Warning
This study guide does not cover box plots. 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.

Study Guide for the Final

Key with Worked-Out Solutions

Final Exam Format

The final is comprehensive and all multiple-choice (Scantron graded).

It will have 100 questions (i.e., 100 Scantron bubbles to fill in) with this approximate breakdown:
Exam 1 material: about 22 questions
Exam 2 material: about 22 questions
Exam 3 material: about 22 questions
Post Exam 3 material: about 34 questions.

Exam 1

Chap 1: 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
- Non-randomized controls usually introduce systematic difference between treatment and control groups that could bias the result. These differences are called *confounders*.

Chap 2: 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 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 are.

Sample Questions on Experimental Design (taken from old finals)

The next 2 questions pertain to the following:
In order to assess the effects of tanning beds on the incidence of skin cancer, a researcher collected and compared mole biopsies from individuals who reported that they utilize tanning beds regularly and compared them to mole biopsies from individuals who reported never using a tanning bed. The researcher analyzed the moles for cancerous cells and was unaware of which samples were from the regular tanning group and the group that had never used a tanning bed.

1) Is this an observational study or a controlled experiment?
[✓ a)] Observational study because the people themselves chose whether or not to tan
b) Randomized controlled experiment because one group did not tan

2) Suppose there were a significantly higher number of cancerous moles for the tanning bed users than the non-tanning bed users. Can you conclude with certainty that using tanning beds leads to cancerous moles?

a) Yes, the researcher was blind to which samples belonged to those who tanned and those who didn't
[✓ b)] No—you can't say for sure. There may be other lifestyle differences between people who choose to tan vs. people who do not—for example sunscreen habits, differences in diet, etc.
c) Yes, you can say with certainty that using tanning beds leads to cancerous moles because there were a significantly higher number of cancerous moles in those individuals who used tanning beds

Study Guide for the Final

The next 2 questions pertain to the following:
Last semester, I compared the average Final Exams of Stat 100 students who fully completed their lecture notes to the Final Exams of students who skipped 3 or more chapters from the lecture notes and found a significantly higher final exam average for those who fully completed the lecture notes.

3) Based on these results which conclusion fits best?
- a) There is no relationship between completing the notes and scoring better on the final exam.
- b) Completing the lecture notes definitely helps students score better on the final exam.
- [✓ c)] Completing the lecture notes is associated with—and might cause—students to score better on the final exam. / we see a correlation / relationship but we cant say one causes the other - there could be other reasons like higher attendance, etc. *confounders
- d) Completing the lecture notes is associated with but definitely does not cause students to score better on the final exam.

4) Which of the following is likely to confound the results?
- a) Students who complete the lecture notes are likely to become over-confident and not study as much for the final exam
- b) Students who complete the lecture notes gain valuable practice that helps them to do better on the Final Exam.
- [✓ c)] Students who complete the lecture notes are more likely to be more serious students who maintain other successful study habits.

this may be true - but this is causal - its how it would help. not a confounder →

The next 3 questions pertain to the following study:
A study followed the diet and health habits of 500,000 Americans ages 50-71 over a 10 year period and found that those who ate the most red meat had about a 20% higher death rate from cancer and heart disease than those who consumed the least red meat.

5) Is this an observational study or a designed experiment?
- [✓ a)] Observational Study - we didnt feed people meat - we just observed
- b) Designed Experiment with non-randomized controls
- c) Designed Experiment with randomized controls

6) Based only on the results of this study, which of the following conclusions is best?
- a) Among Americans, high red meat consumption *causes* an increased risk of heart disease and cancer.
- [✓ b)] Among Americans, high red meat consumption is associated with, and *might* cause, an increased risk of heart disease and can
- c) Among Americans, high red meat consumption is associated with, but *does not*, an increased risk of heart disease and cancer.
- d) Among Americans, high red meat consumption is unrelated to heart disease and cancer.

its related but we arent sure how!

7) Which of the following could confound the results? → those who eat red meat are more likely to have unhealthy habits which could cause heart disease
- [✓ a)] Unhealthy habits- Americans who eat the most red meat may also be more likely to engage in unhealthy habits (like eating less vegetables, smoking and drinking) that would put them at higher risk for heart disease and cancer.
- b) Genetics--some people are more prone to heart disease and cancer regardless of their diet. - no relationship
- c) High meat consumptions--Americans have a much higher consumption of red meat than the rest of the world.
- d) All of the above.
- e) None of the above

The next 2 questions pertain to the following:
To test the effectiveness of a new drug (Provenge) designed to treat advanced stage prostate cancer, researchers conducted a clinical trial. The subjects were 512 adult American male volunteers with advanced prostate cancer. Half were randomly assigned to take the drug and the other half were randomly assigned to take a placebo. Neither the subjects nor the doctors who evaluated them knew who was in which group. After three years, 32 percent of those who got Provenge were still alive, compared with only 23 percent of those who got the placebo. (The difference is statistically significant.)

8) Which of the following statements best describes this study?
- a) This is a non-randomized controlled experiment without a placebo
- b) This is a non-randomized controlled experiment with a placebo
- [✓ c)] This is a randomized controlled double blind experiment — random double-blind because neither subjects or doctors knew groups
- d) This is an observational study with controls

9) Which of the following conclusions is best?
- a) This study is likely to have confounders since the people who received the drug already had cancer.
- b) This study is likely to have confounders since the people who received the placebo already had cancer.
- c) This study only shows Provenge to be effective when people are *randomly* given it, and therefore provides no evidence of how effective it would be if given to similar populations under actual, non-random conditions.
- [✓ d)] This study is strong evidence that Provenge can help increase survival time among those with advanced prostate cancer. because it is randomized + double blind → no confounders

Study Guide for the Final

The next 3 questions pertain to the following study:
A recent study asked 700 randomly selected Illinois adults whether or not they regularly watched reality TV and how happy they were with their lives. Those who regularly watched reality TV shows rated themselves as significantly less happy than those who did not.

10) Is this an observational study or a randomized controlled experiment?

[✓ a)] Observational study because the subjects themselves chose whether or not to watch reality TV — we just sampled what they were already doing

b) Randomized controlled experiment because the subjects were randomly selected to participate in the study.

c) Randomized controlled experiment because the subjects were randomly selected to watch reality TV or not.

d) A non-randomized controlled experiment, because the subjects were non-randomly assigned to watch reality TV or not by the researcher whose intent was to make the 2 groups as alike as possible.

11) The study concluded that, reality TV makes people more dissatisfied and less happy with their own lives. Do you think the evidence supports this conclusion?

a) Yes, since the people were randomly selected for the survey, those who watch reality and those who don't must be similar on all characteristics. Since the **only systematic** difference between the 2 groups is reality TV, it must account for the difference in happiness.

[✓ b)] No, since the people were not randomly assigned to watch reality TV or not, there could be other differences between the 2 groups that could confound the results, making it look like reality TV is to blame for their unhappiness when it's really something else. – could be confounders!

c) Yes, people who choose to watch reality TV tend to be people looking to escape their lives, this study proves that such escape is futile and only leads to more unhappiness.

d) It's impossible to draw any conclusions from this study, since 700 people could not possible be representative of the entire adult population of Illinois.

12) Which of the following could confound the study making it look like reality TV is to blame when it's not?

[✓ a)] The people who choose to watch reality TV may be more unhappy to begin with and reality TV has nothing to do with increasing or decreasing their unhappiness.

b) Reality TV may have a numbing effect on people, turning otherwise active people into passive spectators, causing them to be more unhappy. – this is more causal or how TV could cause unhappiness

c) Both of the above are likely confounders.

d) None of the above are likely confounders.

Study Guide for the Final

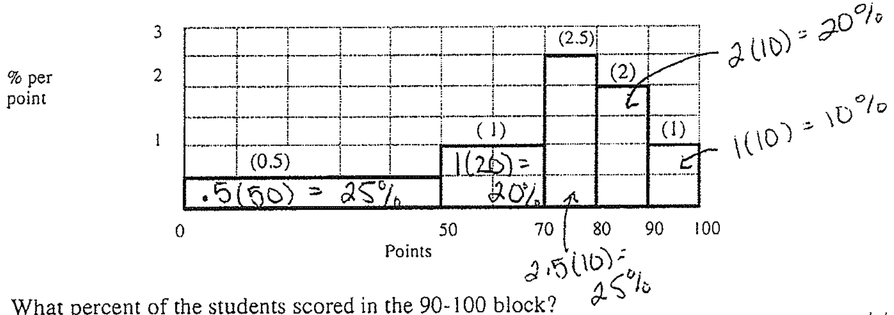
Chapter 3: Histograms

Be able to read histograms, locate the median, and its relation to the average, locate various percentiles, and find the percentage of the area that falls within the various blocks.

Sample Problems:

Questions 1-8 pertain to the histogram below.

The figure below is a histogram for the exam scores of a group of students. The height of each block is given in parentheses Assume even distributions throughout each interval.



area = width × height

14) What percent of the students scored in the 90-100 block?
[✓ a) 10%] b) 15% c) 20% d) 25% e) 30%

1(10) = 10%

15) The average is ____________the median.

[✓ a) less than] b) greater than c) equal to d) cannot be determined

long left hand tail means ave < median

16) Which interval has ***more*** people, 0-70 or 70-90, or are they the same?
a) 0-70 b) 70-90 [✓ c) same]

0-70 has 25% + 20% = 45%
70-90 has 25% + 20% = 45%

17) The percent of students who scored exactly 75 is closest to…..
a) 1% b) 2% c) 3% [✓ d) 2.5%] e) 25%

(just the height of the block – 2.5%)

18) The median score is closest to … a) 50 b) 60 c) 70 d) 80 [✓ e) 72]

(45% below) (50% above + below)

19) What score corresponds to the 70th percentile? a) 50 b) 60 c) 70 [✓ d) 80] e) 90

(70% below)

20) Suppose 10 points were added to all the scores so that the new scores ranged from 10-110. How would that affect the median, average and SD in the histogram above?

- **a)** The average, median and SD would all increase.
- **b)** The average would increase, the median would stay the same and the SD would decrease.
- [✓ **c)** The average and median would increase, but the SD would stay the same.]
- **d)** The average would increase but the median and SD would stay the same.
- **e)** The average and the median would stay the same, but the SD would decrease.

(shifting to the right)
↳ ave + median ↑
SD stays the same b/c spread is the same.

20) Would the normal approximation be appropriate to use to figure out what percentage of the scores fell within various intervals?

- a) Yes, if we knew the average and SD of the data we could use the normal approximation here.
- [✓ b) No, the histogram doesn't follow the normal curve closely so we can't use the normal curve to approximate percentages.]

Study Guide for the Final

Chapter 4: Average, Median and SD

Sample Problems:

The next 3 questions pertain to this list of 4 numbers: 3, 5, 6, 10

21) The median of the list is... [✓ a) 5.5] b) 3 c) 2 d) 4.5 e) 3.5

$\frac{5+6}{2} = 5.5$

22) The average of the list is ... a) 5 b) 4 c) 3 d) 2 [✓ e) 6]

$\frac{3+5+6+10}{4} = 6$

23) The deviations from the average of the list are:

a) 1, -3, 4, 0 b) -1, -2, 2, 6 c) -1, -2, -3, 0 d) 0, 1, 3, 4 [✓ e) -3, -1, 0, 4]

value - ave.

3-6, 5-6, 6-6, 10-6

-3, -1, 0, 4

Question 24

If a list of numbers has a SD of 0 then .... → all deviations must be ∅ so your values = ave

[✓ a) All the numbers on the list must be the same.]
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.

Chapter 5 : Normal Approximation

Sample Problems:

The next 3 questions pertain to the following:

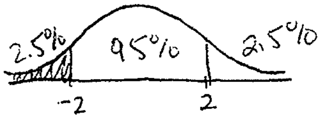
Assume Math SAT scores of a large group of students are **normally distributed** with an average = 500 and a SD = 100.

$Z = \frac{\text{value} - \text{ave}}{\text{SD}}$

25) About what percentage of the students Math SAT scores are *below* 300?

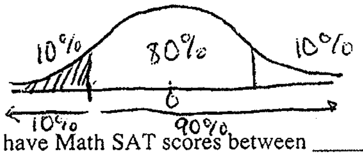
a) 5%
[✓ b) 2.5%]
c) 95%
d) 97.5%

$\frac{300-500}{100} = -2$



26) What Math SAT score corresponds to the 10th percentile? (What score is higher than only 10% of the population?)

a) 335
b) 400
c) 300
[✓ d) 370]



80% → z-score of -1.3 → because on the left side of 0

value = ave + z(SD)

value = 500 - 1.3(100) = 370

27) About 38% of the students have Math SAT scores between ________ and ________.

a) 320 and 680
b) 410 and 590
[✓ c) 450 and 550]
d) 340 and 660
e) 370 and 630

38% → z = ±.5

value = ave + z (SD)

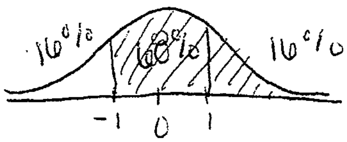
value = 500 + 0.5 (100) = 550

value = 500 - 0.5(100) = 450

28) About what percentage of the students Math SAT scores are **above** 400?

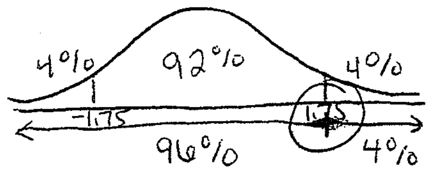
a) 16%
b) 68%
[✓ c) 84%] ← 68+16=84
d) 34%

$z = \frac{400-500}{100} = -1$



29) What Math SAT score corresponds to the 96th percentile? (What score is higher than 96% of the population?)

a) 700
[✓ b) 675]
c) 650
d) 625



z=1.75

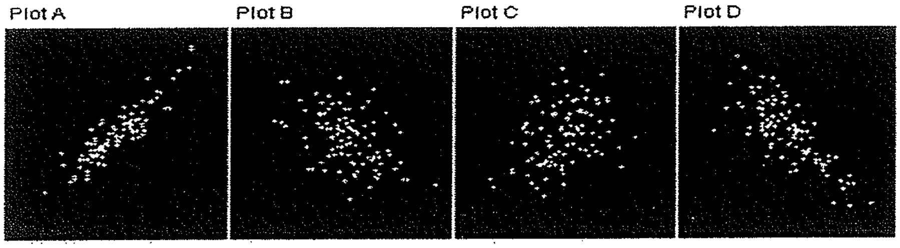
value = ave + z (SD)
= 500 + (1.75)(100)5
= 675

Study Guide for the Final

Exam 2:
Chapter 6: Correlation

The correlation coefficient (r) measures the linear relation between 2 variables.

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



1) Correlation coefficient = -0.79 a) Plot A b) Plot B c) Plot C [✓ d) Plot D]
2) Correlation coefficient = -0.46 a) Plot A [✓ b) Plot B] c) Plot C d) Plot D
3) Correlation coefficient = 0.36 a) Plot A b) Plot B [✓ c) Plot C] d) Plot D
4) Correlation coefficient = 0.90 [✓ a) Plot A] b) Plot B c) Plot C d) Plot D

5) For each of the following pairs of variables, check the box under the column heading that best describes its correlation among typical STAT 100 students: (**Hint:** Every column should have exactly one box checked.)

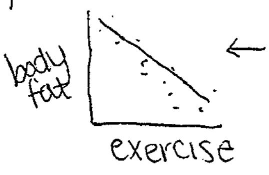
	Correlation		Exactly -1	Between -1 and 0	About 0	Between 0 and 1	Exactly +1
a)	How much you party	How much you drink	□	□	□	☒	□
b)	How much you exercise	Body fat %	□	☒	□	□	□
c)	Height	GPA	□	□	☒	□	□
d)	Number of Stat 100 classes you attended.	Number of Stat 100 classes you missed.	☒	□	□	□	□
e)	Number of Stat 100 classes you attended.	% of Stat 100 classes you attended	□	□	□	□	☒

a. the more you party - the more you drink

drink
party

← its positive
but its not exact
(I cant say if you party 10 hours you will have 20 drinks)

b. The more you exercise, the less body fat you will have



← its negative
but its not exact
because I cant say if you exercise 30 min you will have 20% fat.

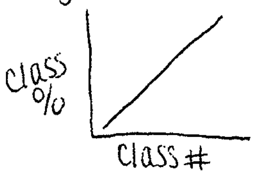
c. height + GPA - no relation!

d. the more classes you attend, the fewer you miss

classes miss
classes attend

negative
this is exact - If I say there were 20 + you missed 2, you know you attended 18.

e. the more classes you attend, the higher % you attended



positive
this is exact - if there were 10 classes$_6$+ you attended 8, I could tell you that you went to $\frac{8}{10} \times 100 = 80\%$!

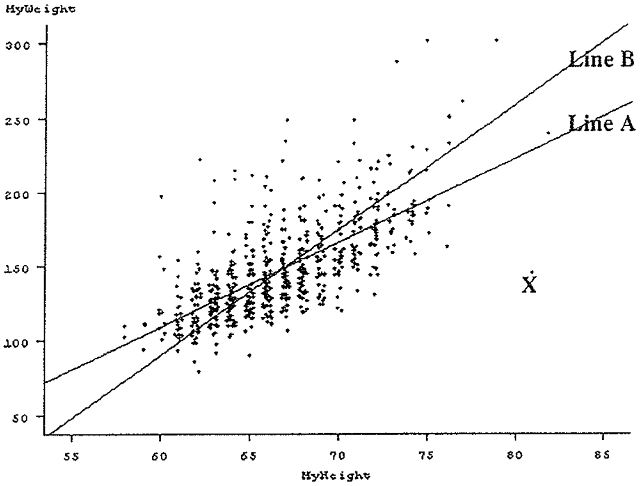
Study Guide for the Final

Chapters 7 – 9: Regression

- The regression line is always flatter than the SD line. (The 2 lines are the same when r= 1 or -1)
- To make a regression estimate first convert the X value or percentile to a Z score, then multiply by r to get the Z score for Y, then convert to the Y value or percentile.
- The slope of the regression line is r (SDy/SDx)
- To find the Y-intercept, plug in the point of averages into this equation: Y = slope X + y-intercept.
- The prediction error = actual value of y – predicted value
- SD of the prediction errors= sqrt (1-r^2) SD_y

Sample Questions:

The next 9 questions pertain to the height and weight survey data from 640 Stat 100 students. The 5 rounded summary statistics and the scatter plot are shown below: (16 pts.)



	Average	*SD*
Height	*67"*	*4"*
Weight	*145 lbs.*	*30 lbs.*

Correlation: r = 0.7

5) Two lines are shown. One is the regression and one is the SD line. Which is the ***regression*** line? (the one with the smaller slope)
Choose one: [✓ a)Line A] b) Line B

6) Which line is the point of averages on? a) Regression Line only b) SD line only [✓ c)Both] d) Neither
(its where the lines intersect)

7) One student is 2.5 SD's above average in height. What is the regression estimate for how many SD's above average he is in weight?
a) 2.1 b) 1.4 [✓ c) 1.75] d) 0

$Z \times r = 2.5 \times 0.7 = 1.75$

8) One student is 63". What is the regression estimate for how much she weighs?
a) 127 lbs. [✓ b) 124 lbs.] c) 129 lbs. d) 115 lbs.

$\frac{63"-67"}{4"} = -1$ $-1 \times .7 = -.7$

value = ave + $z \cdot SD$ = $145 + (-.7)(30) =$ 124

9) The SD of the prediction errors when predicting ***weights*** $\overset{=Y}{}$ from heights is
a) 30 lbs. b) 0 lbs. c) 4" [✓ d) $\sqrt{1 - 0.7^2}$ * 30 lbs.] e) $\sqrt{1 - 0.7^2}$ * 4"

$\sqrt{1-\underset{0.7}{r}^2} \cdot \underset{30}{SDY}$

10) What is the slope of the regression equation when predicting ***weight*** from height? $=Y$
a) 4"/30 lbs. b) 30 lbs./4" [✓ c)0.7 (30 lbs./4")] d) 0.7 e)0.7 (4"/30 lbs.)

$$\text{Slope} = \frac{r \cdot SDY}{SDX} = 0.7\left(\frac{30}{4}\right) = 5.25$$

11) What is the y-intercept of the regression equation when predicting *weight* from height?

weight = slope × height + b
145 = 5.25 (67) + b
b = -206.75

a) 206.75"   [✓ b) -206.75"]   c) 75 "   d) 138.75 "

12) Suppose the 640 heights and weights were converted from inches and pounds to centimeters and kilograms. Would the correlation coefficient change?

[✓ a) No]   b) Yes   c) cannot be determined from the information given

(dont change if multiply all X or Y by same positive #)

13) If point X was removed were deleted the r would … [✓ a) increase]   ii) decrease   iii) stay the same   iv) not enough info

(removing an outlier)

The next 5 questions refer to this situation: A large class took two exams. The scatter plot of the exam scores was roughly football shaped. Here are the 5 summary statistics. (8 pts.)

	Average	*SD*
Exam 1	80	10
Exam 2	70	20

Correlation: r = **0. 6**

14) The slope of the regression equation for predicting $\overset{=Y}{\textit{Exam 2}}$ from Exam 1 is

$m = \frac{r \cdot SD_y}{SD_x} = \frac{.6(20)}{(10)} = 1.2$

i) 0.3   ii) 0.525   [✓ iii) 1.2]   iv) 0.6   v) 0.6857

15) The y-intercept of the regression equation for *predicting Exam 2* from Exam 1 is

$y = mx + b$
$b = y - mx = 70 - 1.2(80)$
$= 70 - 96 = -26$

i) -14   ii) 8   iii) -4   [✓ iv) -26]   v) 59

16) The regression equation for ***estimating Exam 1*** scores from Exam 2 scores is: ***Exam 1 score = 0 . 3 * (Exam 2 score) + 59***
Use the given regression equation to estimate the Exam 1 score of someone who got a 75 on Exam 2?

[✓ a) 81.5]   b) 85   c) 82.5   d) 80

Exam1 = 0.3(75)+59
= 22.5 + 59
= 81.5

17) There's about a 2/3 chance that your estimate in (i) above is right to within ________pts.

$\sqrt{1-r^2} \cdot SD_y$
$\sqrt{1-.6^2} \cdot 10$

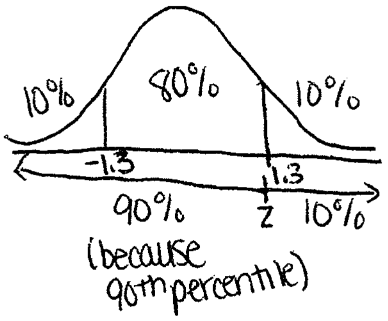
a) $\sqrt{1-0.6^2}$ *20   [✓ b) $\sqrt{1-0.6^2}$ ***10**]   c) 0.6   d) 10   e) 20

18) If 10 points was added to everyone's Exam 2 score the correlation coefficient would…

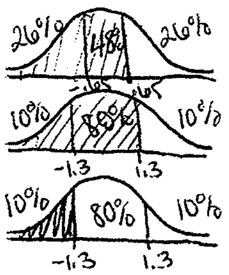
a) Increase   b) Decrease   [✓ c) Stay the Same]

Questions 19-21
Suppose IQ's of husbands and wives follow the normal curve but have different correlations in different countries.
Consider 3 countries where the correlation coefficient between husbands' and wives' IQs are as given in the table below. If a husband has an IQ in the 90th percentile, estimate his wife's IQ percentile for each country.

Husband's IQ Percentile Rank	**r**	**Wife's IQ Percentile Rank**				
19) *90th*	**.5**	a) *10th*	b) *26th*	c) *50th*	[✓ d) *74th*]	e) *90th*
20) *90th*	**1**	a) *10th*	b) *26th*	c) *50th*	d) *74th*	[✓ e) *90th*]
21) *90th*	**-1**	[✓ a) *10th*]	b) *26th*	c) *50th*	d) *74th*	e) *90th*



$z = 1.3 \times \overset{r}{.5} = .65$
$z = 1.3 \times 1 = 1.3$
$z = 1.3 \times -1 = -1.3$



→ 90th

→ 10th

Chapters 10-11 Probability
Sample Questions

The next 6 questions pertain to randomly drawing from the box containing 5 tickets below. → $\text{ave} = \frac{0+2+3+3+7}{5} = 3$

▨0 ▨2 ▨3 $\boxed{3}$ $\boxed{7}$

23) Two tickets are drawn at random **with** replacement. What is the chance that both tickets shaded? "and" = multiply
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

24) Two tickets are drawn at random **without** replacement. What is the chance that both tickets are shaded? "and" = multiply
[✓ 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

25) Five tickets are drawn at random **with** replacement. What is the chance of getting at least one shaded ticket? 1 - none
a) 1 - $(3/5)^5$ b) $(3/5)^5$ c) 1- $(4/5)^5$ d) $(4/5)^5$ [✓ e) 1 -$(2/5)^5$]
↳ b/c 2 white /5 total tickets

26) One ticket is randomly drawn. What is the chance of getting either a shaded ticket or a ticket marked "3"? OR = addition
a) 2/5 [✓ b) 4/5] c) 3/5 d) 1

$\frac{3}{5} + \frac{2}{5} - \frac{1}{5} = \frac{4}{5}$
↳shaded ↳3's ↳shaded & 3

27) 36 draws are made at random with replacement. The EV of the sum of the 36 draws is….
a) 100 b) 50 c) 125 d) 75 [✓ e) 108]
↳ EV sum = n·ave = 36(3) = 108

28) The SD of the box is 2.8. What is the SE of the sum of the 36 draws? → $\text{SE sum} = \text{SD}\sqrt{n} = 2.8\sqrt{36} = 16.8$
[✓ a) 16.8] b) 2.14 c) 21.4 d) 100.8

The next 4 Questions pertain to rolling fair dice. (4 pts.)

29) Two dice are rolled. What is the chance that the sum of the spots is 5? → $\frac{4 \text{ chances}}{36 \text{ total}}$ ⇒ 1 4, 4 1, 2 3, 3 2
i) 2/36 ii) 3/36 [✓ iii) 4/36] iv) 5/36 v) 1/6*1/6 v) 1/6 + 1/6

30) 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

31) One die is rolled 3 times. What is the chance of not 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

not all = 1 - all
(all is your answer to #30)

32) One die is rolled 3 times. What is the chance of getting at least one 6?
i) $(5/6)^3$ ii) $(1/6)^3$ [✓ iii) 1- $(5/6)^3$] iv) 1- $(1/6)^3$ v) 3/6

at least 1 = 1 - none
$= 1 - \left(\frac{5}{6}\right)^3$

Study Guide for the Final

The next 4 questions refers to the following medical test:
Only about 0.1% of young women who participate in routine screening have breast cancer. Suppose **90%** of women who have breast cancer will correctly get a positive result and that **20%** of women ***without*** breast cancer will also get a positive result (false positives). ***Fill in blanks 1 and 2 below*** for a typical sample of 10,000 young women (8 pts)

	Tests Positive	**Tests Negative**	**Total**
Has Breast Cancer	**Cell 1** .90 (10) = [✓ 9]	1	**10**
Does NOT have Breast Cancer	**Cell 2** 9990 − 7992 = [✓ 1998] .20 (9990)	**7992**	**9,990**
Total	**2007**	7993	**10,000**

33) Fill in cell 1 with the correct number a) .1 b) 1 c) 2 [✓ d) 9] e) 9.9

34) Fill in cell 2 with the correct number a) .1 b) 20 c) 99 d) 999 [✓ e) 1998]

35) If a woman gets a positive result, = denominator the chance she really has breast cancer is closest to..

a) 90% b) 20% c) 50% d) .1% [✓ e) 0.45%]

9 ← has cancer / 2007 ← pos. × 100%

36) If a woman gets a negative result, = denominator the chance she really has breast cancer is closest to..

a) 1% b) 0.1% [✓ c) 0.0125%] d) .1% e) 0.45%

1 ← has cancer / 7993 ← neg. × 100%

The next 2 questions refers to the following medical test:
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.
The table below gives the results for 10,000 people.

	Tests Positive	**Tests Negative**	**Total**
Has AIDS	**99**	**1**	**100**
Does Not have AIDS	**594**	**9306**	**9900**
Total	**693**	**9307**	**10,000**

37) What fraction of the people who test negative = denominator truly have AIDs? = numerator
a) 99/100 b) 99/693 c) 9307/10,000 [✓ d) 1/9307] e) 6/100

1 ← have aids / 9307 ← test neg

38) What fraction of the people who test positive truly have AIDs?
a) 99/100 [✓ b) 99/693] c) 693/10,000 d) 1/9307 e) 6/100

99 ← have aids / 693 ← test pos

Study Guide for the Final

Exam 3

Chapters 13-15—EV, SE and histograms for chance numbers

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 on page 162**
- **Know the short-cut formula for the SD of boxes that just have 2 types of tickets on page 156**
- **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

Sample Questions

The next 4 questions pertain to the following situation:

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.

1) 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

1 ← there is 1 correct answer
4 ← there are 4 incorrect answers

$\boxed{\boxed{4}\ \boxed{-1}}$ ← on your tickets is what you will either win or lose

2) The expected value for the student's score is $EV_{sum} = n \cdot ave = 100 \cdot 0 = 0$

ave of box: $\frac{1(4)+4(-1)}{5} = \frac{0}{5} = 0$

[✓ a)] 0 b) 10 c) 20 d) 40 e) 50

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

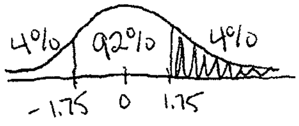
$SD = |4 - -1|\sqrt{1/5 \cdot 4/5} = 2$

[✓ a)] 20 b) .4 c) 2 d) .2 e) not enough info

4) 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).

a) 2% [✓ b)] 4% c) 8% d) 10% e) 20%

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



The next 10 questions pertain to tossing a fair coin and counting the number of heads:

5) 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.

$\boxed{\boxed{1}\ \boxed{0}}$

head "no head"

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

a) 2 b) 2.5 [✓ c)] 5 d) 10 e) 20

$SE_{sum} = SD\sqrt{n}$

$SD = |1-0|\sqrt{1/2 \cdot 1/2} = 1/2$

$SE = (1/2)\sqrt{100} = 5$

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

a) 34% b) 38% [✓ c)] 68% d)95%

8) If you toss a coin 400 times, you'd expect to get 200 heads, give or take _____heads . Fill in the blank with the correct SE.

a) 2 b) 2.5 c) 5 [✓ d)] 10 e) 20

$SE = SD\sqrt{n}$
$= (1/2)\sqrt{400} = 10$

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

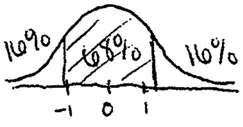
a) 34% [✓ b)] 38% c) 68% d)95%

go to z-scores

$Z = \frac{value - EV}{SE}$

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

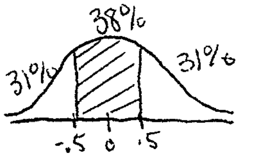
$Z = \frac{55-50}{5} = 1$



go to z-scores

$Z = \frac{195-200}{10} = -.5$

$Z = \frac{205-200}{10} = .5$



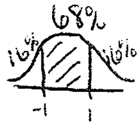
10) If you toss the coin 100 times you'd expect 50% heads, give or take______%. Fill in the blank with the correct SE.

a) 2 b) 2.5 [✓ c) 5] d) 10 e) 20

$$SE\% = \frac{SD}{\sqrt{n}} \times 100 = \frac{.5}{\sqrt{100}} \times 100 = 5$$

11) What's the chance you'd get between 45%-55% heads in 100 tosses?
a) 34% b) 38% [✓ c) 68%] d)95%

$$Z = \frac{value - EV}{SE} \Rightarrow \quad Z = \frac{45-50}{5} = -1 \qquad Z = \frac{55-50}{5} = 1$$

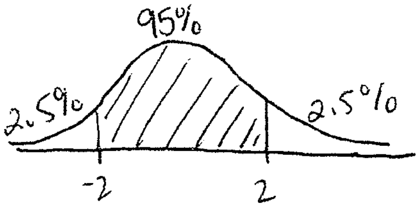


12) If you toss a coin 400 times, you'd expect to get 50%, give or take _____%.
a) 2 [✓ b) 2.5] c) 5 d) 10 e) 20

$$SE\% = \frac{SD}{\sqrt{n}} \times 100 = \frac{.5}{\sqrt{400}} \times 100 = 2.5$$

13) What's the chance you'd get between 45%-55% heads in 400 tosses?
a) 34% b) 38% c) 68% [✓ d) 95%]

$$Z = \frac{value - EV}{SE} \qquad Z = \frac{45-50}{2.5} = -2 \qquad Z = \frac{55-50}{2.5} = 2$$



Question 14

Fill in the blanks to make the statement true

In general the more times you toss a fair coin the _________likely you are to get closer to 50% heads, but the ________likely you are to get closer to *exactly* half heads.

Fill in the first blank with [✓ a) more] b) less c) equally

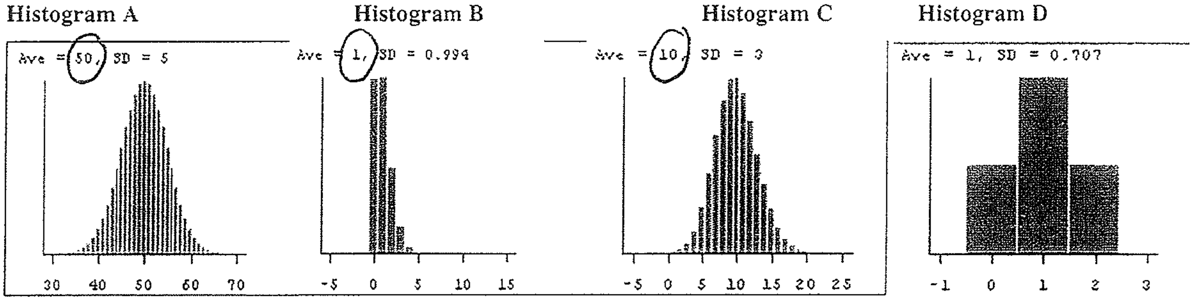
Fill in the second blank with a) more [✓ b) less] c) equally

The next 4 questions pertains to the 3 boxes and 4 histograms below: (10 pts.)

Box 1 [0] [1] $ave = \frac{1+0}{2} = .5$

Box 2 9 [0] 's [1] $ave = \frac{9(0)+1(1)}{10} = 1/10$

Box 3 99 [0] 's [1] $ave = \frac{99(0)+1(1)}{100} = \frac{1}{100}$


Histogram A

Histogram B

Histogram C

Histogram D

Fill in the blanks below to identify the correct histogram. Use each histogram exactly once.

15. The probability histogram for the <u>sum</u> of 2 draws from Box 1 is Histogram *Choose one:* A B C [✓ D]

16. The probability histogram for the <u>sum</u> of 100 draws from Box 1 is Histogram *Choose one:* [✓ A] B C D

EV = n·ave = 100(.5) = 50

17. The probability histogram for the <u>sum</u> of 100 draws from Box 2 is Histogram *Choose one:* A B [✓ C] D

EV = n·ave = 100(1/10) = 10

18. The probability histogram for the <u>sum</u> of 100 draws from Box 3 is Histogram *Choose one:* A [✓ B] C D

EV = n·ave = 100(1/100) = 1

Chapter 16-18

Sample Surveys—

Random Samples are best for the same 2 reasons that randomized experiments are best:

1. **They eliminate selection bias**
2. **They can be translated into box models so you can attach error bars (SE's) to your estimates.**

Box Model for Sample Surveys: (See box model summary on page 188 in the Notes)

- **The box has 1 ticket for every person in the population.**
- **A random sample is of n tickets is drawn from the box without replacement (because you don't want to sample the same person twice).**
- **You know the average or percent of your sample and you use it to estimate the average or percent in the whole population.**
- **Of course, the average or percent in your sample won't be *exactly* the same as that of the population, because of chance error (samples will vary because of the luck of the draw). As long as the sample size is big enough, the probability histogram for the sample average and percent will follow the normal curve so we can attach SE's to our estimates and build confidence intervals.**
- **Note: The size of the population doesn't affect the accuracy of our estimates, only the size of the sample matters. The bigger our sample size, the smaller the SE for averages and percents.**

Sample Questions:

19. City A has **1 million** people and City B has **9 million** people. A simple random sample of **1000** people is taken from City A and a simple random sample of **1000** is taken from City B. Other things being equal the sample from City A is ___________ the sample from city B.

a) 9 times *more* accurate b) 3 times *more* accurate [✓ c)] the same accuracy as d) 9 times *less* accurate e) 3 times *less* accurate

SE only depends on n (your sample size) - so with the same n... your SE (and accuracy) will be the same

Study Guide for the Final

The next 2 questions pertain to the following situation:
A recent Pew Research Center Poll asked a random sample of 1,211 adults nationwide the following question: "Do you think a woman should be able to get an abortion if she decides she wants one no matter what the reason."
I posted the same question on last semester's Bonus Survey. Here's the results of both surveys:

	Yes	No	Sample Size
Pew Research Center	18%	82%	1211
Bonus Survey	46%	54%	631

20) As you can see, the results of the 2 polls are quite different. Which survey gives a better estimate of the percentage of all US adults who would answer "yes" to this question? *Choose one:*
a) The Pew Research survey because the sample size was larger.
b) The Bonus Survey because we can be sure it was an anonymous survey.
[✓ c)] The Pew Research survey because the people were ***randomly*** drawn from all adults nation-wide.

It wont be the bonus survey because of bias (not random) + too narrow!

21) What is SE of the sample percent for the Pew Poll? *Choose one:*
9) It's not possible to calculate a SE for this sample because we don't know the SD of the sample.
10) It's not possible to calculate a SE for this sample because we don't know the size of the population.
11) The SE of the sample percent is approximately 13.4%
[✓ 12)] The SE of the sample percent is approximately 1.1%

$$SE = \frac{SD}{\sqrt{n}} \times 100 = \frac{\sqrt{.18 \times .82}}{\sqrt{1211}} \times 100 = 1.1$$

The next 3 questions pertain to the following:
A recent Gallup poll asked a simple random sample of 900 adults nationwide how much they spent on Black Friday. The sample average was $400 with a SD of $300.

22) What most closely resembles the relevant box model?
a) It has 900 tickets marked with "0"s and "1"s.
b) It has about millions of tickets marked with "0"s and "1"s..
[✓ c)] It millions of tickets. On each ticket is written a $ amount. The exact average and SD are unknown but are estimated from the sample. ↳ we are looking at how much they spent
d) It has 900 tickets. The average of the tickets is $400 and the SD is $300.

23) 900 draws are made ___________replacement.

Choose one: a) With [✓ b)] Without

sampling people = dont replace

24) What is the SE of the sample average?

a) $300 b) $30 [✓ c)] $10 d) $100 e) $0.33

$$SE_{ave} = \frac{SD}{\sqrt{n}} = \frac{300}{\sqrt{900}} = \frac{300}{30} = 10$$

25) If a 95% confidence interval was constructed from this sample to which of the following populations would it apply?
f) All US females
[✓ g)] All US adults
h) All Illinois adults
i) All middle class US adults
j) All of the above

Sampled all aduts nationwide - so applies to all US adults
↳ cant narrow it down any further

The next 5 questions pertain to the following poll:
A CBS News Poll conducted on Oct 24, 2011 asked a random sample of 1,600 adults nationwide the following question: **"Do you think the distribution of money and wealth in this country is fair or you do you think wealth should be more evenly distributed among more people?"** 26% answered "Fair"

26) What most closely resembles the relevant box model?
a) It has 1600 tickets, 26% are marked "1" and 74% are marked "0"
b) It has 1600 tickets with an average of 0.
[✓ c)] It has millions of tickets marked "0" and "1", but the exact percentage of each is unknown and estimated from the sample.

↳ we don't know what all adults say - so we estimate from the sample

27) The draws are made ____________replacement. a) With [✓ b) Without] sampling people = don't replace

28) Which one of the statements below is true?

a) The expected value for the percent of registered Democrats who would answer "Fair" to the question is 26%.
b) The expected value for the percent of corporation executives who would answer "Fair" to the question is 26%.
c) The expected value for the percent of Chicago residents who would answer "Fair" to the question is 26%.
d) All of the above are true.
[✓ e) None of the above are true.] we sampled all adults nationwide- we cant narrow our results down to a specific group

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

EV+/-2SE random sample

a) Yes, a 95% confidence interval is approximately 26% +/- 1.1%
[✓ b) Yes, a 95% confidence interval is approximately 26% +/- 2.2%]
c) No, because we're not given the SD of the sample.
d) No, because we cannot infer with 95% confidence the answers of 200 million Americans from data based on a sampl of only 1,650 randomly selected Americans.

$$SE\% = \frac{SD}{\sqrt{n}} \times 100 = \frac{\sqrt{.26 \times .74}}{\sqrt{1600}} \times 100 = 1.1\%$$

30) If the researcher decreased his sample size by a factor of 4 (to n=400) then the width of the 95% confidence interval would ...

[✓ a) increase by a factor of 2] b) increase by a factor of 4 c) decrease by a factor of 2 d) decrease by a factor of 4

$n \downarrow 4$ $SE \uparrow 2$

$$\frac{SD}{\sqrt{n}} \times 100$$

we are dividing denominator by 2 ($\sqrt{4}$)

If we were to increase the sample size by 4 it would be the opposite - we would divide by 2

You can also plug this into what you did above

$$SE\% = \frac{\sqrt{.26 \times .74}}{\sqrt{400}} \times 100 = 2.2\%!$$

SE doubled!

Study Guide for the Final

Post Exam 3 Chapters 19 - 22 on Significance Tests

All significance tests are based on box models and tell you whether some difference between what you observe and what you expect is likely to be due to chance or not.

Chapters

The one sample Z Test

Null Hypothesis: The population parameter is a particular value (given by the null box) and any difference between our observed sample and what we'd expect is small and just due to chance variation.

Alternative Hypothesis: There is some other reason besides chance that explains the sample data.

Compute the test statistic:

Z = (observed – expected)/ SE

Compute P, the area of the tail. P tells you how likely it would be to get our data or something even further from the null, if the null were right.

The convention is to reject the null when p < 5% and call the result "statistically significant" and when p <1% call the result "highly significant". There's no particular justification for those values. In other words, a p-value of 4.9% isn't really much different than a p-value of 5.1%, people just like to draw the line somewhere.

The next 6 questions pertain to the following situation:

I think I have no musical ability. To test whether I'm right about that, I took a musical memory test online that had 36 questions. For each question I had to choose whether a sequence of notes were the same or different. I answered 24 of the 36 questions correctly. The null hypothesis is that I was just guessing.

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

a) It has 36 tickets, 24 marked "1" and 12 marked "0"

b) It has 36 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" – you are counting the number you got right - and you are making a box for 1 question

2) The draws are made __________ replacement. [✓ a)] with b) without

use sum stats

Assuming the null hypothesis to be true, you would expect me to answer ______questions correct, give or take_______ questions.

$\boxed{1}\ \boxed{0}$

3) Fill in the first blank in the above sentence with the correct expected value. (EVsum)

EV = n · ave
= 36 · 0.5
= 18

a) 12 [✓ b)] 18 c) 21 d) 24 e) 18

ave = $\frac{1+0}{2}$ = 0.5

SD = $|1-0|\sqrt{\frac{1}{2}\cdot\frac{1}{2}}$ = 0.5

4) Fill in the second blank in the above sentence with the correct SE. (SE sum)

SE = SD$\sqrt{n}$
= .5$\sqrt{36}$
= 3

a) 1 b) 2 [✓ c)] 3 d) 4 e) 5

5) The z -statistic for testing the null hypothesis is $\frac{\text{obs-exp}}{\text{SE}} = \frac{24-18}{3} = \frac{6}{3} = 2$

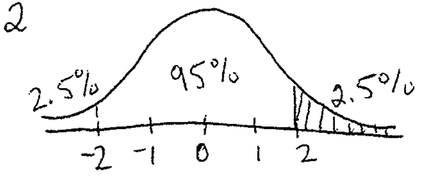
a) 6/SE for the average [✓ b)] 6/ SE for the sum c) 7/SE for sum d) 6/SD of the box

↳ this is 3!

6) The P-value is closest to …

[✓ a)] 2.5% b) 5% c) 16% d) 21% e) 11.5%

Z = 2



$\frac{100-95}{2}$ = 2.5%

Study Guide for the Final

The next 6 questions pertain to the following situation:

An internet access company that serves millions of customers claims that it takes an average of only 1.8 attempts to connect with their service. To test this claim, a consumer advocate looked at a random sample of 400 connections and recorded the number of attempts required to establish each connection. The average of the 400 observations is 2.1 and the SD is 5.0. We want to decide whether the observed difference between 2.1 (the sample average) and the claimed box average of 1.8 is due to chance or not.

7) The null hypothesis box is best described as:

a) containing millions of tickets, each marked 1 or 0, where 1 denotes that a connection was made.

b) containing 400 tickets, each marked 1 or 0, where 1 denotes that a connection was made.

[✓ c)] containing millions of tickets with whole number values such as 1, 3, 5, 2, ...

d) containing 400 tickets with whole number values such as 1, 3, 5, 2 ...

millions of tickets because we are looking at the entire population (n=400 b/c sample 400)
- the values represent the number of tries/attempts it took to connect

8) The average of the null hypothesis box is: [✓ a)] 1.8 b) 2.1

↳ this is what the company claims

9) The SE of the sample average is closest to:
a) 0.05 [✓ b)] 0.25 c) 0.50 d) 5.0 e) 20.0

$$SEave = \frac{SD}{\sqrt{n}} = \frac{5}{\sqrt{400}} = \frac{5}{20} = .25$$

10) The z-statistic is closest to:
a) 0.15 b) 0.12 c) 0.6 [✓ d)] 1.2 e) 6.0

$$Z = \frac{obs-exp}{SEave} = \frac{2.1-1.8}{.25} = 1.2$$

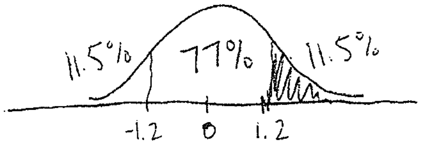
11) The p-value is closest to: a) 77% b) 23% [✓ c)] 11.5%

12) The most reasonable conclusion is that:

[✓ a)] the observed difference could be due to chance

b) the observed difference is real, i.e., the box average is greater than 1.8

we cannot reject the null



$$\frac{100-77}{2} = 11.5$$

The next 3 questions pertain to the following situation:

A computer program is supposed to randomly generate the digits 1, 2, and 3 so that the chance of each is the same, 33.33%. The computer generated 3000 of these numbers, and there turned out to be 1032 1's. Is that too many 1's? You now want to do a hypothesis test to see whether the computer program is really working the way it should.

13. The null box for this hypothesis contains

a) 3000 tickets: 1000 1's, 1000 2's, and 1000 3's.

[✓ b)] Three tickets: one 1 and two 0's.

c) A very large number of tickets with 0's and 1's on them, but we do not know what the percentages of 0's and 1's are.

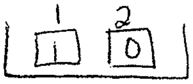
d) 3000 tickets with 0's and 1's on them, but we do not know what the percentages of 0's and 1's are.

e) Three tickets: one 1, one 2, and one 3.

you are making the box to contain "1"s and "not 1s"

14. The standard error for the number of 1's you should get in 3000 tries is about 26. The test statistic for testing the null hypothesis is closest to

a) 0.011 b) -1.23 [✓ c)] 1.23 d) -0.81 e) 0.81



$$ave = \frac{1(1)+2(0)}{3} = \frac{1}{3}$$

$$EV_{sum} = n \cdot ave$$
$$= 3000\left(\tfrac{1}{3}\right) = 1000$$

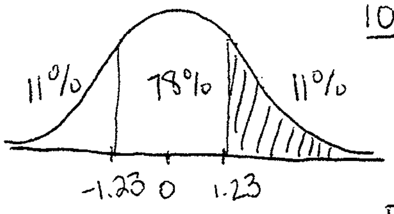
15. What should we conclude from the hypothesis test?

a) We are sure that the computer program is working correctly.

b) We are sure that the computer program is not working the way it should.

[✓ c)] We are not convinced that there is anything wrong with the computer program.

d) We reject the null hypothesis.



$$\frac{100-78}{2} = \frac{22}{2} = 11$$

$$Z = \frac{obs-exp}{SE} = \frac{1032-1000}{26}$$
$$= 1.23$$

P = 11%

Therefore we can't reject the null and it could be due to chance

Study Guide for the Final

Chapter 20 The *t test*

Use the t-test when

1. **You have a small sample n<25**
2. **The contents of the box follows the normal curve**
3. **The SD of the null box is unknown, all you know is the SD of the observed sample.**

The observed sample SD is an underestimate of the box SD, so you have to adjust it by making using SD+ = $\frac{\sqrt{n}}{\sqrt{n-1}} \times$ **SD of sample, and using the t-curves instead of the normal curve.**

t = (observed ave- expected ave)/ SE ave   where SEave = SD+/ $\sqrt{n}$

degrees of freedom = n-1

The next 4 questions refer to the following situation:
A factory that packages corn flakes is supposed to put the flakes in the boxes so that the boxes weigh an average of 16 ounces and a standard deviation of 1 ounce. An inspector randomly chose 12 boxes from one day's output of 2500 boxes. These 12 had an average weight of 15 ounces. (Assume the weights are normally distributed.)
The inspector wishes to test the null hypothesis that the factory is doing what it is supposed to on this day.

16. Which of the following best describes the null box?
a) The box has 12 tickets, with an average of 180/12 = 15 ounces.
b) The box has 12 tickets, with an average of 16 ounces.
c) The box has 2500 tickets, but we do not know exactly the average.
d) The box has 2500 tickets, with 16% 1's and 84% 0's.
[✓ e) The box has 2500 tickets, with an average of 16 ounces]

one day's output is 2500 boxes and they told us the average (it is what the company claims)

17. The standard error for the average of the draws is closest to
a) 0.367   [✓ b) 0.288]   c) 3.46   d) 4   e) .02

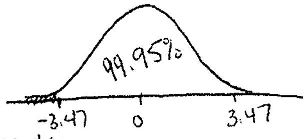
$SEave = \frac{SD}{\sqrt{n}} = \frac{1}{\sqrt{12}} = 0.289$

18. What test statistic would you use?   [✓ a) z-statistic]   b) t-statistic

b/c we know the SD of the box (it is what the company claims) = 1

19. The test statistic is -3.47. What conclusion do you draw?
a) Accept the null hypothesis.
b) There is not enough evidence to suspect there is anything wrong.
[✓ c) Reject the null hypothesis, there is strong evidence that the factory is not doing what it is supposed to.]
d) The p-value is larger than 5%.

P<<5% so reject the null!



$\frac{100-99.95}{2} = 0.025$

20. Now suppose the factory makes the same claim as above, that the boxes weigh 16 ounces on the average, but the factory doesn't make any claim about the SD. Instead, the inspector computes the SD of the 12 boxes and finds the SD =1 ounce
What is the best estimate of the SD of the 2500 boxes?
a) 1 ounce   [✓ b) 1.049 ounces]   c) 1.4 ounces

SD⁺ would be the best estimate!

$SD^+ = SD\sqrt{\frac{n}{n-1}} = 1\sqrt{\frac{12}{11}} = 1,049$

21. What test statistic should the inspector now use?
a) z-statistic   [✓ b) t-statistic]

↳ SNUT!

S = Small sample (n<25)
N = follows normal curve
U = SD of box (factory claims) is unknown
T = DO THE T!

22. If he decides to use the t-statistic, how many degrees of freedom are there?

n-1 = 12-1 = 11

a) 2499   b) 12   [✓ c) 11]   d) 6

23) The instructor computes 2 SE's, one for the z test and one for the t-test. Which SE should be used for the t-test?
[✓ a) .3028]   b) .2887

SD⁺ is larger than SD because we take $SD \cdot \sqrt{\frac{n}{n-1}}$
You can also plug in the number to double check: $SE = \frac{SD}{\sqrt{n}} = \frac{1,049}{\sqrt{12}}$

24) What is the value of the t-statistic?   [✓ a) - 3.3]   b) -3.47   c) – 3. 9

25) Should you reject the null using the t-test?   a) Yes because p>5%   [✓ b) Yes, because p<5%]   b) No

$t = \frac{obs - exp}{SEave} = \frac{15-16}{.3028} = -3.3$

look up 11 degrees of freedom -
I copied it below:

25%	10%	5%	2.5%	1%	0.5%
.70	1.36	1.80	2.20	2.72	3.11

3.3 is way out here so P<0.5%

Chapter 21 The 2 sample z- test

Used to compare averages and percents of 2 populations
Null Hypothesis is that the 2 populations have the SAME average or percent
Alternative is that they're not the same, one is larger than the other

Z = Observed Difference between the 2 Samples/ SE difference
Where SE difference is the square root of the sum of the squares of each sample's SE

The next 4 questions refer to the following study:
A study on the amount of time teenagers spend watching TV took a simple random sample of 100 girls and 64 boys and found the following:

	Girls	Boys
Ave hrs per day spent watching TV	2.5 hours	2.1 hours
SD	1 hour	1 hour

The null hypothesis is that the observed time difference between girls' and boys' TV-watching is due to chance. The alternative hypothesis is that girls watch more TV on the average than boys.

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

a) There is one null box with 164 tickets, 100 marked "1" and 64 marked "0"
b) There is one null box with millions of tickets each marked with the amount of hours spent watching TV.
c) There are 2 null boxes, each with millions of tickets. One box has an average of 2.5 and the other has an average of 2.1
[✓ d)] There are 2 null boxes, each with millions of tickets. The 2 boxes have the same average.
e) There are 2 null boxes, each with millions of tickets marked "0" and "1".

null = "dull"
So we assume both groups are the same

27. The SE for the difference of the 2 samples is closest to
a) 1.41 [✓ b)] .16 c) .1 d) .125 e) .225

$SE_{diff} = \sqrt{SE_A^2 + SE_B^2} = \sqrt{.1^2 + .125^2} = .16$

$\underline{SE_{ave}}$

$SE_{girls} = \frac{SD}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$

$SE_{boys} = \frac{SD}{\sqrt{n}} = \frac{1}{\sqrt{64}} = 0.125$

28. The z statistic for testing the null hypothesis is closest to
[✓ a)] 2.5 b) 1 c) 0

$Z = \frac{obs - exp}{SE}$, or in our case $\frac{obs.diff}{SE_{diff}} = \frac{2.5 - 2.1}{.16} = 2.5$

29. The p-value is .62%. We can conclude

[✓ a)] That it's extremely likely that the average time spent watching TV is greater among girls than boys.
b) That girls watch TV more than boys 99.38% of the time.
c) That it's plausible that there is no real difference between the amount of time boys and girls spend watching TV.

reject the null that stated the difference between boys + girls is due to chance

The next 4 questions refer to the following situation:

A study is done on a new vaccine designed to protect against the common cold. Thousands of people volunteer for the study but the researchers only have enough money to study 85 people. A simple random sample of 49 people are given the vaccine and a simple random sample of 36 people are given a placebo.

After 6 months, the average number of days spent sick from colds for the treatment group was 4 with a SD of 2, while the average for the control group was 6 days with a SD of 3 days . The null hypothesis is that there's no difference between the vaccine and the placebo.

30. Which of the following most accurately describes the null box(es)?
   a) There are 2 boxes, one with 49 people and one with 36 people.
   b) There are 2 boxes each with thousands of tickets. The average of one box is 4 and the average of the other is 6.
   [✓ c)] There are 2 boxes each with thousands of tickets. The average of the 2 boxes is the same.

↳ null = "dull" so we assume both groups are the same

31. What is the SE for the average of the treatment group?

a) 4/7 b) 28 [✓ c)] 2/7 d) 14

$SE_{ave} = \frac{SD}{\sqrt{n}} = \frac{2}{\sqrt{49}}$

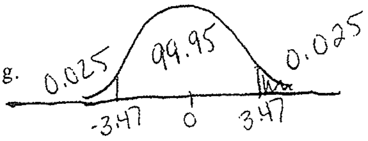
32. What is the SE for the difference between the 2 averages?

a) .285 b) .5 [✓ c)] .576 d).786

$SE_{\substack{control\\(ave)}} = \frac{SD}{\sqrt{n}} = \frac{3}{\sqrt{36}} = \frac{3}{6}$

$SE_{diff} = \sqrt{\left(\frac{2}{7}\right)^2 + \left(\frac{3}{6}\right)^2} = 0.576$

33. The z-statistic is -3.47. What do you conclude?
   [✓ a)] Reject the null because the p-value is less than 5%
   b) Cannot reject the null because the p-value is less than 5%
   c) It's plausible that the 2 day difference in averages could be due to chance variation in sampling.



P=0.025 → P<5% so reject null

Question 34

A researcher discovered a new drug for baldness, and performed an experiment to see if the new drug worked better than the old drug, Rogaine. (Assume the experiment was double-blind and used randomized controls.) The new drug performed better in the experiment. The p-value was 18%.

34) Which of the following statements most accurately describes the implication of that p-value:
   a) The new drug is significantly better than the old.
   b) We can reject the null hypothesis.
   [✓ c)] It is plausible that the superior performance of the new drug was due to chance.

P > 5% so we cannot reject the null

The next 3 questions refer to the following situation:

Gallup asked a random sample of 400 men and 400 women nationwide the following question: "If you were taking a new job and had your choice of a boss, would you prefer to work for a man or a woman?" 50% of the women and 45% of the men said they would prefer a male boss. The null hypothesis is that the 5% difference is due to chance.

35) Which of the following most accurately describes the null box(es)?
   a) There is one null box with 800 tickets, marked with "0"s and "1"s
   b) There is one null box with millions of tickets, marked with "0"s and "1"s
   c) There are 2 null boxes, each with millions of tickets. One box has 45% "1"s and 55% "0"s and the other has 50% "1"s and 50% "0"s
   [✓ d)] There are 2 null boxes, each with millions of tickets. The 2 boxes have the same percentage of "1"s and "0"s.

null= "dull" - we assume both groups are the same

36) The SE for the 2 sample percentages are both about 2.5%.
The SE for the difference of the 2 sample percentages is closest to

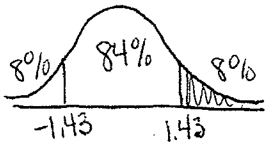
a) 2.5% b) 0% c) 5% [✓ d)] 3.5%

$\sqrt{2.5^2 + 2.5^2} = 3.54$

37) The p-value for testing the null hypothesis is closest to

a) 0% b) 1% c) 5% [✓ d)] 8% e) 84%

$Z = \frac{\text{obs diff}}{SE\%} = \frac{50-45}{3.5} = 1.43$



$\frac{100-84}{2} = 8$

Study Guide for the Final

Chapter 22

The Chi-Square Goodness-of-Fit Test

Used to decide whether the observed data fits a specified model when the model has more than 2 catergories.
With 2 categories (0-1 box) we use the one sample z test.
Null Hypothesis: The observed data fits the model "good". (The difference between the observed and expected is just due to chance.)
Alternative Hypothesis: The observed data does NOT fit the model "good". (The difference between the observed and expected are too big to be due to chance.)

Chi-Square Statistic = sum of (observed frequency – expected frequency) ^{2}/ expected frequency
Degrees of freedom = # of categories -1

The Chi-Square Independence Test

Use to compare the percent composition of 2 or more variables when each variable has 2 or more categories. With 2 variables and 2 categories you can use either a 2 sample z-test or a chi-sq ind test.

(You can think of the Chi-Square Goodness-of-fit Test as a 1 sample test, comparing the percent compostion of the sample to a null box that has multiple categories and you can think of the Chi-Square Independence Test as a 2 sample test, comparing the percent composition of 2 populations when each population has multiple categories.)

Null Hypothesis: The 2 variables are independent. (The 2 populations have the SAME percent composition; the difference between the observed and expected frequencies are just due to chance.)
Alternative Hypothesis: The 2 variables are dependent. (The 2 populations have different percent compositions; the difference between observed and expected are too big to be due to chance.)

Chi-Square Statistic = sum of (observed frequency – expected frequency*) ^{2}/ expected frequency
Degrees of freedom = (# of rows -1) x (# of columns -1)

*To figure the expected frequency for each cell: multiply the row total x column total/overall total

The next 6 questions refer to the following situation:
A certain University has 30% freshman, 25% sophomores, 25% juniors and 20% seniors. A group of 200 students are chosen for a survey. The group has 30 freshman, 40 sophomores, 60 juniors and 70 seniors. The null hypothesis is the students were chosen at random.

	Expected Percents	***Observed #***	***Expected #***
Freshman	10%	30	
Sophomores	20%	40	
Juniors	30%	60	60
Seniors	40%	70	80
Total	100%	200	200

38) To test the null hypothesis that the students were chosen at random we'd do
[✓ a)] the chi-square test for "goodness -of-fit" – there is 1 sample (year in school) that has 3+ categories
b) the chi-square test for independence
c) the one-sample z test
d) the two-sample z test

The table above is missing 3 values. Fill in the missing values by answering the following 3 questions:

39) What is the *expected* number of *freshman*?
a) 10 [✓ b)] 20 c) 30 d) 40 e) 50

10% of 200 = .10(200) = 20

40) What is the *expected* number of *sophomores*?
a) 10 b) 20 c) 30 [✓ d)] 40 e) 50

20% of 200 = .20(200) = 40

41) To compute the proper test statistic you'd sum 4 terms: 5 + 0 + 0 + ____. The term for **seniors** is missing, what should it be?
a) 0 b) 1 [✓ c)] 1.25 d) 1.43 e) 2.5

$$\frac{(obs-exp)^2}{exp} = \frac{(70-80)^2}{80} = \frac{100}{80} = 1.25$$

X^2 table for #43:

D.of F	99%	95%	90%	70%	50%	30%	10%	5%
3	.12	.35	.58	1.42	2.37	3.67	6.25	7.82

Study Guide for the Final

42) The number of degrees of freedom is
a) 2 [✓ b) 3] c) 4 d) 5 e) 6

= #categories − 1
= 4−1 = 3

43) What do you conclude?
a) Reject the null because P < 5% b) Reject the null because P > 5% [✓ c) Cannot reject the null because P > 5%]

X^2 statistic (from ?41) = 5+0+0+1.25 = 6.25

look this up in the table (X^2) - find that this corresponds to P=10%

The next 3 questions refer to the following situation:
A simple random sample of 148 Stat 100 students were asked whether or not they thought they would ever use statistics again in their lives. Assume the students were chosen from a population of 2000. The following table gives the results:

	Would use	Would not use	Total:
Men	47	21	68
Women	64	16	80
Total:	111	37	148

The chi-square statistic to test the null hypothesis that sex and anticipated use are independent is 2.32.

44. To compute this statistic, expected frequencies were calculated. What is the expected frequency for the men who answer "would use"?
[✓ a) 51] b) 47 c) 44

The totals are calculated above - just add the rows + columns.

Total men = 68 Total would use = 111
Overall total: 148

$$\text{Expected} = \frac{(68)(111)}{148} = 51$$

45. How many degree of freedom does the chi-square statistic have?
[✓ a) 1] b) 2 c) 3 d) 4

= (rows − 1)(columns − 1) = (2−1)(2−1) = 1

46. Can you reject the null hypothesis?

Table:

Dof F	50%	30%	10%	5%
1	.46	1.07	2.71	3.84

a) Yes [✓ b) No]

P is between 10% + 30%

2.32 falls here!

47. We did a chi-square independence test and got a statistic of 2.32. Would it be appropriate to do 2-sample z test on this data?
[✓ a) Yes] b) No

(2 categories + 2 populations)

The next 3 questions refer to the following situation:
The table below shows the results of a recent nationwide poll of Hispanic adults who were asked;
"All in all, do you think the situation for the younger generation of Hispanic or Latino Americans is better, worse, or about the same as their parents' situation was when they were the same age?"
You may assume that the data are from a simple random sample of 200 people, of whom 100 were over 35 years old and 100 were 18-34 years old.

	Better	Worse	About the Same	Total
18-34	49	37	14	100
Over 35	39	45	16	100
Total	88	82	30	200

46. To answer the question of whether the sample data reflects a real difference between older and younger Hispanic Americans on this issue, you would use the
a) the one-sample z test
b) the two-sample z test
c) the chi-square test for "goodness-of-fit" which specifies the contents of the box
[✓ d) the chi-square test for independence] – 2 samples with 2 or more categories

49. To compute the test statistic you need to calculate the expected frequencies. What is the expected frequency for the 18-34 year olds who answer "Better"?
a) 39 b) 40 [✓ c) 44] d) 45 e) 50

18-34 year olds = 100
better = 88
overall total = 200

$$\frac{100 \times 88}{200} = 44$$

50. To compute the test statistic you need sum 6 terms: 25/44 + 16/41 + 1/15 + _____ + ______ + ______
The first 3 terms correspond to the 1st row of the table and sum to 1, what is the sum of all 6 terms?
a) 1.5 [✓ b) ~~2.32~~ 2.05] c) 3 d) 3.39 e) 5

1.025

(the population is divided 100/100 for the two rows - so you just have to take 1.025 + 1.025)
22

Expected:
$$\text{Over 35 + Better} = \frac{88(100)}{200} = 44$$
$$\text{Over 35 + Worse} = \frac{82(100)}{200} = 41$$
$$\text{Over 35 + About the Same} = \frac{30(100)}{200} = 15$$

X^2 value: $(\text{obs}-\text{exp})^2/\text{exp}$

$$\frac{25}{44} + \frac{16}{41} + \frac{1}{15} + \frac{(39-44)^2}{44} + \frac{(45-41)^2}{41} + \frac{(16-15)^2}{15} = 2.05$$

Chapter 23

Significance tests can only tell you whether or not a difference is likely to be due to chance, not whether a difference was important or what caused the difference, or whether the experiment was properly designed

By definition, statistically significant results will appear by chance with enough tests. A p-value of 5% means that even when the null is true, you'll reject it 5% of the time.

Question 51

Which of the following does a test of significance deal with?

[✓ a.] Is the difference due to chance?
b. Is the difference important?
c. Was the experiment properly designed?
d. What are the probable causes of the difference?

The next 2 questions refer to the following situation:
100 investigators each set out to test a different null hypothesis. Unknown to them, all the null hypotheses happen to be true.

52. About how many of them would you expect to get statistically significant results?

a. None, if they did the test correctly they would all confirm that the null hypothesis is true.
b. 1
[✓ c.] 5
d. 95
e. Impossible to predict.

$P \leq 5\%$ is statistically significant
→ even if the null is rejected, we would still see the null appearing true 5% of the time

5% of 100 = 0.05(100) = 5!

53. About how many of them would you expect to get highly statistically significant results?

a. None, if they did the test correctly they would all confirm that the null hypothesis is true.
[✓ b.] 1
c. 5
d. 95
e. Impossible to predict.

highly significant - same reasoning as above but highly significant means $P \leq 1\%$

1% of 100 = .01(100) = 1!

Question 54

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%?

a. 0
b. 5
c. 10
[✓ d] 50
e. Not enough information to determine.

5% of 1000

.05(1000) = 50!

Question 55

A new chemical is tested to see if it causes cancer in lab mice. 250 mice are chosen at random and fed the test chemical in their food and 250 mice get the same food without the chemical. After 3 years, cancer rates in the two groups are compared using a 2-sample z-test.
The investigators are looking at about 50 different types of cancer, so they do 50 different 2-sample z-tests. They find statistically significant evidence for lung and liver cancer.

Is it valid to reject the null hypothesis and conclude that the chemical does cause cancer?

a. Yes, since statistically significant results were found in 2 of the 50 types of cancer (lung and liver).
[✓ b.] No, because if you run 50 tests, you're likely to get 2 statistically significant results even if the null hypothesis is true, just due to the luck of the draw.

.05 (50) = 2.5
(same reasoning as above)