Math 3070 § 1. **Benfords Law: Counting Frequencies** Name: Example
Treibergs **and Chi-Squared Test of Proportion.** June 27, 2011

This example is pure numerology! You may suspend your credulity for this one!

If one counts the frequency of first digits of numbers that appear on the first page of a newspaper, one sees a counterintuitive fact: the most frequently occurring first digit is one, and the frequencies fall off for higher first digits. Perhaps one would have guessed that the numbers are uniformly distributed.

This observation was made first by the mathematician Simon Newcombe who noticed a hundred years ago that the first pages of a logarithm table were used more than later pages. Fifty years later, a physicist postulated an empirical distribution for the first significant digit. The distribution is called *Benford's Law* and states that the pmf for the first significant digit $s \in \{1, 2, \ldots, 9\}$

$$p(s) = \mathrm{P}(\text{1st significant digit is } s) = \log_{10}\left(1 + \frac{1}{s}\right).$$

The sum $\sum_{i=1}^{9} p(s)$ telescopes and equals $\log_{10}(10) - \log_{10}(1) = 1$.

Here is my guess of Benford's explanation. Think of numbers coming at all scales, and suppose that an order of magnitude is as likely as any other order of magnitude. In other word, the proportion should be given by the relative density of $\log_{10}(s)$. If this is the case, then $p(s)$ should be the relative density of logarithms of numbers beginning with s, or what is the same, the relative density over one order of magnitude. So if we take numbers of "order zero," the interval $[1, 10)$, the numbers beginning with s fall in the interval $[s, s+1)$. Their logarithms have density

$$p(s) = \frac{\log_{10}(s+1) - \log_{10}(s)}{\log_{10}(10) - \log_{10}(1)} = \log_{10}\left(1 + \frac{1}{s}\right).$$

To test this, lets take a convenient, reasonably large data set of popular numbers. I use the `state.x77` data set canned in **R**. The data consists of population estimate as of July 1, 1975, income per capita income (1974), illiteracy (1970, percent of population), life expectancy in years (1969-71), murder and non-negligent manslaughter rate per 100,000 population (1976), percent high-school graduates (1970), mean number of days with minimum temperature below freezing (1931-1960) in capital or large city and land area in square miles. We tabulated the frequencies of first digits, and ran the chi-squared test of proportion.

R Session:

```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.31 (5538) powerpc-apple-darwin8.11.1]

[Workspace restored from /Users/andrejstreibergs/.RData]

> # To see list of datasets canned in R.
> library(help=datasets)
> state.x77
               Population Income Illiteracy Life Exp
Alabama              3615   3624        2.1    69.05
Alaska                365   6315        1.5    69.31
Arizona              2212   4530        1.8    70.55
Arkansas             2110   3378        1.9    70.66
California          21198   5114        1.1    71.71
Colorado             2541   4884        0.7    72.06
Connecticut          3100   5348        1.1    72.48
Delaware              579   4809        0.9    70.06
Florida              8277   4815        1.3    70.66
Georgia              4931   4091        2.0    68.54
Hawaii                868   4963        1.9    73.60
Idaho                 813   4119        0.6    71.87
Illinois            11197   5107        0.9    70.14
Indiana              5313   4458        0.7    70.88
Iowa                 2861   4628        0.5    72.56
Kansas               2280   4669        0.6    72.58
Kentucky             3387   3712        1.6    70.10
Louisiana            3806   3545        2.8    68.76
Maine                1058   3694        0.7    70.39
Maryland             4122   5299        0.9    70.22
Massachusetts        5814   4755        1.1    71.83
Michigan             9111   4751        0.9    70.63
```

Minnesota	3921	4675	0.6	72.96
Mississippi	2341	3098	2.4	68.09
Missouri	4767	4254	0.8	70.69
Montana	746	4347	0.6	70.56
Nebraska	1544	4508	0.6	72.60
Nevada	590	5149	0.5	69.03
New Hampshire	812	4281	0.7	71.23
New Jersey	7333	5237	1.1	70.93
New Mexico	1144	3601	2.2	70.32
New York	18076	4903	1.4	70.55
North Carolina	5441	3875	1.8	69.21
North Dakota	637	5087	0.8	72.78
Ohio	10735	4561	0.8	70.82
Oklahoma	2715	3983	1.1	71.42
Oregon	2284	4660	0.6	72.13
Pennsylvania	11860	4449	1.0	70.43
Rhode Island	931	4558	1.3	71.90
South Carolina	2816	3635	2.3	67.96
South Dakota	681	4167	0.5	72.08
Tennessee	4173	3821	1.7	70.11
Texas	12237	4188	2.2	70.90
Utah	1203	4022	0.6	72.90
Vermont	472	3907	0.6	71.64
Virginia	4981	4701	1.4	70.08
Washington	3559	4864	0.6	71.72
West Virginia	1799	3617	1.4	69.48
Wisconsin	4589	4468	0.7	72.48
Wyoming	376	4566	0.6	70.29

	Murder	HS Grad	Frost	Area
Alabama	15.1	41.3	20	50708
Alaska	11.3	66.7	152	566432
Arizona	7.8	58.1	15	113417
Arkansas	10.1	39.9	65	51945
California	10.3	62.6	20	156361
Colorado	6.8	63.9	166	103766
Connecticut	3.1	56.0	139	4862
Delaware	6.2	54.6	103	1982
Florida	10.7	52.6	11	54090
Georgia	13.9	40.6	60	58073
Hawaii	6.2	61.9	0	6425
Idaho	5.3	59.5	126	82677
Illinois	10.3	52.6	127	55748
Indiana	7.1	52.9	122	36097
Iowa	2.3	59.0	140	55941
Kansas	4.5	59.9	114	81787
Kentucky	10.6	38.5	95	39650
Louisiana	13.2	42.2	12	44930
Maine	2.7	54.7	161	30920
Maryland	8.5	52.3	101	9891
Massachusetts	3.3	58.5	103	7826
Michigan	11.1	52.8	125	56817
Minnesota	2.3	57.6	160	79289

```
Mississippi      12.5    41.0    50  47296
Missouri          9.3    48.8   108  68995
Montana           5.0    59.2   155 145587
Nebraska          2.9    59.3   139  76483
Nevada           11.5    65.2   188 109889
New Hampshire     3.3    57.6   174   9027
New Jersey        5.2    52.5   115   7521
New Mexico        9.7    55.2   120 121412
New York         10.9    52.7    82  47831
North Carolina   11.1    38.5    80  48798
North Dakota      1.4    50.3   186  69273
Ohio              7.4    53.2   124  40975
Oklahoma          6.4    51.6    82  68782
Oregon            4.2    60.0    44  96184
Pennsylvania      6.1    50.2   126  44966
Rhode Island      2.4    46.4   127   1049
South Carolina   11.6    37.8    65  30225
South Dakota      1.7    53.3   172  75955
Tennessee        11.0    41.8    70  41328
Texas            12.2    47.4    35 262134
Utah              4.5    67.3   137  82096
Vermont           5.5    57.1   168   9267
Virginia          9.5    47.8    85  39780
Washington        4.3    63.5    32  66570
West Virginia     6.7    41.6   100  24070
Wisconsin         3.0    54.5   149  54464
Wyoming           6.9    62.9   173  97203
```

```
> ############### PICK OFF FIRST SIGNIFICANT DIGIT ######################
> # The function  as.character()  converts number to string
> # The function  substr(x,a,b) extracts characters a-b from string x
> substr(as.character(345678),1,1)
[1] "3"
> substr(as.character(state.area),1,1)
 [1] "5" "5" "1" "5" "1" "1" "5" "2" "5" "5" "6" "8" "5"
[14] "3" "5" "8" "4" "4" "3" "1" "8" "5" "8" "4" "6" "1"
[27] "7" "1" "9" "7" "1" "4" "5" "7" "4" "6" "9" "4" "1"
[40] "3" "7" "4" "2" "8" "9" "4" "6" "2" "5" "9"

> # Apply to all state data. The function  table(x)  tabulates number of
> # occurences of each entry in  x.  Note zeros appearing from illiteracy
> #rates.
> table(substr(as.character(state.x77),1,1))

 0  1  2  3  4  5  6  7  8  9
26 89 26 36 59 53 36 52 12 11
```

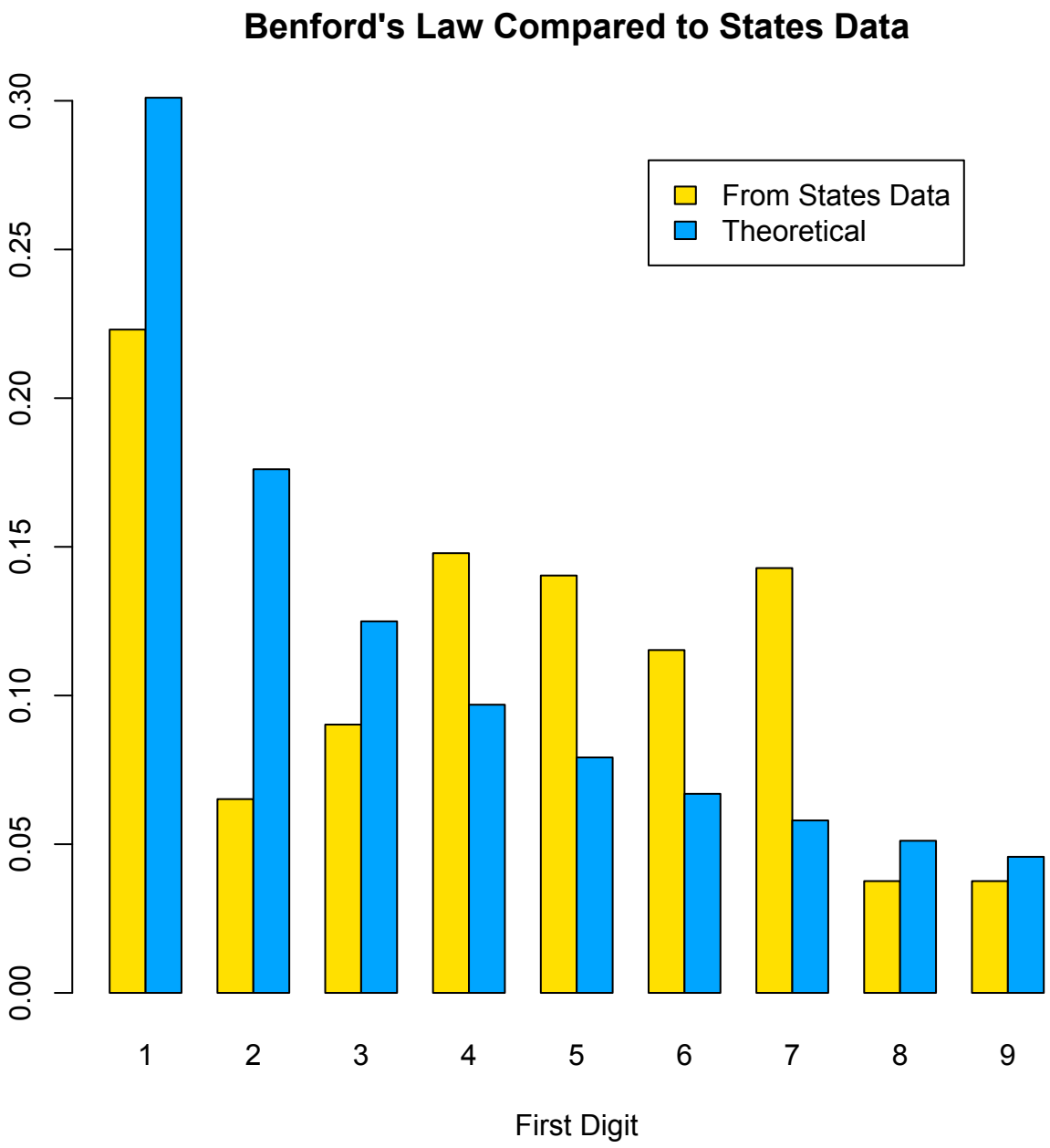
```
> # Multiply by 10 to clear leading zeros.
> table(substr(as.character(10*state.x77),1,1))

 0  1  2  3  4  5  6  7  8  9
 1 89 26 36 59 56 46 57 15 15
> table(substr(as.character(100*state.x77),1,1))

> # One zero remains. it comes from the number of frost days in Hawaii.
> # Convert the zero to NA and tabulate again.
> dat <- state.x77
> dat["Hawaii","Frost"]<- NA
>  ta <- table(substr(as.character(100*dat),1,1)); ta

 1  2  3  4  5  6  7  8  9
89 26 36 59 56 46 57 15 15
>
> ################### MAKE A TABLE COMPARING OBSERVED AND THEORETICAL PROP #####
> sta <- sum(ta)
> # The Benford pmf
> pb <- sapply(1:9, function(x) log10(1+1/x));  pb
[1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679 0.05799195
[8] 0.05115252 0.04575749
> sum(pb)
[1] 1
> m <- cbind(ta/sta,pb)
> colnames(m)<- c("Observed Prop.", "Theoretical Prop.")
> m
  Observed Prop. Theoretical Prop.
1     0.22305764        0.30103000
2     0.06516291        0.17609126
3     0.09022556        0.12493874
4     0.14786967        0.09691001
5     0.14035088        0.07918125
6     0.11528822        0.06694679
7     0.14285714        0.05799195
8     0.03759398        0.05115252
9     0.03759398        0.04575749
```

```
> ############## MAKE SIDE BY SIDE HISTOGTORAM OF OBSERVED VS THEORETICAL #####
> barplot( rbind(ta/sta,pb/sum(pb)), beside = T, col = rainbow(7)[c(2,5)],
+ xlab = "First Digit")
> title("Benford's Law Compared to States Data")
> legend(16,.28, legend = c("From States Data", "Theoretical"),
+ fill = rainbow(7)[c(2,5)],bg="white")
> # M3074Benford1.pdf
```



```
> ############ CHI SQ TEST FOR PROPORTION ######################################
> chisq.test(ta,p=pb)

Chi-squared test for given probabilities

data:  ta
X-squared = 134.8303, df = 8, p-value < 2.2e-16

>
> # Small p-value indicates that these digits don't satisfy Benford's Law.
>
> # same computation by hand:
> ep <- sta*pb
> chisq <- sum((ta-ep)^2/ep); chisq
[1] 134.8303
> nu <- length(ta)-1;nu
[1] 8
> pchisq(chisq,nu,lower.tail=F)
[1] 2.815477e-25
```