Practice Solutions for Exam

1.

(a) Quantitative discrete interval

(b) Quantitative continuous interval

(c) Qualitative discrete

(d) Qualitative discrete

(e) Quantitative continuous interval

(f) Quantitative discrete interval

(g) Qualitative discrete

2. Tabular: Using a cross-tabulation of both variables.

   Graphical: multiple bar graphs. For example, for each party show three bar graphs: one corresponding to parents and child affiliated to that party, another to parents affiliated but child not, and the third for child affiliated but parent not.

3. You can compute the mean and the standard deviation to give an idea of the variation in the distribution.

4. The higher the national debt of a country the lower its GDP is likely to be. Lower national debts are associated with higher GDP.

5. N = 1000

p = .11

Var(p) = p*(1-p)/N = (.11*.89)/1000 = .0000979 s =.0099

$[p - Z_{\alpha/2} * s\ p + Z_{\alpha/2} * s] = [.11-1.96*.0099, .11+1.96*.0099] = [.09,.13]$

6.

(a) Bar graphs for each category.

(b) The median is the category corresponding to "unfair". We know that the first half of the sample provided answers that varied from "very fair" to "unfair", while the second half finds redistribution in Brazil to be unfair or very unfair. The mode is also the "unfair" answer. From all possible answers, respondents were more likely to judge the income distribution in Brazil as unfair.

7. In order to form a 95% confidence interval we need to know the standard errors of the proportions we are interested in.

For the first category we have:

$Var(p_y) = p*(1-p)/N = .06(.94)/927 = .00006$ for Mexico, and

$Var(p_x) = .02(.98)/716 = .00003$ for Colombia

$\sqrt{Var(\bar{p}_x - \bar{p}_y)} = (.00003 + .00006)^{1/2} = .0094$

with confidence interval:

$[p_x-p_y - Z_{\alpha/2} * \sqrt{Var(\bar{p}_x - \bar{p}_y)}, p_x-p_y + Z_{\alpha/2} * \sqrt{Var(\bar{p}_x - \bar{p}_y)}]$

[.02-.06-1.96*.0094, .04 + 1.96*.0094]

[-.058,-.022]

For the extreme right category:

$Var(p_y) = p*(1-p)/N = .05(.95)/927 = .00005$ for Mexico, and

$Var(p_x) = .34(.66)/716 = .0003$ for Colombia

$\sqrt{Var(\bar{p}_x - \bar{p}_y)} = (.0003 + .00005)^{1/2} = .019$

with confidence interval:

[.25, .33]

With respect to the extreme left category, both countries display low proportions, however, Colombia displays a statistically significantly lower proportion if compared to Mexico (the confidence interval around the difference between the two proportions does not include zero, so we reject the null that there is no difference between the two countries at the 5% level).

With respect to the extreme right category, Colombia displays a statistically significantly higher proportion of citizens assuming that position. In this case the difference between the two countries is higher too, between 25% and 33%, at the 5% level.

8. $H_0: \mu = 20{,}000$

$H_A: \mu \neq 20{,}000$

N =21 cars $\bar{X} = 18{,}700$ $\sigma = 8{,}600$

$Z = (\bar{X} - \mu)/ \sigma/\sqrt{N} = -.69$

$2*P(Z \geq .69) = .4885$

Given that .4885 > .05 we fail to reject the null that the car manufacturer is right.

9. Here we employ a test of proportion.

H_0: $\pi = .5$

H_A: $\pi \neq .5$

$p = 0.7$

$Var(p) = [p\ (1-p)\ /N] = [(.7*.3)\ /10] = 0.021$

Std. Error: $[p\ (1-p)\ /N]^{1/2} = 0.021^{1/2} = 0.145$

$Z = (.7-.5)/0.145 = 1.38$

If we employ the 10% level we would find no evidence of ESP. The same is true for tests at the 5% and 1% levels.

10. Here we need a test of population mean.

H_0: $\mu = .25$

H_A: $\mu \neq .25$

$N = 17$, $\sigma = .056$

$Z = (.23 - .25)/\ .056/\sqrt{17} = -1.47$

This corresponds to a p-value of 0.1409. Thus, employing a 5% level of confidence, we would fail to reject the null that the deli is right.

11. Here we need to run a difference of means test on two independent samples.

H_0: $\mu_x = \mu_y$

H_A: $\mu_x \neq \mu_y$

Female (x): $\bar{x} = 11$, $\sigma_x = 1$, $N_x = 11$

Male (y): $\bar{y} = 14$, $\sigma_y = 3$, $N_y = 16$

$Z = (\bar{x} - \bar{y})/ (\sigma_x / \sqrt{\bar{N}_x} + \sigma_y / \sqrt{\bar{N}_y})$

$Z = (11-14) / (1/\sqrt{\overline{11}} + 3/\sqrt{\bar{16}}) = - 3.7$

Since $|Z| > 2.57$ (the critical value at the 1% level) we reject the null of no discrimination at any conventional level of confidence.

12. Here we need a difference in proportions test.

H_0: $\pi_X = \pi_Y$

H_A: $\pi_X \neq \pi_Y$

Female (x): $p_x = 445/(445+675) = 40\%$ $N_x = 445+675 = 1120$

Male (y): $p_y = 515/(515+641) = 45\%$, $N_y = 1156$

$Z = [(p_x - p_y) - (\pi_x - \pi_y)]/[p_x (1-p_x)/N_x + (p_y (1-p_y)/N_y]^{1/2}$

$Z = .05/(.24/1120 + .2475/1156)^{1/2} = - 2.42$

The corresponding p-value is 0.0157

We would, then, reject the null that males and females high school students are equally likely to consume marijuana at the 5% level, but we would not reject the null at the 1% level.

The 95% confidence interval around the estimated difference is:

[-0.0906, -0.0094]

Since it does not include zero and is negative, we find strong evidence that female high school students are less likely than males to consume marijuana at the 5% level.

13. In this case we need a difference in means test on two independent samples.

H_0: $\mu_X = \mu_Y$

H_A: $\mu_X \neq \mu_Y$

Mississippi (x): $\bar{x} = 121$, $\sigma_x = 18$, $N_x = 717$

Alabama (y): $\bar{y} = 114$, $\sigma_y = 13$, $N_y = 834$

$$Z = (\bar{x} - \bar{y}) / (\sigma_x / \sqrt{N_x} + \sigma_y / \sqrt{N_y})$$

$$Z = (121 - 114)/(18/\sqrt{717} + 13/\sqrt{834})$$

$$Z = 6.24$$

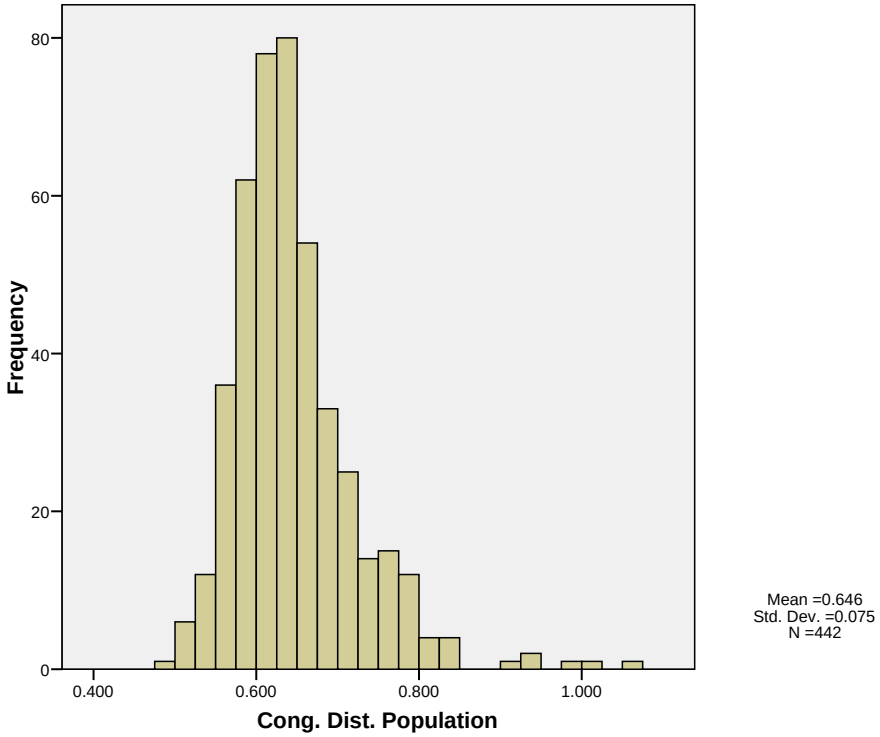
With such a high Z score we reject the null that house prices in Alabama are the same as in Mississippi at any conventional level of confidence. Looking at the 95% confidence interval around the difference in means, we find that houses in Mississippi tend to be more expensive than houses in Alabama.

$[7 - 1.96*(\sigma_x / \sqrt{N_x} + \sigma_y / \sqrt{N_y}) , 7 + 1.96*(\sigma_x / \sqrt{N_x} + \sigma_y / \sqrt{N_y})]$

[4.8, 9.2]

14.

(a) We can plot the distribution of district population using the Graphs → Legacy Dialogs → Histogram command. We obtain the following,



We can determine that the data are unimodal, since the above graph only has a single peak.

(b) Computing the correlation between these two variables, we have,

		Dem. House Cand. Ideology (NPAT)	Rep. House Cand. Ideology (NPAT)
Dem. House Cand. Ideology (NPAT)	Pearson Correlation	1	.304(**)
	Sig. (2-tailed)		.000
	N	355	299
Rep. House Cand. Ideology (NPAT)	Pearson Correlation	.304(**)	1
	Sig. (2-tailed)	.000	
	N	299	365

The correlation is 30.4%, indicating a moderately strong positive relationship. Districts that had Democratic candidates that were more conservative than average also tended to have Republican candidates that were more conservative than average.

(c) We can answer this question by employing an independent samples test.

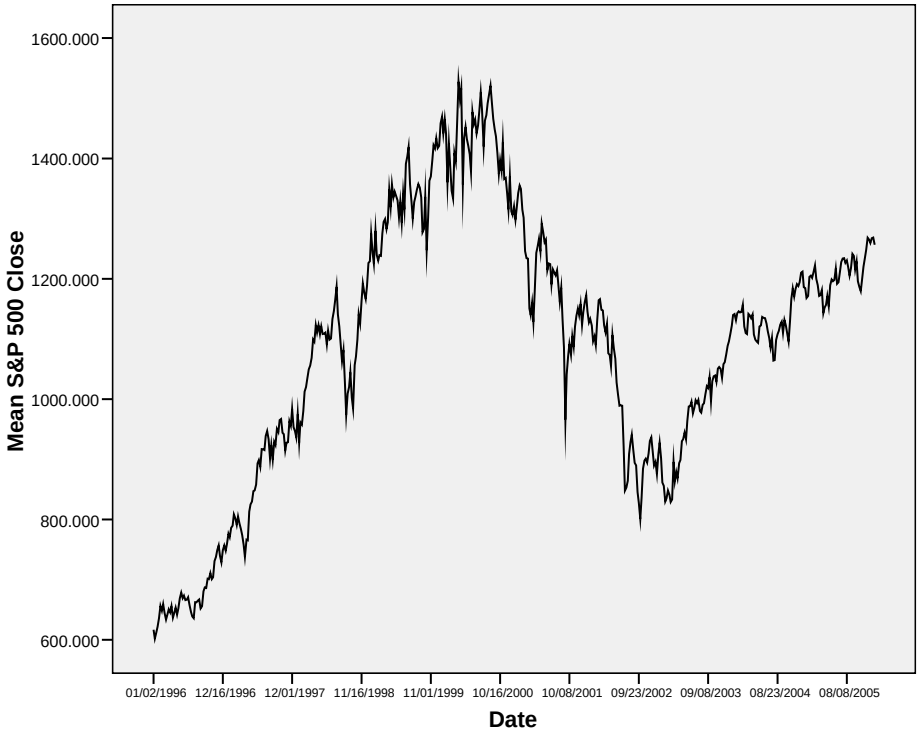
	Cong. Dist. in the South	N	Mean	Std. Deviation	Std. Error Mean
Cong. Dist. Ideology (Survey-based Measure)	0	284	3.10413	.182350	.010820
	1	155	3.27552	.144297	.011590

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Cong. Dist. Ideology (Survey-based Measure)	Equal variances assumed	7.528	.006	-10.101	437	.000	-.171395	.016968	-.204745	-.138046
	Equal variances not assumed			-10.809	381.663	.000	-.171395	.015856	-.202571	-.140219

Focusing on the second row (Equal variances not assumed), we find that the difference is statistically significant at the 5% level. Hence, the average congressional district in the south is more conservative than the average congressional district in the non-Southern United States.

15. A line graph of the time-series can be obtained by using the SPSS command, Graphs → Legacy Dialogs → Line, selecting 'other statistic (e.g. mean), selecting S&P 500 close as the variable, and selecting 'date' as the category axes.



The results indicate that the S&P500 index was rising up until about the beginning of 2001, and then began to rapidly fall. Starting around 2003, the stock index began to rise again.

16. We can analyze this relationship using either a cross tabulation or computing the correlation coefficient. The correlation coefficient can be computed using Analyze → Correlate → Bivariate, and selecting the relevant variables.

		Return Golan to Syria	Right left
Return Golan to Syria	Pearson Correlation	1	.460(**)
	Sig. (2-tailed)		.000
	N	788	774
Right left	Pearson Correlation	.460(**)	1
	Sig. (2-tailed)	.000	
	N	774	1381

We can determine that the correlation is 46%. This is a moderately strong positive correlation, indicating that those with more left-wing ideologies are more likely to support returning the Golan Heights to Syria. This result conforms with our expectations- we expect views on security to dominate in the responses Israelis give to the left-right item, with right-wing views encompassing the beliefs that maintaining the occupied areas is essential for the security of Israel and with left-wing views encompassing the belief that returning some of the occupied areas are in Israelis long-term security interests, because they increase the likelihood of a peaceful settlement between Israel, the Palestinians, and neighboring Arab states.

Using a cross-tabulation, we have,

			Right left							Total
			Right	2	3	4	5	6	Left	Right
Return Golan to Syria	Give back none	Count	97	64	54	96	30	19	19	379
		% within Right left	78.9%	71.1%	65.9%	47.3%	28.6%	22.4%	22.1%	49.0%
	Give back small part	Count	20	21	17	33	27	10	18	146
		% within Right left	16.3%	23.3%	20.7%	16.3%	25.7%	11.8%	20.9%	18.9%
	Give back large part	Count	1	2	11	29	14	16	20	93
		% within Right left	.8%	2.2%	13.4%	14.3%	13.3%	18.8%	23.3%	12.0%
	Give back all	Count	5	3	0	45	34	40	29	156
		% within Right left	4.1%	3.3%	.0%	22.2%	32.4%	47.1%	33.7%	20.2%
Total		Count	123	90	82	203	105	85	86	774
		% within Right left	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

In this cross-tabulation, we see the same pattern. The vast majority of those holding very right wing views don't support giving any of the Golan Heights back to Syria, while those holding the most left-wing views are likely to support giving some or all of the Golan Heights back to Syria.

17.

(a) Income has a distribution that is observable in the population is observed through the census, but it is subject to substantial measurement error and item nonresponse. In fact, even the census measurements of income are subject to item nonresponse, so income is not a particularly good variable to use in weighting.

(b) Education level should be included as a weighting variable because its distribution in the population is also observed through the census.

(c) Partisan identification should not be included as a weighting variable because its distribution in the population is not easily observed. As the purpose of the poll is to predict election day support, we know that the last reliable exit poll data on the voting population is two to four years old, and likely changed over that time. Thus, we cannot obtain the correct distribution of party id for the population and it serves as a poor weighting variable.

(d) Vote in the previous presidential election should not be included as a weighting variable not because it is not easily observed, but because there is a significant degree of respondent error as well as the problem of handling new voters. Respondents tend to claim that they voted for the winner of an election in proportions greater than what was actually observed by a number of percentage points. These misreported respondents and new voters will skew the estimate if we weight by previous presidential vote.

18.

(a) If we want 50 percent turnout, we need to use the 50% of the population most likely to vote. Thus, we begin with all the 40% of respondents with a score of 7, and we need to add the next 10% of the population most likely to vote. This

would be half of the 20% with a score of 6. Thus, we need 100% of 7s, 50% of 6s, and 0 percent of the rest.

(b) If we want 75 percent turnout, we need to use the 75% of the population most likely to vote. Thus, we begin with all the 40% of respondents with a score of 7, and we need to add the next 35% of the population most likely to vote. This requires all of the 20% of respondents with a score of 6, and we still need to add the next 15% of the population most likely to vote. This is satisfied by adding all of the15% with a score of 5. Thus, we need 100% of 7s, 100% of 6s, 100% of 5s, and 0 percent of the rest.

19. If the population has a greater proportion of educated people than the sample, and educated people are more likely to support Democratic candidates than less educated people, and Gallup does not adjust for this through appropriate weighting, then I would agree with this statement, as it would underreport support for Democratic candidates due to survey bias.

20. If party identification fluctuates from one election to the next, and the number of voters identifying with the GOP has decreased to 27% with some degree of confidence, then I would disagree with that statement because the LA Times poll would accurately account for the current state of party identification, not the state existing after the previous election which has since changed.

21.

(a) The sample proportion of respondents who answer “Yes, worth fighting, STRONGLY”?

A frequency table of respondents to item q19 can be obtained by using the SPSS command, Analyze → Descriptive Statistics → Frequencies, selecting selecting q19 as the variable. Ensure that the weight is turned off. We get the following:

Q.19 All in all, considering the costs to the United States versus the benefits to the United States, do you think the war in Iraq was worth fighting?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes, worth fighting, STRONGLY	351	29.2	36.5	36.5
	Yes, worth fighting, SOMEWHAT	110	9.1	11.4	48.0
	No, not worth fighting, SOMEWHAT	87	7.2	9.1	57.0
	No, not worth fighting, STRONGLY	387	32.1	40.3	97.3
	DK/No opinion	26	2.2	2.7	100.0
	Total	961	79.8	100.0	
Missing	System	243	20.2		
Total		1204	100.0		

We see that 36.5 percent answered "Yes, worth fighting, STRONGLY."

(b) What is the weighted sample proportion of respondents who answer "Yes, worth fighting, STRONGLY"?

If we turn the weights on and repeat the same command, we get the table:

Q.19 All in all, considering the costs to the United States versus the benefits to the United States, do you think the war in Iraq was worth fighting?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes, worth fighting, STRONGLY	312	25.9	34.7	34.7
	Yes, worth fighting, SOMEWHAT	107	8.8	11.8	46.5
	No, not worth fighting, SOMEWHAT	76	6.3	8.4	54.9
	No, not worth fighting, STRONGLY	381	31.7	42.3	97.2

	DK/No opinion	25	2.1	2.8	100.0
	Total	900	74.8	100.0	
Missing	System	304	25.2		
Total		1204	100.0		

We see that the weighted sample proportion of respondents who answer "Yes, worth fighting, STRONGLY" is 34.7 percent.

(c) If we look at the variable q3, which measures the respondent's intended vote choice for President, and the weighted and unweighted proportion in support of President Bush, we see that the sample over-represents Bush supporters by about 2 percent, which is nearly the same as the difference between both proportions in parts a and b. However, there could be any number of ways in which the sample over or under represents parts of the population that alters the estimate.

(d) The best estimate is provided by the weighted sample proportion because it corrects survey bias on a number of variables and works to make the sample more representative of the population as a whole.

22.

(a)

(Weight: wtfctr) Congressional vote: Party's candidate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democratic Party's candidate	733	48.4	48.4	48.4
	Republican Party's candidate	536	35.3	35.3	83.7
	Other candidate (VOL)	29	1.9	1.9	85.6
	DK/Undecided (VOL)	148	9.8	9.8	95.4
	REF	43	2.8	2.8	98.2
	Don't plan to vote (VOL)	28	1.8	1.8	100.0

(Weight: wtfctr) Congressional vote: Party's candidate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democratic Party's candidate	733	48.4	48.4	48.4
	Republican Party's candidate	536	35.3	35.3	83.7
	Other candidate (VOL)	29	1.9	1.9	85.6
	DK/Undecided (VOL)	148	9.8	9.8	95.4
	REF	43	2.8	2.8	98.2
	Don't plan to vote (VOL)	28	1.8	1.8	100.0
	Total	1516	100.0	100.0	

(Weight: wtlv35) Congressional vote: Party's candidate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democratic Party's candidate	271	49.3	49.3	49.3
	Republican Party's candidate	223	40.5	40.5	89.7
	Other candidate (VOL)	11	2.0	2.0	91.8
	DK/Undecided (VOL)	28	5.0	5.0	96.8
	REF	18	3.2	3.2	100.0
	Total	550	100.0	100.0	

(Weight: wtlv40) Congressional vote: Party's candidate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democratic Party's candidate	299	49.3	49.3	49.3
	Republican Party's candidate	244	40.3	40.3	89.6
	Other candidate (VOL)	12	2.1	2.1	91.7
	DK/Undecided (VOL)	32	5.2	5.2	96.9
	REF	19	3.1	3.1	100.0
	Total	606	100.0	100.0	

(Weight: wltv45) Congressional vote: Party's candidate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democratic Party's candidate	336	49.3	49.3	49.3
	Republican Party's candidate	274	40.2	40.2	89.5
	Other candidate (VOL)	14	2.1	2.1	91.6
	DK/Undecided (VOL)	37	5.4	5.4	97.0
	REF	20	3.0	3.0	100.0
	Total	682	100.0	100.0	

(Weight: wtlv50) Congressional vote: Party's candidate

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democratic Party's candidate	374	49.4	49.4	49.4
	Republican Party's candidate	304	40.1	40.1	89.5
	Other candidate (VOL)	16	2.1	2.1	91.5
	DK/Undecided (VOL)	42	5.6	5.6	97.1
	REF	22	2.9	2.9	100.0
	Total	757	100.0	100.0	

(b) Why do these estimators differ?

The estimates based on the various weights for likely voters differ because Republicans tend to score higher on likely voter scales than Democrats. Thus, weights with lower turnout put greater weight on respondents with higher likely voter scores, and thus Republicans are given more weight and Democrats get less support.

23. Poor, black, young, and urban individuals tend to vote Democratic, but are also less likely to respond to a survey. This tendency is typically corrected for by

employing demographic weights. Neglecting to correct for demographic imbalances would be one way of inflating her poll numbers.

Individuals who prefer Republican candidates are more likely to vote—particularly in midterm elections. Reporting results for likely voters would help inflate her numbers.

Given that it was widely acknowledged that the Democrats were likely to face loses in the election, Democratic voters were much less enthusiastic about the election than Republican voters. Including an item measuring enthusiasm when constructing the likely voter model would help inflate her results.

All polls will have some undecided voters. When forming a prediction, these voters must be allocated to one of the candidates. One rule is to allocate them evenly among the two candidates. An alternative that would be more favorable to O'Donnell would be to allocate the undecideds to the challenger party—in this case the Republican Party--because the incumbent candidate (Joe Biden) was a Democrat.

Short of actually allocating the undecideds, one could design a question wording that is likely to generate many undecideds. For example,

"Suppose that the election was held to today? Would you vote for the Democratic candidate Chris Coons, the Republican candidate Christine O'Donnell, or are you not <u>100% sure</u> of which candidate you will vote for?"

Such a question wording is likely to lead to many undecided voters. Then, when reporting the results to the media, one could emphasize that Chris Coons (the candidate for the incumbent party) is below 50%.

Finally, one could ask the respondent about their Senate vote immediately after a series of survey items about the economy. This may make the economy more salient in the respondents' minds leading them to vote retrospectively against the Democratic Party based on the economy, which would benefit O'Donnell's numbers.

24. To the extent that surveys systematically miss younger individuals, this problem can be easily corrected for using demographic weighting. It is always possible that coverage error (such as missing cell phone-only households) will lead to biased estimates, but missing young individuals is unlikely to be the cause because such demographic imbalances are easily corrected for and the vast majority of public opinion polls employ such corrections.