Final Project - Camila Alvarez

1992 Presedential Election Results

Problem 1

> Can we conclude that the results from the 51 samples (50 states + DC) is consistent for the three presedential candidates?

Conclusion

> The election results are not consistent between candidates. The mean amount of votes collected varies by candidate when based on the 51 samples.
>
> The election results are not consistent between candidates. The mean amount of votes collected varies by candidate when based on the electoral college.
>
> There are no signigicant reactions between the candidates election results.

Solutions/Work

In [ ]:

```
! pip install pingouin
```

In [ ]:

```
import pandas as pd
import pingouin as pg

election_data = pd.read_csv('/Project CSV Files/TransformedElectionData.csv')

election_data
```

Out[ ]:

	ID	Candidate	Votes Received	Electoral Votes
0	A1	Bush	0.80	9
1	B1	Bush	0.08	3
2	C1	Bush	0.55	8
3	D1	Bush	0.33	6
4	E1	Bush	3.34	54
...	...	...	...	...
148	ES1	Perot	0.34	13
149	ET1	Perot	0.47	11
150	EU1	Perot	0.11	5
151	EV1	Perot	0.54	11
152	EW1	Perot	0.05	3

153 rows × 4 columns

In [ ]:

```
election_data.anova = pg.anova(data=election_data, dv='Votes Received', between=['Candidate','El
election_data.anova
```

Out[ ]:

	Source	SS	DF	MS	F	p-unc	np2
0	Candidate	6.651021	2.0	3.325510	240.890123	7.338398e-37	0.842597
1	Electoral Votes	65.803470	20.0	3.290173	238.330417	5.130520e-69	0.981469
2	Candidate * Electoral Votes	7.965782	40.0	0.199145	14.425441	4.246204e-25	0.865071
3	Residual	1.242458	90.0	0.013805	NaN	NaN	NaN

County Demographic Data

Problem 2(a):

> Are there significant differences in number of counties in the regions?

Conclusion

> H_0: The mean number of counties does not differ across regions
>
> # H_a: At least one of the mean number of counties differs across regions
>
> At an $\alpha = 0.05$ significance level we reject the null hypothesis that all the means are the same since the calculated $P\text{-}value < 0.05$.

Solutions/Work

In [111…

```python
# Read the CSV file
data = pd.read_csv('/Project CSV Files/county_demographics.csv')

# Define the regional groups
northeast = ['ME', 'MA', 'RI', 'CT', 'NH', 'VT', 'NY', 'PA', 'NJ', 'DE', 'MD']
southeast = ['WV', 'VA', 'KY', 'TN', 'NC', 'SC', 'GA', 'AL', 'MS', 'AR', 'LA', 'FL']
midwest = ['OH', 'IN', 'MI', 'IL', 'MO', 'WI', 'MN', 'IA', 'KS', 'NE', 'SD', 'ND']
southwest = ['TX', 'OK', 'NM', 'AZ']
west = ['CO', 'WY', 'MT', 'ID', 'WA', 'OR', 'UT', 'NV', 'CA', 'AK', 'HI']

# Create a new column 'Region' based on the state
data['Region'] = data['State'].apply(lambda x: 'Northeast' if x in northeast
                                     else 'Southeast' if x in southeast
                                     else 'Midwest' if x in midwest
                                     else 'Southwest' if x in southwest
                                     else 'West')

# Group the data by region
grouped_data = data.groupby('Region')
```

In [112…

```python
county_counts = data.groupby(['Region', 'State']).size().reset_index(name='County Count')

# Print the number of counties in each state, grouped by region
print("Number of counties in each state, grouped by region:")
county_counts
```

```
Number of counties in each state, grouped by region:
```

Out[112]:

	Region	State	County Count
0	Midwest	IA	99
1	Midwest	IL	102
2	Midwest	IN	92
3	Midwest	KS	105
4	Midwest	MI	83
5	Midwest	MN	87
6	Midwest	MO	115
7	Midwest	ND	53
8	Midwest	NE	93
9	Midwest	OH	88
10	Midwest	SD	65
11	Midwest	WI	72
12	Northeast	CT	8
13	Northeast	DE	3
14	Northeast	MA	14
15	Northeast	MD	24
16	Northeast	ME	16
17	Northeast	NH	10
18	Northeast	NJ	21
19	Northeast	NY	62
20	Northeast	PA	67
21	Northeast	RI	5
22	Northeast	VT	14
23	Southeast	AL	67
24	Southeast	AR	75
25	Southeast	FL	67
26	Southeast	GA	159
27	Southeast	KY	120
28	Southeast	LA	64
29	Southeast	MS	82
30	Southeast	NC	100
31	Southeast	SC	46
32	Southeast	TN	95
33	Southeast	VA	133
34	Southeast	WV	55
35	Southwest	AZ	15
36	Southwest	NM	33
37	Southwest	OK	77
38	Southwest	TX	254
39	West	AK	27

	Region	State	County Count
40	West	CA	58
41	West	CO	64
42	West	DC	1
43	West	HI	5
44	West	ID	44
45	West	MT	56
46	West	NV	17
47	West	OR	36
48	West	UT	29
49	West	WA	39
50	West	WY	23

In [ ]:

```
# Perform one-way anova to determine if differences in county amounts per region are significant
countycount_anova = pg.anova(data=county_counts, dv='County Count', between='Region', detailed=T
countycount_anova
```

Out[ ]:

	Source	SS	DF	MS	F	p-unc	np2
0	Region	48127.407754	4	12031.851939	9.023611	0.000018	0.43967
1	Within	61335.219697	46	1333.374341	NaN	NaN	NaN

Problem 2(b):

> Are there significant differences in the population distribution in the regions?

Conclusion

> H_0: The mean population per square mile does not differ across regions.
>
> H_a: At least one of the population per square mile differs across regions.

At an $\alpha = 0.05$ significance level we reject the null hypothesis that all the means are the same since the calculated $P\text{-}value < 0.05$.

Solutions/Work

In [ ]:

```
# Group the data by region, county and population per square mile
region_population_density = data[['Region', 'County','Population.Population per Square Mile']].c

# Print the new DataFrame
print("County population per square mile grouped by region:")
region_population_density
```

```
County population per square mile grouped by region:
```

Out[ ]:

	Region	County	Population.Population per Square Mile
0	Southeast	Abbeville County	51.8
1	Southeast	Acadia Parish	94.3
2	Southeast	Accomack County	73.8
3	West	Ada County	372.8
4	Midwest	Adair County	13.5
...	...	...	...
3134	Southwest	Yuma County	35.5
3135	West	Yuma County	4.2
3136	Southwest	Zapata County	14.0
3137	Southwest	Zavala County	9.0
3138	Midwest	Ziebach County	1.4

3139 rows × 3 columns

In [ ]:

```
# Perform one-way anova to determine if differences in population per square mile per region are
population_dist_anova = pg.anova(data=region_population_density, dv='Population.Population per S
population_dist_anova
```

Out[ ]:

	Source	SS	DF	MS	F	p-unc	np2
0	Region	3.103577e+08	4	7.758944e+07	26.930542	5.324957e-22	0.03323
1	Within	9.029350e+09	3134	2.881095e+06	NaN	NaN	NaN

Problem 2(c):

> Are there significant differences in family size across the five regions?

Conclusion

> H_0: The mean family size does not differ across regions.
>
> H_a: At least one of the family sizes differs across regions.

At an $\alpha = 0.05$ significance level we fail to reject the null hypothesis that all the means are the same since the calculated $P\text{-}value > 0.05$.

Solutions/Work

In [ ]:

```
# Group the data by region, county, and family size
household_size = data[['Region', 'County','Housing.Persons per Household']].copy()

# Print the new DataFrame
print("Family size grouped by region:")
household_size
```

```
Family size grouped by region:
```

Out[ ]:

	Region	County	Housing.Persons per Household
0	Southeast	Abbeville County	2.46
1	Southeast	Acadia Parish	2.76
2	Southeast	Accomack County	2.35
3	West	Ada County	2.58
4	Midwest	Adair County	2.17
...	...	...	...
3134	Southwest	Yuma County	2.79
3135	West	Yuma County	2.45
3136	Southwest	Zapata County	3.17
3137	Southwest	Zavala County	3.33
3138	Midwest	Ziebach County	3.70

3139 rows × 3 columns

In [ ]:

```
# Perform one-way anova to determine if differences in family size per region are significant
household_size_anova = pg.anova(data=household_size, dv='Housing.Persons per Household', between
household_size_anova
```

Out[ ]:

	Source	SS	DF	MS	F	p-unc	np2
0	Region	29.422772	4	7.355693	116.017412	2.156937e-92	0.128977
1	Within	198.700711	3134	0.063402	NaN	NaN	NaN

Problem 2(d):

> Does median home owenership differ significantly across the five regions?

Conclusion

> H_0: The mean of the median homeownership rate per region does not differ.
>
> ## H_a: At least one of the means differs between the regions.
>
> At an $\alpha = 0.05$ significance level we reject the null hypothesis that all the means are the same since the calculated $P\text{-}value < 0.05$.

Solutions/Work

In [ ]:

```
# Group the data by region and state, and calculate the mean homeownership rate
region_median_homeownership_rates = data.groupby(['Region', 'State'])['Housing.Homeownership Rat
region_median_homeownership_rates
```

Out[ ]:

	Region	State	Median Homeownership
0	Midwest	IA	74.787879
1	Midwest	IL	74.838235
2	Midwest	IN	74.459783
3	Midwest	KS	72.425714
4	Midwest	MI	77.921687
5	Midwest	MN	76.809195
6	Midwest	MO	71.891304
7	Midwest	ND	72.807547
8	Midwest	NE	72.781720
9	Midwest	OH	72.106818
10	Midwest	SD	71.383077
11	Midwest	WI	73.995833
12	Northeast	CT	68.775000
13	Northeast	DE	72.433333
14	Northeast	MA	66.014286
15	Northeast	MD	70.933333
16	Northeast	ME	74.656250
17	Northeast	NH	71.900000
18	Northeast	NJ	67.752381
19	Northeast	NY	68.641935
20	Northeast	PA	73.043284
21	Northeast	RI	66.440000
22	Northeast	VT	74.350000
23	Southeast	AL	71.643284
24	Southeast	AR	69.749333
25	Southeast	FL	71.828358
26	Southeast	GA	68.167925
27	Southeast	KY	71.839167
28	Southeast	LA	69.659375
29	Southeast	MS	70.480488
30	Southeast	NC	69.887000
31	Southeast	SC	71.236957
32	Southeast	TN	72.514737
33	Southeast	VA	69.590977
34	Southeast	WV	76.329091
35	Southwest	AZ	68.120000
36	Southwest	NM	70.369697
37	Southwest	OK	71.577922
38	Southwest	TX	71.731102
39	West	AK	63.544444

	Region	State	Median Homeownership
40	West	CA	62.144828
41	West	CO	70.279687
42	West	DC	41.600000
43	West	HI	49.420000
44	West	ID	72.800000
45	West	MT	71.598214
46	West	NV	69.023529
47	West	OR	66.238889
48	West	UT	75.186207
49	West	WA	68.420513
50	West	WY	73.130435

In [ ]:

```
# Perform one-way anova to determine if differences in family size per region are significant
median_homeownership_rates_anova = pg.anova(data=region_median_homeownership_rates, dv='Median H
median_homeownership_rates_anova
```

Out[ ]:

	Source	SS	DF	MS	F	p-unc	np2
0	Region	460.296614	4	115.074154	3.990018	0.007332	0.257586
1	Within	1326.663459	46	28.840510	NaN	NaN	NaN

Problem 2(e):

> Does the level of illiteracy (less than high school degree) differ significantly across the five regions?

Conclusion

> H_0: The mean illiteracy rate per region does not differ.
>
> H_a: At least on of the mean illiteracy rate differs between the regions.

At an $\alpha = 0.05$ significance level we reject the null hypothesis that all the means are the same since the calculated $P\text{-}value < 0.05$.

Solutions/Work

In [ ]:

```
data['Less than high school degree'] = 100 - data['Education.High School or Higher']

region_illiteracy_rates = data[['Region','County','Less than high school degree']].copy()
region_illiteracy_rates
```

Out[ ]:

	Region	County	Less than high school degree
0	Southeast	Abbeville County	18.3
1	Southeast	Acadia Parish	21.0
2	Southeast	Accomack County	18.5
3	West	Ada County	4.8
4	Midwest	Adair County	5.8
...	...	...	...
3134	Southwest	Yuma County	26.7
3135	West	Yuma County	11.4
3136	Southwest	Zapata County	38.1
3137	Southwest	Zavala County	33.1
3138	Midwest	Ziebach County	15.9

3139 rows × 3 columns

In [ ]:

```
# Perform one-way anova to determine if differences in rates of illiteracy are significant
region_illiteracy_rates.anova = pg.anova(data=region_illiteracy_rates, dv='Less than high school
region_illiteracy_rates.anova
```

Out[ ]:

	Source	SS	DF	MS	F	p-unc	np2
0	Region	34662.174686	4	8665.543672	307.801161	1.400475e-223	0.28205
1	Within	88231.681153	3134	28.153057	NaN	NaN	NaN