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: 110 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 with point values that may vary from 1 to 3 points each for a total of 70 points. The point value for each question is shown by [1pt], [2pts], [3pts]. 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 aid sheets and statistical table (Standard Normal) from the end of this test. These question papers and the aid sheets will not be collected. Only the BUBBLE FORM is collected.

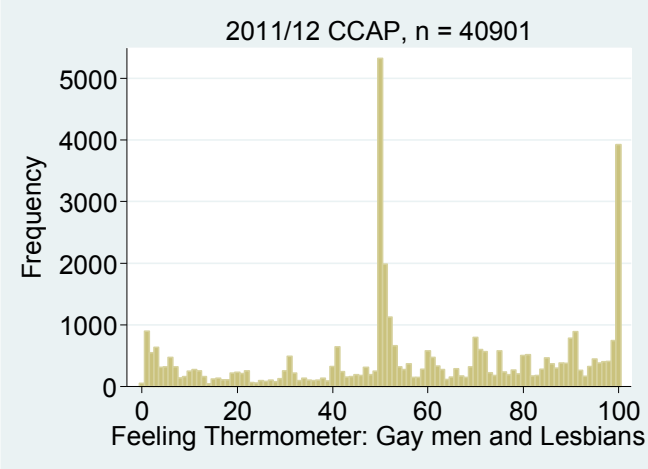
You must record your answers on the BUBBLE FORM. In ALL cases what is (or is not) indicated on the BUBBLE FORM is your FINAL ANSWER. Marks are based SOLEY on the BUBBLE FORM, which must be completed before the end of the test is announced.

- 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. It is entirely *your responsibility* to properly indicate your FORM CODE.
- On the BACK of the BUBBLE FORM:
    - Write in your name, sign, and 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 require rounding, round your final answer to be consistent with the choices offered. Use standard rounding rules.

▶ **Questions (1) – (14):** Consider data collected by YouGov for the 2012 Cooperative Campaign Analysis Project (CCAP). At the end of 2011 and repeatedly in 2012 it randomly sampled Americans. Questions included "feeling thermometers" about various demographic groups. Here is an excerpt from the instructions respondents received.

> We will now ask you to please rate a series of demographic groups. We will display the name of a group, and we would like you to rate the group using the same "feeling thermometer." Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person. Your rating will appear at the end of the slider. How would you rate _____ on a thermometer from 0 to 100. [Blank filled in with: Whites, Gay men and Lesbians, Latinos, Blacks, Mormons, Asian-Americans, Muslims.]

Michael Tesler used these data in "Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed" published in 2014 in the *American Journal of Political Science*. The replication data are posted at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26721. Here are a discrete frequency histogram and STATA summary of the feeling thermometer ratings regarding gay men and lesbians (`gaytherm`).



gaytherm

Percentiles		Smallest		
1%	1	0		
5%	3	0		
10%	10	0	Obs	40901
25%	45	0	Sum of Wgt.	40901
50%	54		Mean	58.70517
		Largest	Std. Dev.	29.36892
75%	85	100		
90%	99	100	Variance	862.5332
95%	100	100	Skewness	-.3156515
99%	100	100	Kurtosis	2.208889

(1) [2pt] For the distribution of feeling thermometer ratings, which *best* describes the shape?

(A) Uniform
(B) Multi-modal
(C) Positively skewed
(D) Negatively skewed
(E) Normal (Bell) shaped

(2) [2pts] In a *density* histogram (with the same bins), what would be the approximate height of the tallest bar?

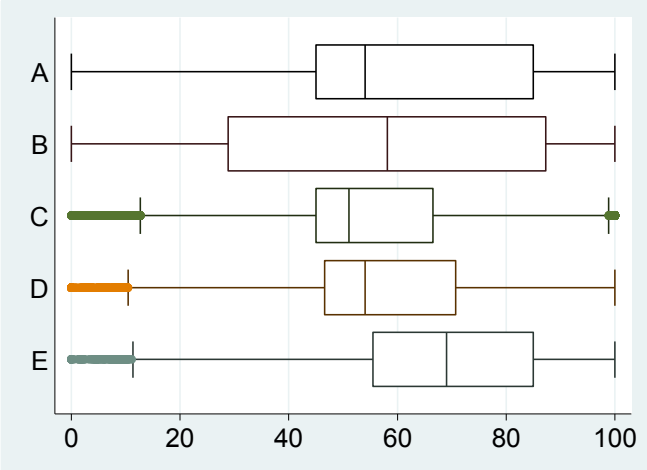
(A) 0.001
(B) 0.01
(C) 0.05
(D) 0.13
(E) 0.21

(3) [2pts] What percent of the 40,901 observations lie within two standard deviations of the mean?

(A) 0%
(B) about 89%
(C) about 95.4%
(D) about 99.9%
(E) 100%

(4) [2pts] Which of these could be the box plot of the feeling thermometer ratings variable?

(A) A
(B) B
(C) C
(D) D
(E) E



(5) [2pts] If a survey participant's *standardized* feeling thermometer is -0.74, what rating did they give?

(A) 29
(B) 31
(C) 33
(D) 35
(E) 37

(6) [2pts] Researchers have explored whether people will change their opinions after speaking with a gay person. Suppose that 100 people give a feeling thermometer (`pre`) and the mean is 60 and the s.d. is 20. After speaking with a gay person they give a feeling thermometer again (`post`) and the mean is 64 and the s.d. is 18. The variables `pre` and `post` have a coefficient of correlation of 0.9. What is the s.d. of the *change* in ratings?

(A) 2
(B) 6.44
(C) 8.72
(D) 9.83
(E) 26.91

▶ **Questions (7) – (14):** Are attitudes of younger generations different from older generations regarding gay men and lesbians? To explore this question, consider an additional variable in the CCAP data. Specifically, the data include a variable measuring year of birth.

```
                          yr_born
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1929            1917
 5%         1937            1919
10%         1943            1919       Obs                40901
25%         1953            1919       Sum of Wgt.        40901

50%         1962                       Mean           1963.144
                       Largest         Std. Dev.      15.40963
75%         1974            1994
90%         1985            1994       Variance       237.4568
95%         1989            1994       Skewness      -.0254416
99%         1992            1994       Kurtosis       2.413504
```

Consider a simple regression of the feelings thermometer on the year the respondent was born:

`gaytherm-hat = -458.04 + 0.26*yr_born; R-squared = 0.0191; n = 40,901`

Also, some questions ask you to consider an alternate measure of age: age in 2012 (in years) instead of year of birth. The variable name of this alternate measure is `age_yrs`.

(7) [2pts] How should you interpret the OLS coefficient 0.26? On average, in 2011/12, Americans who are ___.

(A) one year older have thermometer ratings that are 26% higher
(B) one year older have thermometer ratings that are 0.26% higher
(C) one year younger have thermometer ratings that are 26% higher
(D) one year younger have thermometer ratings that are 0.26 higher on a 100-point scale
(E) one year older have thermometer ratings that are 26 points higher on a 100-point scale

(8) [2pts] What is the covariance between `gaytherm` and `yr_born`? (Note: Your answer may be off a bit depending on rounding and how you solve this question. In that case, choose the closest answer.)

(A) 0.37
(B) 9
(C) 29
(D) 44
(E) 63

(9) [2pts] Based on these data, which is a correct conclusion about how attitudes of younger generations differ from older generations?

(A) There is a weak relationship: older generations are more positive towards gay people
(B) There is a weak relationship: younger generations are more positive towards gay people
(C) There is a strong relationship: older generations are more positive towards gay people
(D) There is a strong relationship: younger generations are more positive towards gay people

(10) [2pts] If you randomly selected 20 respondents out of the 40,901 and re-ran the regression, what should you expect about the R^2?

(A) It will be lower because the sample size is smaller
(B) It will be higher because the sample size is smaller
(C) It will be roughly the same but may be higher or lower given sampling error
(D) It will be lower but the smaller sample size will likely result in heteroscedasticy
(E) It will be higher but the smaller sample size will likely result in heteroscedasticy

(11) [2pts] What are the mean and standard deviation of `age_yrs`?

(A) mean = 46.2 and s.d. = 14.9
(B) mean = 46.2 and s.d. = 15.4
(C) mean = 47.1 and s.d. = 14.7
(D) mean = 48.9 and s.d. = 14.7
(E) mean = 48.9 and s.d. = 15.4

(12) [2pts] What is the coefficient of correlation between `yr_born` and `age_yrs`?

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

(13) [2pts] For a regression of `gaytherm` on `age_yrs`, which of these OLS results is closest to correct?

(A) `gaytherm-hat = 71.57 - 0.26*age_yrs`
(B) `gaytherm-hat = 458.04 - 0.26*age_yrs`
(C) `gaytherm-hat = 1553.96 + 0.26*age_yrs`
(D) `gaytherm-hat = -458.04 + 0.26*age_yrs`
(E) `gaytherm-hat = -2470.04 - 0.26*age_yrs`

(14) [2pts] 120 people are randomly divided into two groups: Group 1 (n=40) and Group 2 (n=80). Both answer the feeling thermometer question for gay men and lesbians. Group 1 then participates in a discussion led by a gay person about gay rights and Group 2 participates in a discussion about various issues (e.g. global warming, health care policy, etc.). Afterwards, each group re-does the feeling thermometer question. The researcher compares the original feeling thermometer responses to those after the discussion. Before the discussion, the s.d. of thermometer ratings is 29.9 in Group 1 and 32.3 in Group 2. Why is the s.d. of Group 2 larger than the s.d. of Group 1?

(A) because of sampling error
(B) because of non-sampling errors
(C) because the sample size of Group 2 is larger than Group 1
(D) because these are observational data and heteroscedasticity is a common issue with observational data
(E) All of the above

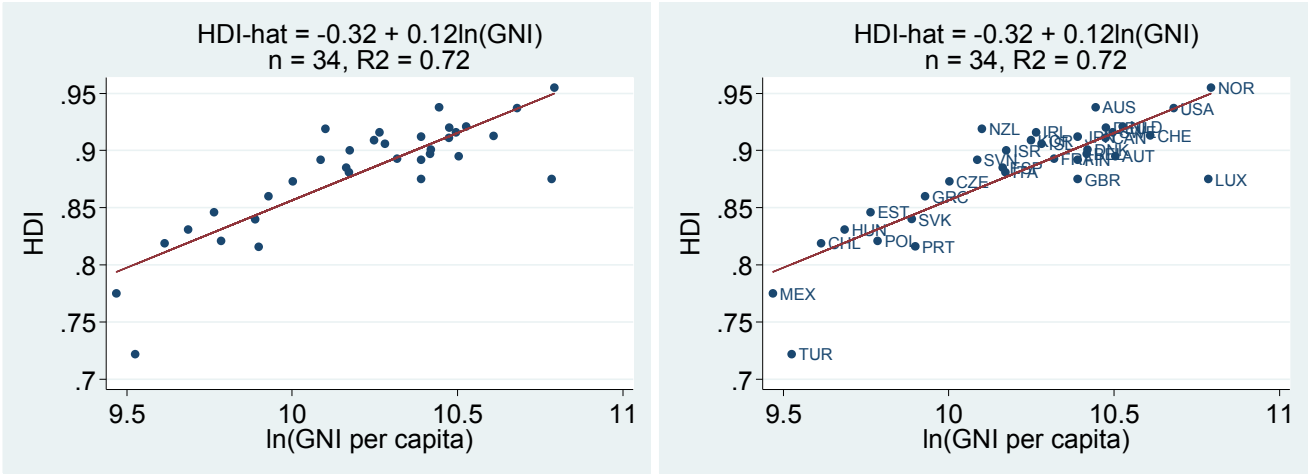
(15) [2pts] If height is Normal with mean 169.8cm and s.d. 8.3cm, how tall do you need to be to be in the top 1%?

(A) a bit taller than 188cm
(B) a bit taller than 189cm
(C) a bit taller than 190cm
(D) a bit taller than 191cm
(E) a bit taller than 192cm

(16) [2pts] If test scores are Normal with mean 65.7 and s.d. 11.2, what percent have a score between 60 and 70?

(A) 26% of test takers
(B) 30% of test takers
(C) 34% of test takers
(D) 45% of test takers
(E) 55% of test takers

▶ **Questions (17) – (22):** From the required reading "Logarithms in Regression Analysis," recall the UN data on the Human Development Index (HDI) and Gross National Income (GNI) for a cross-section of countries in 2012. The results below limit that analysis to the 34 member nations of the OECD. Data and quotes are from the UN website.



(17) [2pts] "Can GNI per capita be used to measure human development instead of the HDI?" The UN's answer:

> No. GNI per capita only reflects average national income. It tells nothing of how that income is spent, whether on universal health, education or military expenditure. Comparing rankings on GNI per capita and the HDI can reveal much about the results of national policy choices. For example, a country with a very high GNI per capita, such as Kuwait which has a relatively low mean years of schooling for its adult population, can have a lower HDI rank than, say, Czech Republic, which has less than 32% of the GNI per capita of Kuwait.

Which course concept does this quote *specifically illustrate*? (Note: All choices are real course concepts. The question asks which is specifically relevant to the UN excerpt.)

(A) Data that are not stationary have trends
(B) Even a very strong correlation does not imply causation with observational data
(C) If many variables affect y, then a plot of y versus a single x variable will show scatter
(D) Heteroskedasticity means the standard deviation of the residuals is a poor measure of fit
(E) When there are diminishing returns use a logarithm (e.g. diminishing returns to increases in the GNI)

(18) [2pts] Approximately, what is the Gross National Income (GNI) per capita of the Czech Republic (CZE)?

(A) 18,000 US$
(B) 22,000 US$
(C) 26,000 US$
(D) 30,000 US$
(E) 34,000 US$

(19) [2pts] How should you interpret the OLS coefficient 0.12? In 2012, on average, OECD member nations with GNI per capita that is ___ higher have an HDI that is approximately ___ higher.

(A) 1 dollar; 12 percent
(B) 1 dollar; 0.12 percent
(C) 1,000 dollars; 0.12 units
(D) 1 percent; 0.12 units
(E) 1 percent; 0.0012 units

(20) [2pts] Turkey (TUR) and Luxembourg (LUX) stand out a bit in these data. If the regression were rerun without these two observations, what would happen to the R^2?

(A) The R^2 would increase
(B) The R^2 would be exactly the same
(C) The R^2 would be about the same (chance differences only)
(D) The R^2 would decline because the sample size would be smaller
(E) The R^2 would decline because TUR and LUX were both parallel to the OLS line

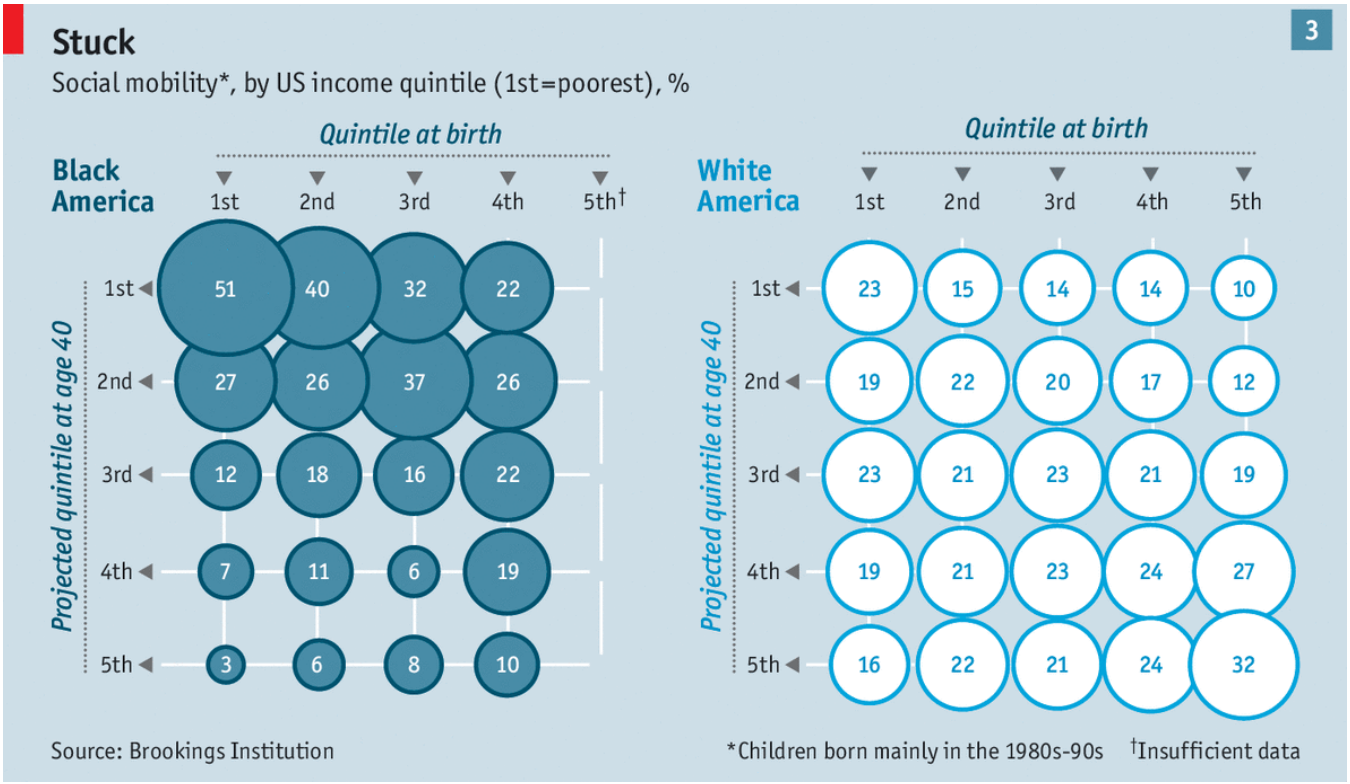
(21) [2pt] Recall this sentence from the UN quote: "For example, a country with a very high GNI per capita, such as Kuwait which has a relatively low mean years of schooling for its adult population, can have a lower HDI rank than, say, Czech Republic, which has less than 32% of the GNI per capita of Kuwait." Kuwait is not a member of OECD and hence is not included in the regression results provided. What would happen to the OLS coefficient of 0.12 if Kuwait were included in the regression?

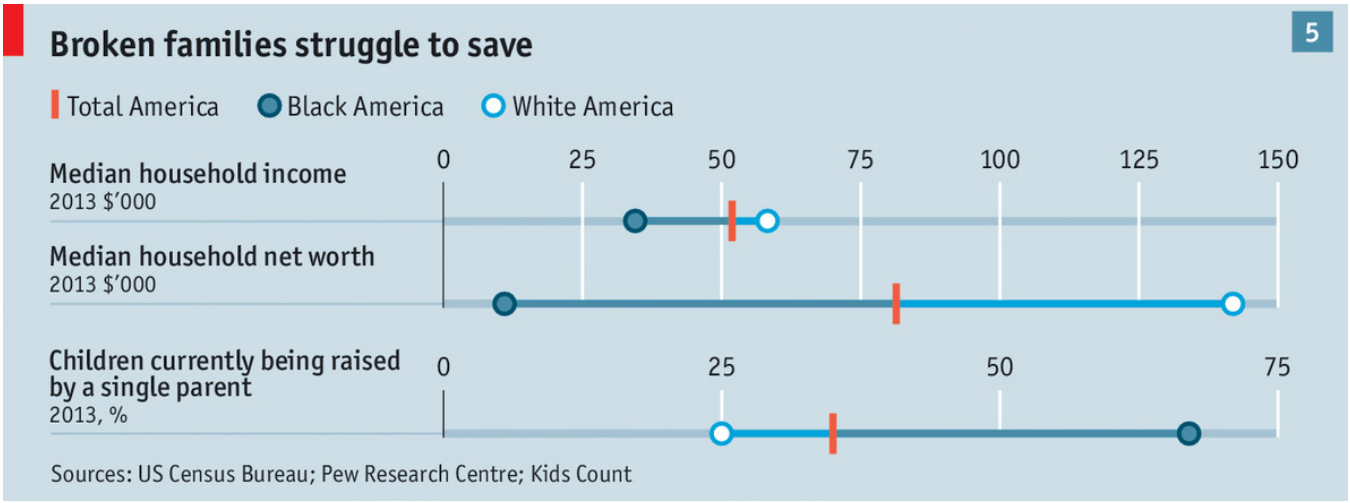
(A) the OLS coefficient of 0.12 would increase
(B) the OLS coefficient of 0.12 would decrease
(C) the OLS coefficient of 0.12 would stay exactly the same
(D) the OLS coefficient of 0.12 may go up or down because of sampling error
(E) the OLS coefficient of 0.12 would not change much at all because it is robust to outliers

(22) [2pts] The HDI is a number between 0 and 1. Given that it is an index, its scale is arbitrary. For example, it could be multiplied by 100 to yield an index that ranges between 0 and 100. If the HDI were between 0 and 100 and the regression were rerun, which of these would stay exactly the same (i.e. same value either way)?

(A) R^2
(B) SST
(C) SSR
(D) OLS slope coefficient
(E) All of the above

▶ **Questions (23) – (28):** Consider Charts 3 and 5 from a May 9, 2015 article in *The Economist* entitled "Black America: The fire and the fuel." It explores underlying factors contributing to recent riots in U.S. cities. Note: Quartiles divide data into fourths; Quintiles divide data into fifths. Hence if you are in the first income quintile that means your income is between the 0th and 20th percentiles of the entire income distribution (all races combined).





(23) [2pts] For a black American, if the event of being born into a family in the first income quintile is BQ1 and the event of being in the first income quintile at age 40 is 40Q1, then which probability (approximately) equals 0.51?

(A) $P(BQ1) \approx 0.51$
(B) $P(40Q1) \approx 0.51$
(C) $P(BQ1 \mid 40Q1) \approx 0.51$
(D) $P(40Q1 \mid BQ1) \approx 0.51$
(E) $P(BQ1 \ \& \ 40Q1) \approx 0.51$

(24) [2pts] For a white American, the biggest number in Chart 3 is 32. What does 32 mean?

(A) 32% of white Americans are in the top income quintile at birth
(B) 32% of white Americans are in the top income quintile at age 40
(C) 32% of white Americans are in the top income quintile both at birth and at age 40
(D) 32% of white Americans in the top income quintile at birth are in the top income quintile at age 40
(E) 32% of white Americans in the top income quintile at age 40 were in the top income quintile at birth

(25) [2pts] Based on Chart 3, what can we conclude about comparing black America with white America?

(A) Most Americans in the first income quintile are black, not white
(B) For both black and white Americans and for each birth quintile, the most likely income quintile at age 40 is the same as the income quintile at birth
(C) Social mobility is similar for black and white Americans: the key difference is that black Americans are more likely to be born into low income quintiles
(D) Compared to black Americans, white Americans are more likely to move up the income distribution and less likely to move down the income distribution
(E) All of the above

(26) [2pts] Why do the first two graphics in Chart 5 report medians and not means? Because the ___.

(A) median is robust to outliers and the mean is not robust
(B) income and wealth distributions are all positively skewed
(C) income and wealth distributions are all negatively skewed
(D) income and wealth distributions are likely bimodal (given the large differences across the two races)
(E) income distributions are all positively skewed and the wealth distributions are all negatively skewed

(27) [2pts] The final graphic in Chart 5 shows that 25% of white children are being raised by a single parent whereas 67% of black children are. This could be restated as: compared to white children, the percent of black children currently being raised by a single parent in 2013 is ___.

(A) 168 percent higher
(B) more than twice as big
(C) 42 percentage points higher
(D) All of the above

(28) [2pts] If you randomly selected four white children, what is the chance that two are being raised by a single parent?

(A) 0.035
(B) 0.125
(C) 0.211
(D) 0.303
(E) 0.448

▶ **Questions (29) – (31):** The next table, which uses U.S. census data for households with a husband and wife, comes from a 2014 NBER working paper "Marry Your Like: Assortative Mating and Income Inequality." (http://www.nber.org/papers/w19829) "Positive assortative mating" means that people tend to marry people with a similar education level. The table has two panels: one for 1960 and another for 2005. A person's *highest level* of education is categorized as: less than high school (HS-), high school degree (HS), some college (C-), college degree (C), or post-college degree (C+). The rows are husbands' education and columns are wives'.

Contingency Table: Marital Sorting by Education

1960:	*Wife*				
Husband	HS-	HS	C-	C	C+
HS-	0.323	0.138	0.019	0.004	0.001
HS	0.076	0.165	0.028	0.008	0.002
C-	0.018	0.051	0.027	0.008	0.002
C	0.005	0.027	0.019	0.018	0.003
C+	0.003	0.016	0.017	0.016	0.008
2005:	*Wife*				
Husband	HS-	HS	C-	C	C+
HS-	0.039	0.031	0.010	0.003	0.001
HS	0.023	0.192	0.082	0.037	0.012
C-	0.005	0.065	0.088	0.047	0.016
C	0.002	0.030	0.045	0.104	0.037
C+	0.001	0.010	0.018	0.050	0.053

(29) [2pts] In 1960, if a husband has a high school degree (HS), what is the chance his wife has a high school degree (HS)?

(A) 0.165
(B) 0.389
(C) 0.404
(D) 0.416
(E) 0.591

(30) [2pts] In 2005, if a wife has a college degree (C), what is the chance her husband has a college degree (C)?

(A) 0.104
(B) 0.389
(C) 0.432
(D) 0.477
(E) 0.485

(31) [3pts] In 2005, if husbands and wives were randomly assigned to each other – in other words, if there is no assortative mating – what number would replace 0.053 (i.e. the cell C+, C+)?

(A) 0.016
(B) 0.019
(C) 0.022
(D) 0.025
(E) 0.027

(32) [2pts] For a randomly selected Ontario teacher, variables measuring class size (number of students) and class performance (mean final grade) are recorded each year from 1995 through 2015. What kind of data are these?

(A) experimental and time series
(B) observational and time series
(C) experimental and cross-sectional
(D) observational and cross-sectional
(E) experimental and panel (longitudinal)

▶ **Questions (33) – (34):** The table below shows the relationship between whether a person was born in Canada and employment status. The table below it (with cells intentionally left blank) focuses on only those in the labor force.

Joint Probability Table: 2012, 25 – 54 Year Olds, Stats Canada				
Place of birth	Employed	Unemployed	Not in Labor Force	Total
Canada	0.6136	0.0351	0.0885	0.7373
Not Canada	0.2002	0.0169	0.0455	0.2627
Total	0.8139	0.0521	0.1341	1.0000

Joint Probability Table: 2012, 25 – 54 Year Olds in the Labor Force			
Place of birth	Employed	Unemployed	Total
Canada			**??**
Not Canada		**?**	
Total			1.0000

(33) [2pts] What value should appear in the cell marked "**?**"?

(A) 0.0137
(B) 0.0169
(C) 0.0195
(D) 0.0476
(E) 0.0643

(34) [2pts] What value should appear in the cell marked "**??**"?

(A) 0.7373
(B) 0.7388
(C) 0.7421
(D) 0.7492
(E) 0.7511

(35) [1pt] What is your FORM CODE? (It is given on Page 1 of these test papers.)

(A) My FORM CODE is A and I have marked it here *and* on the front on my bubble form
(B) My FORM CODE is B and I have marked it here *and* on the front on my bubble form
(C) My FORM CODE is C and I have marked it here *and* on the front on my bubble form
(D) My FORM CODE is D and I have marked it here *and* on the front on my bubble form