Final - STAT 301
Spring 2015

Name:

UIN:

Signature:

Version A:

1. Do not open this test until told to do so.
2. This is a closed book examination, However you may use the cheatsheet provided. You should have no other printed or written material with you on the exam. But scrap paper is allowed.
3. You have 2 hours to work on this exam. There are 25 multiple choice questions.
4. On the scantron please state the version of exam that you have.
5. You may use a calculator in the exam.
6. If there is no correct answer or if multiple answers are correct, select the **best** answer.
7. If you are unsure of what a question is asking for, do not hesitate to ask the instructor or course assistant for clarification.
8. Please only give one answer per question (the one that is closest to the solution).
9. No wearing hats that can cover ones eyes.
10. Good Luck. Have a wonderful summer, it was lovely to teach you.

(1) Adults at a shopping Mall were randomly selected and asked four questions. What type of variable do we associate to each answer:

1	2	3	4
Their Height	Favourite shop	Number of items bought	How far is home

Answer:

	1	2	3	4
A	Numerical continuous	Categorical	Numerical discrete	Numerical continuous
B	Numerical continuous	Categorical	Numerical continuous	Numerical discrete
C	Numerical discrete	Categorical	Categorical	Numerical continuous
D	Numerical discrete	Numerical discrete	Numerical discrete	Numerical continuous
E	Numerical discrete	Categorical	Categorical	Binary

(2) You are given the data set: $-10, -9, 0, 3, 4, 5, 6$. The largest value in this data set, 6, is changed to a larger number (a number greater than 6). How do the summary statistics change?

(A) The median stays the same, but IQR, mean and standard deviation will get larger.

(B) The IQR stays the same, but the mean and standard deviation are likely to get lower.

(C) The IQR stays the same, the mean will get larger, and the standard deviation is likely to get larger.

(D) The median stays the same, but the 1st quartile will get lower/smaller.

(E) The median stays the same, but the 3rd quartile will get lower/smaller.

(3) The mean age of undergraduates at A&M is 20 years old with standard deviation one year. These students will be followed over 30 years. In 30 years time, what will be the mean age and standard deviation of this same set of students (converted to months)?

(A) mean $= 20 + 30 = 50$, sd $= 12 \times 1$.

(B) mean $= 20 + 30 = 50$, sd $= 1 + 1 \times 30 = 31$.

(C) mean $= 12 \times 20 + 12 \times 30 = 600$, sd $= 12 \times 1 + 12 \times 30 = 372$

(D) mean $= 12 \times 20 + 12 \times 30 = 600$, sd $= 1 + 1 \times 30 = 31$

(E) mean $= 12 \times 20 + 12 \times 30 = 600$, sd $= 12 \times 1 = 12$.

(4) Match.com, an online dating website, has spend a decade collecting and analyzing data from people who use the website. The team randomly selected 500 female users and 500 male users. They analyse the difference between attractiveness scores given to the male and female users. This is an example of

(A) Experimental Study. (B) Observational Study.

(C) Comparative Study. (D) [A] and [C]. (E) [B] and [C].

(5) Female heights are known to be normally distributed with mean 64.5 inches and standard deviation 2.5 inches. Male heights are known to be normally distributed with mean 70 inches and standard deviation 4 inches. Peter is 67 inches tall.

Using equivalent percentiles, calculate how tall Peter would be if he were female.

(A) 64.5 inches (B) 66.375 inches (C) 63.75 inches (D) 61.5 inches (E) 62.625 inches

(6-8) The length of parrots beaks is known to be normally distributed with mean $\mu = 5$ inches and standard deviation $\sigma = 2$ inches (use the normal distribution in all calculations).

(6) What is the chance that the beak of a randomly selected parrot will be between $2 - 7$ inches?

(A) 0.5 (B) 0.62 (C) 2 (D) 0.77 (E) 0.977

(7) Two parrots are randomly sampled and the average (sample mean) length of parrot beak is calculated. What is the population mean and standard error of the sample mean (the sample mean)?

(A) $\mu = 5$, se $= \frac{2}{\sqrt{2}}$ (B) μ is unknown and se $= \frac{5}{2}$ (C) $\mu = \frac{5}{2}$ and se $= \frac{2}{2}$
(D) $\mu = 5$, se $= \frac{2.5}{2}$ (E) μ is unknown and se is unknown.

(8) What is the probability the **SUM** of the two beak lengths will be less than 11 inches?

(A) 91.5% (B) 63.7% (C) 35.5% (D) 6% (E) 3%.

(9) A high-tech company wants to know how many portable devices people own. They intend to draw a random sample from the general population and use the data to construct a **95%** confidence interval for the mean number of devices. It is believed that the standard deviation is between $2 - 4$. What is the minimum number of people required to be surveyed such that the margin of error is at **most** 0.25?

(A) 246 (B) 554 (C) 984 (D) 1236 (E) 15735.

(10-12) In one project, students leaving the MRC were asked how many cell phones they owned. 17 students were surveyed, the summary statistics based on the data and a QQplot of the data is given below.

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
CellPhone	17	1.4117647	0.50735294	0.71228712	0.172755	2	2	0	2	1	2

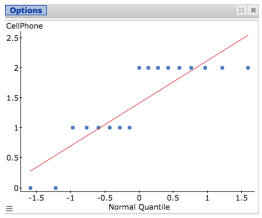


Figure 1: Sibling data

(10) Construct a **99%** confidence interval for the mean number of cell phones a person owns (remember to use the t-distribution with 16df).

(A) $[1.411 \pm 1.761 \times 0.507]$ (B) $[1.411 \pm 2.92 \times 0.1727]$ (C) $[1.411 \pm 2.145 \times 0.507]$

(D) Since the number cell phones is integer valued, the mean has to be integer values. Therefore the CI has no meaning.

(E) $[1.411 \pm 2.145 \times 0.1727]$

(11) 10 years ago the average number of cell phones that a student owned was 1.1. Is there any evidence in the data to suggest that mean number of phones owned by a student has **increased**?

(A) $H_0 : \mu \geq 1.1$ vs $H_A : \mu < 1.1$. $t = 0.43$, the p-value is greater than 5% thus there no evidence

(B) $H_0 : \mu \leq 1.1$ vs $H_A : \mu > 1.1$. $t = 1.795$, the p-value is less than 5% thus there is some evidence that mean number of cell phone has increased.

(C) $H_0 : \mu \leq 1.1$ vs $H_A : \mu > 1.1$. $t = -1.795$, the p-value is greater than 95% thus there is NO evidence that the mean number of cell phone has increased.

(D) $H_0 : \mu \geq 1.41$ vs $H_A : \mu < 1.41$. $t = 1.795$, the p-value is greater than 95% thus there no evidence that the mean number of cell phones has increased.

(E) $H_0 : \mu \leq 1.41$ vs $H_A : \mu > 1.41$. $t = -1.795$, the p-value is less 5% thus there is evidence that the mean number of cell phone has decreased.

(12) Based on the data and the QQplot, comment on the reliablity of the confidence interval (do we really have 95% confidence in it?) and p-values obtained in (10) and (11).

(A) The data is not normally distributed, however, by using the t-distribution we have corrected for the lack of non-normality, thus giving a reliable CI and p-value.

(B) The data is close to normal, thus both the CI and the p-values obtained are reliable.

(C) The data is not normally distributed, therefore with such a small sample size neither the CI or p-value are reliable.

(D) The central limit theorem means that the data will become normal for large sample sizes.

(E) (A) and (D)

(13) A 95% confidence interval for the mean height of a skyscrapper is given below.

95% confidence interval results:
μ : Mean of variable

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Height	143.69905	0.89919143	2104	141.93565	145.46245

Figure 2: Skyscrapper data

What are the conclusion of the following tests for the mean height of a skyscrapper (at the 5% level).

	$H_0 : \mu \leq 143.69$ vs $H_A : \mu > 143.69$	$H_0 : \mu = 143.69$ vs $H_A : \mu \neq 143.69$	$H_0 : \mu \geq 143.69$ vs $H_A : \mu < 143.69$
A	p-value $> 97.5\%$, cannot reject null	p-value $< 5\%$ reject null	p-value $< 2.5\%$ reject null.
B	p-value $< 2.5\%$, reject null	p-value $< 5\%$ reject null	p-value $> 97.5\%$ cannot reject null.
C	p-value $> 5\%$, cannot reject null	p-value $> 5\%$ cannot reject null	p-value $> 5\%$ cannot reject null.
D	p-value $> 97.5\%$, reject null	p-value $< 5\%$ cannot reject null	p-value $< 2.5\%$ cannot reject null.
E	p-value $< 2.5\%$, cannot reject null	p-value $< 5\%$ cannot reject null	p-value $> 97.5\%$ reject null.

(14) An expectant mother is diagnosed with gestational diabetes if her mean glucose level is **over** 140. 4 blood samples from a expectant mother is taken, they are 144.7, 142.8, 144.29 and 145.52. The sample mean is $\bar{x} = 144.33$. The obstetrician wants to see of there is any evidence of gestational diabetes in the expectant mother, based on her blood samples. The results of one particular test is given in

Hypothesis test results:
μ : Mean of variable
H_0 : μ = 140
H_A : μ ≠ 140

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Blood	144.33872	0.56502198	3	7.678843	0.0046

Figure 3: Blood sample data (Glucose)

What is the hypothesis that the obstetrician is interested in, and the result of the test at the **5% level**.

(A) $H_0 : \mu = 140$ vs $H_A : \mu \neq 140$. The p-value is 0.46%, there is NO evidence to suggest she has gestational diabetes.

(B) $H_0 : \mu \leq 140$ vs $H_A : \mu > 140$. The p-value is 0.23% there is evidence to suggest she has gestational diabetes.

(C) $H_0 : \mu \leq 140$ vs $H_A : \mu > 140$. The p-value is 0.23% there is NO evidence to suggest she has gestational diabetes.

(D) $H_0 : \mu \geq 144.38$ vs $H_A : \mu < 144.38$. The p-value is 0.23% there is evidence to suggest she has gestational diabetes.

(E) $H_0 : \mu \leq 144.38$ vs $H_A : \mu > 144.38$. The p-value is 50% there is no evidence to suggest she has gestational diabetes.

(15) If the test is done at the 5% level, what proportion of healthy patients are being falsely diagnosed with gestational diabetes?

(A) More than 95% (B) 95% (C) 50% (D) 5% (E) Less than 5%.

(16) The insurance company feels too many **healthy patients** are being falsely diagnosed with gestational diabetes (your answer to Q15 will be useful in answering this question).

(A) Increase the signifiance level from 5% to 10%

(B) Decrease the significance level from 5% to 1%

(C) Increase the number of blood samples taken.

(D) [A] and [C] (E) [B] and [C].

(Q17) Match the following three experiment with the correct test.

1. To test the efficacy of a hair increasing drug, 1000 people were randomly allocated to either the drug or placebo group (600 in the drug group and 400 in the placebo group). After treatment, thenumber people in each group who saw an increase in hair was counted.
2. To see whether the protein content in bread decreases with age, 30 loaves of breaded were baked. Immediately after baking the amount of protein was measure in each bread and then three days later the amount of protein was measured.
3. To see whether the protein content in bread decreases with age, 60 loaves of breaded were baked. They split the loaves into two groups, each of size 30. In the first group the amount of protein was measured immediately after baking. In the second group the amount of protein was measured 3 days after baking.

	[1]	[2]	[3]
(A)	Matched paired t-test	Independent Sample t-test	Matched paired t-test
(B)	Test on two proportions	Matched paired t-test	Independent sample
(C)	Independent sample	Independent Sample t-test	Matched paired t-test
(D)	Matched paired t-test	Matched paired t-test	Matched paired t-test
(E)	Test on two proportions	Independent sample	Matched paired t-test

Hypothesis test results:
μ_1 : Mean of CalciumLow
μ_2 : Mean of CalciumHigh
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 > 0$
(without pooled variances)

Difference	Sample Diff.	Std. Err.	DF	T-Stat
$\mu_1 - \mu_2$	1.991	0.62343894	17.623898	3.1935766

probability	0.3	0.15	0.10	0.05	0.025	0.01	0.005
t^*	0.53	1.06	1.33	1.73	2.1	2.55	2.88

(18-19) It is believed that a calcium rich/high diet reduces absorption of iron. 20 people were randomly assigned to two groups, one group was given a diet rich in calcium and the other was given a diet low in calcium. An independent sample t-test was done and the results are summarized below.

Critical values for a t-distribution with 17.62 degrees of freedom.

(18) Let μ_H = mean iron absorption in a calcium high diet and μ_L = the mean iron absorption in a calcium low diet. What is the test of interest and the results of the test?

(A) $H_0 : \mu_L - \mu_H \leq 0$ vs $H_A : \mu_L - \mu_H > 0$. The p-value is less than 0.5%, there is evidence that a high calcium diet reduces iron absorption.

(B) $H_0 : \mu_L - \mu_H \leq 0$ vs $H_A : \mu_L - \mu_H > 0$. The p-value is less than 0.5%, there is NO evidence a high calcium diet reduces iron absorption.

(C) $H_0 : \mu_L - \mu_H = 0$ vs $H_A : \mu_L - \mu_H \neq 0$. The p-value is less than 1%, there is NO evidence that calcium makes a difference to diet.

(D) $H_0 : \mu_L - \mu_H \leq 0$ vs $H_A : \mu_L - \mu_H > 0$. The p-value is greater than 99.5%, there is NO evidence a high calcium diet reduces iron absorption.

(E) $H_0 : \mu_L - \mu_H \leq 0$ vs $H_A : \mu_L - \mu_H > 0$. The p-value is greater than 99.5%, there is evidence a high calcium diet reduces iron absorption.

(19) Construct a **90%** confidence interval for the mean difference when using Vitamin C verses Calcium.

(A) $[0.62 \pm 1.73 \times 17.62]$ (B) $[1.99 \pm 1.73 \times \frac{0.62}{\sqrt{20}}]$ (C) $[1.99 \pm 1.73 \times \frac{0.62}{17.62}]$

(D) $[1.99 \pm 0.1 \times 0.62]$ (E) $[1.99 \pm 1.73 \times 0.62]$.

(20) 30 people were put on a weight loss program. It was found that after 20 weeks the average (sample mean) weight loss was 18 pounds.

Blue Cross Blue Shields, the health insurance company, will only pay for the weight loss program if there is **evidence to suggest** the mean weight loss is over 16 pounds. Based on the data will the insurance fund the scheme (hint: write it as a test)?

(A) There is NO evidence to suggest that mean weight loss is over 16 pounds.

(B) Since 18 is bigger than 16 of course the program should be funded. It is clear it works.

(C) It is impossible to be sure without doing a formal statistical test and calculating the p-value.

(D) [B] and [C] (E) [A] and [C].

(21) A question that obstetricians often wonder is whether **heavier twins** have a tendency of popping out **first**. Let μ_F denote the mean weight of the first twin and μ_S the mean weight of the second twin. 19 twins were sampled and weighed. The partial output of one particular test is given below (not necessarily the one we are interested in).

Hypothesis test results:
$\mu_D = \mu_1 - \mu_2$: Mean of the difference between FirstBorn and SecondB
$H_0 : \mu_D = 0$
$H_A : \mu_D \neq 0$

Difference	Sample Diff.	Std. Err.	DF	T-Stat
FirstBorn - SecondBorn	-0.18092105	0.16087062	18	-1.124637

What is the hypothesis that the obstetricians are interested in (see the claim at the start) in and the result of the test?

(A) $H_0 : \mu_F - \mu_S = 0$ vs $H_A : \mu_F - \mu_S \neq 0$. The p-value is between 20-30%, there is SOME evidence in the data to support the view that the heavier twin comes out first.

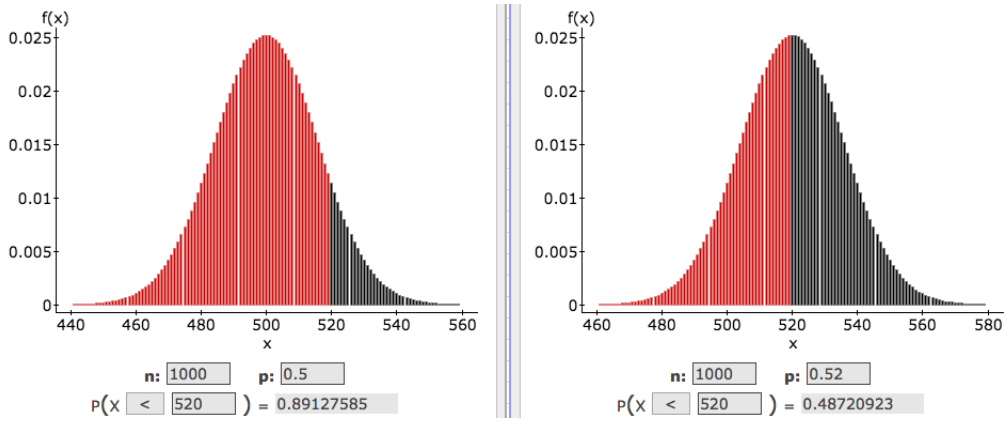
(B) $H_0 : \mu_F - \mu_S \leq 0$ vs $H_A : \mu_F - \mu_S > 0$. The p-value is between 85-90%, there is absolutely no evidence in the data to support the view that the heavier twin comes out first.

(C) $H_0 : \mu_F - \mu_S \geq 0$ vs $H_A : \mu_F - \mu_S < 0$. The p-value is between 85-90%, there is SOME evidence in the data to support the view that the heavier twin comes out first.

(D) $H_0 : \mu_F - \mu_S \geq 0$ vs $H_A : \mu_F - \mu_S < 0$. The p-value is between 10-15%, thus there is evidence to suggest the lighter twin pops out first.

(22) Last Thursday, the United Kingdom held a general election. For the purpose of this question, we will assume there are only two political parties, **Conservatives and Labour**.

In one poll taken before the election, 1000 people were surveyed, 520 (52%) said they would vote conservative (the other 480 saying would support Labour). Using the output below, is there any evidence in this poll that the Conservatives would win (do the test at the 5% level)? Let p denote the proportion of the country who will vote Conservative.



(A) $H_0 : p \leq 0.5$ vs. $H_A : p > 0.5$, the p-value is 10.9%, there is NO evidence that the Conservatives will win.

(B) $H_0 : p \leq 0.52$ vs. $H_A : p > 0.52$, the p-value is 52%, there is evidence that the Conservatives will win.

(C) $H_0 : p \leq 0.5$ vs. $H_A : p > 0.5$, the p-value is 89.1%, there is absolutely no evidence the Conservatives will win.

(D) There is no evidence in the data to suggest that Labour will win.

(E) [A] and [D]

(23-24) Castaneda v Partida is an important court case in which statistical methods were used as part of the legal argument.

In Castaneda the plantiffs alleged that the method for selecting juries in a county in Texas is biased against Mexican Americans. For the period of time of issue, there were 181,000 persons eligible for jury duty, of whom 143,000 were Mexican Americans and 38,000 were non-Mexican American. Of the 870 people selected for jury duty, 340 were Mexican American and 530 were non-Mexican American.

(23) Suppose we want to see whether **there is a bias against** selecting Mexican Americans. Let p_1 denote the proportion of Mexican Americans who are selected in the jury and p_2 the proportion of non-Mexican Americans who are selected in the jury. Which statement(s) are correct.

(A) The hypothesis of interest is $H_0 : p_1 - p_2 \geq 0$ vs $H_A : p_1 - p_2 < 0$.

(B) The hypothesis of interest is $H_0 : p_1 - p_2 = 0$ vs $H_A : p_1 - p_2 \neq 0$.

(C) Since 340 Mexican Americans is not so different to 530 Mexican Americans, it is unclear whether any discrimination took place.

(D) [A] and [C] (E) [B] and [C].

(24) A two sample test on proportions was done and the output given below. What conclusions can be we draw from the output.

Hypothesis test results:
p_1 : proportion of successes for population 1
p_2 : proportion of successes for population 2
p_1 - p_2 : Difference in proportions
H_0 : p_1 - p_2 = 0
H_A : p_1 - p_2 < 0

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
p_1 - p_2	340	143000	530	38000	-0.011569746	0.000399166	-28.984798	<0.0001

Figure 4:

(A) The p-value is so small, the data strongly suggests that Mexican-Americans were under represented.

(B) The p-value is so small, the data is consistent with there being a fair representation of Mexican-Americans.

(C) The data suggests there is a dependence/association between ethnicity (being Mexican-American or not) and selection for jury duty.

(D) [A] and [C] (E) [B] and [C].

(25) The death records of 2000 people were examined. The age of death and whether they were left or right handed is summarized below in the following table. What conclusions can we draw from the table.

Contingency table results:
Rows: Death
Columns: None

Cell format
Count (Row percent) (Column percent) (Percent of total)

	Below 60	Above 60	Total
Left handed	150 (37.5%) (33.33%) (7.5%)	250 (62.5%) (16.13%) (12.5%)	400 (100%) (20%) (20%)
Right handed	300 (18.75%) (66.67%) (15%)	1300 (81.25%) (83.87%) (65%)	1600 (100%) (80%) (80%)
Total	450 (22.5%) (100%) (22.5%)	1550 (77.5%) (100%) (77.5%)	2000 (100%) (100%) (100%)

Figure 5:

(A) The proportion of left handed people who died before 60 is greater than the proportion of right handed people who died before 60 (compare 37.5% with 18.75%).

(B) The proportion of left handed people who died before 60 is less than the proportion of right handed people who died before 60 (compare 7.5% with 15%).

(C) Clearly, if you are left handed you have an increased chance of dying young.

(D) [A] and [C] (E) [B] and [C].