Lecture 7: Boost

Last time

We started by finishing up our work on random number generation -- We examined two kinds of techniques popular today: So-called “true” random numbers based on a physical phenomenon and pseudo-random numbers that involve some deterministic mathematical formula

We then took a look at another kind of randomized trial -- A/B testing from web design is an incredibly popular technique for optimizing the layout and operation of sites

Today

We will finish our example from the NY Times, this time considering transformations of variables (a dangling thread from our early discussions of skewed variables)

We will then consider a somewhat novel analysis of the 2003 California Recall Election -- It will be a kind of “natural experiment” that will let us apply our new found analysis skills

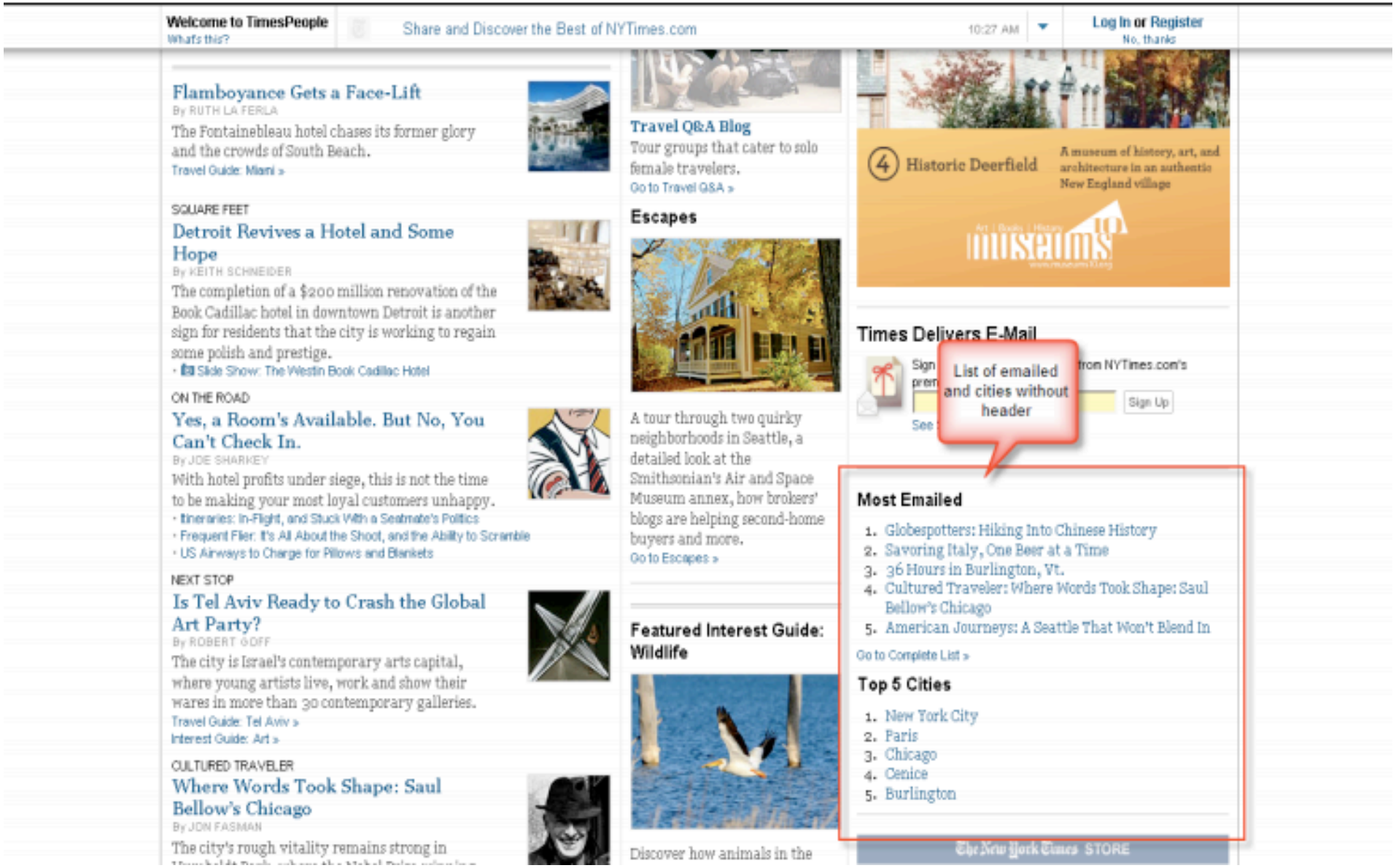
An experiment at `nytimes.com`

We will now consider a more recent example of an A/B test for **The Travel Section** of `nytimes.com` (we'll save the movie test for lab or your midterm or...)

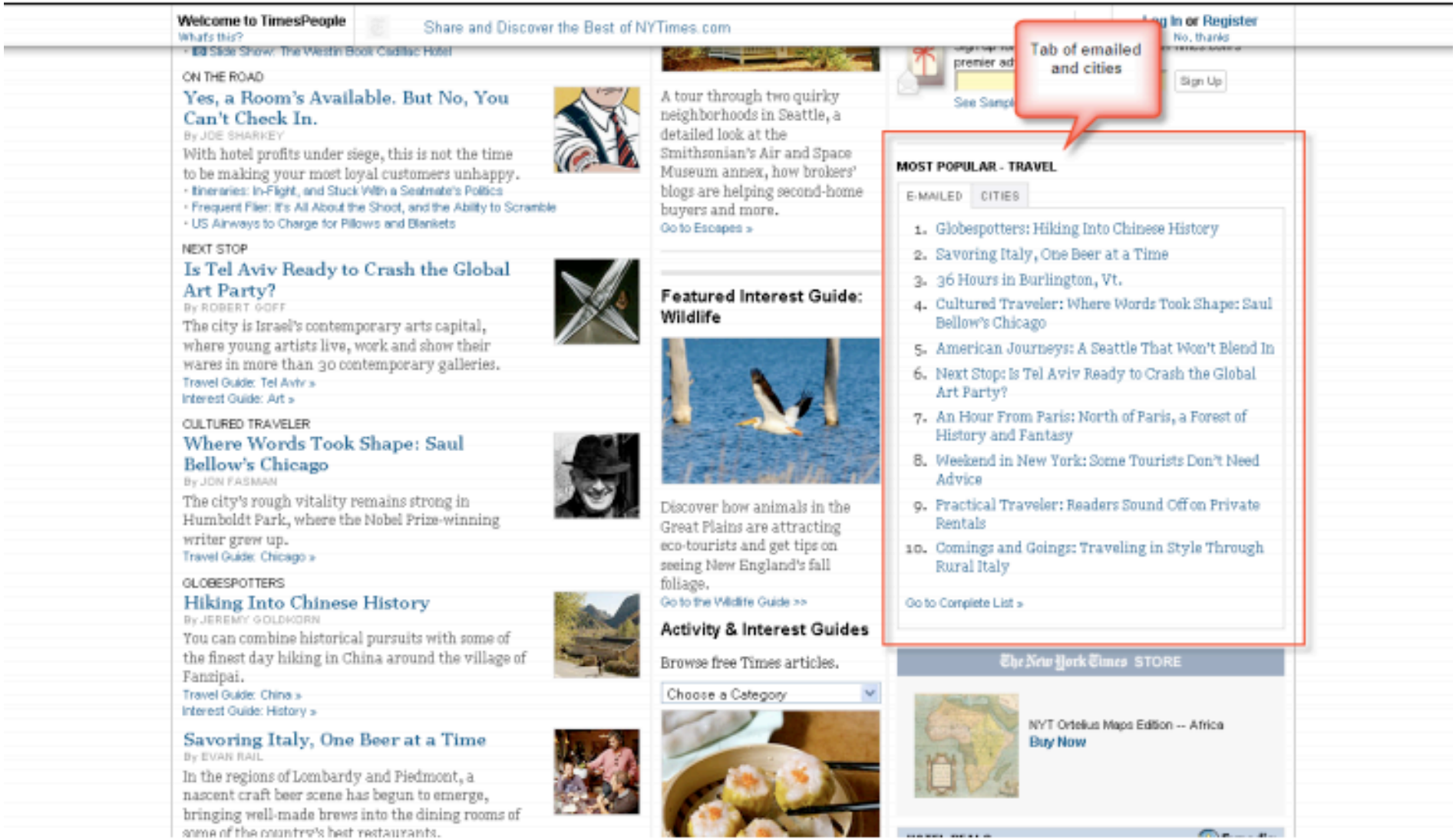
On the next two slides, we present samples of the A and B pages; the changes applied to all pages in The Travel Section, so **as a visitor browsed the site, they would consistently see either A or B**

Have a look at the two designs -- What differences do you see in terms of layout and content? What questions might the Times ask about how visitors react to these two options?

List: Variation 10858



Tabs: Variation 10859



The variables

Recall from last time the variables we have at our disposal; in all data were collected from about 130K users over a period of 6 weeks at the end of 2008

`User_ID` A unique number for each visitor

`UserVisit_ID` A unique number for each visit

`StartTime_SSE` Unix time for the start of the visit

`StartTime_English` A more human readable version of the time

`VisitLength` The number of seconds the visitor was reading Travel Section pages

`Variation` The version of the page they received

`RefererUrl` The page they clicked on (if any) to get to the page

`EntryPageUrl` The first page on `nytimes.com` they visited

`Pageviews` The number of page views to in the Travel Section

`TotalVisits` The total number of visits to the site

`TimeSinceFirstVisit` `(days)` How long it had been since their last visit

`UserAgent` Their browser

`TotalClicks` How many times did they click on the "most popular" field

`IfClicked` 0/1 did they click on the "most popular field" at least once

Another test

We will look at the data in a lot more detail, but to emphasize a concept, **we'll consider a second test that the people who provided the data were interested in** -- Is there a difference between Tabs and Lists in terms of the number of Pageviews?

Let's quickly see how we can address that question...

Hypothesis testing

Before we propose anything formal, let's recall the steps...

1. We begin with **a null hypothesis**, a plausible statement (a model or scenario) which may explain some pattern in a given set of data but made for the purposes of argument; we also select a complementary alternative hypothesis

2. We then define **a test statistic**, some quantity calculated from our data that is used to evaluate how compatible the results are with those expected under the null hypothesis

3. We specify a **threshold or significance level,** α , of the test; at the end of the experiment, this threshold will be applied to determine if we can reject the null

4. We then consider **the distribution of the test statistic under the null hypothesis**; we can get at it either with some probability calculation (remember the table fun from last time) or through computer simulation

5. And finally, after the data are collected, we compute the P-value and apply the threshold α : if our P-value is less than α we reject the null, finding that the data contain evidence for the alternative; if not, we say that we cannot reject the null, and that the data do not contain sufficient evidence for the alternative

Pages per visit

To repeat, the question at hand is, is there some difference in the number of Page Views between groups A and B, between treatment and control, between Tab and Lists?

To turn this into a hypothesis testing exercise, we need a null hypothesis and an alternative; t**he null will be that there is no difference between treatment and control measured in terms of Page Views per visit**

Our **alternative will be that there is a difference,** and that one condition or the other tends to produce more Page Views per visit; notice that in the language we've been evolving over the last couple of lectures, this alternative is two-sided in the sense that we would be convinced of a difference if the Page Views turned out to be much larger for either treatment or control

In terms of significance, we will will choose the default 0.05; if this result proves to be significant, the management at the New York Times (business people) will want a traditional number like that

Now, a test statistic...

Page views per visit

We will work with the absolute value of the difference between the average Page Views per visit computed for each group; **we will use the absolute value to indicate that we are looking for a big difference in either direction**

When judging how extreme our data are, **we will consider how likely it is to see an absolute difference as big or bigger** than the one we see in our experimental data if the null distribution is true

The difference in averages is a reasonable metric for capturing a shift in one group or the other; **the average itself is something that is focused on in the "science" of web traffic, which also recommends it**

Page views per visit

With all that set, let's look at the data we collected

```
> mean((travel$Pageviews)[travel$Variation=="Tabs"])
[1] 1.997261
> mean((travel$Pageviews)[travel$Variation=="List"])
[1] 1.980060
```

The mean number of Pages viewed per visit for Lists is 1.980 while it is 1.997 for the Tabs option; the absolute difference, 0.017 is, well, tiny; and as a practical matter, it might not amount to anything important (although, as we have said, small differences multiplied over millions of visits might prove to be important

Page views per visit

Finally, we need to come up with some way to evaluate the distribution of our test statistic under the null; again, the null is that there is no difference between Tabs and Lists

If there really is no difference, then the value of 0.017 we saw is simply the result of the randomization that took place on the web server

So, if the labels really have nothing to do with the number of Page Views per visit, then **we can simulate other values of the test statistic under the null by simply reassigning visitors to treatment and control, to Tab and Lists**

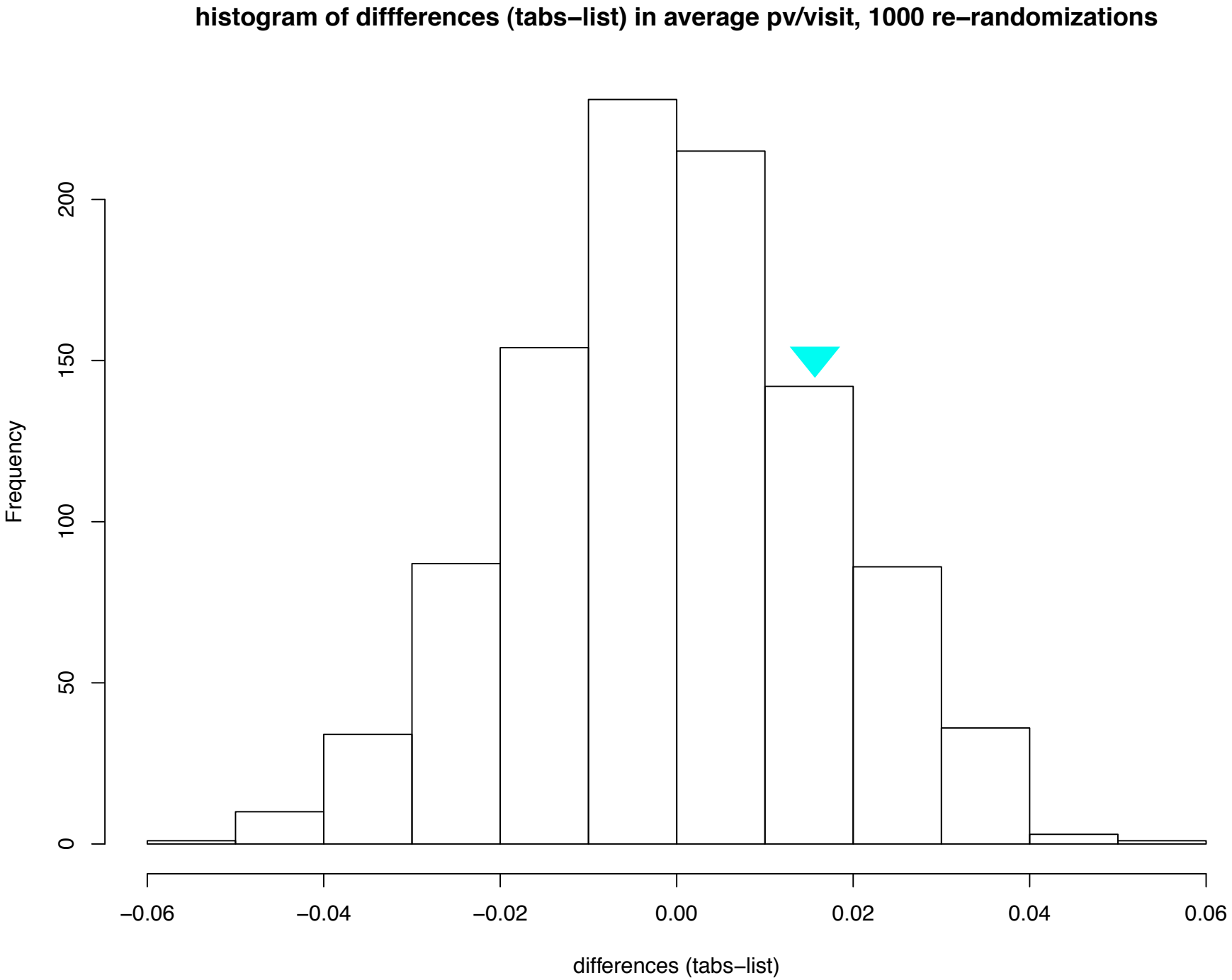
Again we re-randomize...

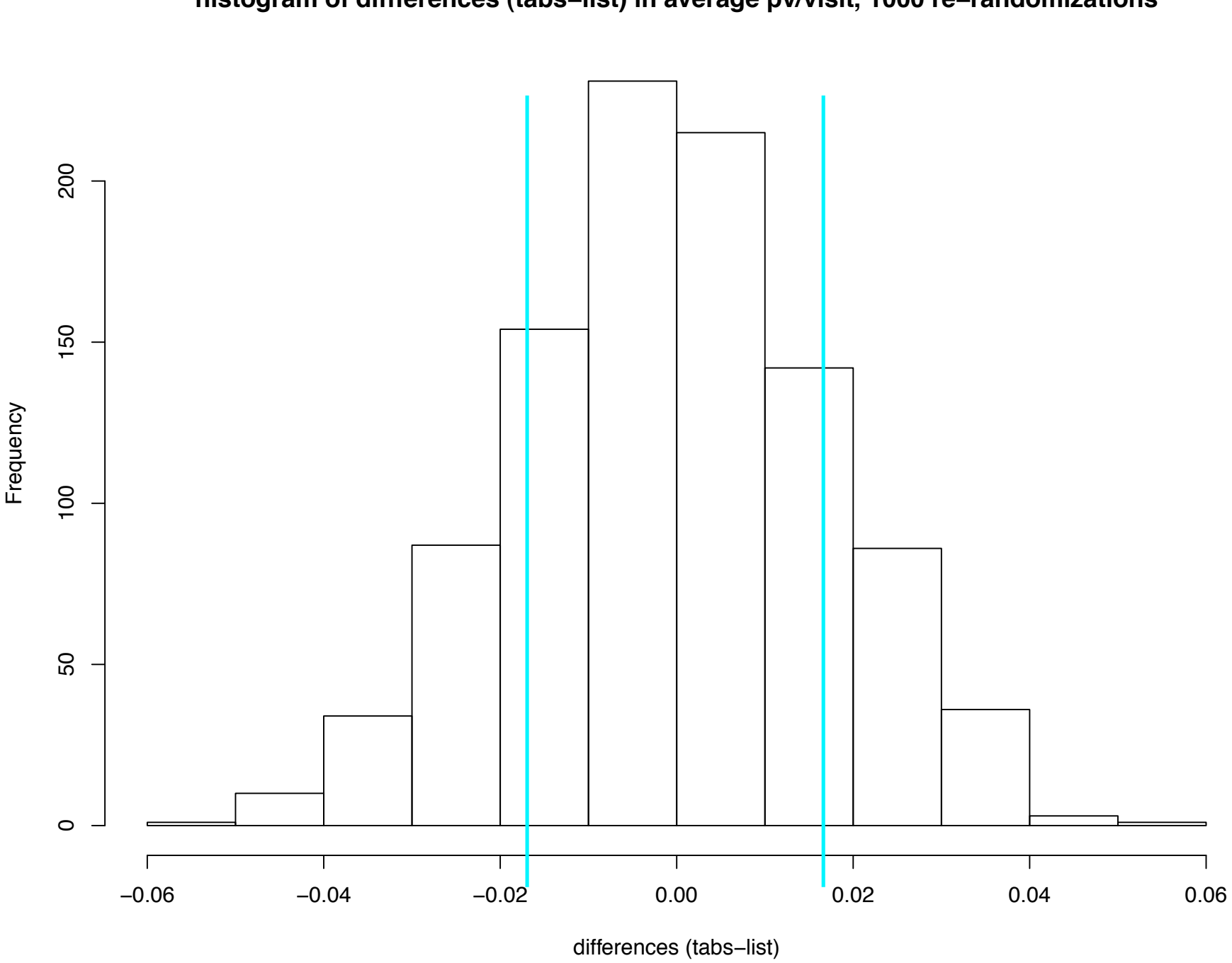
Page views per visit

On the next slide, we present the results of re-randomization; **here we plot the difference between the average Page Views per visit in the List group and the average Page Views per visit in the Tab group**

Our test statistic was really **the absolute value of the difference, but we present the signed value** (before the absolute value) to correspond with what we have been doing so far

Our observed value is 0.017 and for a difference in our null distribution to be more extreme, it has to have an absolute value of 0.017 or greater (meaning -0.017 or smaller and 0.017 or larger)...


histogram of diffferences (tabs−list) in average pv/visit, 1000 re−randomizations
Frequency
0
50
100
150
200
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
differences (tabs−list)


histogram of diffferences (tabs−list) in average pv/visit, 1000 re−randomizations
Frequency
0
50
100
150
200
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
differences (tabs−list)

Page views per visit

In this case, we don’t even have to be very formal about the fact that the observed difference of 0.017 is well within the null distribution, meaning any difference in mean we observed “looks” like it could be the result of our randomization process

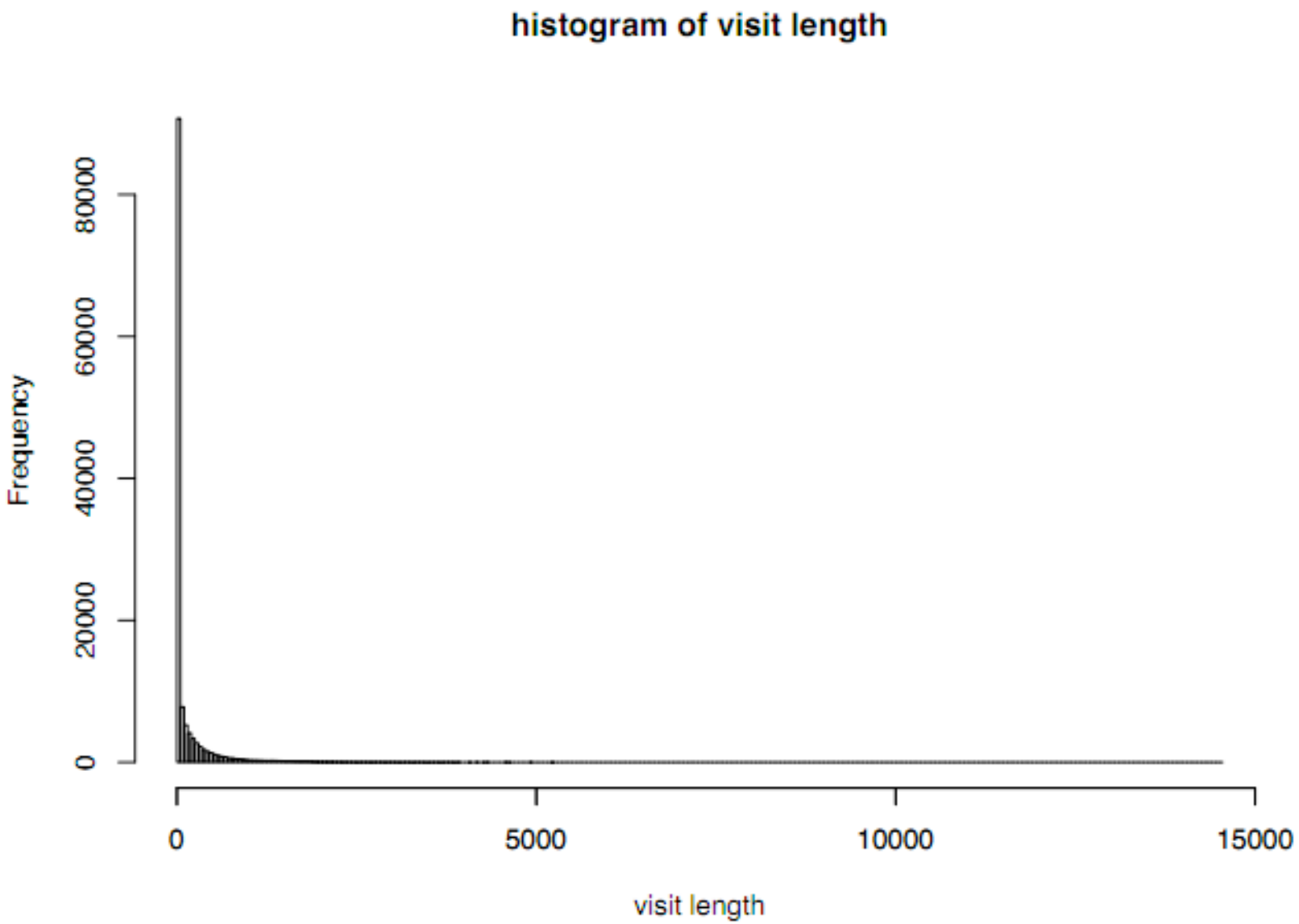
Formally, we would consider the proportion of tables having an absolute difference as large or larger than 0.017 -- That turns out to be 0.32 or 32% of our 1,000 re-randomized tables

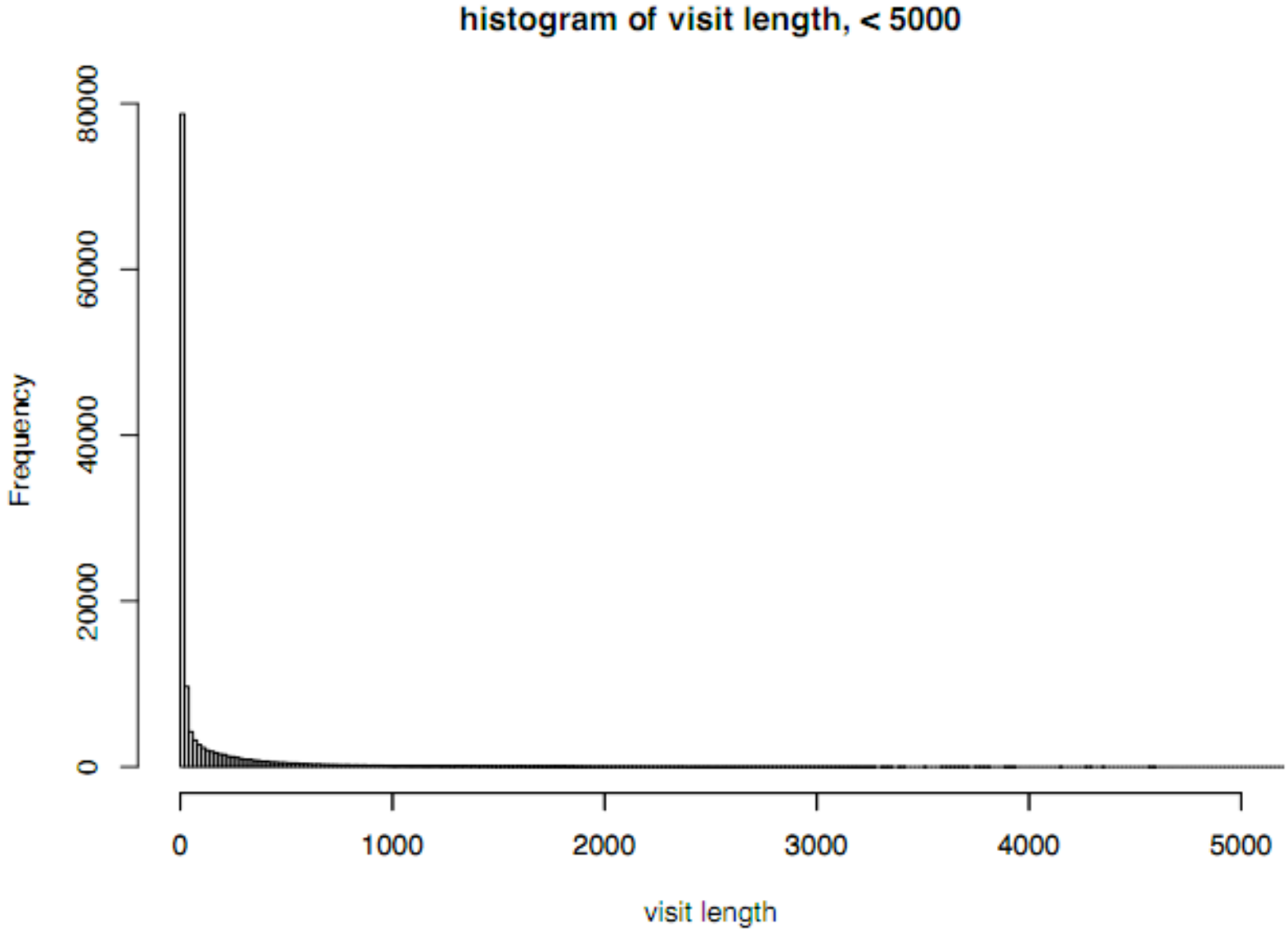
That means we cannot reject the null hypothesis that Tabs and List are performing differently in terms of Pageviews per visit

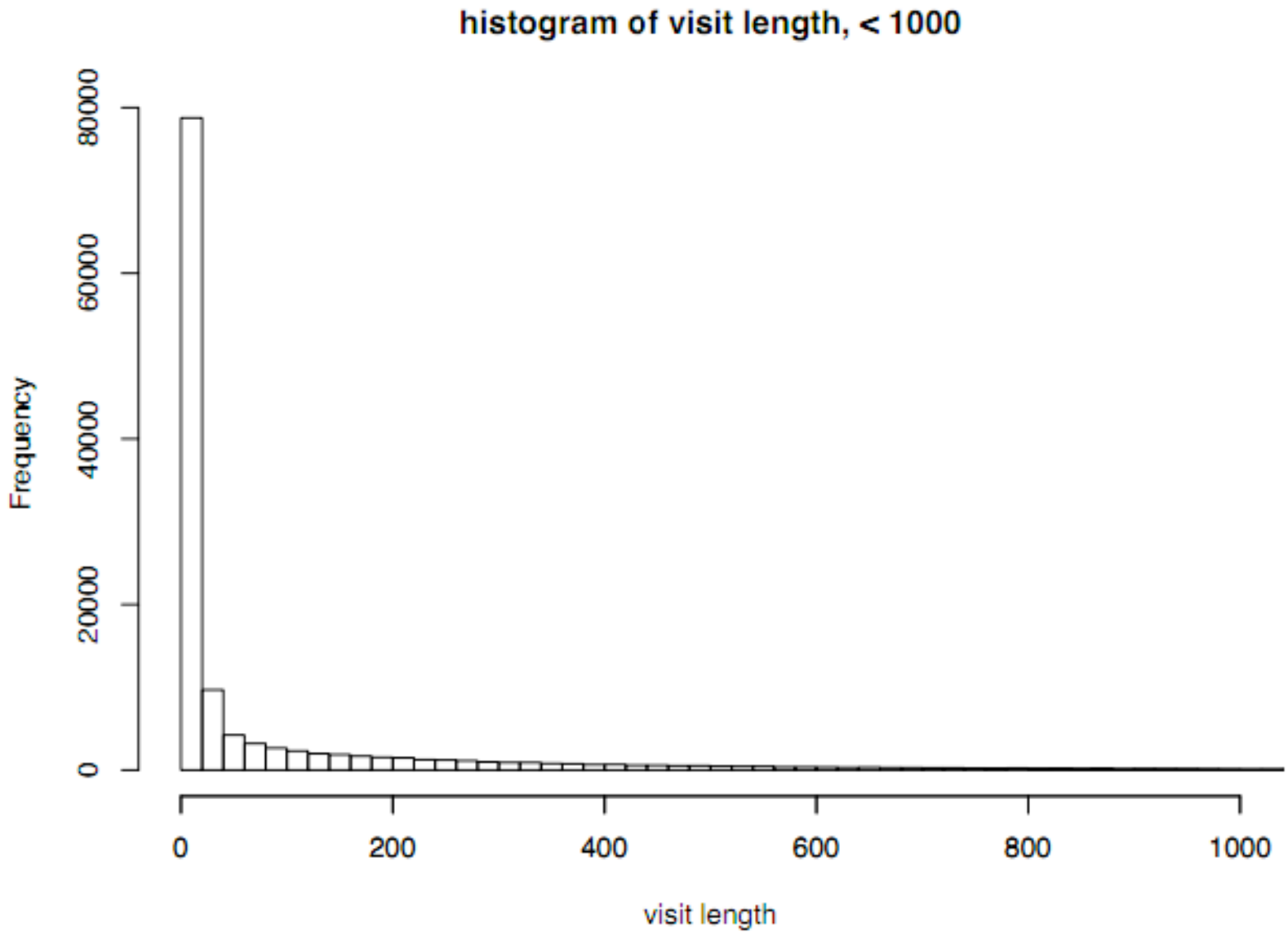
The data

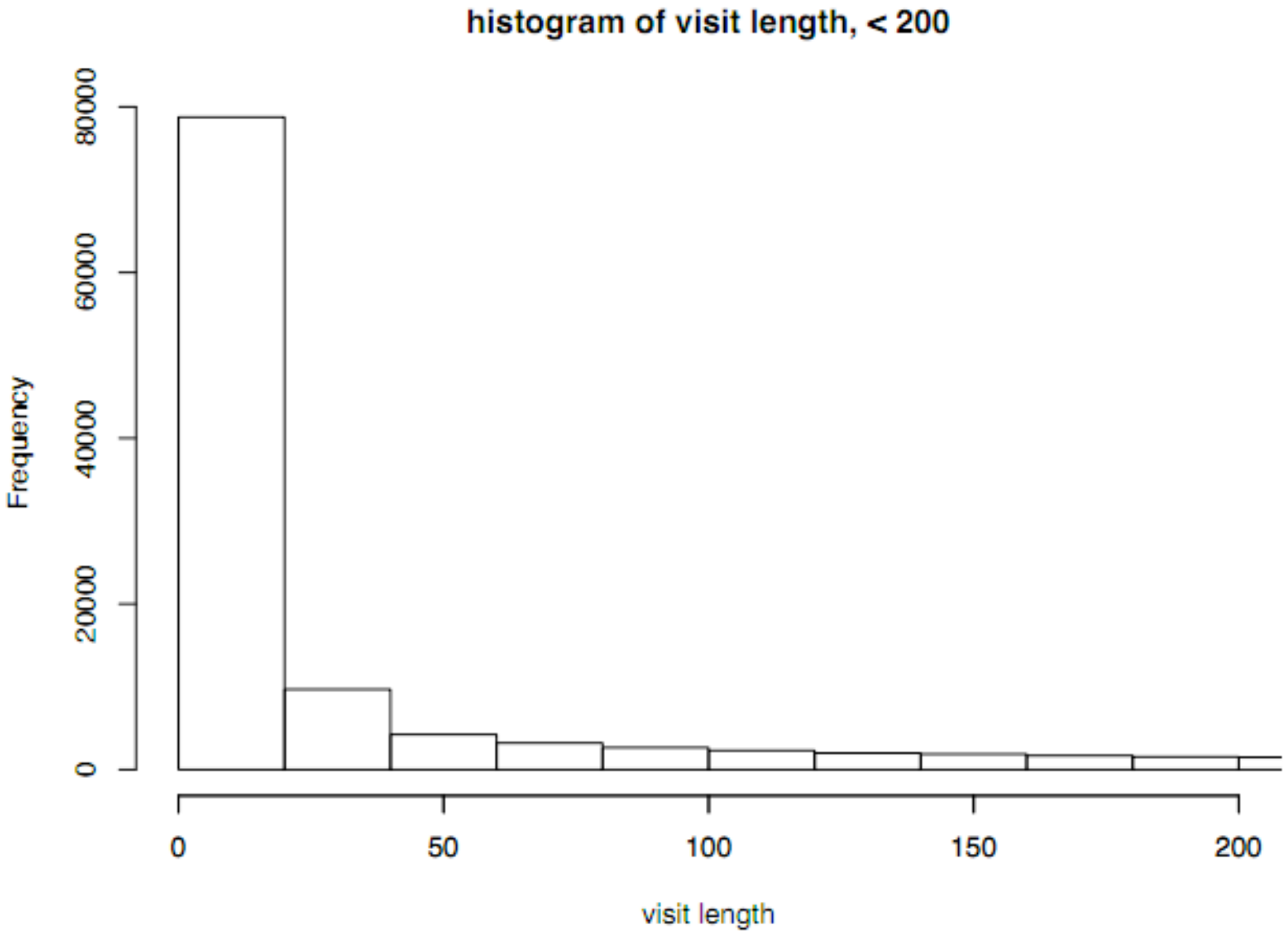
Testing is becoming a little tedious, so let's now move on to have a look at the rest of the data -- We will have one more test for you to perform in lab, but for now, let's consider some other topics

So, let's have a look at visit lengths -- As you see on the next page, no matter how tightly we restrict the x axis, we aren't getting a lot of new information; primarily we see a large number of points on the left and then a very long tail to the right


histogram of visit length
Frequency
0
20000
40000
60000
80000
0
5000
10000
15000
visit length


histogram of visit length, < 5000
Frequency
0
20000
40000
60000
80000
0
1000
2000
3000
4000
5000
visit length


histogram of visit length, < 1000
Frequency
0
20000
40000
60000
80000
0
200
400
600
800
1000
visit length


histogram of visit length, < 200
Frequency
0
20000
40000
60000
80000
0
50
100
150
200
visit length

Transformations

All this chopping is **not really making the distributional shape any clearer** because the dominant feature is simply the skew

To help us see more, we can consider **a family of monotone transformations** like square roots or the logarithm* -- These transformations that have a greater effect on larger values, bringing them in “closer”

* Here we’ll use the natural logarithm -- The choice is made largely to agree with statistical convention

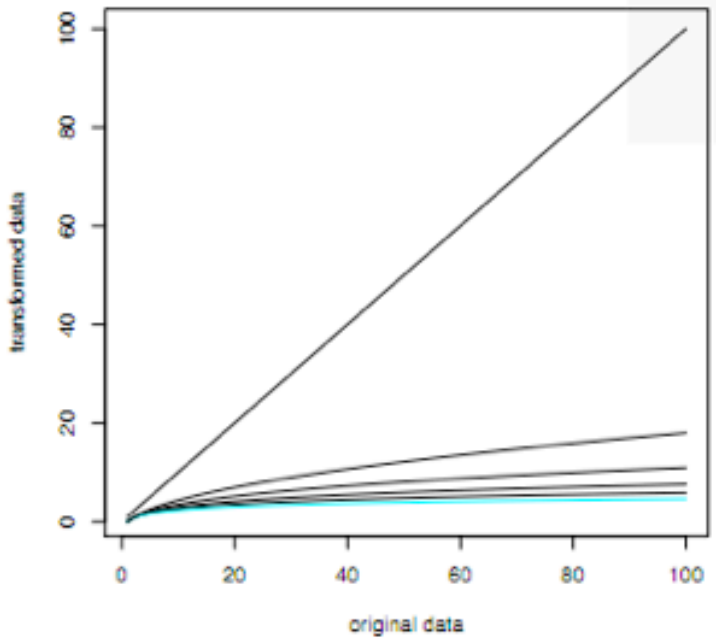
Transformations

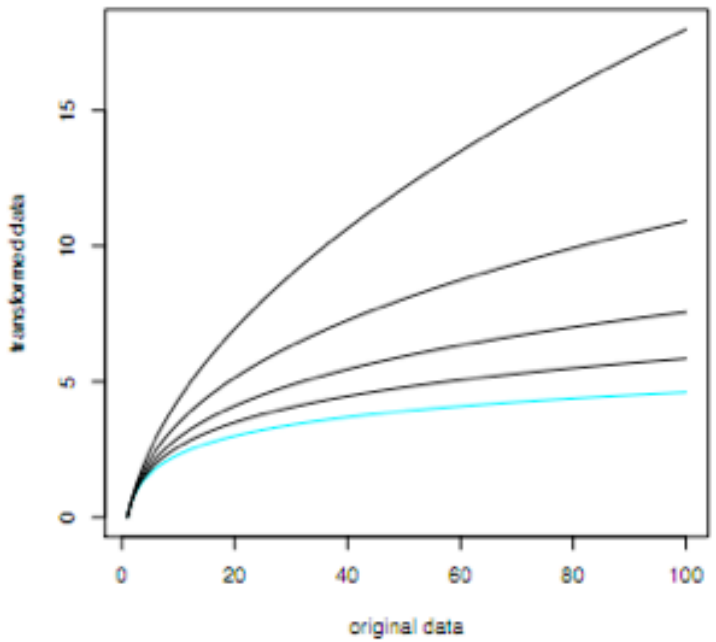
For data with a right skew, the square root, cube root, fourth root, and so on, can be used to define **a family of transformations** that all have the effect of taking big values and bringing them in closer to the rest of the data, to the smaller values

At the right we have a graph of this family (using the the square root, cube root and so on) to help you see what they are doing to the data; the top plot has the original scale (the straight line just being y=x) and the bottom plot is zoomed in on the curves -- What do you see?

Let's see how this works with our visit length data...

* technically the family is given by $f(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda > 0 \\ \log(x) & \lambda = 0 \end{cases}$





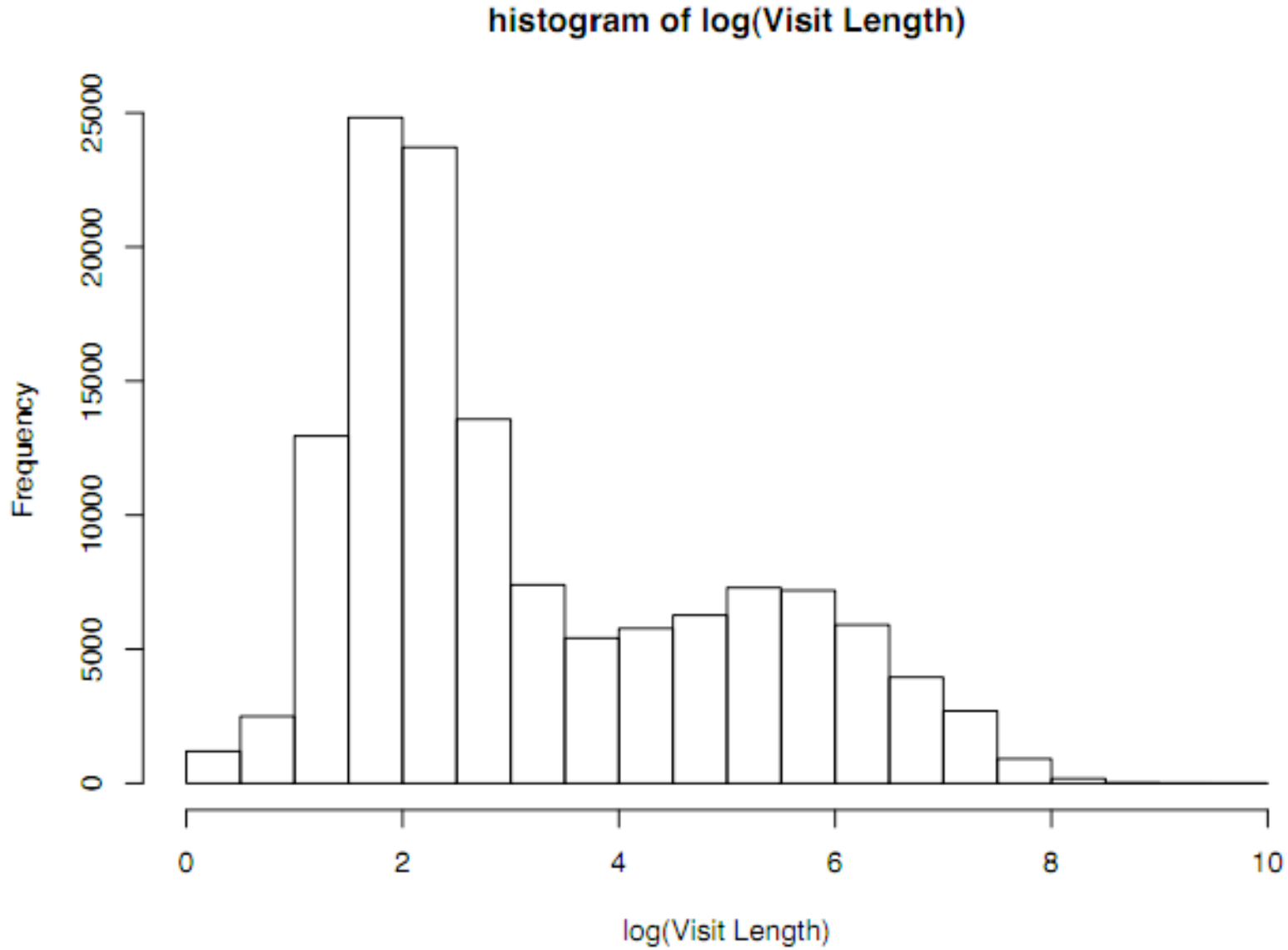
An aside

Notice that the members of this family are all monotonic transformations; that means if x < y then f(x) < f(y); put another way, **these transformation do not change the order of the data**

... and their effect on the median now becomes obvious: let $\tilde{x}$ be the median of the data points $x_1, x_2, \ldots, x_n$; the median of the transformed points $f(x_1), f(x_2), \ldots, f(x_n)$ is just $f(\tilde{x})$

Notice also, that as you work from the top curve to the bottom, the effect on the big values becomes more and more extreme; the square root of 10,000 is 100, the cube root is about 22, and the fifth root is about 6 -- in short, we are pulling the big values in closer and closer*

* To get the logarithm as a limit you need to use not just the simple roots but the family spelled out on the previous page

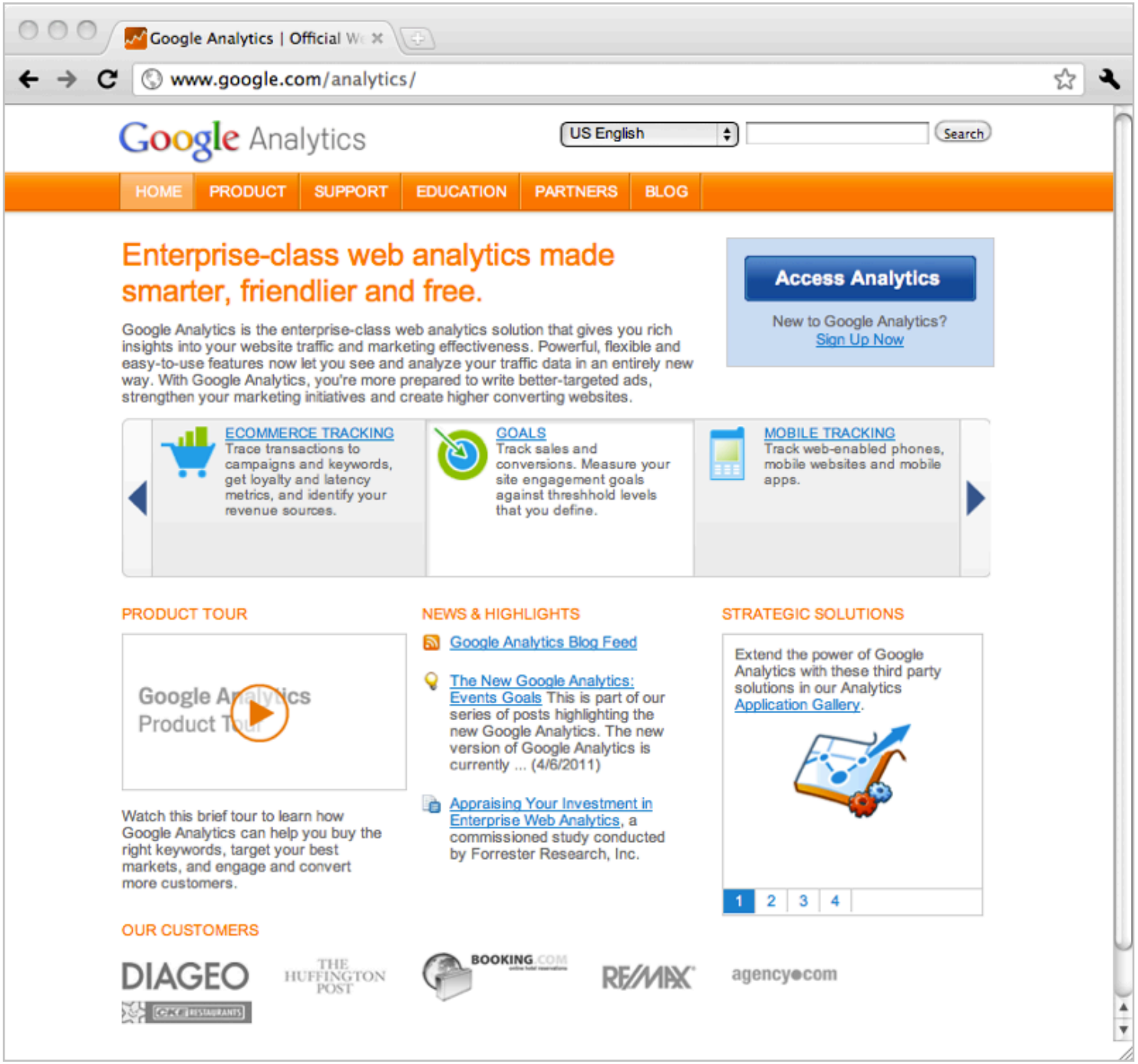

histogram of log(Visit Length)
Frequency
0
5000
10000
15000
20000
25000
0
2
4
6
8
10
log(Visit Length)

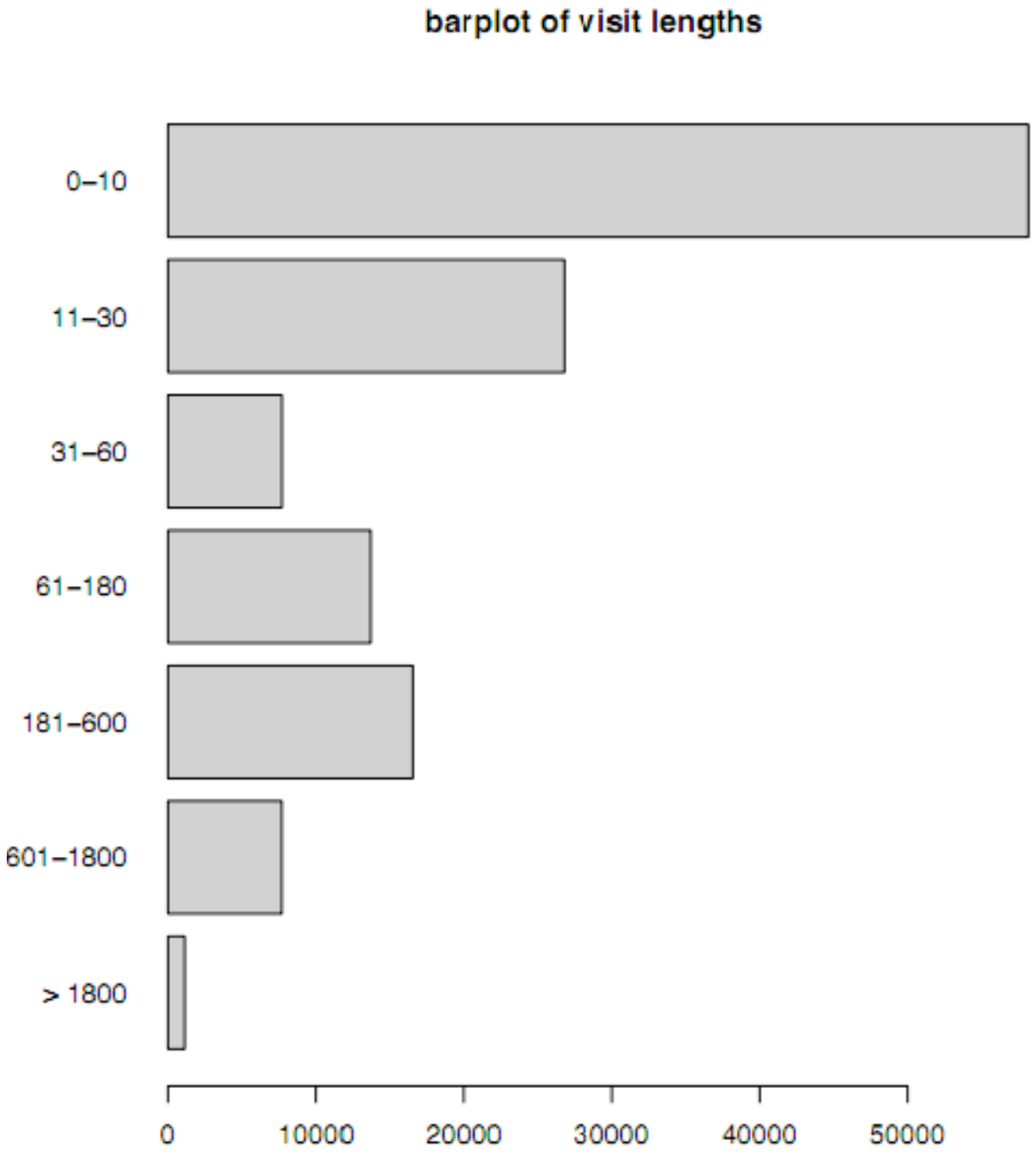
Alternatives

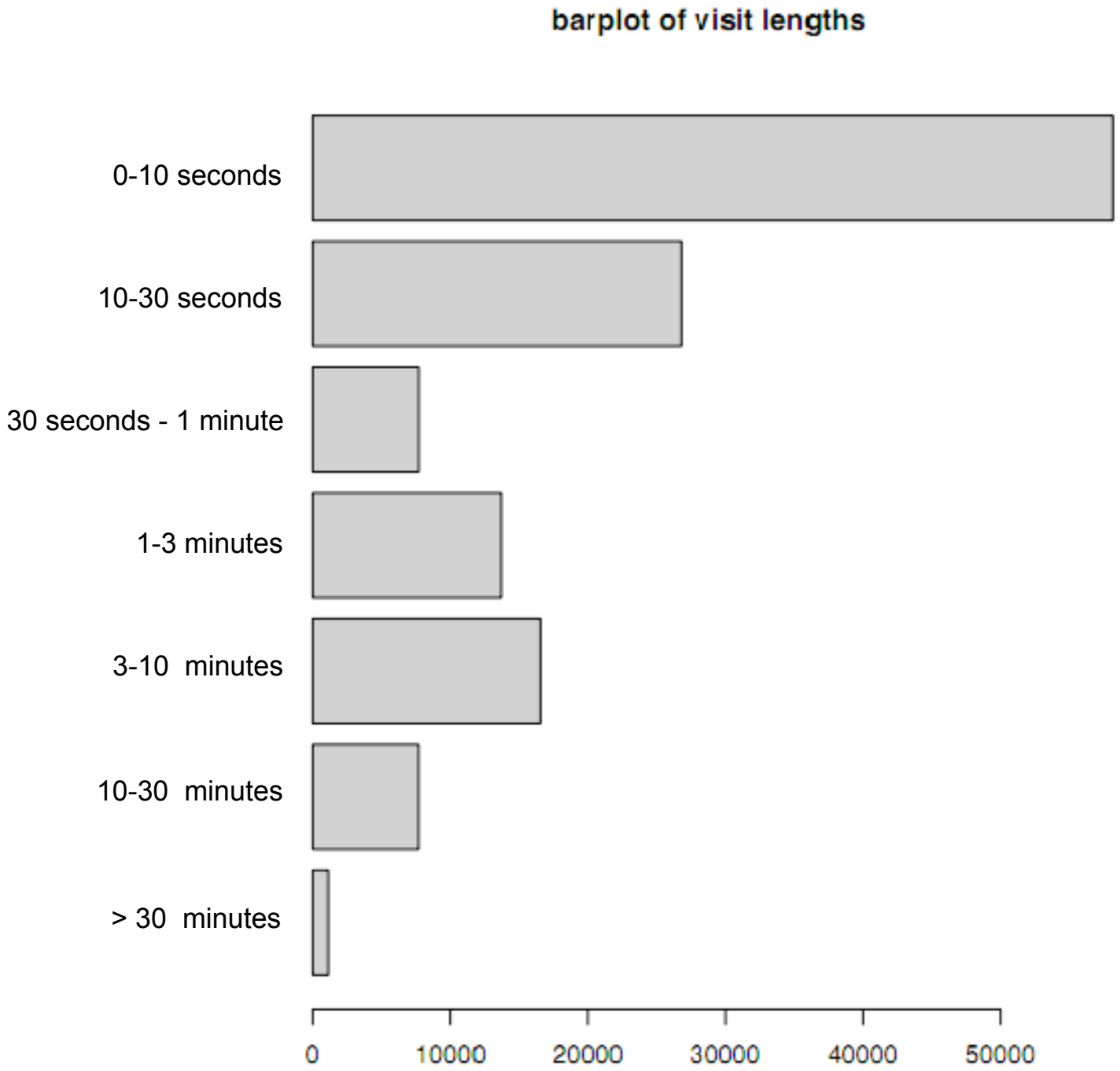
Yesterday, while dutifully analyzing data at the local Starbucks, I started talking to a man who was busy pouring over web site statistics; he owns a site related to retirement communities (although he is probably 40 years away from having to make use of his services)

He was using a platform offered by Google; by putting a piece of code on each of your web pages, you can have Google collect information about who is visiting your site, how long they stay and so on -- in short, the data we have from the New York Times

His view of Visit Length looked like this...


Google Analytics | Official W
www.google.com/analytics/
Google Analytics
US English
Search
HOME
PRODUCT
SUPPORT
EDUCATION
PARTNERS
BLOG
Enterprise-class web analytics made smarter, friendlier and free.
Google Analytics is the enterprise-class web analytics solution that gives you rich insights into your website traffic and marketing effectiveness. Powerful, flexible and easy-to-use features now let you see and analyze your traffic data in an entirely new way. With Google Analytics, you're more prepared to write better-targeted ads, strengthen your marketing initiatives and create higher converting websites.
Access Analytics
New to Google Analytics?
Sign Up Now
ECOMMERCE TRACKING
Trace transactions to campaigns and keywords, get loyalty and latency metrics, and identify your revenue sources.
GOALS
Track sales and conversions. Measure your site engagement goals against threshhold levels that you define.
MOBILE TRACKING
Track web-enabled phones, mobile websites and mobile apps.
PRODUCT TOUR
Google Analytics
Product Tour
Watch this brief tour to learn how Google Analytics can help you buy the right keywords, target your best markets, and engage and convert more customers.
NEWS & HIGHLIGHTS
Google Analytics Blog Feed
The New Google Analytics: Events Goals This is part of our series of posts highlighting the new Google Analytics. The new version of Google Analytics is currently ... (4/6/2011)
Appraising Your Investment in Enterprise Web Analytics, a commissioned study conducted by Forrester Research, Inc.
STRATEGIC SOLUTIONS
Extend the power of Google Analytics with these third party solutions in our Analytics Application Gallery.
1
2
3
4
OUR CUSTOMERS
DIAGEO
THE HUFFINGTON POST
BOOKING.COM
RE/MAX
agency●com

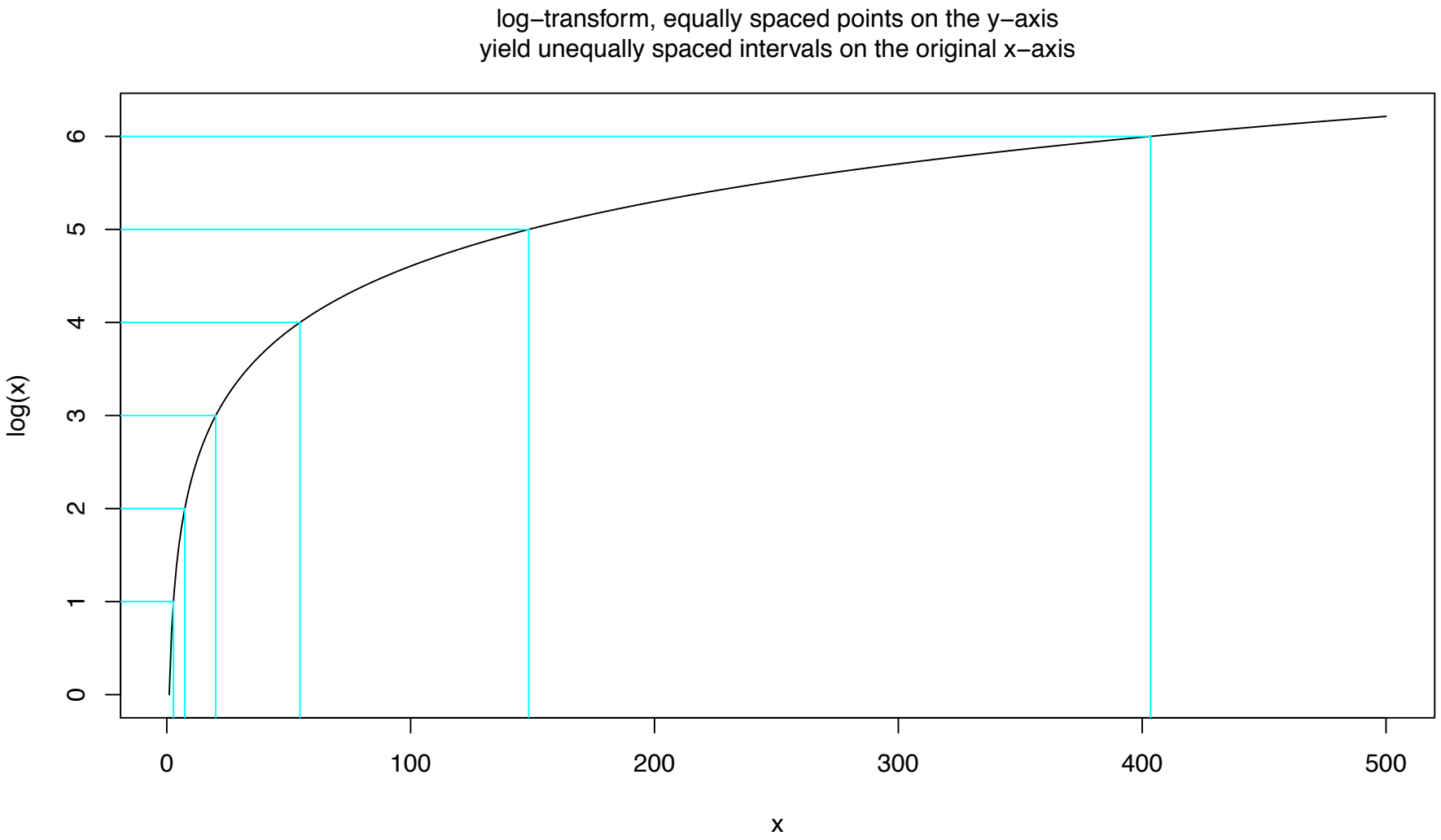

barplot of visit lengths
0–10
11–30
31–60
61–180
181–600
601–1800
> 1800
0
10000
20000
30000
40000
50000


barplot of visit lengths
0-10 seconds
10-30 seconds
30 seconds - 1 minute
1-3 minutes
3-10 minutes
10-30 minutes
> 30 minutes
0
10000
20000
30000
40000
50000

Alternatives

The cutoffs used by common web site analysis packages have a (sort of) **logarithmic feel** to them -- from 10 seconds to 30 seconds to a minute to 1.5 minutes to 10 minutes to 30 minutes; notice that even in this display, the bimodal nature of the data are clear

My hope is that this kind of presentation will demystify the log-transform -- we can think of creating a display by either c**reating bins that get longer as you go into the right tail**, or using equally spaced bins on the transformed data


log−transform, equally spaced points on the y−axis
yield unequally spaced intervals on the original x−axis
log(x)
0
1
2
3
4
5
6
0
100
200
300
400
500
x

Transformations

With that out of the way...

Our observational unit is a visit, and specifically the first visit to the Travel Section; in the distribution on the previous page we see one mode centered around 2 log-seconds -- which in our original units is exp(2) = 7.4 seconds -- and the other at around 5.5 log-seconds -- which is exp(5.5) = 245 seconds or about 4 minutes

Notice what we've done here; **we've transformed back to the original units** when talking about the log-Visit Lengths

Transformations

In some situations, **it is easier to work on one particular scale or another**; think about currency or measures of weight or volume -- fairly straightforward transformations

In some applications a logarithm might be the de facto measurement standard; but it hasn't really caught on for Visit Length -- instead we appeal to a transformation to help us see things about the data, but we have to back-transform to the original scale for presentation

1 U.S. dollar = 5.78338895 Danish kroner

1 US gallon = 3.78541178 litres

1 short ton = 2000 pounds

Transformations

In fact, non-linear scaling is used across science

Should we measure acidity in the concentration of H^+ ions (linear) or PH (logarithmic)?

Should we measure the magnitude of an earthquake in mm of amplitude (linear) or on the Richter scale (logarithmic)?

Should eyeglass lenses be measured in terms of focal length in cm (linear) or dioptres (a reciprocal)?

Transformations

Now, let's consider what it means to work with the log of Visit Lengths -- In particular, let's assume we have some number of data values $x_1, \ldots, x_n$

$$\frac{\log x_1 + \log x_2 + \cdots + \log x_n}{n} = \frac{1}{n} \log x_1 x_2 \cdots x_n$$

$$= \log \left[(x_1 x_2 \cdots x_n)^{\frac{1}{n}} \right]$$

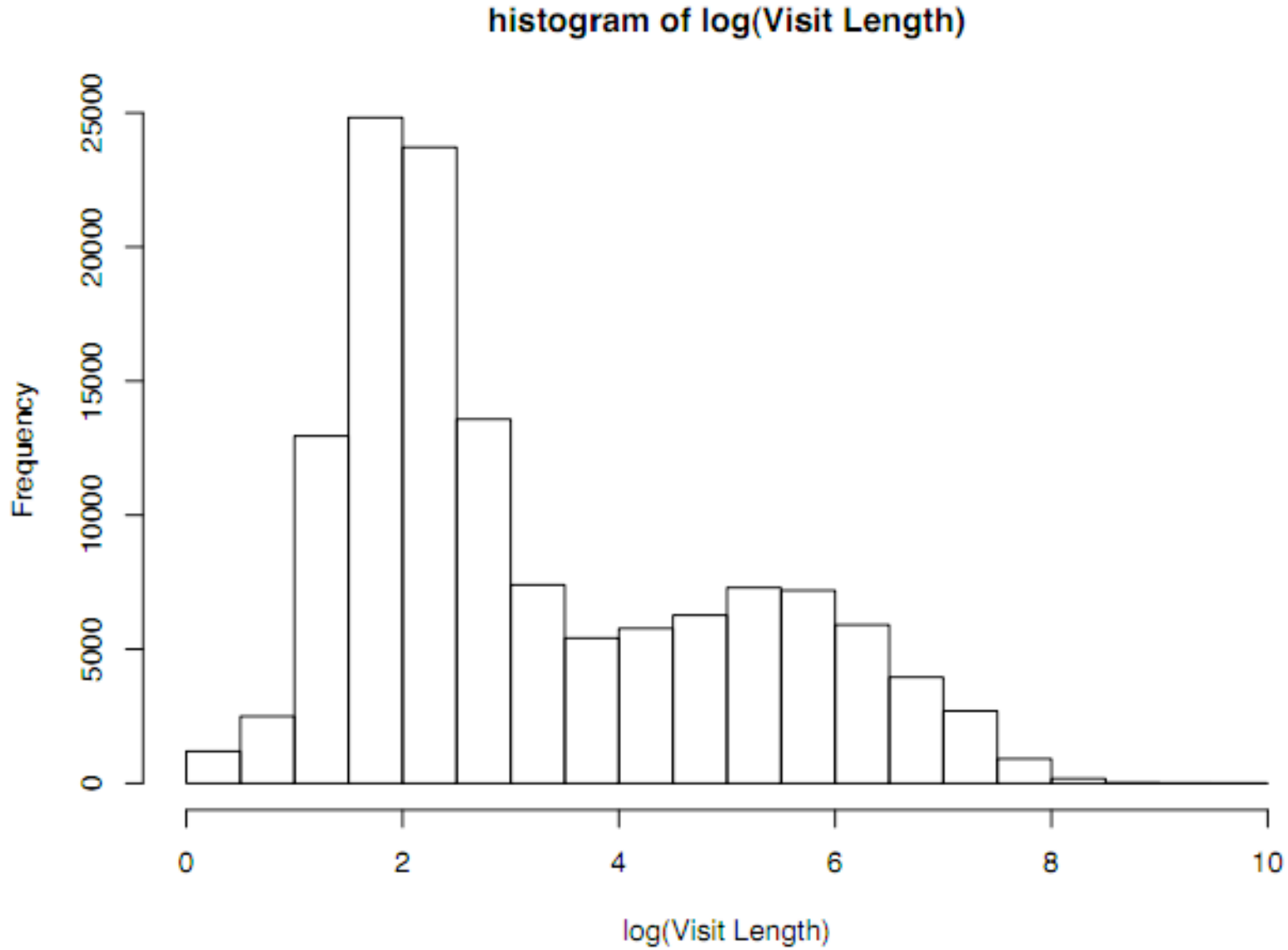
This means that the average of the logged values gives us the logarithm of the geometric mean -- when we back-transform (by exponentiating) we have the geometric mean of $x_1, x_2, \ldots, x_n$, a quantity that is back in the original units

* here log = natural log!

Question

Now, back to the log-Visit Length distribution: We have seen two modes, which implies that there are two kinds of visits taking place -- when faced with a distribution like this, there is really only one question that comes to mind...

"Why?" What variables in our data set might help us explain the two peaks?

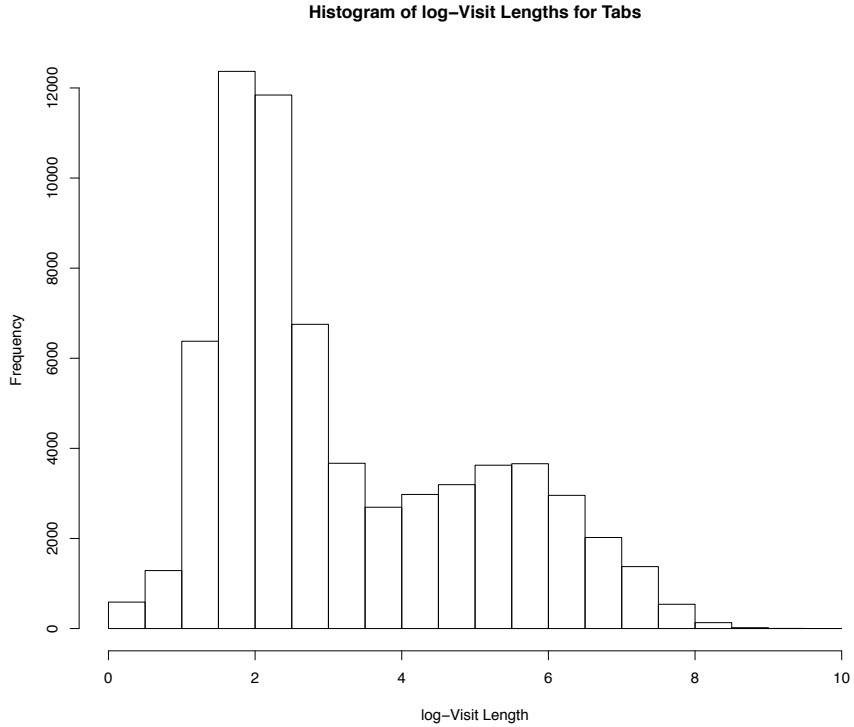

histogram of log(Visit Length)
Frequency
0
5000
10000
15000
20000
25000
0
2
4
6
8
10
log(Visit Length)

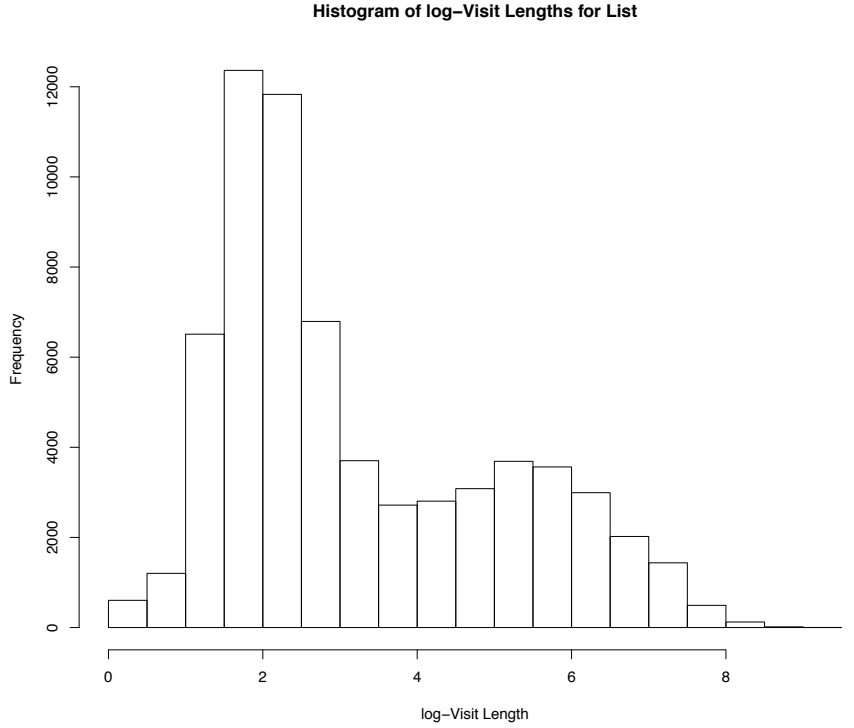
Two groups?

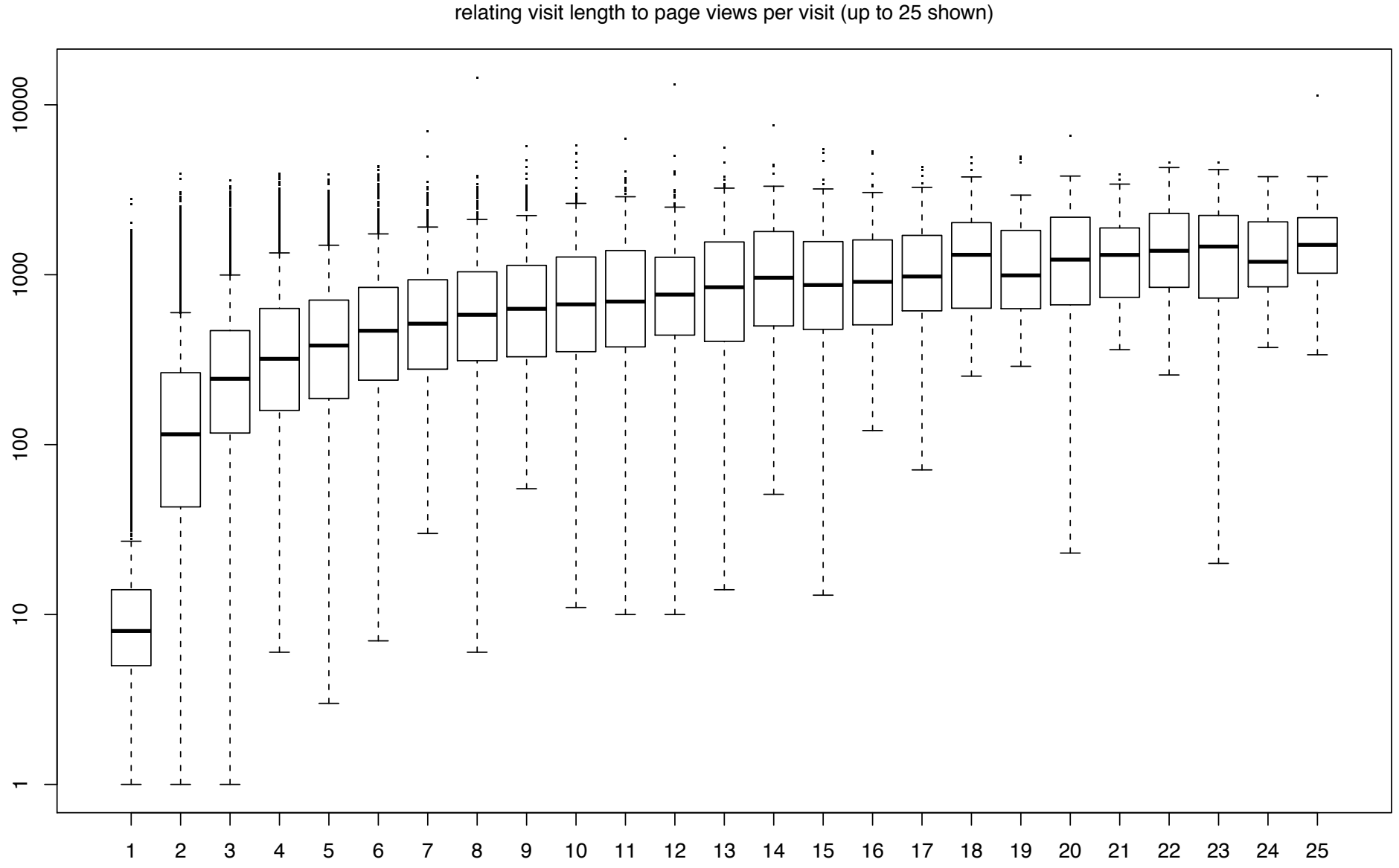
An obvious answer is that somehow the groups are the result of our experiment -- That one bump represents Tabs and the other Lists

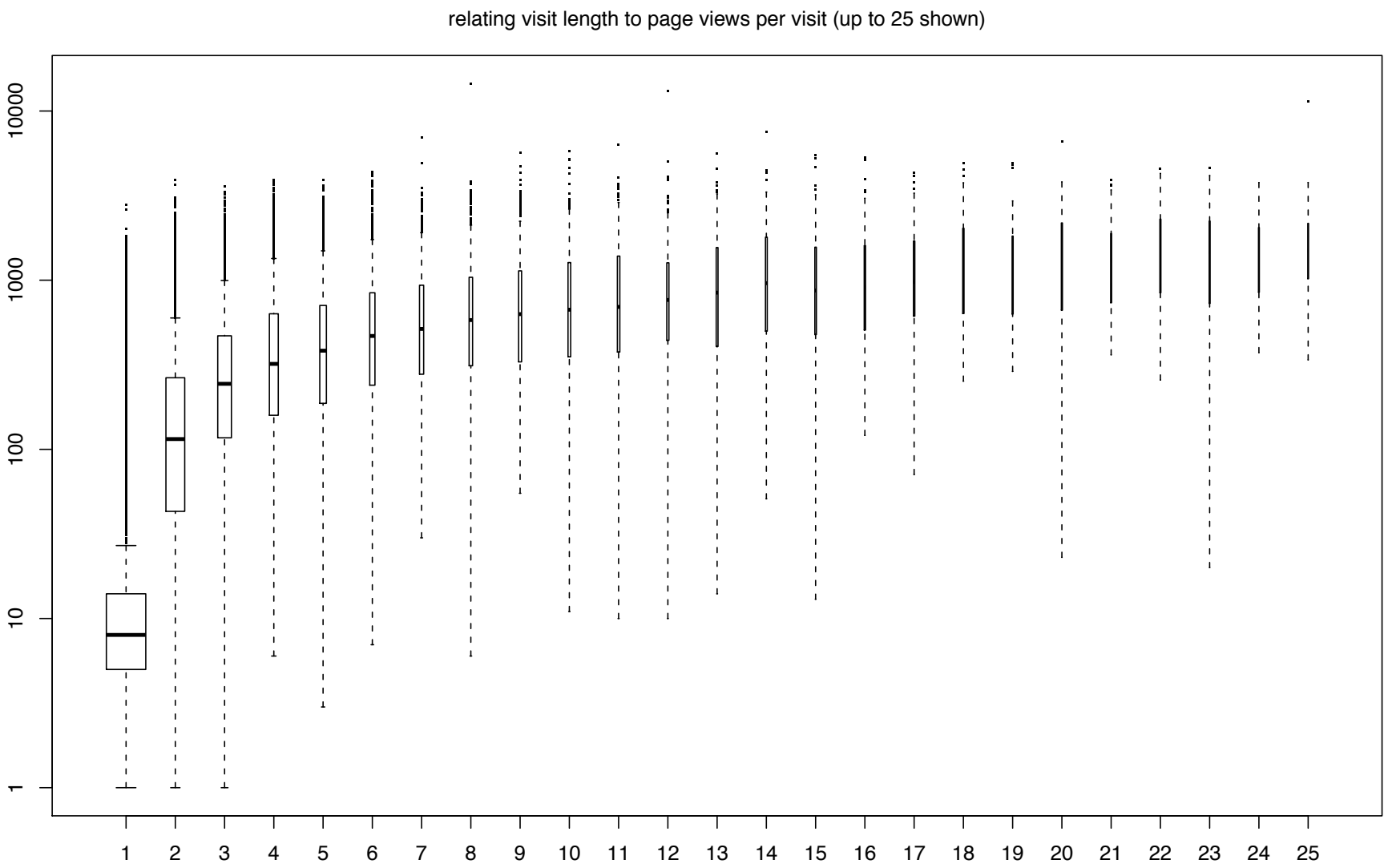
On the right we have two histograms, the top for just the log-visit lengths for those people seeing Tabs and the bottom for those seeing Lists

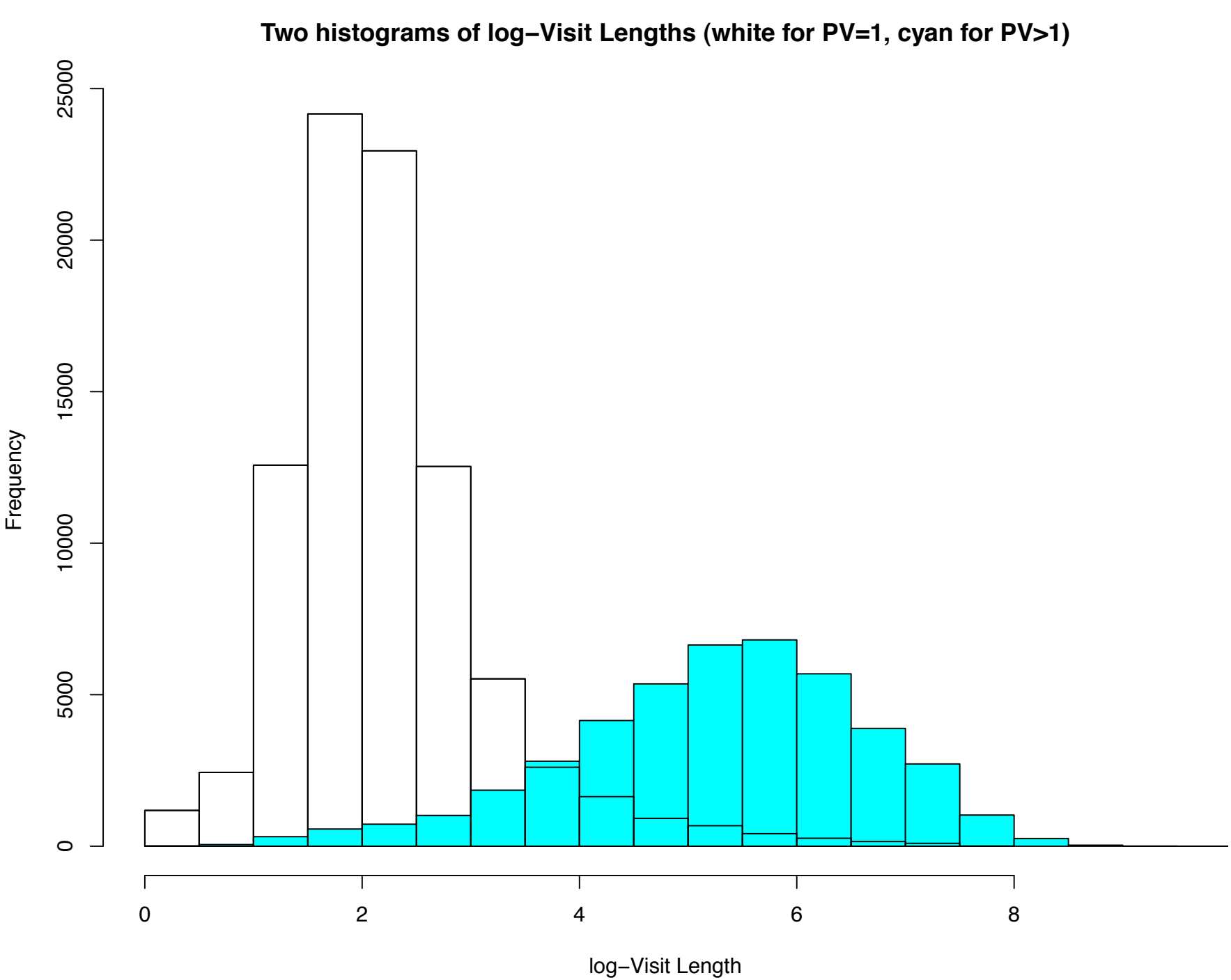
Any difference?






relating visit length to page views per visit (up to 25 shown)
10000
1000
100
10
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25


relating visit length to page views per visit (up to 25 shown)
10000
1000
100
10
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25


Two histograms of log−Visit Lengths (white for PV=1, cyan for PV>1)
Frequency
0
5000
10000
15000
20000
25000
0
2
4
6
8
log−Visit Length

Two groups

What we can see from this is that there are essentially two groups in the data -- **Those that visit only one page** (and their visit lengths correspond to the time it takes to load the page) and **those that visit more pages**

In general, when we identify groups or clusters in the data, we want to see what causes them -- It would have been amazing if our experiment created the groups, but in this case it has more to do with how visit length is calculated

With that out of the way, we can still legitimately ask if there is a difference between the two groups based on visit length or log-visit length...

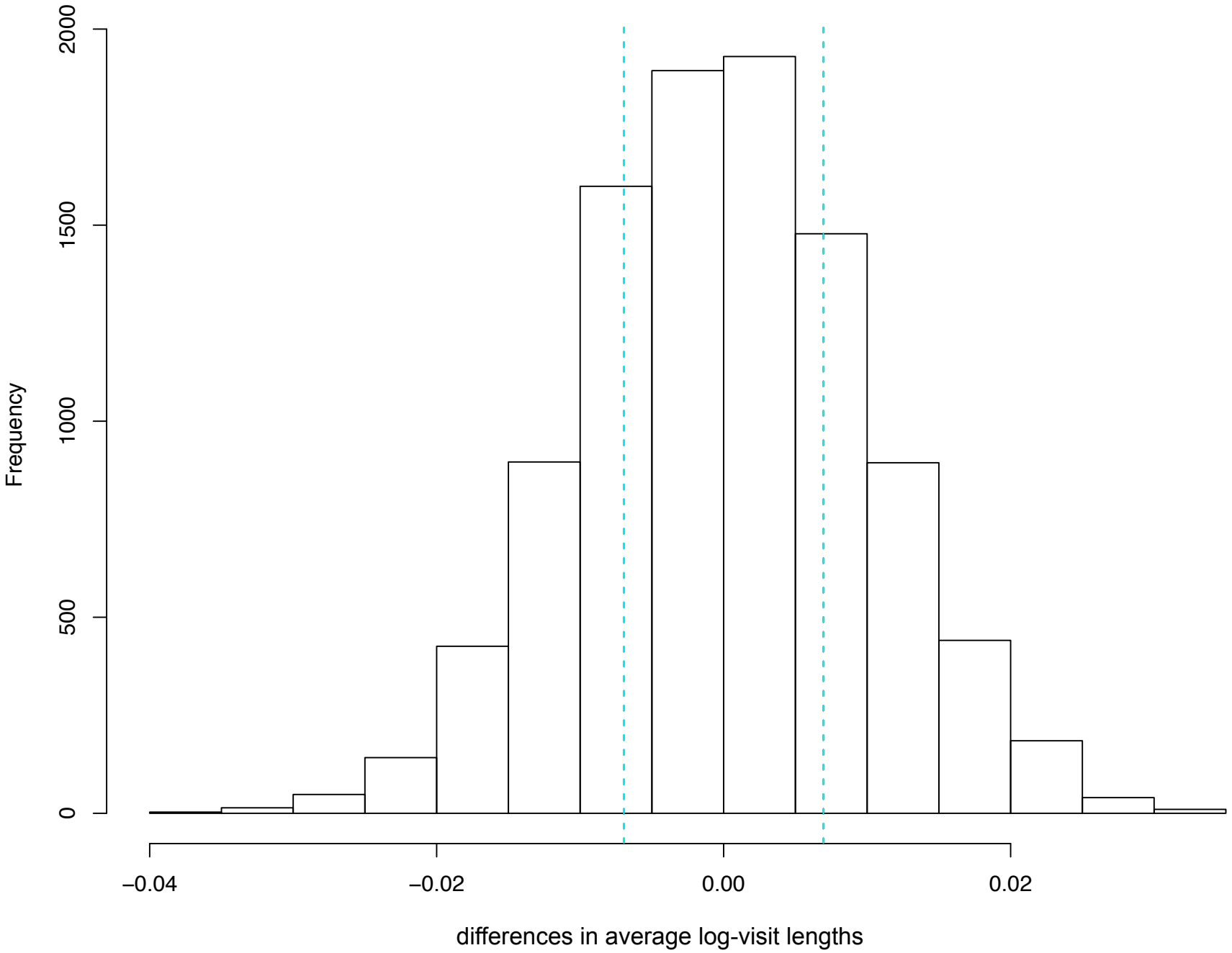
Log-Visit length

We can now have a look at whether the change had an impact on the average log-visit length -- The **null hypothesis** (the strawman) is that there is **no difference and we'll take as an alternative that there is a difference**

As with pageviews, we will take as our test statistic **the absolute value of the difference between the average log-visit lengths under "Tabs" and "Lists"** -- Large values of this statistic will indicate a difference between the two designs (and "large" means large positive or negative differences)

Under the null hypothesis of no difference between "Tabs" and "Lists", **each visitor would have remained on the site for the same number of seconds** no matter which design they were shown -- This means we can re-randomize and **look at the variations in our test statistic that are typical under the null**...

```
> mean(log(subset(travel$VisitLength,travel$Variation=="List")))
[1] 3.225758

> mean(log(subset(travel$VisitLength,travel$Variation=="Tabs")))
[1] 3.232709

> d <- abs(mean(log(subset(travel$VisitLength,travel$Variation=="List"))) -
          mean(log(subset(travel$VisitLength,travel$Variation=="Tabs"))))

# hold the results from 10,000 re-randomizations

> out <- rep(0,10000)

> for(i in 1:10000){

     treat <- sample(travel$Variation)
     out[i] <- mean(log(subset(travel$VisitLength,treat=="List")))
                mean(log(subset(travel$VisitLength,treat=="Tabs")))
 }

# and have a look!

> hist(out)
> abline(v=c(-d,d),lty="dashed")

# the p-value
> mean(abs(out)> d)
[1] 0.4846
```


Histogram of re−randomized Tabs/Lists, differences in average log−visit lengths
Frequency
0
500
1000
1500
2000
−0.04
−0.02
0.00
0.02
differences in average log-visit lengths

Some details

When facing badly skewed data, and a question like the one we've been discussing, evaluating the difference between treatment and control where the alternative is that one group or the other tends to have larger values, you have several options

1. Use something like the mean on the original data scale

2. Apply a "tail-compression" transformation like the log before analysis

3. Use a more "robust" statistic like the median, the trimmed mean or even the Windsorized mean

4. If you are just dealing with one or two extreme outliers, remove them then use the mean

5. Reduce everything to ranks

Some details

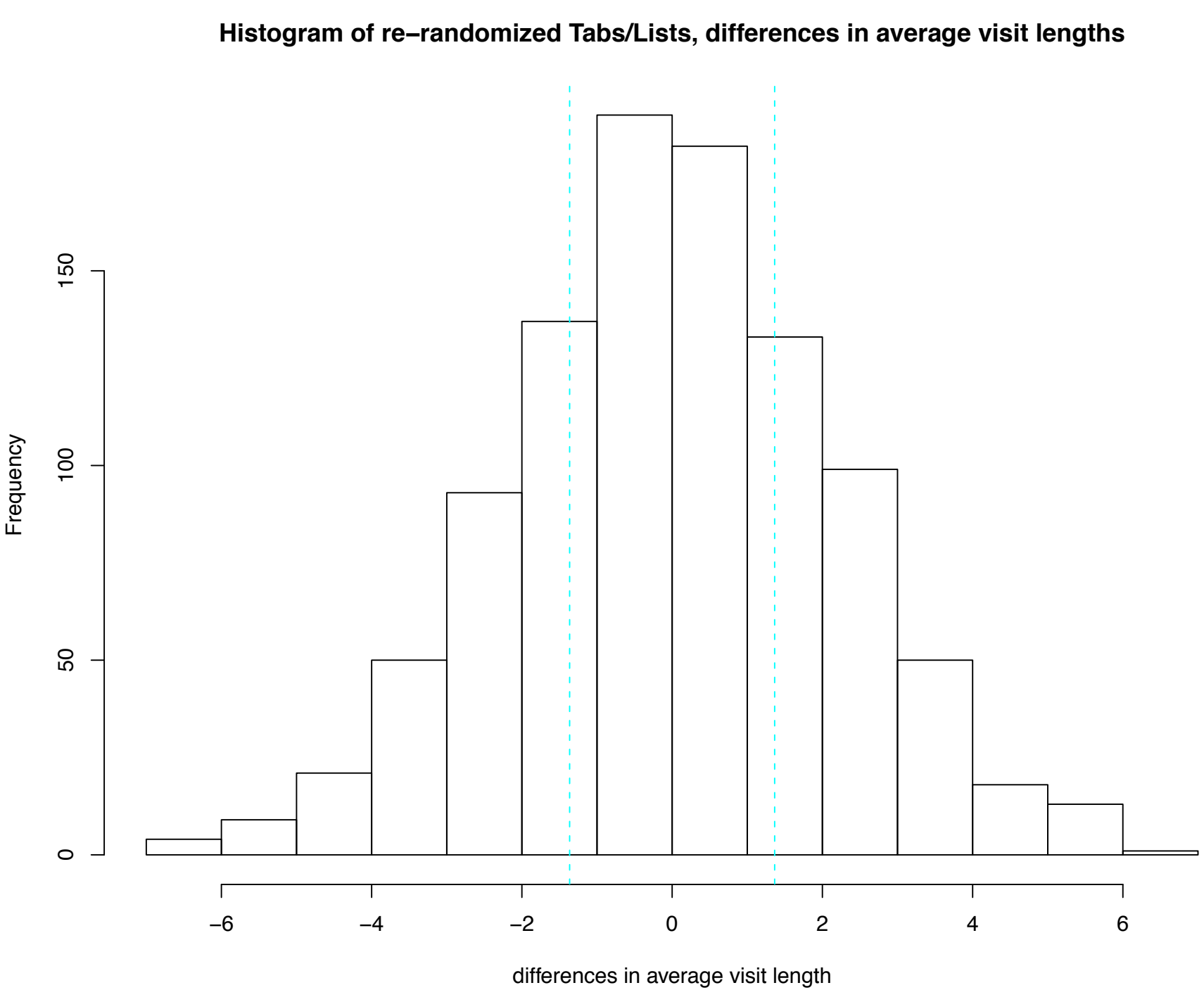
The typical introductory statistics class will focus almost exclusively on the mean and the differences between group means as a test statistic -- It is done primarily because the overall “standard” framework being taught works best with these kinds of statistics

With re-randomization, you will probably not go very wrong if you take the data on the original scale and perform the test -- the framework will provide you with a reasonable null distribution that you can compare against

Keep in mind that no matter what test statistic we chose, **our tests will always have the exact significance level we set in advance** -- By the design of the test, we are **always going to make a mistake 5% of the time** (this isn’t true for other procedures you might have heard about in other classes)

On the flip side, **there are questions of power**; how well does our test statistic perform at **being able to correctly spot a false null hypothesis**? Some test statistics will be better than others, many will perform about the same

```
# repeat our analysis without the log's

> mean(subset(travel$VisitLength,travel$Variation=="List"))
[1] 149.4573

> mean(subset(travel$VisitLength,travel$Variation=="Tabs"))
[1] 150.8219

> d <- abs(mean(subset(travel$VisitLength,travel$Variation=="List")) -
           mean(subset(travel$VisitLength,travel$Variation=="Tabs")))

# hold the results from 10,000 re-randomizations

> out <- rep(0,10000)

> for(i in 1:10000){

     treat <- sample(travel$Variation)
     out[i] <- mean(subset(travel$VisitLength,treat=="List"))
                 mean(subset(travel$VisitLength,treat=="Tabs"))
 }

# and have a look!

> hist(out)
> abline(v=c(-d,d),lty="dashed")

# the p-value
> mean(abs(out)> d)
[1] 0.525
```


Histogram of re–randomized Tabs/Lists, differences in average visit lengths
Frequency
0
50
100
150
–6
–4
–2
0
2
4
6
differences in average visit length

The bottom line

The bottom line is that from the standpoint of testing, **we seek test statistics that will respond well to departures from the null hypothesis** -- the re-randomization approach is different from the usual framework you might have seen that focuses purely on means or differences in means as the only option

During this class you will be primarily told which statistic to use and why; so don't worry -- it seemed important to give you a sense of the choices involved and that statistics is bigger than just means and proportions

Or, more to the point, that you are now capable of analyzing much more elaborate questions with this re-randomization toolbox -- Let's have a look at one!

Natural experiments

We're now going to spend a little time on one more example -- This is not a clinical trial but has been termed **a "natural experiment" that leads to an interesting set of questions** we are now prepared to answer

The example comes from a paper that appeared on one of the "journals of record" for statistics -- The author and volume information is given below, and I'll hold off on the full record until we've worked through the problem (the title reveals a bit too much)

The authors, **Ho and Imai have graciously provided us with the data** so you can have a hand at the analysis as well!

Daniel E. Ho and Kosuke Imai,
Journal of the American Statistical Association, Vol. 101, No. 475.

The 2003 recall election

In 2003, California held its **first gubernatorial recall election,** and ultimately the incumbent Democratic Governor Gray Davis was replaced by Republican candidate Arnold Schwarzenegger

To refresh your memories, the California recall system was established in the early 1900s and was introduced to **enact the "will of the people"** who might collectively decide that **an elected official was not doing their job** -- Prior to Davis, many governors had faced the threat of a recall, but in each case had managed to avoid the vote

THE RECALL ELECTION

Gov. Davis Is Recalled; Schwarzenegger Wins

'I Will Not Fail You,' the Republican Victor Promises

October 08, 2003 | Michael Finnegan | Times Staff Writer

Arnold Schwarzenegger won the historic California recall election Tuesday as a tide of voter anger toppled Gray Davis just 11 months after the Democrat had been reelected governor.

In a popular revolt unmatched in the 92 years that Californians have held the power to recall elected officials, voters chose a Republican film star with no government experience to replace an incumbent steeped for three decades in state politics.

* OK, the ad paired with this particular story is amusing...

2003 recall election

The recall ballot consisted of **two parts** -- In the first you were asked if you thought **Gray Davis should be recalled** and in the second you were asked to **choose a successor** (assuming a majority of voters responded “yes” to the first question)

The list of choices that year, however, was quite long, and in all, **135 candidates qualified to run as a Davis’ replacement --** Here is a list of their names and how they appeared on sample ballots from assembly districts in Sonoma and Orange Counties

ADAM	ADAMS	ALEXSTJAMES	ANDERSON
ANGELYNE	ARIF	BADIOZAMANI	BAJWA
BEARD	BEYER	BHOLA	BLYCHESTER
BOCK	BRITTON	BROWN	BURTON
BUSTAMANTE	CAMEJO	CARSON	CHAMBERS
CHELI	CLEMENTS	COLEMAN	COOK
CULLENBINE	DAVIS	DOLE	EDWARDS
FARRELL	FEINSTEIN	FLYNT	FONTANES
FORTE	FOSS	FRIEDMAN	GALLAGHER
GORMAN	GOSSE	GREEN	GRISHAM
GRUENER	GUZZARDI	HALL	HAMIDI
HANLON	HENDERSON	HERNANDEZ	HICKEY
HOFFMANN	HUFFINGTON	ISSA	JACKSON
KELLY	KENNEDY	KESSINGER	KIMBALL
KNAPP	KOREVAAR	KUNZMAN	LANE
LEONARD	LEWIS	LOUIE	MACALUSO
MAILANDER	MANNHEIM	MARGOLIN	MARIANO
MARTORANA	MCCARTHY	MCCLAIN	MCCLINTOCK
MCMAHON	MCNEILLY	MEDNICK	MEHR
MILLER	MOBLEY	MOCK	MORTENSEN
MUSILLI	NAVE	NEWMANII	PADILLA
PALMIERI	PAWLIK	PETERS	PINEDA
PRADY	PRICE	QUINN	RAINFORTH
RAMIREZ	RANKEN	RENZ	RICHARDS
RICHTER	RIGHTMYER	ROBINSON	ROSCOE
RUSHFORD	RUSSELL	SAFFORD	SAMS
SCHEIDLE	SCHMIER	SCHWARTZMAN	SCHWARZENEGGER
SIMMONS	SIMON	SMITH	SPRAGUE
SPROUL	STRAUSS	SYLVESTER	TAYLOR
TEMPLIN	TILLEY	TRACY	TSANGARES
UEBERROTH	VALDEZ	VANDEVENTER	VANN
VAUGHN	VO	WALKERC	WALKERM
WALTON	WATTS	WEBER	WEIR
WINTERS	WOZNIAK	ZELLHOEFER	

Statewide Special Election
Orange County, California
October 07, 2003

OFFICIAL BALLOT

Instruction Note:

HOW TO VOTE:
- To vote, fill in and BLACKEN completely the rectangle to the left of any candidate or to the left of the word "YES" or "NO".
- Vote for only ONE of the 135 candidates, OR enter a write-in candidate in the space provided.
- Use only the special marking device provided. (Absentee voters should use a dark pen or a #2 pencil.)

Shall GRAY DAVIS be recalled (removed) from the office of Governor?

- ☐ YES
- ☐ NO

Candidates to succeed GRAY DAVIS as Governor if he is recalled:

Vote for One

- ☐ **B.E. SMITH** *Independent* - Lecturer
- ☐ **DAVID RONALD SAMS** *Republican* - Businessman/Producer/Writer
- ☐ **JAMIE ROSEMARY SAFFORD** *Republican* - Business Owner
- ☐ **LAWRENCE STEVEN STRAUSS** *Democratic* - Lawyer/Businessperson/Student
- ☐ **ARNOLD SCHWARZENEGGER** *Republican* - Actor/Businessman
- ☐ **GEORGE B. SCHWARTZMAN** *Independent* - Businessman
- ☐ **MIKE SCHMIER** *Democratic* - Attorney
- ☐ **DARRIN H. SCHEIDLE** *Democratic* - Businessman/Entrepreneur
- ☐ **BILL SIMON** *Republican* - Businessman
- ☐ **RICHARD J. SIMMONS** *Independent* - Attorney/Businessperson
- ☐ **CHRISTOPHER SPROUL** *Democratic* - Environmental Attorney
- ☐ **RANDALL D. SPRAGUE** *Republican* - Discrimination Complaint Investigator
- ☐ **TIM SYLVESTER** *Democratic* - Entrepreneur
- ☐ **STEPHEN L. KNAPP** *Republican* - Engineer
- ☐ **KELLY P. KIMBALL** *Democratic* - Business Executive
- ☐ **D.E. KESSINGER** *Democratic* - Paralegal/Property Manager
- ☐ **EDWARD "ED" KENNEDY** *Democratic* - Businessman/Educator
- ☐ **TREK THUNDER KELLY** *Independent* - Business Executive/Artist
- ☐ **JERRY KUNZMAN** *Independent* - Chief Executive Officer
- ☐ **PETER V. UEBERROTH** *Republican* - Businessman/Olympics Advisor
- ☐ **BILL PRADY** *Democratic* - Television Writer/Producer
- ☐ **DARIN PRICE** *Natural Law* - University Chemistry Instructor
- ☐ **GREGORY J. PAWLIK** *Republican* - Realtor/Businessman
- ☐ **LEONARD PADILLA** *Independent* - Law School President
- ☐ **RONALD JASON PALMIERI** *Democratic* - Gay Rights Attorney
- ☐ **CHARLES "CHUCK" PINEDA JR.** *Democratic* - State Hearing Officer
- ☐ **HEATHER PETERS** *Republican* - Mediator
- ☐ **ROBERT "BUTCH" DOLE** *Republican* - Small Business Owner
- ☐ **SCOTT DAVIS** *Independent* - Business Owner
- ☐ **RONALD J. FRIEDMAN** *Independent* - Physician
- ☐ **GENE FORTE** *Republican* - Executive Recruiter/Entrepreneur
- ☐ **DIANA FOSS** *Democratic* -
- ☐ **LORRAINE (ABNER ZURD) FONTANES** *Democratic* - Film Maker
- ☐ **WARREN FARRELL** *Democratic* - Fathers' Issues Author
- ☐ **DAN FEINSTEIN** *Democratic* -
- ☐ **LARRY FLYNT** *Democratic* - Publisher
- ☐ **DARRYL L. MOBLEY** *Independent* - Businessman/Entrepreneur
- ☐ **JEFFREY L. MOCK** *Republican* - Business Owner
- ☐ **BRUCE MARGOLIN** *Democratic* - Marijuana Legalization Attorney
- ☐ **GINO MARTORANA** *Republican* - Restaurant Owner
- ☐ **PAUL MARIANO** *Democratic* - Attorney
- ☐ **ROBERT C. MANNHEIM** *Democratic* - Retired Businessperson
- ☐ **FRANK A. MACALUSO, JR.** *Democratic* - Physician/Medical Doctor
- ☐ **PAUL "CHIP" MAILANDER** *Democratic* - Golf Professional
- ☐ **DENNIS DUGGAN MCMAHON** *Republican* - Banker
- ☐ **MIKE MCNEILLY** *Republican* - Artist
- ☐ **MIKE P. MCCARTHY** *Independent* - Used Car Dealer
- ☐ **BOB MCCLAIN** *Independent* - Civil Engineer
- ☐ **TOM MCCLINTOCK** *Republican* - State Senator
- ☐ **JONATHAN MILLER** *Democratic* - Small Business Owner
- ☐ **CARL A. MEHR** *Republican* - Businessman
- ☐ **SCOTT A. MEDNICK** *Democratic* - Business Executive
- ☐ **DORENE MUSILLI** *Republican* - Parent/Educator/Businesswoman
- ☐ **VAN VO** *Republican* - Radio Producer/Businessman
- ☐ **PAUL W. VANN** *Republican* - Financial Planner
- ☐ **JAMES M. VANDEVENTER, JR.** *Republican* - Salesman/Businessman
- ☐ **BILL VAUGHN** *Democratic* - Structural Engineer
- ☐ **MARC VALDEZ** *Democratic* - Air Pollution Scientist
- ☐ **MOHAMMAD ARIF** *Independent* - Businessman

MEASURES SUBMITTED TO THE VOTERS

STATE

Proposition 53

FUNDS DEDICATED FOR STATE AND LOCAL INFRASTRUCTURE. LEGISLATIVE CONSTITUTIONAL AMENDMENT.

Generally dedicates up to 3% of General Fund revenues annually to fund state and local (excluding school and community college) infrastructure projects. Fiscal Impact: Dedication of General Fund revenues for state and local infrastructure. Potential transfers of $850 million in 2006-07, increasing to several billions of dollars in future years, under specified conditions.

- ☐ YES
- ☐ NO

Proposition 54

CLASSIFICATION BY RACE, ETHNICITY, COLOR, OR NATIONAL ORIGIN. INITIATIVE CONSTITUTIONAL AMENDMENT.

Prohibits state and local governments from classifying any person by race, ethnicity, color, or national origin. Various exemptions apply. Fiscal Impact: The measure would not result in a significant fiscal impact on state and local governments.

- ☐ YES
- ☐ NO

SAMPLE BALLOT

- ☐ **JACK LOYD GRISHAM** *Independent* - Musician/Laborer
- ☐ **JAMES H. GREEN** *Democratic* - Firefighter Paramedic/Nurse
- ☐ **GARRETT GRUENER** *Democratic* - High-Tech Entrepreneur
- ☐ **GEROLD LEE GORMAN** *Democratic* - Engineer
- ☐ **RICH GOSSE** *Republican* - Educator
- ☐ **LEO GALLAGHER** *Independent* - Comedian
- ☐ **JOE GUZZARDI** *Democratic* - Teacher/Journalist
- ☐ **JON W. ZELLHOEFER** *Republican* - Energy Consultant/Entrepreneur
- ☐ **PAUL NAVE** *Democratic* - Businessman/Prizefighter/Father
- ☐ **ROBERT C. NEWMAN II** *Republican* - Psychologist/Farmer
- ☐ **BRIAN TRACY** *Independent* - Businessman/Consultant
- ☐ **A. LAVAR TAYLOR** *Democratic* - Tax Attorney
- ☐ **WILLIAM TSANGARES** *Republican* - Businessperson
- ☐ **PATRICIA G. TILLEY** *Independent* - Attorney
- ☐ **DIANE BEALL TEMPLIN** *American Independent* - Attorney/Realtor/Businesswoman
- ☐ **MARY "MARY CAREY" COOK** *Independent* - Adult Film Actress
- ☐ **GARY COLEMAN** *Independent* - Actor
- ☐ **TODD CARSON** *Republican* - Real Estate Developer
- ☐ **PETER MIGUEL CAMEJO** *Green* - Financial Investment Advisor
- ☐ **WILLIAM "BILL" S. CHAMBERS** *Republican* - Railroad Switchman/Brakeman
- ☐ **MICHAEL CHELI** *Independent* - Businessman
- ☐ **ROBERT CULLENBINE** *Democratic* - Retired Businessman
- ☐ **D. (LOGAN DARROW) CLEMENTS** *Republican* - Businessman
- ☐ **S. ISSA** *Republican* - Engineer
- ☐ **BOB LYNN EDWARDS** *Democratic* - Attorney
- ☐ **ERIC KOREVAAR** *Democratic* - Scientist/Businessman
- ☐ **CALVIN Y. LOUIE** *Democratic* - CPA
- ☐ **DICK LANE** *Democratic* - Educator
- ☐ **TODD RICHARD LEWIS** *Independent* - Businessman
- ☐ **GARY LEONARD** *Democratic* - Photojournalist/Author
- ☐ **DAVID LAUGHING HORSE ROBINSON** *Democratic* - Tribal Chairman
- ☐ **NED ROSCOE** *Libertarian* - Cigarette Retailer
- ☐ **DANIEL C. "DANNY" RAMIREZ** *Democratic* - Businessman/Entrepreneur/Father
- ☐ **CHRISTOPHER RANKEN** *Democratic* - Planning Commissioner
- ☐ **JEFF RAINFORTH** *Independent* - Marketing Coordinator
- ☐ **KURT E. "TACHIKAZE" RIGHTMYER** *Independent* - Middleweight Sumo Wrestler
- ☐ **DANIEL W. RICHARDS** *Republican* - Businessman
- ☐ **KEVIN RICHTER** *Republican* - Information Technology Manager
- ☐ **REVA RENEE RENZ** *Republican* - Small Business Owner
- ☐ **SHARON RUSHFORD** *Independent* - Businesswomen
- ☐ **GEORGY RUSSELL** *Democratic* - Software Engineer
- ☐ **MICHAEL J. WOZNIAK** *Democratic* - Retired Police Officer
- ☐ **DANIEL WATTS** *Green* - College Student
- ☐ **NATHAN WHITECLOUD WALTON** *Independent* - Student
- ☐ **MAURICE WALKER** *Green* - Real Estate Appraiser
- ☐ **CHUCK WALKER** *Republican* - Business Intelligence Analyst
- ☐ **LINGEL H. WINTERS** *Democratic* - Consumer Business Attorney
- ☐ **C.T. WEBER** *Peace and Freedom* - Labor Official/Analyst
- ☐ **JIM WEIR** *Democratic* - Community College Teacher
- ☐ **BRYAN QUINN** *Republican* - Businessman
- ☐ **MICHAEL JACKSON** *Republican* - Satellite Project Manager
- ☐ **JOHN "JACK" MORTENSEN** *Democratic* - Contractor/Businessman
- ☐ **ANGELYNE** *Independent* - Entertainer
- ☐ **DOUGLAS ANDERSON** *Republican* - Mortgage Broker
- ☐ **IRIS ADAM** *Natural Law* - Business Analyst
- ☐ **BROOKE ADAMS** *Independent* - Business Executive
- ☐ **ALEX-ST. JAMES** *Republican* - Public Policy Strategist
- ☐ **JIM HOFFMANN** *Republican* - Teacher
- ☐ **KEN HAMIDI** *Libertarian* - State Tax Officer
- ☐ **SARA ANN HANLON** *Independent* - Businesswoman
- ☐ **IVAN A. HALL** *Green* - Custom Denture Manufacturer
- ☐ **JOHN J. "JACK" HICKEY** *Libertarian* - Healthcare District Director
- ☐ **RALPH A. HERNANDEZ** *Democratic* - District Attorney Inspector
- ☐ **C. STEPHEN HENDERSON** *Independent* - Teacher
- ☐ **ARIANNA HUFFINGTON** *Independent* - Author/Columnist/Mother
- ☐ **ART BROWN** *Democratic* - Film Writer/Director
- ☐ **JOEL BRITTON** *Independent* - Retired Meat Packer
- ☐ **AUDIE BOCK** *Democratic* - Educator/Small Businesswoman
- ☐ **VIK S. BAJWA** *Democratic* - Businessman/Father/Entrepreneur
- ☐ **BADI BADIOZAMANI** *Independent* - Entrepreneur/Author/Executive
- ☐ **VIP BHOLA** *Republican* - Attorney/Businessowner
- ☐ **JOHN W. BEARD** *Republican* - Businessman
- ☐ **ED BEYER** *Republican* - Chief Operations Officer
- ☐ **JOHN CHRISTOPHER BURTON** *Independent* - Civil Rights Lawyer
- ☐ **CRUZ M. BUSTAMANTE** *Democratic* - Lieutenant Governor
- ☐ **CHERYL BLY-CHESTER** *Republican* - Businesswoman/Environmental Engineer
- ☐ Write-In _____

Statewide Special Election
Orange County, California
October 07, 2003

OFFICIAL BALLOT

Instruction Note:

HOW TO VOTE:
·To vote, fill in and BLACKEN completely the rectangle to the left of any candidate or to the left of the word "YES" or "NO".
·Vote for only ONE of the 135 candidates, OR enter a write-in candidate in the space provided.
·Use only the special marking device provided.
·(Absentee voters should use a dark pen or a #2 pencil.)

Shall GRAY DAVIS be recalled (removed) from the office of Governor?

- ☐ YES
- ☐ NO

Candidates to succeed GRAY DAVIS as Governor if he is recalled:

Vote for One

- ☐ **B.E. SMITH** *Independent* - Lecturer
- ☐ **DAVID RONALD SAMS** *Republican* - Businessman/Producer/Writer
- ☐ **JAMIE ROSEMARY SAFFORD** *Republican* - Business Owner
- ☐ **LAWRENCE STEVEN STRAUSS** *Democratic* - Lawyer/Businessperson/Student
- ☐ **ARNOLD SCHWARZENEGGER** *Republican* - Actor/Businessman
- ☐ **GEORGE B. SCHWARTZMAN** *Independent* - Businessman
- ☐ **MIKE SCHMIER** *Democratic* - Attorney
- ☐ **DARRIN H. SCHEIDLE** *Democratic* - Businessman/Entrepreneur
- ☐ **BILL SIMON** *Republican* - Businessman
- ☐ **RICHARD J. SIMMONS** *Independent* - Attorney/Businessperson
- ☐ **CHRISTOPHER SPROUL** *Democratic* - Environmental Attorney
- ☐ **RANDALL D. SPRAGUE** *Republican* - Discrimination Complaint Investigator
- ☐ **TIM SYLVESTER** *Democratic* - Entrepreneur
- ☐ **STEPHEN L. KNAPP** *Republican* - Engineer
- ☐ **KELLY P. KIMBALL** *Democratic* - Business Executive
- ☐ **D.E. KESSINGER** *Democratic* - Paralegal/Property Manager
- ☐ **EDWARD "ED" KENNEDY** *Democratic* - Businessman/Educator
- ☐ **TREK THUNDER KELLY** *Independent* - Business Executive/Artist
- ☐ **JERRY KUNZMAN** *Independent* - Chief Executive Officer
- ☐ **PETER V. UEBERROTH** *Republican* - Businessman/Olympics Advisor
- ☐ **BILL PRADY** *Democratic* - Television Writer/Producer
- ☐ **DARIN PRICE** *Natural Law* - University Chemistry Instructor
- ☐ **GREGORY J. PAWLIK** *Republican* - Realtor/Businessman
- ☐ **LEONARD PADILLA** *Independent* - Law School President
- ☐ **RONALD JASON PALMIERI** *Democratic* - Gay Rights Attorney
- ☐ **CHARLES "CHUCK" PINEDA JR.** *Democratic* - State Hearing Officer
- ☐ **HEATHER PETERS** *Republican* - Mediator
- ☐ **ROBERT "BUTCH" DOLE** *Republican* - Small Business Owner
- ☐ **SCOTT DAVIS** *Independent* - Business Owner
- ☐ **RONALD J. FRIEDMAN** *Independent* - Physician
- ☐ **GENE FORTE** *Republican* - Executive Recruiter/Entrepreneur
- ☐ **DIANA FOSS** *Democratic* -
- ☐ **LORRAINE (ABNER ZURD) FONTANES** *Democratic* - Film Maker
- ☐ **WARREN FARRELL** *Democratic* - Fathers' Issues Author
- ☐ **DAN FEINSTEIN** *Democratic* -
- ☐ **LARRY FLYNT** *Democratic* - Publisher
- ☐ **DARRYL L. MOBLEY** *Independent* - Businessman/Entrepreneur
- ☐ **JEFFREY L. MOCK** *Republican* - Business Owner
- ☐ **BRUCE MARGOLIN** *Democratic* - Marijuana Legalization Attorney
- ☐ **GINO MARTORANA** *Republican* - Restaurant Owner
- ☐ **PAUL MARIANO** *Democratic* - Attorney
- ☐ **ROBERT C. MANNHEIM** *Democratic* - Retired Businessperson
- ☐ **FRANK A. MACALUSO, JR.** *Democratic* - Physician/Medical Doctor
- ☐ **PAUL "CHIP" MAILANDER** *Democratic* - Golf Professional
- ☐ **DENNIS DUGGAN MCMAHON** *Republican* - Banker
- ☐ **MIKE MCNEILLY** *Republican* - Artist
- ☐ **MIKE P. MCCARTHY** *Independent* - Used Car Dealer
- ☐ **BOB MCCLAIN** *Independent* - Civil Engineer
- ☐ **TOM MCCLINTOCK** *Republican* - State Senator
- ☐ **JONATHAN MILLER** *Democratic* - Small Business Owner
- ☐ **CARL A. MEHR** *Republican* - Businessman
- ☐ **SCOTT A. MEDNICK** *Democratic* - Business Executive
- ☐ **DORENE MUSILLI** *Republican* - Parent/Educator/Businesswoman
- ☐ **VAN VO** *Republican* - Radio Producer/Businessman
- ☐ **PAUL W. VANN** *Republican* - Financial Planner
- ☐ **JAMES M. VANDEVENTER, JR.** *Republican* - Salesman/Businessman
- ☐ **BILL VAUGHN** *Democratic* - Structural Engineer
- ☐ **MARC VALDEZ** *Democratic* - Air Pollution Scientist
- ☐ **MOHAMMAD ARIF** *Independent* - Businessman

MEASURES SUBMITTED TO THE VOTERS

STATE

Proposition 53

FUNDS DEDICATED FOR STATE AND LOCAL INFRASTRUCTURE. LEGISLATIVE CONSTITUTIONAL AMENDMENT.

Generally dedicates up to 3% of General Fund revenues annually to fund state and local (excluding school and community college) infrastructure projects. Fiscal Impact: Dedication of General Fund revenues for state and local infrastructure. Potential transfers of $850 million in 2006-07, increasing to several billions of dollars in future years, under specified conditions.

- ☐ YES
- ☐ NO

Proposition 54

CLASSIFICATION BY RACE, ETHNICITY, COLOR, OR NATIONAL ORIGIN. INITIATIVE CONSTITUTIONAL AMENDMENT.

Prohibits state and local governments from classifying any person by race, ethnicity, color, or national origin. Various exemptions apply. Fiscal Impact: The measure would not result in a significant fiscal impact on state and local governments.

- ☐ YES
- ☐ NO

SAMPLE BALLOT

003

A

OFFICIAL BALLOT

Statewide Special Election

Sonoma County

October 7, 2003

This ballot stub shall be removed and retained by the voter.

MARK YOUR CHOICE(S) IN THIS MANNER ONLY:

VOTING AREA

STATE

Shall **GRAY DAVIS** be recalled (removed) from the office of Governor?	Yes	
	No	

Candidates to succeed GRAY DAVIS as Governor if he is recalled.

Vote for One

KURT E. "TACHIKAZE" RIGHTMYER, Independent Middleweight Sumo Wrestler	
DANIEL W. RICHARDS, Republican Businessman	
KEVIN RICHTER, Republican Information Technology Manager	
REVA RENEE RENZ, Republican Small Business Owner	
SHARON RUSHFORD, Independent Businesswoman	
GEORGY RUSSELL, Democratic Software Engineer	
MICHAEL J. WOZNIAK, Democratic Retired Police Officer	
DANIEL WATTS, Green College Student	
NATHAN WHITECLOUD WALTON, Independent Student	
MAURICE WALKER, Green Real Estate Appraiser	
CHUCK WALKER, Republican Business Intelligence Analyst	
LINGEL H. WINTERS, Democratic Consumer Business Attorney	
C.T. WEBER, Peace and Freedom Labor Official/Analyst	
JIM WEIR, Democratic Community College Teacher	
BRYAN QUINN, Republican Businessman	
MICHAEL JACKSON, Republican Satellite Project Manager	
JOHN "JACK" MORTENSEN, Democratic Contractor/Businessman	
DARRYL L. MOBLEY, Independent Businessman/Entrepreneur	
JEFFREY L. MOCK, Republican Business Owner	
BRUCE MARGOLIN, Democratic Marijuana Legalization Attorney	
GINO MARTORANA, Republican Restaurant Owner	
PAUL MARIANO, Democratic Attorney	

49-A007R CONTINUED OTHER SIDE

A

I HAVE VOTED—HAVE YOU?

MARK YOUR CHOICE(S) IN THIS MANNER ONLY:

VOTING AREA

(CANDIDATES CONTINUED)

ROBERT C. MANNHEIM, Democratic Retired Businessperson
FRANK A. MACALUSO, JR., Democratic Physician/Medical Doctor
PAUL "CHIP" MAILANDER, Democratic Golf Professional
DENNIS DUGGAN MCMAHON, Republican Banker
MIKE MCNEILLY, Republican Artist
MIKE P. MCCARTHY, Independent Used Car Dealer
BOB MCCLAIN, Independent Civil Engineer
TOM MCCLINTOCK, Republican State Senator
JONATHAN MILLER, Democratic Small Business Owner
CARL A. MEHR, Republican Businessman
SCOTT A. MEDNICK, Democratic Business Executive
DORENE MUSILLI, Republican Parent/Educator/Businesswoman
VAN VO, Republican Radio Producer/Businessman
PAUL W. VANN, Republican Financial Planner
JAMES M. VANDEVENTER, JR., Republican Salesman/Businessman
BILL VAUGHN, Democratic Structural Engineer
MARC VALDEZ, Democratic Air Pollution Scientist
MOHAMMAD ARIF, Independent Businessman
ANGELYNE, Independent Entertainer
DOUGLAS ANDERSON, Republican Mortgage Broker
IRIS ADAM, Natural Law Business Analyst
BROOKE ADAMS, Independent Business Executive
ALEX-ST. JAMES, Republican Public Policy Strategist
JIM HOFFMANN, Republican Teacher
KEN HAMIDI, Libertarian State Tax Officer

49-A008R CONTINUED NEXT CARD

A

Sample Ballot

Sonoma County

October 7, 2003

This ballot stub shall be removed and retained by the voter.

MARK YOUR CHOICE(S) IN THIS MANNER ONLY:

VOTING AREA

I HAVE VOTED—HAVE YOU?

MARK YOUR CHOICE(S) IN THIS MANNER ONLY:

VOTING AREA

STATE		
Shall **GRAY DAVIS** be recalled (removed) from the office of Governor?	Yes	
	No	
Candidates to succeed GRAY DAVIS as Governor if he is recalled. **Vote for One**		
KURT E. "TACHIKAZE" RIGHTMYER, Independent Middleweight Sumo Wrestler		
DANIEL W. RICHARDS, Republican Businessman		
KEVIN RICHTER, Republican Information Technology Manager		
REVA RENEE RENZ, Republican Small Business Owner		
SHARON RUSHFORD, Independent Businesswoman		
GEORGY RUSSELL, Democratic Software Engineer		
MICHAEL J. WOZNIAK, Democratic Retired Police Officer		
DANIEL WATTS, Green College Student		
NATHAN WHITECLOUD WALTON, Independent Student		
MAURICE WALKER, Green Real Estate Appraiser		
CHUCK WALKER, Republican Business Intelligence Analyst		
LINGEL H. WINTERS, Democratic Consumer Business Attorney		

(CANDIDATES CONTINUED)	
ROBERT C. MANNHEIM, Democratic Retired Businessperson	
FRANK A. MACALUSO, JR., Democratic Physician/Medical Doctor	
PAUL "CHIP" MAILANDER, Democratic Golf Professional	
DENNIS DUGGAN MCMAHON, Republican Banker	
MIKE MCNEILLY, Republican Artist	
MIKE P. MCCARTHY, Independent Used Car Dealer	
BOB MCCLAIN, Independent Civil Engineer	
TOM MCCLINTOCK, Republican State Senator	
JONATHAN MILLER, Democratic Small Business Owner	
CARL A. MEHR, Republican Businessman	
SCOTT A. MEDNICK, Democratic Business Executive	
DORENE MUSILLI, Republican Parent/Educator/Businesswoman	
VAN VO, Republican Radio Producer/Businessman	
PAUL W. VANN, Republican Financial Planner	
JAMES M. VANDEVENTER, JR., Republican Salesman/Businessman	

San

2003 recall election

There were so many names that most voting districts had to **split them up into multiple pages** -- In addition to that formatting question, what else do you notice about the official ballots on the previous pages?

In particular, what can you tell me about the order of the names?

KURT E. "TACHIKAZE" RIGHTMYER, Independent
Middleweight Sumo Wrestler

DANIEL W. RICHARDS, Republican
Businessman

KEVIN RICHTER, Republican
Information Technology Manager

REVA RENEE RENZ, Republican
Small Business Owner

SHARON RUSHFORD, Independent
Businesswoman

GEORGY RUSSELL, Democratic
Software Engineer

MICHAEL J. WOZNIAK, Democratic
Retired Police Officer

DANIEL WATTS, Green
College Student

NATHAN WHITECLOUD WALTON, Independent
Student

MAURICE WALKER, Green
Real Estate Appraiser

CHUCK WALKER, Republican
Business Intelligence Analyst

LINGEL H. WINTERS, Democratic
Consumer Business Attorney

C.T. WEBER, Peace and Freedom
Labor Official/Analyst

JIM WEIR, Democratic
Community College Teacher

BRYAN QUINN, Republican
Businessman

MICHAEL JACKSON, Republican
Satellite Project Manager

JOHN "JACK" MORTENSEN, Democratic
Contractor/Businessman

DARRYL L. MOBLEY, Independent
Businessman/Entrepreneur

JEFFREY L. MOCK, Republican
Business Owner

BRUCE MARGOLIN, Democratic
Marijuana Legalization Attorney

GINO MARTORANA, Republican
Restaurant Owner

PAUL MARIANO, Democratic
Attorney

ROBERT C. MANNHEIM, Democratic
Retired Businessperson

FRANK A. MACALUSO, JR., Democratic
Physician/Medical Doctor

PAUL "CHIP" MAILANDER, Democratic
Golf Professional

DENNIS DUGGAN MCMAHON, Republican
Banker

MIKE MCNEILLY, Republican
Artist

MIKE P. MCCARTHY, Independent
Used Car Dealer

BOB MCCLAIN, Independent
Civil Engineer

TOM MCCLINTOCK, Republican
State Senator

JONATHAN MILLER, Democratic
Small Business Owner

CARL A. MEHR, Republican
Businessman

SCOTT A. MEDNICK, Democratic
Business Executive

DORENE MUSILLI, Republican
Parent/Educator/Businesswoman

VAN VO, Republican
Radio Producer/Businessman

PAUL W. VANN, Republican
Financial Planner

California alphabet soup

Prior to 1975, the state tended to place incumbents at the top of its ballots -- In that year, however, **the state Supreme Court decided that listing candidates according to incumbency status or even in a straight alphabetical order was unconstitutional**

> ... the superior court's finding that placement in a top ballot position affords a candidate a substantial advantage over lower-placed candidates is supported by abundant expert testimony introduced at trial and is consistent with parallel findings rendered in similar litigation throughout the country. In light of this finding, we explain that **any procedure which allocates such advantageous positions to a particular class of candidates inevitably discriminates against voters supporting all other candidates**, and accordingly can only be sustained if necessary to further a compelling governmental interest. Applying this test, we conclude that the city [Santa Monica] has demonstrated no compelling interest which necessitates the provision's discriminatory classification scheme and thus we uphold the trial court's determination of invalidity. Finally, with respect to a subsidiary matter, we conclude that the allocation of advantageous ballot positions on the basis of "alphabetical order" is similarly unconstitutional.

http://scocal.stanford.edu/opinion/gould-v-grubb-27913

California alphabet soup

To address this, the state legislature developed a randomization process in which the Secretary of State creates **a new alphabetical ordering** to be used when sorting candidates

Quite literally, this means putting all **26 letters in a hat (well, something hat-like) and drawing them out one at a time** -- The order in which each letter is drawn specifies its precedence when sorting candidate names

Here is Section 13112 of the California Elections Code...

The Secretary of State shall conduct a drawing of the letters of the alphabet, the result of which shall be known as a randomized alphabet. The procedure shall be as follows:

(a) **Each letter of the alphabet shall be written on a separate slip of paper, each of which shall be folded and inserted into a capsule.** Each capsule shall be opaque and of uniform weight, color, size, shape, and texture. **The capsules shall be placed in a container, which shall be shaken vigorously in order to mix the capsules thoroughly. The container then shall be opened and the capsules removed at random one at a time.** As each is removed, it shall be opened and the letter on the slip of paper read aloud and written down. **The resulting random order of letters constitutes the randomized alphabet, which is to be used in the same manner as the conventional alphabet in determining the order of all candidates in all elections.** For example, if two candidates with the surnames Campbell and Carlson are running for the same office, their order on the ballot will depend on the order in which the letters M and R were drawn in the randomized alphabet drawing.

(b) (1) There shall be six drawings, three in each even-numbered year and three in each odd-numbered year. Each drawing shall be held at 11 a.m. on the date specified in this subdivision. The results of each drawing shall be mailed immediately to each county elections official responsible for conducting an election to which the drawing is applicable, who shall use it in determining the order on the ballot of the names of the candidates for office.

(A) The first drawing under this subdivision shall take place on the 82nd day before the April general law city elections of an even-numbered year, and shall apply to those elections and any other elections held at the same time.

(B) The second drawing under this subdivision shall take place on the 82nd day before the direct primary of an even-numbered year, and shall apply to all candidates on the ballot in that election.

(C) (i) The third drawing under this subdivision shall take place on the 82nd day before the November general election of an even-numbered year, and shall apply to all candidates on the ballot in the November general election.

(ii) In the case of the primary election and the November general election, the Secretary of State shall certify and transmit to each county elections official the order in which the names of federal and state candidates, with the exception of candidates for State Senate and Assembly, shall appear on the ballot. The elections official shall determine the order on the ballot of all other candidates using the appropriate randomized alphabet for that purpose.

(D) The fourth drawing under this subdivision shall take place on the 82nd day before the March general law city elections of each odd-numbered year, and shall apply to those elections and any other elections held at the same time.

(E) The fifth drawing under this subdivision shall take place on the 82nd day before the first Tuesday after the first Monday in June of each odd-numbered year, and shall apply to all candidates on the ballot in the elections held on that date.

(F) The sixth drawing under this subdivision shall take place on the 82nd day before the first Tuesday after the first Monday in November of the odd-numbered year, and shall apply to all candidates on the ballot in the elections held on that date.

(2) In the event there is to be an election of candidates to a special district, school district, charter city, or other local government body at the same time as one of the five major election dates specified in subparagraphs (A) to (F), inclusive, and the last possible day to file nomination papers for the local election would occur after the date of the drawing for the major election date, theprocedure set forth in Section 13113 shall apply.

(c) Each randomized alphabet drawing shall be open to the public. At least 10 days prior to a drawing, the Secretary of State shall notify the news media and other interested parties of the date, time, and place of the drawing. The president of each statewide association of local officials with responsibilities for conducting elections shall be invited by the Secretary of State to attend each drawing or send a representative. The state chairman of each qualified political party shall be invited to attend or send a representative in the case of drawings held to determine the order of candidates on the primary election ballot, the November generalelection ballot, or a special election ballot as provided for in subdivision (d).

California alphabet soup

As an example, suppose the lottery produced this sequence of letters

O G F W N T Q B M U Y R S D K H V E X J Z I P A C L

Then, using this sequence, we can produce the following sorted list of names

GORMAN	GOSSE	GUZZARDI	GRUENER
GREEN	GRISHAM	GALLAGHER	FONTANES
FORTE	FOSS	FRIEDMAN	FEINSTEIN
FARRELL	FLYNT	WOZNIAK	WEBER
WEIR	WINTERS	WATTS	WALTON
WALKERM	WALKERC	NEWMANII	NAVE
TRACY	TSANGARES	TEMPLIN	TILLEY
TAYLOR	QUINN	BOCK	BURTON
BUSTAMANTE	BROWN	BRITTON	BHOLA
BEYER	BEARD	BADIOZAMANI	BAJWA
BLYCHESTER	MOBLEY	MORTENSEN	MOCK
MUSILLI	MEDNICK	MEHR	MILLER
MANNHEIM	MARGOLIN	MARTORANA	MARIANO
MAILANDER	MACALUSO	MCNEILLY	MCMAHON
MCCARTHY	MCCLINTOCK	MCCLAIN	UEBERROTH

California alphabet soup

Or, suppose the lottery produced this sequence of letters

F R I A Z U M Y J N X L V O B D S T E Q K W H C P G

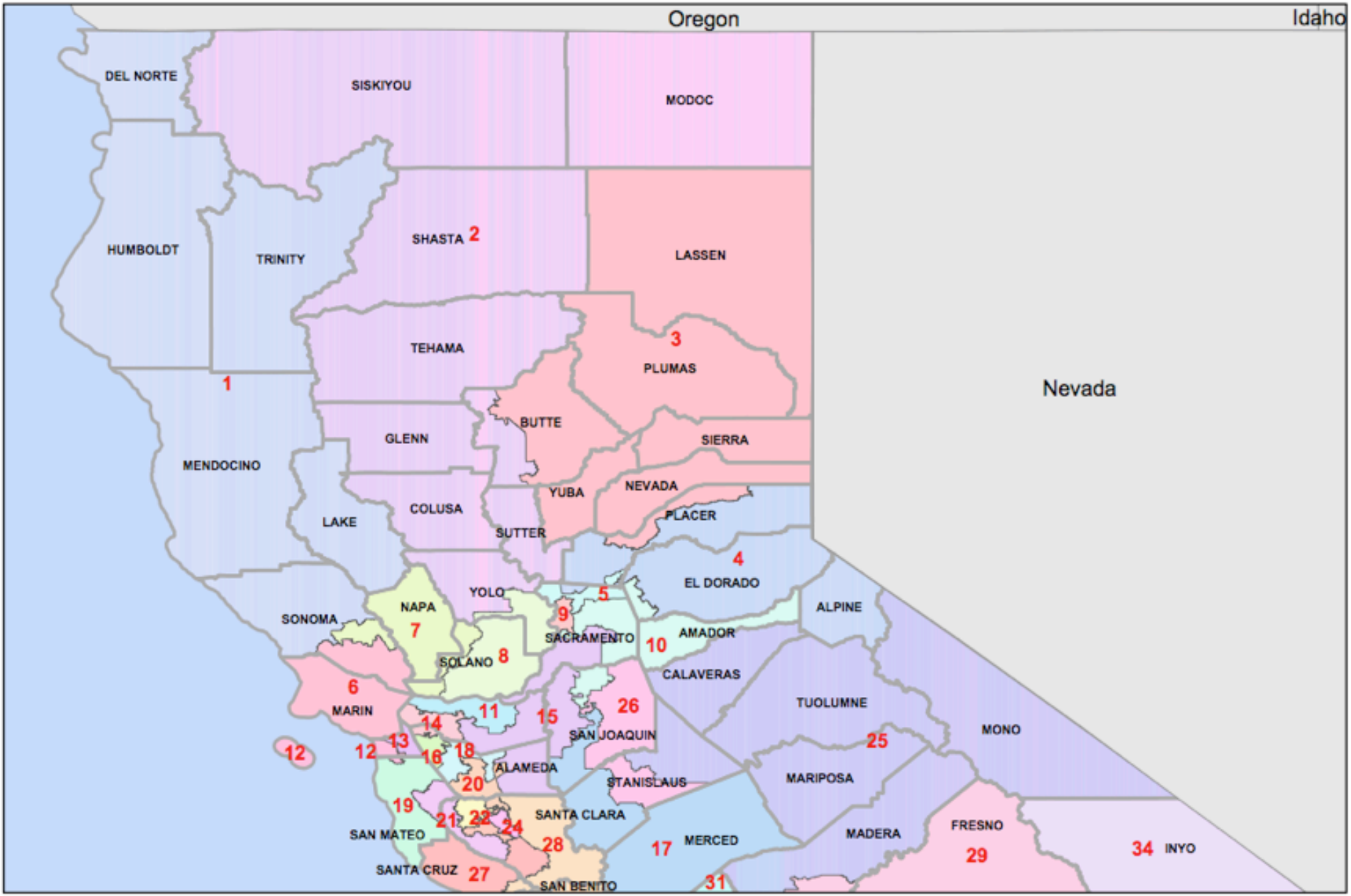
Then, using this sequence, we can produce the following sorted list of names

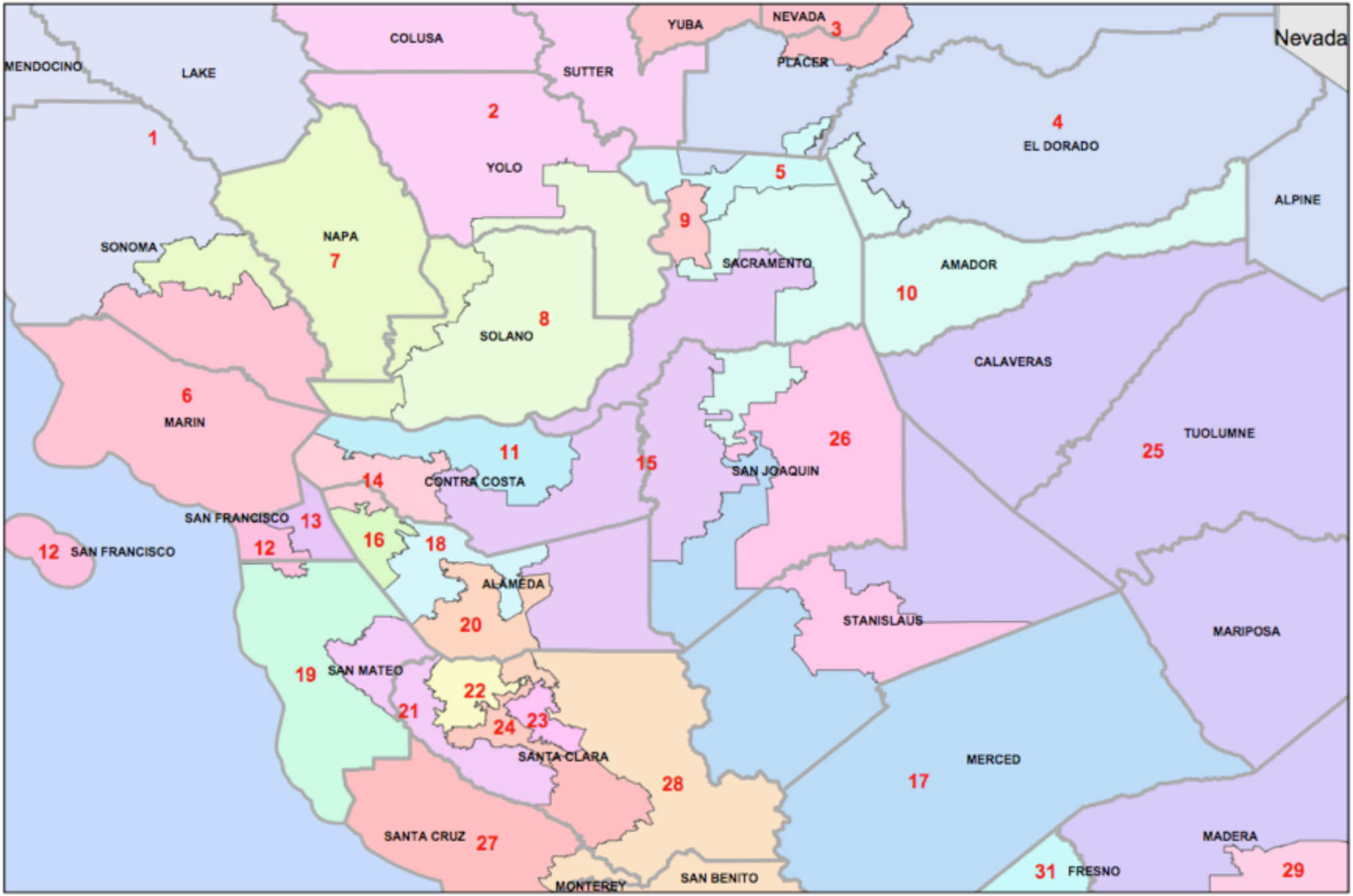
FRIEDMAN	FARRELL	FLYNT	FORTE
FONTANES	FOSS	FEINSTEIN	RICHARDS
RICHTER	RIGHTMYER	RAINFORTH	RAMIREZ
RANKEN	RUSSELL	RUSHFORD	ROBINSON
ROSCOE	RENZ	ISSA	ARIF
ANDERSON	ANGELYNE	ALEXSTJAMES	ADAM
ADAMS	ZELLHOEFER	UEBERROTH	MILLER
MARIANO	MARTORANA	MARGOLIN	MAILANDER
MANNHEIM	MACALUSO	MUSILLI	MORTENSEN
MOBLEY	MOCK	MEDNICK	MEHR
MCMAHON	MCNEILLY	MCCARTHY	MCCLINTOCK
MCCLAIN	JACKSON	NAVE	NEWMANII
LANE	LOUIE	LEONARD	LEWIS
VAUGHN	VANN	VANDEVENTER	VALDEZ
VO	BRITTON	BROWN	BAJWA

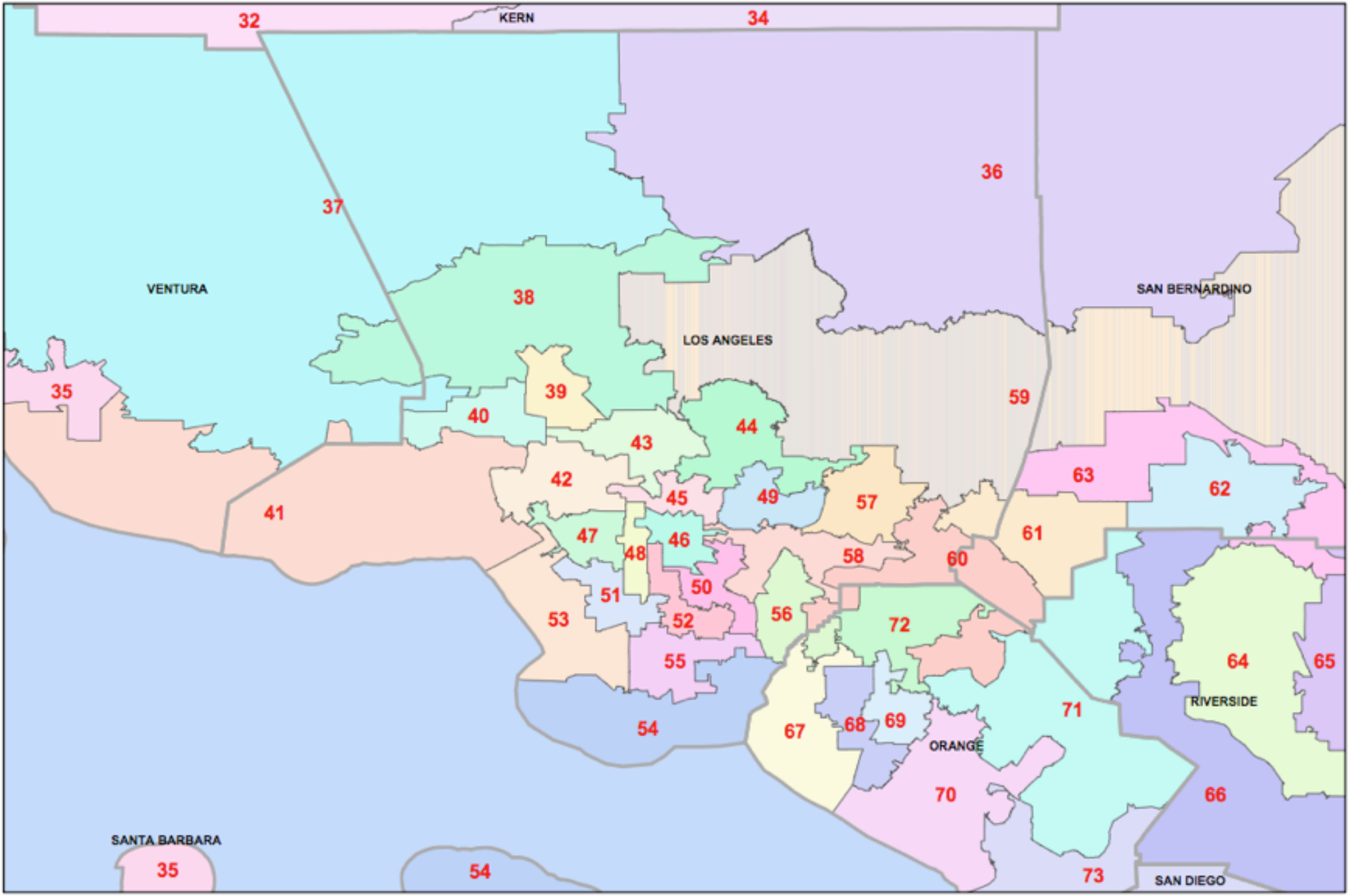
Randomize and rotate

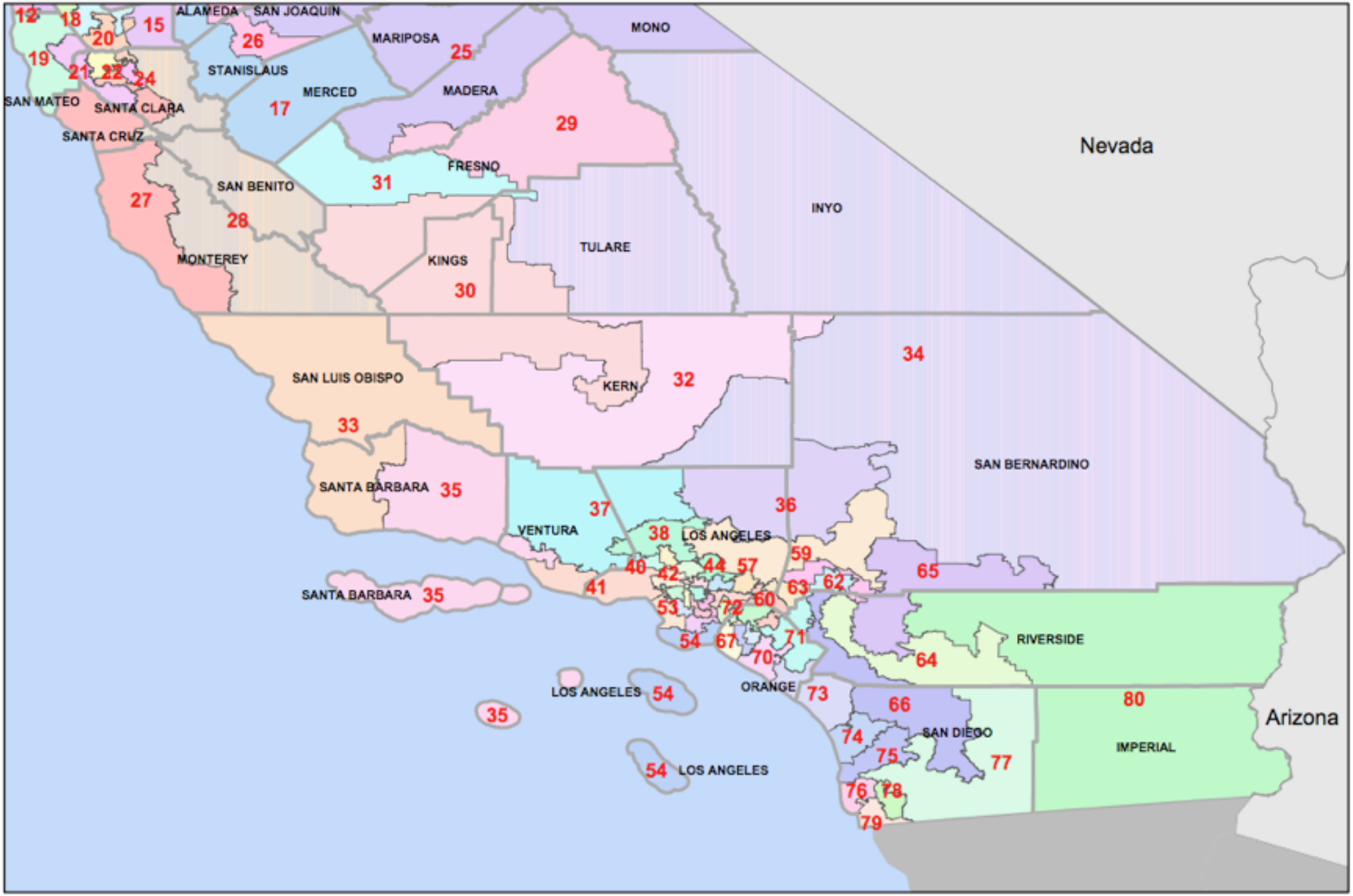
Once an alphabet is set, the candidates ordered but only the first assembly district uses this as its ballot order -- For each subsequent assembly district (we have 80 districts in the 2003 election), **the name at the top of the list is moved to the bottom and all the other names are moved up one place**

The idea behind this rotation is that candidates each spent time near the top of a ballot in at least some of the districts -- This all seems sensible when **the number of candidates is small**









Randomize and rotate

Assembly districts intersect counties -- Some districts contain multiple counties (District 1 contains Del Norte, Humboldt, Lake, Mendocino, Sonoma and Trinity counties) while some counties are split among many districts (Alameda county is in Districts 14, 15, 16, 18, and 20)

While the order of candidates' names is set per assembly district, **the ballot format (the number of names per page) is set by each county** -- In all there are 158 district-county combinations

2003 recall election

The alphabet used in 2003 was

R W Q O J M V A H B S G Z X N T C I E K U P D Y F L

The order and the breakdown by page number for the first district (in Del Norte County) is given on the next slide and then the breakdown by page number for the last district (in Riverside County) is given after that

District 1,
Del Norte county

Page 1

ROBINSON	ROSCOE	RAMIREZ	RANKEN
RAINFORTH	RIGHTMYER	RICHARDS	RICHTER
RENZ	RUSHFORD	RUSSELL	WOZNIAK
WATTS	WALTON	WALKERM	WALKERC
WINTERS	WEBER	WEIR	QUINN
JACKSON	MORTENSEN	MOBLEY	

Page 2

MOCK	MARGOLIN	MARTORANA	MARIANO
MANNHEIM	MACALUSO	MAILANDER	MCMAHON
MCNEILLY	MCCARTHY	MCCLAIN	MCCLINTOCK
MILLER	MEHR	MEDNICK	MUSILLI
VO	VANN	VANDEVENTER	VAUGHN
VALDEZ	ARIF	ANGELYNE	ANDERSON

Page 3

ADAM	ADAMS	ALEXSTJAMES	HOFFMANN
HAMIDI	HANLON	HALL	HICKEY
HERNANDEZ	HENDERSON	HUFFINGTON	BROWN
BRITTON	BOCK	BAJWA	BADIOZAMANI
BHOLA	BEARD	BEYER	BURTON
BUSTAMANTE	BLYCHESTER	SMITH	SAMS

Page 4

SAFFORD	STRAUSS	SCHWARZENEGGER	SCHWARTZMAN
SCHMIER	SCHEIDLE	SIMON	SIMMONS
SPROUL	SPRAGUE	SYLVESTER	GRISHAM
GREEN	GRUENER	GORMAN	GOSSE
GALLAGHER	GUZZARDI	ZELLHOEFER	NAVE
NEWMANII	TRACY	TAYLOR	TSANGARES

Page 5

TILLEY	TEMPLIN	COOK	COLEMAN
CARSON	CAMEJO	CHAMBERS	CHELI
CULLENBINE	CLEMENTS	ISSA	EDWARDS
KOREVAAR	KNAPP	KIMBALL	KESSINGER
KENNEDY	KELLY	KUNZMAN	UEBERROTH
PRADY	PRICE	PAWLIK	PADILLA

Page 6

PALMIERI	PINEDA	PETERS	DOLE
DAVIS	FRIEDMAN	FORTE	FOSS
FONTANES	FARRELL	FEINSTEIN	FLYNT
LOUIE	LANE	LEWIS	LEONARD

District 80,
Riverside County

Page 1

SPROUL	SPRAGUE	SYLVESTER	GRISHAM
GREEN	GRUENER	GORMAN	GOSSE
GALLAGHER	GUZZARDI	ZELLHOEFER	NAVE
NEWMANII	TRACY	TAYLOR	TSANGARES
TILLEY	TEMPLIN	COOK	COLEMAN

Page 2

CARSON	CAMEJO	CHAMBERS	CHELI
CULLENBINE	CLEMENTS	ISSA	EDWARDS
KOREVAAR	KNAPP	KIMBALL	KESSINGER
KENNEDY	KELLY	KUNZMAN	UEBERROTH
PRADY	PRICE	PAWLIK	PADILLA

Page 3

PALMIERI	PINEDA	PETERS	DOLE
DAVIS	FRIEDMAN	FORTE	FOSS
FONTANES	FARRELL	FEINSTEIN	FLYNT
LOUIE	LANE	LEWIS	LEONARD
ROBINSON	ROSCOE	RAMIREZ	RANKEN

Page 4

RAINFORTH	RIGHTMYER	RICHARDS	RICHTER
RENZ	RUSHFORD	RUSSELL	WOZNIAK
WATTS	WALTON	WALKERM	WALKERC
WINTERS	WEBER	WEIR	QUINN
JACKSON	MORTENSEN	MOBLEY	MOCK

Page 5

MARGOLIN	MARTORANA	MARIANO	MANNHEIM
MACALUSO	MAILANDER	MCMAHON	MCNEILLY
MCCARTHY	MCCLAIN	MCCLINTOCK	MILLER
MEHR	MEDNICK	MUSILLI	VO
VANN	VANDEVENTER	VAUGHN	VALDEZ

Page 6

ARIF	ANGELYNE	ANDERSON	ADAM
ADAMS	ALEXSTJAMES	HOFFMANN	HAMIDI
HANLON	HALL	HICKEY	HERNANDEZ
HENDERSON	HUFFINGTON	BROWN	BRITTON
BOCK	BAJWA	BADIOZAMANI	BHOLA

Page 7

BEARD	BEYER	BURTON	BUSTAMANTE
BLYCHESTER	SMITH	SAMS	SAFFORD
STRAUSS	SCHWARZENEGGER	SCHWARTZMAN	SCHMIER
SCHEIDLE	SIMON	SIMMONS	

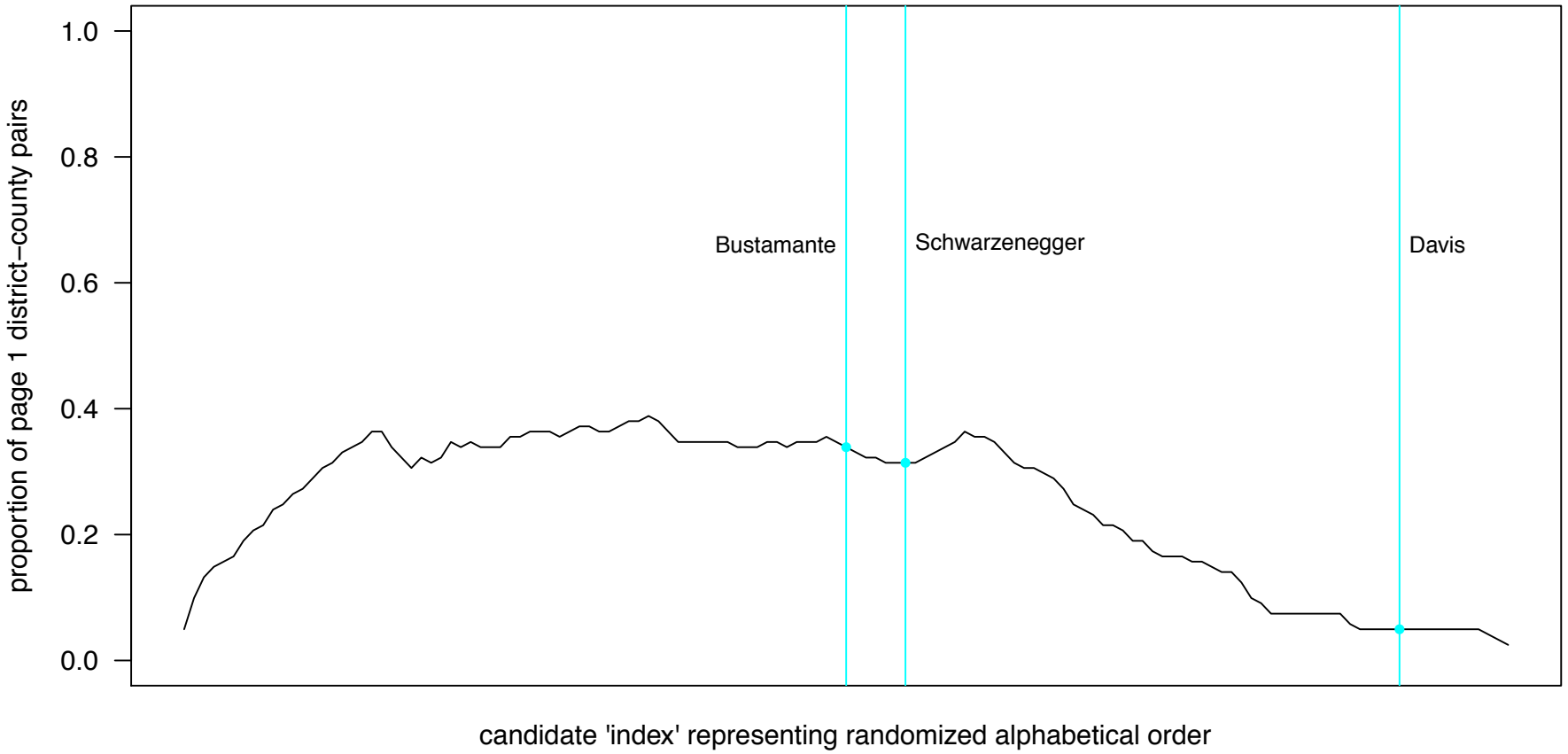
Location, location, location

With so many candidates on the ballot, the division into **multiple pages** was inevitable -- In all, **121 of the 158 district-county pairs produced ballots with more than one page**

Recalling the Supreme Court's decision that ballot placement has an effect on voting behavior, the rotation process was meant to **even out the number of times people are banished to the back pages**

We can have a look at how often each of the 135 candidates were on the first page of a ballot in the 121 district-county combinations...

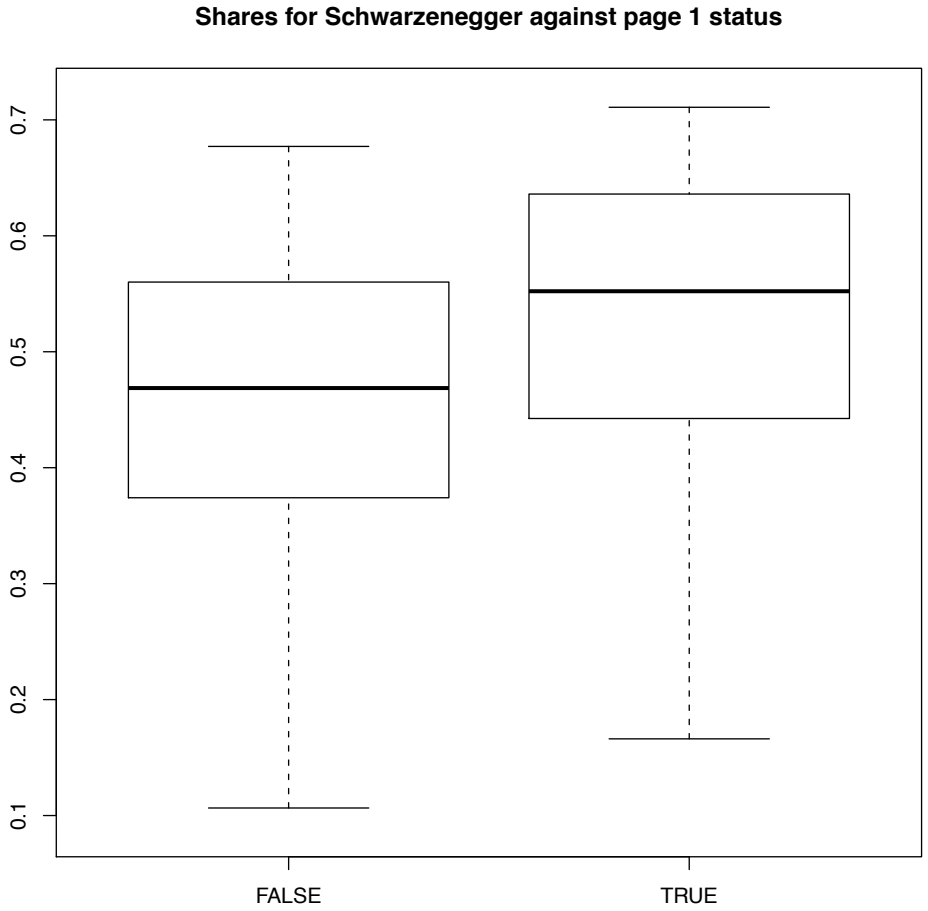
```
# read in a csv (comma separated values) file provided to us by prof daniel ho
# at stanford university

> ballot <- read.csv(url("http://www.stat.ucla.edu/~cocteau/stat13/data/ballot.csv"))

> table(ballot$Schwarzenegger)

 1  2  3  4  6  7
38 29 30 16  2  6

> table(ballot$Bustamante)

 1  2  3  4  6  7
41 31 35  3  2  9

> table(ballot$Davis)

 1  2  3  4  5  6  7 11
 6 24 11 22 33 21  3  1
```

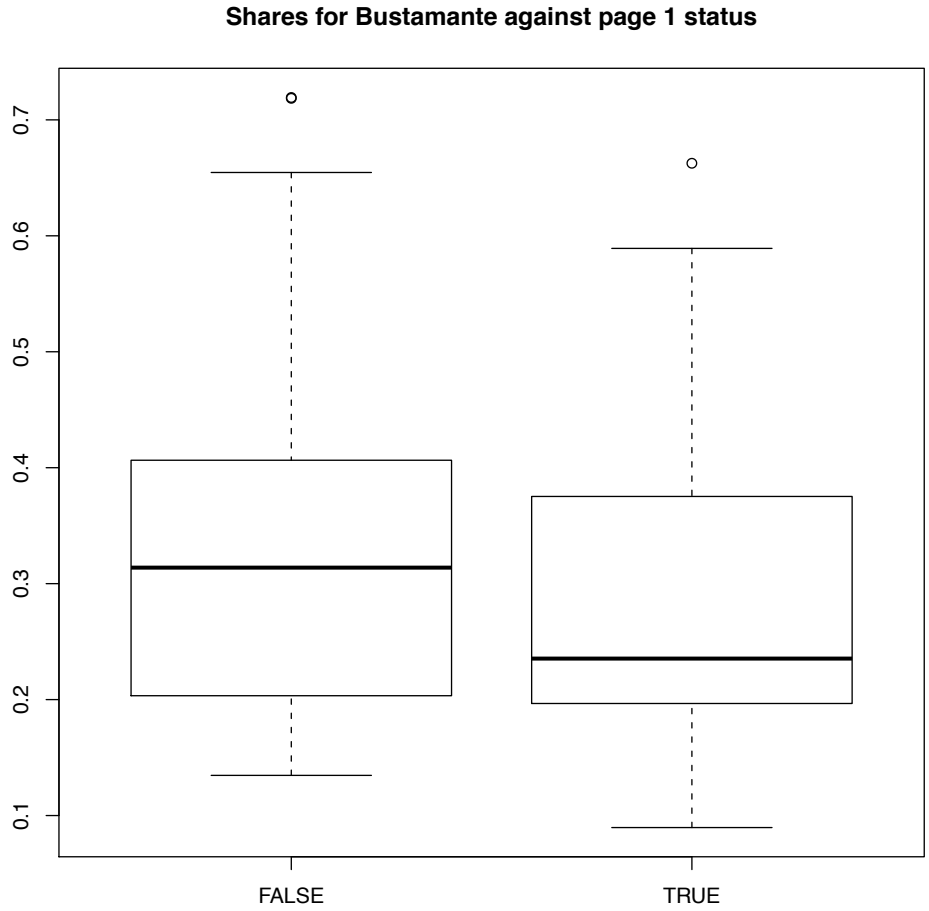

1.0
0.8
0.6
0.4
0.2
0.0
proportion of page 1 district–county pairs
Bustamante
Schwarzenegger
Davis
candidate 'index' representing randomized alphabetical order

Location, location, location

Given the number of candidates in 2003, each was on the front page in less than half of the county-district pairs -- We can reasonably ask whether or not **being on the first page of a ballot has an effect** on the share of the vote a candidate receives

How might we measure this?






```
> shares <- read.csv(url("http://www.stat.ucla.edu/~cocteau/stat13/data/shares.csv"))

> head(shares$Schwarzenegger)
[1] 0.5497971 0.4709309 0.3719113 0.3448414 0.6331666 0.6420732

> head(ballot$Schwarzenegger)
[1] 4 4 4 4 4 1

> boxplot(shares$Schwarzenegger ~ (ballot$Schwarzenegger == 1))
```

Vote shares

One natural measure is just the difference between the average share (proportion) of the vote a candidate received in districts where they were on the first page of the ballot and the average computed over districts where they were on a later page

In symbols, if we let y_i denote one candidate's vote share in district i, then we can capture the effect of being on the first page with

T = (mean y_i for districts i where candidate is listed on page 1) -
(mean y_i for districts i where candidate is on later page)

Vote shares: Examples

As an example, **Arnold Schwarzenegger** received an average of 52% of the vote in districts where he was listed on the first page of the ballot, but only 46% when he was on page 2 or higher -- That is, **T = 0.06**

Cruz Bustamante received an average of 29% of the vote in districts where he was listed on the first page of the ballot, and an average of 33% of the vote in the remaining districts, so that **T = -0.04**

Ariana Huffington (who eventually dropped out of the race although her name remained on the ballot) received an average of 0.54% of the vote in districts where she was on page one of the ballot, but an average of 0.52% in other districts, implying **T = 0.0002**

Finally, **Bruce Margolin** was a relatively small player earning an average of 0.2% of the vote when on the first page, but 0.1% when listed on later pages so that **T = 0.001**

Starting to look familiar

What do we make of these differences? Does our analysis end with observing the differences or can we say more? Given that we've spent a week talking about "statistically significant" differences, is there some way for us apply that framework here?

Hint: I wouldn't have burned 30 slides on this story if we couldn't do more! So, what do you think?

Re-randomization

Given that the State of California creates **a randomized alphabet** to start the ballot placement process, we have a kind of **“natural experiment”** that we can exploit to assess whether the differences T we see for a given candidate could be **purely the result of randomization**

Let’s start with the **null hypothesis** that **page placement has no effect on the vote share earned by a candidate** -- Under this assumption, we can **generate new alphabets, create new ballot layouts and compute a new difference in mean vote shares**

Repeating this many times for a given candidate, we create a null distribution for the difference in mean vote shares T, to which we can **compare the difference that was actually recorded on election day**...

One iteration

In the original alphabet

R W Q O J M V A H B S G Z X N T C I E K U P D Y F L

Schwarzenegger appeared in the 74th position in District 1 -- After rotation, here are the page numbers he appeared on in the 121 districts with multi-page ballots

```
  [1] 4 4 4 4 4 1 4 4 4 4 4 3 1 6 4 3 4 4 4 3 1 4 3 4 3 3 1 3 3 1 1 1 1 3 1 3 1
 [38] 3 3 1 1 3 1 3 3 1 3 3 3 3 3 3 1 1 1 1 1 3 3 3 3 2 3 2 2 1 1 2 1 2 3 1 3 2
 [75] 3 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1
[112] 1 1 1 7 7 7 7 7 6 7
```

His average vote share in those districts where he was on page 1 was 52% while it was 46% in the other districts so that **T = 0.06**

One iteration

Now, drawing a new alphabet (literally calling `sample` in R), we come up with this

U K N R V H P S G Q J B O D E T Y W L X I F M C A Z

With this order, Schwarzenegger appears in the 51st position in District 1 -- After rotation, here are the page numbers he appeared on in the 121 districts with multi-page ballots

```
  [1] 3 3 3 3 3 1 3 3 3 3 3 3 1 5 3 3 3 3 3 2 1 3 2 3 2 3 1 2 2 1 1 1 1 2 1 2 1
 [38] 2 2 1 1 2 1 2 2 1 2 2 2 2 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2 1 1 1 1 1 2 1 2 1
 [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 7 7 7 7 7 7 7 7 2 7 2 7 2 2 7 2 7 7 2 7 7
[112] 6 6 6 6 6 6 6 6 5 6
```

In this case, his average vote share in those districts where he was on page 1 is 44% while it is 50% in the other districts so that **T = -0.06**

One (more) iteration

Drawing yet another new alphabet, we come up with the following

W P B D T X N K L Q E A R J Y G M C O S Z V U I F H

Here, Schwarzenegger appears in the 108th position in District 1 -- After rotation, here are the page numbers he appeared on in the 121 districts with multi-page ballots

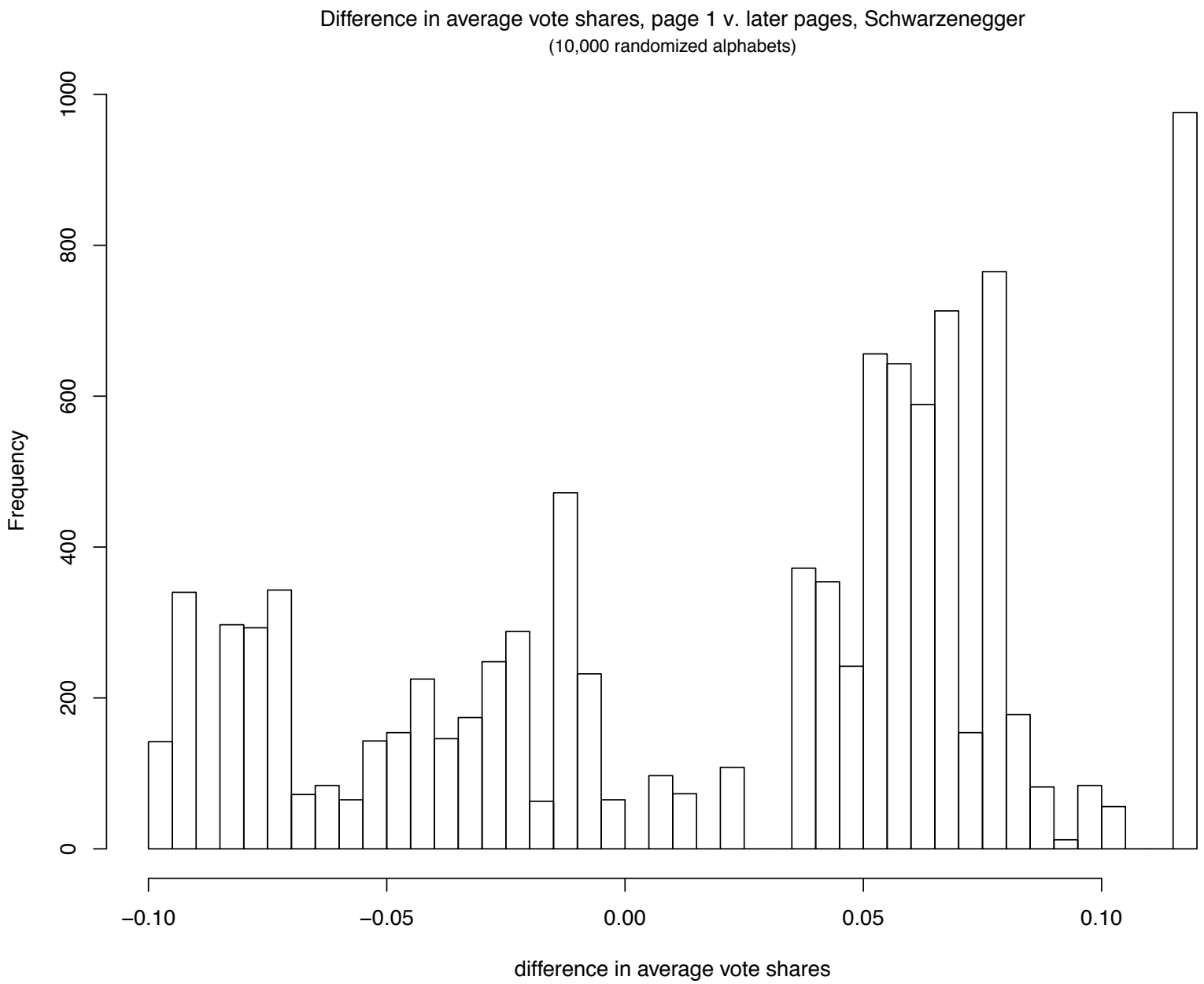
```
  [1] 5 5 5 5 5 2 5 6 5 5 5 5 2 9 5 5 5 5 5 5 2 6 5 6 5 5 2 5 5 2 2 2 2 4 2 4 1
 [38] 5 4 1 1 4 2 4 5 2 5 5 5 5 4 4 1 1 1 1 1 4 5 4 4 4 5 4 4 2 2 4 2 4 4 1 4 3
 [75] 4 3 4 4 4 3 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 3 3 3 3 1 3 1 3 1 1 3 1 3 3 1 3 3
[112] 2 3 2 2 2 2 2 2 2 2
```

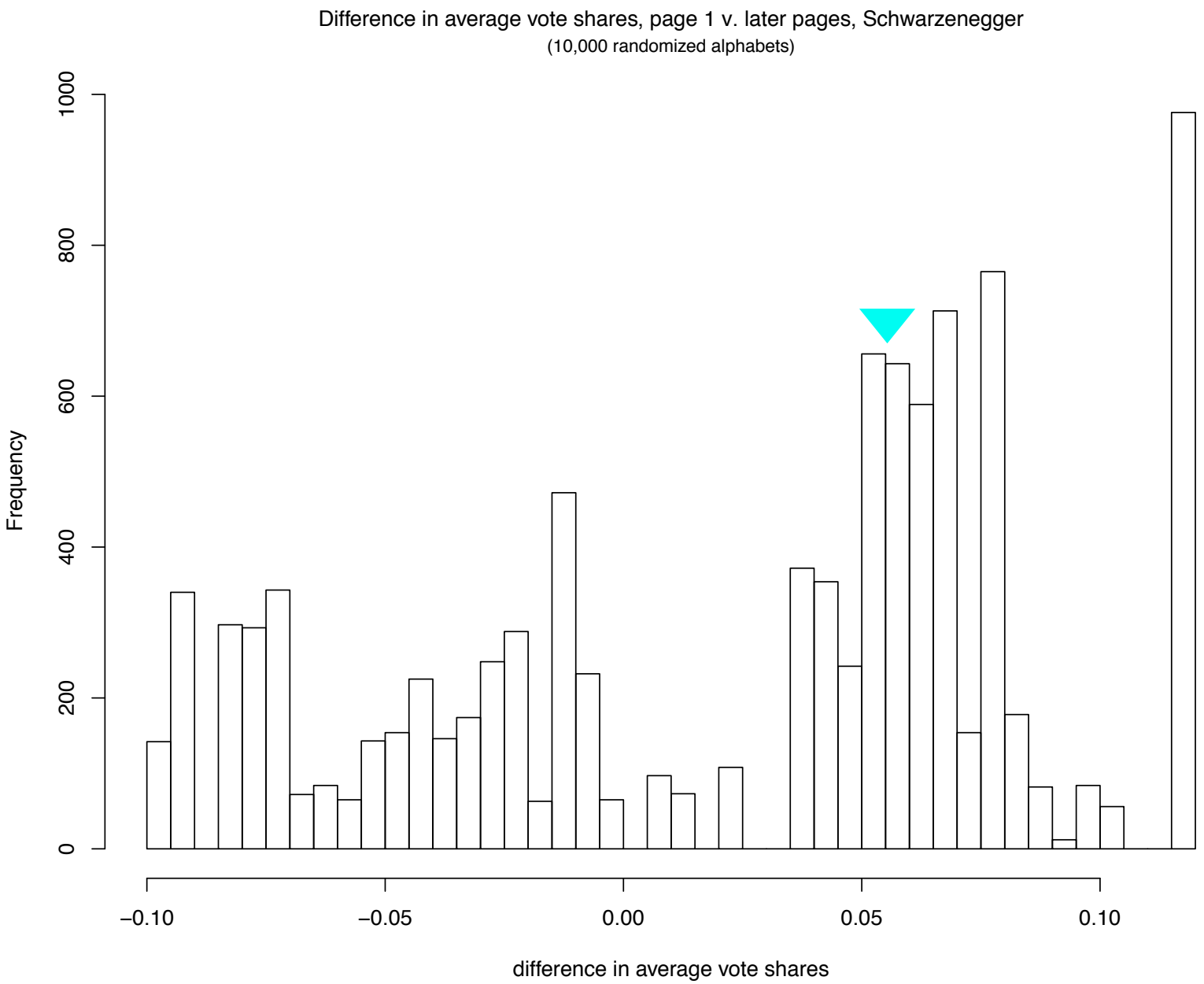
In this case, his average vote share in those districts where he was on page 1 is 55% while it is 47% in the other districts so that **T = 0.08**

Repeating

If we carry this out 10,000 times, we come up with **the histogram of T values** on the next page -- These represent the variability in the difference in mean vote shares solely due to the randomization process

Recalling that **the difference for Schwarzenegger in the 2003 election was 0.06**, what do you make of this histogram? What can you say about the impact of ballot placement for Schwarzenegger?


Difference in average vote shares, page 1 v. later pages, Schwarzenegger
(10,000 randomized alphabets)
Frequency
0
200
400
600
800
1000
−0.10
−0.05
0.00
0.05
0.10
difference in average vote shares


Difference in average vote shares, page 1 v. later pages, Schwarzenegger
(10,000 randomized alphabets)
Frequency
0
200
400
600
800
1000
−0.10
−0.05
0.00
0.05
0.10
difference in average vote shares

The alternative

Ho and Imai used **a one-sided test** associated with **the alternative hypothesis** that **the difference in average vote shares is bigger than zero** -- Interestingly, they also reference work that suggests candidates might exhibit "recency effects" of gaining votes when listed later in the ballot

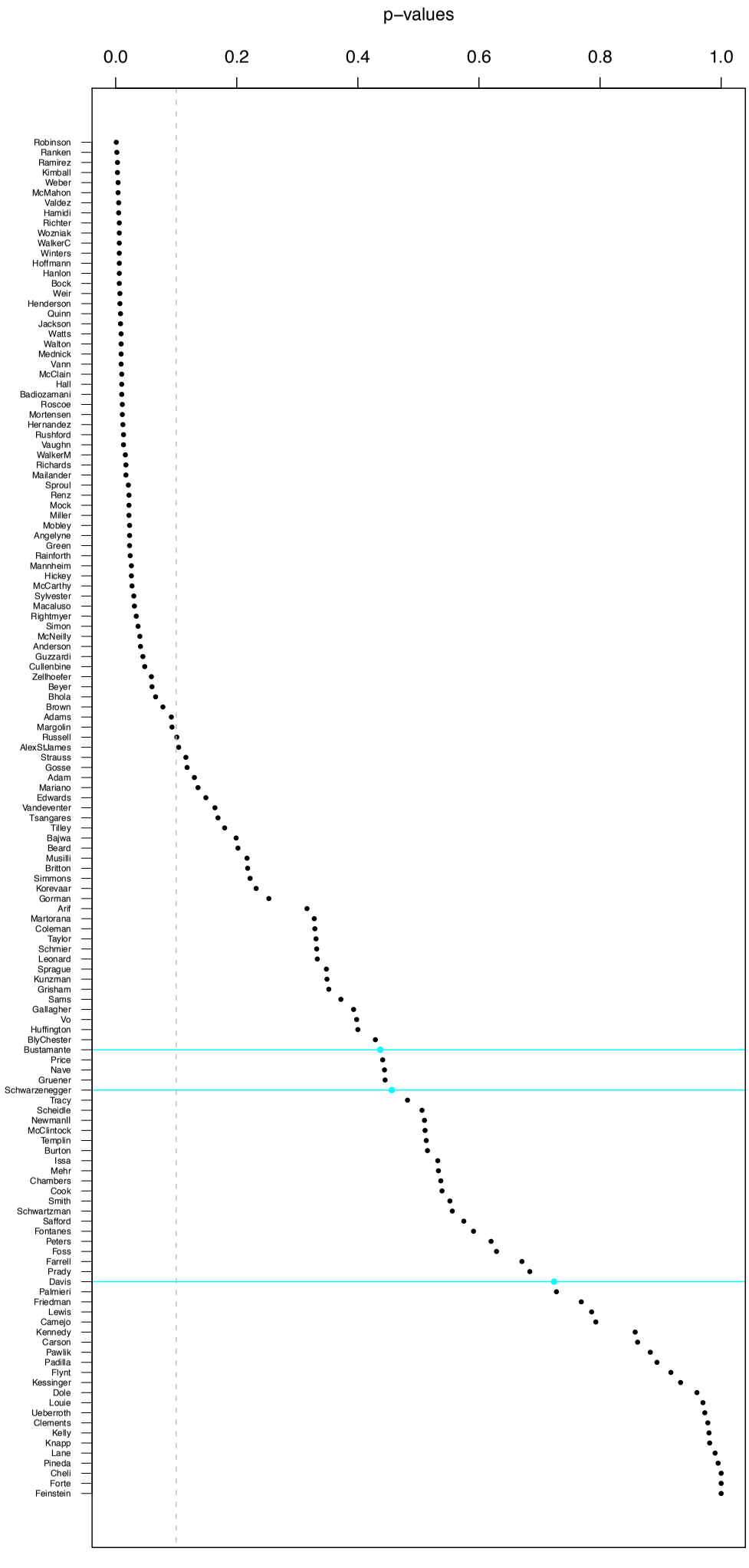
Still, **we'll repeat their analysis** and stick with one sided tests, meaning our alternative is that a candidate sees a boost from being listed on the first page of a ballot

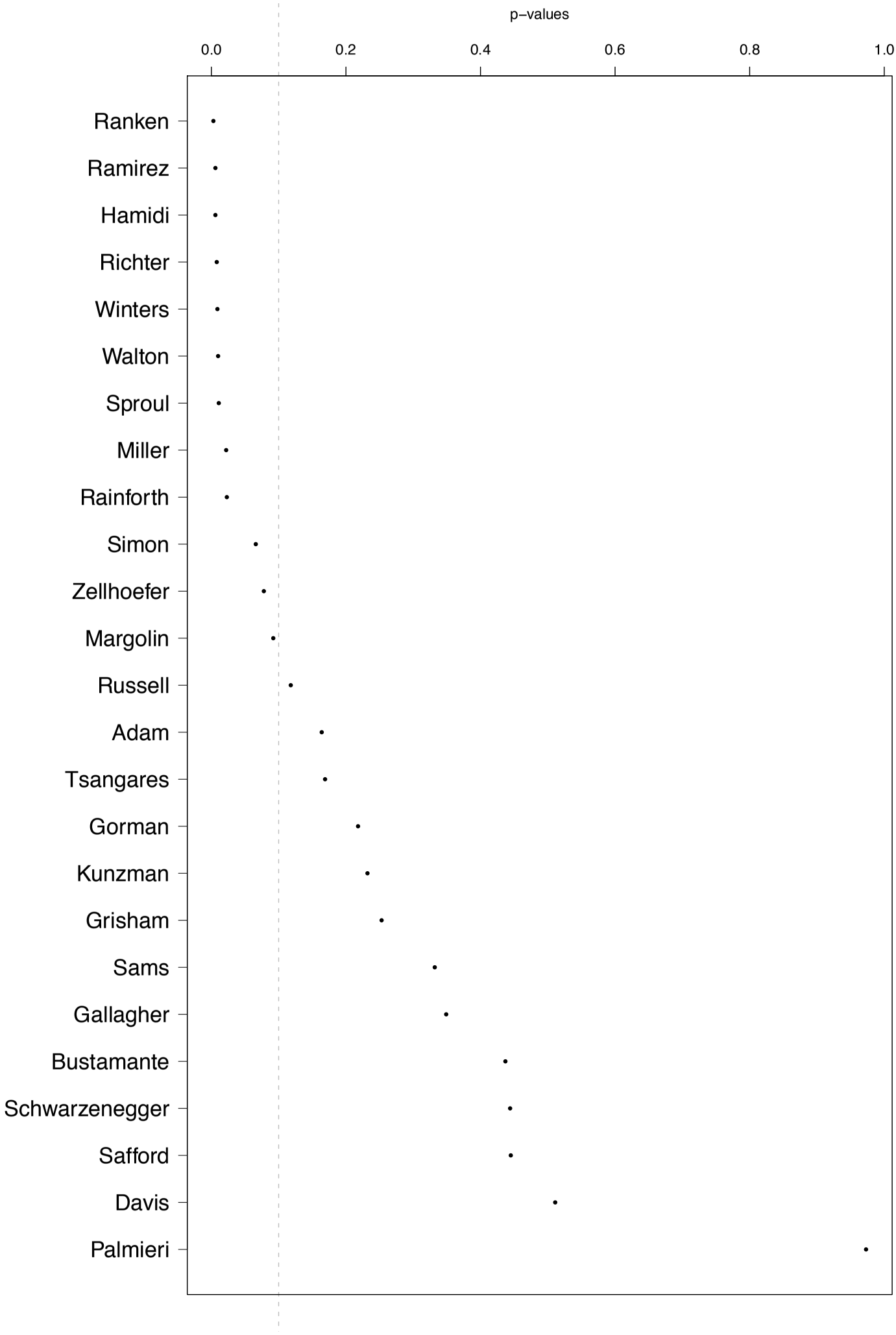
In the Schwarzenegger case, 45% of our samples are larger than the observed value of 0.06, meaning **he did not see a significant boost relative to the uncertainty present due to the randomized alphabets**

And the rest of the pack

We can **repeat this analysis** for the remaining 134 candidates from the 2003 ballot and see how they perform -- On the next two pages, we present first the complete list and then a smaller subset (that can actually be read when projected!)

What do you see in these images?


p−values
0.0
0.2
0.4
0.6
0.8
1.0
Robinson
Ranken
Ramirez
Kimball
Weber
McMahon
Valdez
Hamidi
Richter
Wozniak
WalkerC
Winters
Hoffmann
Hanlon
Bock
Weir
Henderson
Quinn
Jackson
Watts
Walton
Mednick
Vann
McClain
Hall
Badiozamani
Roscoe
Mortensen
Hernandez
Rushford
Vaughn
WalkerM
Richards
Mailander
Sproul
Renz
Mock
Miller
Mobley
Angelyne
Green
Rainforth
Mannheim
Hickey
McCarthy
Sylvester
Macaluso
Rightmyer
Simon
McNeilly
Anderson
Guzzardi
Cullenbine
Zellhoefer
Beyer
Bhola
Brown
Adams
Margolin
Russell
AlexStJames
Strauss
Gosse
Adam
Mariano
Edwards
Vandeventer
Tsangares
Tilley
Bajwa
Beard
Musilli
Britton
Simmons
Korevaar
Gorman
Arif
Martorana
Coleman
Taylor
Schmier
Leonard
Sprague
Kunzman
Grisham
Sams
Gallagher
Vo
Huffington
BlyChester
Bustamante
Price
Nave
Gruener
Schwarzenegger
Tracy
Scheidle
NewmanII
McClintock
Templin
Burton
Issa
Mehr
Chambers
Cook
Smith
Schwartzman
Safford
Fontanes
Peters
Foss
Farrell
Prady
Davis
Palmieri
Friedman
Lewis
Camejo
Kennedy
Carson
Pawlik
Padilla
Flynt
Kessinger
Dole
Loule
Ueberroth
Clements
Kelly
Knapp
Lane
Pineda
Cheli
Forte
Feinstein


p−values
0.0
0.2
0.4
0.6
0.8
1.0
Ranken
Ramirez
Hamidi
Richter
Winters
Walton
Sproul
Miller
Rainforth
Simon
Zellhoefer
Margolin
Russell
Adam
Tsangares
Gorman
Kunzman
Grisham
Sams
Gallagher
Bustamante
Schwarzenegger
Safford
Davis
Palmieri

Interpretation

It seems that **for the major candidates, the page-1 effect is undetectable** relative to the uncertainty introduced in the randomization, but that **among minor candidates, we see a fair number of significant effects**

Recall that our testing procedures have a built-in error rate that we can control -- If we were to reject a null hypothesis of no page effect at the 0.05 level, for example, **we would expect to be mistaken 5% of the time**

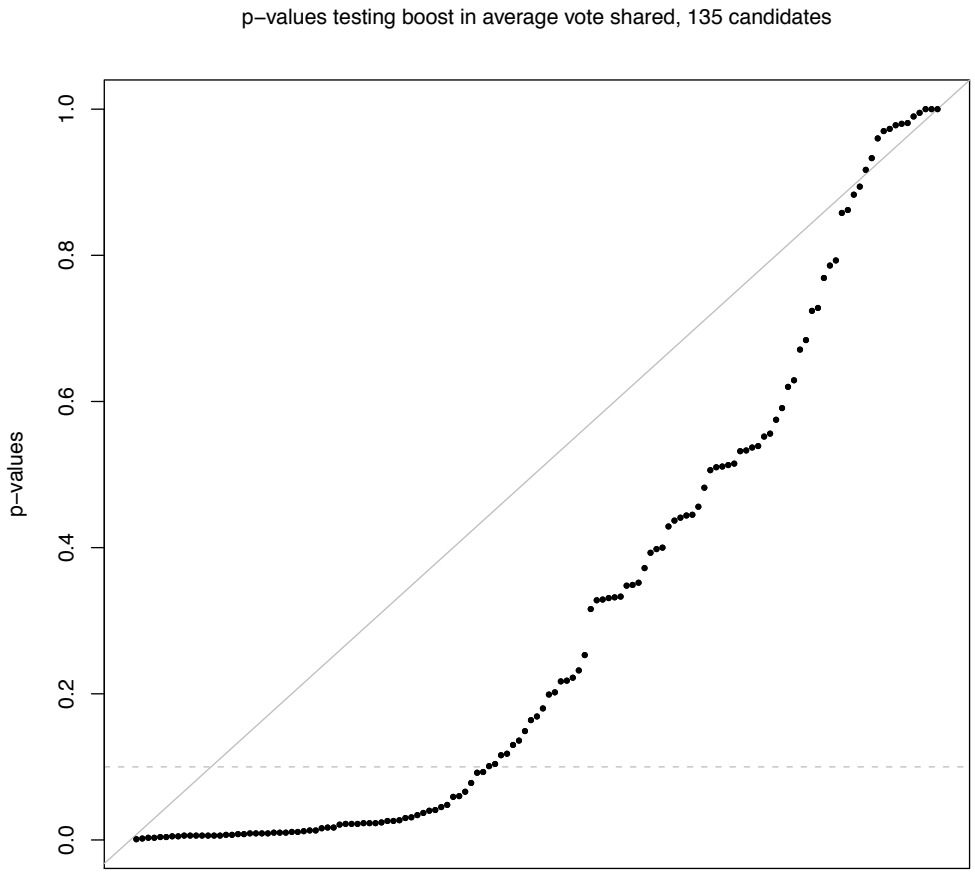
in the figure we see 53 out of 135 or **40% of the candidates with P-values smaller than 0.05** -- What this means is that we are seeing more significant effects than we would expect if the null hypothesis was true for all 135 candidates

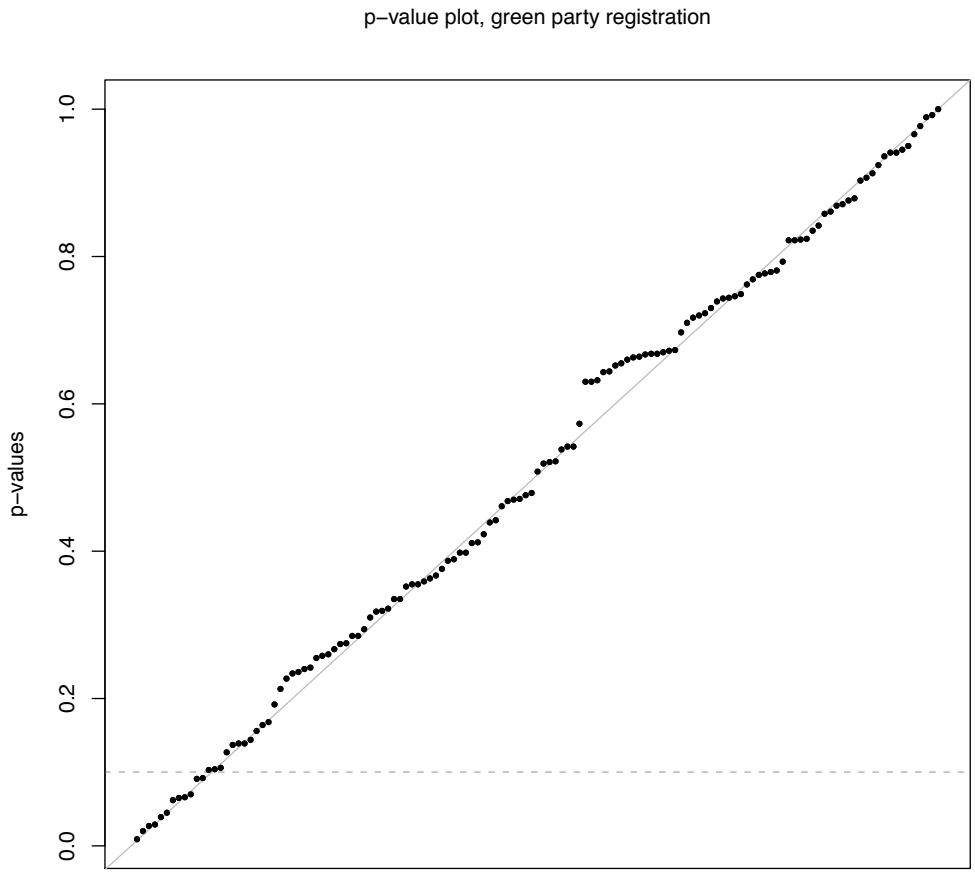
A little advanced

To check that we haven't made a mistake (either in programming or in our thinking), we can try the same experiment but on so-called **"pretreatment" variables** -- These are variables that, unlike vote shares, should **have absolutely no relation to ballot position**

For example, suppose that for each candidate's position relative to our random alphabets, we **compute the difference between average Democratic registration totals for the district-county pairs in which the candidate is listed on page 1 and those for which the candidate is on on a later page**-- This variable had no relationship to the ballot order in the 2003 election and so the null hypothesis is undoubtedly true

Let's look at the sorted P-values for those tests...


p−values testing boost in average vote shared, 135 candidates
p−values
0.0
0.2
0.4
0.6
0.8
1.0


p−value plot, green party registration
p−values
0.0
0.2
0.4
0.6
0.8
1.0

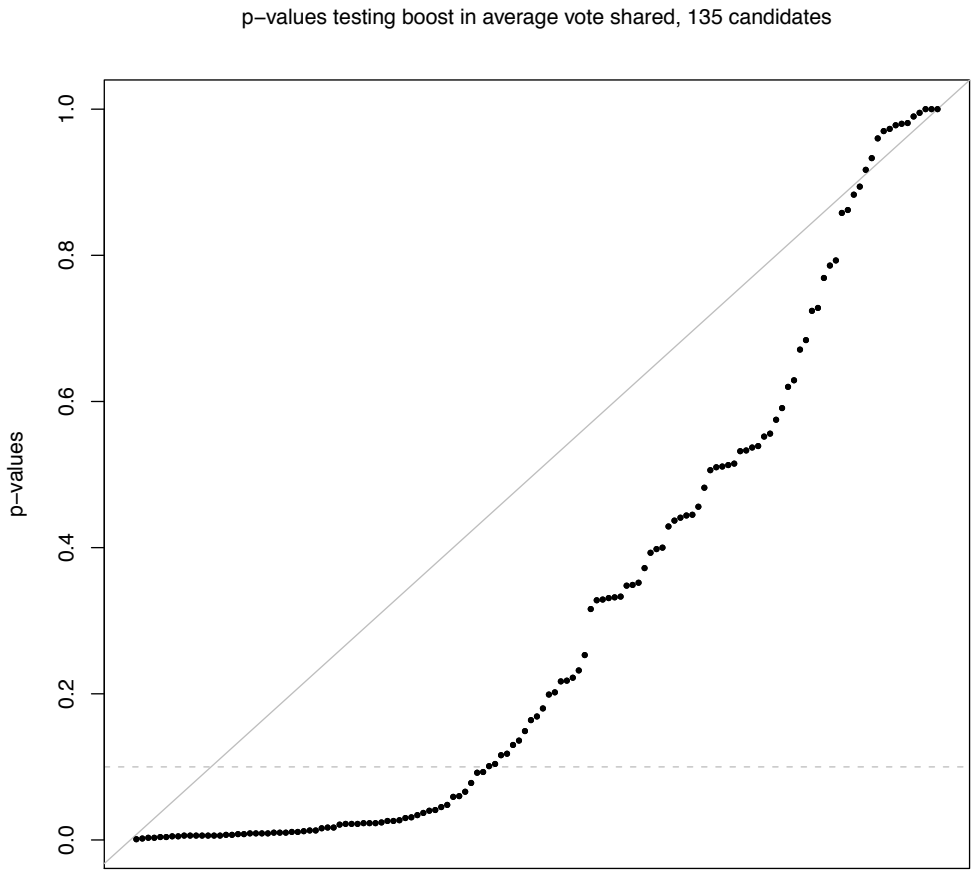
A little advanced

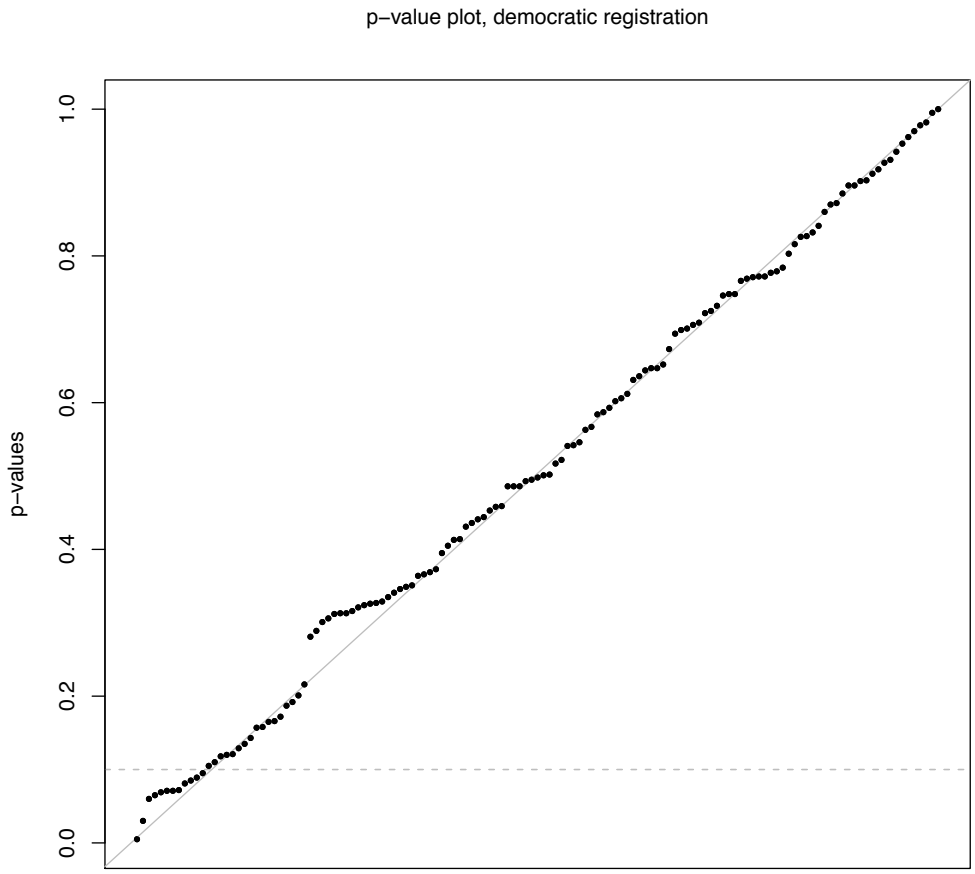
If you think about it a little, if the null hypothesis is true for all 135 candidates, then 5% of P-values should be less than 0.05 -- Similarly, 10% should be less than 0.1 and so on

This explains why, in the previous plot, we looked at **the proportion of candidates** rather than their number and we added **a line with unit slope** -- If all 135 null hypotheses were correct, we'd expect to see our sorted P-values track the line with unit slope

This is the case for Democratic registration totals and (next page) Green Party totals -- In short, t**hese pictures don't look like the ones we drew for 2003 vote shares**

Based on the P-value plots a few slides back and these "sanity check" plots, Ho and Imai conclude that the **none of the major candidates exhibited significant page-1 effects**, but that about **40% of the minor candidates saw a significant increase in vote shares** by being listed on the first page


p−values testing boost in average vote shared, 135 candidates
p−values
0.0
0.2
0.4
0.6
0.8
1.0


p−value plot, democratic registration
p−values
0.0
0.2
0.4
0.6
0.8
1.0

To sum

This example illustrates our tag line **“analyze as you randomized”** -- While the California Legislature was not thinking about analysis when they crafted their ballot placement protocol, **it presents us with an interesting “natural experiment”**

The initial randomization opens the possibility of analyzing the results statistically -- We are suddenly able **to make a more refined comment** on the effect of a page-1 spot on a ballot

Ho and Imai take this analysis farther and incorporate “covariates” to adjust for characteristics of the district-county pairs -- This is a bit beyond what we’ve covered so far in class but you can read the paper below if you like

Randomization Inference With Natural Experiments:
An Analysis of Ballot Effects in the 2003 California Recall Election

Daniel E. Ho and Kosuke Imai,
Journal of the American Statistical Association, Vol. 101, No. 475.