c) 1.5% d) 2% (e) 2.5%
↳ Look at height of 75 block!

6) If everyone lived 1 more year, the average would ….
(a) Increase by 1 year. b) Increase by 0.01 years c) Increase 1 5% d) Stay the Same e)Decrease
Every number on the list +1 → Average +1

7) and the SD would ….
a) Increase by 1 year. b) Increase by 0.01 years c) Increase 1 5% (d) Stay the Same e)Decrease
Histogram shifts to the right by 1 year - deviations from ave. stay the same!

8) If you knew the average and SD of the ages displayed in the histogram above, would it be appropriate to use the normal approximation to figure what percentage of the ages fell within various intervals?
a) Yes, because we know that the histogram represents the age at death of a large population.
b) Yes, because the ages at death range from 0 to 100.
(c) No, because the histogram of the ages is not close enough to following the normal curve; it has a long left-hand tail.
d) Maybe, depending on whether the ages were randomly drawn from a larger population.

The next 3 questions pertain to this list of 4 numbers: 2 , -2 , 0, 8

$\frac{2 + (-2) + 0 + 8}{4} = 2$

9) The average of the list is … a) 0 b) 1 (c) 2 d) 3 e) 4

10) The median of the list is… a) 0 (b) 1 c) 2 d) 3 e) 4

-2 0 | 2 8
Median = $\frac{0+2}{2} = 1$

11) The deviations from the average of the list are:
a) -2, 2, 0, -8 b) 0, 0, 2, 6 c) -2, -6, -4, 4 d) 1, -3, -1, 7 (e) 0, -4, -2, 6
Subtract the average (2) from each number on the list.

Question 12

A list has 6 numbers. Five of the deviations from the average are: 1, 2, 3, 4, 5. What is the fifth deviation from the average?
a) 0 b) 6 c) -5 (d) -15 e) cannot be determined

All deviations must add to 0!
1+2+3+4+5 = 15
So the missing number must be -15.

Questions 13-17 pertain to the following study:

A study published in the journal *Autism Research* found that older mothers are more likely to have autistic children than younger mothers. The study examined nearly 5 million birth records from the 1990's in California and found 12,159 cases of autism. The median maternal age was 30 for the autistic group and 27 for all other births. Mothers over 40 were 77% more likely to have given birth to an autistic child than mothers under 25 were.

13) Based only on the information above, this study is an example of .... *Choose one:*

(a) Observational Study
b) Randomized Controlled Experiment without a placebo
c) Randomized Double-Blind Controlled Experiment
d) Non-Randomized Controlled Experiment

Researchers observed Autism rates - They did not apply any treatment.

14) Which of the following statements is best? *Choose one:*

a) This study is strong evidence that maternal age *causes* autism.
(b) This study only shows that maternal age is *associated with* autism: it doesn't show whether or not maternal age *causes* autism. Cannot prove causation through observations.
c) This study shows that maternal age is *associated with but definitely does not* cause autism.

Below are either **confounders** that *mix up* the study described above making it look like maternal age is responsible for autism when it really isn't, **causation links** that explain *how* maternal aging *causes* autistic births, or neither.

15) Heredity—Mothers who have autistic tendencies themselves are both more likely to have autistic kids and more likely to be older when they give birth (since they lack social skills.)

Choose one: (a) Confounder b) Causation Link c) Neither

Heritable Autistic Social Tendencies → Autistic children; Older when they get pregnant

16) Damaged DNA—As people age their DNA deteriorates which makes them more likely to have babies with birth defects including autism.

Choose one: a) Confounder (b) Causation Link c) Neither

Older Mothers → DNA damaged with Age → Autistic Children

17) Poor Diet—Poor nutrition can damage neurological development.

Choose one: a) Confounder b) Causation Link (c) Neither

Poor Diet in Mothers → Autistic Children ...but what does this have to do with AGE?

The next 3 questions pertain to this study:

A study was done to test the effectiveness of an allergy medicine designed to reduce allergic reactions to dust. The subjects were 1000 adult volunteers with allergies to dust. Half were *randomly* injected with the medicine and half with salt water. A week later the subjects were all exposed to the same levels of dust and their allergic reactions were measured. *Neither the subjects nor those who evaluated them knew who was in which group.* Those who received the medicine had significantly milder allergic reactions than those who received the salt water.

→ Placebo! → Double Blind!

18) Which of the following statements is best? *Choose one:*

a) This was a randomized controlled experiment without a placebo.
b) This was an observational study.
c) This was a non-randomized controlled experiment with a placebo.
(d) This was a randomized controlled double-blind experiment.

19) Which of the following statements is best? *Choose one:*

(a) This study is very strong evidence that the medicine reduces allergic reactions to dust.
b) This study has cause and effect reversed. It's more likely that the dust exposure protected people from the effects of the medicine than that the medicine protected people from the dust exposure.
c) This study only shows that there is an ***association*** between the medicine and fewer dust allergies. It does not show that taking the medicine actually caused fewer allergies. A likely confounder is pre-existing health problems.

Randomized Experiment eliminates confounders. This is a GREAT study!

20) Which of the following are likely to confound the results? *Choose one:*

a) Age- older people are more susceptible to allergic reactions.
b) Weak Immune systems-people with poor immune responses are more susceptible to allergies.
c) Inflated Sense of Protection —Those who got the medicine might have felt protected and so had fewer allergic responses.
d) All of the above are likely confounders.
(e) None of the above are likely confounders.

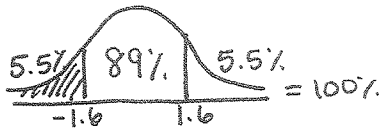
Because this is a randomized experiment, there should not be any differences between the Medicine Group and the Placebo Group. Randomization eliminates confounders & the Double Blindness eliminates Human Bias in the patient evaluations. A++ Experiment!!!

The next 3 questions pertain to the following:
According to our survey data Stat 100 students have mothers whose ages are **normally distributed** with an average = 48 and a SD = 5. **(Use the normal table at the end of this exam to answer these questions.)**

21) About what percentage of the students have mothers younger than 40 ? Value → Z-Score → Middle Area → Percentages

(a) 5.5%
b) 15%
c) 20%
d) 89%

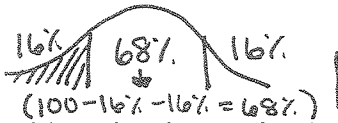
$Z = \frac{V - A}{SD}$ $Z = \frac{40-48}{5} = -1.6$ → Use Table to find Middle Area!



22) What mother's age corresponds to the 16th percentile? (16th percentile means older than 16% of the mothers.)

a) 40
b) 41
c) 42
(d) 43
e) 44

Percentile → Mid. Area → Z-Score → Value



68% Middle Area → Z = -1

Value = ave − Z(SD) = 48 − 1(5) = 43

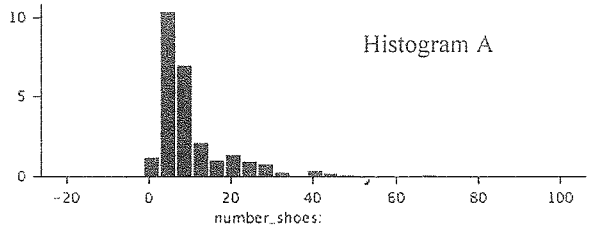
23) About 95% of the students have mothers between _____ years and _____ years old.

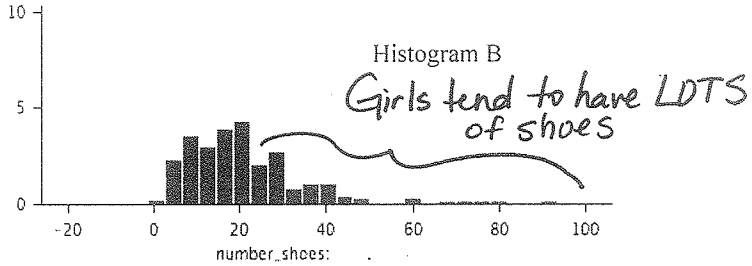
a) 36 and 60
(b) 38 and 58
c) 45 and 55
d) 43 and 53
e) 42 and 54

95% Middle Area → Z-scores −2 & 2

Value = 48 − 2(5) = 38 Value = 48 + 2(5) = 58

The next 2 questions pertain to the 2 histograms below that depict the male and female Stat 100 answers to the survey question: "How many pairs of shoes do you own?"





24) Which histogram represents the *female* answers? (You can assume females generally have more shoes than males)

Circle one: a) Histogram A (b) Histogram B c) Not enough Information to determine

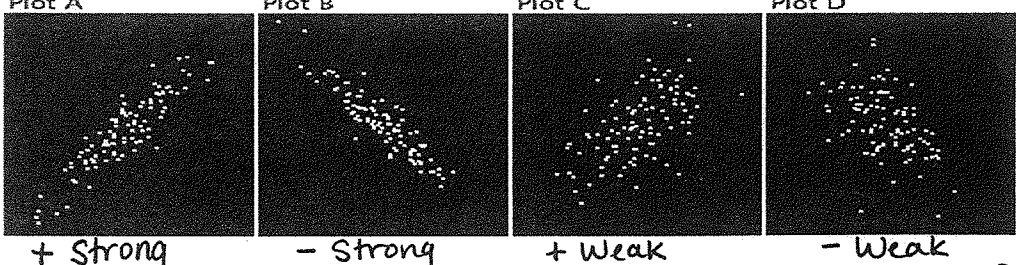
25) The following 4 numbers (in no particular order) are the averages and medians of the 2 histograms: **7, 19, 10.152, 20.168.**
Which number is the median of Histogram A?

High density of low #'s on Histogram A

(a) 7 b) 19 c) 10.152 d) 20.160

Whole #'s are most likely Medians — These long decimals are averages

For the next 4 questions match the scatter plots with their corresponding correlation coefficients



+ Strong − Strong + Weak − Weak

26) Correlation coefficient = -0.49	a) Plot A	b) Plot B	c) Plot C	(d) Plot D
27) Correlation coefficient = -0.92	a) Plot A	(b) Plot B	c) Plot C	d) Plot D
28) Correlation coefficient = 0.59	a) Plot A	b) Plot B	(c) Plot C	d) Plot D
29) Correlation coefficient = 0.91	(a) Plot A	b) Plot B	c) Plot C	d) Plot D

Questions 30-32

Suppose the scores on a test measuring artistic skills and a test measuring athletic skills follow the normal curve but have different correlations in different populations. Consider 3 populations where the correlation coefficient between artistic scores and athletic skills scores are as given in the table below. If someone scores in the 10^{th} percentile on the artistic skills test, estimate his or her percentile on the athletic skills test.

Percentile on Artistic Skills Test	r	Estimated Percentile on Athletic Skills Test
30) 10^{th} Random! Go with 50th!	0	a) 10^{th} b) 26^{th} (c) 50^{th} d) 74^{th} e) 90^{th}
31) 10^{th} Positive correlation - not perfect	0.5	a) 10^{th} (b) 26^{th} c) 50^{th} d) 74^{th} e) 90^{th}
32) 10^{th} Exact Opposite!	-1	a) 10^{th} b) 26^{th} c) 50^{th} d) 74^{th} (e) 90^{th}

r=+1 r≈0.5 r=0 r≈-0.5 r=-1

Questions 33-38

In a certain class of 500 students, scores on the midterm and final have the following 5 summary statistics:

	Average	SD
Midterm	80	10
Final	70	15
Correlation	r = 0.6	

Midterm | Z_1 | $\times r$ | Z_2 | Final

60 $\frac{60-80}{10} = -2 \times .6 = -1.2 \Rightarrow 70 - 1.2(15) = 52$

90 $\frac{90-80}{10} = 1 \times .6 = 0.6 \Rightarrow 70 - 0.6(15) = 79$

In the table below you're given the midterm scores of 2 students. For each student circle the regression estimate for the Final.

Midterm Score (Hint: change to Z score)	r	Regression Estimate for the Final
33) *60*	.6	a) 40 (b) *52* c) *55* d) *60* e) *65*
34) *90*	.6	a) *76* (b) *79* c) *85* d) *88* e) *90*

35) What is the slope of the regression equation for predicting Finals (Y) from midterms (X)?

$m = \frac{SD_Y}{SD_X} \times r = \frac{15}{10} \times 0.6 = 0.9$

a) 0.4 b) 0.6 c) .67 (d) 0.9 e) 1.5

(Remember – whatever we are Predicting is Y!)

36) The regression equation for predicting Midterm scores from Finals is: Midterms= 0.4 (Final Score) + _____. Fill in the blank with the correct y-intercept. Plug in averages!

$80 = 0.4(70) + b$

$b = 52$

a) -25 b) 4 c) 38 d) -50 (e) 52

37) The SD of the prediction errors when predicting Finals (Y) from midterms is …

$SD_{error} = \sqrt{1-r^2} \times SD_Y$

$= \sqrt{1-(.6)^2} \times 15 = 12$

a) 8 b) 10 (c) 12 d) 15 e) 18

38) If 2 points were added to everyone's Final Exam score the correlation coefficient would …

a) increase b) decrease (c) stay the same d) cannot be determined

Questions 39-40

39) The regression line is the same as the SD line when…

$m_{SD\,line} = \frac{SD_Y}{SD_X}$ $m_{reg\,line} = \frac{SD_Y}{SD_X} \cdot r$

a) The correlation is 0
(b) The correlation is 1
c) The average and SD of both variables are the same

40) The regression line is a horizontal line through the average of Y when…

(a) The correlation is 0
b) The correlation is 1
c) The average and SD of both variables are the same

The next 6 questions pertain to randomly drawing from the box containing the 6 tickets below.

41) Two tickets are drawn at random **with** replacement. What is the chance that the first ticket is a square and the second is a circle?
a) 2/6 x 2/6 b) 2/6 x 4/5 c) 2/6 x 2/5 (d) 2/6 x 4/6 e) 2/6 + 4/6 $\frac{2}{6} \times \frac{4}{6}$

42) Two tickets are drawn at random **without** replacement. What is the chance that the first ticket is a square and the second is a circle?
a) 2/6 x 2/6 (b) 2/6 x 4/5 c) 2/6 x 2/5 d) 2/6 x 4/6 e) 2/6 + 4/6 $\frac{2}{6} \times \frac{4}{5}$

43) Four tickets are drawn at random with replacement. What is the chance of getting at least one square ticket?
(a) $1 - (4/6)^4$ b) $(4/6)^4$ c) $1 - (2/6)^4$ d) $(2/6)^4$ e) 2/6 At least 1 □ = 1 − No □ $1 - \left(\frac{4}{6}\right)^4$

44) One ticket is randomly drawn. What is the chance of getting either a square ticket or a ticket marked "1"?
a) 2/6 b) 3/6 (c) 4/6 d) 5/6 e) 1 $P(\square) + P(1) - P(\square 1) = \frac{2}{6} + \frac{3}{6} - \frac{1}{6} = \frac{4}{6}$

45) What's the chance of getting a "2" if you draw only from the circular tickets?
a) 1/6 (b) 1/4 c) 1/3 d) 1/2 e) 2/3 $\frac{1}{4}$

46) What's the chance of getting a circular ticket if you draw only from the tickets marked "2"?
a) 1/6 b) 1/4 c) 1/3 (d) 1/2 e) 2/3 $\frac{1}{2}$

The next 3 questions refer to the following screening test for bus drivers:
Bus drivers are given random drug tests. If they test positive for drugs, they fail the test and face losing their jobs. Suppose only 1% of drivers who get tested for drugs are really using drugs. If a driver is **using drugs**, then **95%** of the time he'll correctly fail the test, but **10%** of the drivers ***not*** **using drugs** will also (incorrectly) fail the test. The table below gives the results for 10,000 people.

	Fail Test	Pass Test	Total
Drug Users	.95(100) = 95	+ 5 =	100
Not Drug Users	.10(9900) = 990	+ 8910 =	9900
Total	1085	8915	10,000

Fill in the four missing cells in the table to answer the following 2 questions:

47) What fraction of those who fail the test are not drug users?
a) 5/100 b) 990/9900 c) 10/100 d) 5/8915 (e) 990/1085 $\frac{990 \text{ NonDrug Fail}}{1085 \text{ Fail Total}}$

48) What fraction of those who pass the test are drug users?
a) 5/100 b) 95/10,000 c) 10/100 (d) 5/8915 e) 990/1085 $\frac{5 \text{ Drug Pass}}{8915 \text{ Pass Total}}$

The next 2 questions pertain to tossing a fair coin repeatedly.

49) Which of the following is more likely?
(a) Getting exactly 1 head in 2 tosses
b) Getting exactly 1000 heads in 2000 tosses
c) Both of the above are equally likely because they're both exactly half heads.

Easy to see "remarkable" result in short term, like getting EXACTLY half. Things get more normal with more draws.

50) Which of the following is more likely?
a) Getting between 45%-55% heads in 100 tosses
(b) Getting between 45%-55% heads in 1000 tosses
c) Both of the above are equally likely.

Things move closer to "normal" with more tosses.

Think about SE!

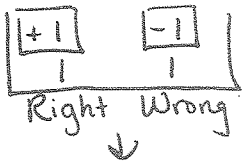
$\downarrow SE_{\%} = \frac{SD}{\uparrow\sqrt{n}} \times 100$ means less error with more tosses! $\uparrow SE_{sum} = SD\sqrt{n}\uparrow$ means more error with more tosses!

The next 4 questions pertain to the following situation:
A 100-question true/false test awards 1 point for each correct answer and **subtracts 1 point** for each incorrect answer.

+1 −1

51) Suppose a student guesses at random on each question and his score is computed, what is the corresponding box model?
- a) It has two tickets: 1 and 0
- b) It has 100 tickets: half 1's and half 0's
- c) It has three tickets: 1, 0, -1
- d) It has two tickets: 1, -1
- e) It has hundreds of tickets marked with 1's and 0's, but the exact percentages are unknown.

Draw box based on One Question:



52) How many draws do we make from the box above? – AKA – how many questions on the test?
a) 2 b) 20 c) 50 (d) 100

ave of box = $\frac{1(1)+1(-1)}{2} = 0$

53) The expected value for the student's score is EVsum = n × ave
(a) 0 b) 10 c) 20 d) 30 e) 40

EV = 100 × 0 = 0

54) The standard error of the student's score is …(Hint: First calculate the SD of the box using the short-cut formula)
a) .1 b) 5 (c) 10 d) 15 e) 20

SEsum = SD√n = 1√100 = 10

SD: $1-(-1)\sqrt{\frac{1}{2}\times\frac{1}{2}} = 1$

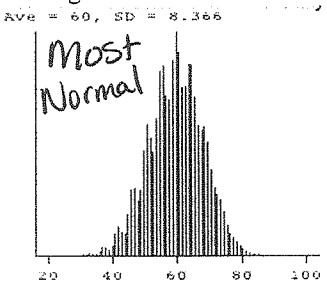
55) Now suppose you're just interested in how many correct answers the student would get by guessing, not his score. Then the **EV = 50 and the SE = 5.** Suppose the student needs to get 60 answers correct in order to pass. What's the probability the student will pass if he guesses on all the questions? (Hint: convert to a Z score, and use the normal curve. Round percents given in the table to the nearest whole number). Value → Z-Score → Mid Area → Percentage

a) 1% (b) 2.5% c) 5% d) 16% e) 32%

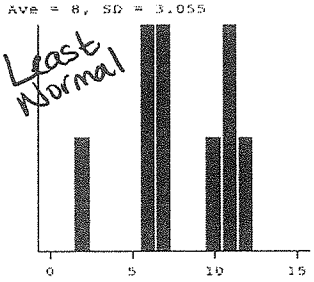
1) $\frac{60-50}{5} = 2$ 2) Mid Area for Z=2 is 95% 3) 2.5% | 95% | 2.5%

Probability of getting a 60 or higher by guessing

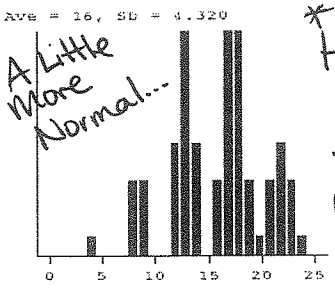
The next 4 questions pertain to the following:
The histograms below (in scrambled order) depict the sums of 2, 4, and 15 draws from the same box. Match the number of draws to the histogram. **Each answer will be used only once.**



Histogram 1



Histogram 2



Histogram 3

* Histograms look more & more like the normal curve with more draws.

56) Histogram 1 is the probability histogram for how many draws? ….
a) 2 b) 4 (c) 15

Most normal → Most draws

57) Histogram 2 is the probability histogram for how many draws? ….
(a) 2 b) 4 c) 15

Least normal → Least # of draws

58) Histogram 3 is the probability histogram for how many draws? ….
a) 2 (b) 4 c) 15

59) The 3 histograms above represent the sum of 2, 4, and 15 draws from the same box. Which of these boxes is it?

Box A	Box B	Box C
a) [1] [5]	(b) [1] [5] [6]	c) [0] [1] [5]
ave: 3	ave: 4	ave: 2

There are several ways to attack this problem…
Look at Histogram A
The EV is 60 with 15 draws. Only a box with an ave of 4 could satisfy this.
EV = n × ave
60 = 15 × 4

The next 5 questions pertain to the following poll:

A recent CNN poll asked a random sample of 900 adults nationwide the following question: **"Would you approve of raising the national debt ceiling at this time?"** 48% answered "Yes" The same question was also recently asked on an open poll via the Fox News website where anyone who wishes to do so can vote on the poll. About 90,000 people voted on the site and 40% answered "Yes".

*CNN: Random ☺ * Fox: Self-Selected ☹

60) Which poll better represents the US population's attitude toward raising the national debt ceiling? → Random!!!

a) CNN Poll, because the Fox poll has too many responders and could artificially make results significant.
(b) CNN Poll, because it used a randomized sample whereas Fox did not.
c) Fox Poll, because it has 100 times more responders than the CNN poll.
d) The two polls will have about the same degree of accuracy because the advantages and disadvantages of each will balance out. The advantage of large size is offset by the disadvantage of selection bias for one poll while the advantage of random selection is offset by the disadvantage of small size for the other.

61) What is the SE of the percentage of YES's in the Fox Poll? Need randomization to calculate SE.

a) $\sqrt{\frac{.40\times.60}{90{,}000}}\times 100\%$ b) $\sqrt{.40\times.60}\times 100\%$ c) $\sqrt{90{,}000}\times\sqrt{.40\times.60}\%$ (d) Not possible to compute a SE

Questions 62-64 refer to the CNN poll:

62) What is the SE of the percentage of YES's in the CNN Poll? $SE_{\%} = \frac{SD}{\sqrt{n}}\times 100 \leq \frac{(1-0)\sqrt{.48\times.52}}{\sqrt{900}}\times 100$

(a) $\sqrt{\frac{.48\times.52}{900}}\times 100\%$ b) $\sqrt{.48\times.52}\times 100\%$ c) $\sqrt{900}\times\sqrt{.48\times.52}\%$ d) Not possible to compute a SE

$= 1.67\%$

63) A 95% confidence interval for the % of all American adults who would answer "Yes" to this question is about 48 +/- 2SE

48 +/- 2(1.67) =

a) 46.3%-49.7% (b) 44.7%-51.3 c) 33%-63% d) Not possible to compute a confidence interval

64) A 95% confidence interval for the % of all American *democrats* who would answer "Yes" to this question is about

CNN Randomly asked USA adults – "democrats" is too specific!

a) 46.3%-49.7% b) 44.7%-51.3 c) 33%-63% (d) Not possible to compute a confidence interval

This data only applies to USA adults – nothing more, nothing less.

The next 4 questions pertain to the following poll:

A USA Today poll asked a nation-wide random sample of 1000 adults the question: "In addition to scholarships, do you think college athletes should be paid for playing on college teams, or not?" 24% answered "Yes" and 76% answered "No"

65) What most closely resembles the relevant box model?

a) It has 1000 tickets, 24% are marked "1" and 76% are marked "0"
b) It has 1000 tickets marked 1 and 0, but the exact amounts are unknown.
c) It has millions of tickets. 24% are marked "1" and 84% are marked "0".
(d) It has millions of tickets marked 1 and 0, but the exact amounts are unknown and estimated from our sample.

* Every USA adult is in the box – we can only estimate the opinion of American adults.

66) Which one of the statements below is true?

a) The expected value for the % of all college athletes who would answer "Yes" to the question is 24%. Too Specific!
(b) The expected value for the % of all US adults who would answer "Yes" to the question is 24%. Just right!!
c) The expected value for the % of all US college students would answer Yes" to the question is 24%. Too specific!
d) All of the above are true.
e) None of the above are true.

Survey asked USA adults so data only applies to USA adults.

67) Is it possible to compute a 95% confidence interval for the percent of all US adults who would answer "Yes" to the question?

a) Yes, a 95% confidence interval is approximately 24% +/- 8%
(b) Yes, a 95% confidence interval is approximately 24% +/- 3%
c) No, because we're not given the SD of the sample.
d) No. The sample and population don't match. One has only 900; The other has millions.

Yes! Randomized data!

24 +/- 2SE%

$SE_{\%} = \frac{\sqrt{.24\times.76}}{\sqrt{1000}}\times 100$

68) If the sample size was multiplied by 4 (from 1000 to 4000) then the SE of the sample percent and the width of the confidence interval would be

a) multiplied by 4 b) multiplied by 2 c) divided by 4 (d) divided by 2 e) Not changed

$SE_{\%} = \frac{SD}{\sqrt{n}}\times 100$

As n increases by 4, $\sqrt{n}$ increases by 2. When $\sqrt{n}$ increases by 2 $SE_{\%}$ decreases by 2 (since $\sqrt{n}$ is in denominator)

Question 69

A research wants to conduct a pre-election poll (using random sampling) in each state. He polled 1000 people in Maine to estimate the percentage of people in Maine who favor Obama over Romney. Now he wants to conduct the same poll in California where the population is 25 times larger. How many people does he need to poll in California to keep about the same level of accuracy as the Maine poll?

a) 200 (b) 1000 c) 5000 d) 25,000 e) 625,000

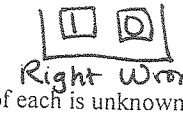
Population size does not matter! In order to compare polls, n needs to be the same.

The next 6 questions pertain to the following situation:

A mother thinks her 2 year-old child is particularly gifted in statistics. To test her claim I give the child a 64 question true-false exam about statistics. The child answers 39 of the 64 questions correctly. The null hypothesis is that the child is just guessing.

70) Which of the following most accurately describes the null box?

a) It has 64 tickets, 39 marked "1" and 25 marked "0"
b) It has 64 tickets marked either "1" or "0" but the exact percentage of each is unknown.
(c) It has 2 tickets, 1 marked "1" and 1 marked "0" → 2 options: the correct answer or incorrect answer on each question.



ave: .5
SD: .5

71) The draws are made _____ replacement. (a) with b) without

Probability of guessing correct answer are the same for each question.

Assuming the null hypothesis to be true, you would expect the child to answer _____questions correct, give or take_____ questions.

72) Fill in the first blank in the above sentence with the correct expected value.

a) 25 b) 30 (c) 32 d) 35 e) 38

We'd expect her to answer half of the 64 correctly.

EV = n × ave = 64 × .5

73) Fill in the second blank in the above sentence with the correct SE.

a) 3 (b) 4 c) 5 d) 6 e) 8

SEsum: $SD\sqrt{n}$

$.5\sqrt{64} = 4$

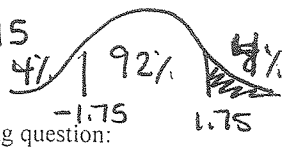
74) The z -statistic for testing the null hypothesis is

a) 7/SE for the average (b) 7/ SE for the sum c) 7/SE for percent d) 7/SD of the box

75) The P-value is closest to …

a) 1% b) 2% c) 3% (d) 4% e) 5%

$\frac{39-32}{4} = 1.75$



The next 5 questions pertain to the following survey:

Last August, Gallup asked a nation-wide random sample of 800 white adults and 400 black adults the following question:
"Do you think that relations between blacks and whites will always be a problem for the United States?"
55% of the blacks answered "Yes" and 44% of the whites answered "Yes".
The null hypothesis is that the 11% difference between black and white responses is just due to chance and doesn't reflect a real difference between blacks and whites in the population.

76) Which of the following most accurately describes the null box(es)?

a) There are 2 null boxes, one with 800 tickets marked with "0"s and "1"s and one with 400 tickets with 0's and 1's.
(b) There are 2 null boxes, each with millions of tickets, and each with the same percentage of "1"'s.
c) There are 2 null boxes, each with millions of tickets. One box has 55% "1"s, and the other has 44% "1"s.

77) The draws are made _____replacement. a) with (b) without

You never want to survey the same person twice!

78) Assuming the null to be true, the SE for the blacks' sample percentage is about 2.5% and the SE for the whites' sample percentage is about 1.75%. The SE for the difference of the 2 sample percentages is closest to .

$SE_{difference} = \sqrt{SE_A^2 + SE_B^2}$

a) 0.75% b) 1.75% c) 2.5% (d) 3.05% e) 4.25%

$= \sqrt{2.5^2 + 1.75^2} = 3.05$

79) The Z statistic for testing the null hypothesis is closest to …

a) .75 b) 1.5 c) 2.8 (d) 3.6 e) 11

$Z = \frac{55\% - 44\%}{3.05\%} = 3.6$

80) Suppose the P-value is about 0.016%, what do you conclude?

a) Cannot reject the null. It's plausible that there is no black/white difference on this question among US adults
(b) Reject the null and conclude that there is overwhelming evidence that our sample difference reflects a real black/white difference on this question among US adults.

This means there is a 0.016% chance this difference in responses was simply due to chance – HIGHLY unlikely! So Reject that Null!

The next 6 questions refer to the following situation:
Suppose the instructor of a large class with an enrollment of over a thousand students claims to always grade on a curve so that 20% of the students receive A's, 40% B's, 30% C's and 10% D's or F's. To test the claim I take a random sample of 50 students from the course . Here are the results:

Grade	Percents Claimed by Instructor	*Observed #*	*Expected #*	*Obs -Exp*	*(Obs-Exp)²*	$\frac{(Obs - Exp)^2}{Exp}$
A	20%	5	10	-5	25	25/10 = 2.5
B	40%	20	20	0	0	0
C	30%	20	15	5	25	25/15 = 1.67
D or F	10%	5	5	0	0	0
Total	100%	50	50	0	50	4.167

← always

81) To test the null hypothesis that our observed data fits the letter grade percentages claimed by the instructor we'd do ..
- a) the one-sample z test
- b) the two-sample z test
- (c) the chi-square test for "goodness -of-fit"
- d) the chi-square test for independence

How "good" does our data "fit" his claim?!

82) The table above is missing all 4 expected numbers, which of the following is the missing column?

(a)	b)	c)	d)
10	20	25	12.5
20	40	25	12.5
15	30	25	12.5
5	10	25	12.5

- .20(50) = 10
- .40(50) = 20
- .30(50) = 15
- .10(50) = 5 ← expected

↑ claim ↑ # of students

83) The value for C is missing in the *Obs –Exp* column, fill in the missing blank.
a) 0 (b) 5 c) 25 d) not enough information to determine

20 − 15 = 5 → Also, all Obs-Exp must add up to 0!

84) To compute the proper test statistic you'd have to sum the 4 values in the last column. The test statistic is closest to

a) 0 b) 2.5 (c) 4.17 d) 5 e) not enough information to determine

85) The number of degrees of freedom is
a) 2 (b) 3 c) 4 d) 5 e) 6

4 categories − 1 = 3

86) What is the p-value?
- a) < 1%
- b) between 1% and 5%
- c) between 5% and 10%
- (d) between 10% and 30%
- e) 30% and 50%

Use Chi-Squared Table

1. Go down the ~~row~~ column to 3 degrees of freedom

2. Go across the row to a χ^2 of 4.17

3. Go up the column to find the p value (between 10%-30%)

The next 6 questions pertain to our survey question on Occupy Wall Street:

The table below shows the survey responses of the 246 male and 471 female students from our class to the question: "What is your opinion of the Occupy Wall Street demonstrations?"

Here are the results in percentages:

	Favor	Oppose	Unsure	Never Heard of them
Male	26%	14.2%	32.9%	26.8%
Female	20.2%	4.7%	27.8%	47.3%

Here are the same results as frequencies (or counts):

	Favor	Oppose	Unsure	Never Heard of them	Total
Male	64	35	81	66	246
Female	95	22	131	223	471
Total	159	57	212	289	717

87) Which significance test should we use to test the null hypothesis that Stat 100 students opinions about the Occupy Wall Street Demonstrations are independent of whether they're male or female?

a) one sample z test b) 2 sample z test c) t-test d) chi-square test for goodness-of-fit (e) chi-square test for independence

88) To compute the chi-square test statistic, we need to calculate the sum of the (observed-expected)^2/expected? Should we use the observed percentages from the above table or the observed frequencies from the above table in that calculation?

a) We should use the percentages (b) We should use the frequencies c) We can use either

Always use the frequencies!

89) How many degrees of freedom are there for the chi-square independence test?

a) 1 b) 2 (c) 3 d) 4 e) 5 $(2-1)(4-1) = 3$

90) Assuming the null hypothesis is true, what is the expected number of males who would answer "Never Heard of them"?

a) $\frac{246\times159}{717}$ (b) $\frac{246\times289}{717}$ c) $\frac{246\times212}{717}$ d) $\frac{471\times246}{717}$ e) $\frac{471\times289}{717}$

$\frac{\#\text{ of Males} \times \text{"Never Heard"}}{\text{Total \# of People}}$

91) The test statistic is 39.36 . What do you conclude?

(a) Reject the null and conclude that there is very strong evidence that the distribution of male and female responses to this question is really different among typical Stat 100 students.

b) Cannot reject the null, it looks like the difference between male and female Stat 100 students is just due to the luck of the draw.

P value close to 0! HIGHLY unlikely due to chance.

92) If the survey question was changed to a Yes/No question, asking "Have you ever heard of the Occupy Wall Street demonstrations?", what significance test(s) could be used?

(a) Either a chi square test for independence or a 2 sample z test

b) Only a chi-square test for independence

c) Only a 2 sample z test

The next 4 questions pertain to the following situation:
Suppose a doctor claims that the average body temperature for healthy adults is 98.6 degrees, but I think it's really less than that. To test the doctor's claim I randomly sample 9 healthy adults and find their average temperature to be only 98 degrees with a SD=1 degree. (Assume body temperatures are normally distributed.)

Small n
Normal
Unknown SD pop
T-Test!

93) What test statistic should I use?
(a) The t-statistic since the sample size is small and I only know the SD of the sample, not of the whole population.
b) The 1 sample z-statistic since the temperatures are normally distributed.
c) The 2 sample z-statistic since I'm comparing 2 quantities, 98 degrees to 98.6 degrees
d) The chi square test of independence since I'm testing whether the 0.6 difference depends on chance or not.

94) If I used the t-test, I'd have to use SD+ to estimate the SD of the population. What is SD+?
a) 1 b) $\sqrt{8/9}\times 1$ (c) $\sqrt{9/8}\times 1$ d) $\frac{\sqrt{9/8}\times 1}{3}$ $SD^+ = \sqrt{\frac{n}{n-1}} * SD$

95) If I used the t-test, how many degrees of freedom would there be? $= n-1$
a) 3 b) 4 c) 5 (d) 8 e) 9
$9-1=8$

96) I computed these 2 test statistics: -1.7 and -1.8. One is the t-statistic and one is the z-statistic. Which is the t-statistic?
(a) -1.7 is the t-stat b) -1.8 is the t-stat c) not enough info to determine
Because of larger SD^+, t-stat is always smaller (absolute value!)

97) Would the t-test and the z-test both reject the null?
a) Yes, both would yield p-values <5%
b) No, only the t-test would get a p-value < 5%
(c) No, only the z-test would get a p-value < 5%
d) No, neither would yield p-values < 5%

t test	2 sample z
t-stat = -1.7	z-score = -1.8
p-value ≥ 5%.	pvalue ≤ 5% (3.6%)

Question 98
Suppose a well-designed randomized controlled double-blind experiment is done to test a new drug.
The null hypothesis is that the drug works no better than a placebo. A significance test is done and P is computed to be 5%.
Which statement **best** describes what is meant by a P value of 5%?
(a) It means that even if the null hypothesis was true and the drug didn't work, we would still see evidence this strong or stronger 5% of the time just by chance.
b) It means there is a 5% chance that the null is true.
c) It means that there is a 5% chance that the null is false.
d) It means that we have proof the drug certainly works.

Question 99
An experiment on ESP is repeated **1000** times. Suppose there is no ESP, and the experiment is done correctly with no cheating. About how many of the experiments would you expect to find statistically significant evidence for ESP, that is how many of the results would get p-values < 5%? (Note, answer *how many, **not** what percent.)*

a) 0 experiments b) 0.05 experiments c) 5 experiments d) 10 experiments (e) 50 experiments
5% of 1000 → .05(1000) = 50

Question 100
The convention is to reject the null when p<5% and call the result "statistically significant". Is there any particular mathematical justification for this?
a) Yes, the shape of the normal curve, the t-curves and the chi-square curves all have sharp dropping off points that make 5% a natural dividing line.
b) Yes, 5% is the most likely percent to avoid the mistake of rejecting the null when the null is really true. All other percents would yield a higher likelihood of making that mistake.
(c) No, there's no particular mathematical justification for choosing 5%.
The line had to be drawn somewhere! 5% sounded just as good as any other number for a threshold!

Good Luck on the Final!