

Lecture 5: It's almost beautiful

Last time

We gave you a little tour of R in an attempt to better prepare you for doing something really innovative with the Registrar's data! We talked a little about the history of R and then operations on vectors (we'll do something like this every other lecture or so)

Then, we looked at inference, examining a simple randomized controlled trial that comes up in A/B testing -- We took a small historical detour and talk about the first such trial (which came up in a medical application)

Randomized controlled trials

We began our discussion of inference with a simple experimental design -- Subjects are randomized into treatment and control groups and then some outcome is measured at the end of the experiment

The results are displayed in the form of a simple table -- We started last lecture with A/B testing on web sites but took a slight detour with a medical application

Let's pick up there...

Hill's tuberculosis trial

Recall the data from Hill's trial -- Here we have patient's general condition on admission to his clinic and on the next page we present mosaic plots of these data

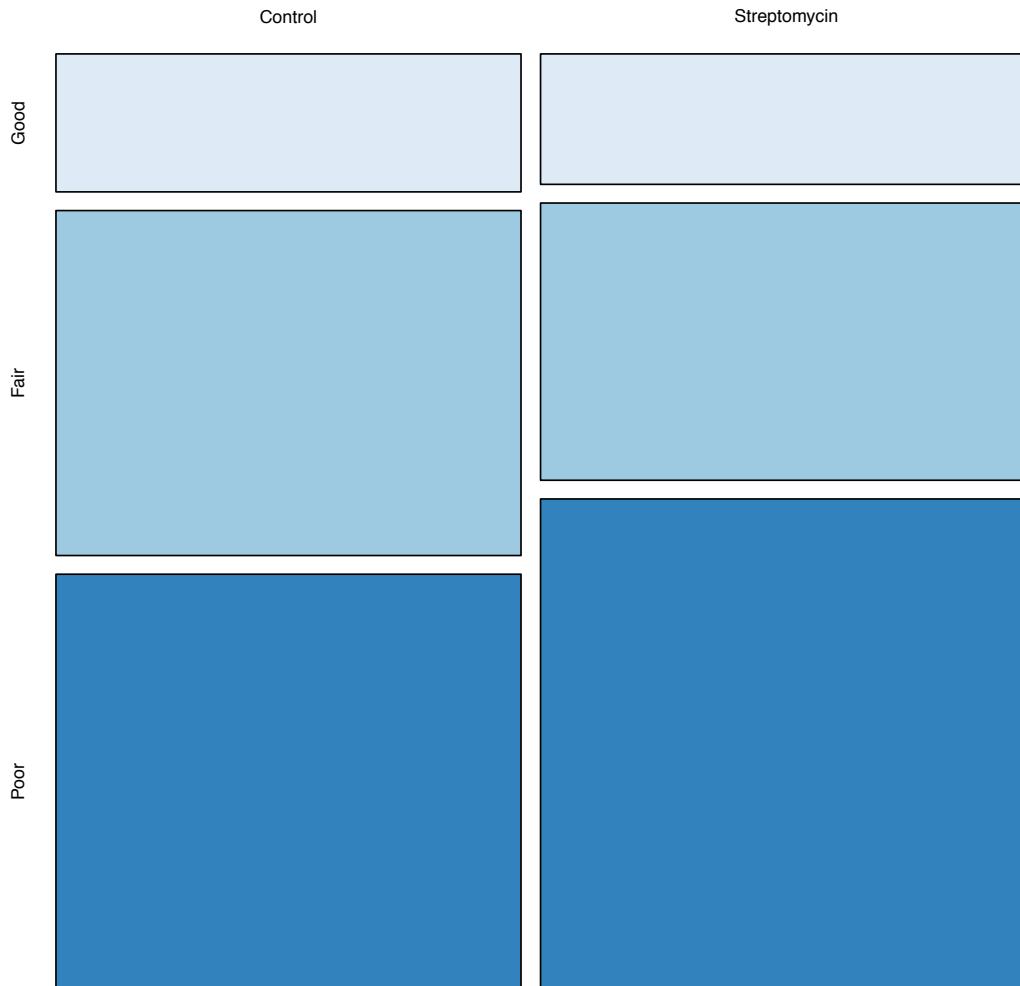
What do we notice? What should we see if the design was working correctly?

TABLE I.—Condition on Admission

General Condition	S Group	C Group	Max. Evening Temp. in First Week*	S Group	C Group	Sedimentation Rate	S Group	C Group
Good ..	8	8	98-98.9° F. (36.7-37.15° C.)	3	4	0-10	0	0
Fair ..	17	20	99-99.9° F. (37.2-37.75° C.)	13	12	11-20	3	2
Poor ..	30	24	100-100.9° F. (37.8-38.25° C.)	15	17	21-50	16	20
			101° F. (38.3° C.) †	24	19	51+	36	29
Total	55	52	Total	55	52	Total	55	51†

* Temperature by mouth in all but six cases. †Examination not done in one case.

Condition on admission



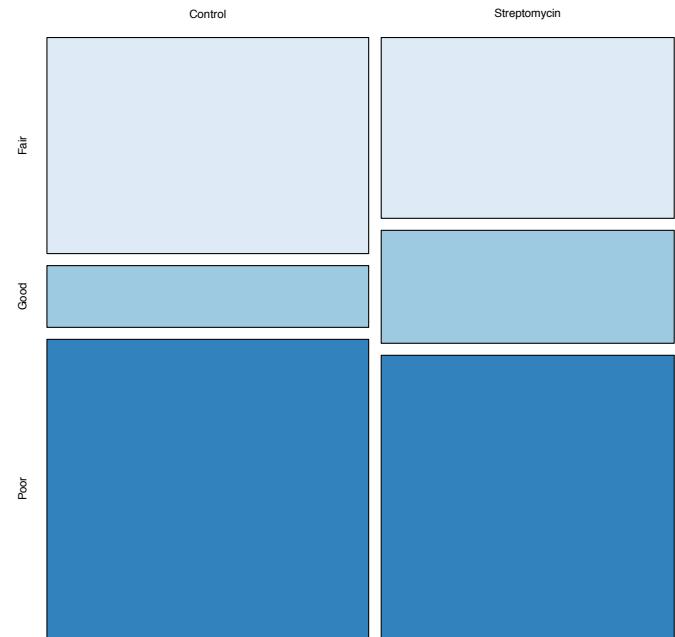
Admission status

This division was purely the result of the randomization into treatment and control -- For each of the 107 patients, Hill essentially tossed a coin and routed them to treatment or control

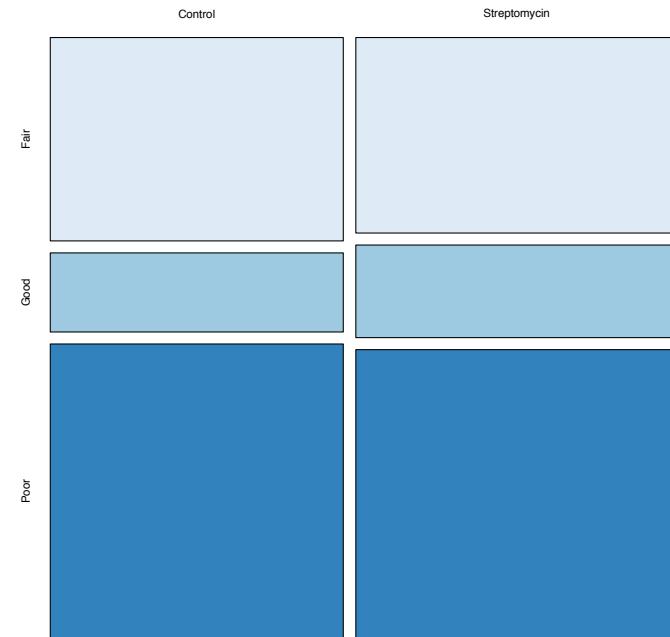
Because of that, any pattern we see in this plot is purely the result of randomization and hence “noise” -- Again, this evaluation of patients’ conditions was done prior to any treatment

We can imagine repeating Hill’s randomization at entry and re-randomize patients into treatment and control and look at the plots of their conditions in each arm of the study -- Again, any pattern we see is purely the result of randomization as we are tossing a coin for each patient

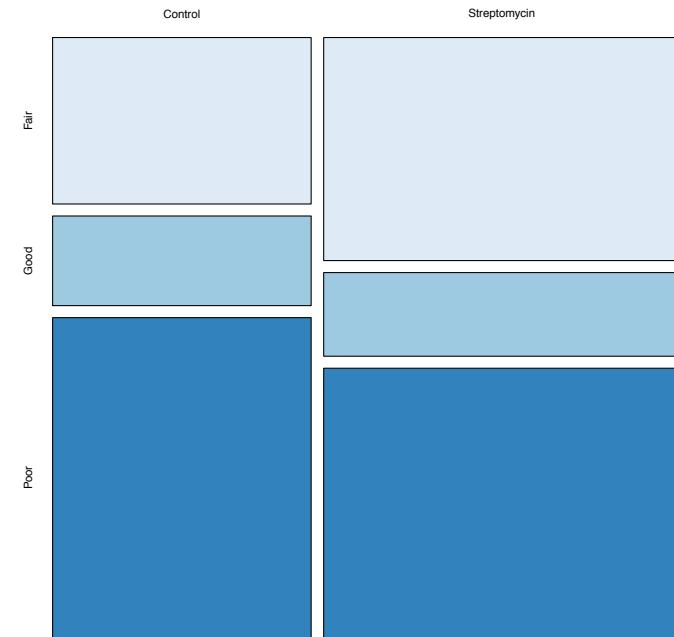
Condition on admission, alternate randomization



Condition on admission, alternate randomization



Condition on admission, alternate randomization



Hill's data

The plots on the previous page show us possible alternative experiments that could have been run -- Each is a division into treatment and control

Hill's data came from just one of these -- After the randomization he then treated patients with Streptomycin or bed rest (which is the control)

Hill's tuberculosis trial

And here are Hill's original results from his 1948 paper; what do we see?

TABLE II.—*Assessment of Radiological Appearance at Six Months as Compared with Appearance on Admission*

Radiological Assessment	Streptomycin Group		Control Group	
Considerable improvement ..	28	51%	4	8%
Moderate or slight improvement	10	18%	13	25%
No material change	2	4%	3	6%
Moderate or slight deterioration	5	9%	12	23%
Considerable deterioration ..	6	11%	6	11%
Deaths	4	7%	14	27%
Total	55	100%	52	100%

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ 1948; 2: 769-782.

Some analysis with Hill's data

Here we create a 2x2 table for Hill's data; we will focus on whether or not patients survived to the end of the trial

		Treatment		
		C	S	
Status	Survived	38	51	
	Died	14	4	18
		52	55	107

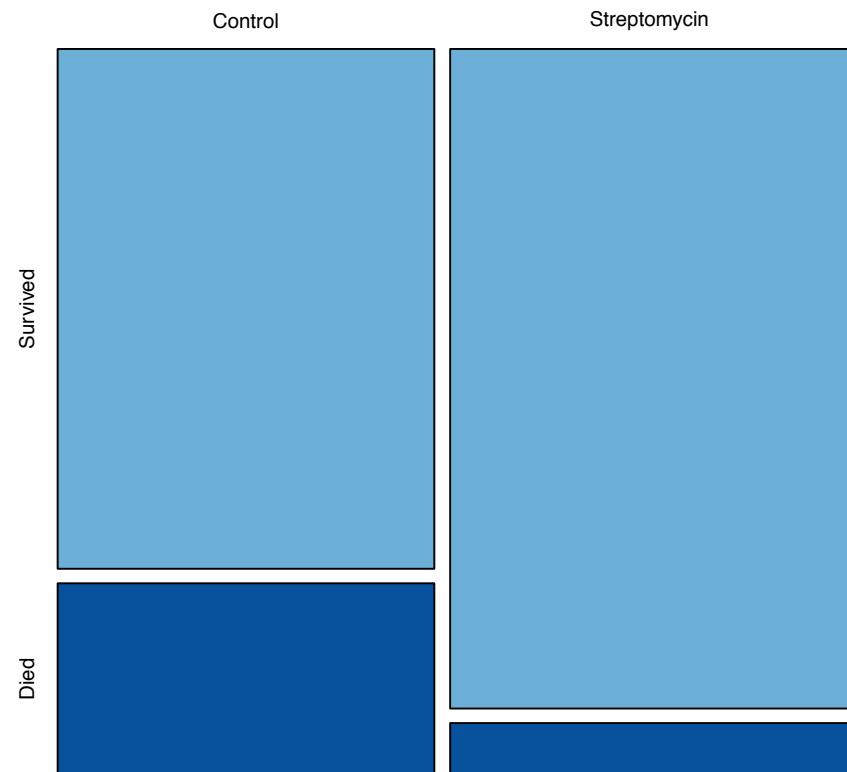
Another view

Here is a mosaic plot of Hill's tuberculosis study; it's worth taking a second look at the computations that go into this plot

Recall that the columns are sized according to the proportion of people receiving Streptomycin and the Control (slightly more received the treatment, 52 v. 55)

Then, within each column, the proportion of participants that survived is shaded yellow; we will call this the **conditional proportion who survived** given that a participant received either Streptomycin or the control

Hill's tuberculosis study



Some analysis with Hill's data

To work this out, we see that 14/52 or 27% of the patients receiving the control died; whereas 4/55 = 7.3% of those receiving Streptomycin died

What do we think?

		Treatment		
		C	S	
Status	Survived	38	51	
	Died	14	4	18
		52	55	107

Statistical analysis

On the face of it, things look promising for Streptomycin relative to the standard therapy, bed rest, but is that where our analysis stops?

How do we judge the size of an effect? In particular, could these results have occurred “by pure chance”?

And what is the model for chance here?

Randomized controlled trials

Below we present a simple cartoon of the steps that go into “analyzing” data from a trial such as this

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 made for the purposes of argument -- A good null hypothesis is a statement that would be interesting to reject
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 (if the hypothesized statement - or model or scenario - was true)
3. We then simulate values of the test statistic using the null hypothesis -- Today this will mean **simulating a series of data sets assuming the null hypothesis is true**, and for each computing the test statistic (the ensemble of simulated test statistics is often called a null distribution, but we'll talk about this more formally when we review probability)
4. And finally, **we compare** the value of the test statistic calculated for our data and compare it to the values that were obtained by simulation -- If they are very different, we have evidence that the null hypothesis is wrong

Hill's tuberculosis study

So let's talk about each of these components in the context of Hill's randomized trial -- When testing the efficacy of a new medical procedure, **the natural null hypothesis is that it offers no improvement over the standard therapy**

Under this “model” we assume that the two treatments are the same, so that patients would have had the same chance of survival under either -- Put another way, their outcome, whether they lived or died, would have been the same regardless of which group they were placed in

Under this hypothesis, the table we see is merely the result of random assignment -- That is, 18 people would have died regardless of what group we assigned them to, and **the fact that we saw 4 in the Streptomycin group and 14 in the control group was purely the result of chance**

Hill's tuberculosis study

Therefore, under the null hypothesis, if we had chosen a different random assignment of patients, **we would still have 18 people who died and 89 who survived, but they would appear in different cells of the table**

We can simulate under this “model” pretty easily -- That is, we take the 18 people who died and the 89 who survived and we re-randomize, **assigning 52 of them to the control group and 55 to the treatment group**

Let's see what that produces...

Simulating random assignments

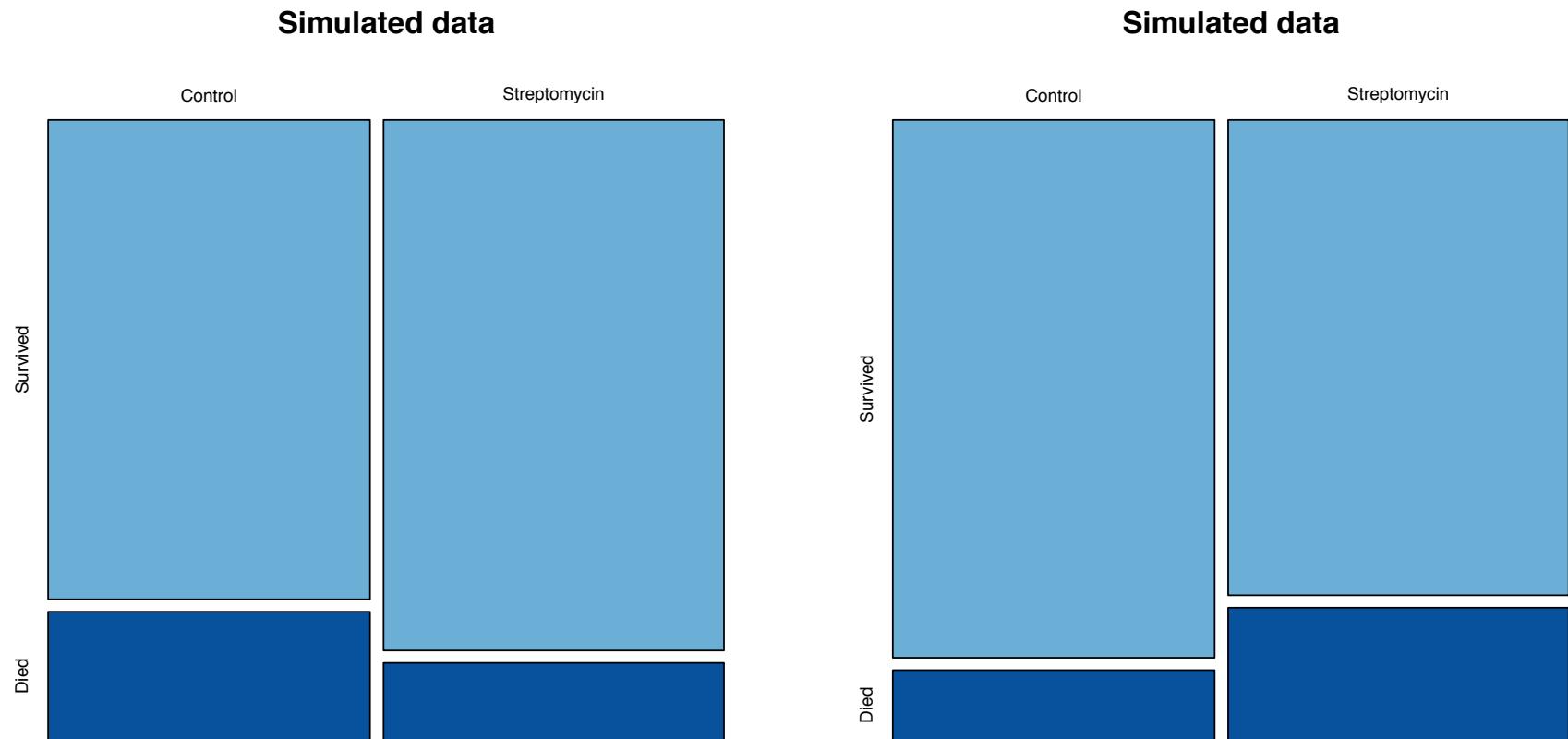
In this simulated table, we have 11/52 or 21% chance of dying under the control, and a 7/55 or 12% chance under Streptomycin; the treatment reduced the mortality rate among the participants by nearly 60%

		Treatment		
		C	S	
Status	Survived	41	48	89
	Died	11	7	18
		52	55	107

Simulating random assignments

In this simulated table, we have the opposite, with 6/52 or 12% chance of dying under the control, and a 12/55 or 22% chance under Streptomycin; the treatment almost doubled the mortality rate among the participants

		Treatment		
		C	S	
Status	Survived	46	43	89
	Died	6	12	18
		52	55	107



Simulating random assignments

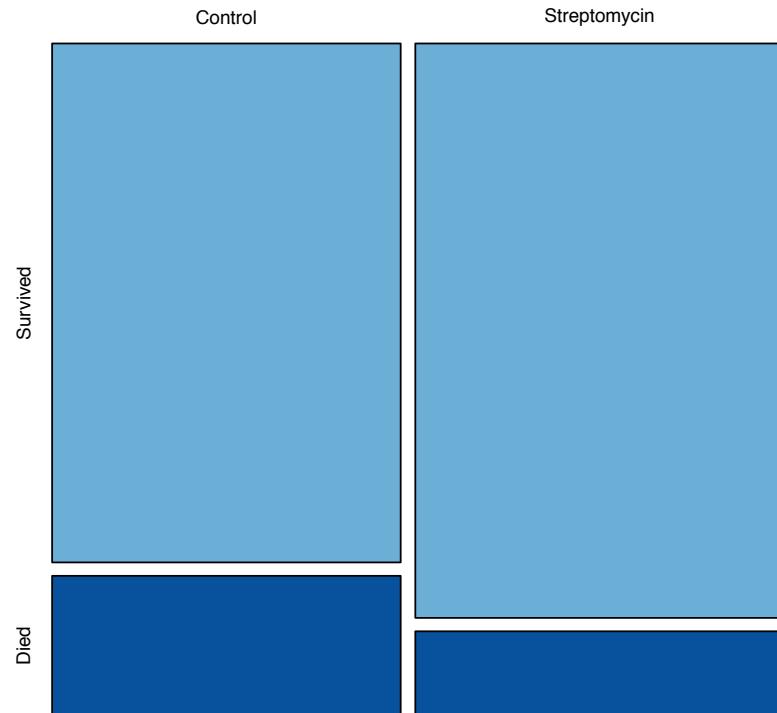
Notice that we only need to record **one piece of information for each trial, the number of deaths under Streptomycin** -- Knowing that we know all the other entries in the table

Using the language of hypothesis testing, we will take **the number of patients in the Streptomycin group that died as our test statistic**

Therefore, the question becomes, under the random assignment patients to treatments, **how common is it for us to see 4 or fewer deaths in the Streptomycin group?**

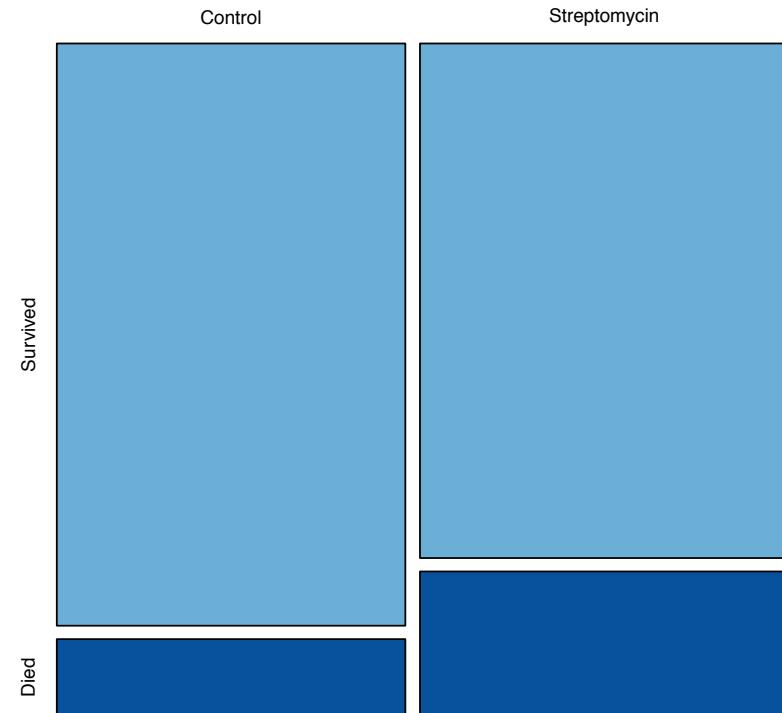
How would we figure this out?

Simulated data

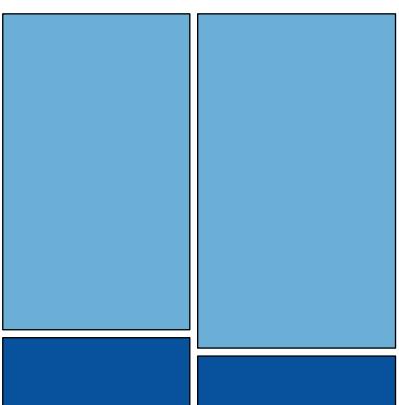
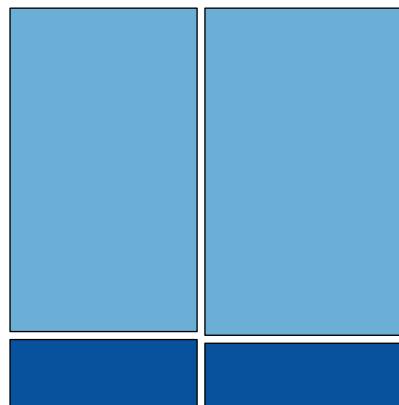
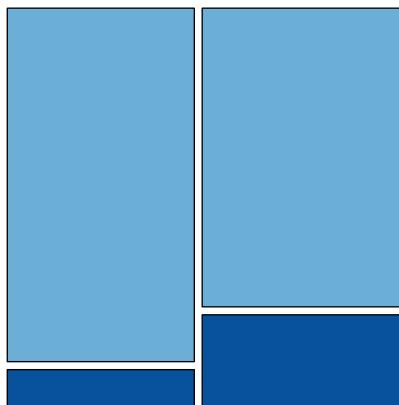
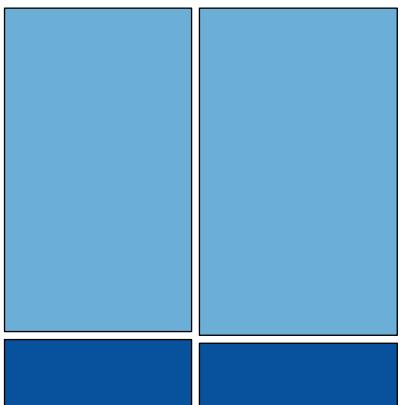
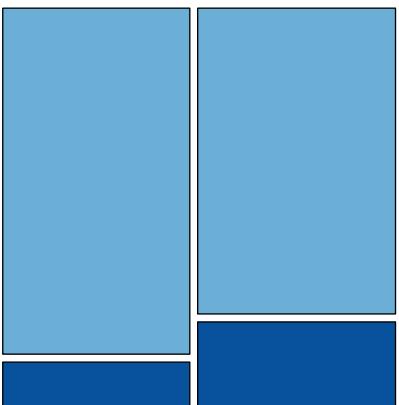
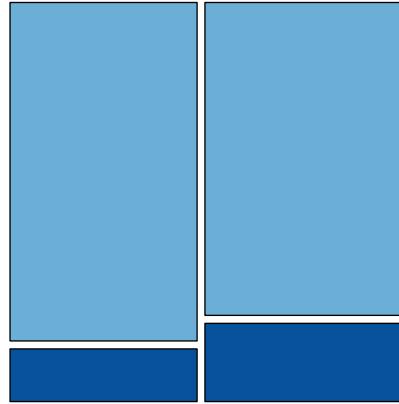
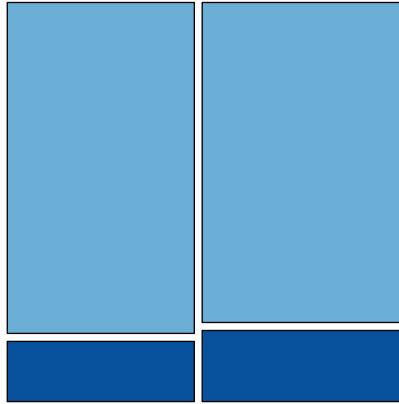
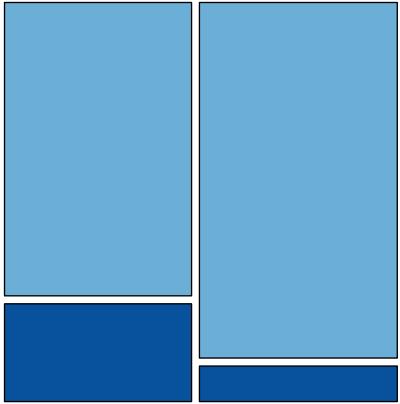
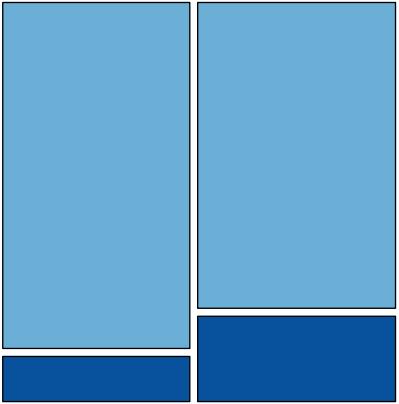


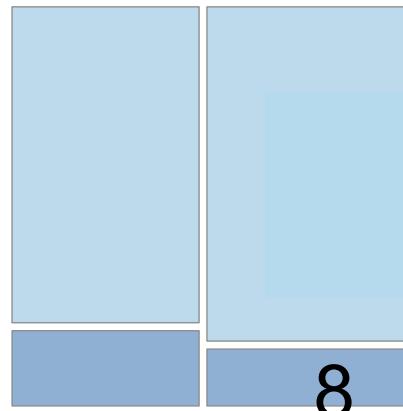
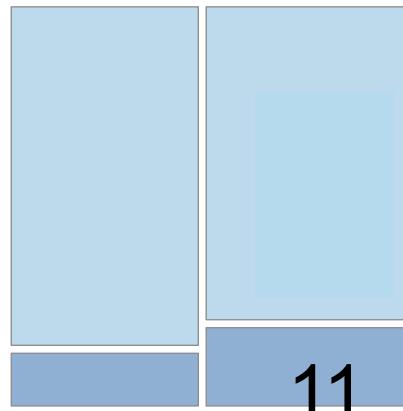
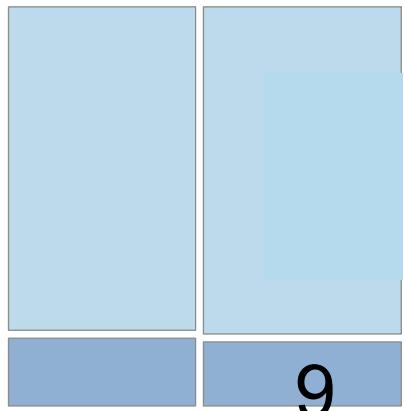
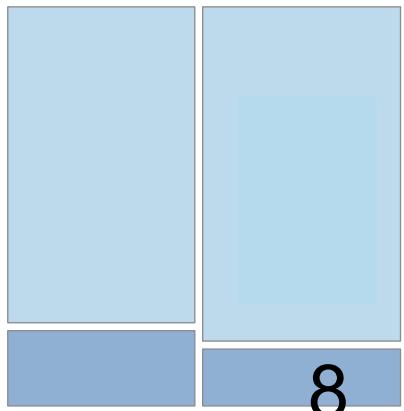
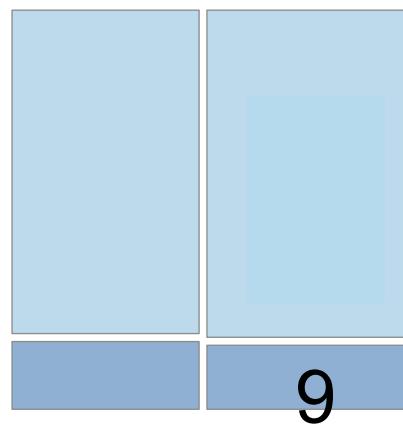
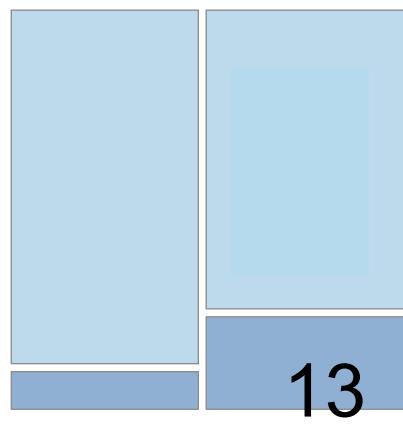
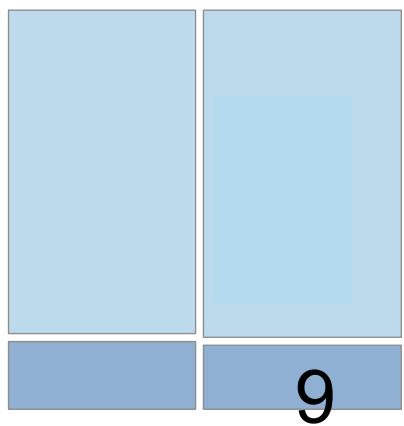
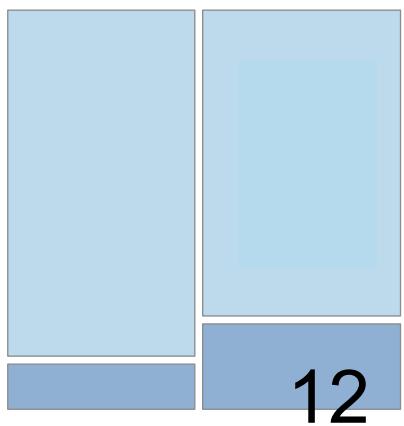
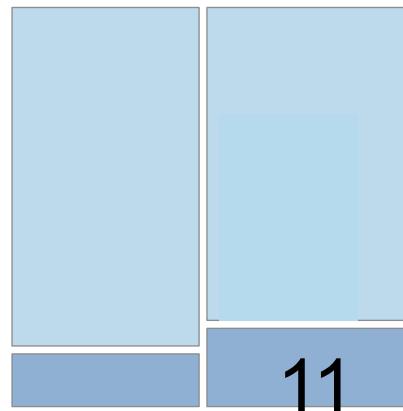
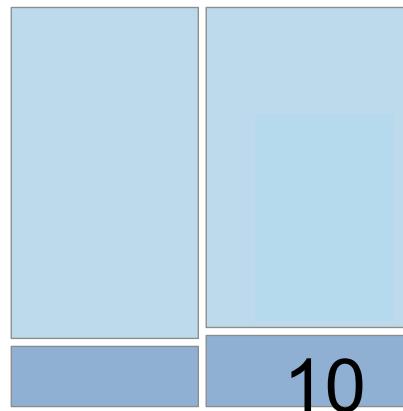
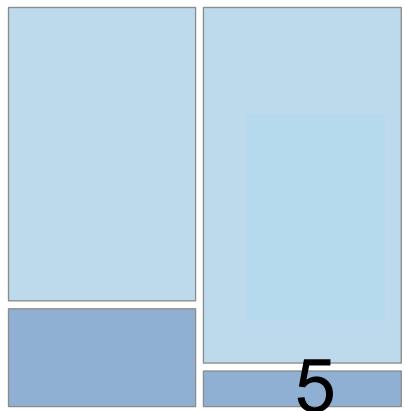
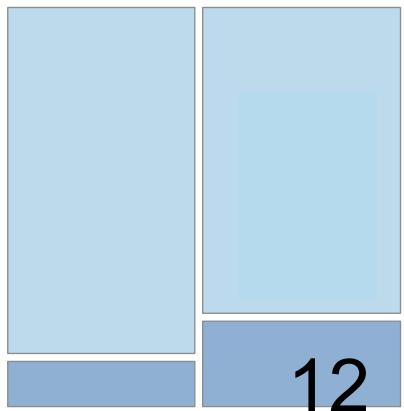
7 deaths in the Streptomycin group

Simulated data



12 deaths in the Streptomycin group





12

5

10

11

12

9

13

9

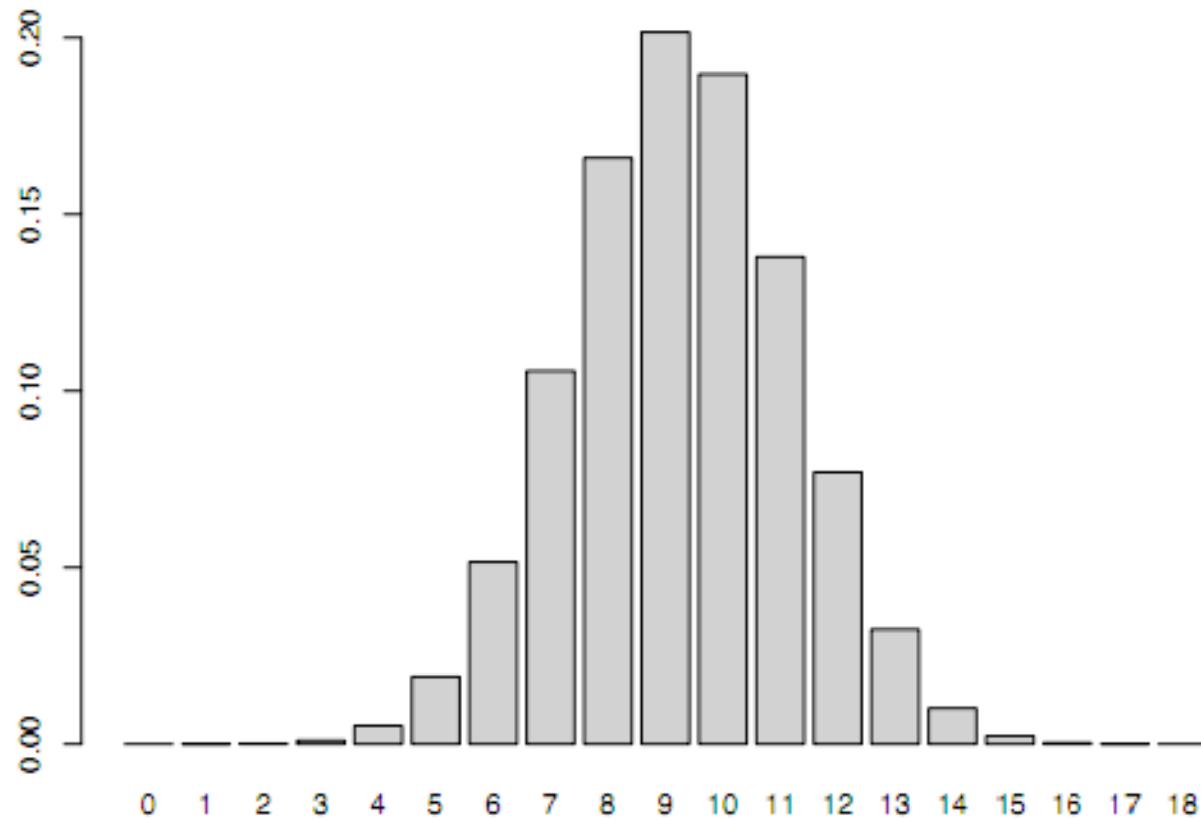
8

9

11

8

Proportion of simulated tables with n deaths under Streptomycin

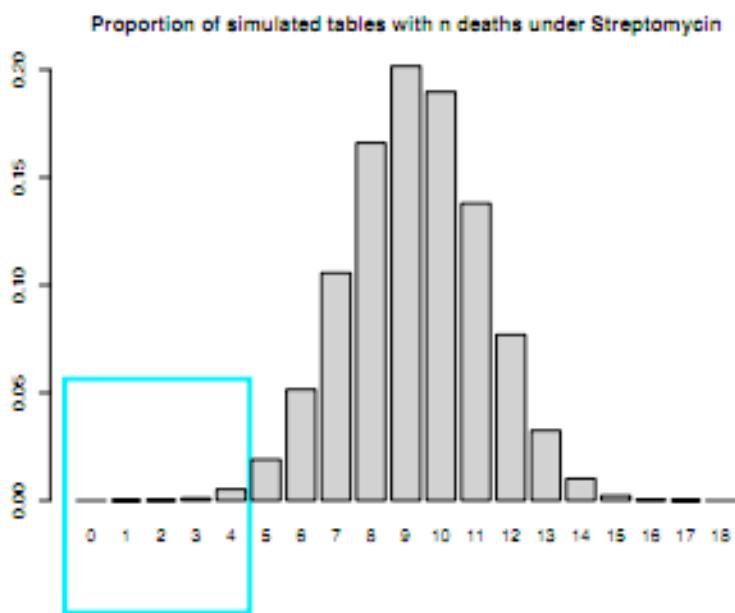


Simulating random assignments

In this plot we see that a value as small or smaller than four is fairly rare; to be precise, only 0.6% of the tables have 4 or fewer deaths in the Streptomycin group

This, then, provides us with evidence that there is something more at work here than random assignment

If we believed the null hypothesis, that there was no difference between Streptomycin and bed rest, the results Hill observed would have been extremely rare, coming up a very small fraction of the time



Hypothesis testing

The value 0.006 (the proportion of random tables with 4 or fewer deaths in the Streptomycin group) is also known as a P-value -- In general, **the P-value for a hypothesis test refers to the chance that we see a test statistic as or more extreme than the one you computed for your data**

Fisher proposed this measure to express **the weight of evidence against a null hypothesis -- the smaller the value, the stronger the evidence**; it was meant to be combined with other sources of information as you reason about the phenomenon you're studying

Keep in mind, however, that rare things do happen (but only rarely) -- It is possible that the null hypothesis is correct (and Streptomycin is no more effective than bed rest) and **Hill could have been incredibly unlucky** in selecting a division into groups that produced 4 deaths in the Streptomycin group

This is the nature of statistical reasoning, and this is why Fisher advocated performing many experiments as you study a phenomenon

Hill's tuberculosis trial

To sum up:

1. The null hypothesis for Hill's experiment was that Streptomycin and the standard therapy, bed rest, would perform the same when treating pulmonary tuberculosis
2. We took as our test statistic the number of patients that died in the Streptomycin group
3. Under the null hypothesis of no difference, we repeated Hill's randomization a large number of times, with each one we recorded the number of deaths assigned to the Streptomycin group
4. We then looked at the fraction of random assignments that gave us 4 or fewer deaths in the Streptomycin group and determined it was extremely rare; we took this as evidence that the null hypothesis is wrong, that something other than simple chance assignment could explain the data he collected

Back to our A/B test

Recall the data for the Times Travel Section experiment -- Here we were comparing a series of two short lists (control, most e-mailed stories and most popular cities) versus a proposed tab layout (treatment, twice the number of items in each list, but overlaid)

Here are the two designs again...

List: Variation 10858

Welcome to TimesPeople
What's this?

Share and Discover the Best of NYTimes.com

10:27 AM Log In or Register
No, thanks

Flamboyance Gets a Face-Lift
By RUTH LA FERLA

The Fontainebleau hotel chases its former glory and the crowds of South Beach.

[Travel Guide: Miami >](#)

SQUARE FEET

Detroit Revives a Hotel and Some Hope
By KEITH SCHNEIDER

The completion of a \$200 million renovation of the Book Cadillac hotel in downtown Detroit is another sign for residents that the city is working to regain some polish and prestige.

• [Slide Show: The Westin Book Cadillac Hotel](#)

ON THE ROAD

Yes, a Room's Available. But No, You Can't Check In.
By JOE SHARKEY

With hotel profits under siege, this is not the time to be making your most loyal customers unhappy.

• [Itineraries: In-Flight](#), and [Stuck With a Seatmate's Politics](#)
• [Frequent Flier: It's All About the Shoot](#), and the Ability to Scramble
• [US Airways to Charge for Pillows and Blankets](#)

NEXT STOP

Is Tel Aviv Ready to Crash the Global Art Party?
By ROBERT GOFF

The city is Israel's contemporary arts capital, where young artists live, work and show their wares in more than 30 contemporary galleries.

[Travel Guide: Tel Aviv >](#)
[Interest Guide: Art >](#)

CULTURED TRAVELER

Where Words Took Shape: Saul Bellow's Chicago
By JON FASMAN

The city's rough vitality remains strong in

Travel Q&A Blog
Tour groups that cater to solo female travelers.

[Go to Travel Q&A >](#)

Escapes

A tour through two quirky neighborhoods in Seattle, a detailed look at the Smithsonian's Air and Space Museum annex, how brokers' blogs are helping second-home buyers and more.

[Go to Escapes >](#)

Featured Interest Guide: Wildlife

Discover how animals in the

④ Historic Deerfield
A museum of history, art, and architecture in an authentic New England village

Art | Books | History 
[www.deerfield.org](#)

Times Delivers E-Mail

Sign up | Sign in | See what's new | Sign Up

List of emailed and cities without header

Most Emailed

1. Globespotters: Hiking into Chinese History
2. Savoring Italy, One Beer at a Time
3. 36 Hours in Burlington, Vt.
4. Cultured Traveler: Where Words Took Shape: Saul Bellow's Chicago
5. American Journeys: A Seattle That Won't Blend In

[Go to Complete List >](#)

Top 5 Cities

1. New York City
2. Paris
3. Chicago
4. Venice
5. Burlington

The New York Times STORE

Tabs: Variation 10859

Welcome to TimesPeople

What's this?

Share and Discover the Best of NYTimes.com

Log In or Register

No, thanks

Sign Up

See Sample

Tab of emailed and cities

ON THE ROAD

Yes, a Room's Available. But No, You Can't Check In.

By JOE SHARKEY

With hotel profits under siege, this is not the time to be making your most loyal customers unhappy.

- Itineraries: In-Flight, and Stuck With a Seatmate's Politics
- Frequent Flier: It's All About the Seat, and the Ability to Scramble
- US Airways to Charge for Pillows and Blankets

NEXT STOP

Is Tel Aviv Ready to Crash the Global Art Party?

By ROBERT GOFF

The city is Israel's contemporary arts capital, where young artists live, work and show their wares in more than 30 contemporary galleries.

Travel Guide: Tel Aviv »

Interest Guide: Art »

CULTURED TRAVELER

Where Words Took Shape: Saul Bellow's Chicago

By JON FASMAN

The city's rough vitality remains strong in Humboldt Park, where the Nobel Prize-winning writer grew up.

Travel Guide: Chicago »

GLOBESPOTTERS

Hiking Into Chinese History

By JEREMY GOLDKORN

You can combine historical pursuits with some of the finest day hiking in China around the village of Fanzipai.

Travel Guide: China »

Interest Guide: History »

Savoring Italy, One Beer at a Time

By EVAN RAIL

In the regions of Lombardy and Piedmont, a nascent craft beer scene has begun to emerge, bringing well-made brews into the dining rooms of some of the country's best restaurants.

A tour through two quirky neighborhoods in Seattle, a detailed look at the Smithsonian's Air and Space Museum annex, how brokers' blogs are helping second-home buyers and more.

Go to Escapes >

Featured Interest Guide: Wildlife

Discover how animals in the Great Plains are attracting eco-tourists and get tips on seeing New England's fall foliage.

Go to the Wildlife Guide >

Activity & Interest Guides

Browse free Times articles.

Choose a Category





MOST POPULAR - TRAVEL

E-MAILED CITIES

1. Globespotters: Hiking Into Chinese History
2. Savoring Italy, One Beer at a Time
3. 36 Hours in Burlington, Vt.
4. Cultured Traveler: Where Words Took Shape: Saul Bellow's Chicago
5. American Journeys: A Seattle That Won't Blend In
6. Next Stop: Is Tel Aviv Ready to Crash the Global Art Party?
7. An Hour From Paris: North of Paris, a Forest of History and Fantasy
8. Weekend in New York: Some Tourists Don't Need Advice
9. Practical Traveler: Readers Sound Off on Private Rentals
10. Comings and Goings: Traveling in Style Through Rural Italy

Go to Complete List >

The New York Times STORE

NYT Ortelius Maps Edition -- Africa
Buy Now

Back to our A/B test

And here are the data from the experiment

```
> table(nyt$Variation,nyt$IfClicked)
      0      1
List 65181    766
Tabs 64836   1244

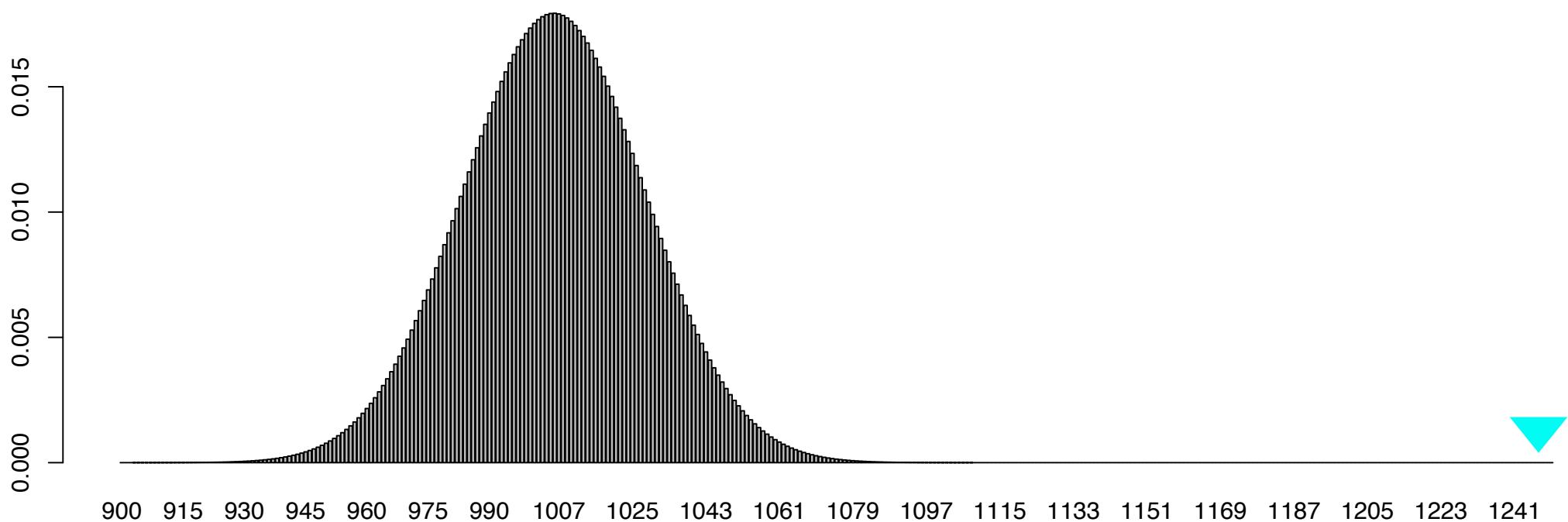
> 766/(766+65181)
[1] 0.01161539

> 1244/(1244+64836)
[1] 0.01882567

> (1244/(1244+64836))/(766/(766+65181))
[1] 1.620752
```

Following the re-randomization procedure we've just introduced, let's have a look at the reference distribution under the null hypothesis that Lists and Tabs are clicked on equally...

proportion of tables



Generalities

In this case the P-value is vanishingly small -- It is clear that the difference in usage between tabs and lists is not due to the randomization alone and there must be some effect of the treatment

Notice that in Hill's trial and in our A/B testing example, **random assignment gives our analysis its validity** -- The way we collect data dictates the kinds of inferences we are allowed to make

However, have have not said anything about what the effect of Streptomycin might have on patients outside the study or how attractive the tabs and lists are to NYT visitors generally -- For that, we have to make more assumptions about how people were recruited into these trials

Fisher and randomization

"The theory of estimation presupposes a process of random sampling. All our conclusions within that theory rest on this basis; without it our tests of significance would be worthless. ... In controlled experimentation it has been found not difficult to introduce explicit and objective randomisation in such a way that the tests of significance are demonstrably correct. In other cases we must still act in faith that Nature has done the randomisation for us.... We now recognise randomisation as a postulate necessary to the validity of our conclusions, and the modern experimenter is careful to make sure that this postulate is justified."



Fisher RA. *Development of the theory of experimental design*. Proceedings of the International Statistical Conferences 1947;3:434–39

Significance testing

Our discussion of P-values and our examination of the null distribution are in line with the methodology advocated by Fisher throughout his career; **the null hypothesis plays the role of devil's advocate, and a P-value provides evidence against the null** -- this is often called **significance testing**

There are a few obvious questions facing practitioners, the first of which involves evaluating the evidence provided by a P-value -- **Is there a rule which helps you decide when you should “reject” the null hypothesis, or, rather, decide that it’s not true?**

Fisher wrote: *If [the P-value] is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05....*" (Fisher 1950) -- and certainly in his own work on agricultural field trials, used thresholds of 0.05 and 0.01 as guides to “reject” a null hypothesis

Still, Fisher believed that **the individual researcher should interpret a P-value** (a value of 0.05 might not lead to either belief or disbelief in the null, but to a decision to conduct another experiment); he wrote that the rigid use of thresholds was **the “result of applying mechanically rules laid down in advance; no thought is given to the particular case, and the tester’s state of mind, or his capacity for learning, is inoperative.”** (Fisher 1955, p.73-4).

Significance testing

Broadly, Fisher treated P-values as a tool for inductive inference -- His practice is expansive in the sense that suggested P-values be interpreted not through cutoffs but rather on **a case-by-case basis, weighing the investigator's other "evidence and ideas"**

But this stance meant he was at times inconsistent in his use of P-values and the article below does a nice job of pooling together some of his writings (I am not sure, however, that I agree with the conclusion of the article, but the background quotes from Fisher are worth it)

http://www.webpages.uidaho.edu/~brian/why_significance_is_five_percent.pdf

There are many theories and stories to account for the use of $P=0.05$ to denote statistical significance. All of them trace the practice back to the influence of R.A. Fisher. In SMRW (the book we mentioned in the first lecture) Fisher states

The value for which $P=0.05$... is convenient to take... **as a limit in judging whether a deviation ought to be considered significant or not....**

Similar remarks can be found in Fisher (1926, 504).

... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." ...

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, **the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.** A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

However, Fisher's writings might be described as inconsistent. On page 80 of SMRW, he offers a more flexible approach

In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed [test statistic], but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: Either the hypothesis is untrue, or the value of [the test statistic] has attained by chance an exceptionally high value. The actual value of P obtainable from the table... indicates the strength of the evidence against the hypothesis. **A value of [the test statistic] exceeding the 5 per cent point is seldom to be disregarded.**

These apparent inconsistencies persist when Fisher dealt with specific examples. On page 137 of SMRW, Fisher suggests that values of P slightly less than 0.05 are not conclusive.

[T]he results... show that P is between .02 and .05.

The result must be judged significant, though barely so; in view of the data we cannot ignore the possibility that on this field, and in conjunction with the other manures used, nitrate of soda has conserved the fertility better than sulphate of ammonia; **the data do not, however, demonstrate this point beyond the possibility of doubt.**

On pages 139-140 of SMRW, Fisher dismisses a value (0.008) greater than 0.05 but less than 0.10.

The difference between the regression coefficients, though relatively large, **cannot be regarded as significant**. There is not sufficient evidence to assert that culture B was growing more rapidly than culture A.

while in Fisher [19xx, p 516] he is willing pay attention to a value not much different.

...P=.089. Thus a larger value of [the test statistic] would be obtained by chance only 8.9 times in a hundred, from a series of values in random order. **There is thus some reason to suspect that the distribution of rainfall in successive years is not wholly fortuitous**, but that some slowly changing cause is liable to affect in the same direction the rainfall of a number of consecutive years.

Yet in the same paper another such value is dismissed!

[paper 37, p 535] ...P=.093 from Elderton's Table, showing that although there are signs of association among the rainfall distribution values, such association, if it exists, **is not strong enough to show up significantly** in a series of about 60 values.

Part of the reason for the apparent inconsistency is the way Fisher viewed P values. When Neyman and Pearson proposed using P values as absolute cutoffs in their style of fixed-level testing, Fisher disagreed strenuously. Fisher viewed P values more as measures of the evidence against a hypotheses, as reflected in the quotation from page 80 of SMRW above and this one from Fisher (1956, p 41-42)

The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. **A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions.** For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. **However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.**

Another test

Before we leave the A/B testing example, we can ask more of these data -- We have established that people in the study clicked on tabs more frequently than lists but what is the NY Times really after?

Pages per visit

The Times would like to see that one display is promoting greater engagement with its content -- One common metric is **the number of pages viewed** (stories read) while the visitor is on the Times' site

So our question is now, can we determine if there's a difference in, say, average page views per visit between groups A and B, between treatment and control, between tabs and lists?

Let's take as our statistic the difference in average page views between the two groups...

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 that for tabs is 1.997 which means the difference is 0.017 and is, well, tiny -- As a practical matter it probably doesn't amount to much (although small difference multiplied over millions of visits might prove important)

Page views per visit

As with the click-through statistic, we could ask if this 0.017 could have been **the result of randomization or if it represents a statistically significant difference** between the designs

Following the template before, we begin with a null hypothesis that the two layouts prompt people to visit the same number of pages

Under this assumption, people visiting the site would have viewed the same number of pages no matter what version of the site they were shown (tabs or lists) -- That suggests we can re-randomize visitors, allocating them again to treatment and control, and expect they would have viewed the same number of pages

Re-randomizing many times, each time computing the difference in average page views between the new tab and list groups, gives us a sense of the uncertainty under the null hypothesis

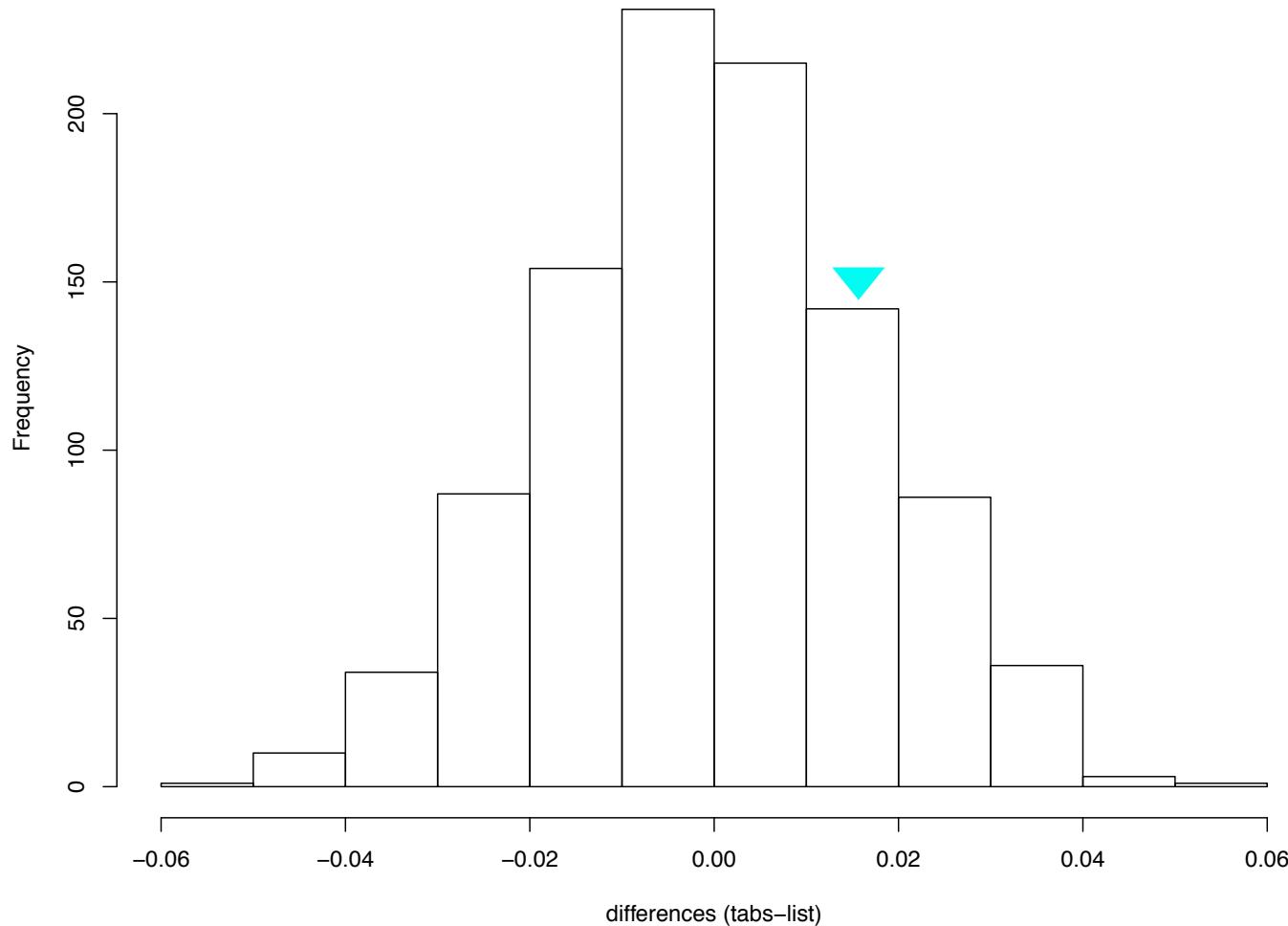
Some code

Here is some code that implements one re-randomization

```
> new_split <- sample(travel$Variation)
> new_diff <- mean(travel$Pageviews[new_split == "Tabs"]) -
  mean(travel$Pageviews[new_split == "List"])
```

These lines can then be wrapped into a “for” loop and repeated as many times as we want (a few thousand is sufficient)

histogram of differences (tabs–list) in average pv/visit, 1000 re-randomizations



Page views per visit

In this case, we don't have to be very formal about the fact that the observed difference of 0.017 is well within what we see from the null distribution

Therefore our observed difference in means “looks” like it could simply be the result of the randomization process -- In technical terms, we fail to reject the null hypothesis

Inductive inference v. inductive behavior

Let's return for a moment to Fisher's comments about fixed thresholds being absurdly academic -- What we see here is Fisher reacting to an alternative approach to inference one advocated by Jerzy Neyman and Egon Pearson

Neyman and Pearson disagreed with the subjective interpretation inherent in Fisher's approach and developed instead a procedure (which they termed hypothesis testing) based on **hard decisions about when to reject a null hypothesis** -- In effect they imposed a threshold called **the significance level**

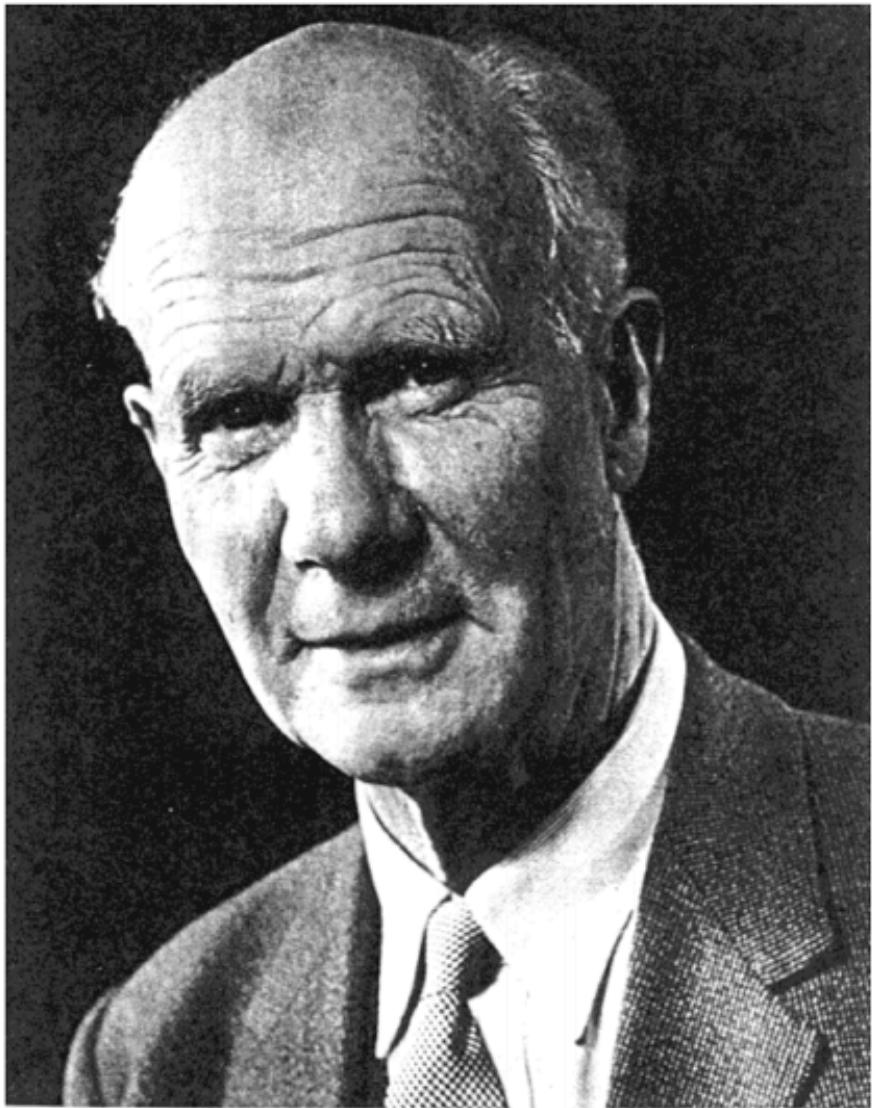
Hypothesis testing, as it is covered in most introductory texts (including the one we are using), is a slightly uncomfortable synthesis of Fisher's ideas (the P-value) together with the Neyman-Pearson framework

Here's how extreme the Neyman-Pearson approach was...



J. Neyman

Courtesy of University Statistics Department Archive



EGON SHARPE PEARSON

*"No test based upon a theory of probability, can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, **we may search for rules to govern our behaviour** with regard to them, in following which **we insure** that, in the long run of experience, **we shall not often be wrong.**"*

Neyman and Pearson, 1933

Hypothesis testing

The hypothesis testing setup begins with **both a null hypothesis as well as a complementary alternative hypothesis**

1. For Hill's study, our null was that Streptomycin was the same as bed rest when it came to treating pulmonary tuberculosis; and the alternative was that it was better, saving more lives
2. For the Travel Section trial, the null is that the two designs produce the same number of page views, with the alternative that they are different

In the first case, the experimenter would not have believed that Streptomycin was worse than bed rest because there had been plenty of studies prior that suggested improvement (he would have thought something went wrong with the trial rather than accept that the treatment did harm) -- In the second case, the experimenters had no idea which would be better and left the alternative open

One- or two-sided?

The choice of alternative dictates which tail of the null distribution we look to for evidence, whether we conduct a one- or two-sided test

The decision in clinical trials is hotly debated; some suggest one criterion would be the **impossibility or extreme improbability of a difference in one of the directions** ('Before the data are examined one should decide to use a one-sided test only if it is quite certain that departures in one particular direction will always be ascribed to chance, and therefore regarded as nonsignificant however large they are'. - Armitage)

A second criterion involves **anticipated actions that would result** from knowing the true value of the effect (one sided tests are appropriate 'when a large difference in one direction would lead to the same action as no difference at all'. - Bland and Altman)

Sometimes the decision is not in the hands of the researcher; several medical and epidemiological journals insist on the use of a two-sided test (for example, in the Instructions to Authors, the Journal of the National Cancer Institute states: 'All tests of significance should be two-sided' -- remember, one-sided tests have a relatively easier time achieving significance, the P-value being half of its two-sided friend)

Hypothesis testing

In addition to the null and alternative hypotheses, we also have to determine a threshold for the P-value, the significance level α ; these are all to be decided before you collect any data

On what grounds should you choose α ?

Hypothesis testing

Neyman and Pearson discuss the consequences of making a decision based on data; in forcing you to make a choice, you are subject to two kinds of errors

Type I: You reject a null hypothesis that is in fact true (for example, we conclude that Vioxx has a greater CE risk than naproxen, when it doesn't)

Type II: You fail to reject the null hypothesis when it is in fact false (for example, we conclude that Streptomycin is no more effective than bed rest, when in fact it is)

Let's consider these two separately...

Type I error

If the null hypothesis is true, then our null distribution (the simulations we have been working with) **accurately reflects the variation in our data** (coming from randomization in the cases we have seen so far)

If our test statistic is in the “extremes” of this null distribution we will incorrectly reject the null hypothesis and commit a Type I error; we can “control” the probability of this happening by setting the significance level α

That is, if we set a threshold at $\alpha = 0.05$ and only reject when our P-value is 0.05 or less, then when the null is true, **we will be making a mistake 5% of the time** (or in 1 out of 20 experiments we run); if we want to guard against this, we might consider lower thresholds, say $\alpha = 0.02$ or $\alpha = 0.01$

Type II error

In this case, the null hypothesis is not correct, but we fail to reject it; we often refer to the probability of making a Type II error as β where $1 - \beta$ is also called the power of the test ($1 - \beta$ is the chance that we correctly reject a null hypothesis that is not true)

The Type II error is much harder to work out in general; obviously, it will relate to **the size of the difference from the null** (are people twice as likely to improve on Streptomycin? 1.2 times more likely?)

When we get to more complicated settings in the next lecture, we will see that there are other measures of noise in the data that will impact Type II errors

Finally, **the sample size plays an enormous role in determining power or hence Type II error**; in general, the larger your sample size, the greater the power to detect a difference; we will see an example of that later in the quarter

Hypothesis testing

Here, then, are the steps for conducting a hypothesis test -- They are only a little different than what we presented for Fisher's approach

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 through computer simulation or some more precise mathematical calculation (upcoming lectures)
5. And finally, after the data are collected, we compare the probability of seeing 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**

Hypothesis testing

Importantly, in the hypothesis testing framework, **we don't report P-values at the end of the analysis** and instead we report the result of our decision -- The actual P-value doesn't contain useful information for Neyman and Pearson

Again, Neyman and Pearson are more interested in **behaviors and decision making** than the “strength of evidence” concept from Fisher

As a practical matter, researchers (because of the murkiness of most textbooks) seem to subscribe to both schools of thought and report a significance test result as well as a P-value (sigh)

A problem with thresholds

In my mind, this decision-oriented framework has a lot of problems -- First off, by setting a hard threshold for what is and what is not “significant”, **researchers routinely face publication barriers that favor significant results** (if not require them outright)

Hard limits define which effects are reported and which are not -- This sets up a situation in which **researchers are incentivized to be on one side of that line or the other**

There is something more human about Fisher’s view of a researcher taking their results in context and not blindly applying a threshold; in lab you have been looking at one example when this kind of mechanistic thinking led to real problems..

An alternative

Many disciplines are transitioning to publishing “effect sizes” and confidence intervals over simple significance tests -- These allow us to **assess the practical importance of a result** along with some notion of **the precision or reliability of the result**

In short, the “new view” in many fields emphasizes **estimation over testing** --
And this is where we are headed!

A template

This simple framework provides you with the basic motivation for a great deal of statistical inference -- We will typically create a reference distribution (a “sampling distribution”) of some test statistic and then use that distribution to quantify the randomness in our data

Here, the randomness was the result of the randomized nature of the design -- In material we'll start Monday, the randomness comes from sampling from a population

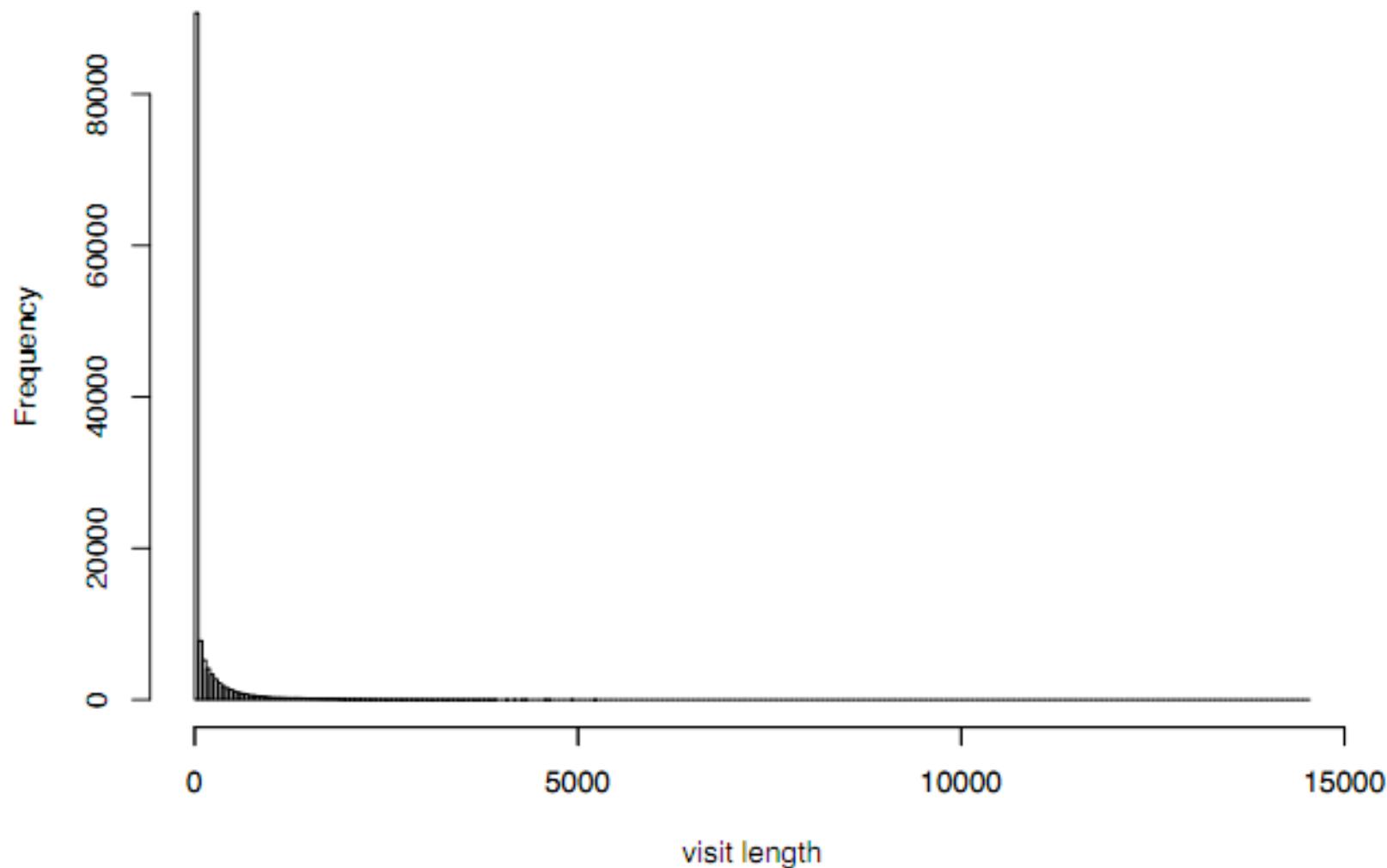
The techniques we start with are simple and computational and, frankly, provide you with a framework to analyze a large number of situations -- In many cases, we can leverage effects like the central limit theorem to come up with analytical (um, mathematical) approximations (under certain assumptions) to sampling distributions

The data

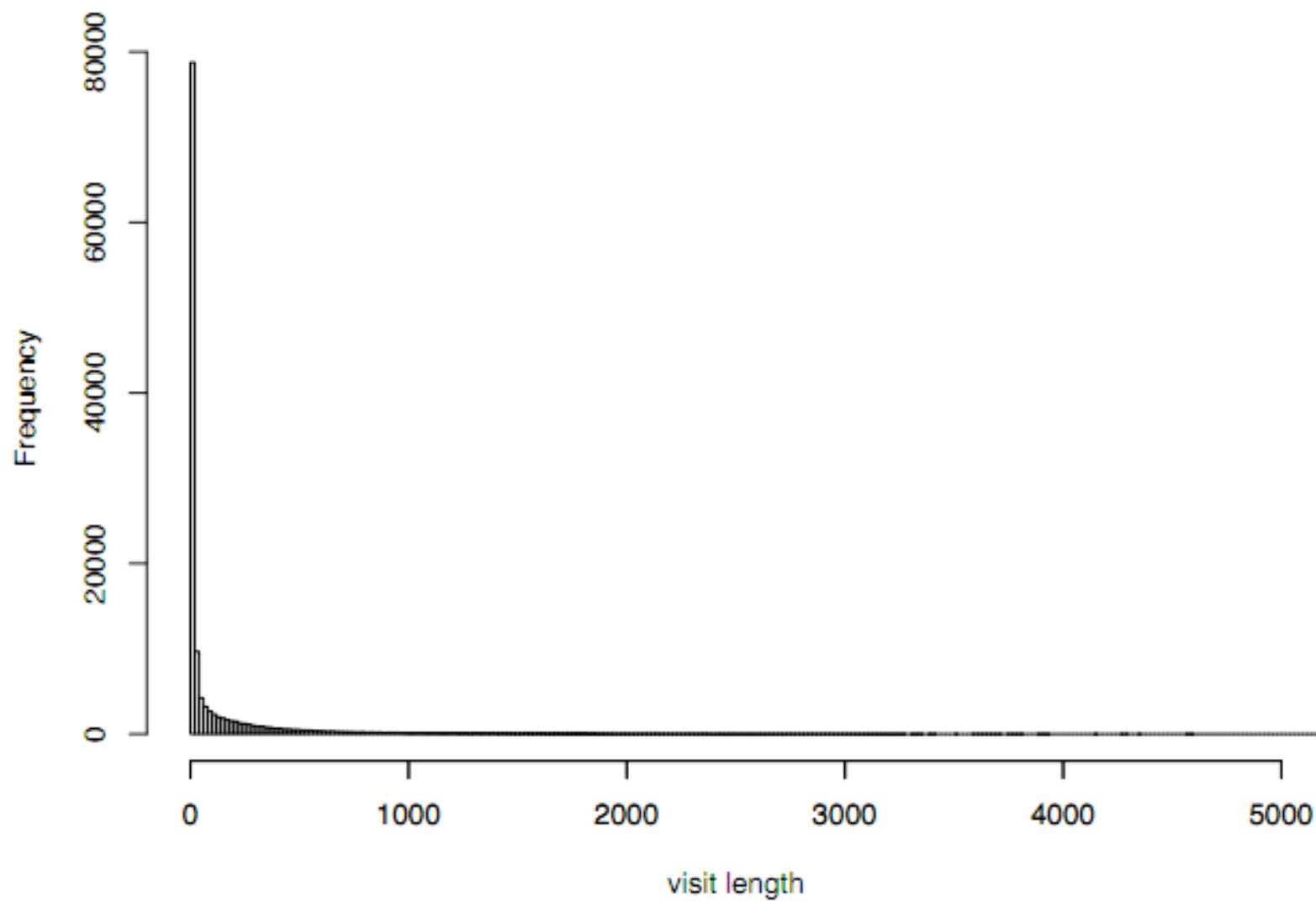
Testing is becoming a little tedious, so let's now move on to have a look at the rest of the data -- We will have one more test for you to perform in lab, but for now, let's consider some other topics

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

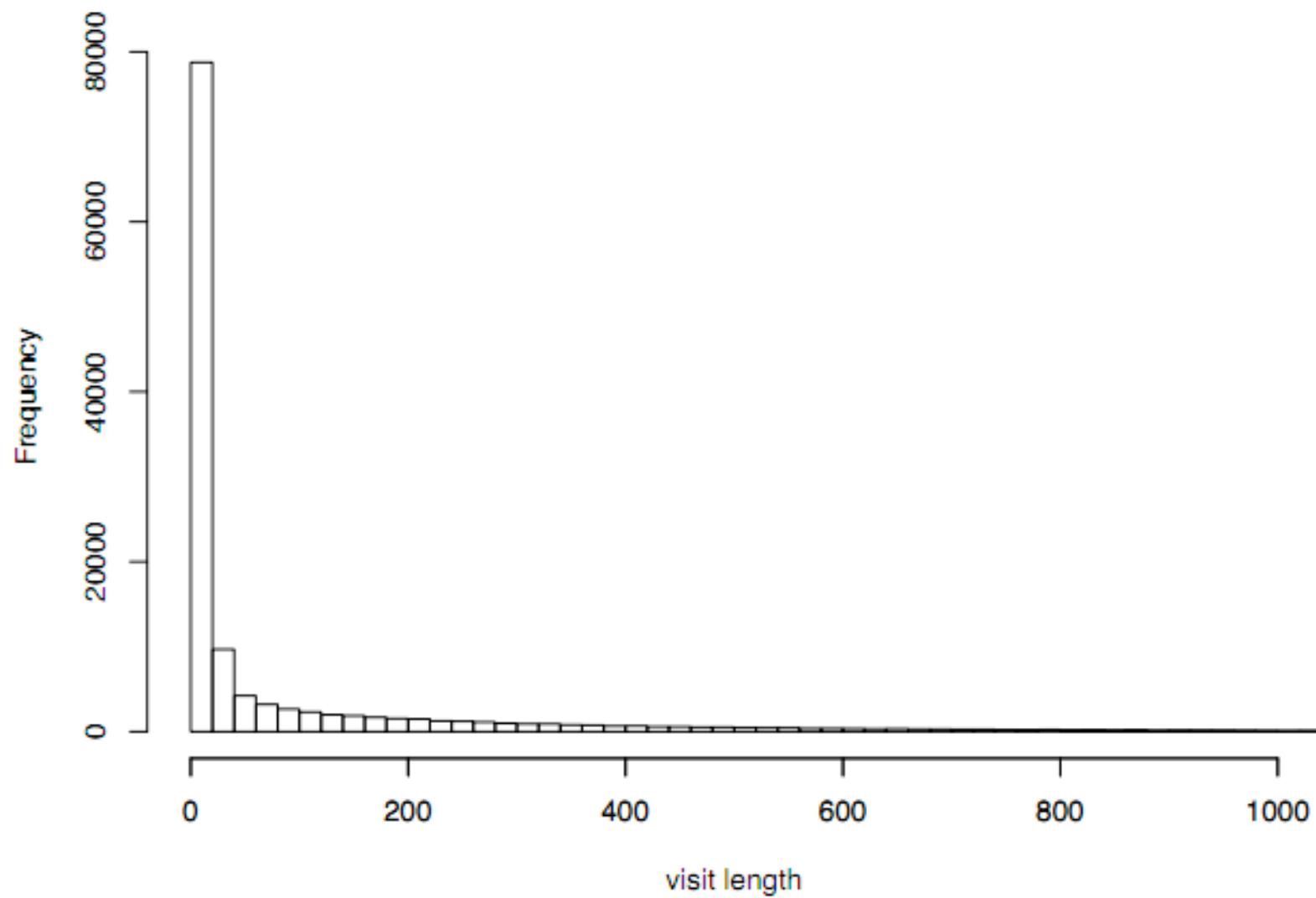
histogram of visit length



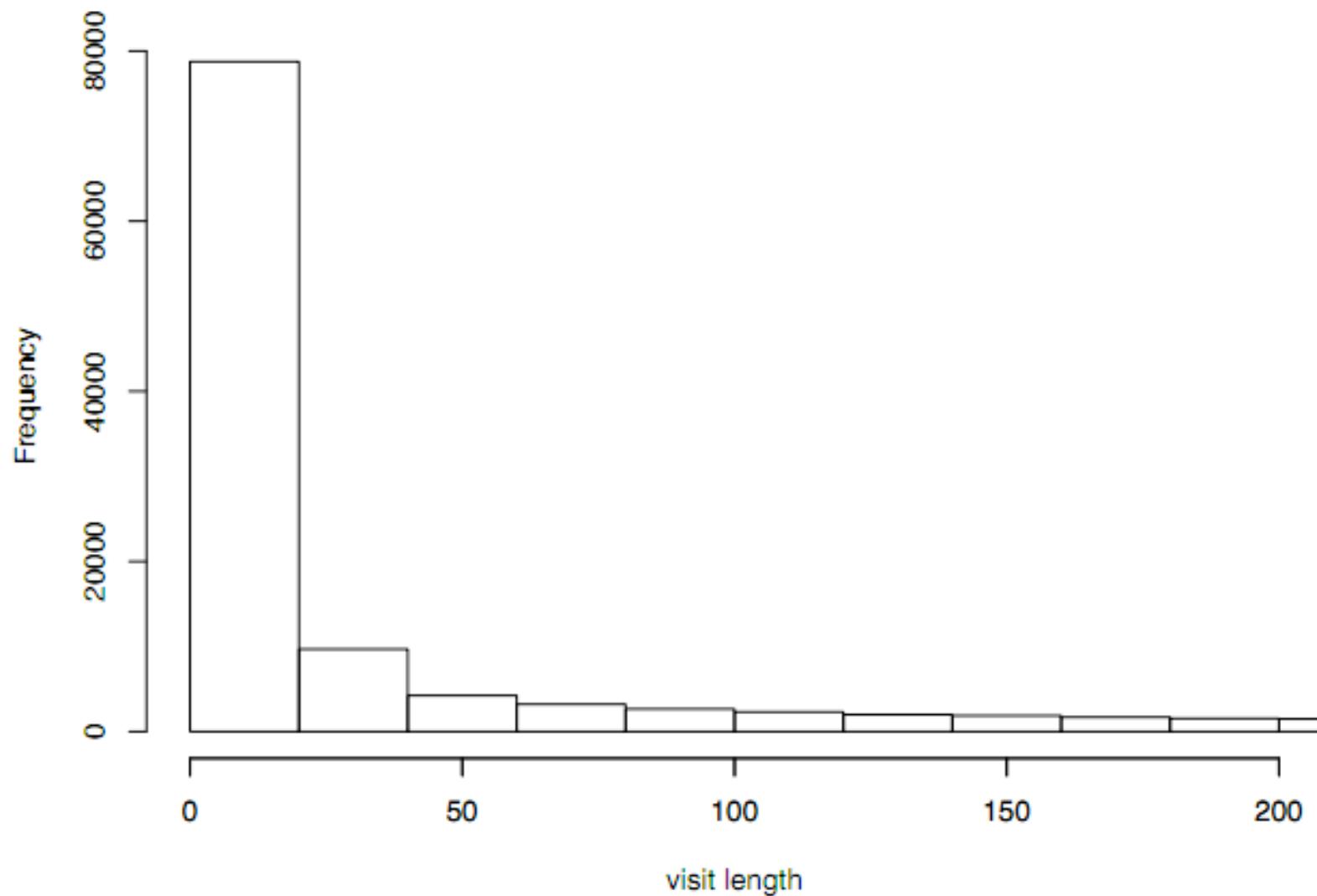
histogram of visit length, < 5000



histogram of visit length, < 1000



histogram of visit length, < 200



Transformations

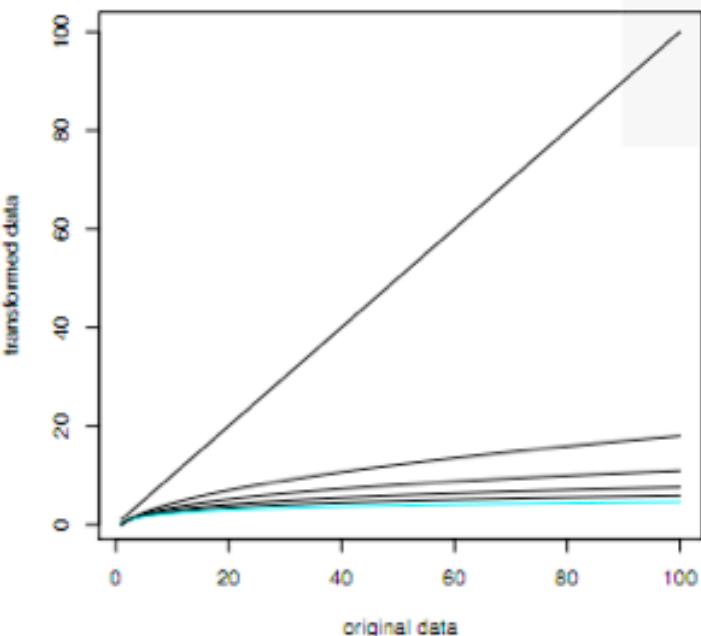
At a very practical level, transforming these data help us see more; that is, monotone transforms like square roots or the logarithm*, have a (relatively) **greater effect on large values, bringing them in closer**

With strongly skewed data, we want to consider transformations of this type simply to see what's going on...

* here log = natural log!

Transformations

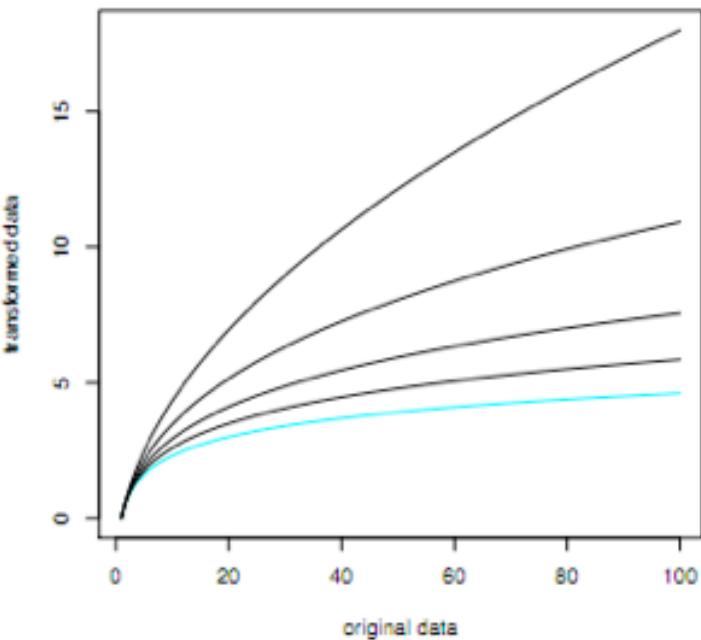
For data with a right skew, the square root, cube root, fourth root, and so on, can be used to define a family of transformations that all have the effect of taking big values and bringing them in closer to the rest of the data, to the smaller values



At the right we have a graph of this family (using the 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

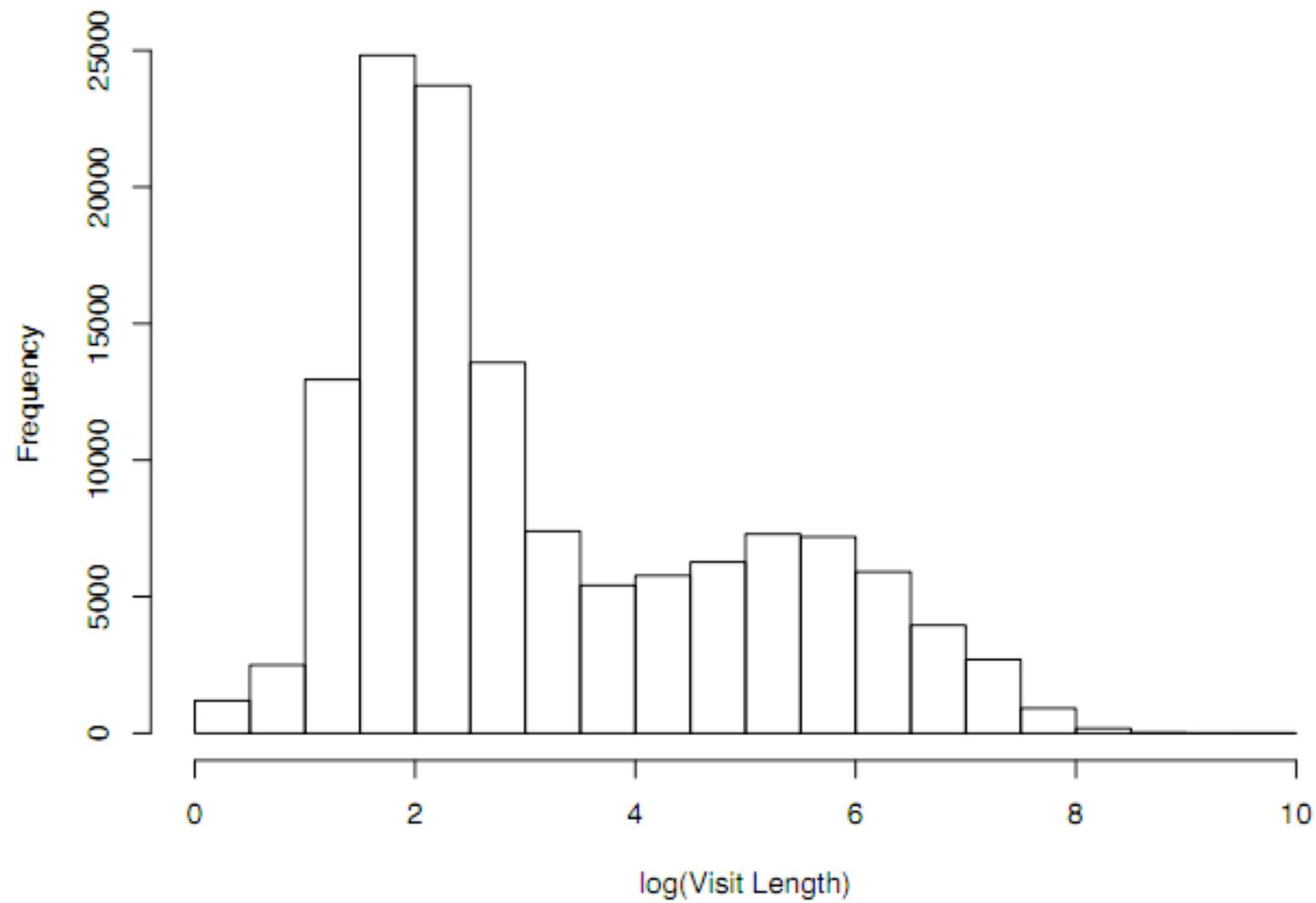
Notice that the members of this family are all monotonic transformations; that means if $x < y$ then $f(x) < f(y)$; put another way, **these transformation do not change the order of the data**

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



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 X

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.



ECOMMERCE TRACKING
Trace transactions to campaigns and keywords, get loyalty and latency metrics, and identify your revenue sources.

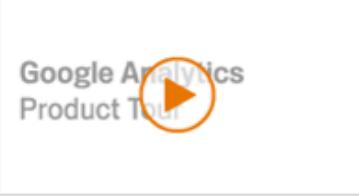


GOALS
Track sales and conversions. Measure your site engagement goals against threshold levels that you define.



MOBILE TRACKING
Track web-enabled phones, mobile websites and mobile apps.

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

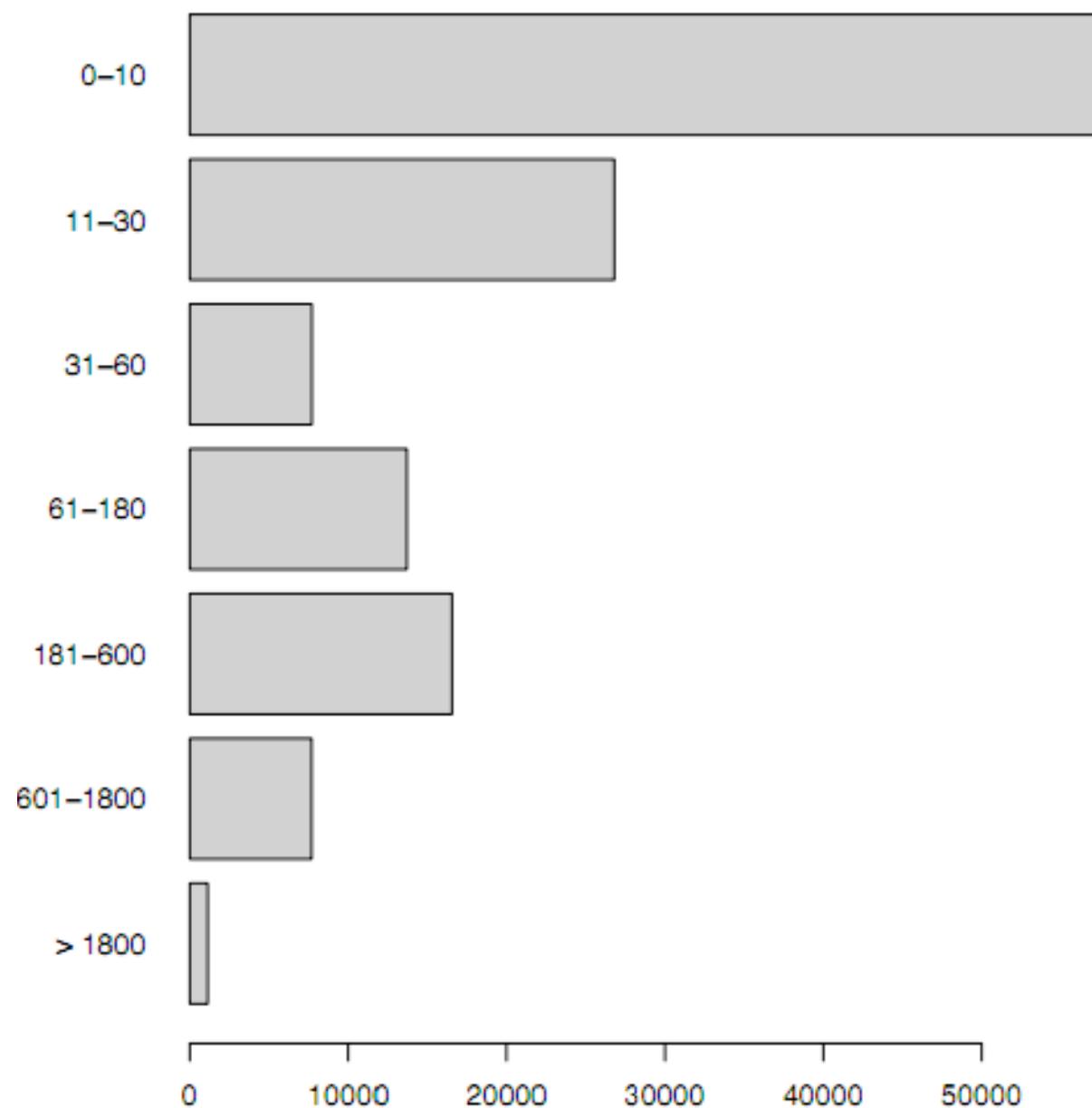
Extend the power of Google Analytics with these third party solutions in our Analytics [Application Gallery](#).



1 2 3 4

THE HUFFINGTON POST


barplot of visit lengths



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 **creating bins that get longer as you go into the right tail**, or using equally spaced bins on the transformed data

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, \dots, x_n

$$\begin{aligned}\frac{\log x_1 + \log x_2 + \dots + \log x_n}{n} &= \frac{1}{n} \log x_1 x_2 \dots x_n \\ &= \log \left[(x_1 x_2 \dots x_n)^{\frac{1}{n}} \right]\end{aligned}$$

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, \dots, x_n , a quantity that is back in the original units

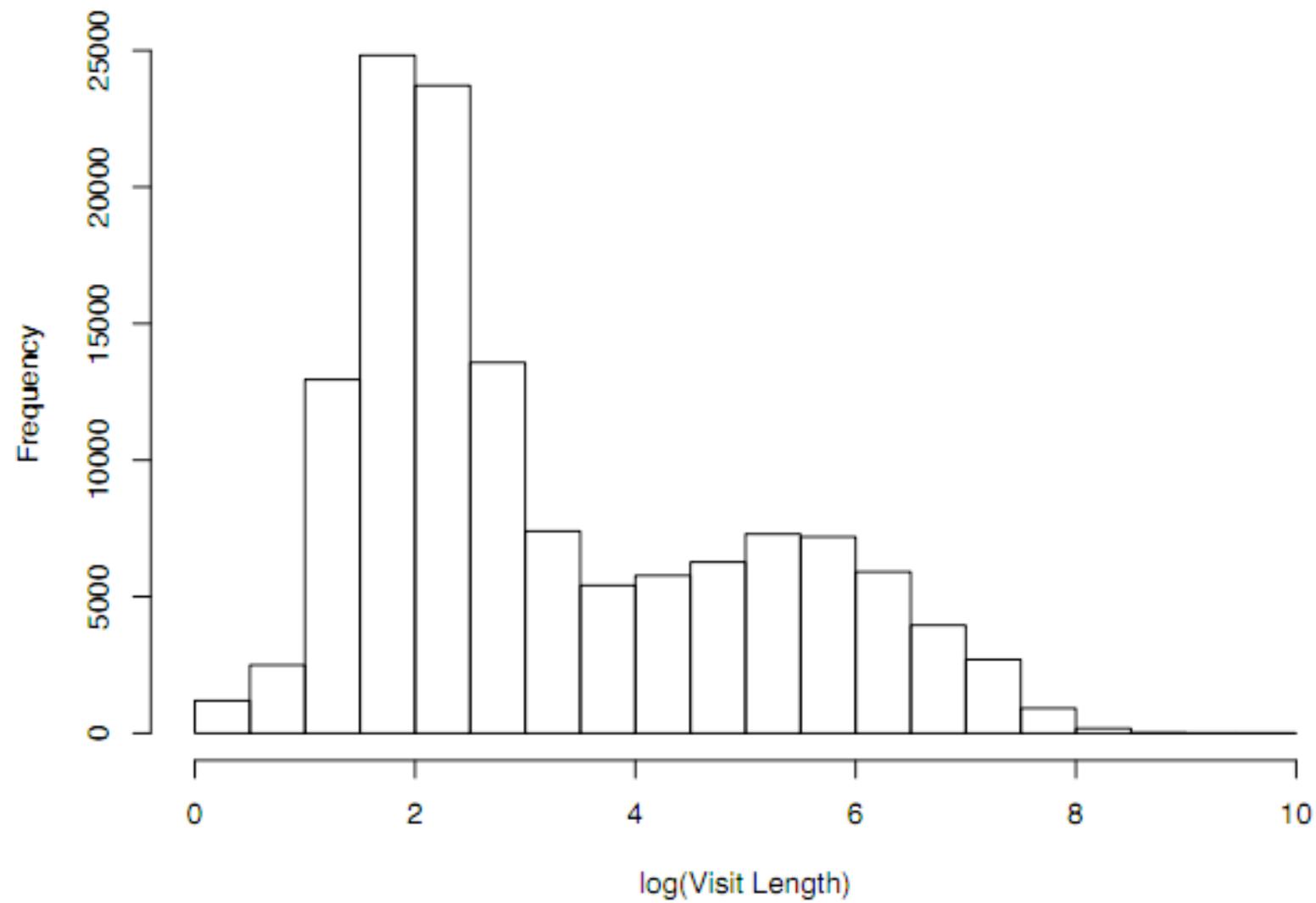
* here \log = natural \log !

Question

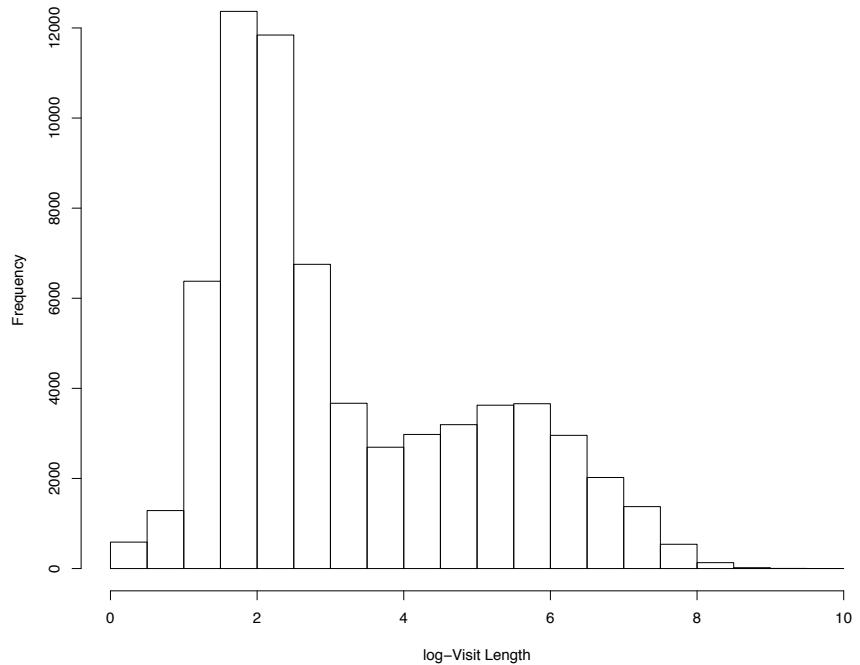
Now, back to the log-Visit Length distribution: We have seen two modes, which implies that there are two kinds of visits taking place -- when faced with a distribution like this, there is really only one question that comes to mind...

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

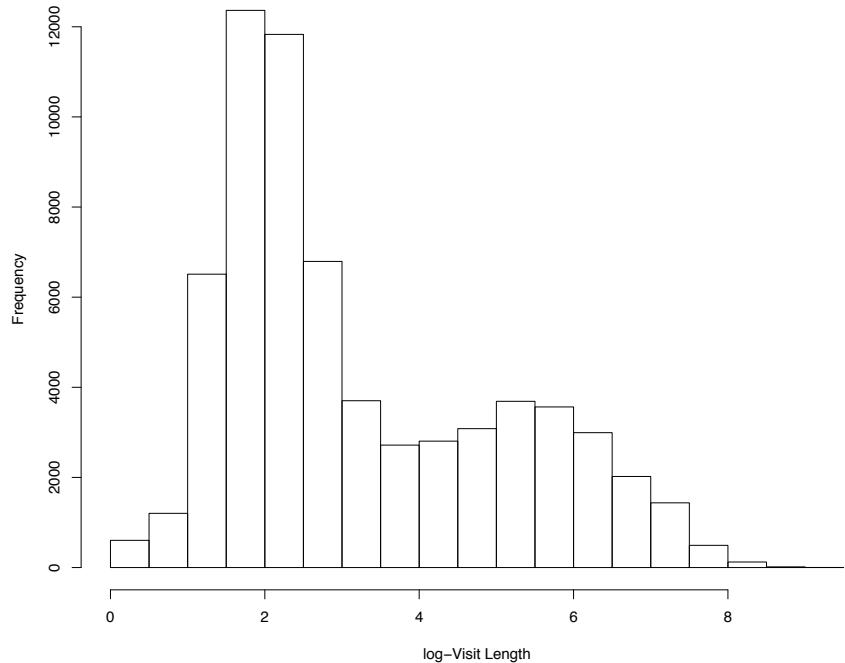
histogram of log(Visit Length)



Histogram of log-Visit Lengths for Tabs



Histogram of log-Visit Lengths for List



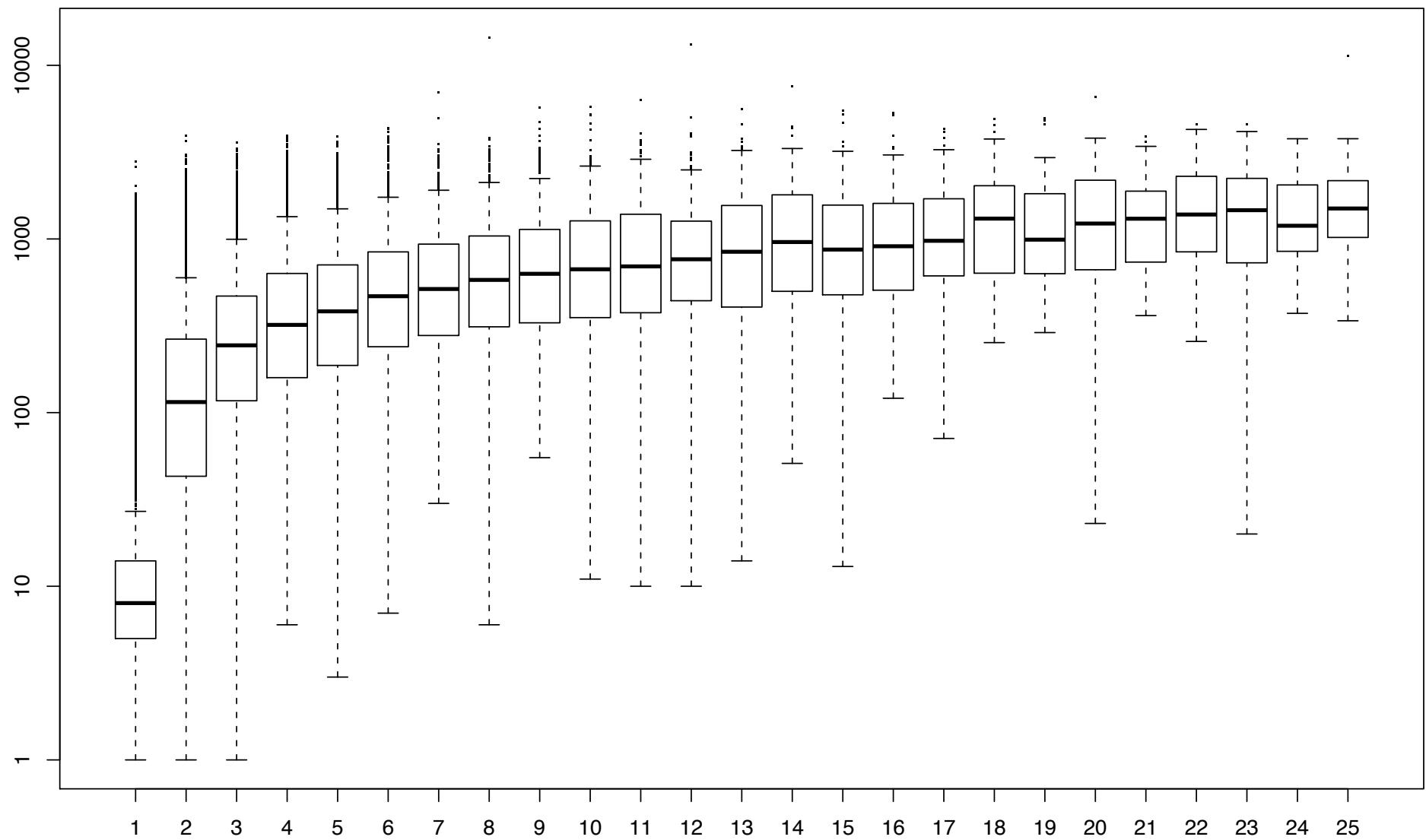
Two groups?

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

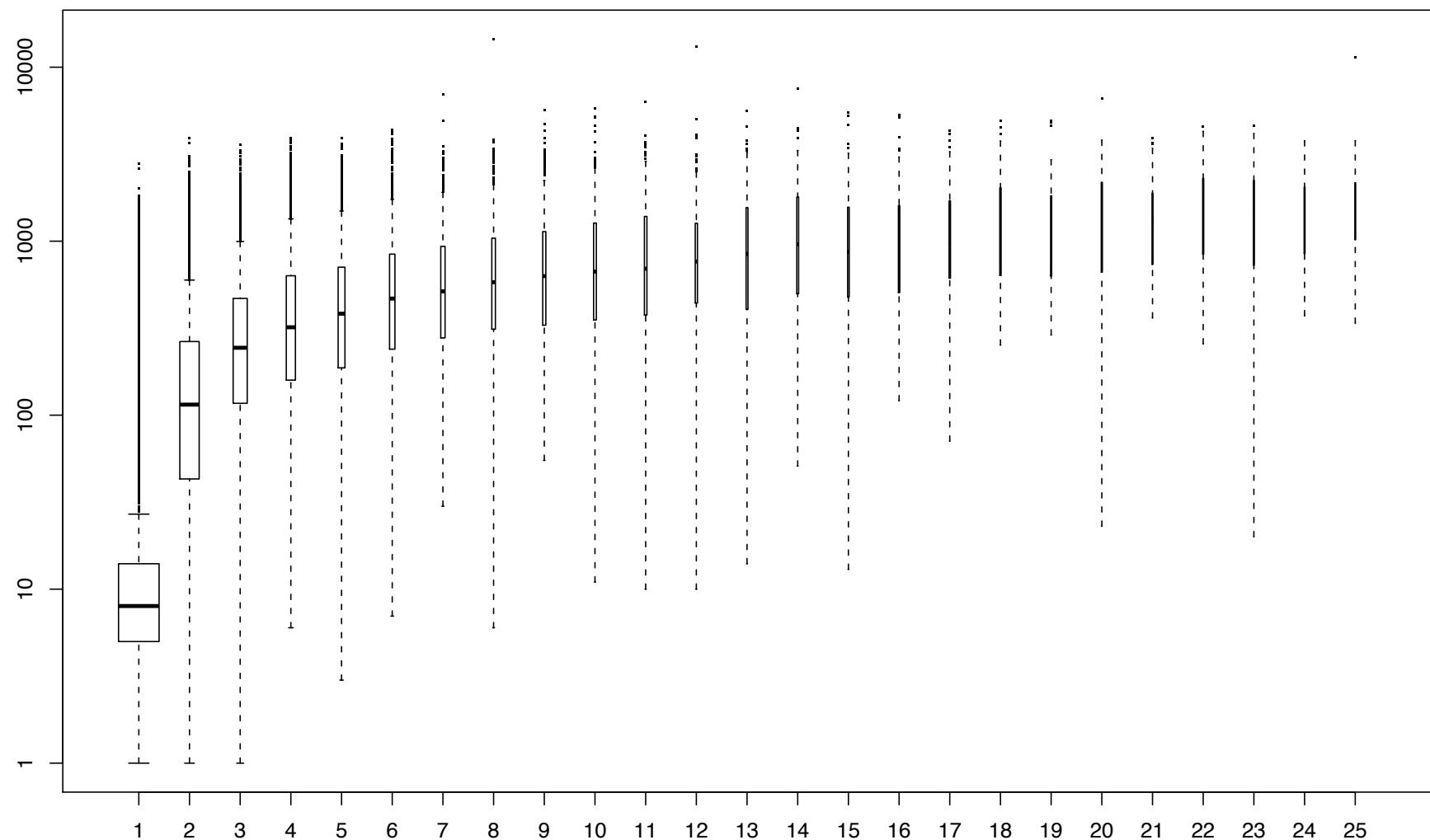
On the right we have two histograms, the top for just the log-visit lengths for those people seeing Tabs and the bottom for those seeing Lists

Any difference?

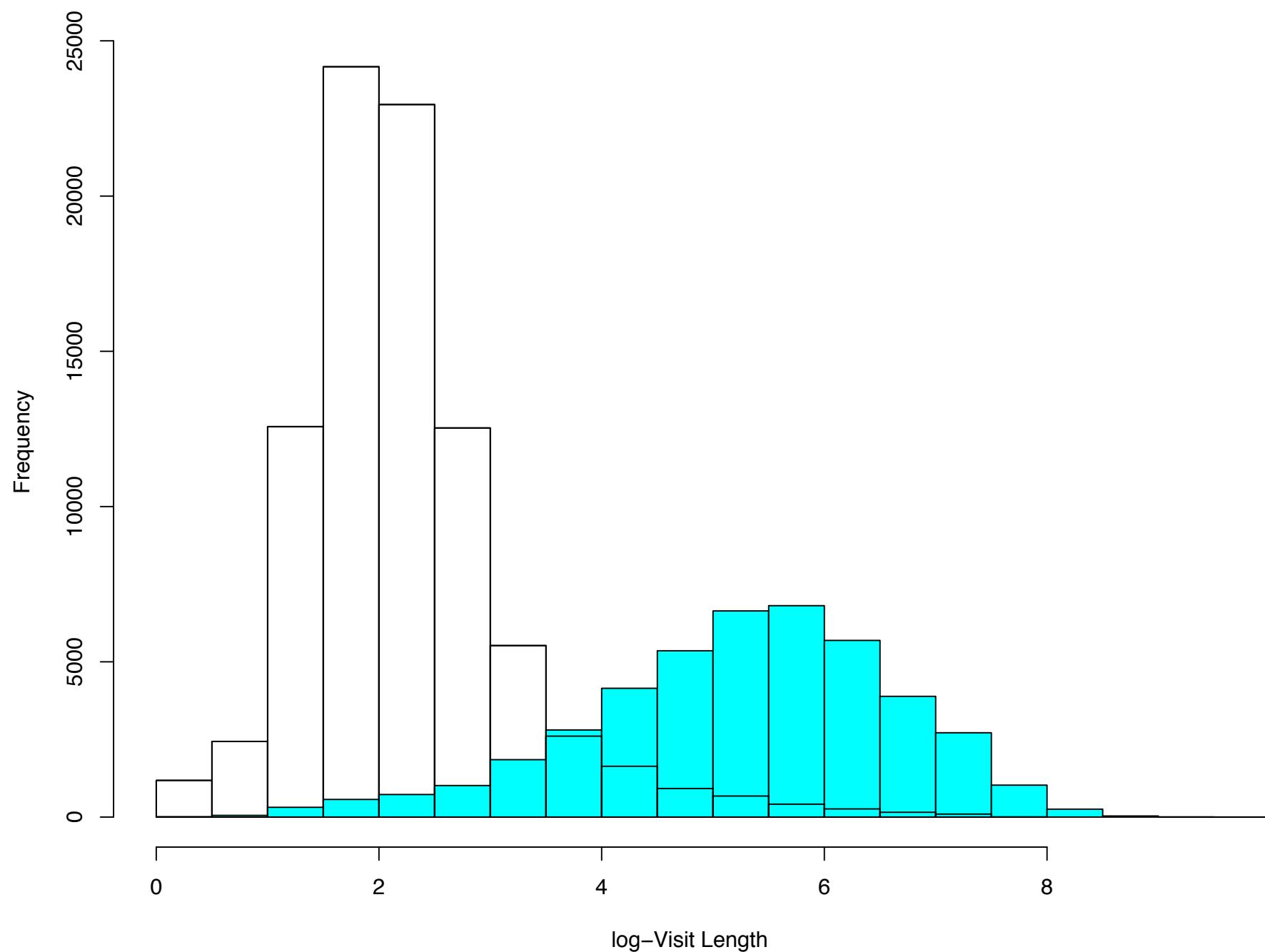
relating visit length to page views per visit (up to 25 shown)



relating visit length to page views per visit (up to 25 shown)



Two histograms of log–Visit Lengths (white for PV=1, cyan for PV>1)



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...

Random number generation

So far, we have emphasized the use of **graphics and simple simulation** (re-randomization) to analyze a data set -- We will circle back to talk about probability more formally next lecture (probably), but before that we should (probably) **put simulation on firmer footing**

For example, we seem to be trusting that R can generate “randomizations” for us, dividing or redividing patients into treatment and control -- In short, this means that **we can depend on the computer to toss coins for us**

How does it do this?

Random number generation

In Hill's trial, the samples were small enough that you cold rely on actual "random" mechanisms like **drawing tickets from a hat** (and we'll see lots of examples of the heroic work behind pre-computer simulation!)

Even **Francis Galton** (right, and we'll see a fair bit from him later too!) in the late 1880s recognized the need for statisticians to have access to simulated data (and suggested various physical mechanisms)

Besides randomized trials, toward what other purposes might we apply a sequence of random numbers?

DICE FOR STATISTICAL EXPERIMENTS.

EVERY statistician wants now and then to test the practical value of some theoretical process, it may be of smoothing, or of interpolation, or of obtaining a measure of variability, or of making some particular deduction or inference. It happened not long ago, while both a friend and myself were trying to find appropriate series for one of the above purposes, that the same week brought me letters from two eminent statisticians asking if I knew of any such series suitable for their own respective and separate needs. The assurance of a real demand for such things induced me to work out a method for supplying it, which I have already used frequently, and finding it to be perfectly effective, take this opportunity of putting it on record.

The desideratum is a set of values taken at random out of a series that is known to conform strictly to the law of frequency of error, the probable error of any single value in the series being also accurately known. We have (1) to procure such a series, and (2) to take random values out of it in an expeditious way.

Suppose the axis of the curve of distribution (whose ordinates at 100 equidistant divisions are given in my "Natural Inheritance," p. 205) to be divided into n equal parts, and that a column is erected on each of these, of a + or a - height as the case may be, equal to the height of the ordinate at the middle of each part. Then the values of these heights will form a series that is strictly conformable to the law of frequency when n is infinite, and closely conformable when n is fairly large. Moreover the probable error of any one of these values irrespectively of its sign, is 1.

As an instrument for selecting at random, I have found nothing superior to dice. It is most tedious to shuffle cards thoroughly between each successive draw, and the method of mixing and stirring up marked balls in a bag is more tedious still. A teetotum or some form of roulette is preferable to these, but dice are better than all. When they are shaken and tossed in a basket, they hurtle so variously against one another and against the ribs of the basket-work that they tumble wildly about, and their positions at the outset afford no perceptible clue to what they will be after even a single good shake and toss. The chances afforded by a die are more various than are commonly supposed; there are 24 equal possibilities, and not only 6, because each face has four edges that may be utilized, as I shall show.

I use cubes of wood $1\frac{1}{4}$ inch in the side, for the dice.
A carpenter first planes them off.

... at one time in the not too distant past, this problem was addressed in a very direct way!

A MILLION
Random Digits

WITH

100,000 Normal Deviates

RAND

TABLE OF RANDOM DIGITS

1

00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	68065	74717	34072	76850	36897	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64776	35806	34282	60935	20344	35273	88433
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	98015	45318	23374	21115	76253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96288	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12880	74697	96644	89439	28707	25815
00024	83606	48329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	63394	59593	42582	60527
00026	15474	45266	95270	78953	59367	83848	82396	10118	33211	59466
00027	94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
00028	42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
00029	23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
00030	04493	32494	75246	33824	45862	51025	61962	79335	65337	12472
00031	00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
00032	35963	15307	26898	09354	33351	35462	77974	50024	90103	39333
00033	59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
00034	46058	85236	01390	92286	77281	44077	83810	83647	70617	42941
00035	32179	00597	87379	25241	05567	07007	86743	17157	85394	11838
00036	69234	61406	20117	45204	15956	60000	18743	92423	97118	96338
00037	19565	41430	01758	75379	40419	21585	65674	36806	84962	85207
00038	45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
00039	94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
00040	98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
00041	33185	16232	41941	50949	89435	48581	88695	41994	37548	73043
00042	80951	00406	96382	70774	20151	23387	25016	25298	94624	61171
00043	79752	49140	71961	28296	69861	02591	74852	20539	00387	59579
00044	18633	32537	98145	06571	31010	24674	05455	61427	77938	91936
00045	74029	43902	77557	32270	97790	17119	52527	58021	80814	51748
00046	54178	45611	80993	37143	05335	12969	56127	19255	36040	90324
00047	11664	49883	52079	84827	59381	71539	09973	33440	88461	23356
00048	48324	77928	31249	64710	02295	36870	32307	57546	15020	09994
00049	69074	94138	87637	91976	35584	06401	10518	21615	01848	76938

00050	09188	20097	32825	39527	04220	86304	83389	87374	64278	58044
00051	90045	85497	51981	50654	94938	81997	91870	76150	68476	84659
00052	73189	50207	47677	26289	62290	64464	27124	67018	41361	82780
00053	75768	76490	20971	87749	90429	12272	95375	05871	93823	43178
00054	54016	44056	66281	31003	00682	27398	20714	53295	07706	17813
00055	08358	69910	78542	42785	13661	58873	04618	97553	31223	08420
00056	28306	03264	81333	10591	40510	07893	32604	60475	94119	01840
00057	53840	86233	81594	13628	51215	90290	28466	68795	77762	20791
00058	91757	53741	61613	62269	50263	90212	55781	76514	83483	47055
00059	89415	92694	00397	58391	12607	17646	48949	72306	94541	37408
00060	77513	03820	86864	29901	68414	82774	51908	13980	72893	55507
00061	19502	37174	69979	20286	55210	29773	74287	75251	65344	67415
00062	21818	58313	93278	81757	65686	73156	07082	85046	31853	38452
00063	51474	66499	68107	23621	94049	91345	42836	09191	08007	43449
00064	99659	68331	62535	24170	69777	12830	74819	78142	43860	72834
00065	33713	48007	93584	72869	51926	64721	58303	29822	93174	93972
00066	85274	86893	11303	22970	28834	34137	73515	90400	71148	43643
00067	84133	89640	44035	52166	73852	70091	61222	60561	62327	18423
00068	56732	16234	17395	96131	10123	91622	85496	57560	81604	18880
00069	65138	56806	87648	85261	34313	65861	45875	21069	85644	47277
00070	38001	02176	81719	11711	71602	92937	74219	64049	63384	49698
00071	37402	96397	01304	77586	56271	10086	47324	62605	40030	37438
00072	97125	40348	87083	31417	21815	39250	75237	62047	15501	29578
00073	21826	41134	47143	34072	64638	85902	49139	06441	03856	54552
00074	73135	42742	95719	09035	85794	74296	06789	88156	64691	19202
00075	07638	77929	03061	18072	96207	44156	23821	99538	04713	66994
00076	60528	83441	07954	19814	59175	20695	05533	52139	61212	06455
00077	83596	35635	06958	92983	05128	09719	77433	53783	92301	50498
00078	10850	62746	99599	10507	13499	06319	53075	71839	06410	19362
00079	39820	98932	43622	63147	64421	80814	43800	09351	31024	73167
00080	59580	06478	75569	78890	88835	54486	23768	06156	04111	08408
00081	38508	07341	23793	48763	90822	97022	17719	04207	95954	49953
00082	30692	70668	94688	16127	56196	80091	82067	63400	05462	69200
00083	65443	95659	18288	27437	49632	24041	08337	65676	96299	90836
00084	27267	50264	13192	72294	07477	44806	17985	48911	97341	30358
00085	91307	06991	19072	24210	36699	53728	28825	35793	28976	66252
00086	68434	94688	84473	13622	62126	98408	12843	82590	09815	93146
00087	48908	15877	54745	24591	35700	04754	83824	52692	54130	55160
00088	06913	45197	42672	78861	11883	09628	63011	98901	14974	40344
00089	10455	16019	14210	33712	91342	37821	88325	80851	43667	70883
00090	12883	97343	65027	61184	04285	01392	17974	15077	90712	26769
00091	21778	30976	38807	36961	31649	42096	63281	02023	68816	47449
00092	18523	58515	65122	59659	86283	68258	69572	13798	16435	91529
00093	67245	52670	35583	16563	79246	86686	76463	34222	26655	90802
00094	60584	47377	07500	37992	45134	26529	26760	83637	41326	44344
00095	53853	41377	36066	94650	58838	73859	49364	73331	96240	43642
00096	24637	38736	74384	88342	52623	07992	12369	18601	03742	83873
00097	83080	12451	38992	22815	07759	51777	97377	27585	51972	37887
00098	16444	24334	36151	99073	27493	70939	85130	32552	54846	54759
00099	60790	18157	57178	65762	11161	78576	45819	52979	65130	04860

TABLE OF RANDOM DIGITS

3

00100	03991	10461	93716	16894	66083	24653	84609	58232	88618	19161
00101	38555	95554	32886	59780	08355	60860	29735	47762	71299	23853
00102	17546	73704	92052	46215	55121	29281	59076	07936	27954	58909
00103	32643	52861	95819	06831	00911	98936	76355	93779	80863	00514
00104	69572	68777	39510	35903	14060	40619	29549	69616	33564	60780
00105	24122	66591	27699	06494	14845	46872	61958	77100	90899	75754
00106	61196	30231	92962	61773	41839	55382	17267	70943	78038	70267
00107	30532	21704	10274	12202	39685	23309	10061	68829	53986	66485
00108	03788	97599	75867	20717	74416	53166	35208	33374	87539	08823
00109	48228	63379	85783	47619	53152	67433	35663	52972	16818	60311
00110	60365	94653	35075	33949	42614	29297	01918	28316	98953	73231
00111	83799	42402	56623	34442	34894	41374	70071	14736	09858	18065
00112	32960	07405	36409	83232	99385	41600	11133	07586	15917	06253
00113	19322	53845	57620	52606	66497	68646	78138	66559	19640	99413
00114	11220	94747	07399	37408	48509	23929	27482	45476	85244	35159
00115	31751	57260	68980	05339	15470	46355	88651	22596	03152	19121
00116	88492	99382	14454	04504	20094	98977	74843	93413	22109	78508
00117	30934	47744	07481	83828	73788	06533	28597	20405	94205	20380
00118	22888	48893	27499	98748	60530	45128	74022	84617	82037	10268
00119	78212	16993	35902	91386	44372	15486	65741	14014	87481	37220
00120	41849	84547	46850	52326	34677	58300	74910	64345	19325	81549
00121	46352	33049	69248	93460	45305	07521	61318	31855	14413	70951
00122	11087	96294	14013	31792	59747	67277	76503	34513	39663	77544
00123	52701	05837	56303	87315	16520	69676	11654	99893	02181	68161
00124	57275	36898	81304	48585	68652	27376	92852	55866	88448	03584
00125	20857	73156	70284	24328	79375	95220	01159	63267	10622	48391
00126	15633	84924	90415	93614	33521	26665	55823	47641	86225	31704
00127	92694	48297	39804	02115	59589	49067	66821	41575	49767	04037
00128	77613	19019	88152	00080	20554	91409	96277	48257	50816	97616
00129	38688	32486	45134	63545	59404	72059	43947	51680	43852	59693
00130	25163	01889	70014	15021	41290	67312	71657	15957	68971	11403
00131	65251	07629	37239	33295	05870	01119	92784	26340	18477	65622
00132	36815	43625	18637	37509	82444	99005	04921	73701	14707	93997
00133	64397	11692	05327	82162	20247	81759	45197	25332	83745	22567
00134	04515	25624	95096	67946	48460	85558	15191	18782	16930	33361
00135	83761	60873	43253	84145	60833	25983	01291	41349	20368	07126
00136	14387	05345	80554	09279	43529	06318	38384	74761	41196	37480
00137	51321	92246	80088	77074	88722	56736	66164	49431	66919	31678
00138	72472	00008	80890	18002	94813	31900	54155	83436	35352	54131
00139	05466	55306	93128	18464	74457	90561	72848	11834	79982	68416
00140	39528	72484	82474	25593	48545	35247	18619	13674	18611	19241
00141	81616	18711	53342	44276	75122	11724	74627	73707	58319	15997
00142	07586	16120	82641	22820	92904	13141	32392	19763	61199	67940
00143	90767	04235	13574	17200	69902	63742	78464	22501	18627	90872
00144	40188	28193	29593	88827	94972	11598	62095	36787	00441	58997
00145	34414	82157	86887	55087	19152	00023	12302	80783	32624	68691
00146	63439	75363	44989	16822	36024	00867	76378	41605	65961	73488
00147	67049	09070	93399	45547	94458	74284	05041	49807	20288	34060
00148	79495	04146	52162	90286	54158	34243	46978	35483	59362	95938
00149	91704	30552	04737	21031	75051	93029	47665	64382	99782	93478

[Click to LOOK INSIDE!](#)

A MILLION Random Digits

WITH
100,000 Normal Deviates

RAND

[Click to LOOK INSIDE!](#)

A Small Book Of Random Numbers

Volume One

James McNalley

A MILLION **Random Digits** THE SEQUEL

with
Perfectly Uniform Distribution



Classical Computing

David Dubowski

kindle edition



Random number generation

These days, there are two dominant techniques for generating random numbers

One is not really random according to any romantic notions of the word and are the result of **a mathematical formula** which is entirely predictable and repeatable -- These are often called **pseudo-random numbers**

The second, on the other hand, is often touted as “**true” random numbers** and are generated by **observing some physical process** -- You can think of a small coin-tossing device attached to your computer although the physical phenomena used tend to be more exotic

Let's have a look at both, starting with the latter as it will give us to talk about data and the publication of data (a theme for this course)

HotBits: Genuine Random Numbers

www.fourmilab.ch/hotbits/



HotBits: Genuine random numbers, generated by radioactive decay

Click on the icon to turn off the sound effects. If your browser doesn't do Java, you won't hear the sound effects anyway. If you like, you can [download source code](#) for the applet.

People working with computers often sloppily talk about their system's "random number generator" and the "random numbers" it produces. But numbers calculated by a computer through a deterministic process, cannot, by definition, be random. Given knowledge of the algorithm used to create the numbers and its internal state, you can predict all the numbers returned by subsequent calls to the algorithm, whereas with genuinely random numbers, knowledge of one number or an arbitrarily long sequence of numbers is of no use whatsoever in predicting the next number to be generated.

Computer-generated "random" numbers are more properly referred to as *pseudorandom numbers*, and *pseudorandom sequences* of such numbers. A variety of clever algorithms have been developed which generate sequences of numbers which pass every statistical test used to distinguish random sequences from those containing some pattern or internal order. A [test program](#) is available at this site which applies such tests to sequences of bytes and reports how random they appear to be, and if you run this program on data generated by a high-quality pseudorandom sequence generator, you'll find it generates data that are indistinguishable from a sequence of bytes chosen at random. Indistinguishable, but not genuinely random.

HotBits is an Internet resource that brings *genuine* random numbers, generated by a process fundamentally governed by the inherent uncertainty in the quantum mechanical laws of nature, directly to your computer in a variety of forms. *HotBits* are generated by timing successive pairs of radioactive decays detected by a Geiger-Müller tube interfaced to a computer. You order up your serving of HotBits by [filling out a request form](#) specifying how many random bytes you want and in which format you'd like them delivered. Your request is relayed to the HotBits server, which flashes the random bytes back to you over the Web. Since the [HotBits generation hardware](#) produces data at a modest rate (about 100 bytes per second), requests are filled from an "inventory" of pre-built HotBits. Once the random bytes are delivered to you, they are immediately discarded—the same data will never be sent to any other user and no records are kept of the data at this or any other site.

How HotBits Works

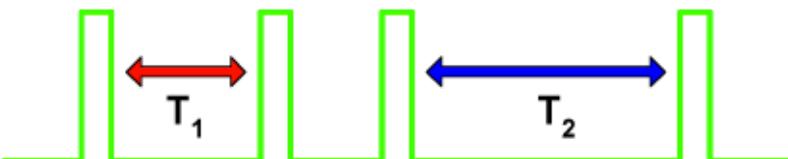
www.fourmilab.ch/hotbits/how3.html

Bit from It

This inherent randomness in decay time has profound implications, which we will now exploit to generate random numbers—HotBits. For if there's no way to know when a given Cæsium-137 nucleus will decay then, given an collection of them, there's no way to know when the *next* one of them will shoot its electron bolt and settle down to a serene eternity as Barium. That's uncertainty, with its origins in the deepest and darkest corners of creation—precisely what we're looking for to make genuinely random numbers.

If we knew the precise half-life of the radioactive source driving our detector (and other details such as the solid angle to which our detector is sensitive, the energy range of decay products and the sensitivity of the detector to them, and so on), we could generate random bits by measuring whether the time between a pair of beta decays was more or less than the time expected based on the half-life. But that would require our knowing the average beta decay detection time, which depends on a large number of parameters which can only be determined experimentally. Instead, we can exploit the inherent uncertainty of decay time in a parameter-free fashion which requires less arm waving and fancy footwork.

The trick I use was dreamed up in a conversation in 1985 with John Nagle, who is doing some fascinating things these days with [artificial animals](#). Since the time of any given decay is random, then the *interval* between two consecutive decays is also random. What we do, then, is measure a pair of these intervals, and emit a zero or one bit based on the relative length of the two intervals. If we measure the same interval for the two decays, we discard the measurement and try again, to avoid the risk of inducing bias due to the resolution of our clock.



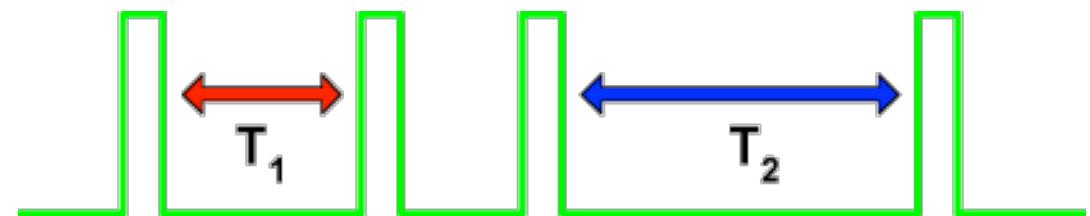
To create each random bit, we wait until the first count occurs, then measure the time, T_1 , until the next. We then wait for a second pair of pulses and measure the interval T_2 between them, yielding a pair of durations. If they're the same, we throw away the measurement and try again. Otherwise if T_1 is less than T_2 we emit a zero bit; if T_1 is greater than T_2 , a one bit. In practice, to avoid any residual bias resulting from non-

Bits

A “bit” stands for a “binary digit” and takes on the value 0 or 1 -- You can think of it as a coin toss where we map “heads” to 1, say, and “tails” to 0 (The term “bit” was actually coined by our man John Tukey; he also came up with the term “software”)

HotBits uses radioactive decay as a means for generating physically random or “true” random bits (coin tosses) -- You can imagine listening to a Geiger counter and group the ticks into pairs

If $T_1 > T_2$ they produce a 1, otherwise they generate a 0 -- Voila! “true” random numbers!



Bits

Another project, this run through the web site `random.org` uses atmospheric noise, suitably filtered, to accomplish the same task

Our interest in this site was because of the mechanism they used to “publish” their random bits, **offering a service that scientists around the world** -- They have implemented a **public API (application programming interface)** that let’s programs (like R) request data (random bits)

RANDOM.ORG – True Random

www.random.org

Home Games Numbers Lists & More Drawings Web Tools Statistics Testimonials Learn More Login

RANDOM.ORG

True Random Number Service

What's this fuss about *true* randomness?

Perhaps you have wondered how predictable machines like computers can generate randomness. In reality, most random numbers used in computer programs are *pseudo-random*, which means they are generated in a predictable fashion using a mathematical formula. This is fine for many purposes, but it may not be random in the way you expect if you're used to dice rolls and lottery drawings.

RANDOM.ORG offers *true* random numbers to anyone on the Internet. The randomness comes from atmospheric noise, which for many purposes is better than the pseudo-random number algorithms typically used in computer programs. People use RANDOM.ORG for holding drawings, lotteries and sweepstakes, to drive games and gambling sites, for scientific applications and for art and music. The service has existed since 1998 and was built and is being operated by [Mads Haahr](#) of the [School of Computer Science and Statistics at Trinity College, Dublin](#) in Ireland.

As of today, RANDOM.ORG has generated 958.5 billion random bits for the Internet community.

FREE services Games and Gambling

Lottery Quick Pick is perhaps the Internet's most popular with over 130 lotteries
Keno Quick Pick for the popular game played at many casinos
Coin Flipper will give you heads or tails in many currencies
Dice Roller does exactly what it says on the tin
Playing Card Shuffler will draw cards from multiple shuffled decks
Birdie Fund Generator will create birdie holes for golf courses

PAID service Random Drawings

Q3.1 in the [FAQ](#) explains how to pick a winner for your giveaway for FREE
Third-Party Draw Service is the premier solution to holding random drawings online
Step by Step Guide explains how to hold a drawing with the Third-Party Draw Service
Step by Step Video shows how to hold a drawing with the Third-Party Draw Service
Price Calculator tells exactly how much your drawing will cost
Drawing FAQ answers common questions about holding drawings

True Random Number Generator

Min: 1 Max: 100

Generate Result:

Powered by RANDOM.ORG

Like RANDOM.ORG?
[Get the Newsletter](#)

Another service

With the advent of Web 2.0, the dominant method of data distribution has changed **from simply serving up web pages** and, along with it, we have experienced a massive shift in our view of the web itself -- In 2005, the term Web 2.0 emerged to represent this new view, one of **“collective intelligence”** of **“data sharing”** and **“web services”**

To deliver on this version of the web as a system of **cooperating data services**, we need techniques to specify the kind of data we want and specify the format we expect to see it in -- For `random.org`, it's pretty easy...

RANDOM.ORG – HTTP Interface

www.random.org/clients/http/

Home Games Numbers Lists & More Drawings Web Tools Statistics Testimonials Learn More Login

RANDOM.ORG

True Random Number Service

HTTP Interface Description

RANDOM.ORG is a true random number service that generates randomness via atmospheric noise. This page explains how to interface to the service via the Hyper-Text Transfer Protocol (HTTP). There is also the [HTTP Client Archive](#), which contains clients that other people have written.

Important note!

If you access RANDOM.ORG via an automated client, please make sure you observe the [Guidelines for Automated Clients](#) or your computer may be banned.

If you are writing a general-purpose client, please make sure it is easy for your users to run it in accordance with the guidelines.

This page contains documentation for the [Integer Generator](#), the [Sequence Generator](#), the [String Generator](#) and the [Quota Checker](#), which allows you to examine your current bit allowance.

All the interfaces on this page return HTTP status code 503 (Service Unavailable) in the case of errors and code 200 (OK) when successful. Not all languages allow you to access the HTTP status codes in a straightforward manner. A reasonable workaround is to look for the string "Error:" (don't forget the colon) as the first line of the response. This will work for all the generators on this page, including the [String Generator](#) (which could by chance produce the string "Error" in a successful response, but which cannot produce the colon character).

Please note that the old CGI scripts (randbyte, randnum, etc.) are no longer supported and you should use the ones described below instead. In particular, the old scripts do not return the 503 status code in case of errors (they return the 200 response code in all cases), so please use the new ones instead.

Integer Generator

The [Integer Generator](#) will generate truly random integers in configurable intervals. It is pretty easy to write a client to access the integer generator. The integer generator accepts only HTTP GET requests, so parameters are passed via encoding in the URL.

Parameters

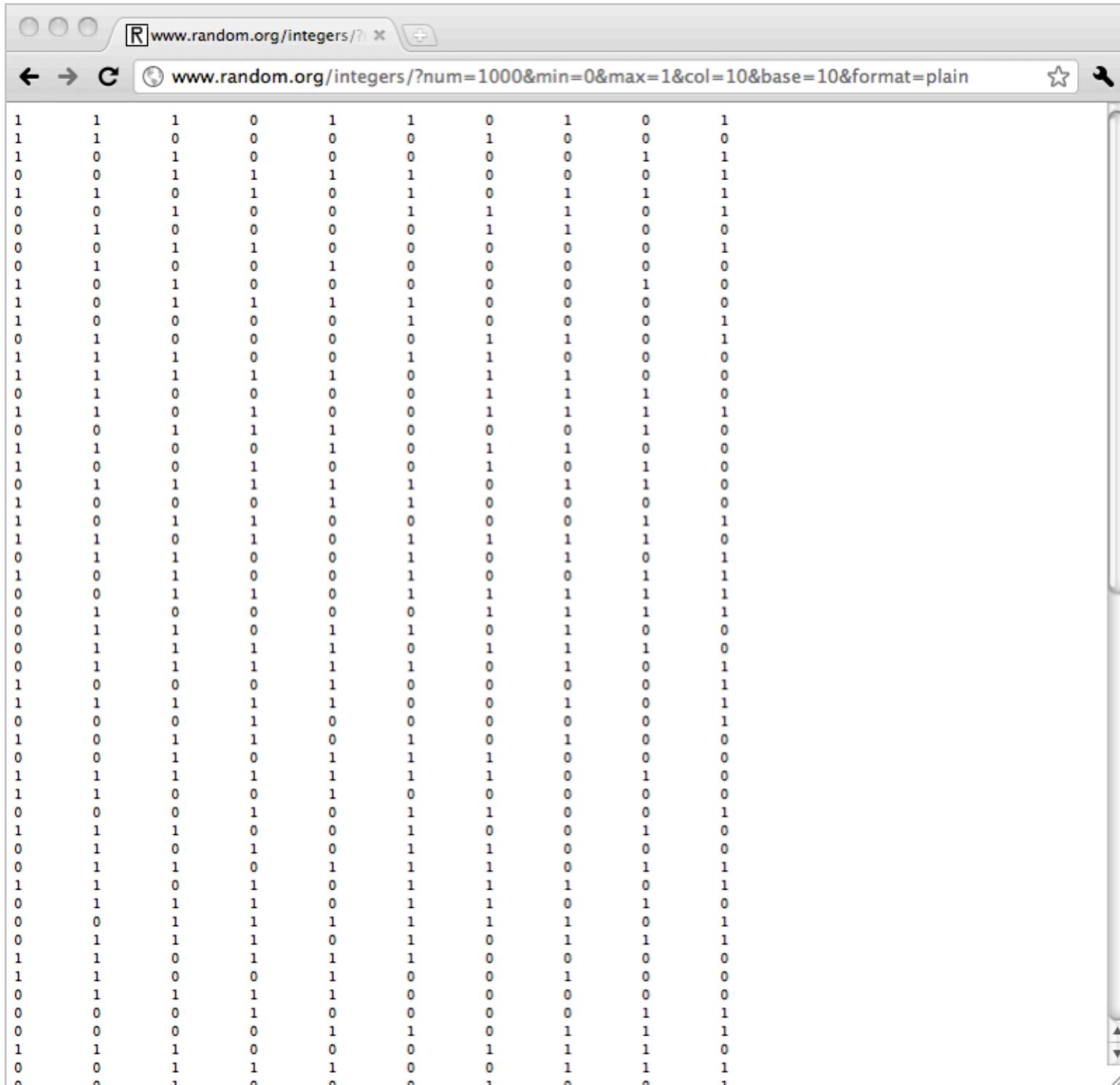
Name	Possible Values	Description
------	-----------------	-------------

Integer Generator

The Integer Generator will generate truly random integers in configurable intervals. It is pretty easy to write a client to access the integer generator. The integer generator accepts only HTTP GET requests, so parameters are passed via encoding in the URL.

Parameters

Name	Possible Values	Description
num	[1,1e4]	The number of integers requested.
min	[-1e9,1e9]	The smallest value allowed for each integer.
max	[-1e9,1e9]	The largest value allowed for each integer.
col	[1,1e9]	The number of columns in which the integers will be arranged. The integers should be read (or processed) left to right across columns.
base	2 8 10 16	The base that will be used to print the numbers, i.e., binary, octal, decimal or hexadecimal.
format	html plain	Determines the return type of the document that the server produces as its response. If html is specified, the server produces a nicely formatted XHTML document (MIME type <code>text/html</code>), which will display well in a browser but which is somewhat cumbersome to parse. If plain is specified, the server produces a minimalistic document of type plain text (MIME type <code>text/plain</code>) document, which is easy to parse. If you are writing an automated client, you probably want to specify plain here.
rnd	new id.identifier date.iso-date	Determines the randomization to use to generate the numbers. If new is specified, then a new randomization will be created from the truly random bitstream at RANDOM.ORG. This is probably what you want in most cases. If id.identifier is specified, the identifier is used to determine the randomization in a deterministic fashion from a large pool of pregenerated random bits. Because the numbers are produced in a deterministic fashion, specifying an id basically uses RANDOM.ORG as a pseudo-random number generator. The third (date.iso-date) form is similar to the second; it allows the randomization to be based on one of the daily pregenerated files. This form must refer to one of the dates for which files exist, so it must be the current day (according to UTC) or a day in the past. The date must be in ISO 8601 format (i.e., <code>YYYY-MM-DD</code>) or one of the two shorthand strings today or yesterday .



APIs

I mention this because many organizations offer their data via web services like this -- Specific HTTP requests yield not web pages designed for human consumption (reading), but instead **structured data that can be processed easily in an autonomous way by computers**

In this way, services build “mash-up” style, with data flowing easily between different organizations -- This is really an amazing development that opens up incredible possibilities for statistics, for computer science, for machine learning for the data geeks out there!

Graph API - Facebook Developers

developers.facebook.com/docs/reference/api/

facebook DEVELOPERS Documentation Forum Showcase Blog My Apps Search for documentation

Getting Started

Core Concepts >

- Social Plugins
- Graph API
- Social Channels
- Authentication
- Open Graph protocol

Advanced Topics

SDKs & Tools

Objects

- Album
- Application
- Checkin
- Comment
- Event
- FriendList
- Group
- Insights
- Link
- Message
- Note
- Page
- Photo

Graph API

Core Concepts > Graph API

At Facebook's core is the social graph; people and the connections they have to everything they care about. The Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags).

Every object in the social graph has a unique ID. You can access the properties of an object by requesting <https://graph.facebook.com/ID>. For example, the official page for the Facebook Platform has id 19292868552, so you can fetch the object at <https://graph.facebook.com/19292868552>:

```
{  
    "name": "Facebook Platform",  
    "type": "page",  
    "website": "http://developers.facebook.com",  
    "username": "platform",  
    "founded": "May 2007",  
    "company_overview": "Facebook Platform enables anyone to build...",  
    "mission": "To make the web more open and social.",  
    "products": "Facebook Application Programming Interface (API)...",  
    "fan_count": 449921,  
    "id": 19292868552,  
    "category": "Technology"  
}
```

Alternatively, people and pages with usernames can be accessed using their username as an ID. Since "platform" is the username for the page above, <https://graph.facebook.com/platform> will return what you expect. All responses are JSON objects.

API Documentation | dev.twitter.com

← → C dev.twitter.com/doc

twitter developers

Begin | Documentation | Discussions | Sign in

Search

API Documentation

REST API · Streaming API · Search API

Twitter exposes its data via an Application Programming Interface (API). These documents are the official reference for that functionality.

Getting started with the Twitter API is easy. Jump right into the resource documentation or read some of the fine literature below.

As we prepare this new developer portal, you might find some pieces of documentation missing. Consider looking on the Twitter API Wiki if you can't find what you're looking for while we are in transition.

Guidelines and Terms

- Rules of the Road
- Display Guidelines
- Geo Developer Guidelines
- Guidelines for Use of Tweets in Broadcast or Other Offline Media

Authentication

- Which authorization path should I choose?
- Authentication
- Transitioning from Basic Auth to OAuth
- OAuth Libraries
- OAuth FAQ
- Overview of "Sign in with Twitter"

Timeline resources

- statuses/public_timeline
- statuses/home_timeline
- statuses/friends_timeline
- statuses/user_timeline
- statuses/mentions
- statuses/retweeted_by_me
- statuses/retweeted_to_me
- statuses/retweets_of_me

Tweets resources

User resources

Trends resources

Local Trends resources

List resources

List Members resources

List Subscribers resources

Direct Messages resources

REST API & General

March 30, 2011 Bring interactivity to Tweets that you display on the web with Web Intents. x

LinkedIn Developer Network developer.linkedin.com/docs/DOC-1286

LinkedIn Developers

Home Plugins API Developers Discuss Documentation

[Up to Documentation in JavaScript API](#)

JSAPI Tutorials

When learning a new framework, sometimes it's easier to understand how things behave by having working examples explained to you. We've built several tutorials at various levels, from simple applications with a single function all the way up to complex applications bringing together several different types of functionality. Each tutorial points to the appropriate places in the JSAPI documentation so you can see how to adapt and extend them for your own applications

Basic

The first step on your journey

[Authentication](#) - learn how to get started with a JSAPI app by authenticating a user

Simple Applications

Reading data from LinkedIn

[Login and Registration](#) - Quick and easy integration between LinkedIn and your site's registration system

[Profile](#) - Display the user's profile information

[Connections](#) - Display the user's connections

[Network Stream](#) - Show the activity stream from the user's network

[Search](#) - Find people matching a specified filter

Extended Applications

Write data to LinkedIn

[Like](#) - "Like" entries in the network stream

[Post message](#) - Send a message

[Post update](#) - Post an update to the user's activity stream

Integrated Applications

Take data from one API and use it in another

[Streamin'](#) - Authentication, Profile, Network Stream, Like

Actions

[View as PDF](#)
 [View print preview](#)

More Like This

[LinkedIn Platform Guidelines](#)
 [Post Network Update](#)
 [Get Network Updates and Statistics API](#)
 [Exchange JSAPI Tokens for REST API OAuth Tokens](#)
 [JavaScript API Tutorial](#)

Incoming Links

[Getting Started with the JavaScript API](#)

The screenshot shows a web browser window for the New York Times Developer Network. The title bar reads "APIs" and the address bar shows "developer.nytimes.com/docs". The main content area features the "The New York Times" logo and the "Developer Network" heading with a "BETA" badge.

API Documentation and Tools

The Times Developer Network is our API clearinghouse and community. Get the latest news about New York Times APIs, read the API documentation, browse the application gallery and connect with other developers in the forum.

Overview

APIs

- APIs
- The Article Search API
- The Best Sellers API
- The Campaign Finance API
- The Community API
- Reference
- The Congress API
- Frequently Asked Questions
- The Districts API
- The Most Popular API
- The Movie Reviews API
- The NY State Legislature API
- The Real Estate API
- The Times Newswire API
- The TimesPeople API
- The TimesTags API
- Campaign Finance API

Examples

- Constructing a Request
- Requesting a Key
- Standard Errors
- Congress API Examples
- Times Newswire API: Version 2

Tools

Forum

Gallery

APIs

Terms of Use
Before you can use New York Times APIs, you must agree to the Terms of Use.

Attribution Guidelines and Restrictions
Please review these guidelines before you use our APIs.

API Key Registration
Ready to start coding? Request a key for each API you want to use.

API FAQ
Learn more about the hows and whys of Times APIs.

Available APIs

The Article Search API
Search Times articles from 1981 to today, retrieving headlines, abstracts and links to associated multimedia.

The Best Sellers API
Get data from all New York Times best-seller lists, including rank history for specific best sellers.

The Campaign Finance API
Get presidential campaign contribution and expenditure data based on United States Federal Election Commission filings.

The Community API
Get comments by NYTimes.com users.

The Congress API
Get U.S. Congressional vote data, including information about specific House and Senate members.

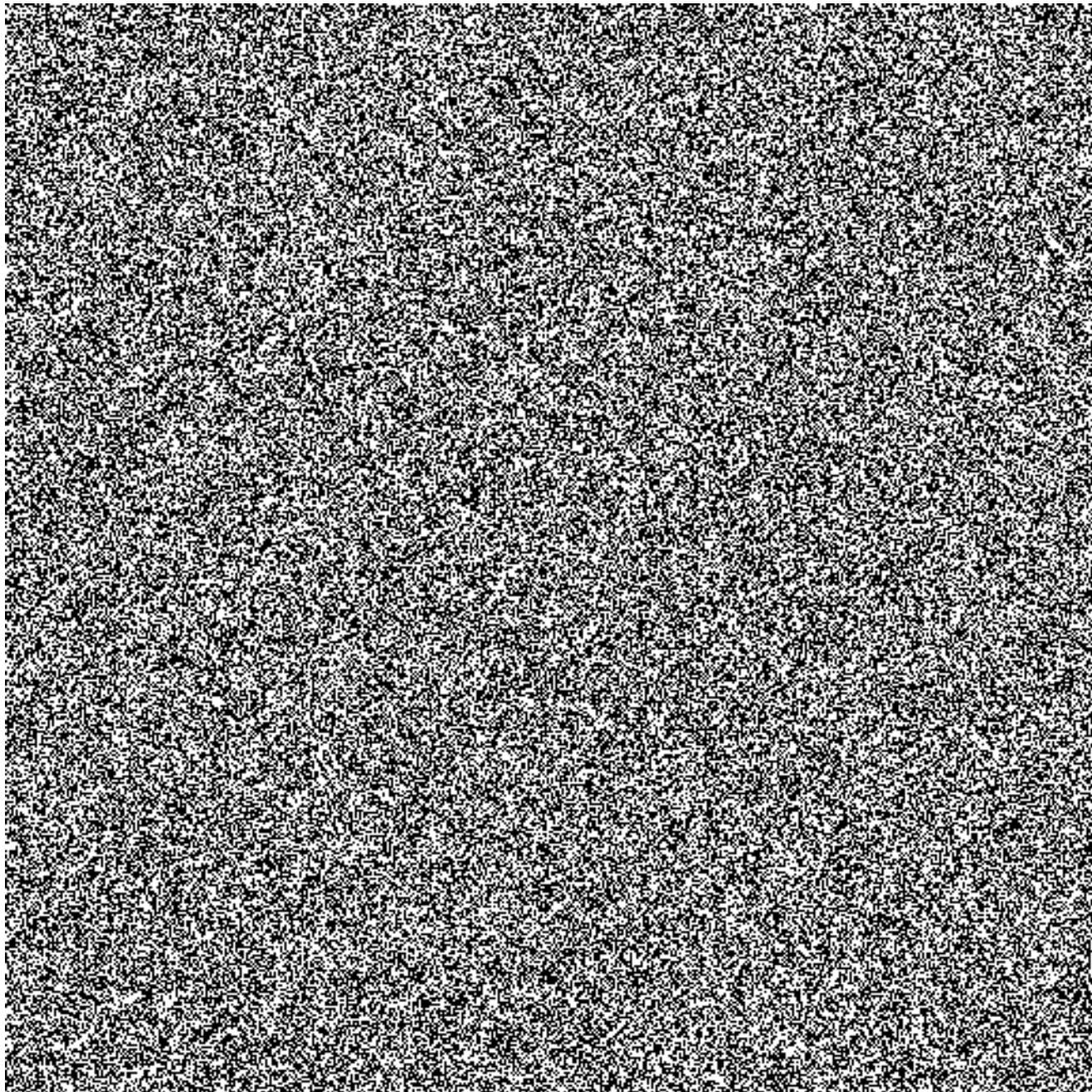
Testing?

OK, back from that distraction...

Just because a service advertises random bits (and they have a good story to go along with it) doesn't mean that it works -- To get a little technical, **we probably shouldn't think about a random number in isolation (is 1 random?) but instead talk about a sequence of random numbers**

Even this is a little vague -- What properties would we expect from a sequence of random bits (random coin tosses)? Intuitively, **what do you expect to see as you look across and down the web page on the previous slide?**

On the next slide we mapped each 1 to a black pixel and each 0 to a white one -- We then asked for $512 \times 512 = 262,144$ random bits from `random.org` and displayed them as a 512 by 512 image (putting the first bit in the upper lefthand corner, the second bit just below it and so on, filling up each column one at a time from left to right)



Testing?

Formally, there are a set of **classical statistical tests** (yes, tests of hypothesis!) that could help us assess if a random number generator (true or otherwise) is performing as expected

In this case, we can use the mathematics of probability to determine the null distribution for test statistics like **the fraction of 1's in the sequence (it should be about a half)** or **the length of “runs” of bits of the same kind (we shouldn't see long runs of 1s or 0s)**

These mathematical results help us avoid **a chicken and egg problem** -- If we needed simulation to test a null hypothesis, how could we ever test the simulator?!

Random number generation

The second kind of random number generation comes from **a mathematical formula, a deterministic algorithm** that produces a repeatable, predictable series of numbers -- These are called **pseudo-random numbers**

Here is a snippet of code that implements a “classical” example of one of these algorithms -- We start by setting the variable `seed` to some number we choose (here I picked 200)

```
# initialize

> seed <- 200

# we then update the seed and generate
# a "random" number

> a <- 16807
> m <- 2147483647
> seed <- (a*seed)%%m
> random <- seed/m

# the values random will be in the
# interval (0,1)
```

```
# iteration 1
> seed <- 200
> (a*seed)
[1] 3361400
> (a*seed)%%m
[1] 3361400
> ((a*seed)%%m)/m
[1] 0.001565273851885122

# iteration 2
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 56495049800
> (a*seed)%%m
[1] 660474978
> ((a*seed)%%m)/m
[1] 0.3075576286332484

# iteration 3
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 11100602955246
> (a*seed)%%m
[1] 259983903
> ((a*seed)%%m)/m
[1] 0.1210644390066454

# iteration 4
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 4369549457721
> (a*seed)%%m
[1] 1567719723
> ((a*seed)%%m)/m
[1] 0.7300263846898575

# iteration 5
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 26348665384461
> (a*seed)%%m
[1] 1188519418
> ((a*seed)%%m)/m
[1] 0.553447482433844

# iteration 6
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 19975445858326
> (a*seed)%%m
[1] 1700457579
> ((a*seed)%%m)/m
[1] 0.7918372656180697

# iteration 7
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 28579590530253
> (a*seed)%%m
[1] 878155977
> ((a*seed)%%m)/m
[1] 0.408923242897225

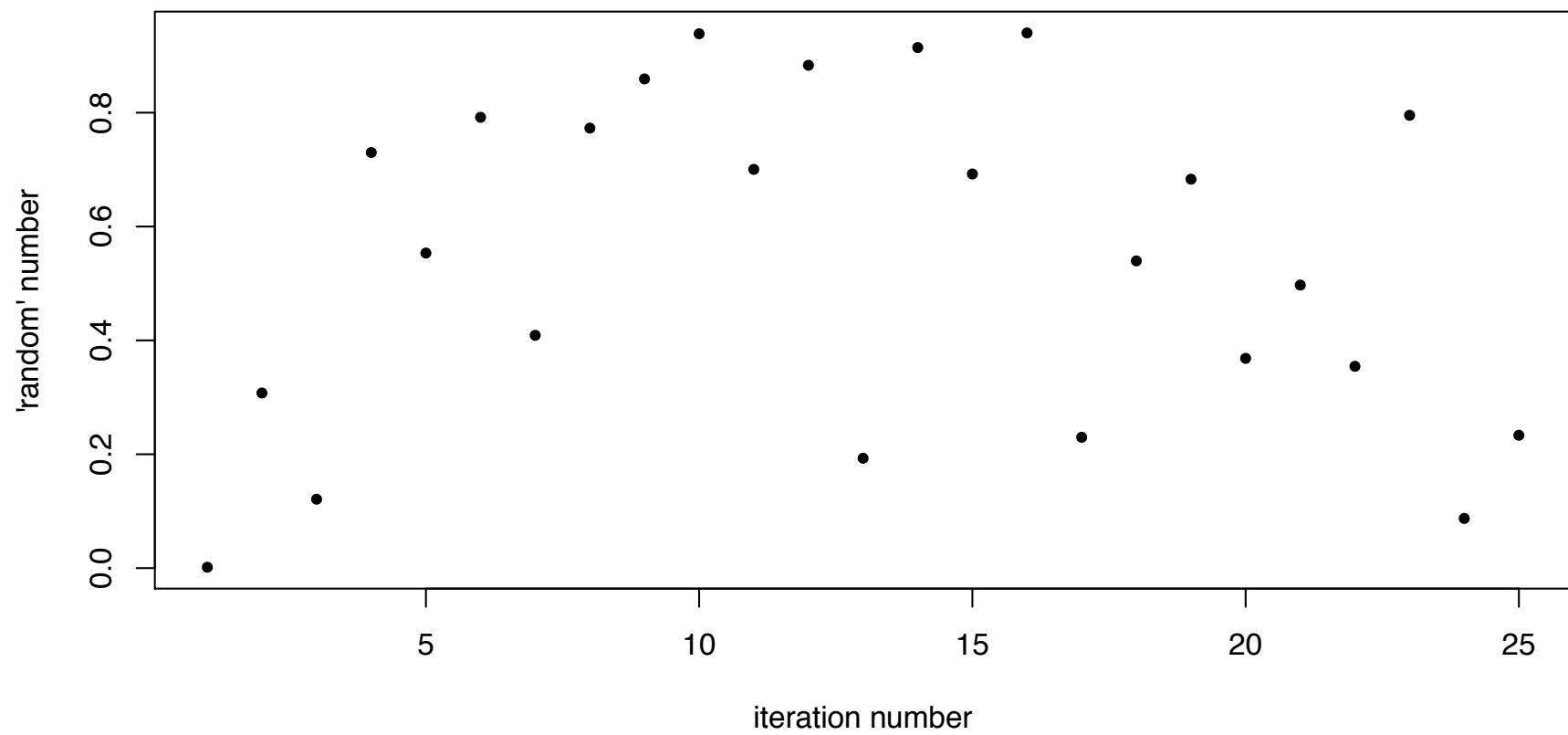
# iteration 8
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 14759167505439
> (a*seed)%%m
[1] 1659883255
> ((a*seed)%%m)/m
[1] 0.77294337366379

# iteration 9
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 27897657866785
> (a*seed)%%m
[1] 1845292255
> ((a*seed)%%m)/m
[1] 0.859281167322435

# iteration 10
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 31013826929785
> (a*seed)%%m
[1] 2015583458
> ((a*seed)%%m)/m
[1] 0.938579188165525

# iteration 11
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 33875911178606
> (a*seed)%%m
[1] 1504130828
> ((a*seed)%%m)/m
[1] 0.700415497971892

# iteration 12
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 25279926826196
> (a*seed)%%m
[1] 1896817359
> ((a*seed)%%m)/m
[1] 0.883274413590913
```



Uniform random numbers

This procedure is also known by the mouthful of a name “**prime modulus multiplicative linear congruential generator**” (and often shortened to the equally difficult PMMLCG)

Technically the algorithm leaves the constants (a and m) unspecified, but our choice can be shown to have good properties relative to the statistical tests I alluded to

By construction, the numbers we highlighted in red are all between 0 and 1 -- They can be used anywhere we might need observations from the so-called **uniform distribution** on the interval $[0,1]$

Uniform random numbers

As its name suggests, we expect to see observations from the uniform distribution distributed, well, uniformly over $[0,1]$ -- To be a little more precise, if we have a sample of 1200 such observations, we'd expect about 600 to be less than 0.5

Going farther, if we divided $[0,1]$ into four equally sized subintervals (from 0 to 0.25, 0.25 to 0.5, 0.5 to 0.75 and 0.75 to 1) we would expect to see 300 observations of the 1200 in each bin (or so)

In general, under the uniform distribution, we expect **the proportion of our sample that falls in some subinterval we specify to be equal to the length of that interval** (the uniform distribution is a mathematical construction that we will examine more closely when we discuss probability in a future lecture)

So, using the algorithm two slides back, let's create some random bits -- We'll generate $512 \times 512 = 262,144$ numbers in $[0,1]$ using this algorithm and **coloring a square black if the associated number is larger than 0.5 and color it white if it is less than or equal to 0.5...**

Uniform random numbers

Here is the sequence of seed values we get are

```
[1] 3361400 660474978 259983903 1567719723 1188519418 1700457579  
[7] 878155977 1659883255 1845292255 2015583458 1504130828 1896817359  
[13] 414612998 1963706518 1486760930 2018717665 493656702 1158862153  
[19] 1467010828 791234989 1067717899 761374161 1707955101 187473058  
[25] 501175657
```

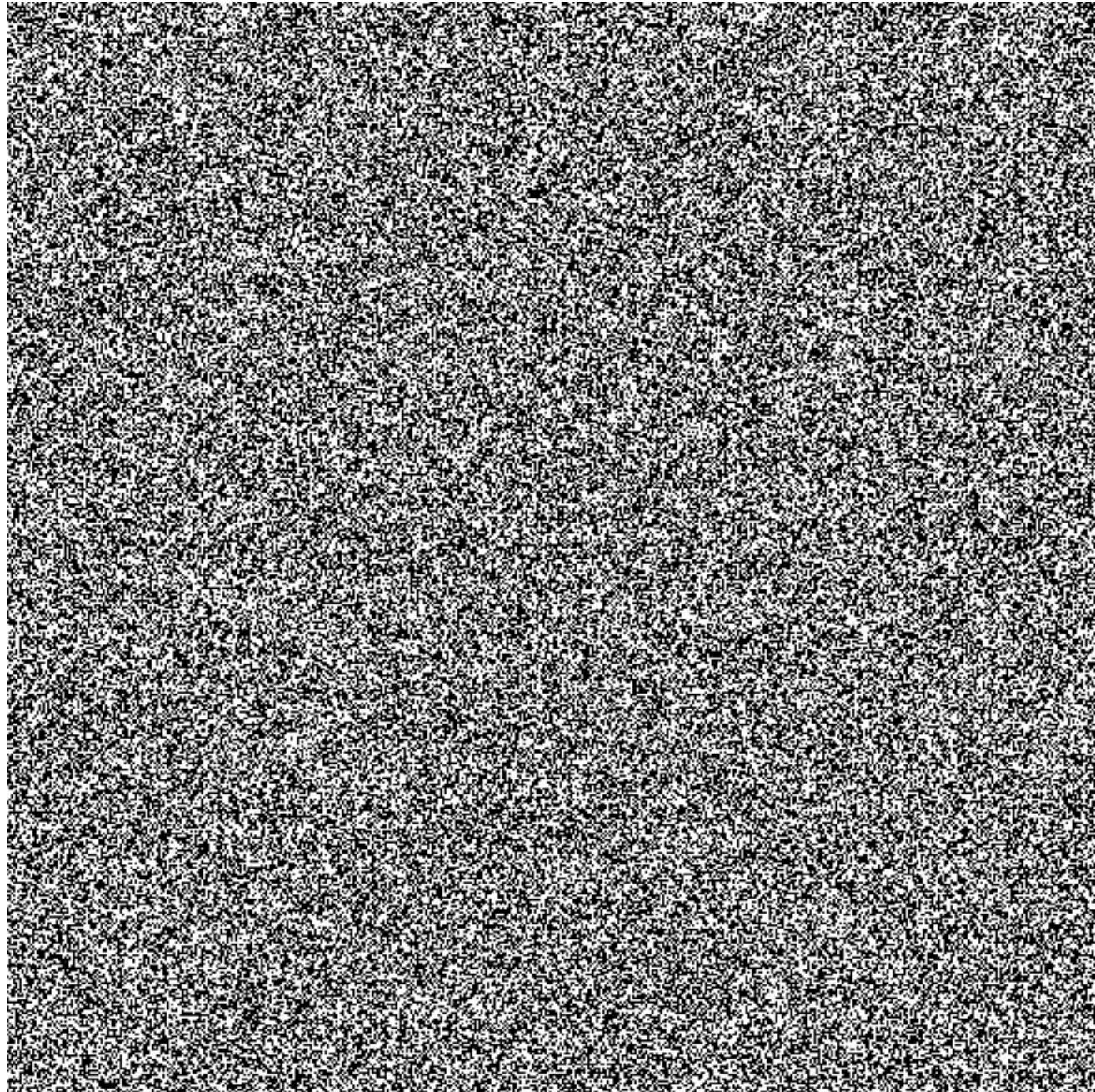
which when divided by $2^{31}-1$ gives the pseudo-random uniform observations

```
[1] 0.001565273851885122 0.307557628633248425 0.121064439006645430  
[4] 0.730026384689857477 0.553447482433844118 0.791837265618069663  
[7] 0.408923242897225203 0.772943373663790179 0.859281167322434980  
[10] 0.938579188165524547 0.700415497971892176 0.883274413590912855  
[13] 0.193069222473105984 0.914422105492289194 0.692327008905041508  
[16] 0.940038667032513153 0.229876815448457755 0.539637242229486502  
[19] 0.683130150978979778 0.368447503712236668 0.497194891561379138  
[22] 0.354542472099206640 0.795328571365833570 0.087298945564450212  
[25] 0.233378101714597136
```

and the random bits

```
[1] 0 0 0 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 0 0 1 0 0
```

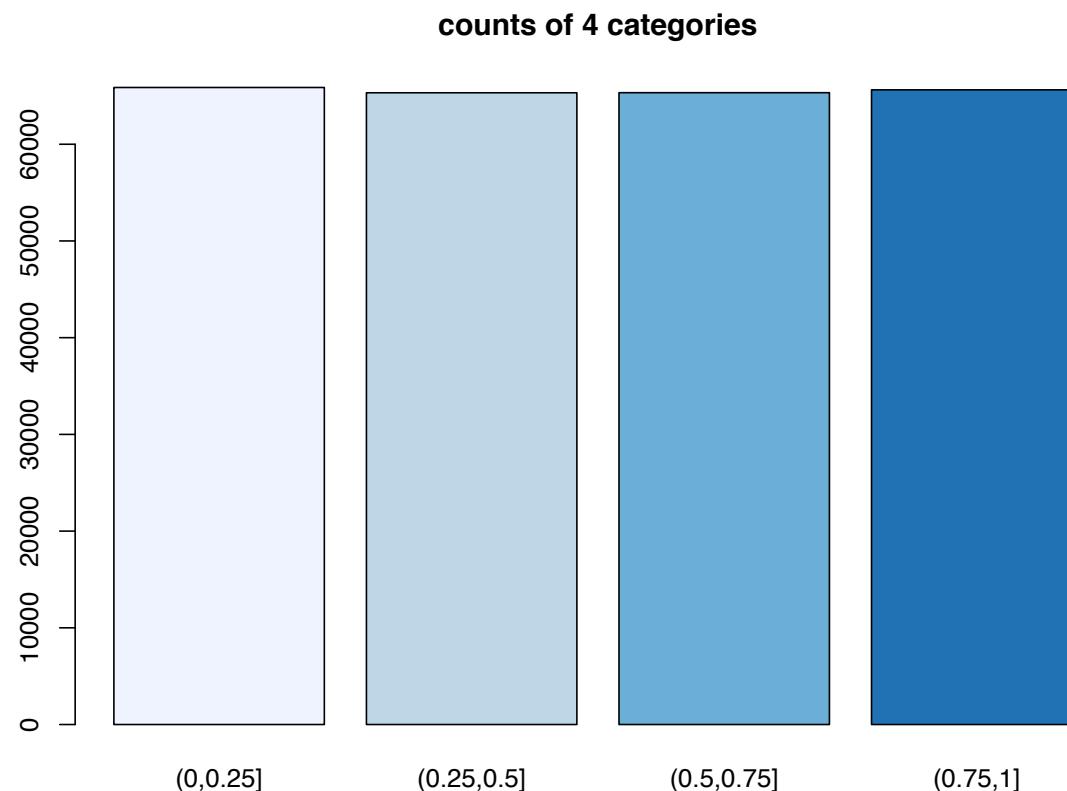
which, when continued for the full 512*512 values and arranged in an image gives us...

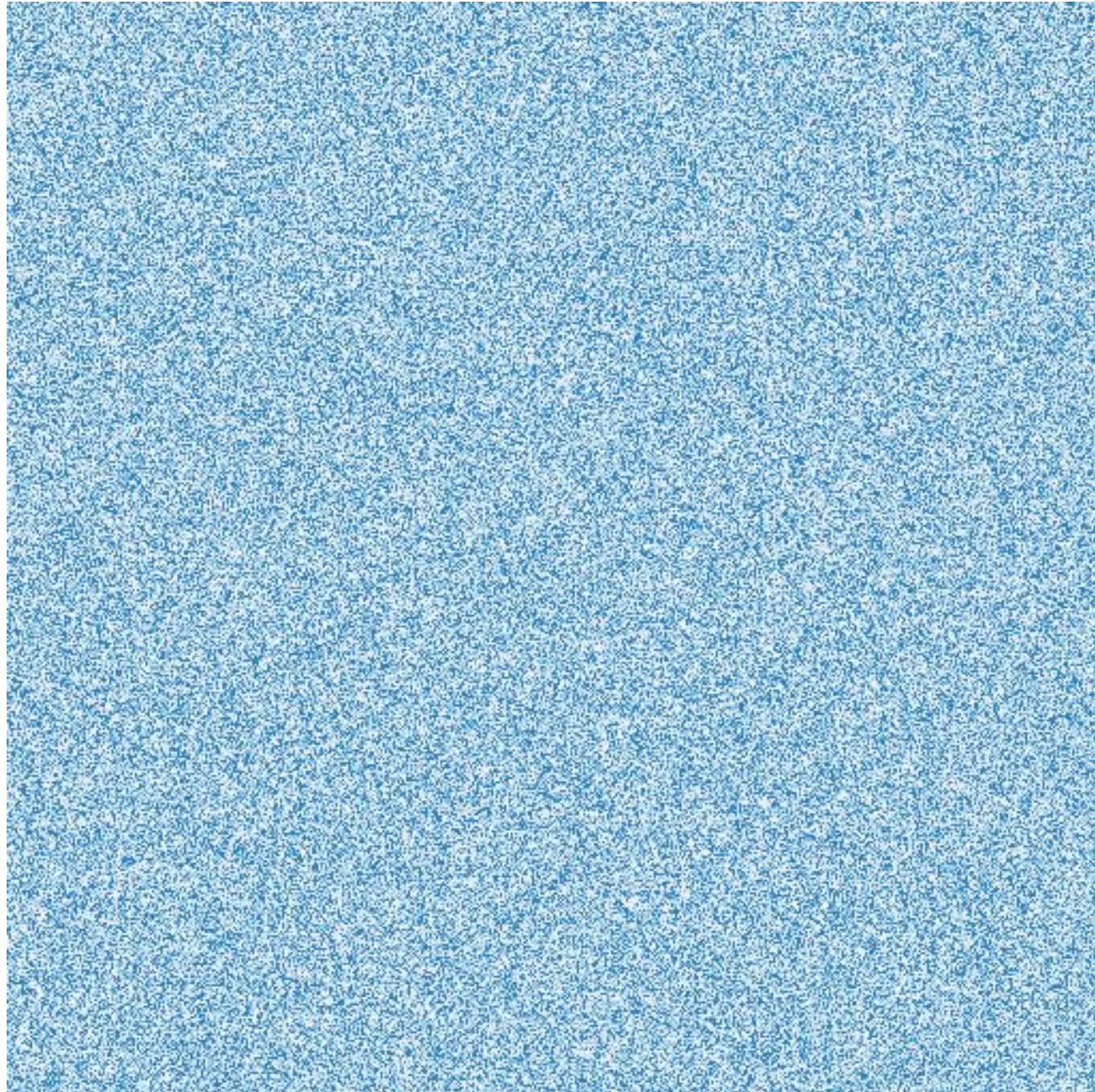


Uniform random numbers

The proportion of 1's constructed this way is 0.49955 -- We can take this farther and **divide the data into four pieces** (those between 0 and 0.25, between 0.25 and 0.5, between 0.5 and 0.75 and then larger than 0.75) and below we have a barplot of counts in each interval (and yes, it looks essentially flat)

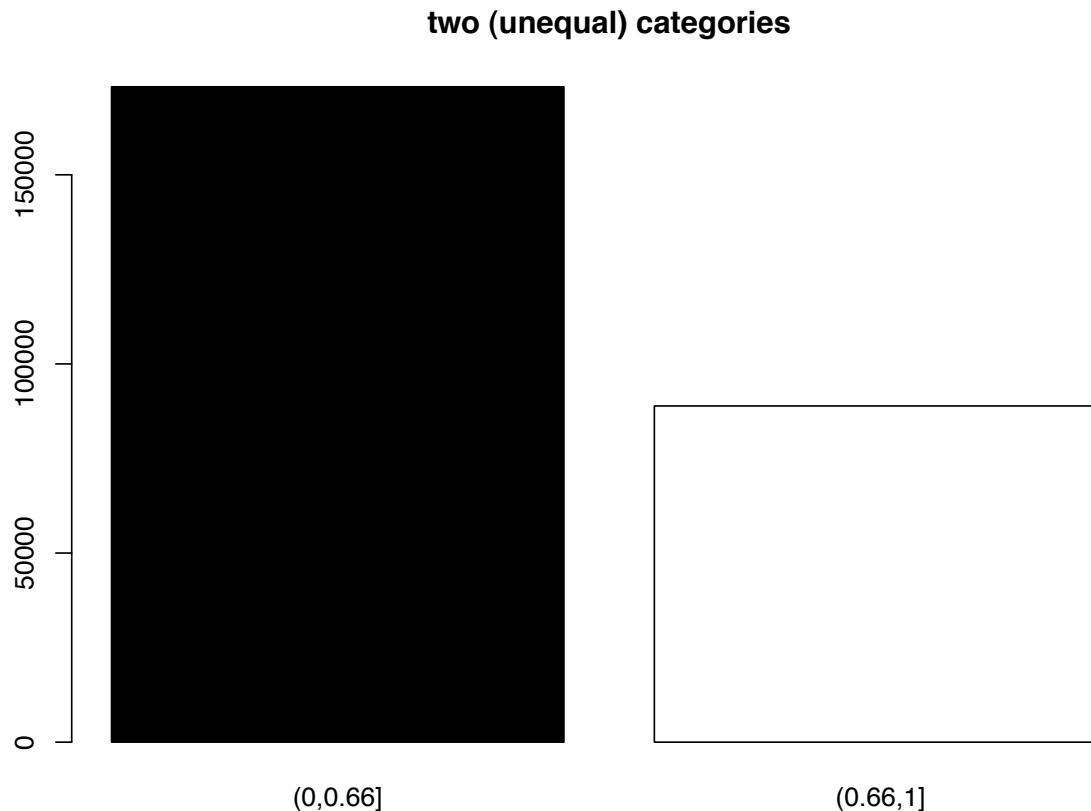
Essentially, an equal number of points falls in each interval -- We can color points falling into the four pieces using different colors (the colors of the four bars on the right) and pack them into an image again

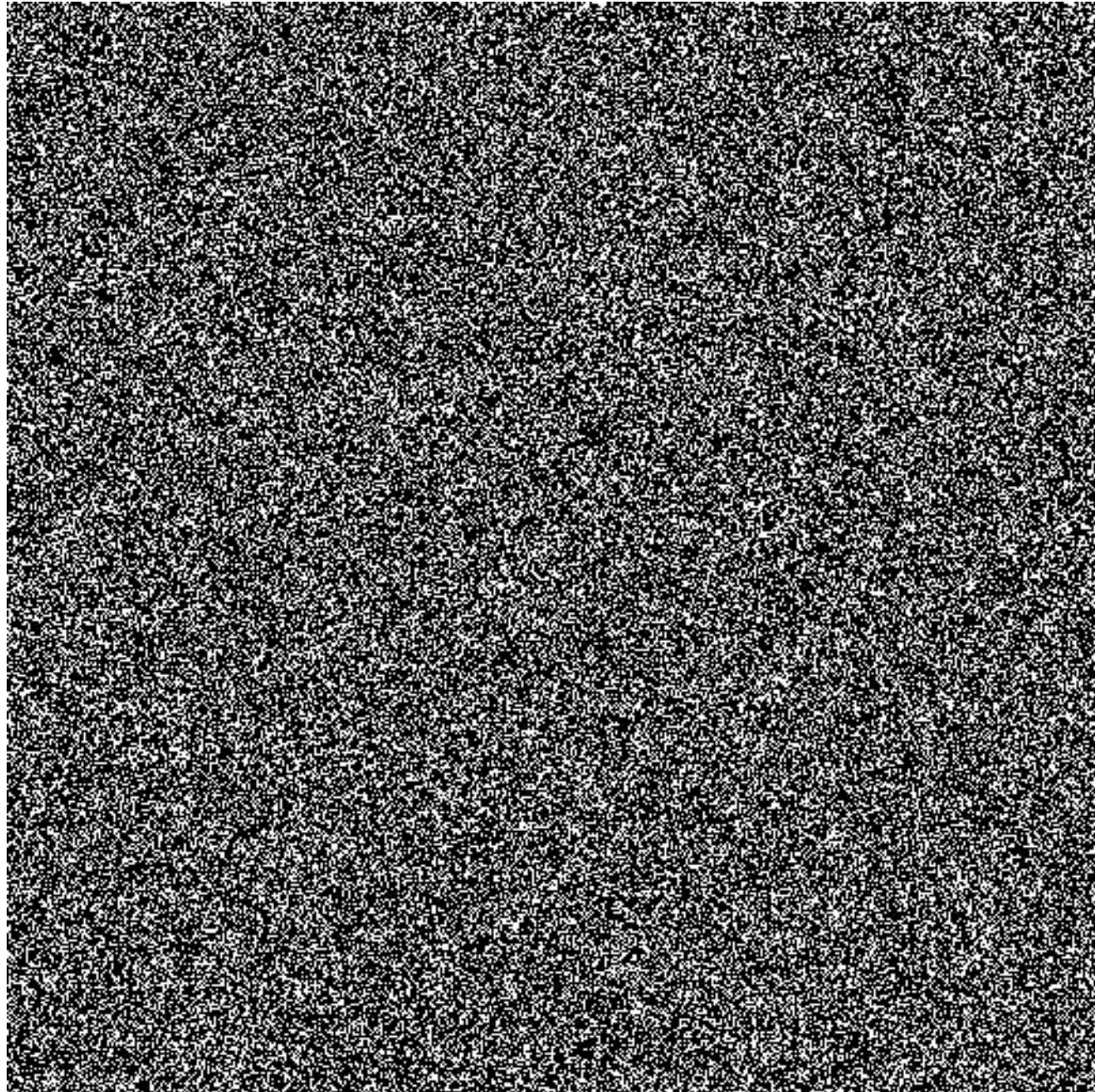




Uniform random numbers

Dividing the data using the intervals [0,0.66) and [0.66,1], we see essentially 2/3 of the data fall into the first category and 1/3 in the second -- On the next slide we play the image game again





To sum

There are **deterministic mathematical algorithms that allow us to generate observations that have many of the characteristics we'd expect to see from truly random data** (where our expectations are set by statistical tests like proportions falling into intervals or “runs” of particular kinds)

Starting with the uniform distribution, we can simulate a host of random phenomena, from permuting our data as in the case of our re-randomization analysis to tossing coins (a la Arbuthnot) to making observations from the normal (bell-shaped) distribution (later)

This idea might take some getting used to, but **the use of pseudo-random numbers is common practice** and even R depends on one such algorithm (the default is the so-called Mersenne-Twister)

Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.

John von Neumann

Some advantages

While pseudo-random numbers are entirely deterministic, there are some advantages for scientific uses

Chief among them is reproducibility! If our analysis depends on simulation (like our re-randomization procedures for inference) we would like to be able to reproduce our results exactly (this comes in hand, say, when you want to debug more complex algorithms)

In R you can use the function `set.seed()` at any point to reset your sequence of random numbers (R does not use the algorithm described here, but it shares the properties of being a mathematical formula, deterministic and predictable)

```
> treat
[1] "T" "T" "T" "T" "T" "C" "C" "C" "C" "C"

# treat holds a division into treatment and
# control -- let's use sample() to permute
# them or re-randomize

> set.seed(1000)
> sample(treat)
[1] "T" "C" "T" "T" "C" "C" "T" "T" "C" "C"
> sample(treat)
[1] "T" "C" "T" "C" "T" "T" "T" "C" "C" "C"
> sample(treat)
[1] "T" "C" "T" "T" "T" "C" "C" "C" "C" "T"

# we can repeat our rerandomizations by
# resetting the seed

> set.seed(1000)
> sample(treat)
[1] "T" "C" "T" "T" "C" "C" "T" "T" "C" "C"
> sample(treat)
[1] "T" "C" "T" "C" "T" "T" "T" "C" "C" "C"
> sample(treat)
[1] "T" "C" "T" "T" "T" "C" "C" "C" "C" "T"
```

Testing?

One final note -- **Not every service or system or program that advertises random numbers is any good!**

Here is the same bit picture for a combination of the programming language **PHP running on a Windows machine** -- Evidently this was the result of a bug that has since been corrected, but it does serve as a cautionary example

