Math 263

Exam I PRACTICE

February 5th 2015

The University of Arizona

Name: ______________________________

Answers without adequate justification will not receive full credit, including multiple choice. Include units with your answer when appropriate, and box all answers unless an answer line is provided. By signing below I am agreeing to abide by the University of Arizona academic integrity policies and that all work done on this test is my own.

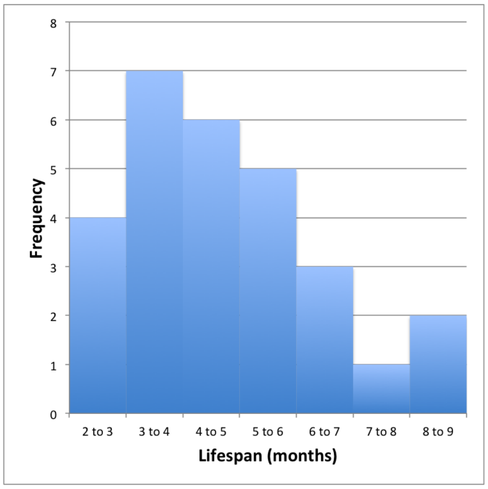
Signature: ______________________________

Tips for Success:

- Look through the entire test before starting to prioritize questions.
- If you get stuck on a question, move on and come back to it later.
- Do a quick reality check after each question: does my answer make sense? Did I include units? Did I show all my work?
- Read over the entire test at the end to make sure you didn't miss anything.
- For each question: take a deep breath, think slowly and deliberately at first, then work quickly once you see what to do.

Special note for 263 exams: if you use a calculator to compute statistical quantities such as a 5 number summary, linear regression, normal proportion etc, please indicate which function you used i.e. 1-Var Stats, LinReg(ax+b), normalcdf().

1. The life span of 28 ant colonies is displayed in the histogram below:



- How many ant colonies survived less than 3 months? **The histogram shows 4 ant colonies lived between 2 and 3 months.**
- Approximately what was the median lifespan of the ant colonies? **Since there are 28 total ant colonies, we expect the median to be between 4 and 5 (about 14 observations to the left and right).**
- Based on the histogram, which do you expect to be higher, the mean or the median? **The histogram is skewed slightly right, so we expect the mean to be larger than the median**

Suppose now that the ant colony life span is normally distributed with a mean of 5 months and a standard deviation of 1 month. (Is this reasonable given the empirical distribution?)

- What proportion of ant colonies survive between 3 and 6 months? Find the z-scores, then draw a sketch on a standard normal before finding the value. **To find the proportion that survived between 3 and 6 months, we use z-scores and cumulative proportion table. The z scores for 3 and 6 months are:**

$$z_3 = \frac{3-5}{1} = -2, \quad z_6 = \frac{6-5}{1} = 1$$

Thus we compute $F(1) - F(-2)$ where $F(z)$ is the cumulative proportion function for the standard normal (i.e. the z-table values). We have

$$F(1) - F(-2) \approx 0.8413 - 0.0228 = \boxed{0.8185}$$

On the TI84, you can do this by

normalcdf(min,max,mean,standard deviation) = normalcdf(3,6,5,1)

- What is the lifespan of an ant colony at the 80th percentile? **Here, we want to find the z-value such that 80% of the population is below that z-score. To do this, we evaluate $F^{-1}(0.8)$ where F^{-1} is the 'inverse cumulative proportion' function. This can be done on the table by looking for a value near 0.8, for instance $z = 0.84$ has $F(z) = 0.7995$. This is a z score, however, so we need to convert it to colony lifespan:**

$$0.84 = \frac{x-5}{1} \Rightarrow \boxed{x = 5.84}$$

2. (Mult. Choice) When using a histogram to display categorical values, you should make sure the categories are in alphabetical order.

   - True - Histograms are not useful if the categories are not in order.
   - True - Histograms can be used on any type of data.
   - False - You cannot use histograms to display categorical data.
   - False - The categories can be in any order when displaying categorical data.

   This is False because you cannot use histograms to display categorical data, only quantitative data.

3. For describing the distribution of a set of data, when is the five-number summary preferred over the mean and standard deviation?

   (a) When the distribution is reasonably symmetric.

   (b) When the distribution has little skewness and there are no outliers.

   (c) If the data are provided in increasing order of magnitude.

   (d) If the data exhibit skewness and there are strong outliers.

   (e) Never, because the mean and standard deviation are always more reliable.

   We typically prefer the 5 number summary when the distribution exhibits skewness and has outliers. For symmetric or near symmetric distributions with no outliers, we prefer mean and std dev.

4. Suppose the duration of a particular hand surgery is normally distributed with a mean of 75 minutes and a standard deviation of 15 minutes.

   (a) What proportion of surgeries last longer than 90 minutes? Sketch the region on the normal distribution below corresponding to the proportion of surgeries lasting longer than 90 minutes. Compare the answer you get using the 68-95-99.7 rule and the z-table.

   We would shade the region under the curve and to the *right* of 90 minutes. Using the 68-95-99.7 rule, we would have 68% of surgeries lasting between 60 and 90 minutes ('within 1 std dev of the mean'). Thus we have about 32% of surgeries lasting shorter than 60 and longer than 90 minutes. Since normal distributions are symmetric, we must have about 15.5% of surgeries lasting longer than 90 minutes. Alternately, we would use z scores - the z-score for 90 minutes is

   $$z = \frac{90 - 75}{15} = 1$$

   Going to the table, we have $F(1) = 0.8413$, and so about 84.13% of surgeries last *less* than 90 minutes, leaving about 15.87% lasting longer than 90 minutes. Using the TI84, we would enter

   normalcdf(min,max,mean,std dev) = normalcdf(90,1000,75,15)

   to get the same result.

   (b) Sketch the region corresponding to surgeries lasting between 50 and 60 minutes. **we would sketch the region under the curve between 50 and 60 minutes.**

   (c) Find the z-scores for surgeries of 50 and 60 minutes **the z-scores are**

   $$z_{50} = \frac{50 - 75}{15} \approx -1.66, \quad z_{60} = \frac{60 - 75}{15} = -1$$

(d) What percentage of surgeries last between 50 and 60 minutes? **Using a z-table, we would find $F(-1)$ and $F(-1.66)$. The area in question is $F(-1) - F(-1.66)$, or about**

$$F(-1) - F(-1.66) \approx 0.1587 - 0.0485 = 0.1102$$

So about $\boxed{11.02\%}$ of surgeries last between 50 and 60 minutes.

(e) Suppose you have a surgery that lasts 100 minutes. What percentage of people have a *longer* surgery than you? **We again use z-scores: $z_{100} = (100 - 75)/15 \approx 1.66$. We want the area to the *right* of $z = 1.66$, so we find $1 - F(1.66)$ resulting in**

$$1 - F(1.66) \approx 1 - 0.9515 = 0.0485$$

so about 4.85%.

(f) Suppose that the doctor tells you 'that was easy, 99% of these surgeries last longer than yours did'. How long was yours? **We use the inverse normal cdf again - we want a z score so that $F(z) = 1 - 0.99 = 0.01$. Using either the z-table or the invNorm function on the calculator, we obtain**

$$z \approx -2.326$$

Converting this z score to a surgery time, ('un-standardizing') we have

$$-2.326 = \frac{x - 75}{15} \Rightarrow x = -2.326 \cdot 15 + 75 = 40.11$$

So our surgery took about $\boxed{40.11\textbf{min}}$.

5. True or false:

(a) When analyzing correlation, the explanatory variable and response variable can be interchanged and the correlation coefficient will remain the same. **True - looking at the formula for correlation, we see that if we interchange the roles of X and Y, the formula is no different. This is because $z_x z_y = z_y z_x$.**

(b) When analyzing regression, the explanatory variable and response variable can be interchanged and the least-squares regression line will remain the same. **False - the slope and vertical intercept of a line will change if the roles of X and Y are swapped.**

6. The government wants to investigate average weekly household spending on tobacco products, T, and average household spending on alcohol, A for a group of domestic offenders. The data is below:

Offender ID	Alcohol ($)	Tobacco ($)
1	0	30
2	15	25
3	30	45
4	40	85
5	85	75
6	95	70
7	120	70
8	130	65
9	150	20

   (a) Use average household spending on alcohol as the explanatory variable and find the equation of the least-squares regression line. **Using a TI84 and LinReg(ax+b), I get**

   $$\textbf{Tobacco} = \textbf{0.0915 (Alcohol)} + \textbf{47.125}$$

   (b) What does your regression line predict the average weekly spending on tobacco products will be for someone who spends an average of $100 per week on alcohol? **It predicts a spending of about $56.279 per week.**

   (c) Is it reasonable to conclude from this data that there is a causal relationship between the data? **Not really - the R^2 goodness of fit value is only 0.04!**

   (d) Comment on any / all of the following. Outliers, influential data, strength of correlation, etc etc etc. **There is a possible outlier at (150,20), but even without this data point the correlation is pretty low.**

7. The thorax lengths of a population of fruit flies follow a normal distribution with mean $\mu = 0.800$mm and a standard deviation of $\sigma = 0.078$ mm.

   (a) What percent of flies have thorax lengths smaller than 1mm? **Use z-scores or the normalcdf function on the calculator to arrive at about 99.48%.**

   (b) What percent of flies have thorax length greater than the mean? **50% - normal distributions are symmetric, so the mean is the same as the median.**

   (c) What thorax lengths make up the bottom 15% of all thorax lengths? **Use invNorm(0.15,0.8,0.078) on the calculator (or inverse z-table method) to arrive at a length of about 0.719mm**

   (d) Suppose we collect 600 fruit flies. About how many of these do we expect to have thorax lengths greater than 1mm? **We can use normalcdf(1,1000,0.8,0.078) to arrive at a *proportion* of about 0.00517. This means that we will get about 600*0.00517 = 3.1032, or about 3 flies.**

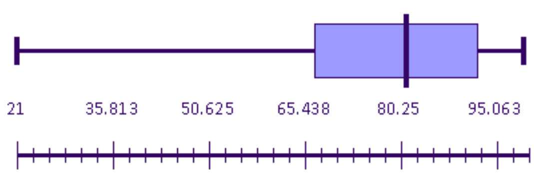
8. The following 20 test scores were recorded on an exam:

   21, 29, 48, 55, 67, 70, 71, 75, 79, 81, 81, 86, 89, 90, 90, 92, 97, 98, 99, 100

   (a) Find the mean and the 5-number summary for the data and draw a box-and-whiskers plot, including the mean **Enter the data in your calculator and use 1-Var Stats to obtain**

   $$\bar{x} = \mathbf{75.9,\ min = 21,\ Q1 = 68.5,\ med = 81,\ Q3 = 91,\ max = 100}$$

   The box and whiskers plot looks like this:



   (It should have a little mark to indicate the mean as well!)

   (b) Based on your box-and-whiskers plot, do you think the data is skewed? Why? **Yes - the mean is to the left of the median, and the box-and-whiskers is centered to the right. The distribution is likely skewed left.**

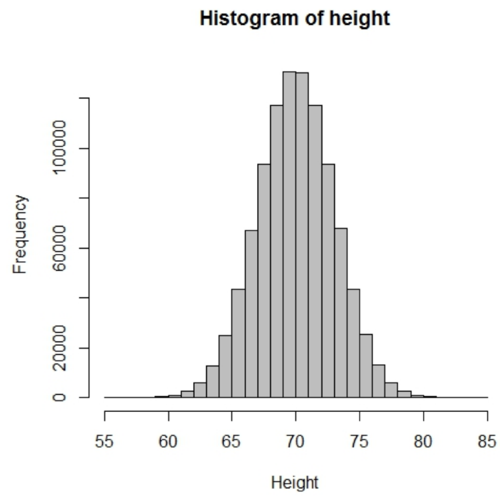
   (c) Use the 1.5IQR rule to identify possible outliers. **The IQR is 22.5, so 1.5IQR is 33.75. Thus Q1 - 1.5IQR is 34.75 and Q3+1.5IQR is 124.75. We have two data points that qualify - 21 and 29.**

   (d) Find the standard deviation for the data. **Again, we use 1-Var Stats to get** $S_x = 22.487$**.**

   (e) Compute *and interpret* the z-score for the following exam scores: 21, 55, 79, 89, 99 **Use the formula**

   $$z = \frac{x - \bar{x}}{s_x}$$

9. The following histogram shows the distribution of heights among US men:



(a) Describe the form of the distribution in height:
- One peak and is approximately symmetric.
- Two peaks and is skewed to the right
- Two peaks and is skewed to the left.
- One peak and is skewed to the right.
- One peak and is skewed to the left.

The distribution clearly has one peak and is approximately symmetric.

(b) Where is the mean location in relationship to the median? **The mean and median are the same since it is symmetric.**