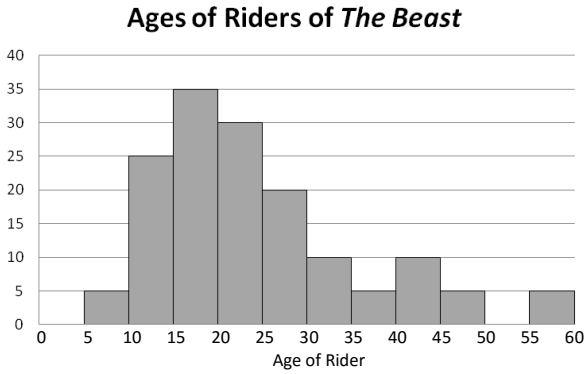
AP Statistics Semester I Exam Review

This is a ***required*** study guide for the upcoming semester exam, covering chapters 1-15 in your textbook. Do these problems on a separate sheet of paper. You may use your book, notes, old tests, or each other for help. This review may not be completely comprehensive, but combined with your old tests and quizzes, it will give you a good idea of the material that will appear on the final exam. The answers will be posted on my website for you to use. Good luck!

List of terms and concepts that you need to be familiar with:

- **Displaying and Describing Data (Ch. 1-4)**
  - Categorical/quantitative data
  - Dotplot
  - Stemplot
  - Histogram
  - Ogive
  - Shape - Symmetrical/Skewed/etc.
  - Center - Mean /Median/What is more appropriate
  - Spread/Range
  - Boxplot
  - Quartiles/interquartile range (IQR)
  - Outliers
  - Variance
  - Standard deviation
- **Describing a Position in a Distribution (Ch. 5)**
  - Percentile
  - Z-score
  - Shifting and Scaling Data
  - Density curve
  - Normal distribution/normal curve
  - Empirical rule (68-95-99.7)
  - Standard normal curve
  - Normal probability plot
- **Scatterplots, Linear Regression (Ch. 6-9)**
  - Response variable
  - Explanatory variable
  - Scatterplot
  - Correlation coefficient
  - Regression line
  - Slope
  - Y-intercept
  - Coefficient of determination (r^2)
  - Residual plot
  - Influential observation
  - Linear transformation
  - Extrapolation
  - Lurking variable
  - Confounding variable
- **Simulations, Surveys and Experiments (Ch.10-12)**
  - Population/Parameter
  - Sample/Statistics
  - Census
  - Simple random sample
  - Stratified random sample
  - Cluster Sample
  - Systematic Sample
  - Bias:
    - ◊ Voluntary response
    - ◊ Convenience
    - ◊ Undercoverage
    - ◊ Non-response
    - ◊ Response
  - Observational/experimental study
  - Cause and effect
  - Experimental unit/subject
  - Randomization
  - Factor/Level
  - Treatment
  - Placebo/placebo effect
  - Statistically significant
  - Blind/double-blind
  - Block-design
  - Simpson's Paradox
- **Probability (Ch. 13-15)**
  - Random
  - Probability
  - Sample space
  - Independent
  - Disjoint
  - Conditional probability
  - Random variable
  - Probability histogram
  - Expected value
  - Standard Deviation
  - Law of Large Numbers

Displaying and Describing Data (Chapters 1-4)



This histogram shows the ages of the last 150 people who rode *The Beast* at Kings Island. Use it to answer questions #1-3.

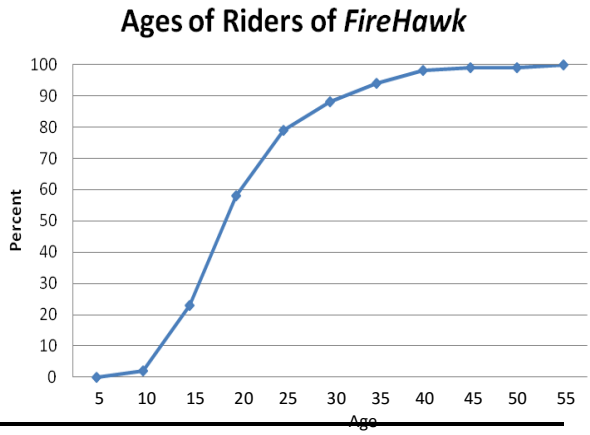
1. Describe the distribution.
2. Is the mean greater than, less than, or equal to the median?
3. In which range will the median of this data lie?
   In which range will the mean of this data lie?
4. Give two major differences between the mean and the median as measures of the center of a distribution.

The following data is a list of the ages of the last 30 people to ride the carousel at Kings Island. Use it to answer questions #5-10.

35 71 24 54 55 68 4 29 31 6 10 73 45 48 52 27 3 43 52 81 78 36 39 11 8 63 60 29 35 9

5. Organize the data into a stemplot. Describe the shape.
6. Find the mean, standard deviation, and five-number summary of the data? Based on the values you get what can you say about the shape of the data set?
7. What is the interquartile range?
8. Does the data have any outliers? How can you tell?
9. Organize the data into a boxplot. 10. Organize the data into a histogram.

The following ogive shows the ages of the last 1000 people to ride the *FireHawk* at Kings Island. Use it to answer questions #11-15.



11. If a 32-year-old rides *FireHawk*, in what percentile would he be?
12. What age corresponds to the 40th-percentile?
13. What is the IQR for this set of data, approximately?
14. Should a *FireHawk* rider that is 55 be considered an outlier in this set of data? Why or why not?
15. If you were to draw a histogram of this data, would it symmetrical or skewed? How can you tell?

---

16. Describe the two main differences between a bar graph and a histogram.

Use the following table to answer questions 17-20.

	Smoking Status			
Education	Never smoked	Smoked, but quit	Smokes	
Did not complete high school	82	19	113	214
Completed high school	97	25	103	225
1 to 3 years of college	92	49	59	200
4 or more years of college	86	63	37	186
	357	156	312	825

17. What percent of those with 4 or more years of college have quit smoking?
18. Find the conditional distribution of levels of education for people who never smoked.
19. Create a segmented bar chart comparing education levels for people who never smoked and people who smoked.
20. Do these data suggests that there is an association between education level and smoking status? Give statistical evidence to support your conclusion.

Describing a Position in a Distribution (Ch. 5)

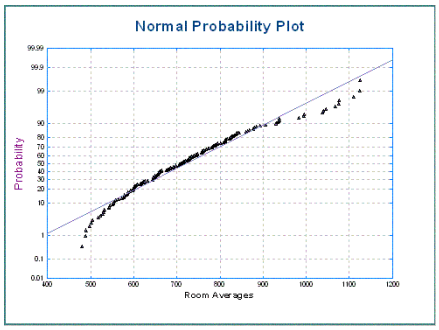
21. Suppose that the mean of a set of data is 55.8 and the standard deviation of a set of data is 12.2.
    a. What would the new mean and standard deviation be if you added 10 to each data point?
    b. What would the new mean and standard deviation be if you multiplied each data point by 5?
    c. What is the variance of this set of data?

The lengths of time (in hours) an incandescent light bulb can stay lit solidly are normally distributed, with a mean of 62 hours and a standard deviation of 4.4 hours. Use this information to answer questions #22-26.

22. What percent of light bulbs can last within one standard deviation of the mean, between 57.6 hours and 66.4 hours?
23. A company considers a bulb defective if it can only last 50 hours straight. What portion of light bulbs would this company consider defective?
24. An energy-efficient bulb can stay lit for at least 68 hours. What percent of incandescent bulbs can stay lit for that long?
25. What portion of light bulbs can stay lit for between 55 and 65 hours straight?
26. The top 2% of light bulbs can stay lit for at least how long?

---

27. Sophia got a 95% on her Statistics mid-term and a 91% on her Calculus mid-term. The grades on both tests were normally distributed. The Statistics grades had a mean of 87%, with a standard deviation of 7%, while the Calculus grades had a mean of 85% with a standard deviation of 4%. On which test did Sophia do better, compared to the rest of her class? How can you tell?
28. A set of data has the following normal probability plot. Is the data normal? How can you tell?



29. A bad statistician heard that the mean age of the riders of *Invertigo* at Kings Island is 23, with a standard deviation of 5.5 years. He concluded that 95% of the riders of *Invertigo* must be between the ages of 12 and 34. What is incorrect about his conclusion?

---

30. Describe similarities and differences between z-scores and percentiles.

Scatterplots, Linear Regression (Ch. 6-9)

The table below compares the average weight and average life span of several common dog breeds. Use it to answer questions #31-39.

Breed	Avg. weight (lbs)	Avg. life span (yrs)		Breed	Avg. weight (lbs)	Avg. life span (yrs)
Beagle	26	13.0		Golden retriever	70	11.0
Boxer	70	12.5		Labrador retriever	73	12.0
Bulldog	50	11.1		Pomeranian	5	13.9
Chihuahua	4	14.3		Poodle	42	12.7
Dachshund	19	13.2		Rottweiler	113	10.8
German shepherd	82	12.3		Yorkshire terrier	6	13.5

31. What are the response variable and the explanatory variable here?
32. Calculate r for this set of data. What can you conclude about the relationship between average weight and average life span of a dog breed?

33. Calculate the equation of the least-squares regression line for this set of data. How accurate will the regression line be at predicting the average life span of a dog breed from its weight? How do you know?
34. Interpret the slope and the intercept.
35. Construct a residual plot and comment on the appropriateness of your model.
36. A Shih Tzu weighs an average of 13 pounds. Approximately how long is a Shih Tzu's average life span?
37. A Shih Tzu's actual average life span is 12.5 years. What is the residual of that data point?
38. A Mastiff weighs an average of 200 pounds. Approximately how long is a Mastiff's average life span? Are you confident in your answer as the true estimate of a Mastiff's average life span? Why or why not?
39. What percent of a dog breed's average life span can be explained by its average weight?

---

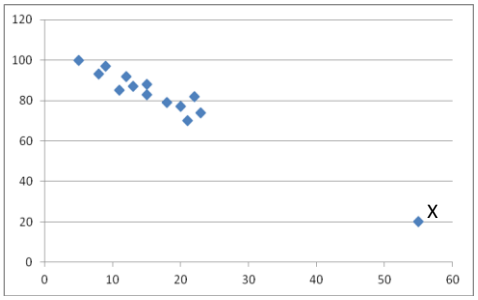
40. A researcher runs a detailed study and concludes the following: "The correlation between the age when a child first walks and the age when a child says their first word appears to be approximately zero." Describe what this means to someone who does not know anything about statistics.
41. Complete the statement: "Every LSRL passes through the point (,)".
42. A study was conducted to see if a baby's birth weight was related to their birth length. A sample of 200 babies was measured and the following data was gathered:

	Mean	Standard deviation
Birth weights:	8.7 lbs	2.8 lbs
Birth lengths:	15.3 in	4.2 in

Correlation coefficient: $r = 0.895$

A scatterplot revealed that the data was fairly linear. Use all of that information to write the regression line that predicts *birth length* from *birth weight*.

43. In the scatterplot to the right, what would happen to the value of *r* if point X was removed? What would happen to the slope of the regression line? Based on that, is point X an outlier? Why or why not?



---

44. This regression analysis examines the relationship between the number of years of formal education a person has and their annual income.

Write the regression equation and define the variables of your equation in context. What is the correlation coefficient? Interpret this value in context.

Dependent variable is **Income**
R-squared = 25.8%
s = 3888 with 57 degrees of freedom

Variable	Coefficient	s.e. of Coeff
Constant	3984.45	6600
Education	2668.45	600.1

---

45. The following table shows the federal debt for a short period of time from 1980 through 1991.

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Debt (in trillions)	0.909	0.994	1.1	1.4	1.6	1.8	2.1	2.3	2.6	2.9	3.2	3.6

a) Construct a scatterplot, determine if the data appears linear or exponential.
b) Transform the data (x, log y) and perform a least squares regression on the transformed data.
c) Transform your least squares regression back into an exponential model.
d) Use your exponential model to predict the national debt in the year 2000.

Simulations, Surveys and Experiments (Ch.10-12)

46. Studies have indicated that there is a strong, negative correlation between the number of times per week that you brush your teeth and the number of cavities you get each year. Which of the following can be concluded from this information?
    a. Not brushing your teeth causes you to get cavities.
    b. People who don't brush their teeth often are more likely to have cavities than people who do.
    c. There is statistically significant evidence that brushing your teeth prevents cavities.
    d. None of these can be concluded.

---

47. When asked what their favorite sport is, 44% of American men say football, 26% say baseball, 22% say basketball, 5% say soccer, 2% say hockey, and 1% say something else. Use the following sequence of random numbers to simulate asking 30 men about their favorite sport. Clearly explain the process you used in your simulation.

    14459 26056 31429 80371 65103 62253 50490 61181 38967 98532 62183 70632 23417 26185

48. A report on a new brand of headache medicine, *Probanol*, is published that says, "After extensive research, there is statistically significant evidence that *Probanol* reduces the likelihood of getting a migraine headache." Explain what that means to someone who doesn't know anything about statistics.

An eye doctor believes that he has invented a new drop that improves eyesight. He randomly selects 200 people, and he has them read an eye chart. He then administers the drops and has them read a similar eye chart, noting any improvement. However, despite telling everyone that they will be receiving the new drops, he only gives 100 of the subjects the actual drops; the other half of the sample is simply given water drops.

49. What is the purpose of the water drops? Why couldn't the eye doctor simply have administered his drops and noted improvement?
50. What is the factor in this experiment? What are the treatments?
51. Is this experiment blind? Is it double-blind?

The principal wants to know if the BHS student body would like the library to stay open longer during the day. She gathers a random sample of 100 students from each grade level.

52. What is the *population* of this study?
53. What is the *parameter of interest*?
54. Which of these most accurately describes this sample: simple random sample, stratified random sample, or census?

---

55. Define confounding variable. Give an example.
56. Describe the difference between stratified and cluster sampling methods.
57. What is the purpose of blocking?
58. What is the purpose of randomization?
59. What is a bias?

60. A politician wants to know how the residents of his district will react to a bill that lowers the driving age to 15 years old. He runs an ad during the evening news on a local television station that says:

Let us know what you think! Would you be in favor of allowing 15-year-old children to get their driver's license, or would you rather keep the driving age at 16-years-old, when they are more mature and ready to take on the responsibility of driving? Give us a call at 555-7834, and give us your opinion!

Which of the following types of bias are present in this ad: voluntary response bias, convenience sampling, undercoverage bias, non-response bias, poor wording effect bias?

Probability (Ch. 13-15)

61. Event A is that you will complete this exam review. Event B is that you will get an A on the semester exam. Are Events A and B disjoint? Are they independent? Explain.

You have a large bag of marbles, with proportions of each color listed below. Use that chart to answer questions #47-54.

Color	Red	Yellow	Blue	Green	Orange	Purple	Black	White
Prob.	0.13	0.06	0.21	0.08	0.18	0.05	0.11	???

62. What is the probability of drawing a white marble from the bag?
63. What is the probability of drawing either a red or blue marble from the bag?
64. What is the probability of drawing a marble from the bag that is **not** yellow?
65. I draw a marble from the bag and tell you that it isn't black or white. What is the probability that it is green?
66. Imagine that you draw two marbles from the bag, replacing the first before drawing the second. What is the probability that they are both orange?
67. What is the probability that neither of them is orange?
68. Imagine that you draw five marbles from the bag, replacing them each time before drawing the next. What is the probability that you get *at least one* purple?

Suppose that, in Mr. Willets' AP Calculus class, 71% of the students are seniors, 26% are juniors, and 4% are sophomores. 36% of the seniors and 15% of the juniors are also in AP Statistics, but none of the sophomores are. Use that information to answer questions #55-60.

69. Draw a tree diagram to organize the above information. Make sure to include the final probability of each branch.
70. What is the probability that a randomly-selected student from Mr. Willets' class is a junior in AP Statistics?
71. What is the probability that a randomly-selected student from Mr. Willets' class is a senior who is **not** in AP Statistics?
72. What is the probability that a randomly-selected student from Mr. Willets' class is a sophomore who is in AP Statistics?
73. What is the probability that a randomly-selected student from Mr. Willets' class is also in AP Statistics?
74. A student is randomly selected from Mr. Willets' class, and that student is also in AP Statistics. What is the probability that they are a senior?

---

75. Which of the following sequences of heads/tails is *most* likely to occur, if you flipped a fair coin 6 times?
    a. H, H, H, H, H, H
    b. H, T, H, T, H, T
    c. H, H, H, T, T, T
    d. T, H, H, T, H, T
    e. None of these

---

76. A fair die is rolled 600 times. Label each of the following statements as true or false:
    a. Exactly 100 of the rolls will be 1s.
    b. As the number of rolls approaches 600, the proportion of 5s rolled will get closer to 1/6.
    c. A run of 8 odd numbers in a row is impossible, since the proportion of evens and odds has to stay close to 50%.
    d. The number of 3s rolled should be approximately equal to the number of 6s rolled, by the end.
    e. The first twelve rolls will include two of each number, to keep the proportions equal.

In the following probability distribution, X = the number of 10s rolled on three ten-sided dice. Use it to answer questions #63-65.

X	0	1	2	3
P(X)	.729	.243	.027	.001

77. Find the mean and standard deviation of X.
78. If you multiplied each X-value above by 3 and added 5, what would be the new mean and standard deviation of X?

79. Suppose you are given a standard 6-sided die and told that that die is loaded in such a way that while the numbers 1, 3, 4, and 6 are equally likely to turn up ,the numbers 2 and 5 are three times as likely to turn up as any of the other numbers.
    a) The die is rolled once and the number turning up (X) is observed. Use the information given above to fill in the following table:

X	1	2	3	4	5	6
P(X)						

    b) Let A be event: “The number rolled is a prime number (note that 1 is not a prime)”. Find P(A).
    c) Let B be event: “The number rolled is even”. Find P(B).
    d) Are A and B disjoint?
    e) are A and B independent?

---

80. A man on the street offers you a wager. He’ll fan out a deck of cards and let you pick one at random. If it’s a face card, he’ll give you $2. If it’s an ace, he’ll give you $4. If it’s anything else, you give him $1. What is the amount you are expected to win each time you play? Should you take his offer?

---

81. What is the general conditional probability equation?
82. What two equations can you use to test for independence of two variables?

Answers to AP Statistics Semester Exam Review:

1. shape-outliers-center-spread in context
2. greater than
3. 20-25; 25-30
4. Median – splits data in half, resistant to outliers. Mean – balancing point, not resistant to outliers.
5. Stemplot
6. μ = 39.3; M = 37.5; σ = 23.52; five-number summary – (3 24 37.5 55 81); mean > Median - right skewed
7. IQR = 31
8. No, because 1.5xIQR yields a range from -22.5 to 101.5, and none of the data is outside of that range.
9. Boxplot
10. Histogram
11. 90th-percentile
12. 18 years old
13. 10
14. Yes, because 1.5xIQR is 15, and 55 is more than 15+Q_3=40.
15. Skewed to the right, because most of the data is on the left side of the age axis.
16. Bar graphs are for categorical data, have gaps between bars; histograms – quantitative data, no gaps.
17. 0.462
18. No HS - 23%, HS- 27%, College 1-3 - 26%, College 4+ - 24%
19. Segmented bar graph
20. Compare the graphs
21. a. μ = 65.8; σ = 12.2
    b. μ = 279; σ = 61
    c. 148.84
22. 68%
23. 0.0032
24. 0.0869
25. 0.6958
26. 71.02 hours
27. Calculus exam; her z-score for Statistics (1.14) was lower than her z-score for Calculus (1.5)
28. No, because the normal probability plot isn't straight.
29. The data probably is not normal, as it is most likely skewed toward younger riders.
30. Both are the measures of position; z-score measures st.dev from the mean, percentile measures 5 below.
31. Response – life span, explanatory – average weight
32. r = -0.849; there is a strong negative relationship between a dog breed's average weight and average life span.
33. $\widehat{\text{life span}} = 13.8 - 0.027\text{weight}$; it will be pretty accurate, since the correlation is so strong.
34. Interpretations
35. Residual plot
36. 13.5 years
37. -1
38. 8.4 years; not very confident, since the Mastiff is far outside of our range of data.
39. 72.2%
40. There is no relationship between the age when a child first walks and the age when a child first says their first word. You cannot predict one based off of the other.
41. $\bar{x}, \bar{y}$
42. $\widehat{weight} = 3.161 + 1.343length$
43. r would most likely decrease, as the scatter would increase. The slope of the least-squares regression line would be unaffected, as the line would follow a similar pattern. Point X is an outlier, as it affects the value of r.
44. $\widehat{income} = 3984.45 + 2668.45education$; r=0.508 - there is positive, moderately strong association between the income and the education level

45. a) scatterplot
    b) $\log(\widehat{debt}) = -0.094 + 0.056(years\ since\ 1980)$
    c) $\widehat{debt} = 10^{-0.094}10^{0.056years}$
    d) $\widehat{debt} = 12.08\ trillion\ \$$
46. b
47. I broke the numbers 1-100 into the following categories: 01-44 = football, 45-70 = baseball, 71-92 = basketball, 93-97 = soccer, 98-99 = hockey, 00 = other. Then I created two-digit numbers from the random number sequence and came up with the following results: 15 said football, 8 said baseball, 5 said basketball, 1 said soccer, and 1 said hockey
48. There is enough evidence, gathered through many experiments, to conclude that the relationship between taking *Probanol* and the reduction of migraine is not happening by random coincidence. Thus, it can be concluded that *Probanol* causes migraine reduction.
49. He needed a placebo, to control for the placebo effect. The drops will be considered effective only if participants who receive the drops have better outcomes than participants who receive the placebo.
50. The factor is the eye drops, while the treatments are the actual drops versus the placebo group.
51. It is blind, as the subjects do not know which group they are in, but it is not double-blind, as the eye doctor himself does know.
52. The population is the BHS student body.
53. The parameter is "Do students want the library to be open longer"
54. Stratified random sample
55. When levels of one factor are associated with the levels of another factor in such a way that their effects cannot be separated. For example, feeding dogs and cats the same new food and comparing health benefits. We will not be able to tell whether any differences in animals' health are due to the new food or to the differences in how species respond.
56. Stratified – divide population by a certain characteristic you feel will affect the response into stratus, the pick a proportional number to survey from each strata. Cluster – all clusters are similar to each other and to the whole population, pick a cluster at random, survey everyone in the cluster.
57. To isolate variability attributable to the differences between the blocks, so we can see the difference caused by treatments more clearly.
58. We want to be sure each treatment has mixture/variety of individuals; randomization allows us to equalize the effects of unknown/uncontrollable sources of variation
59. Systematically favoring certain outcomes
60. voluntary response bias, undercoverage bias, poor wording bias
61. They are neither disjoint nor independent.
62. 0.18
63. 0.34
64. 0.94
65. 0.113
66. 0.0324
67. 0.6724
68. 0.2262
69. Tree diagram
70. 0.039
71. 0.454
72. 0
73. 0.295
74. 0.8676
75. e (They are all equally likely!)
76. a. false
    b. true
    c. false
    d. true
    e. false
77. μ = 0.3; σ = 0.520
78. μ = 5.9; σ = 1.559
79. a) table
    b) 0.7
    c) 0.5
    d) not disjoint, 2 is in both events
    e) not independent (prove)
80. \$0.08; yes, you should take his bet, because you should expect to win 8 cents every bet, in the long run
81. $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$
82. $P(A|B) = P(A)$ and $P(A) * P(B) = P(A\ and\ B)$