

Show all work and justify all answers. Always specify units. Avoid "fake precision" (report no more than 3 significant digits). Write clearly --- neatness counts. I will ignore anything I cannot read (or find). Closed books, no notes; calculators allowed.

I expect you to exhibit a level of individual academic integrity that is commensurate with being a part of the Honors College; please acknowledge that integrity by signing the honor statement at the end of the test.

1. The prices paid for haircuts by students in our class are given below, in dollars, omitting non-responses and free haircuts. Note: this problem can be completed without the help of technology! [15 points]

10 12 12 13 14 15 15 15 15 17 20 25 30 30 50

- 1) Compute the five-number summary:

Min	Q1	Median	Q3	Max

- 2) Depict the distribution of the 15 haircut prices by a stem-plot (describe stems and leaves; split stems) or by a histogram (describe the number and the width of the bins).

- 3) Describe the distribution in words (address its shape, symmetry, skew, and outliers). Explain its shape.

- 4) Is the \$50 haircut an outlier, based on your picture of the distribution? Explain your reason.

- 5) Is the \$50 haircut an outlier, based on the 1.5-IQR rule? Show your computations.

- 6) How expensive is the average haircut: *at*, *above*, or *below* the median reported in (a)? Why do you think so?

2. Swedish sport medics studied arthritis among soccer players of various proficiency levels. They suspected that the more serious soccer players have higher risk of arthritis later in life. Does the data below support their suspicion? Explain. [5]

	Did not play	Local level	Elite level
Have arthritis	24	9	10
No arthritis	548	206	61

3. The height of young people in the US is distributed Normally, with mean 64 inches and standard deviation 2.7 inches for women and mean 69.3 and standard deviation 2.8 for men. [15]
- 1) Circle your gender (Male or Female) and enter your own height (in inches) _____.
 - 2) How many standard deviations (with sign) are you away from the mean of your gender?

 - 3) What fraction of your gender is taller than you? Describe your computation.

 - 4) The top 2.5% of your gender are taller than _____ inches (round to 0.1).

 - 5) The top 10% of your gender are taller than _____ inches (round to 0.1).
4. Even though heights of people of each gender are distributed Normally, weights are not. In fact, the distribution of weights is skewed to one side. To which side? Carefully explain why. [5]
5. Black and white: are juries racially biased? The sentences of 326 convicted murderers are summarized in the table below as fractions of death sentences (number of death sentences per total number of sentences) in each category. [10]
Argue *for* or *against* each of the claims below. Base your argument *only on the data below* and not on your own beliefs.

Death / Convicted	Black victim	White victim	Total
Black convict	6 / 103	11 / 63	
White convict	0 / 9	19 / 151	

- 1) Claim: Interracial violence is predominant.

- 2) Claim: Blacks are sentenced to death more often than whites (hint: look at each convict/victim combination).

- 3) Claim: Whites are sentenced to death more often than blacks (hint: look at the convict totals).

The paradox in the last two claims is know in statistics as _____.

6. After analyzing data about new cars, a student concluded that engine power (in horse powers) is a good predictor for the new car sticker price (in US dollars) and computed the following least-squares regression line: [15]

$$price = 150 (power) - 1000.$$

- 1) Fill in the blanks with numbers and units:
 The intercept is _____. The slope is _____.
- 2) Predict the price of a 100 hp compact car.
- 3) Explain the meaning of the number 150 in the regression equation, in terms of dollars and horse powers.
- 4) The average car price was \$20,000. What was the average engine power (in horse powers)?
- 5) If the engine power is measured in kilowatts instead (1 hp = 0.75 kW), how would the following change:
 - i. Intercept
 - ii. Slope
 - iii. Correlation coefficient
- 6) The student's calculator reported $r^2 = 0.64$. This means that (circle **all** that apply):
 - i. the regression line fits the data points fairly well, since 0.80 is close to 1.
 - ii. the regression line does not fit the data points well, since 0.64 is not close to 1.
 - iii. 64% of all prices are accurately predicted by the regression line; 36% of the prices do not fit the line.
 - iv. 64% of the price variation is explained by the engine power; the other 36% is due to other factors.

7. Pollutants in drinking water are often measured indirectly: an added chemical reacts with the pollutant to produce a dye whose optical absorbance (color intensity) is measured in a spectroscope. Beer's law in chemistry states that the absorbance is proportional to the amount of dye --- the darker the dye, the larger the amount of pollutant. [10]

- 1) Sketch a scatter plot of the following measurements:

Pollutant, mg	100	200	400	800	2000
Absorbance	13	24	47	93	230

- 2) Find the least-squares regression line, the correlation coefficient r , and its square.
- 3) How well do the measurements above fit the theory? Quantify your answer.
- 4) Is the last measurement an outlier of the correlation? Why?
- 5) Is the last measurement influential, that is, if it changed, would the line change very much? Why?

Show all work and justify all answers. Always specify units. Avoid "fake precision" (report no more than 3 significant digits). Write clearly --- neatness counts. I will ignore anything I cannot read (or find). Closed books, no notes; calculators allowed.

I expect you to exhibit a level of individual academic integrity that is commensurate with being a part of the Honors College; please acknowledge that integrity by signing the honor statement at the end of the test.

1. [5 points] Give an example (with complete description of all conditions and parameters) of:
 - a. a survey question that might bias the survey results (in which direction?);

 - b. a random variable with a binomial distribution;

 - c. a *stratified* random sample (describe the sampling procedure)

 - d. and a *simple* random sample in the same situation as above;

 - e. an observational study that is *not* a random sample (label its type).

2. [5] Suppose that X is a random variable that can take any decimal value between 0 and 5. All values are equally likely.
 - a. Fill in the blanks: the random variable X has a _____ distribution.

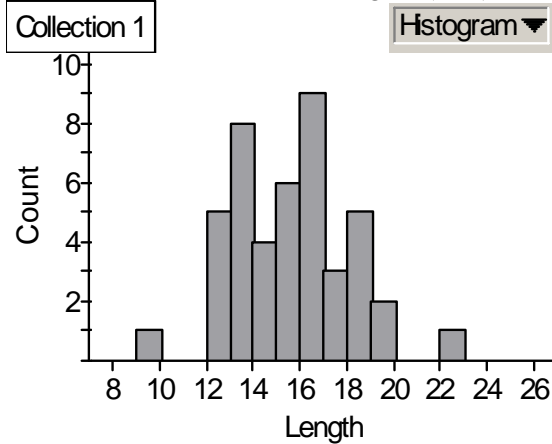
 - b. Sketch and label the probability density curve for X .

 - c. Describe the shape of the density curve above in words, including its maximal height.

 - d. The probability that X is greater than 3 is _____. Shade this area on you sketch above.

 - e. Excel and TI-83 have a function **rand()** that returns a random number between 0 and 1. Explain how you can use this function to generate random samples of X (between 0 and 5).

3. [20] The distribution of the lengths (in ft) of several great white sharks is pictured and summarized below.



Collection 1	Summary Table
↓	⇒ Length
	15.586364
	44
	15.75
	2.5499285
	3.65
	S1 = mean ()
	S2 = count ()
	S3 = median ()
	S4 = stdDev ()
	S5 = iqr ()

- How many sharks are in the sample?
- Is it reasonable to assume that the length of *all* white sharks has Normal distribution? Explain.
- Describe the *sample* distribution of the shark lengths. Is it likely that this is a small sample from a normal population? Explain your judgment.
- Your roommate, who has taken courses in Statistics and Marine Biology, says: "Oh, I remember well that the standard deviation for the lengths of all great white sharks was $\sigma = 3$ feet, but I cannot remember their mean length; it was something like $\mu = 18$ feet, but I am not sure." Give a 95% confidence interval for the mean length of great white sharks (based on the sample). Describe your result in a complete sentence with simple words to somebody who does not know statistics.
- How many (randomly selected) white sharks do you need to measure in order to estimate the mean white shark length with the same confidence (95%) and a margin of error less than 0.5 feet?
- Based on your 95% confidence interval, is there significant statistical evidence (at the 5% level) to reject your roommate's guess (that great white sharks average 18 feet in length)? Explain briefly.
- (Extra credit) You decide to test your roommate's hypothesis at the 1% level. What is the power of your test against the alternative hypothesis that the average great white shark is 16 feet long? Sketch.

4. [10] Do cancer support groups help women with breast cancer? In a recent study (2001), the researchers assigned 235 breast cancer patients randomly to two groups: 158 to "expressive group therapy" and 77 to a control group. The survival time and pain level were recorded for each patient. The study found that women in the treatment group had less pain ($P=0.04$) but insignificant increase of survival time ($P=0.72$).
- Describe the type of this study.
 - Describe the factors, levels, treatments, and response variables in the study.
 - Formulate the null hypothesis for survival time.
 - Formulate the alternative hypothesis for pain.
 - Explain in simple words how $P=0.04$ is *significant evidence* that support groups do alleviate pain.
 - Explain in simple words how $P=0.72$ means there is *no evidence* that support groups prolong life.
5. [10] Only 0.0004 of young women (20-29 years old) have breast cancer. The *diagnostic sensitivity* (fraction of positive results when ill) of a mammogram is 0.8 and its *diagnostic specificity* (fraction of negative results when healthy) is 0.9. Compute the probability of each event below for a randomly chosen young woman:
- Has cancer and gets a positive mammogram result
 - Has no cancer, but gets a positive mammogram result
 - Gets a positive mammogram result
 - Has cancer, provided she got a positive mammogram result (hint: use conditional prob. formula)
 - Based on your computations above, decide whether tests with small diagnostic power (sensitivity and specificity) are very useful for diagnosing rare diseases.

7. [15] Do antipyretics (temperature-reducing drugs) work? In a randomized comparative experiment, the mean temperature reduction of a new drug was in the 95% confidence interval between 0.2°C and 0.4°C.
- What was the average temperature reduction in the sample? Include units.
 - What was the margin of error?
 - Is the 99.5% confidence interval *narrower* or *wider* than the 95% interval? (Circle one.)
 - Describe two different ways to increase the confidence level from 95% to 99.5% (fill in the blank and cross the wrong alternative in each case).
 - Increase/decrease the margin of error to _____, keeping the same sample size. Explain your computations.
 - Increase/decrease the sample size _____ times, keeping the same margin of error. Explain your computations.
 - Formulate the null hypothesis in the study (in symbols and in words).
 - Formulate the alternative hypothesis in the study (in symbols and in words).
 - Conclude (at the 5% level) whether the drug reduces temperature. Explain.
 - Conclude (at the 10% level) whether the drug reduces temperature.
 - (Extra credit) Conclude (at the 0.5% level) whether the drug reduces temperature.
 - In fact, there is *significant* ($P < 0.01$) *statistical* evidence that the drug is better than placebo and indeed reduces temperature. Explain why this fact may be of almost *no practical significance*.

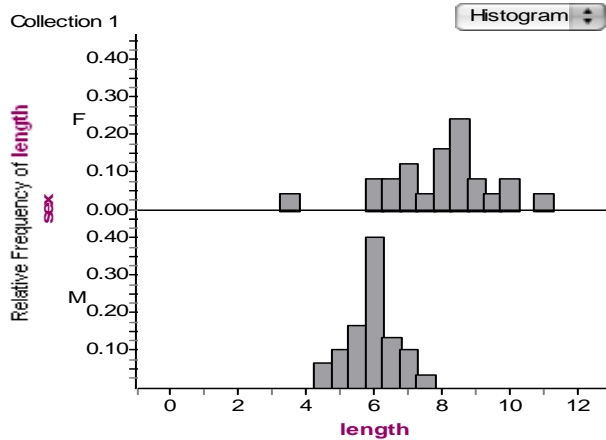
I have abided by the Honor Code in completing this test. _____

Signature

All computations must be reproducible; report the technology and method used. Every computation must result in a clearly stated conclusion (in terms of the problem). Closed books. Approved notes, calculators allowed. Write clearly and concisely.

I expect you to exhibit a level of individual academic integrity that is commensurate with being a part of the Honors College; please acknowledge that integrity by signing the honor statement at the end of the test.

8. [20 points] A biologist wants to compare the lengths (in mm) of male and female green lynx spiders. Summary statistics and plots for a sample of male and female spiders are shown below.



Spiders		length
sex	F	7.874
		25
		1.5461727
sex	M	5.9166667
		30
		0.66323831
Column Summary		6.8063636
		55
		1.5053932

S1 = mean ()
 S2 = count ()
 S3 = stdDev ()

- Without looking at the histograms, explain why it is reasonable to expect that length is Normally distributed among green lynx spiders (in each gender).
- Which gender appears to have more predictable length? Justify your reason.
- Look at the data. Is it likely that the samples come from Normal distributions? Justify your reason.
- State the technical conditions for statistical inference in this case and decide if each is fulfilled.
- Compute a 95% confidence interval for the gender difference in spider lengths and state your result.

9. [15 points] In each of the following situations, specify both the **parameter of interest** and the type of the appropriate **inference procedure**, as an abbreviated code and in full words, for example:

Mean shark length

M 1 C

Population **Mean**, **One**-sample **Confidence Interval**, or

Prop. of married, by sex

P 2 T

Population **Proportion**, **Two**-sample **Significance Test**.

- a. How strong is the support for legalizing marijuana in an upcoming election?
 - b. Do men cheat more than women on exams?
 - c. How much older are husbands than their wives?
 - d. Do trick-or-treaters prefer candy or toys?
 - e. Are women hotter than men (in terms of body temperature)?
10. [15 points] In a very close election race between two opponents (such as the 2000 Bush-Gore race) you decide to poll 1000 randomly selected voters about their preference and report a 95% confidence interval.
- a. Approximately how many interviewed voters do you expect to support the incumbent?
 - b. What parameter are you trying to estimate from the poll data?
 - c. Can you trust the inference results? Why?
 - d. What is the margin of error of your estimate?
 - e. Since it is a close race, your boss wants to decrease the margin of error to 0.01 (i.e., 1 percentage point). How many answers do you need to collect now? Explain your computation.

11. [35 points] A random sample of 12,178 Europeans and 863 Americans were asked about the use of genetically modified foods (GMF); 64% of Europeans and 52% of Americans were skeptical of GMF.
- How many people were skeptical in each continent? _____ Europeans, _____ Americans.
 - Explain why it makes no sense to judge GMF skepticism by comparing your answers above. What should be compared instead?
 - Without computing any scores and performing formal tests, explain why there is overwhelming statistical evidence that Europeans are much more skeptical than Americans about GMF.
 - The attitude towards GMF on which continent can be estimated more precisely? Why?
 - Explain why the "Plus Four" method, which usually improves inference about proportions, will not make a dramatic difference in this situation.
 - Estimate the percentage of GMF-skeptical Europeans with 99% confidence. Clearly state your result and the computation method you used (Large Sample or Plus Four).
 - Suppose you estimate the percentage of American GMF skeptics with 99% confidence as well. Which of the two estimates, European or American, will have a better chance of being correct?
 - Estimate the difference between the percentages of GMF-skeptics in Europe and America with 99% confidence. State your result and method.
 - Having learned of the study, the local TV station displays a shocking splash screen on the evening news: *Most Americans Fear Genetically Modified Foods!* State and test this claim.

