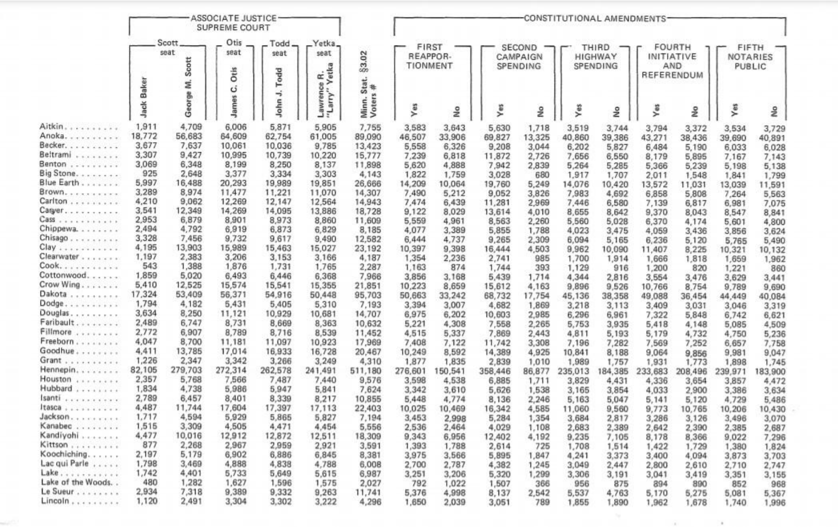
Digitizing Maine’s Voting History with a Statistical Analysis of Error Rate

By Gillian King
Bowdoin College ‘22
Kufe Family Research Fellowship

Advisor: Associate Professor of Mathematics, Jack O’Brien
Related Project: “A Statistical, Historical, and Political Approach to Analyzing Maine’s Election Data” by Jack Olcott, Bowdoin College ‘22

Digitizing Raw Data

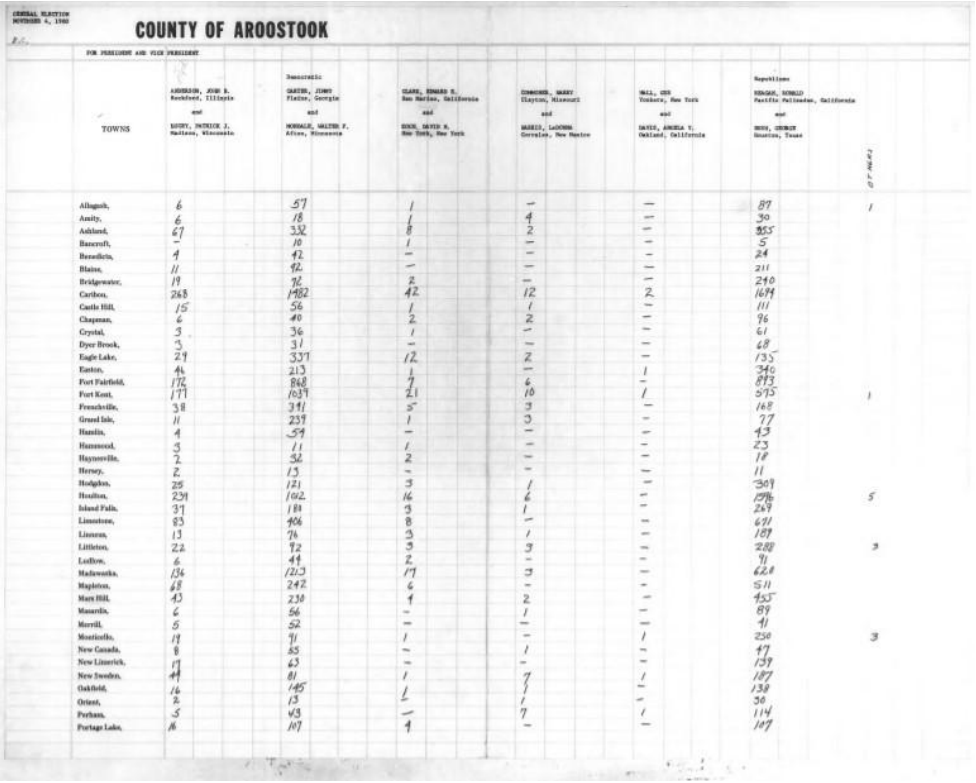
- **Goals:**
  - **Create an online database that stores as much digitized, publicly accessible data as possible in the state of Maine by using an Optical Character Recognition (OCR) software.**
  - **Analyze the efficiency of three statistical tests used to correct errors of the OCR process.**
    - Digital Maine Repository (handwritten data from 1984 to 1860s and before)
      - Digitization process varies state by state.
    - ABBYY Finereader PDF 15 OCR software to convert handwritten documents to searchable PDFs.



	ASSOCIATE JUSTICE SUPREME COURT						CONSTITUTIONAL AMENDMENTS									
	Scott seat		Otis seat	Todd seat	Yetka seat		FIRST REAPPORTIONMENT		SECOND CAMPAIGN SPENDING		THIRD HIGHWAY SPENDING		FOURTH INITIATIVE AND REFERENDUM		FIFTH NOTARIES PUBLIC	
	Jack Baker	George M. Scott	James C. Otis	John J. Todd	Lawrence R. "Larry" Yetka	Minn. Stat. §3.02 Voters #	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Aitkin	1,911	4,709	6,006	5,871	5,905	7,755	3,583	3,643	5,630	1,718	3,519	3,744	3,794	3,372	3,534	3,729
Anoka	18,772	56,683	64,609	62,754	61,005	89,090	46,507	33,906	69,827	13,325	40,860	39,386	43,271	38,436	39,690	40,891
Becker	3,677	7,637	10,061	10,036	9,785	13,423	5,558	6,326	9,208	3,044	6,202	5,827	6,484	5,190	6,033	6,028
Beltrami	3,307	9,427	10,995	10,739	10,220	15,777	7,239	6,818	11,872	2,726	7,656	6,550	8,179	5,895	7,167	7,143
Benton	3,069	6,348	8,199	8,250	8,137	11,898	5,620	4,888	7,942	2,839	5,264	5,285	5,366	5,239	5,198	5,138
Big Stone	925	2,648	3,377	3,334	3,303	4,143	1,822	1,759	3,028	680	1,917	1,707	2,011	1,548	1,841	1,799
Blue Earth	5,997	16,488	20,293	19,989	19,851	26,666	14,209	10,064	19,760	5,249	14,076	10,420	13,572	11,031	13,039	11,591
Brown	3,289	8,974	11,477	11,221	11,070	14,307	7,490	5,212	9,052	3,826	7,983	4,692	6,858	5,808	7,264	5,563
Carlton	4,210	9,062	12,269	12,147	12,564	14,943	7,474	6,439	11,281	2,969	7,446	6,580	7,139	6,817	6,981	7,075
Carver	3,541	12,349	14,269	14,095	13,886	18,728	9,122	8,029	13,614	4,010	8,655	8,642	9,370	8,043	8,547	8,841
Cass	2,953	6,879	8,901	8,973	8,860	11,609	5,559	4,961	8,563	2,260	5,560	5,028	6,370	4,174	5,601	4,800
Chippewa	2,494	4,792	6,919	6,873	6,829	8,185	4,077	3,389	5,855	1,788	4,023	3,475	4,059	3,436	3,856	3,624
Chisago	3,328	7,456	9,732	9,617	9,490	12,582	6,444	4,737	9,265	2,309	6,094	5,165	6,236	5,120	5,765	5,490
Clay	4,195	13,903	15,989	15,463	15,027	23,192	10,397	9,398	16,444	4,503	9,962	10,090	11,407	8,225	10,321	10,132
Clearwater	1,197	2,383	3,206	3,153	3,166	4,187	1,354	2,236	2,741	985	1,700	1,914	1,666	1,818	1,659	1,962
Cook	543	1,388	1,876	1,731	1,765	2,287	1,163	874	1,744	393	1,129	916	1,200	820	1,221	860
Cottonwood	1,859	5,020	6,493	6,446	6,368	7,966	3,856	3,168	5,439	1,714	4,344	2,816	3,554	3,476	3,629	3,441
Crow Wing	5,410	12,525	15,574	15,541	15,355	21,851	10,223	8,659	15,612	4,163	9,896	9,526	10,766	8,754	9,789	9,690
Dakota	17,324	53,409	56,371	54,916	50,448	95,703	50,663	33,242	68,732	17,754	45,136	38,358	49,088	36,454	44,449	40,084
Dodge	1,794	4,182	5,431	5,405	5,310	7,193	3,394	3,007	4,682	1,869	3,218	3,113	3,409	3,031	3,046	3,319
Douglas	3,634	8,250	11,121	10,929	10,681	14,707	6,975	6,202	10,603	2,985	6,296	6,961	7,322	5,848	6,742	6,621
Faribault	2,489	6,747	8,731	8,669	8,363	10,632	5,221	4,308	7,558	2,265	5,753	3,935	5,418	4,148	5,085	4,509
Fillmore	2,772	6,907	8,789	8,716	8,539	11,452	4,515	5,337	7,869	2,443	4,811	5,193	5,179	4,732	4,750	5,236
Freeborn	4,047	8,700	11,181	11,097	10,923	17,969	7,408	7,122	11,742	3,308	7,196	7,282	7,569	7,252	6,657	7,758
Goodhue	4,411	13,785	17,014	16,933	16,728	20,467	10,249	8,592	14,389	4,925	10,841	8,188	9,064	9,856	9,981	9,047
Grant	1,226	2,347	3,342	3,266	3,249	4,310	1,877	1,835	2,839	1,010	1,989	1,757	1,931	1,773	1,898	1,745
Hennepin	82,105	279,703	272,314	262,578	241,491	511,180	276,601	150,541	358,446	86,877	235,013	184,385	233,683	208,496	239,971	183,900
Houston	2,357	5,768	7,566	7,487	7,440	9,576	3,598	4,538	6,885	1,711	3,829	4,431	4,336	3,654	3,857	4,472
Hubbard	1,834	4,738	5,986	5,947	5,841	7,624	3,342	3,610	5,626	1,538	3,165	3,854	4,033	2,900	3,386	3,634
Isanti	2,789	6,457	8,401	8,339	8,217	10,855	5,448	4,774	8,136	2,246	5,163	5,047	5,141	5,120	4,729	5,486
Itasca	4,487	11,744	17,604	17,397	17,113	22,403	10,025	10,469	16,342	4,585	11,060	9,560	9,773	10,765	10,206	10,430
Jackson	1,717	4,594	5,929	5,865	5,827	7,194	3,453	2,998	5,284	1,354	3,684	2,817	3,286	3,126	3,496	3,070
Kanabec	1,515	3,309	4,505	4,471	4,454	5,556	2,536	2,464	4,029	1,108	2,683	2,389	2,642	2,390	2,385	2,687
Kandiyohi	4,477	10,016	12,912	12,872	12,511	18,309	9,343	6,956	12,402	4,192	9,235	7,105	8,178	8,366	9,022	7,296
Kittson	877	2,268	2,967	2,959	2,921	3,591	1,393	1,788	2,614	725	1,708	1,514	1,422	1,729	1,380	1,824
Koochiching	2,197	5,179	6,902	6,886	6,845	8,381	3,975	3,566	5,895	1,847	4,241	3,373	3,400	4,094	3,873	3,703
Lac qui Parle	1,798	3,469	4,888	4,838	4,788	6,008	2,700	2,787	4,382	1,245	3,049	2,447	2,800	2,610	2,710	2,747
Lake	1,742	4,401	5,733	5,649	5,615	6,987	3,251	3,206	5,320	1,299	3,306	3,191	3,041	3,419	3,351	3,155
Lake of the Woods	480	1,282	1,627	1,596	1,575	2,027	792	1,022	1,507	366	956	875	894	890	852	968
Le Sueur	2,934	7,318	9,389	9,332	9,263	11,741	5,376	4,998	8,137	2,542	5,537	4,763	5,170	5,275	5,081	5,367
Lincoln	1,120	2,491	3,304	3,302	3,222	4,296	1,650	2,039	3,051	789	1,855	1,890	1,962	1,678	1,740	1,996

1980 General Election Results for Minnesota (Growe 1980, 16).

Here, each column header represents yes/no vote totals for each question, and each row represents one town within a given county.



GENERAL ELECTION
NOVEMBER 4, 1980

COUNTY OF AROOSTOOK

FOR PRESIDENT AND VICE PRESIDENT

TOWNS	ANDERSON, JOHN B. Rockford, Illinois and LUCEY, PATRICK J. Madison, Wisconsin	Democratic CARTER, JIMMY Plains, Georgia and MONDALE, WALTER F. Afton, Minnesota	CLARK, EDWARD E. San Marino, California and KOCH, DAVID H. New York, New York	COMMONER, BARRY Clayton, Missouri and HARRIS, LaDONNA Corrales, New Mexico	HALL, GUS Yonkers, New York and DAVIS, ANGELA Y. Oakland, California	Republican REAGAN, RONALD Pacific Palisades, California and BUSH, GEORGE Houston, Texas	OTHERS
Allagash,	6	57	1	–	–	87	1
Amity,	6	18	1	4	–	30	
Ashland,	67	332	8	2	–	355	
Bancroft,	–	10	1	–	–	5	
Benedicta,	4	42	–	–	–	24	
Blaine,	11	92	–	–	–	211	
Bridgewater,	19	72	2	–	–	240	
Caribou,	268	1482	42	12	2	1694	
Castle Hill,	15	56	1	1	–	111	
Chapman,	6	40	2	2	–	96	
Crystal,	3	36	1	–	–	61	
Dyer Brook,	3	31	–	–	–	68	
Eagle Lake,	29	337	12	2	–	135	
Easton,	46	213	1	–	1	340	
Fort Fairfield,	172	868	7	6	–	893	
Fort Kent,	177	1039	21	10	1	575	1
Frenchville,	38	391	5	3	–	168	
Grand Isle,	11	239	1	3	–	77	
Hamlin,	4	54	–	–	–	43	
Hammond,	3	11	1	–	–	23	
Haynesville,	7	32	2	–	–	18	
Hersey,	2	13	–	–	–	11	
Hodgdon,	25	121	3	1	–	309	
Houlton,	239	1012	16	6	–	1596	5
Island Falls,	37	181	3	1	–	269	
Limestone,	83	406	8	–	–	691	
Linneus,	13	76	3	1	–	189	
Littleton,	22	92	3	3	–	288	3
Ludlow,	6	44	2	–	–	91	
Madawaska,	136	1213	17	3	–	628	
Mapleton,	68	242	6	–	–	511	
Mars Hill,	43	230	4	2	–	455	
Masardis,	6	56	–	1	–	89	
Merrill,	5	52	–	–	–	41	
Monticello,	19	91	1	–	1	250	3
New Canada,	8	55	–	1	–	47	
New Limerick,	17	63	–	–	–	139	
New Sweden,	44	81	1	7	1	187	
Oakfield,	16	145	1	1	–	138	
Orient,	2	13	–	1	–	30	
Perham,	3	43	–	7	1	114	
Portage Lake,	16	107	4	–	–	107	

1980 General Election Results for Maine (Bureau of Corporations, Elections and Commissions 1980).

Analyzing the Error Rates of Data using OCR

I. Calculating the p-values of the data
- Examining patterns down a column of data (examining the totals by county for each question)

II. Summing the columns of the scanned files
- Looking for correct vote totals in the scanned files (comparing to raw data totals)

III. Random Spot-Check on R-Studio
- Running the code to produce one random column and row in scanned data to compare to raw sheet.

What is a p-value?

- In hypothesis testing, there exists a **null hypothesis** and an **alternative hypothesis.**
    - In general, the null hypothesis suggests that an outcome of an experiment is random, while the alternative hypothesis suggests otherwise.
    - “Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value” (Wasserstein 2016, 131).
    - Further discussion on its potential limitations later in presentation.

P-Values in this presentation

- Goal: Is there a correlation between a county’s proportion of results for one question when compared to their results for another?
    - How likely is the same county likely to vote yes/no to one question given that they voted yes/no to another?
- **In this presentation, the null hypothesis suggests that the entries after the OCR are correct, and that the proportion of yes/no votes per county stays consistent despite a difference in turnout over time.**
- **The alternative hypothesis suggests that for any individual entry after the OCR process, that entry is likely to be an error.**
- Thus, the lower the p-value for a row of data, the more likely that an entry was an error.

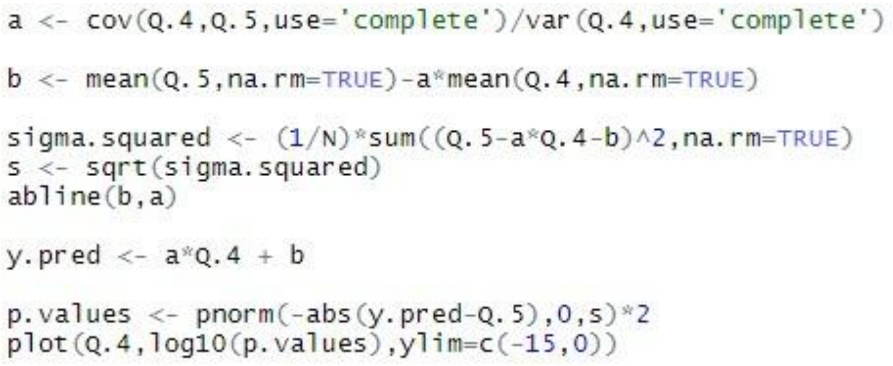
I. Calculate the p-values of the data

Approach using R-Studio:

I. Created a list of yes/no vote totals for on entire question across all all counties for two questions at at time (in log form).
II. Plotted questions against each other to get a general sense of the strength of correlation.
III. Used a simple linear regression to create a line of best fit.
IV. Calculated p-values for each column of data using pnorm in RStudio.
V. Based on calculations in Part I, determined whether or not to re-analyze certain data points.

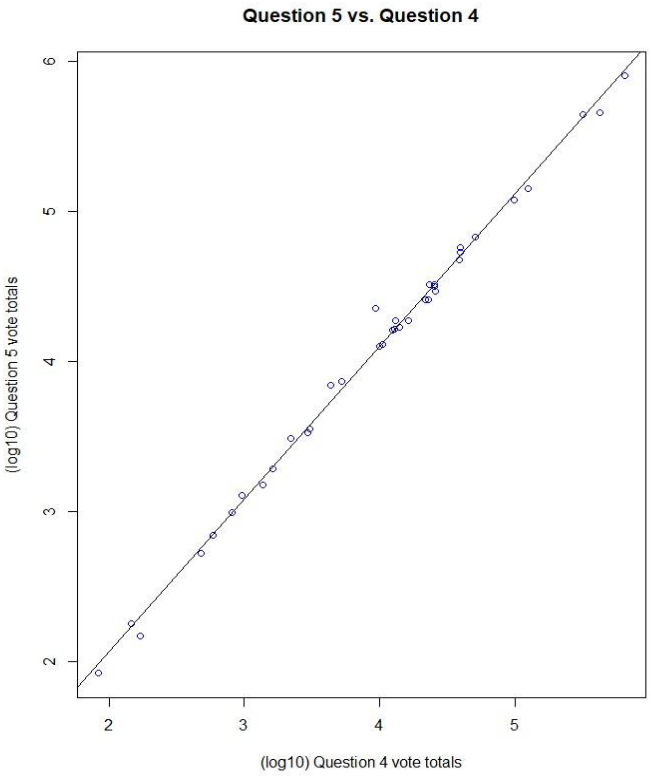
Calculating the Least Squares Estimate (LSE)

- Use the LSE to create a line of best-fit for the data.
- Here, this is merely used to get a sense of large outliers in the data.
- LSE attempts to minimize “the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model” (Kelton and Kindness 2020).

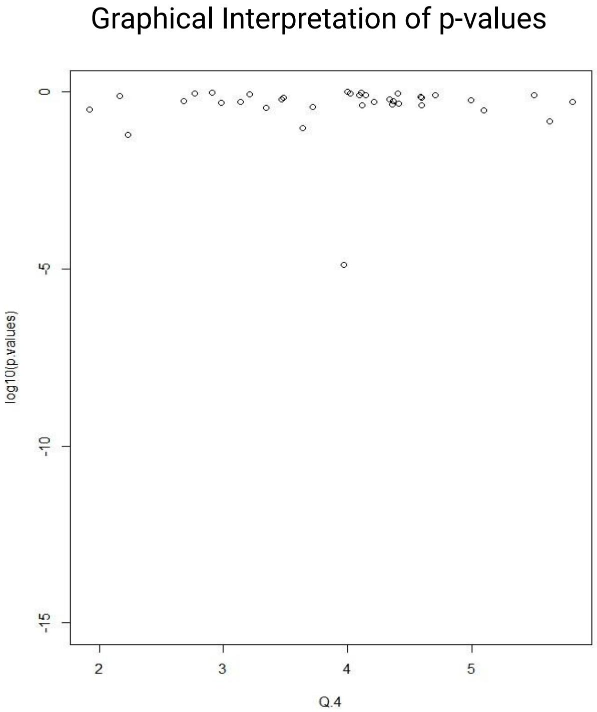


```
a <- cov(Q.4,Q.5,use='complete')/var(Q.4,use='complete')

b <- mean(Q.5,na.rm=TRUE)-a*mean(Q.4,na.rm=TRUE)

sigma.squared <- (1/N)*sum((Q.5-a*Q.4-b)^2,na.rm=TRUE)
s <- sqrt(sigma.squared)
abline(b,a)

y.pred <- a*Q.4 + b

p.values <- pnorm(-abs(y.pred-Q.5),0,s)*2
plot(Q.4,log10(p.values),ylim=c(-15,0))
```

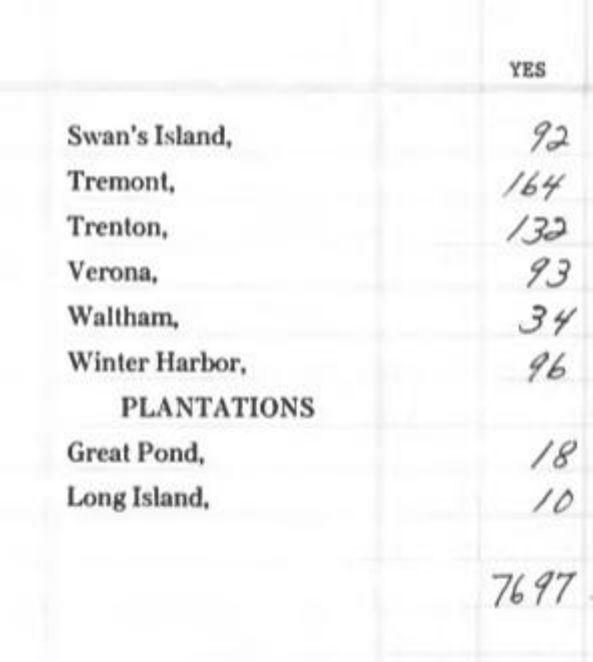
Sample linear regression, data from Hancock County Sheets (1978)



Graphical Interpretation of p-values



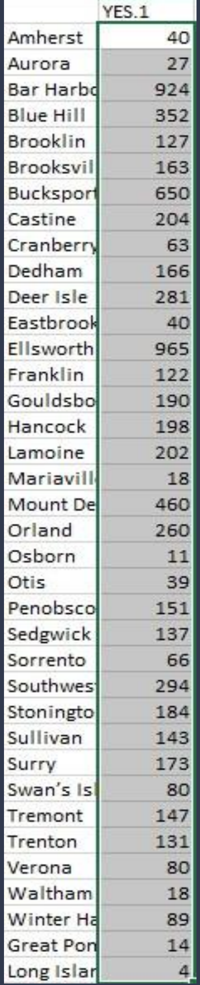
Summing the Columns



	YES
Swan's Island,	92
Tremont,	164
Trenton,	132
Verona,	93
Waltham,	34
Winter Harbor,	96
PLANTATIONS	
Great Pond,	18
Long Island,	10
	7697

(Figure 1)

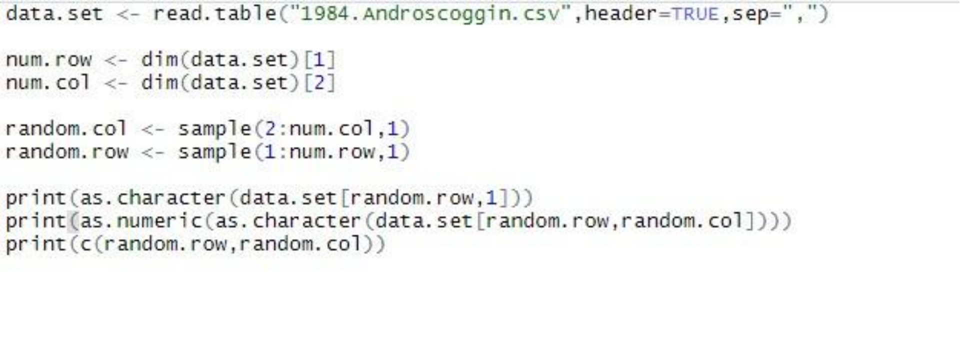
(Bureau of Corporations, Elections and Commissions, 1978)



	YES.1
Amherst	40
Aurora	27
Bar Harbo	924
Blue Hill	352
Brooklin	127
Brooksvil	163
Bucksport	650
Castine	204
Cranberry	63
Dedham	166
Deer Isle	281
Eastbrook	40
Ellsworth	965
Franklin	122
Gouldsbo	190
Hancock	198
Lamoine	202
Mariavill	18
Mount De	460
Orland	260
Osborn	11
Otis	39
Penobsco	151
Sedgwick	137
Sorrento	66
Southwes	294
Stoningto	184
Sullivan	143
Surry	173
Swan's Is	80
Tremont	147
Trenton	131
Verona	80
Waltham	18
Winter Ha	89
Great Pon	14
Long Islar	4

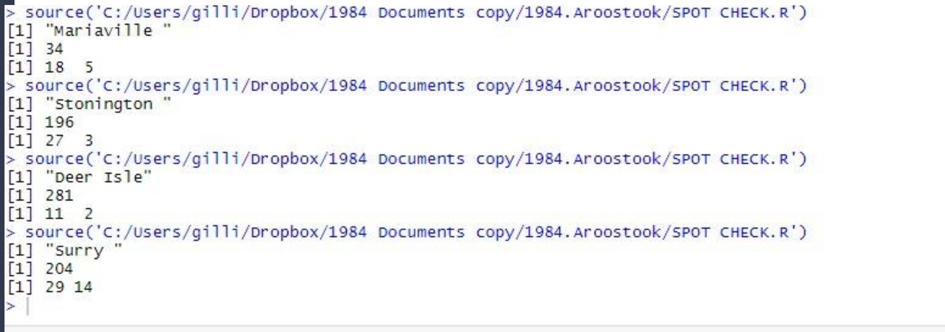
(Figure 2)

Random Spot-Check in RStudio



```
data.set <- read.table("1984.Androscoggin.csv",header=TRUE,sep=",")

num.row <- dim(data.set)[1]
num.col <- dim(data.set)[2]

random.col <- sample(2:num.col,1)
random.row <- sample(1:num.row,1)

print(as.character(data.set[random.row,1]))
print(as.numeric(as.character(data.set[random.row,random.col])))
print(c(random.row,random.col))
```

(Figure 3)



```
> source('C:/Users/gilli/Dropbox/1984 Documents copy/1984.Aroostook/SPOT CHECK.R')
[1] "Mariaville "
[1] 34
[1] 18  5
> source('C:/Users/gilli/Dropbox/1984 Documents copy/1984.Aroostook/SPOT CHECK.R')
[1] "Stonington "
[1] 196
[1] 27  3
> source('C:/Users/gilli/Dropbox/1984 Documents copy/1984.Aroostook/SPOT CHECK.R')
[1] "Deer Isle"
[1] 281
[1] 11  2
> source('C:/Users/gilli/Dropbox/1984 Documents copy/1984.Aroostook/SPOT CHECK.R')
[1] "Surry "
[1] 204
[1] 29 14
> 
```

(Figure 4)

Discussions Around the Limitations of p-Values

Per the American Statistical Association (ASA):

- *P*-values do not directly “measure the probability that the studies hypothesis is true, or the probability that the data were produced by random chance alone” (Wasserstein and Lazar 2020, 131).
- Arbitrary cut-off point
    - Limitations in statistical studies of human behavior such as voting patterns (Wasserstein and Lazar 2020, 131).
- Effect size can provide a false sense of deviation from the null hypothesis (Wasserstein and Lazar 2020, 132).

General Results and Research Design Improvements

General Results:

- Time spent: Test #2 > Test #1 > Test #3
- Errors Caught: Test #2 > Test #1 > Test #3

Research Design Improvements:

- **Crowd-sourcing**
  - Weighing the efficiency of voting results.
  - Will crowdsourcing be more efficient than using the OCR Software?
- Working with Maine State Archives to order printed voting data.

Bibliography

Bureau of Corporations, Elections and Commissions. “1980 General Election: Presidential,” 1980.

Growe, Joan Anderson. *Minnesota Election Results 1980: Primary Election and General Election, Presidential Electors, Representatives in Congress, Senators in Legislature, Representatives in Legislature, Supreme Court, Constitutional Amendments*. St. Paul: Office of the Secretary of State, Election Division, 1980.

Kenton, Will, and David Kindness. “How the Least Squares Method Works.” Investopedia. Investopedia, September 16, 2020. https://www.investopedia.com/terms/l/least-squares-method.asp.

Wasserstein, Ronald L. “ASA Statement on Statistical Significance and P-Values.” *The American Statistician* 70, no. 2 (2016).

Wasserstein, Ronald L., and Nicole A. Lazar. “The ASA Statement on p-Values: Context, Process, and Purpose.” *The American Statistician* 70, no. 2 (2016): 129–33. https://doi.org/10.1080/00031305.2016.1154108.