While you wait for the start of this test, you may fill in the FRONT AND BACK of the BUBBLE FORM and read this cover. BUT, keep these test papers face up and flat on your desk.

Instructor: Prof. Murdock
Duration: 90 minutes. You MUST STAY for at least 60 minutes
Allowed aids: A non-programmable calculator and the aid sheets provided with this test
Format: This test includes these question papers and a BUBBLE FORM. There are **35** multiple choice questions, each worth from 2 to 5 points for a total of **128** points. Point value for each question shown by [2pts], [3pts], [4pts] or [5pts]. Most questions have choices **(A) – (E)**. For questions with fewer choices, the correct answer is ALWAYS one of those offered (e.g. if the choices are **(A) – (D)**, then **(E)** is NOT a possible correct answer.)

Once the start of the test is announced, you may detach the formula sheets and statistical table (Standard Normal) from the end of this test. These question papers and the aid sheets will not be collected.

****** You MUST record your answers on the BUBBLE FORM. In ALL cases, what is (or is not) indicated on the BUBBLE FOR is your FINAL ANSWER. Marks are based SOLEY on the BUBBLE FORM, which must be completed BEFORE the end of the test is announced. ******

- On the FRONT of the BUBBLE FORM:
  - Print your 9 (or 10) digit student number in the boxes AND darken each number in the corresponding circles.
  - Print your last name and initial in the boxes AND darken each letter in the corresponding circles.
  - Fill in the upper-left region of the form.
  - ******Your FORM CODE is A. Darken the circle with the letter A.******
    - Failing to indicate your FORM CODE means that your answers will be out of sync compared to the solution key used to mark your paper. There is NO REMEDY for the resulting failing mark. It is entirely your responsibility to properly indicate your FORM CODE.
- On the BACK of the BUBBLE FORM:
  - Write in your name.
  - Sign.
  - Record your answers.
- Use a pencil and make dark solid marks that fill the bubble completely.
- Erase completely any marks you want to change.
  - Crossing out a marked box is incorrect.
- Choose the best answer for each question.
  - If more than one answer is selected, that question earns 0 points.
- For questions with numeric answers that involve rounding, ***round your final answer to be consistent with the offered choices.***
  - Use standard rounding rules. For example, 94.2649 rounds to 94.3 to the first decimal place, 94.26 to the second decimal place, and 94.265 to the third decimal place.

▸ **<u>Questions (1) – (7)</u>**: Recall the publically available data for all ON public sector employees with salaries of $100,000 or more (http://www.fin.gov.on.ca/en/publications/salarydisclosure/pssd/). Consider <u>only</u> the 42,977 employees whose names, employers, and positions match in the disclosures of 2013 and 2012 salaries. Below are STATA summaries of this subset's salaries in 2012 and 2013 and the change from 2012 to 2013. All are measured in $1000s of dollars.

```
                         salary_2012
-------------------------------------------------------------
      Percentiles      Smallest
 1%     100.3253            100
 5%     101.3999            100
10%     102.3233            100       Obs               42977
25%     107.1425       100.0002       Sum of Wgt.       42977

50%     117.5108                      Mean            130.069
                        Largest       Std. Dev.       42.1083
75%      136.566        843.095
90%     166.9642       935.2365       Variance       1773.109
95%     198.6258        1036.74       Skewness       5.350628
99%     309.1872           1720       Kurtosis       81.53036

                         salary_2013
-------------------------------------------------------------
      Percentiles      Smallest
 1%     100.5254       100.0002
 5%      102.186       100.0016
10%     103.8206        100.005       Obs               42977
25%     109.5235       100.0102       Sum of Wgt.       42977

50%     120.3702                      Mean           132.5981
                        Largest       Std. Dev.      40.33814
75%     141.0905        772.547
90%     171.5426       903.9706       Variance       1627.166
95%     203.3144        915.851       Skewness       4.987547
99%      293.576           1714       Kurtosis       83.48862

                      change_2013_2012
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -45.50845      -877.2404
 5%    -10.10197      -474.1941
10%    -4.718613       -387.802       Obs               42977
25%    -.1520386      -369.0496       Sum of Wgt.       42977

50%     2.443321                      Mean           2.529083
                        Largest       Std. Dev.      16.09937
75%     5.849045       244.8884
90%     12.09851       245.5927       Variance       259.1898
95%     18.57905        289.095       Skewness      -8.258693
99%      37.1982       318.9551       Kurtosis       319.8883
```

(1) [2pts] In <u>2012</u>, what percent had salaries below $120,000?

(A) 10% or less
(B) more than 10% but less than 25%
(C) between 25% and 50%
(D) more than 50% but less than 75%
(E) 75% or more

(2) [3pts] In 2013, what percent had salaries within one standard deviation of the mean?

(A) less than 10%
(B) about 50%
(C) about 68.3%
(D) about 75%
(E) more than 90%

(3) [2pts] In 2012, what is the interquartile range?

(A) $29,424
(B) $38,500
(C) $39,577
(D) $40,983

(4) [3pts] From 2012 to 2013, what percent had their salaries go down (in absolute dollar value)?

(A) 10% or less
(B) more than 10% but less than 25%
(C) between 25% and 50%
(D) more than 50% but less than 75%
(E) 75% or more

(5) [5pts] What is the coefficient of correlation between salaries in 2012 and 2013?

(A) 0.888
(B) 0.891
(C) 0.909
(D) 0.916
(E) 0.925

(6) [4pts] In the highly unrealistic scenario where there is no relationship between salaries in 2012 and 2013, what would be the standard deviation of the change in salaries from 2012 to 2013?

(A) $1,770
(B) $12,081
(C) $16,099
(D) $58,312
(E) $82,446

(7) [3pts] In the highly unrealistic scenario where each employee's salary in 2013 is simply $2,000 higher than in 2012, what would be the coefficient of correlation between salaries in 2012 and 2013?

(A) -1
(B) -0.5
(C) 0
(D) 0.5
(E) 1

▸ **<u>Questions (8) – (9)</u>**: Junior employees at a very large firm must pass a performance review. Those that score in the top 10% earn a bonus. Scores are Normally distributed with a mean of 69.2% and a standard deviation of 8.7%.

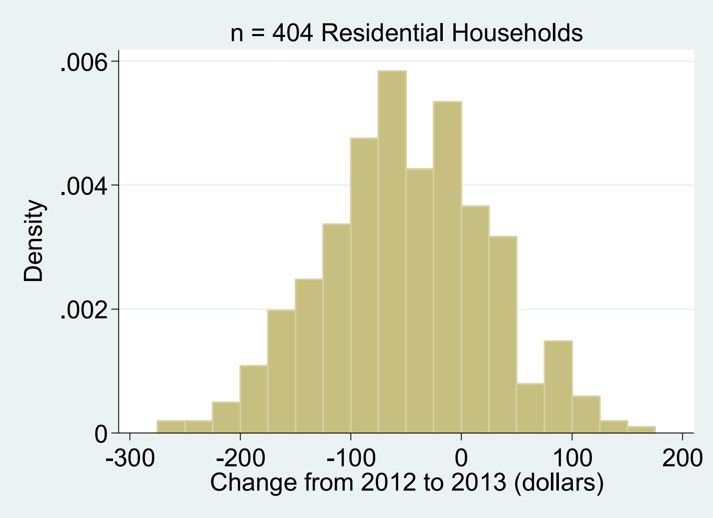
(8) [4pts] If a passing score is 50%, what percent of employees passed the performance review?

(A) less than 95%
(B) 95.4%
(C) 97.3%
(D) 98.6%
(E) more than 99.9%

(9) [4pts] To earn a bonus, how high of a score does an employee need?

(A) a score above 80.3%
(B) a score above 86.2%
(C) a score above 92.8%
(D) a score above 94.5%
(E) a score above 95.1%

▸ **<u>Questions (10) – (11)</u>**: For a random sample of 404 residential households, the histogram below describes the change in their utility bills (dollars) from 2012 to 2013.



(10) [4pts] Roughly, how many households fall in the bin with the tallest bar?

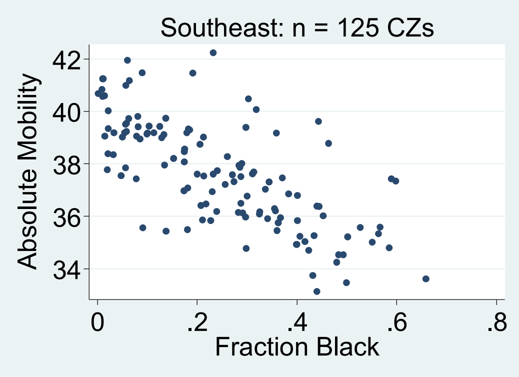
(A) less than 50
(B) between 50 and 75
(C) between 75 and 100
(D) between 100 and 125
(E) more than 125

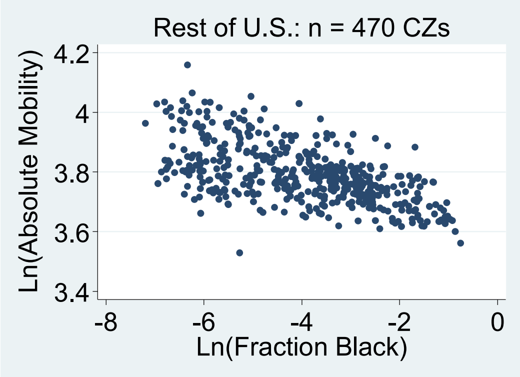
(11) [4pts] Roughly, what is the standard deviation?

(A) $25
(B) $50
(C) $75
(D) $125
(E) $150

▸ **<u>Questions (12) – (14)</u>**: Recall the *The Equality of Opportunity Project* (Chetty et al (2014)). It provides data and analysis exploring intergenerational income mobility in the U.S. The data include many variables for each of 741 "Commuting Zones" (CZ). The analysis below considers just two variables: absolute mobility and the fraction of the population that are black. The unit of observation is a CZ, which is a collection of geographically adjacent counties. In total, there are 595 CZs that have non-missing values for the two variables and have a population of at least 25,000 people. Further, the "Southeast" of the U.S. includes the states of Mississippi (MS), Alabama (AL), Georgia (GA), South Carolina (SC), Tennessee (TN), North Carolina (NC), and Florida (FL). The "Rest of the U.S." includes all states outside the Southeast.

Variable name	Description
absolute_mobility	The average percentile in the national income distribution of a child who is born to parents at the 25th percentile in the national income distribution: the higher it is, the higher mobility is
frac_black	Fraction of the population whose race is black





Southeast Regression: absolute_mobility-hat = 39.82 – 8.77*frac_black, n = 125, R^2 = 0.51

Rest of U.S. Regression: ln(absolute_mobility)-hat = 3.65 – 0.04*ln(frac_black), n = 470, R^2 = 0.35

(12) [4pts] How to interpret the slope coefficient in the Southeast Regression? Within the Southeast, geographic areas where the portion that are black is ___ higher have absolute mobility that is ___ percentiles lower on average.

(A) one percent; 8.77
(B) ten percent; 8.77
(C) ten percent; 87.7
(D) one percentage point; 8.77
(E) ten percentage points; 0.877

(13) [4pts] How to interpret the slope coefficient in the Rest of U.S. Regression? Outside of the Southeast, geographic areas where the portion that are black is ___ higher have absolute mobility that is ___ percent lower on average.

(A) one percent; 4
(B) ten percent; 0.4
(C) one percentage point; 0.04
(D) one percentage point; 4
(E) ten percentage points; 0.4

(14) [4pts] For the Southeast, if both the absolute mobility and fraction black variables are standardized, what would be the slope coefficient for a regression on the *standardized* data?

(A) -8.77
(B) -0.71
(C) -0.26
(D) 0.26
(E) 0.71

▸ **Questions (15) – (16)**: Recall the *Freakonomics* story about teachers cheating by changing students' answers on tests. Suppose that 1 percent of teachers are cheaters. Cheating-detection software spots patterns in the answer data and flags potential cheaters. It catches 98 percent of cheaters but there is a 7 percent chance an innocent teacher is flagged. Let C be the event that a teacher is a cheater. Let F be the event that the software flags a teacher.

(15) [3pts] What is $P(F \mid C')$, which also may be written as $P(F \mid C^c)$?

(A) 0.01
(B) 0.02
(C) 0.07
(D) 0.93
(E) 0.98

(16) [4pts] What is the chance that a flagged teacher is a cheater?

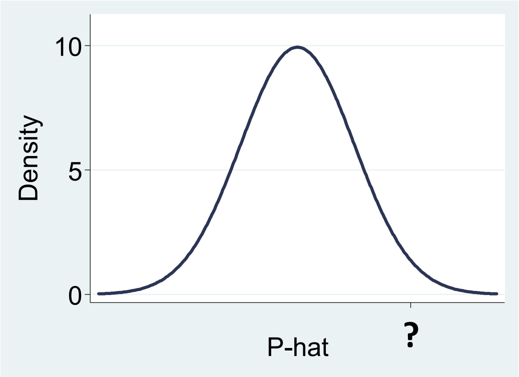
(A) 0.0098
(B) 0.1239
(C) 0.1851
(D) 0.3333
(E) 0.9800

(17) [4pts] If X is a Binomial random variable, which of these statements are TRUE?

(A) $P(X > 5 \mid n = 10, p = 0.7) = P(\hat{P} > 0.5 \mid n = 10, p = 0.7)$
(B) $P(X > 6 \mid n = 10, p = 0.7) = P(\hat{P} > 0.6 \mid n = 100, p = 0.7)$
(C) $P(X > 7 \mid n = 100, p = 0.7) = P(\hat{P} > 0.7 \mid n = 10, p = 0.7)$
(D) $P(X > 8 \mid n = 100, p = 0.7) = P(\hat{P} > 0.8 \mid n = 100, p = 0.7)$
(E) All of the above

(18) [5pts] In the incomplete graph below of the sampling distribution of $\hat{P}$ for $n = 155$ and $p = 0.5$, what value goes in the spot marked "**?**" on the x-axis?

(A) 0.58
(B) 0.63
(C) 0.66
(D) 0.68
(E) 0.75



▸ **Questions (19) – (23)**: A manager has many files to be reviewed by analysts. From experience, the manger knows that:

- 1/3 of the files are trivial and require 1 hour each
- 1/2 of the files are standard and require 4 hours each
- 1/6 of the files are highly complex and require 16 hours each

However, the type of file is only known once an analyst reviews it. The manager randomly assign files to analysts. Time to review is independent across files and they are in random order.

(19) [2pts] If an analyst receives 100 files, what is the expected total hours?

(A) 300 hours
(B) 350 hours
(C) 400 hours
(D) 450 hours
(E) 500 hours

(20) [5pts] If an analyst receives 100 files, what is the standard deviation of total hours?

(A) 5.1 hours
(B) 7.1 hours
(C) 51 hours
(D) 510 hours
(E) 710 hours

(21) [3pts] For the first five files, what is the chance that *none are trivial*?

(A) 0.0041
(B) 0.0188
(C) 0.0671
(D) 0.1044
(E) 0.1317

(22) [3pts] For the first five files, what is the chance that *one of them is highly complex*?

(A) 0.0804
(B) 0.1667
(C) 0.3312
(D) 0.4019
(E) 0.8333

(23) [4pts] For the first five files, what is the chance that *two or more are highly complex*?

(A) 0.1550
(B) 0.1633
(C) 0.1776
(D) 0.1881
(E) 0.1962

▸ **Questions (24) – (28)**: Consider a 20 die-roll Monte Carlo simulation. In each simulation draw 20 fair dice are rolled and the sample mean and sample median are computed. 500,000 simulation draws are used. Below are STATA summaries of the simulation results.

```
                     Sample Mean (X-bar)
--------------------------------------------------------------
      Percentiles      Smallest
 1%          2.6           1.75
 5%         2.85           1.85
10%            3            1.9       Obs              500000
25%         3.25            1.9       Sum of Wgt.      500000

50%          3.5                      Mean           3.500223
                        Largest       Std. Dev.      .3818714
75%         3.75           5.05
90%            4            5.1       Variance       .1458258
95%         4.15            5.1       Skewness      -.0009758
99%          4.4            5.1       Kurtosis       2.937278
```

```
                        Sample Median
--------------------------------------------------------------
      Percentiles      Smallest
 1%            2              1
 5%          2.5              1
10%            3              1       Obs              500000
25%            3              1       Sum of Wgt.      500000

50%          3.5                      Mean           3.499456
                        Largest       Std. Dev.      .6615879
75%            4              6
90%            4              6       Variance       .4376986
95%          4.5              6       Skewness       .0046752
99%            5              6       Kurtosis       2.829294
```

(24) [3pts] The simulation is not necessary to find the standard deviation of the sample mean. Using theory, when 20 fair dice are rolled, what is the s.d. of the sample mean?

(A) $\sqrt{\frac{(6-1)^2}{12}}$

(B) $\sqrt{\frac{(1-3.5)^2}{6}+\frac{(2-3.5)^2}{6}+\frac{(3-3.5)^2}{6}+\frac{(4-3.5)^2}{6}+\frac{(5-3.5)^2}{6}+\frac{(6-3.5)^2}{6}}$

(C) $\sqrt{\frac{(1-3.5)^2}{20*6}+\frac{(2-3.5)^2}{20*6}+\frac{(3-3.5)^2}{20*6}+\frac{(4-3.5)^2}{20*6}+\frac{(5-3.5)^2}{20*6}+\frac{(6-3.5)^2}{20*6}}$

(D) $\sqrt{\frac{(1-3.5)^2}{\sqrt{20}*6}+\frac{(2-3.5)^2}{\sqrt{20}*6}+\frac{(3-3.5)^2}{\sqrt{20}*6}+\frac{(4-3.5)^2}{\sqrt{20}*6}+\frac{(5-3.5)^2}{\sqrt{20}*6}+\frac{(6-3.5)^2}{\sqrt{20}*6}}$

(25) [3pts] Why is there a discrepancy between the theoretical s.d. of the sample mean and the simulation results?

(A) simulation error
(B) the extreme skew of the population
(C) the presence of outliers in the population
(D) the sample size ($n = 20$) is not sufficiently large
(E) All of the above

(26) [4pts] In a roll of 20 dice, which of the following is the LEAST plausible result?

(A) a sample mean less than 3
(B) a sample median less than 3
(C) a sample mean greater than 4.5
(D) a sample median greater than 4.5
(E) a sample median exactly equal to 3

(27) [4pts] If the Monte Carlo simulation is repeated but this time each sample has 100 observations (i.e. 100 dice rolled) instead of 20, what should you expect as the value of the 99th percentile in the new STATA summary of the simulation results for the *sample mean*? (Hint: $V[X] = \sigma^2 = 2.9167$ if X records the value on a single rolled die.)

(A) 3.74
(B) 3.78
(C) 3.82
(D) 3.86
(E) 3.90

(28) [4pts] If the Monte Carlo simulation is repeated but this time each sample has 100 observations (i.e. 100 dice rolled) instead of 20, which should you expect to INCREASE in the new STATA summary of the simulation results for the *sample median*?

(A) the 5th percentile
(B) the 50th percentile
(C) the 95th percentile
(D) the standard deviation
(E) the interquartile range

(29) [5pts] Suppose 60 percent of all unionized workers are in favor of a strike. If 132 unionized workers are randomly sampled and surveyed, what is the chance that less than half are in favor of a strike?

(A) less than 1%
(B) between 1% and 2%
(C) between 2% and 5%
(D) between 5% and 10%
(E) more than 10%

(30) [5pts] Wait time for the "Airport Rocket" bus (X) is Uniformly distributed from 0 to 20 minutes. An airport monitor records wait times for 19 randomly selected travelers. (The chosen sampling procedure ensures that the independence assumption is reasonable.) What is the chance that the sample mean wait time is longer than 13 minutes?

(A) less than 1%
(B) between 1% and 2%
(C) between 2% and 5%
(D) between 5% and 10%
(E) more than 10%

➡

(31) [3pts] The CI estimator of a proportion is found with $\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$. For a 97.5% CI, what is $z_{\alpha/2}$?

(A) 2.24
(B) 2.31
(C) 2.43
(D) 2.45
(E) 2.48

(32) [4pts] Which of these would reduce the margin of error of the CI estimator of a proportion?

(A) decrease the sample size
(B) decrease the confidence level
(C) decrease the standard deviation of p
(D) decrease the extent of non-sampling errors
(E) All of the above

▸ **Questions (33) – (35)**: A telemarketing firm has two different scripts. Both are designed to get respondents to complete a long questionnaire. To test which script is more effective, the computer prompt for each respondent randomly displays either Script 1 or 2 for a random sample of 1,123 respondents. Of the 562 respondents that heard Script 1, 18 completed the questionnaire. Of the 561 respondents that heard Script 2, 24 completed the questionnaire.

(33) [2pts] What is the point estimate of the difference between the population proportions?

(A) 0.0005
(B) 0.0108
(C) 0.0217
(D) 0.0374

(34) [4pts] With a 95% confidence level, what is the margin of error for the point estimate in the previous question?

(A) 0.0003
(B) 0.0113
(C) 0.0186
(D) 0.0222
(E) 0.0318

(35) [4pts] What does the 95% CI of the difference between proportions help make an inference about?

(A) the mean effectiveness of Scripts 1 and 2: $(\hat{P}_2 - \hat{P}_1)/2$
(B) the difference between Script 2 and Script 1 in the completion rate for the 1,123 respondents: $(p_2 - p_1)$
(C) the difference between Script 2 and Script 1 in the completion rate for the 1,123 respondents: $(\hat{P}_2 - \hat{P}_1)$
(D) the difference between Script 2 and Script 1 in the completion rate for all potential respondents: $(p_2 - p_1)$
(E) the difference between Script 2 and Script 1 in the completion rate for all potential respondents: $(\hat{P}_2 - \hat{P}_1)$

******Double-check that you darkened the circle for FORM CODE A on the front-side of your BUBBLE FORM.******

You may keep these question papers and aid sheets: we will collect only your BUBBLE FORM.