Lecture 5: Noise

Last time

We had a look at how we would generalize the notion of a box plot -- It led us to the idea of “data depth” and a new way to look at things like the median

We then introduced a particular kind of study, the randomized controlled trial, and discussed how simulation, and in particular re-randomization, could be used as a basis for drawing conclusions about the study

Today

We are going to finish up the clinical trial discussion by examining a competing strategy for analysis from Neyman and Pearson -- We'll see two very different views of how one learns (or doesn't) from data

We will then discuss random number generation -- Because we rely so heavily on simulation it's worth discussing how a computer generates random numbers in the first place

We end with a second kind of randomized trial, but one with a much more recent history...

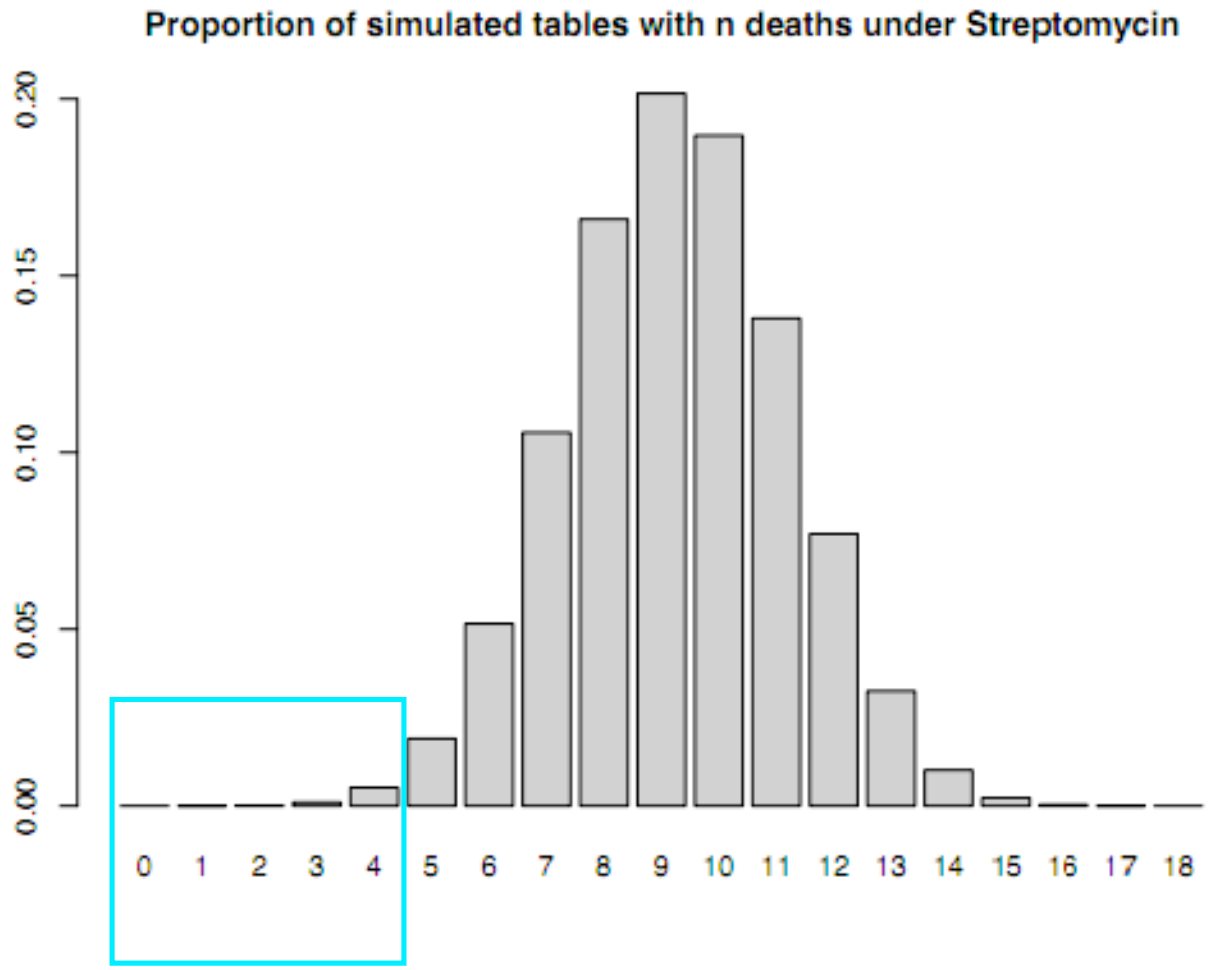
Significance testing

In Hill's Tuberculosis study, of the 55 patients receiving Streptomycin, 4 died (4 out of 55 or 7% of those patients) -- While in of the 52 patients in the bed rest group, 14 died (14 out of 52 or 27%)

To formally test this result, we started with **the null hypothesis that Streptomycin was no more effective at preventing death from tuberculosis than bed rest**

Under that assumption, we computed a P-value of 0.006 -- This is the chance of seeing 4 or fewer deaths assigned to the Streptomycin group at random

This, one could argue, is **very strong evidence against the null** -- It is possible that the null is true, and if it is, Hill witnessed an event that should occur in only 1 out of 157 random assignments


Proportion of simulated tables with n deaths under Streptomycin
0.00
0.05
0.10
0.15
0.20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

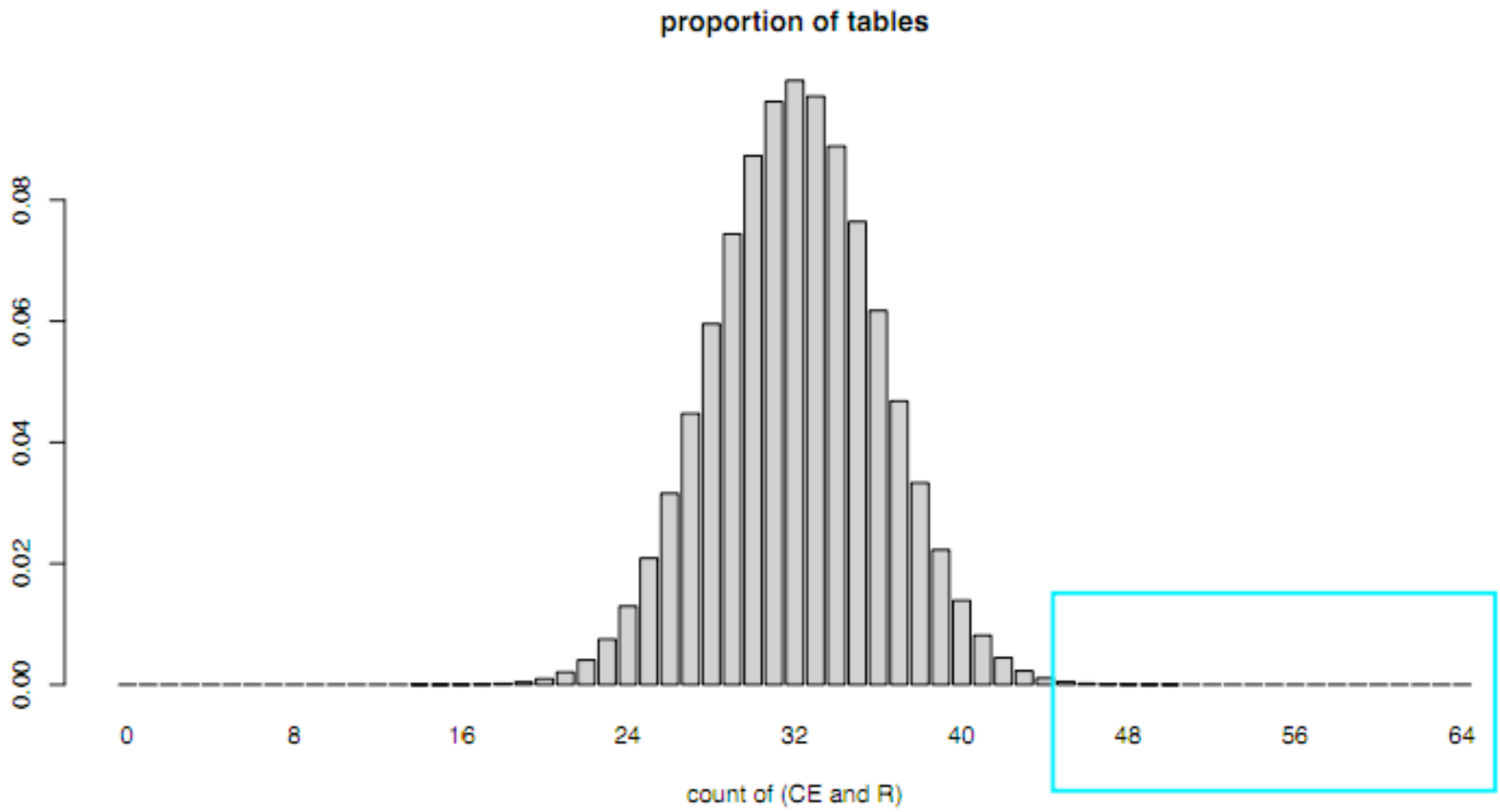
Significance testing

The setup for the Vioxx trials was largely the same, but with one important difference -- **Here we were assuming that the control was safe and we were not expecting Vioxx to reduce the number of cardiovascular adverse events (CE)**

In the Bombardier study from a previous lecture, 19 out of 4029 (0.5%) of the patients in the control group experienced CE, while 45/4047 (1.1%) of patients in the Vioxx group experienced CE

Our null hypothesis was that Vioxx and the control had the same propensity to cause CE -- Under that assumption, we computed a P-value of 0.0000004, the chance of seeing 45 or more deaths assigned to the Vioxx group at random

This, one could argue, is **amazingly strong evidence against the null** -- It is possible that the null is true, and if it is, the researchers witnessed an event that should occur in only 1 out of 2.5M random assignment


proportion of tables
0.00
0.02
0.04
0.06
0.08
0
8
16
24
32
40
48
56
64
count of (CE and R)

Significance testing

In both the Vioxx and the Hill trials, we conducted a one-sided significance test, that is, **we defined the notion of "extreme" in one direction** -- for Hill, we considered extreme to mean fewer deaths, while for Vioxx, we took it to mean more

In both cases, you could argue that **our notion of extreme came from the problem we were considering**: In Hill's case, there had been previous laboratory work and a few small patient trials that suggested Streptomycin should be better than bed rest; and for Vioxx, the entire class of COX-2 inhibitors had been suspected of causing heart problems and it is generally believed that taking the commonly prescribed doses of naproxen carries no increased risk of CE

Therefore, when we sought evidence against the null (that bed rest and Streptomycin were the same in terms of their ability to treat tuberculosis, or that Vioxx and naproxen had the same impacts on the cardiovascular system) **we looked for evidence against the null in one direction** -- Had we seen many more deaths under Streptomycin or many fewer incidents of CE under Vioxx, **the researchers might have questioned the experimental setup rather than take it as evidence against the null**

Significance testing

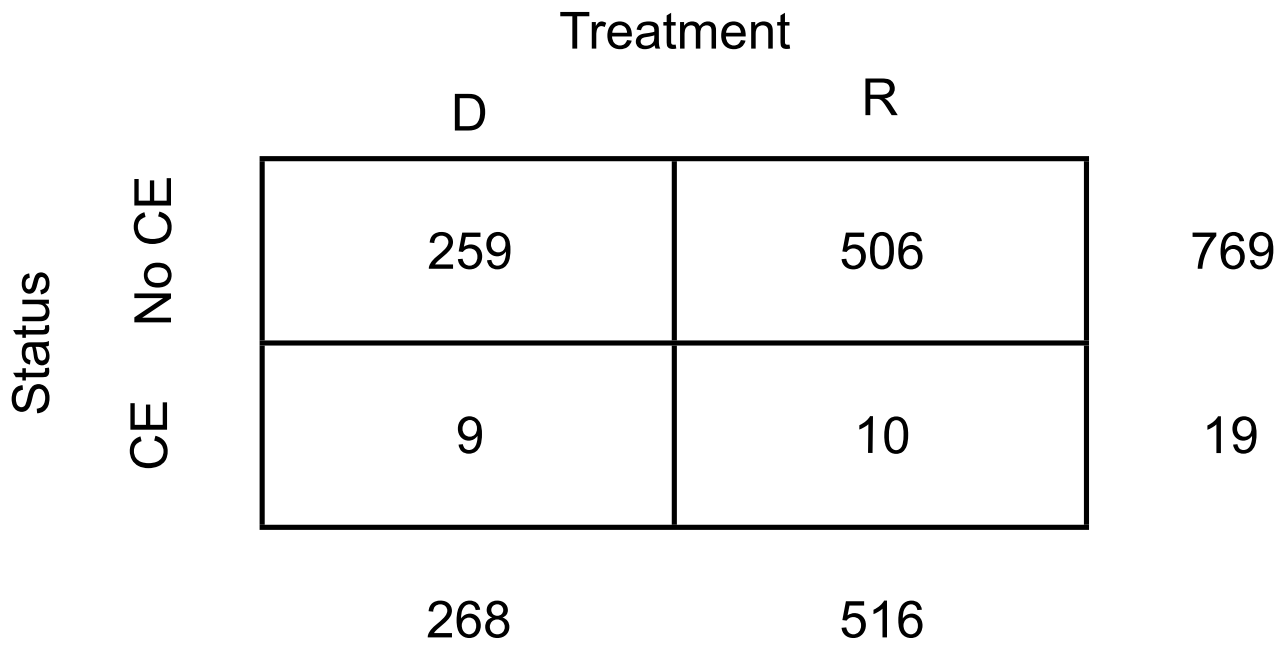
Our decision to introduce tests in this way was largely for clarity -- In many (most?) cases, we don't have quite as much information about the situation and we would be willing to accept departures in either direction as evidence against the null

That brings us to theCannon study for Vioxx (our last one, I promise) -- Unlike the Bombardier trial, the control here was diclofenac, **a substance that many believe increases the chance of adverse cardiovascular events**

In this case, we aren't sure what direction to expect (does diclofenac carry a higher risk of CE than Vioxx?) **and so we would conduct a two-sided test** -- Again, we would take as evidence against the null an extreme event that sees either drug carrying a higher risk

Significance testing

In this study, 9 out of 268 (3.4%) of the patients in the control group experienced CE, while 10/516 (1.9%) of patients in the Vioxx group experienced CE -- **Our null hypothesis is that Vioxx and diclofenac had the same propensity to induce cardiovascular adverse events**

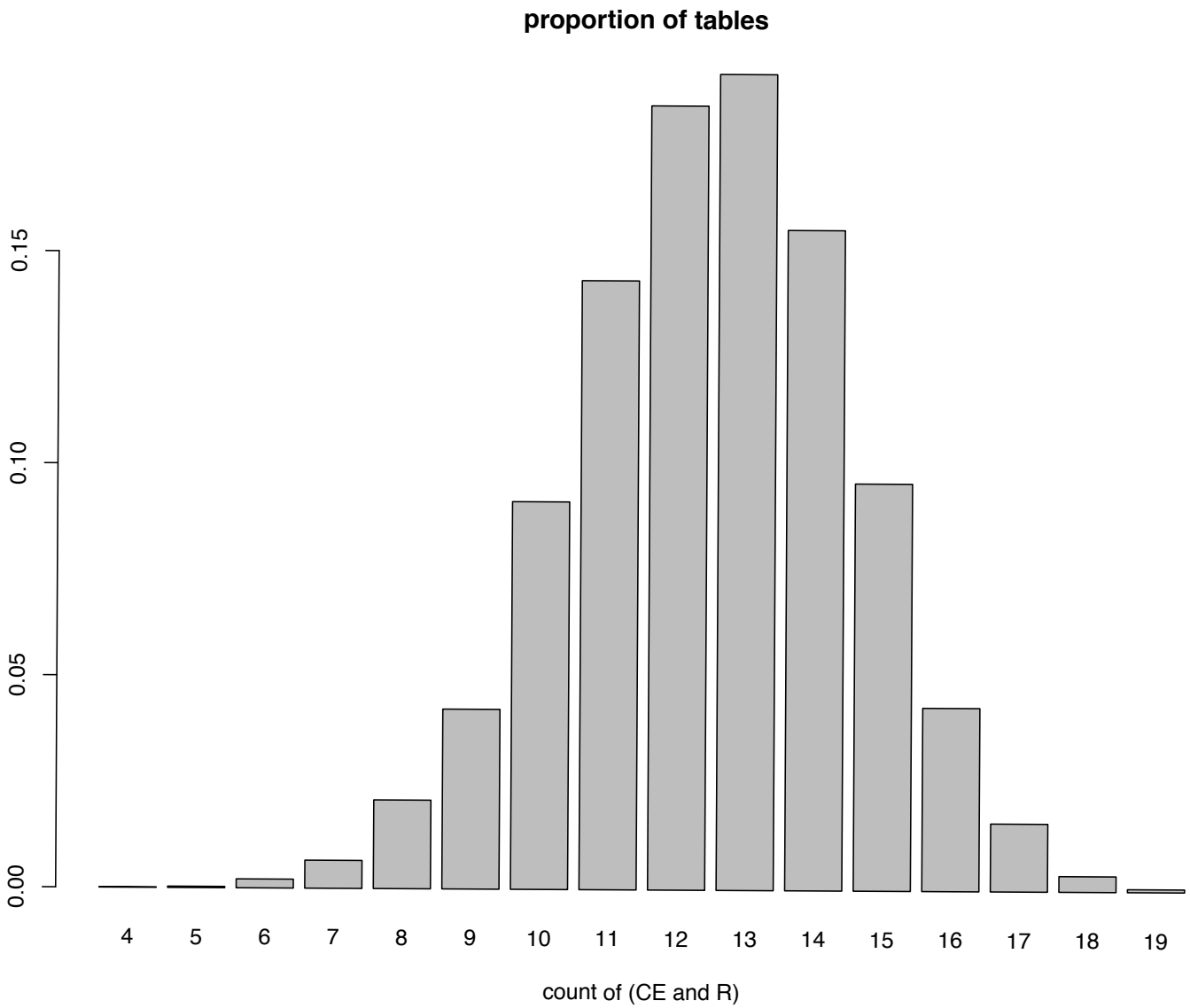


Significance testing

Our null hypothesis is that both treatment and control carry the same risk of heart attack and we again take as our test statistic the number of heart attacks in the Vioxx group

Under the null hypothesis, the 19 heart attacks would have occurred no matter which group they had been randomized into and our observed data (10 deaths in the Vioxx group) is simply the result of random assignment -- To test this we generate a series of re-randomizations and compare our observed test statistic (10 deaths in the Vioxx group) to the values from the different re-randomized tables

On the next slide we give the null distribution of our test statistic -- Identify where our observed test statistic is on the x-axis...


proportion of tables
0.00
0.05
0.10
0.15
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
count of (CE and R)

Significance testing

The very small numbers in this distribution (1, 2, 3...) correspond to small numbers of heart attacks in the Vioxx group and a correspondingly large number in the control group -- This would mean the control drug diclofenac is associated with a greater risk of heart attack

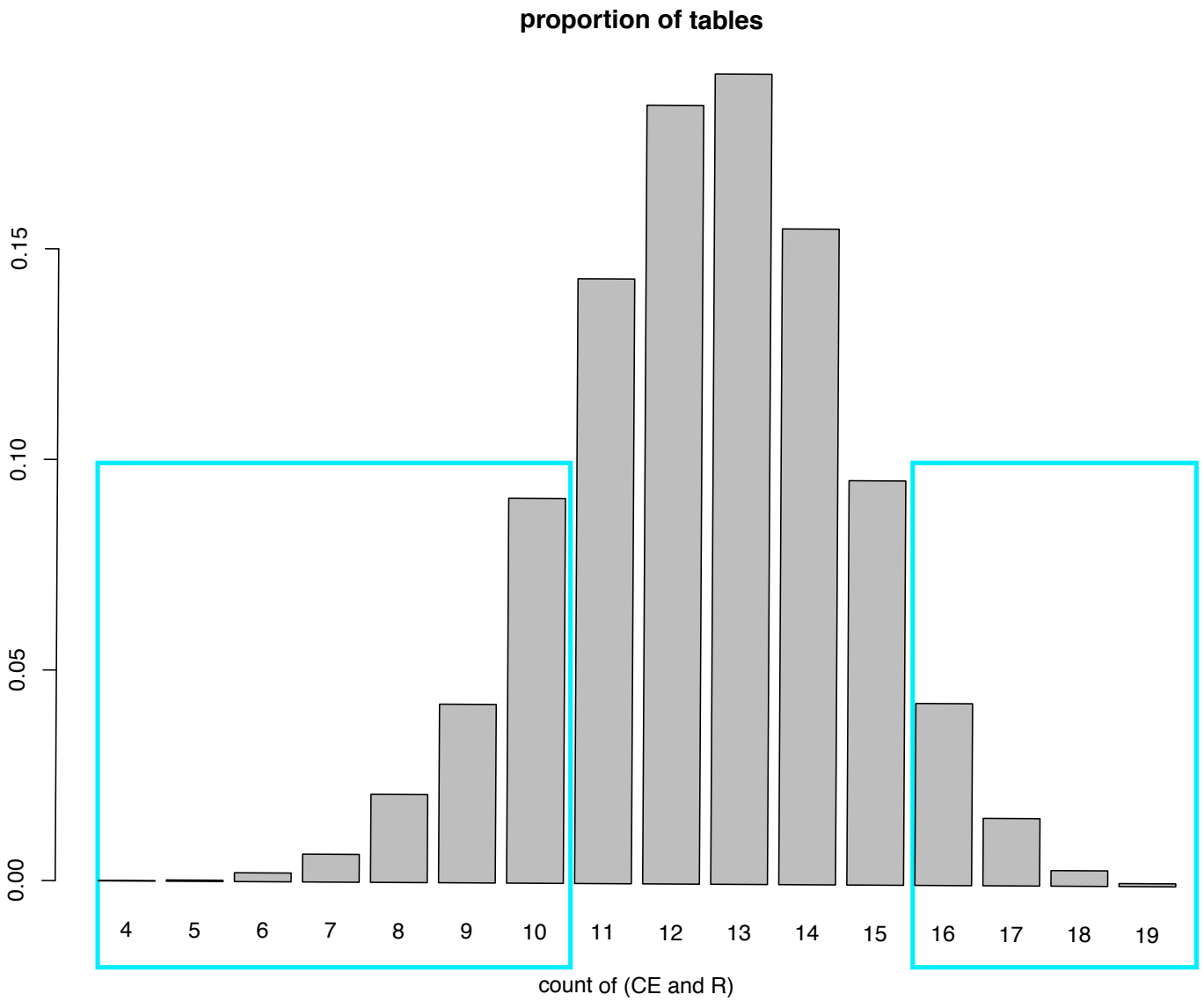
Large numbers (16, 17, 18...) correspond to more deaths under Vioxx than under the control -- This would mean Vioxx is associated with a greater risk of heart attack than the control drug, diclofenac

If we have no reason to suspect which drug should be more harmful, both large and small numbers (both tails of the null distribution) would provide evidence against the null hypothesis that the treatment and control are the same

Significance testing

We now define as extreme counts in the Vioxx group that have **probability not larger than the table we observed** (with 10 deaths) -- This notion of extreme will look for **very large or very small numbers of CE due to Vioxx** as evidence for a difference (in the former, Vioxx has increased risk of CE; in the latter diclofenac)

The probabilities we add are now in boxes that hover over **both tails of the distributions** -- The P-value (the sum of the probabilities highlighted) is 0.23 meaning an event as extreme as the one we saw occurs in about 1 in 4 randomizations and **seems consistent with what we expect from the null**


proportion of tables
0.15
0.10
0.05
0.00
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
count of (CE and R)

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.**

Inductive inference v. inductive behavior

In these comments, we also find fisher reacting against an alternative approach to the problem of inference -- This was advocated by two well-known statisticians **Jerzy Neyman and Egon Pearson**

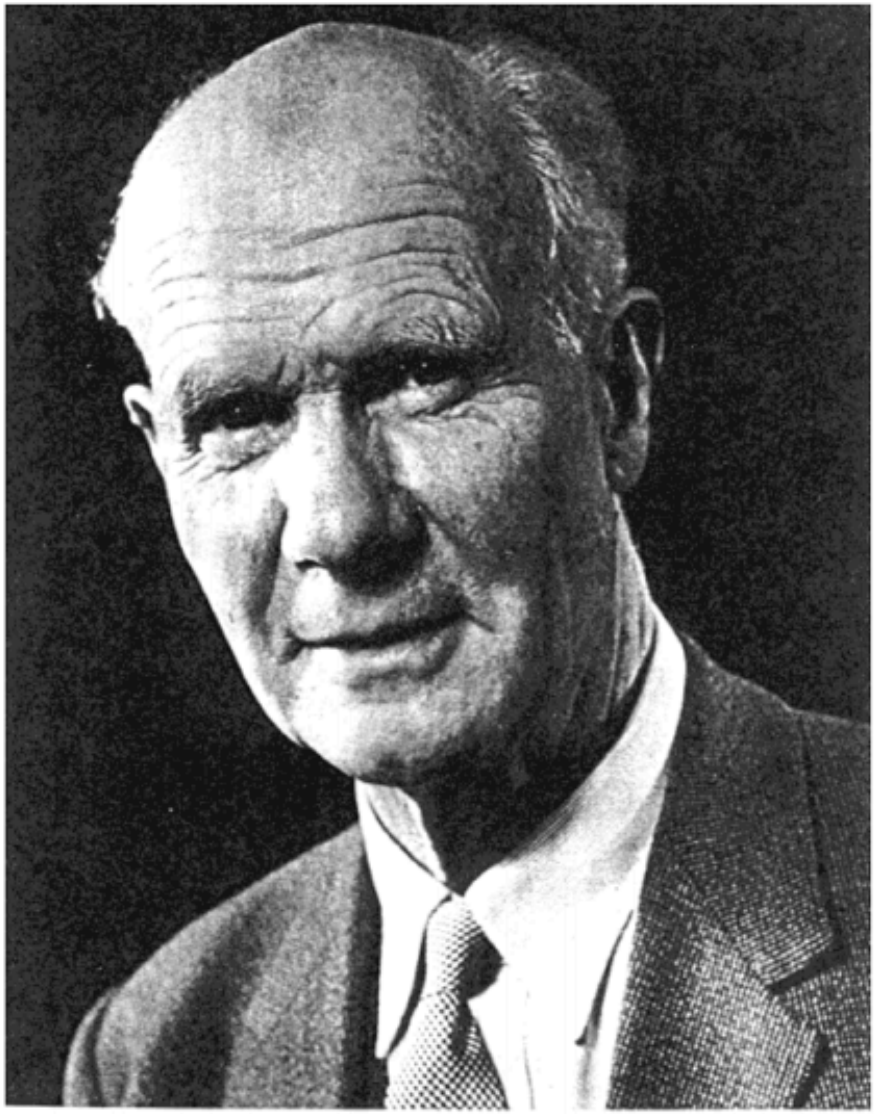
Neyman and Pearson disagreed with the subjective interpretation inherent in Fisher's approach and developed instead a procedure (which they termed hypothesis testing) based on **hard decisions about when to reject a null hypothesis** -- In effect they imposed a threshold called **the significance level**

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 Berkeley Statistics Department Archives



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**; we have seen the need for this alternative in our previous examples (the alternative makes explicit our choice of one- versus two-sided tests, for example)

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 Bombardier Vioxx trial, we took the null to be that Vioxx and naproxen represented the same risk of cardiovascular adverse events; and the alternative was that Vioxx posed a greater risk

3. Finally, for the Cannon study of Vioxx against diclofenac, our null was that both drugs carried the same risk of CE; and the alternative was that the were different (with either posing a greater risk)

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 trails 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..

The New York Times

Evidence in Vioxx Suits Shows Intervention by Merck Officials

By ALEX BERENSON
Published: April 24, 2005

In 2000, amid rising concerns that its painkiller Vioxx posed heart risks, Merck overruled one of its own scientists after he suggested that a patient in a clinical trial had probably died of a heart attack.

In an e-mail exchange about Vioxx, the company's most important new drug at the time, a senior Merck scientist repeatedly urged the researcher to change his views about the death "so that we don't raise concerns." In later reports to the Food and Drug Administration and in a paper published in 2003, Merck listed the cause of death as "unknown" for the patient, a 73-year-old woman.

The discussion of the death is contained in several previously undisclosed Merck records, including e-mail messages from Dr. Edward M. Scolnick, Merck's top scientist from 1985 until 2002, and from Dr. Alise S. Reicin, a vice president for clinical research, that indicate Merck's concerns about data contradicting its view that Vioxx was safe.

The New York Times

Evidence in Vioxx Suits Shows Intervention by Merck Officials

By ALEX BERENSON
Published: April 24, 2005

... Mr. Mayer said the doctors involved in the e-mail exchanges could not be certain whether the woman who died was taking Vioxx or an older painkiller, naproxen, that was used in the trial, because information about which participant was taking which drugs was kept confidential. But at the time, there was widespread concern within the company about the relationship between Vioxx and heart attacks as a result of troubling earlier research.

During the Advantage trial, eight people taking Vioxx suffered heart attacks or sudden cardiac death, compared with just one taking naproxen, according to data released by the F.D.A. earlier this year. **The difference was statistically significant, but Merck never disclosed the data that way.**

Back in 2000, Merck was already struggling to explain the results of another study, called Vigor, which also indicated that patients taking Vioxx had more heart attacks than those taking naproxen, which is found in over-the-counter drugs like Aleve. Unlike the Advantage results, the Vigor results had been publicly disclosed by Merck....

The problem with thresholds

The decision-driven nature of hypothesis testing has its downsides -- We all want a simple procedure that tells us when we are deciding truth from fiction

The problem is that the objective security may **blind us from important results, or have us fixate on effects that are statistically significant but uninteresting** -- either way, many disciplines have felt the sting of having researchers incentivized to be on one side or another of a very hard threshold

Over the last few decades, there have been many attempts to improve how scientific results are reported, how evidence is presented -- In the next few lectures we will come across constructions like **confidence intervals** that many insist are more sensible summaries than P-values or hypothesis tests

An alternative

As an example, rather than report the results of a significance or hypothesis test, we could instead compute the relative risk in each of our Vioxx trials and examine its size

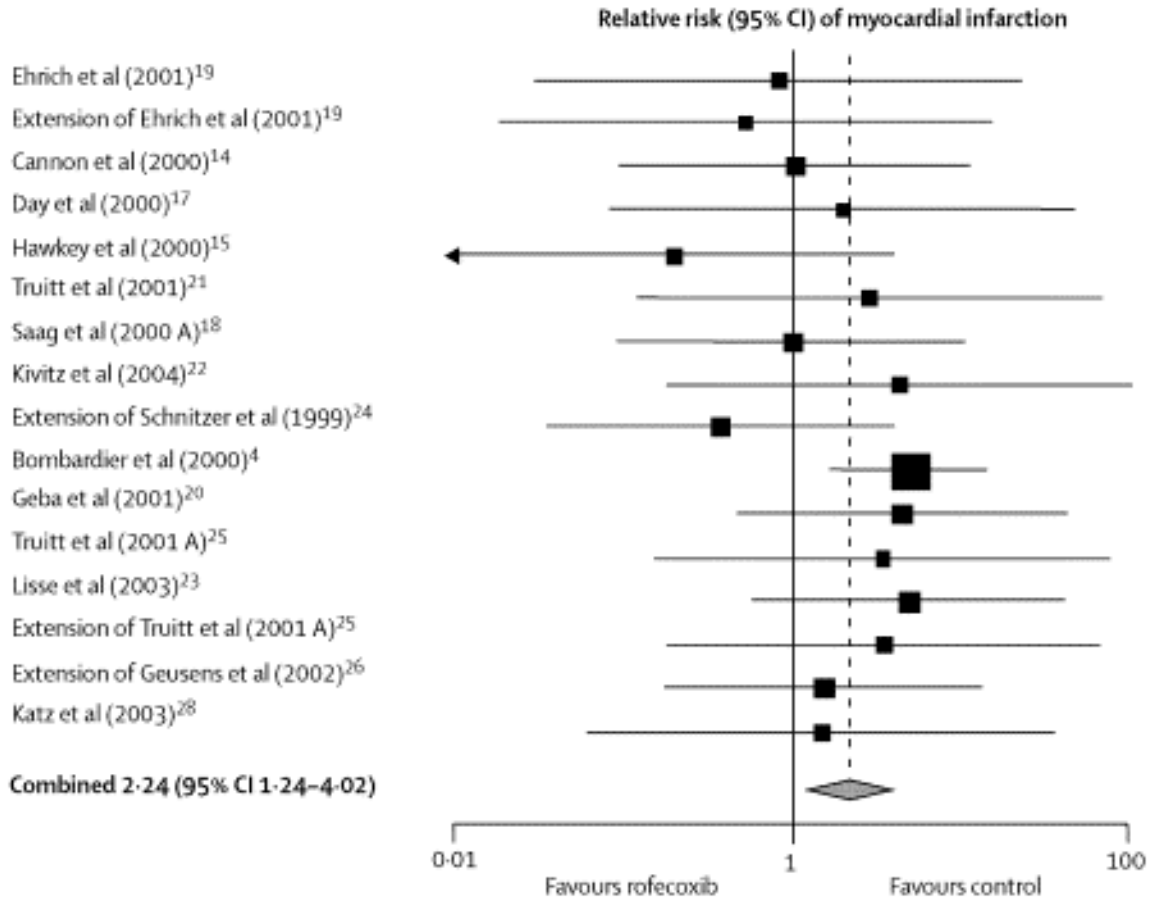
In this case, for each study we've considered, we can compute **the ratio of the chance you have a heart attack taking Vioxx relative to the chance under a control like Aleve** -- With this definition, what values of the relative risk are "interesting" or important?

For each study, we can also **summarize the uncertainty in the data** (again, coming from the randomization process) using a confidence interval -- We will say a lot about these later in the term but for now it's sufficient to recognize that these represent **plausible values for the relative risk given the randomness in our data**

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**

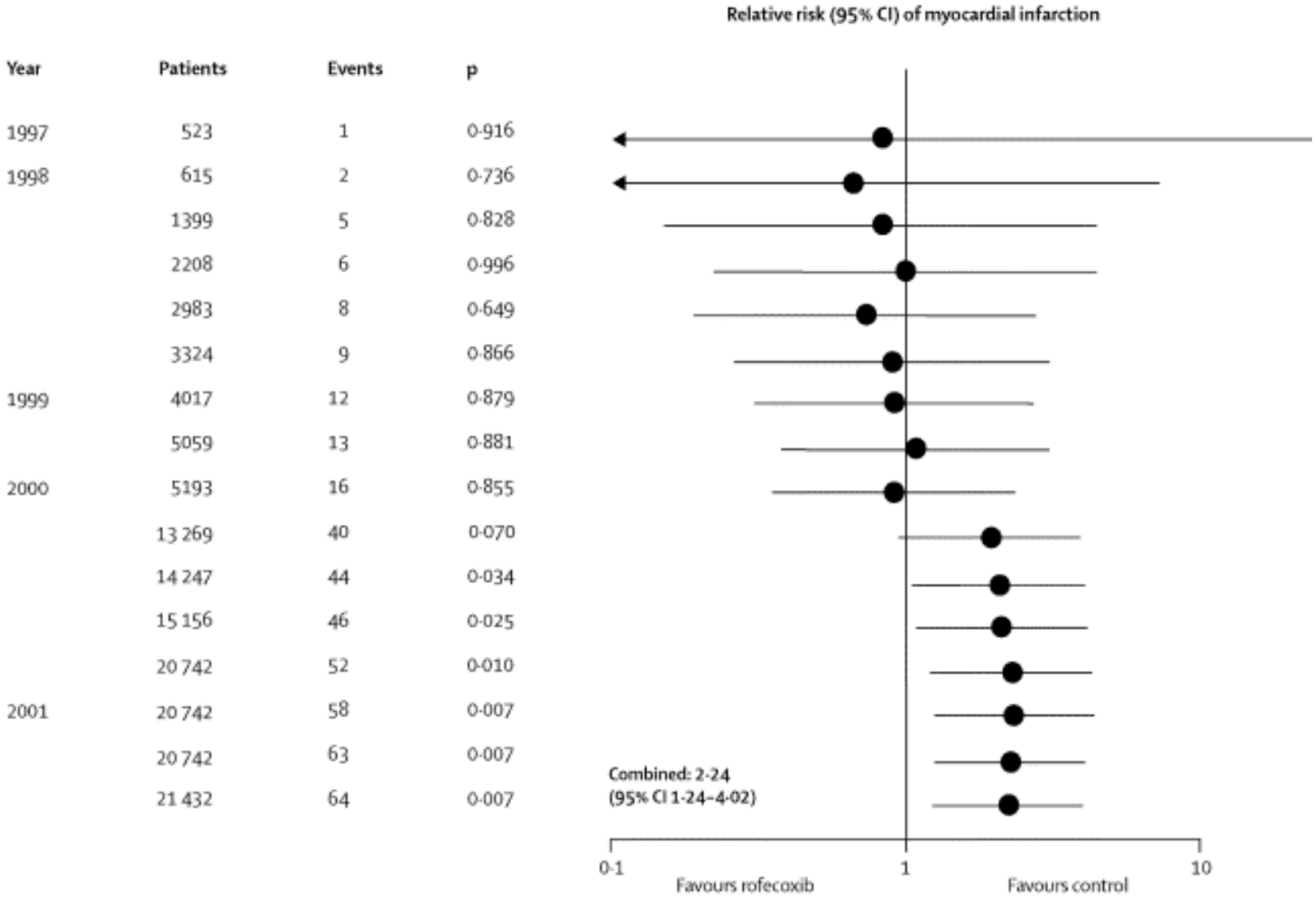
In short, the “new view” in many fields emphasizes **estimation over testing** -- And this is where we are headed!



Risk of cardiovascular events and rofecoxib: cumulative meta-analysis

Peter Jüni MD,Linda Nartey DipMD,Stephan Reichenbach MD,Rebekka Sterchi,Prof Paul A Dieppe MD,Prof Matthias Egger MD

The Lancet - 4 December 2004 (Vol. 364, Issue 9450, Pages 2021-2029)
DOI: 10.1016/S0140-6736(04)17514-4



Risk of cardiovascular events and rofecoxib: cumulative meta-analysis

Peter Jüni MD,Linda Nartey DipMD,Stephan Reichenbach MD,Rebekka Sterchi,Prof Paul A Dieppe MD,Prof Matthias Egger MD

The Lancet - 4 December 2004 (Vol. 364, Issue 9450, Pages 2021-2029)
DOI: 10.1016/S0140-6736(04)17514-4

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, redividing patients into treatment and control, or from the point of an initial study design, 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 tee-totum 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 planed a bar of mahogany squarely and

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

A MILLION
Random Digits

WITH

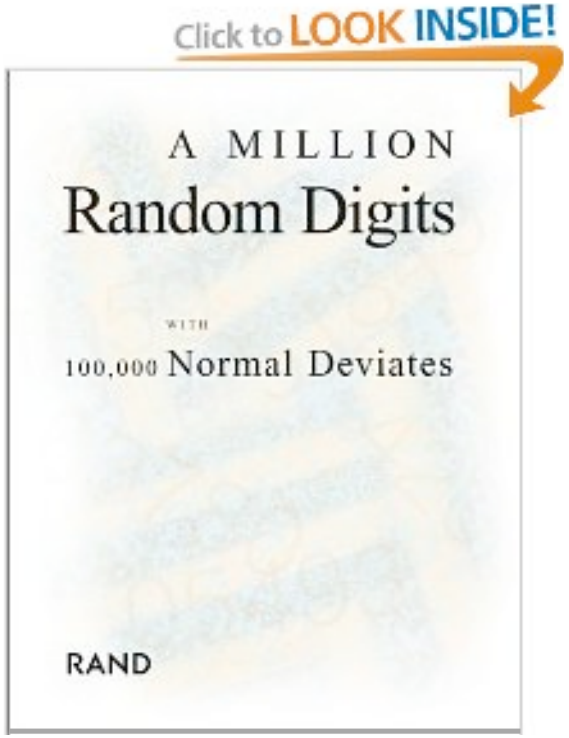
100,000 Normal Deviates

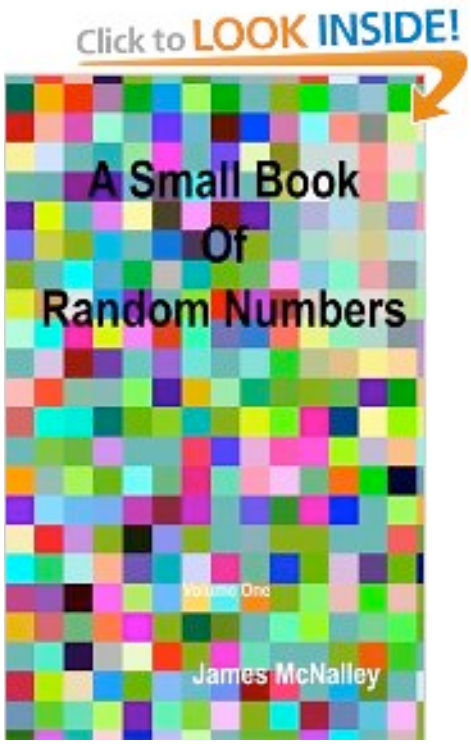
RAND

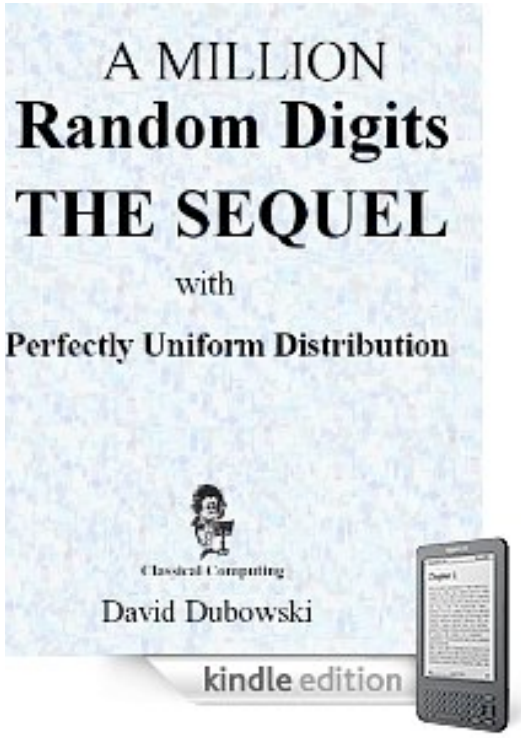
00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	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	64778	35808	34282	60935	20344	35273	88435
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	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	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	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
00026	15474	45266	95270	79953	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	52494	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	93910	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	66674	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	04401	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	64659
00052	73189	50207	47677	26269	62290	64464	27124	67018	41361	82760
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	20288	55210	29773	74287	75251	65344	67415
00062	21818	59313	93278	81757	05686	73156	07082	85046	31853	38452
00063	51474	66499	68107	23621	94049	91345	42836	09191	08007	45449
00064	99559	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	65584	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	08789	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	35655	06958	92983	05128	09719	77433	53783	92301	50498
00078	10850	62746	99599	10507	13499	06319	53075	71839	06410	19362
00079	39820	98952	43622	63147	64421	80814	43800	09351	31024	73167
00080	59580	06478	75569	78800	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	44606	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	78601	11883	09528	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	08816	47449
00092	19523	59515	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	94850	58838	73859	49364	73331	96240	43642
00096	24637	38736	74384	89342	52623	07992	12369	18601	03742	83873
00097	83080	12451	38992	22815	07759	51777	97377	27585	51972	37867
00098	16444	24334	36151	99073	27483	70939	85130	32552	54846	54759
00099	60790	18157	57178	65762	11161	78576	45819	52979	65130	04860

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	35905	14060	40619	29549	69616	33564	60780
00105	24122	66591	27699	06494	14845	46672	61958	77100	90899	75754
00106	61196	30231	92962	61773	41839	55382	17267	70943	78038	70267
00107	30532	21704	10274	12202	39665	23309	10061	68829	55986	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	34994	41374	70071	14736	09958	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	48355	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	08337	56303	87315	16520	69676	11654	99893	02181	68161
00124	57275	36898	81304	48585	68652	27376	92852	55866	88448	03584
00125	20857	73156	70284	24326	79375	95220	01159	63267	10622	48391
00126	15633	84924	90415	93614	33521	26665	55823	47641	86225	31704
00127	92694	48297	39904	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	71857	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	06345	80854	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	88627	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	35482	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
James McNalley


A MILLION
Random Digits
THE SEQUEL
with
Perfectly Uniform Distribution
David Dubowski
kindle edition

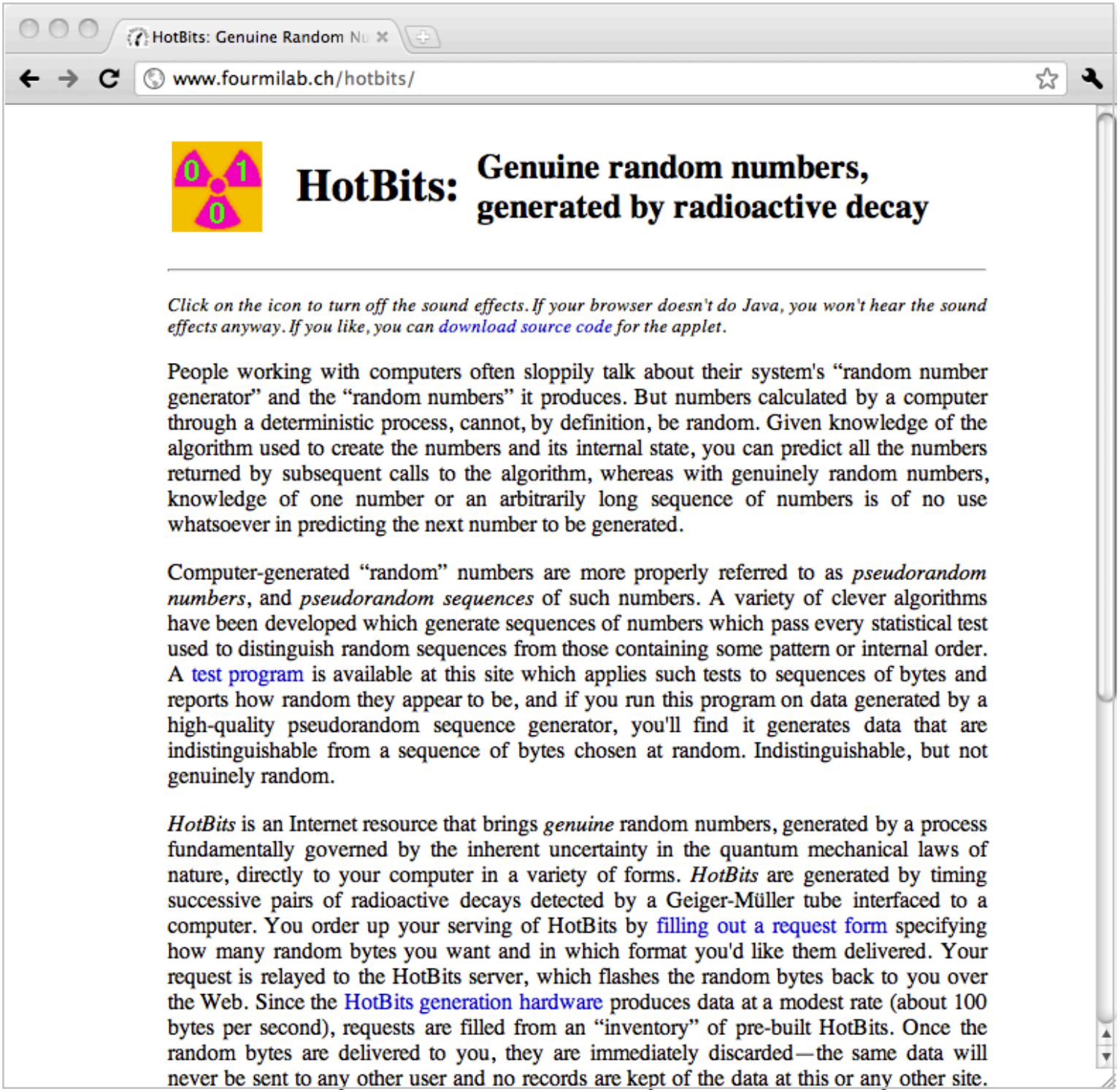
Random number generation

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 **pseduo-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)

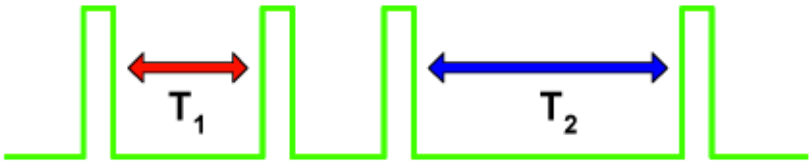

HotBits: Genuine Random Nu
www.fourmilab.ch/hotbits/
HotBits: Genuine random numbers, generated by radioactive decay
Click on the icon to turn off the sound effects. If your browser doesn't do Java, you won't hear the sound effects anyway. If you like, you can download source code for the applet.
People 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.

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) -- `random.org` uses atmospheric noise, suitably filtered, to accomplish the same task...


RANDOM.ORG – True Rando
www.random.org
Home Games Numbers Lists & More Drawings Web Tools Statistics Testimonials Learn More Login
RANDOM.ORG
Search RANDOM.ORG
Search
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 a 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.
True Random Number Generator
Min: 1
Max: 100
Generate
Result:
Powered by RANDOM.ORG
Like RANDOM.ORG?
Get the Newsletter
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

Another service

`random.org` also provides a service that generates random numbers on command -- They make their data available via **a public API** (application programming interface)

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...

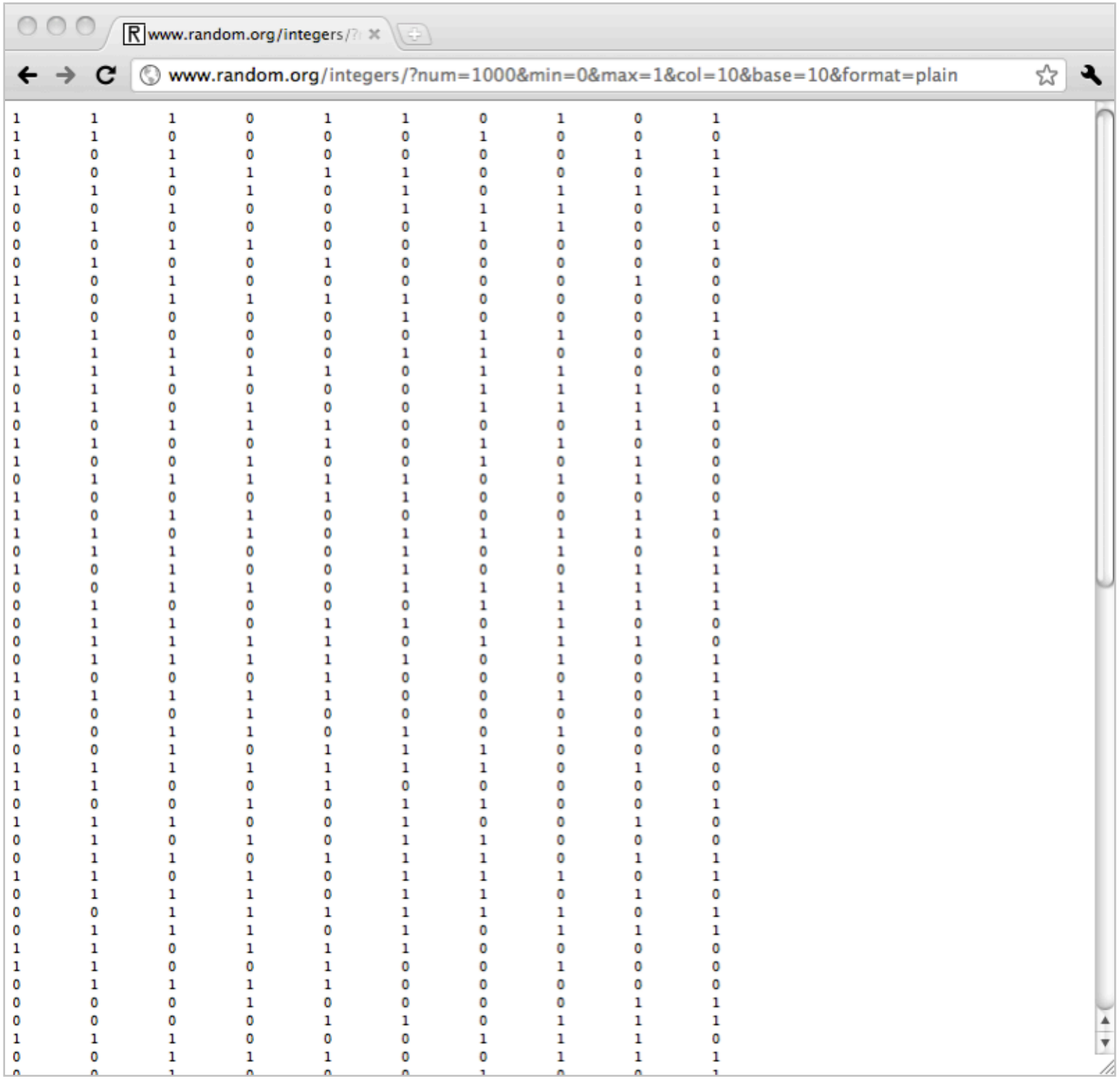

RANDOM.ORG – HTTP Interf
www.random.org/clients/http/
Home Games Numbers Lists & More Drawings Web Tools Statistics Testimonials Learn More Login
RANDOM.ORG
Search RANDOM.ORG
Search
True Random Number Service
HTTP Interface Description
RANDOM.ORG is a true random number service that generates randomness 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 errrors (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 text/html), which will display well in a browser but which is somewhat cumbersome to parse. If **plain** is specified, the server produces as minimalistic document of type plain text (MIME type text/plain) 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 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., YYYY-MM-DD) or one of the two shorthand strings **today** or **yesterday**.

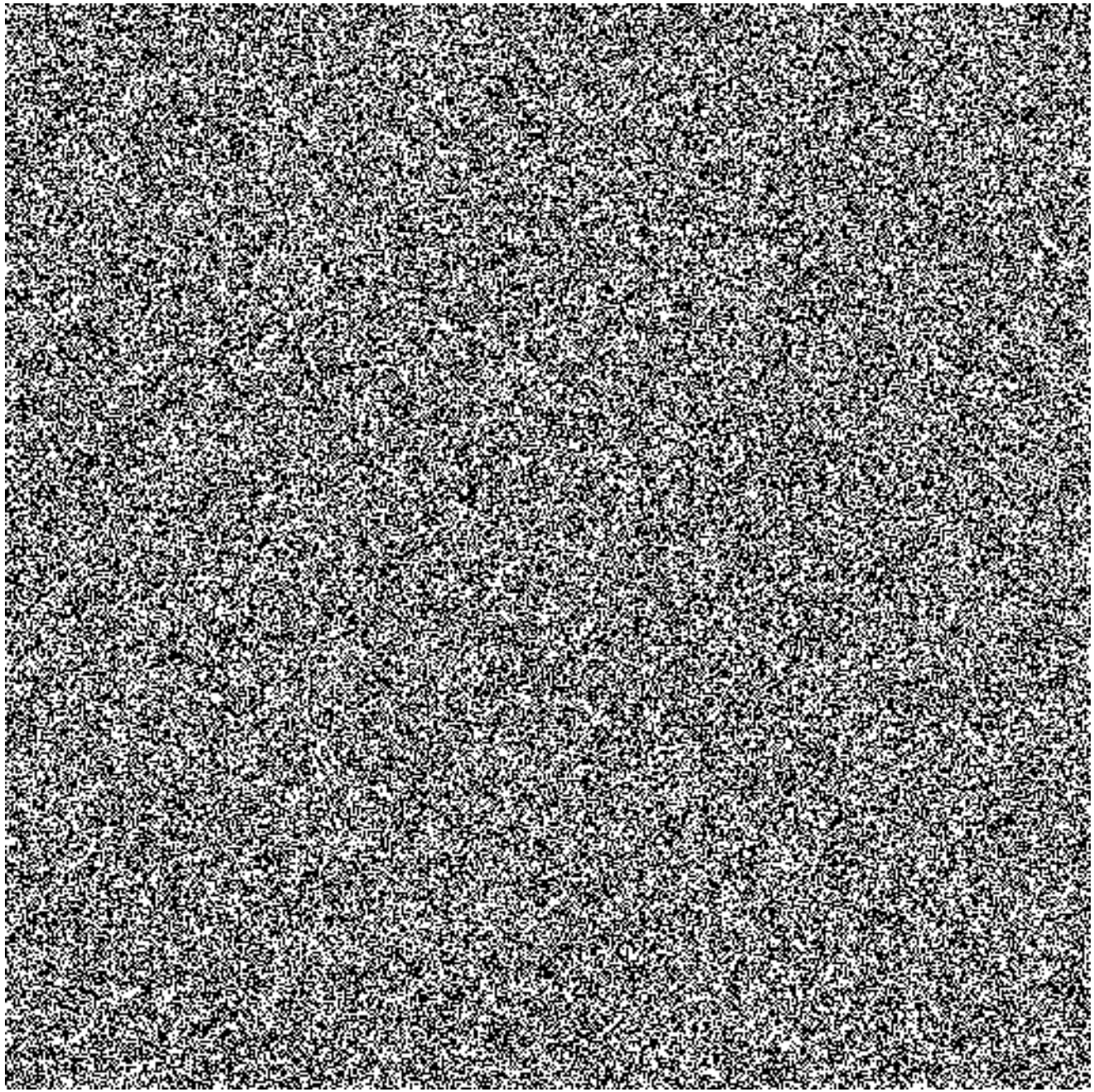

www.random.org/integers/
www.random.org/integers/?num=1000&min=0&max=1&col=10&base=10&format=plain
1 1 1 0 1 1 0 1 0 1
1 1 0 0 0 0 1 0 0 0
1 0 1 0 0 0 0 0 1 1
0 0 1 1 1 1 0 0 0 1
1 1 0 1 0 1 0 1 1 1
0 0 1 0 0 1 1 1 0 1
0 1 0 0 0 0 1 1 0 0
0 0 1 1 0 0 0 0 0 1
0 1 0 0 1 0 0 0 0 0
1 0 1 0 0 0 0 0 1 0
1 0 1 1 1 1 0 0 0 0
1 0 0 0 0 1 0 0 0 1
0 1 0 0 0 0 1 1 0 1
1 1 1 0 0 1 1 0 0 0
1 1 1 1 1 0 1 1 0 0
0 1 0 0 0 0 1 1 1 0
1 1 0 1 0 0 1 1 1 1
0 0 1 1 1 0 0 0 1 0
1 1 0 0 1 0 1 1 0 0
1 0 0 1 0 0 1 0 1 0
0 1 1 1 1 1 0 1 1 0
1 0 0 0 1 1 0 0 0 0
1 0 1 1 0 0 0 0 1 1
1 1 0 1 0 1 1 1 1 0
0 1 1 0 0 1 0 1 0 1
1 0 1 0 0 1 0 0 1 1
0 0 1 1 0 1 1 1 1 1
0 1 0 0 0 0 1 1 1 1
0 1 1 0 1 1 0 1 0 0
0 1 1 1 1 0 1 1 1 0
0 1 1 1 1 1 0 1 0 1
1 0 0 0 1 0 0 0 0 1
1 1 1 1 1 0 0 1 0 1
0 0 0 1 0 0 0 0 0 1
1 0 1 1 0 1 0 1 0 0
0 0 1 0 1 1 1 0 0 0
1 1 1 1 1 1 1 0 1 0
1 1 0 0 1 0 0 0 0 0
0 0 0 1 0 1 1 0 0 1
1 1 1 0 0 1 0 0 1 0
0 1 0 1 0 1 1 0 0 0
0 1 1 0 1 1 1 0 1 1
1 1 0 1 0 1 1 1 0 1
0 1 1 1 0 1 1 0 1 0
0 0 1 1 1 1 1 1 0 1
0 1 1 1 0 1 0 1 1 1
1 1 0 1 1 1 0 0 0 0
1 1 0 0 1 0 0 1 0 0
0 1 1 1 1 0 0 0 0 0
0 0 0 1 0 0 0 0 1 1
0 0 0 0 1 1 0 1 1 1
1 1 1 0 0 0 1 1 1 0
0 0 1 1 1 0 0 1 1 1

Testing?

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*512 = 262,144 random bits from `random.org` and displayed them as an image...



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 does the work -- 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)
```

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 unspecified, but our choice (not the `seed`, but the others), has good properties relative to the statistical tests I alluded to)

The result are pseudo-random numbers in the interval [0,1] and they are used anywhere we might need observations from the so-called **uniform distribution** on that interval

Uniform distribution

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 100 such observations, we'd expect about 50 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 25 observations of the 100 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*512 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**...

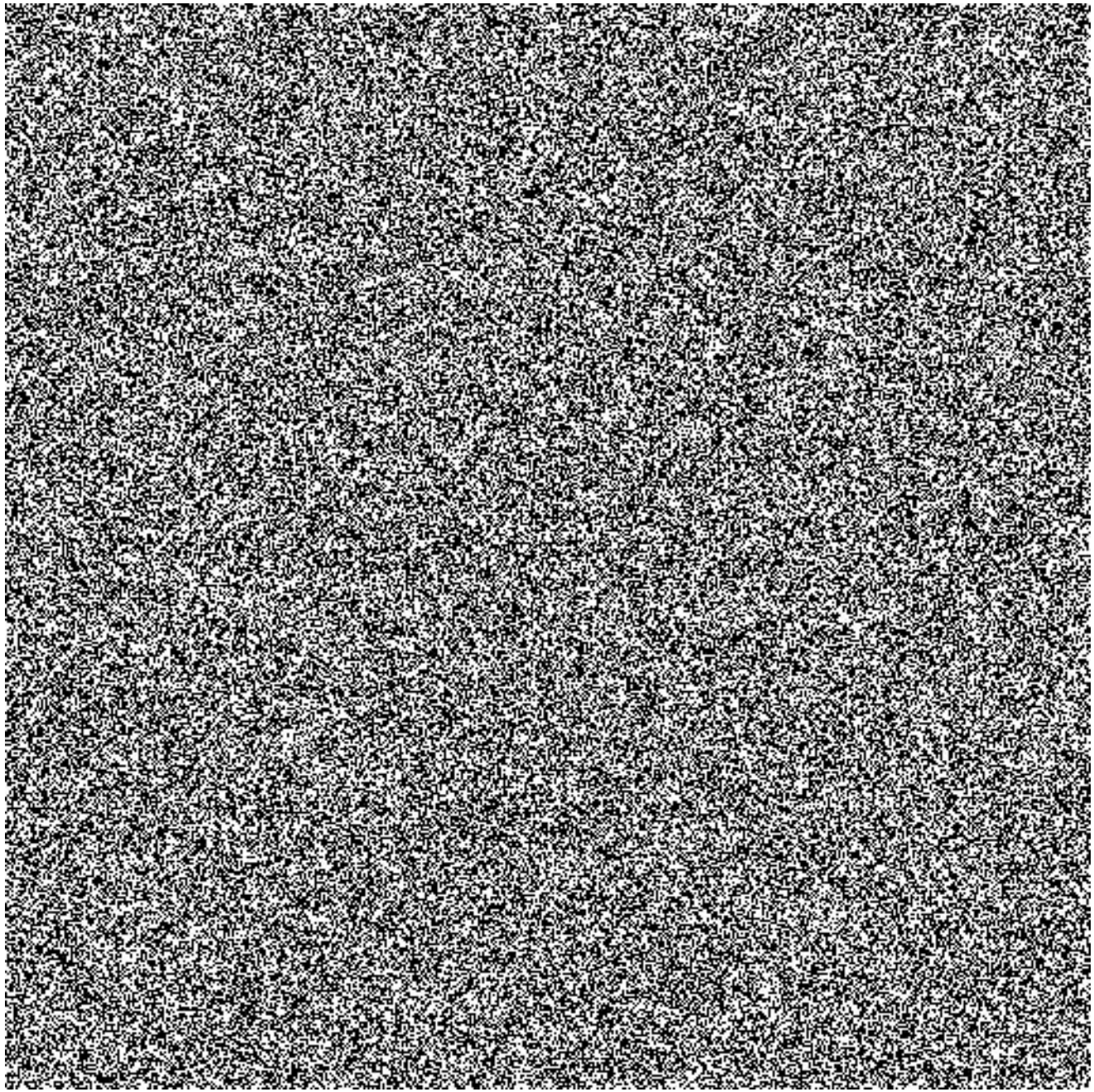
Generating random numbers

Here is the sequence of numbers we get values

```
 [1]  332812214 1527463910 1066419132  407833662 1840039657 1781998399
 [7] 1240150931 1887903182  917895449 1693774942  232225562 1037233935
[13] 1665982846 1281903136 1390060048  264631023  214970624  943783314
[19]  851941656 1309937843  122978257 1016296985 1965982304 1081190586
[25] 1711041635  523242468  190625211 1939803600 1329860093 2096268722
```

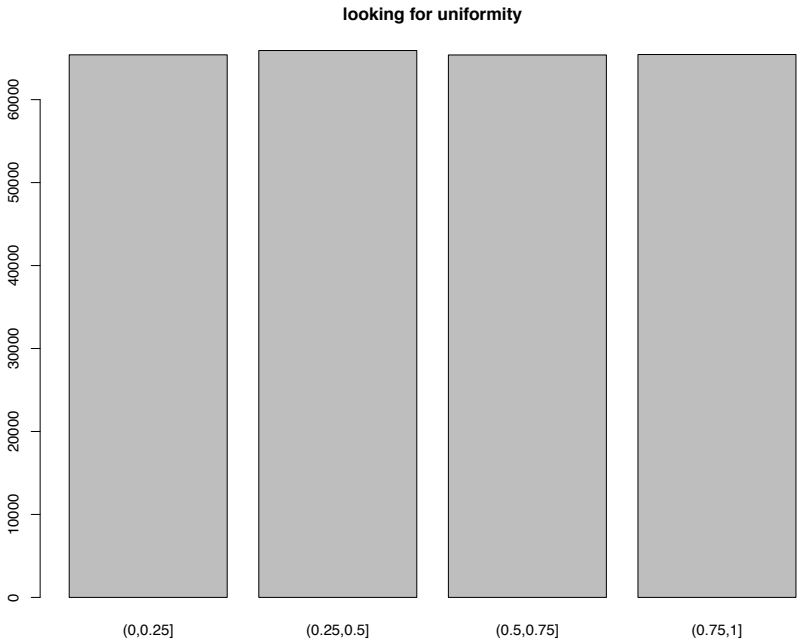
which when divided by `2147483647 = 2^31-1` gives the pseudo-random uniform observations

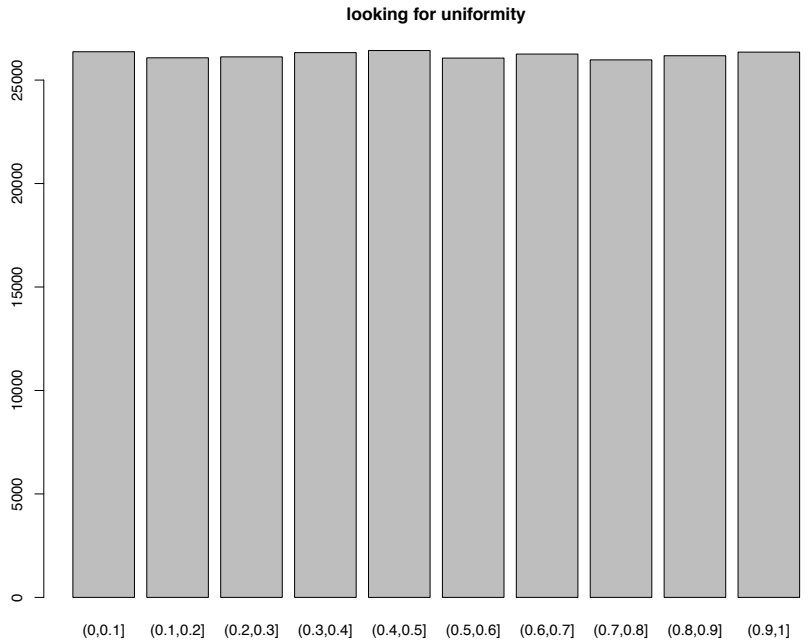
```
 [1] 1.549778e-01 7.112808e-01 4.965901e-01 1.899123e-01 8.568352e-01
 [6] 8.298077e-01 5.774903e-01 8.791234e-01 4.274284e-01 7.887254e-01
[11] 1.081385e-01 4.829997e-01 7.757837e-01 5.969327e-01 6.472972e-01
[16] 1.232284e-01 1.001035e-01 4.394834e-01 3.967162e-01 6.099873e-01
[21] 5.726621e-02 4.732502e-01 9.154819e-01 5.034686e-01 7.967659e-01
[26] 2.436538e-01 8.876678e-02 9.032914e-01 6.192644e-01 9.761512e-01
```



Testing?

How about looking at other intervals? Here we have a barplot for 4 and 10 intervals using the same 512*512 observations we created the bit-image for on the previous page





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)

Testing?

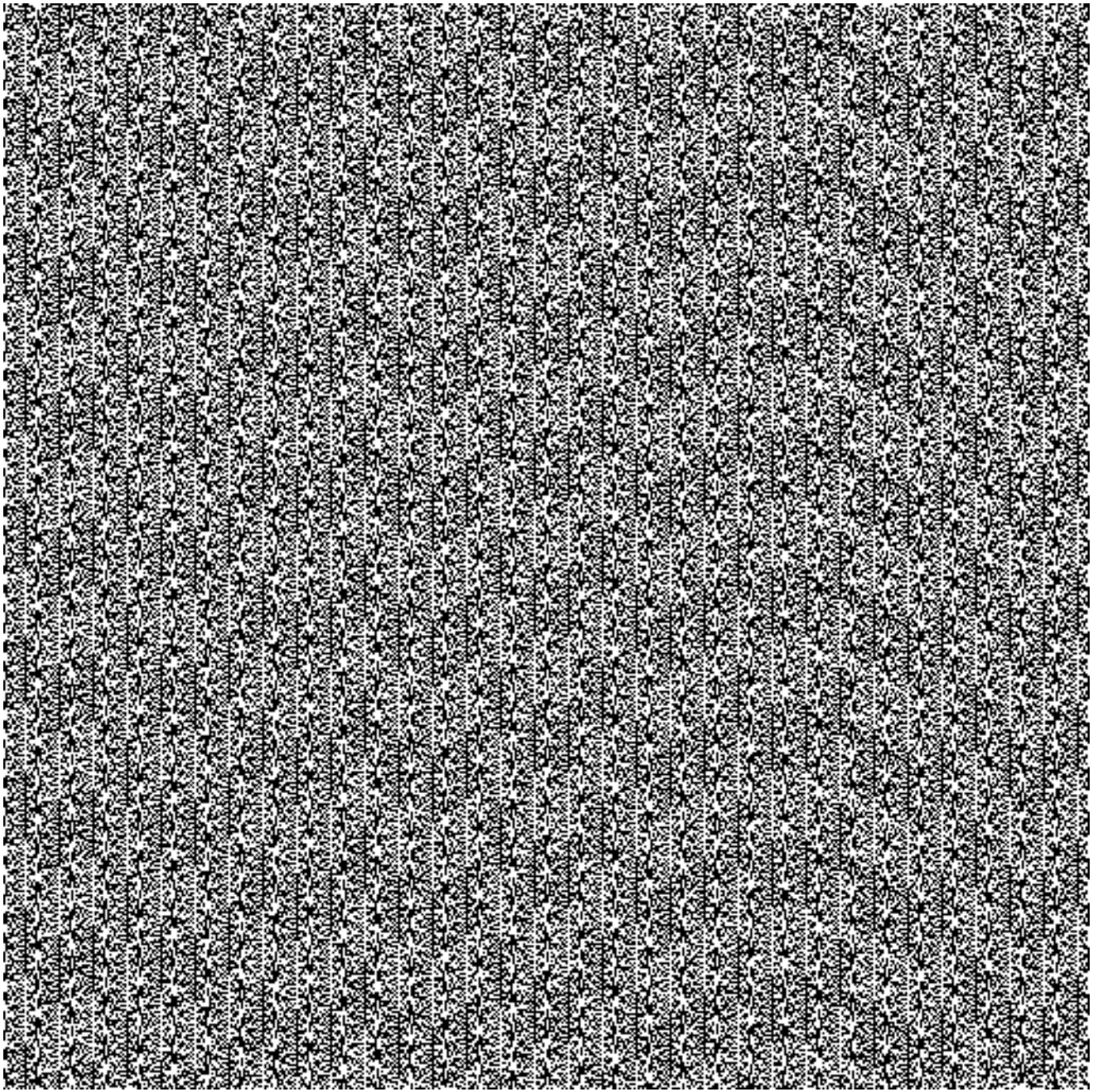
Formally, there are a set of **classical statistical tests (yes, tests of hypothesis!) that could help us assess if a random number generator (true or otherwise) is performing as expected**

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 or the length of "runs" of bits of the same kind**

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?!

And so while the simulation we've been doing is fantastic pedagogically, we do have to cover a little probability in upcoming lectures -- At least enough to cover the basics of coin tossing

As an aside, **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...

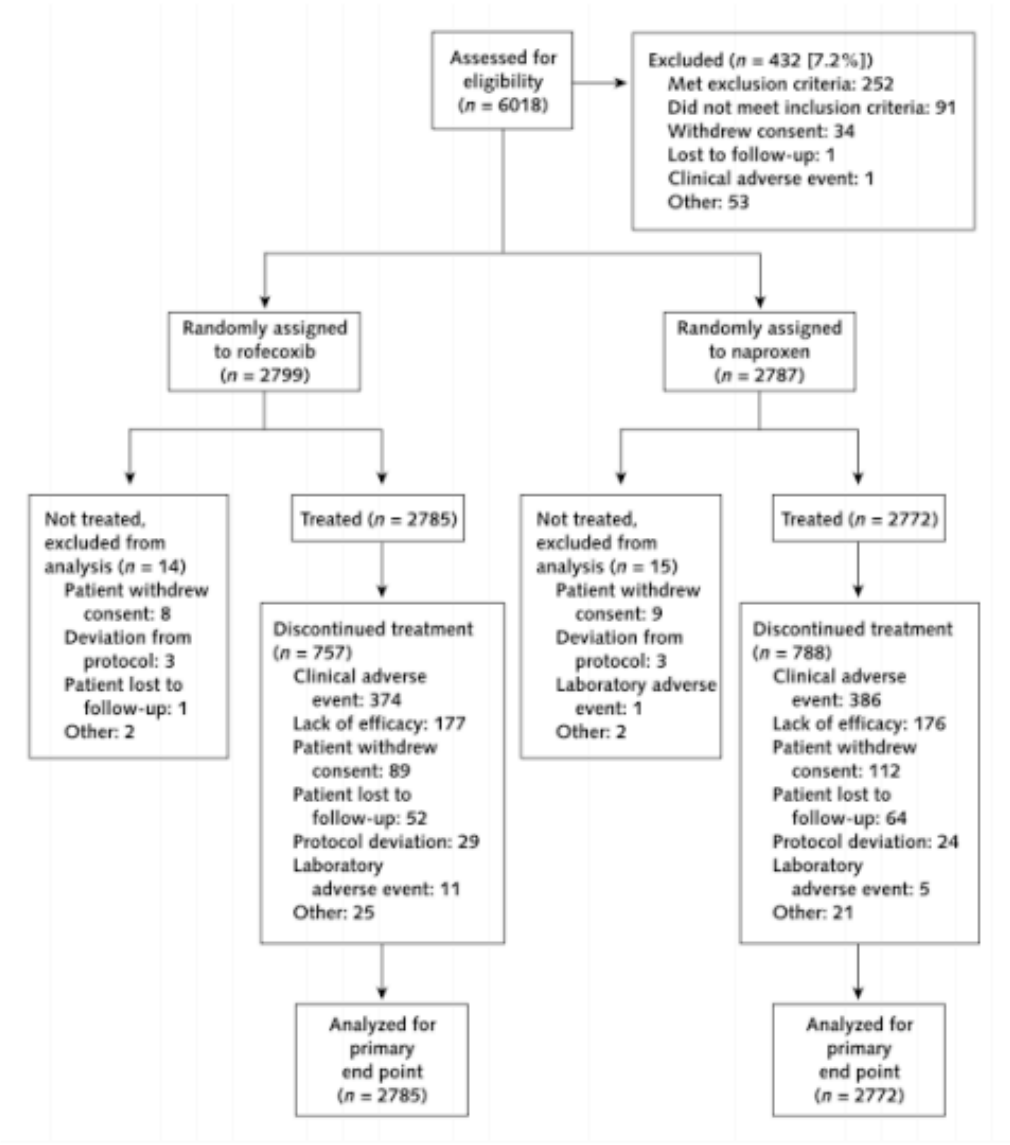


Other randomized experiments

So far, we have examined mainly experimental setups in health studies; the randomized controlled trial is a big deal and nicely highlights many of the features of hypothesis testing

But randomization in experimentation is extremely common; as we mentioned previously, Fisher was an aggressive advocate of its use in general

We'll now consider a more modern application of randomization, in experiments that "optimize" the operation of web sites

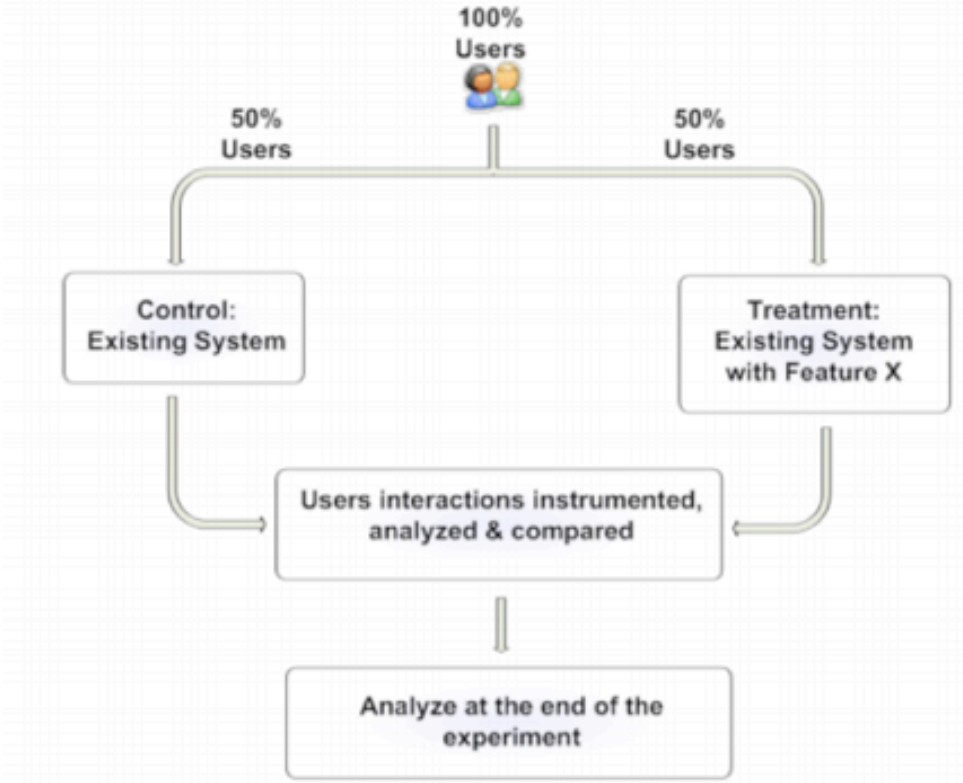


A/B testing

At the right we present a diagram for conducting an experiment on a web site; the structure is not that dissimilar from that for the ADVANTAGE trial (minus all the comments about "exclusions" and "discontinued" treatments

In an A/B test, visitors are presented with two different versions of the site (in some cases we literally have a treatment and control); sometimes the differences have to do with **the content on the page**, while in other cases they have to do with **the location of or visual arrangement of the content**

The outcomes that are measured depend on the site's overall objectives; an informational site might be interested in **the number of pages visitors read**, while commerce sites would be concerned with the **number of transactions made**

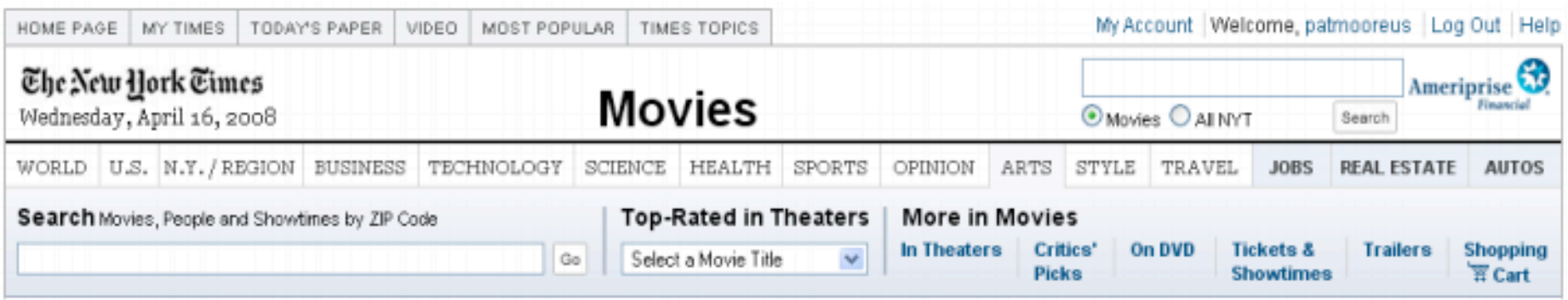


Kohavi, Henne and Sommerfield, *"Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO,"* KDD 2007

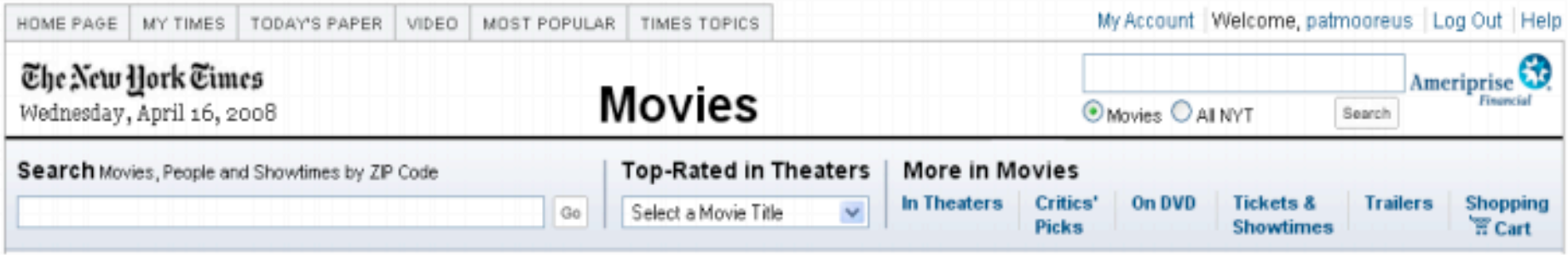
An example

Here is a simple experiment conducted by the New York Times web site (`nytimes.com`) in 2008 -- What is the difference between these two pages and what differences in visits might you be interested in comparing?

A: Control



B: Test

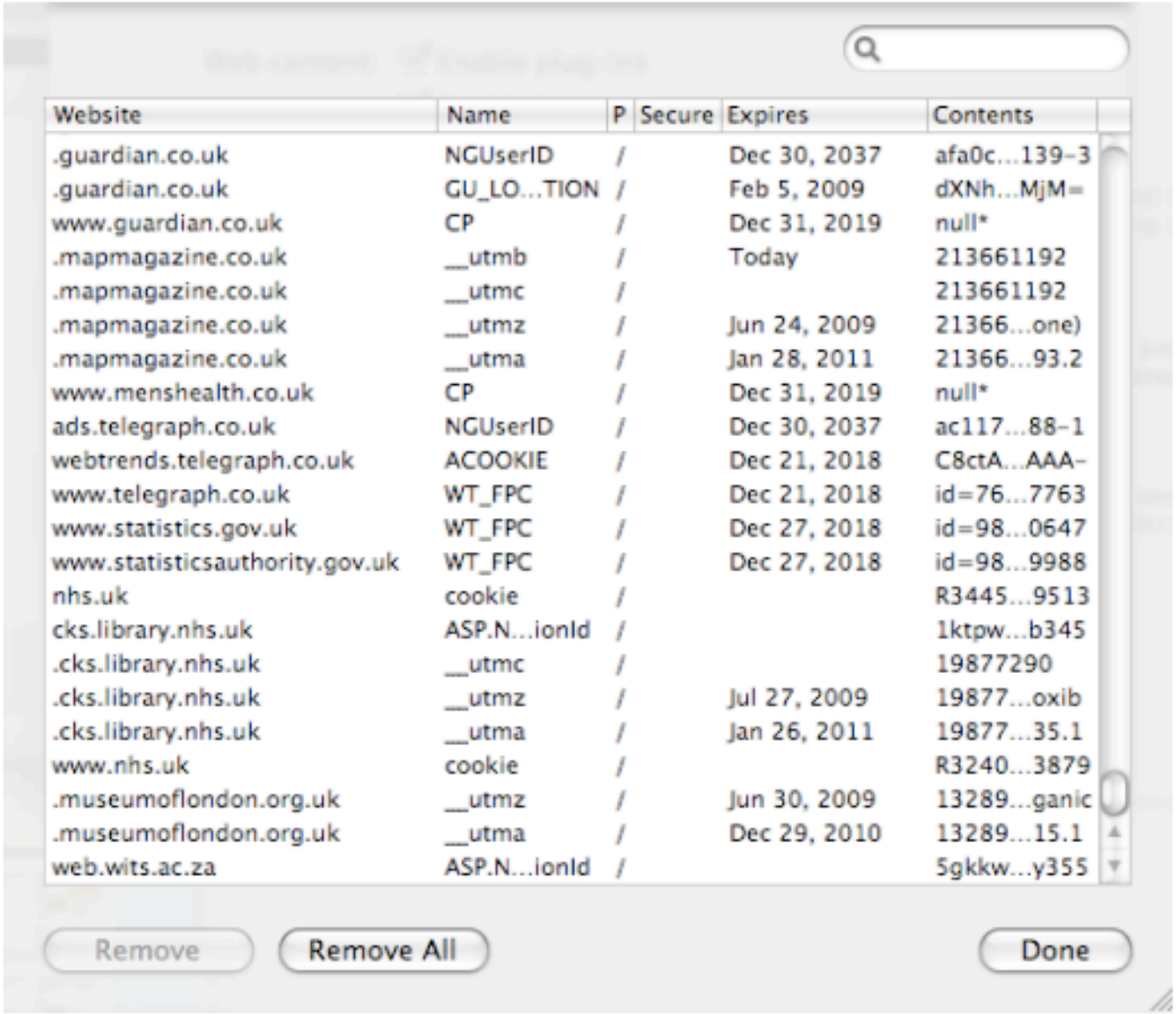


A/B testing

At a technical level (technical meaning from the standpoint of web servers), an A/B test means that visitors are assigned at random to group A or B (or C or D or E, depending on how many changes are being tested at one time)

Visitors assigned to group B, for example, will **see one version of the site consistently during the experimental period**; in the case of then Times example on the previous slide, it means that whenever a visitor in group B requests pages from The Movie Section, those pages are missing the middle navigation bar

What information does a web site need to know to make that happen? How is that information gathered?



Website	Name	P	Secure	Expires	Contents
.guardian.co.uk	NGUserID	/		Dec 30, 2037	afa0c…139-3
.guardian.co.uk	GU_LO…TION	/		Feb 5, 2009	dXNh…MjM=
www.guardian.co.uk	CP	/		Dec 31, 2019	null*
.mapmagazine.co.uk	__utmb	/		Today	213661192
.mapmagazine.co.uk	__utmc	/			213661192
.mapmagazine.co.uk	__utmz	/		Jun 24, 2009	21366…one)
.mapmagazine.co.uk	__utma	/		Jan 28, 2011	21366…93.2
www.menshealth.co.uk	CP	/		Dec 31, 2019	null*
ads.telegraph.co.uk	NGUserID	/		Dec 30, 2037	ac117…88-1
webtrends.telegraph.co.uk	ACOOKIE	/		Dec 21, 2018	C8ctA…AAA-
www.telegraph.co.uk	WT_FPC	/		Dec 21, 2018	id=76…7763
www.statistics.gov.uk	WT_FPC	/		Dec 27, 2018	id=98…0647
www.statisticsauthority.gov.uk	WT_FPC	/		Dec 27, 2018	id=98…9988
nhs.uk	cookie	/			R3445…9513
cks.library.nhs.uk	ASP.N…ionId	/			1ktpw…b345
.cks.library.nhs.uk	__utmc	/			19877290
.cks.library.nhs.uk	__utmz	/		Jul 27, 2009	19877…oxib
.cks.library.nhs.uk	__utma	/		Jan 26, 2011	19877…35.1
www.nhs.uk	cookie	/			R3240…3879
.museumoflondon.org.uk	__utmz	/		Jun 30, 2009	13289…ganic
.museumoflondon.org.uk	__utma	/		Dec 29, 2010	13289…15.1
web.wits.ac.za	ASP.N…ionId	/			5gkkw…y355

Remove | Remove All | Done

... the cookie jar

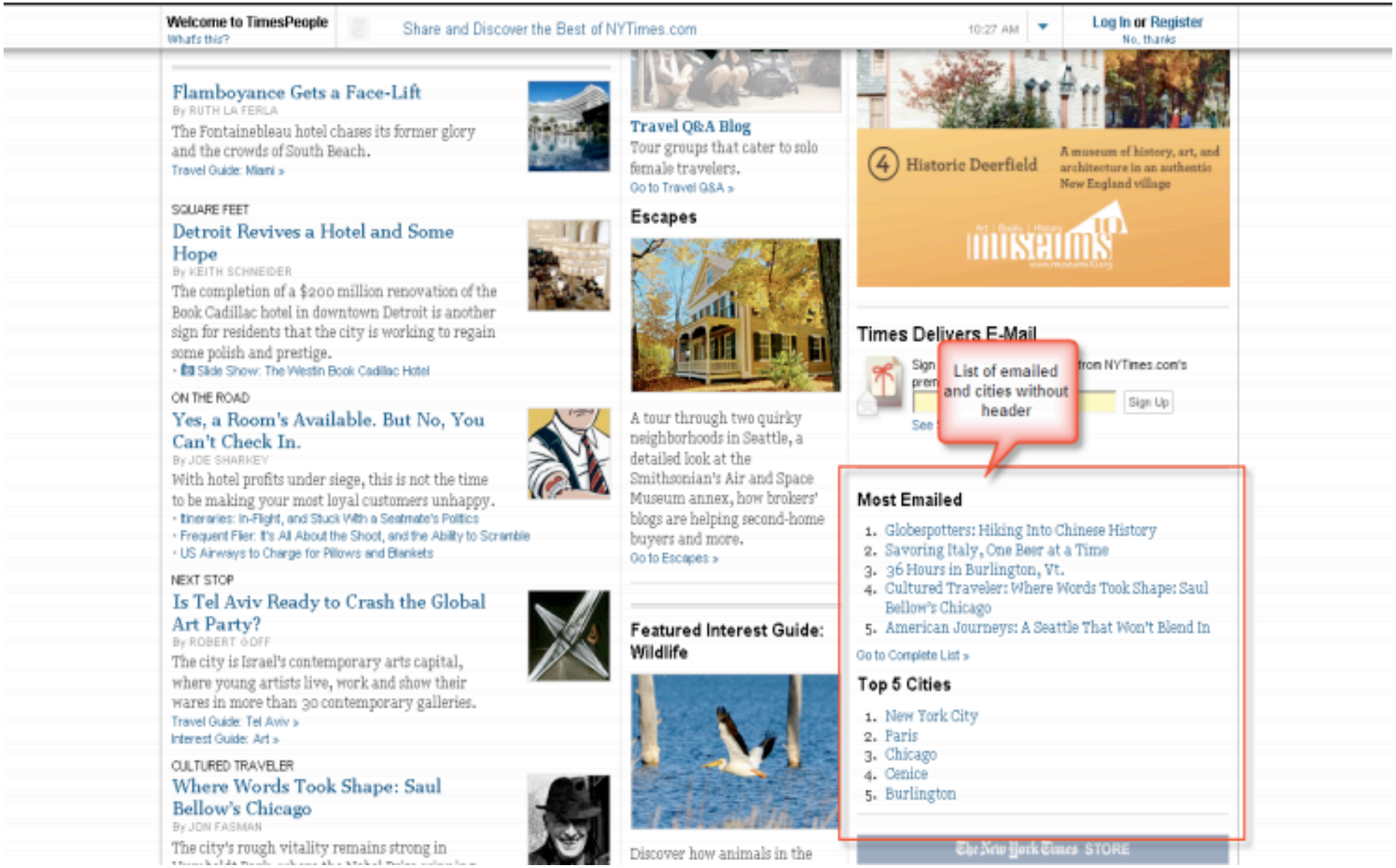
An experiment at `nytimes.com`

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

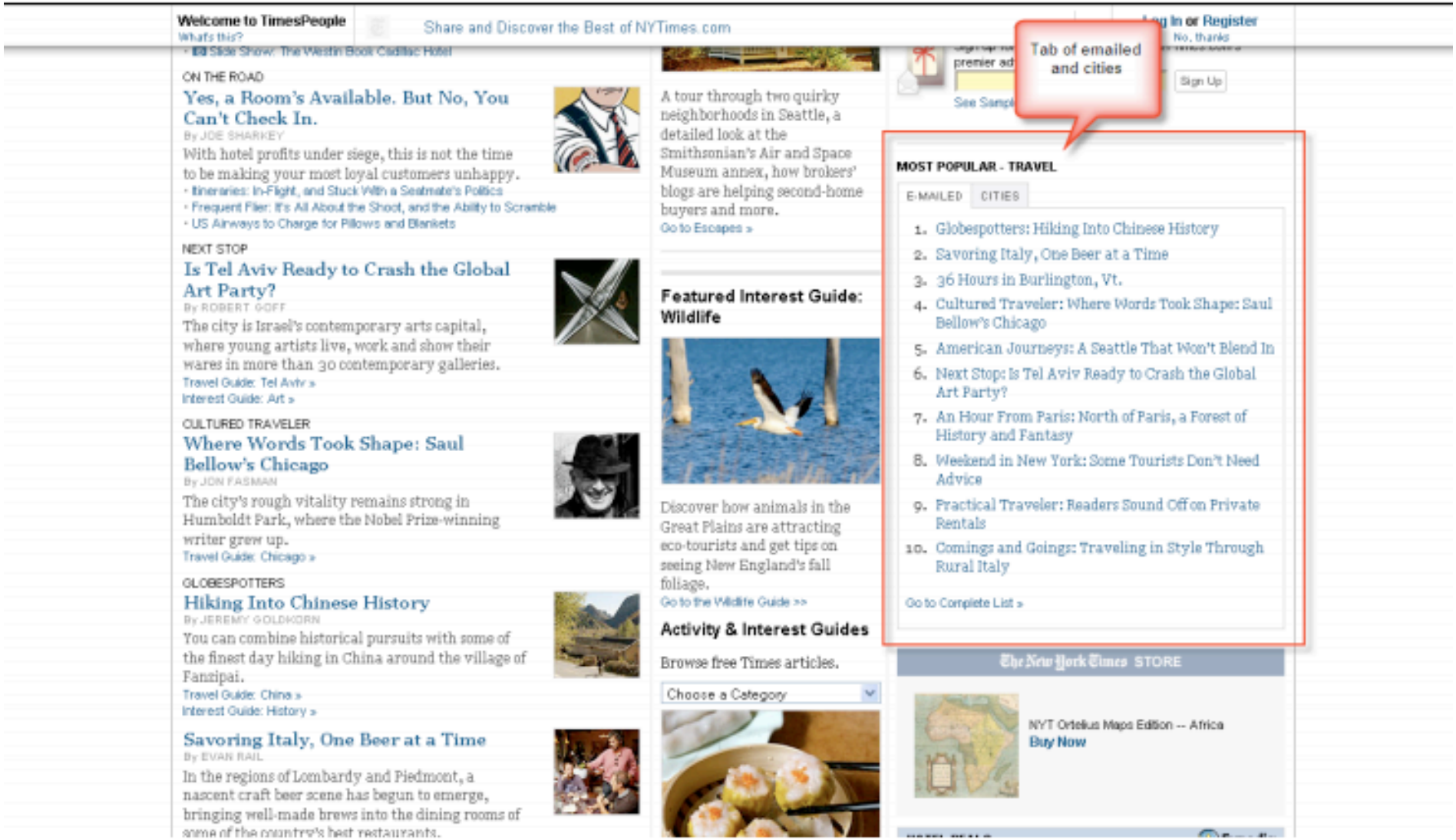
On the next two slides, we present samples of the A and B pages; the changes applied to all pages in The Travel Section, so **as a visitor browsed the site, they would consistently see either A or B**

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

List: Variation 10858



Tabs: Variation 10859



An experiment at `nytimes.com`

The Travel Section experiment lasted **five weeks**; the first visitor during the experiment arrived on Tuesday November 18 of 2008 at 8:40 am PST 2008, and the experiment closed with a final visit on Monday December 29 at about 8 pm PST

During that period, **nearly 200K visits were recorded as part of the experiment** (the experiment involved just a fraction of their overall traffic)

Given the nature of the change between A and B, and some of the objectives you wanted to examine, what sort of information would we like to collect about each visit? About each visitor?

The variables

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

`User_ID` A unique number for each visitor

`UserVisit_ID` A unique number for each visit

`StartTime_SSE` Unix time for the start of the visit

`StartTime_English` A more human readable version of the time

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

`Variation` The version of the page they received

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

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

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

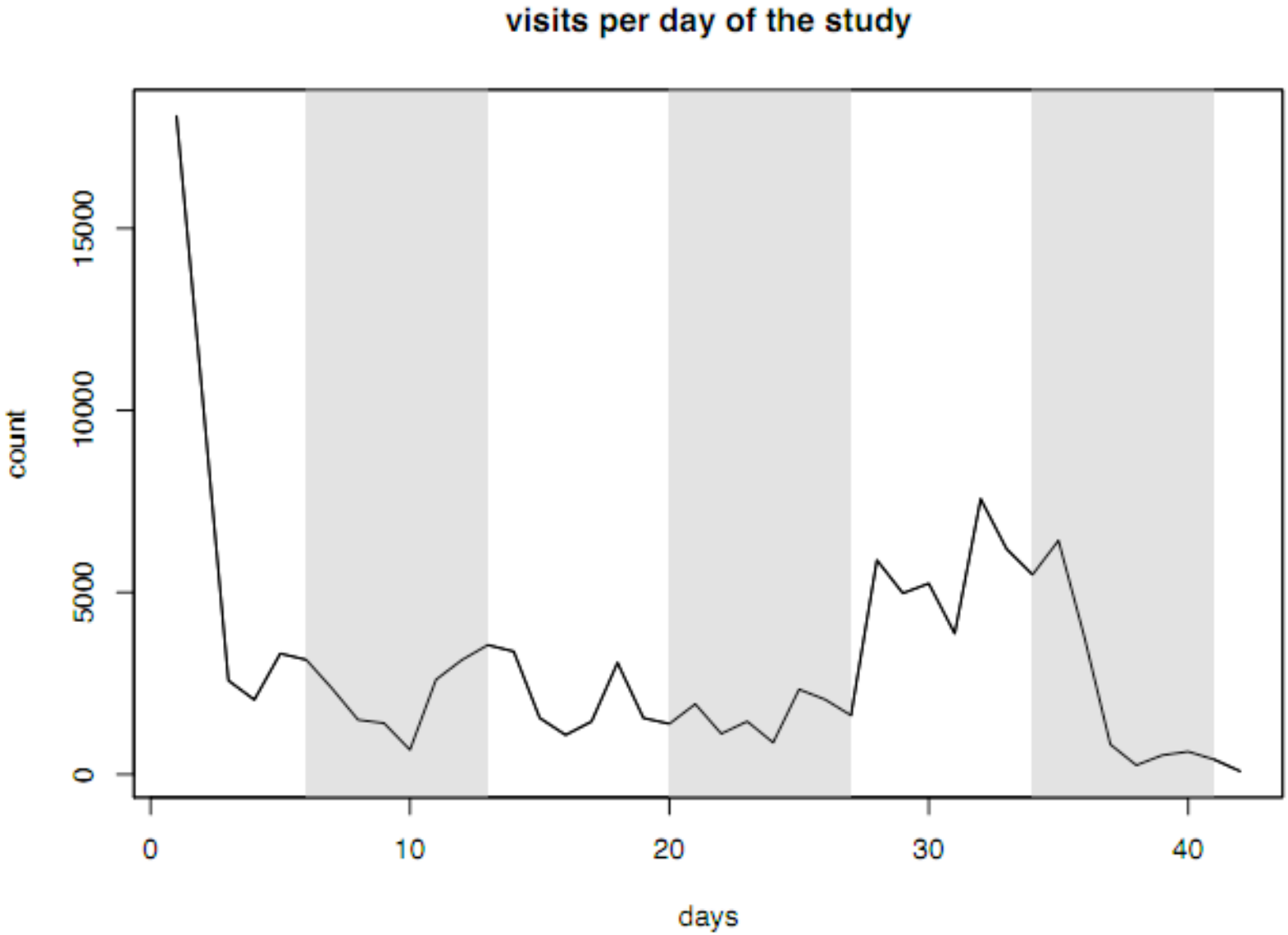
`TotalVisits` The total number of visits to the site

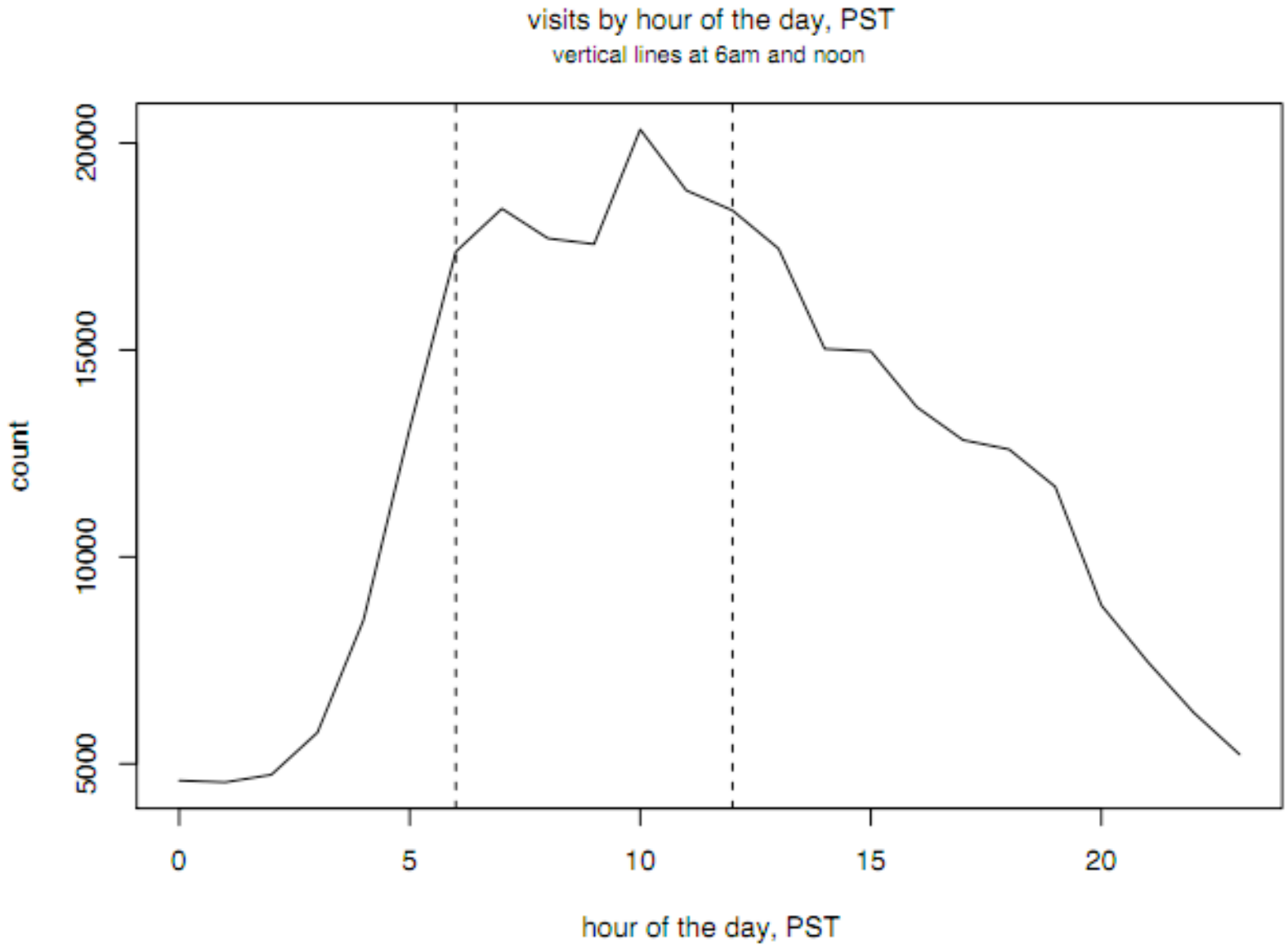
`TimeSinceFirstVisit` `(days)` How long it had been since their last visit

`UserAgent` Their browser

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

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

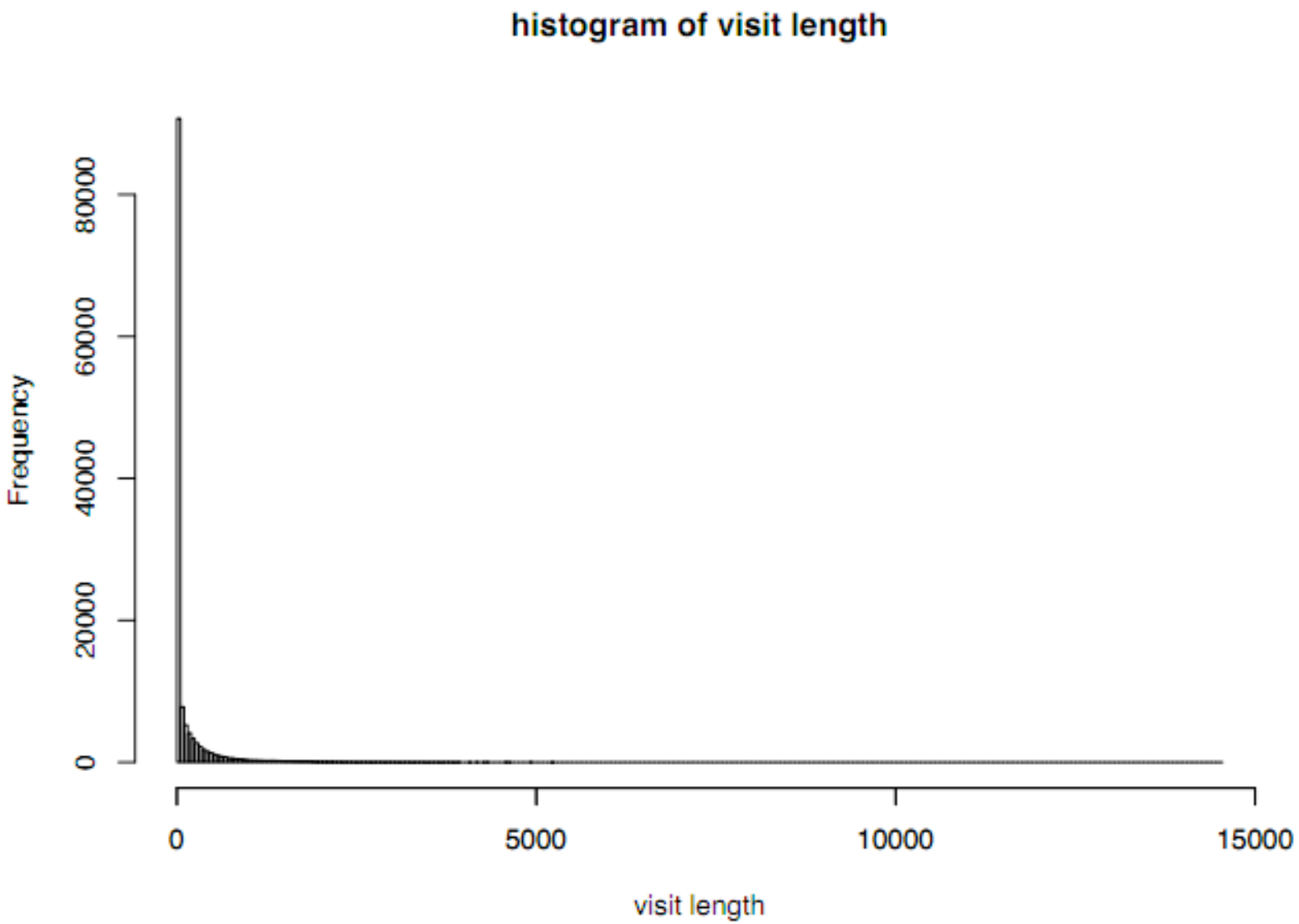

visits per day of the study
count
0
5000
10000
15000
0
10
20
30
40
days

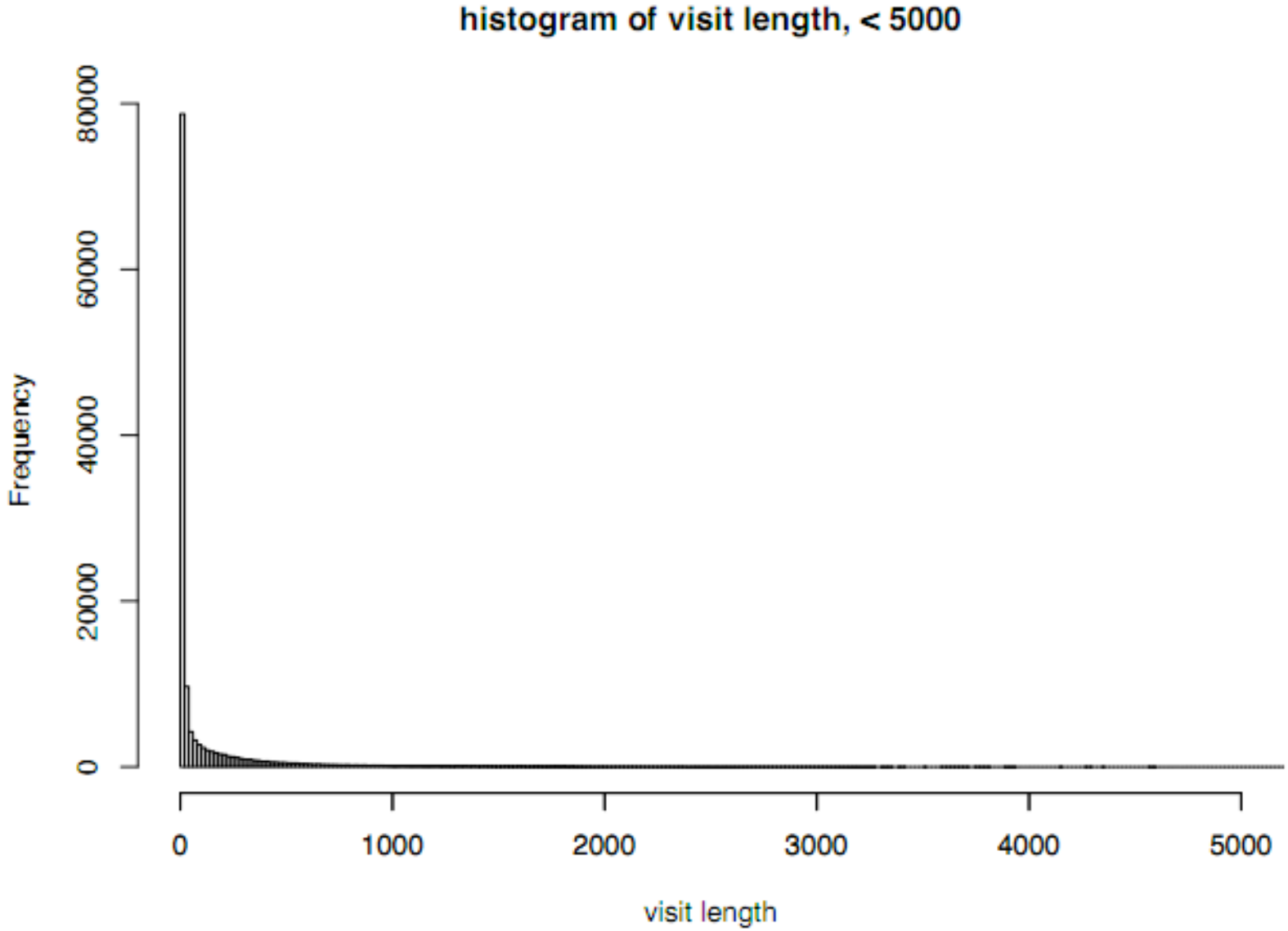

visits by hour of the day, PST
vertical lines at 6am and noon
count
5000
10000
15000
20000
0
5
10
15
20
hour of the day, PST

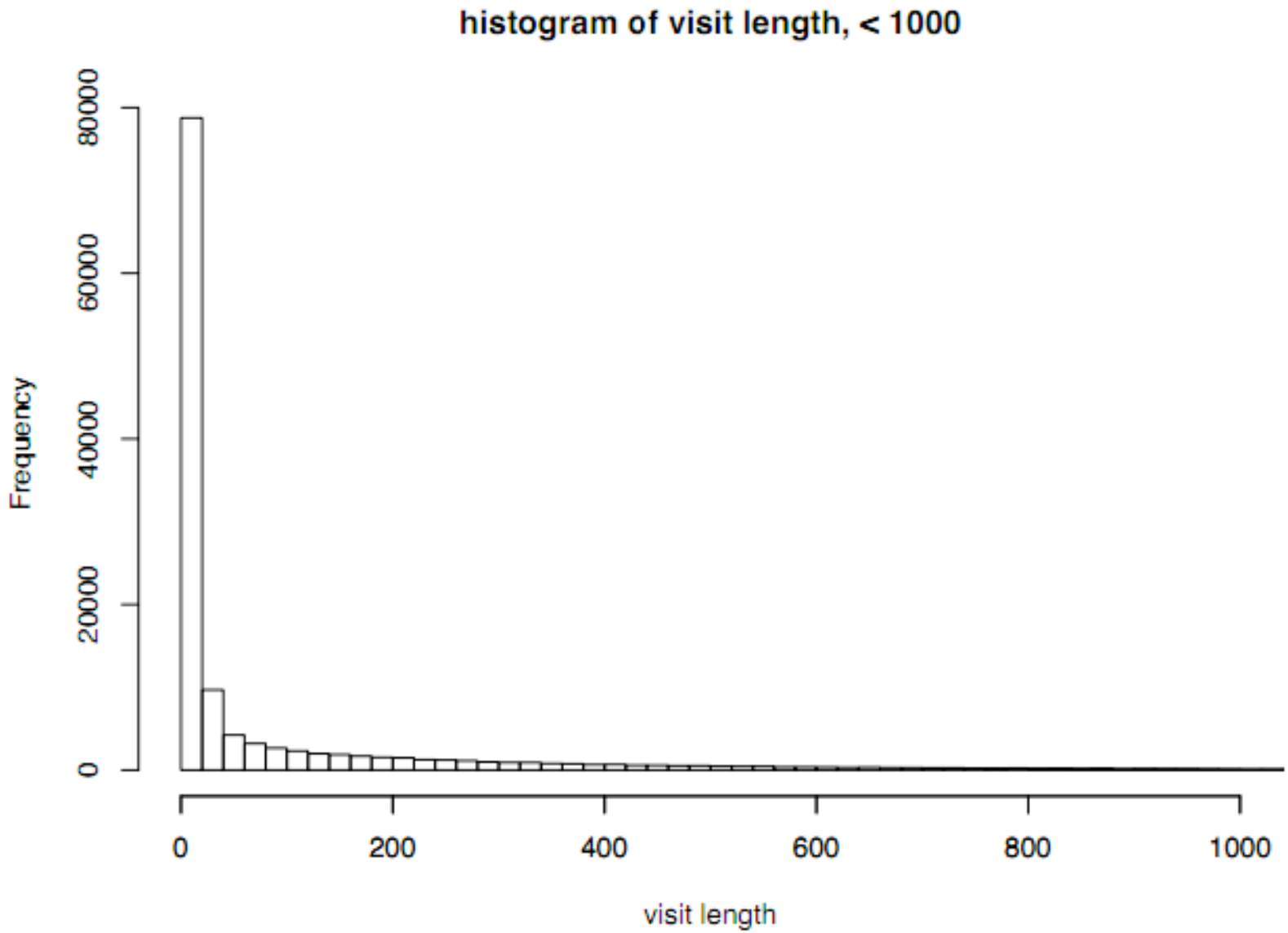
The data

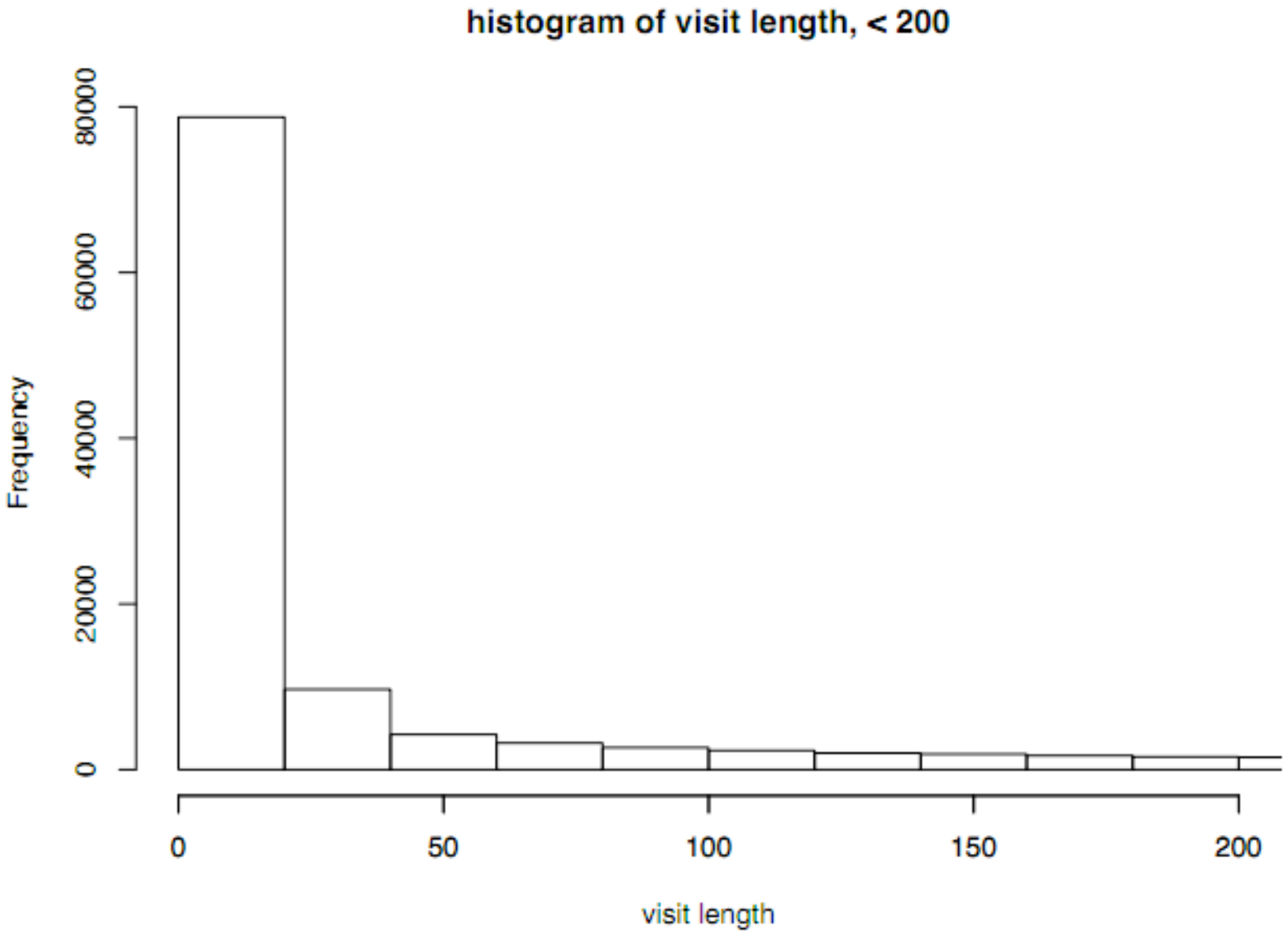
An obvious study metric here is how long people spend on the site -- So, let's have a look at visit lengths

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


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


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


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


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

Transformations

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!

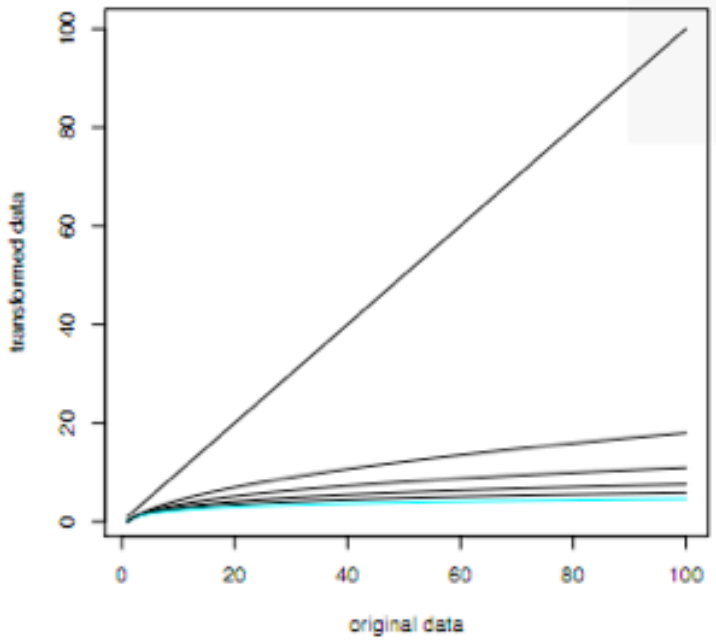
Transformations

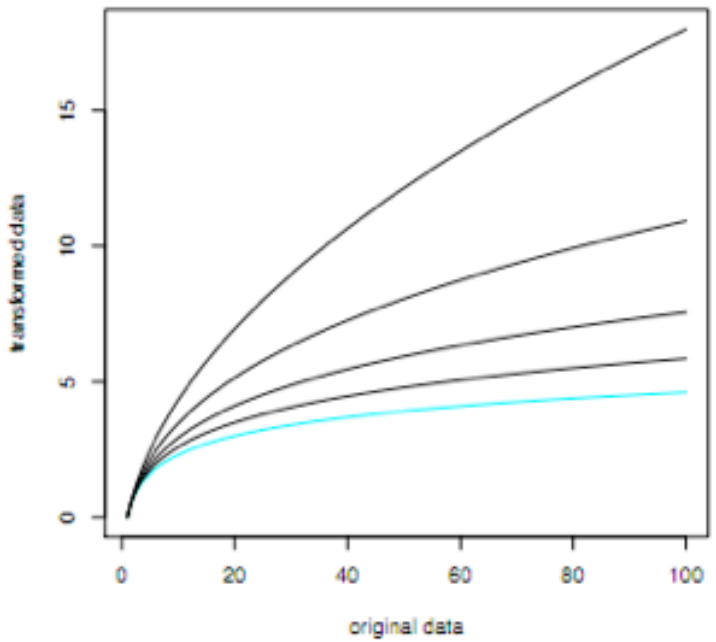
For data with a right skew, the square root, cube root, fourth root, and so on, can be used to define **a family of transformations** that all have the effect of taking big values and bringing them in closer to the rest of the data, to the smaller values

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

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

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





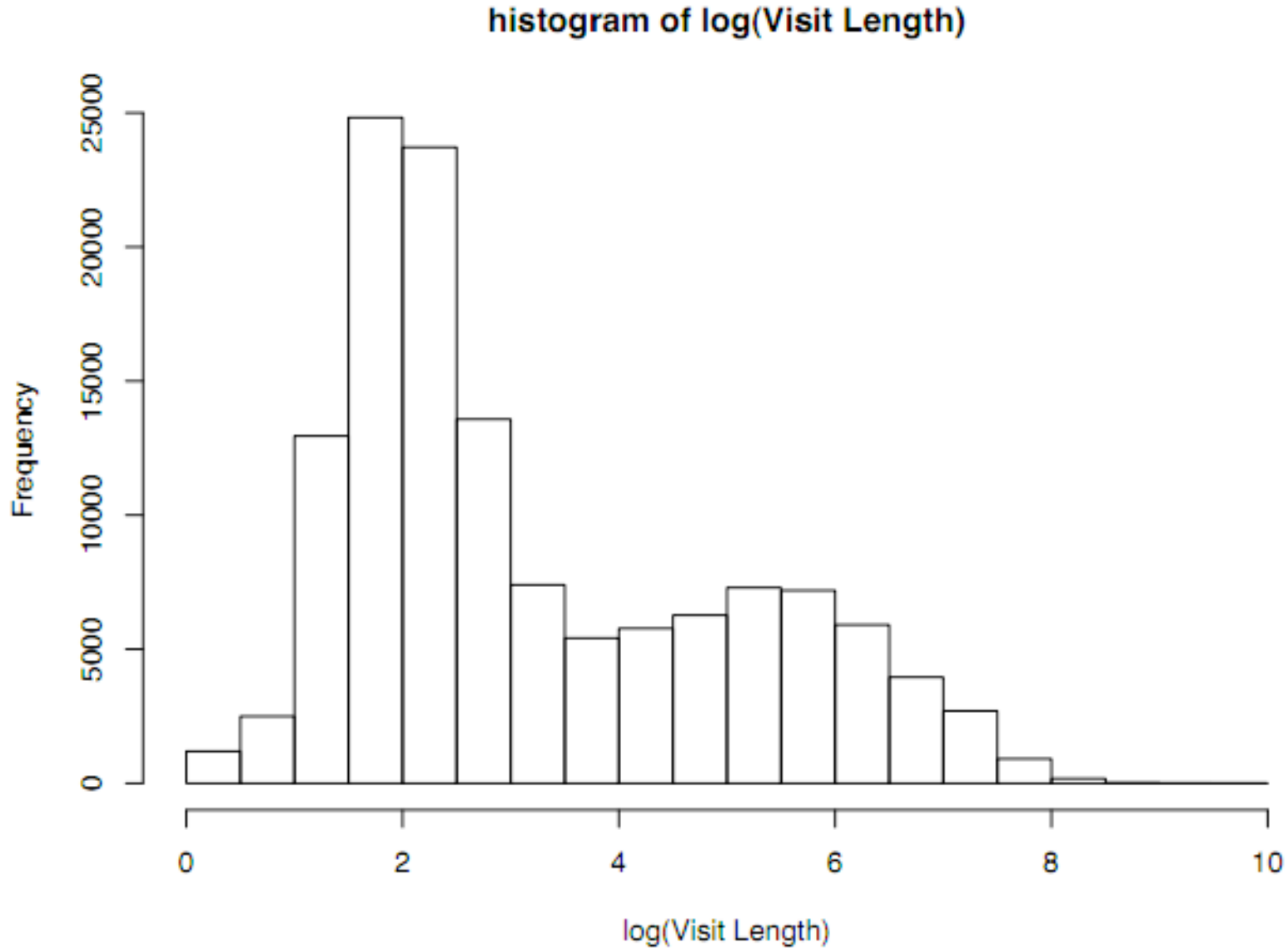
An aside

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

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

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

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

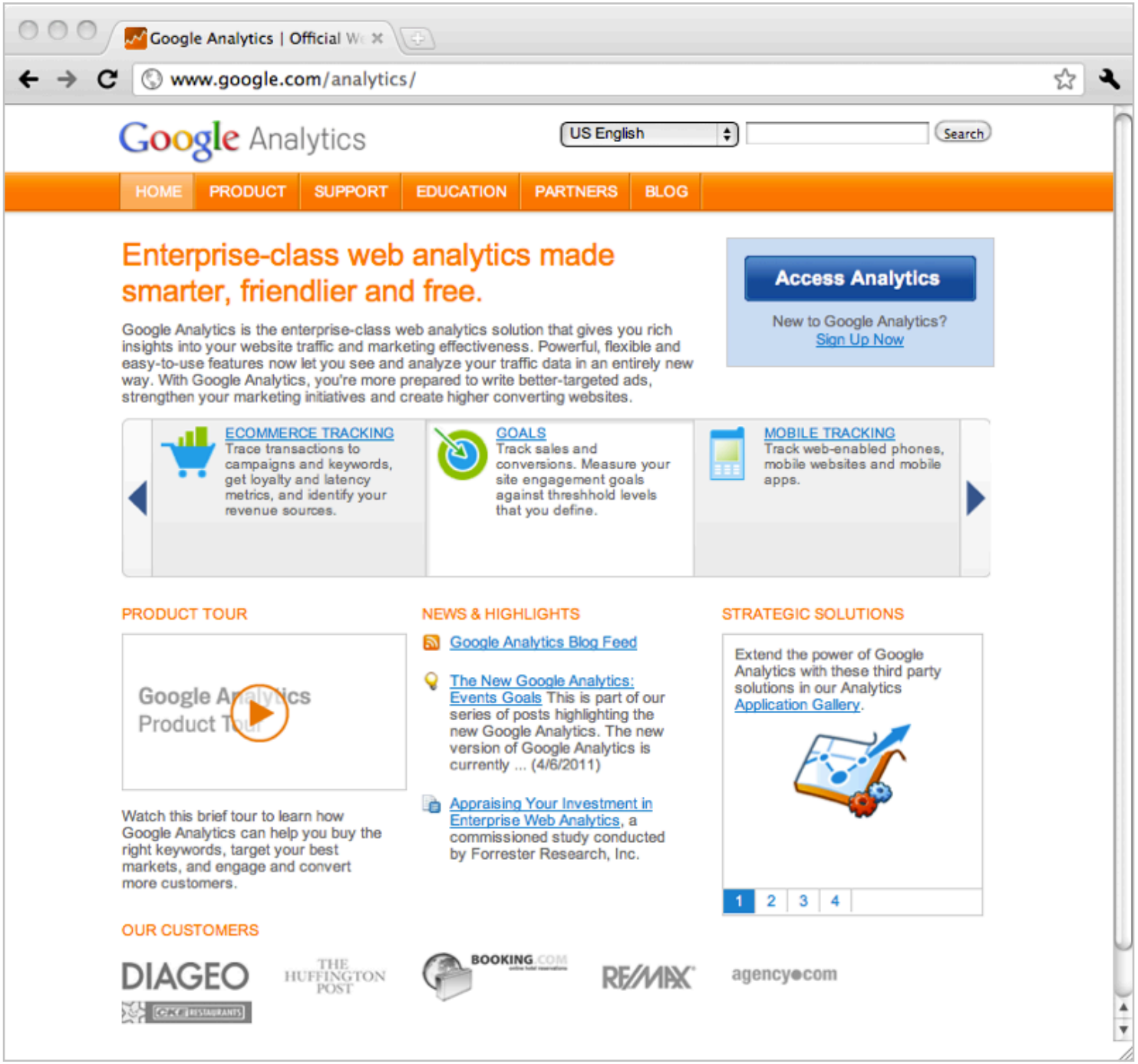

histogram of log(Visit Length)
Frequency
0
5000
10000
15000
20000
25000
0
2
4
6
8
10
log(Visit Length)

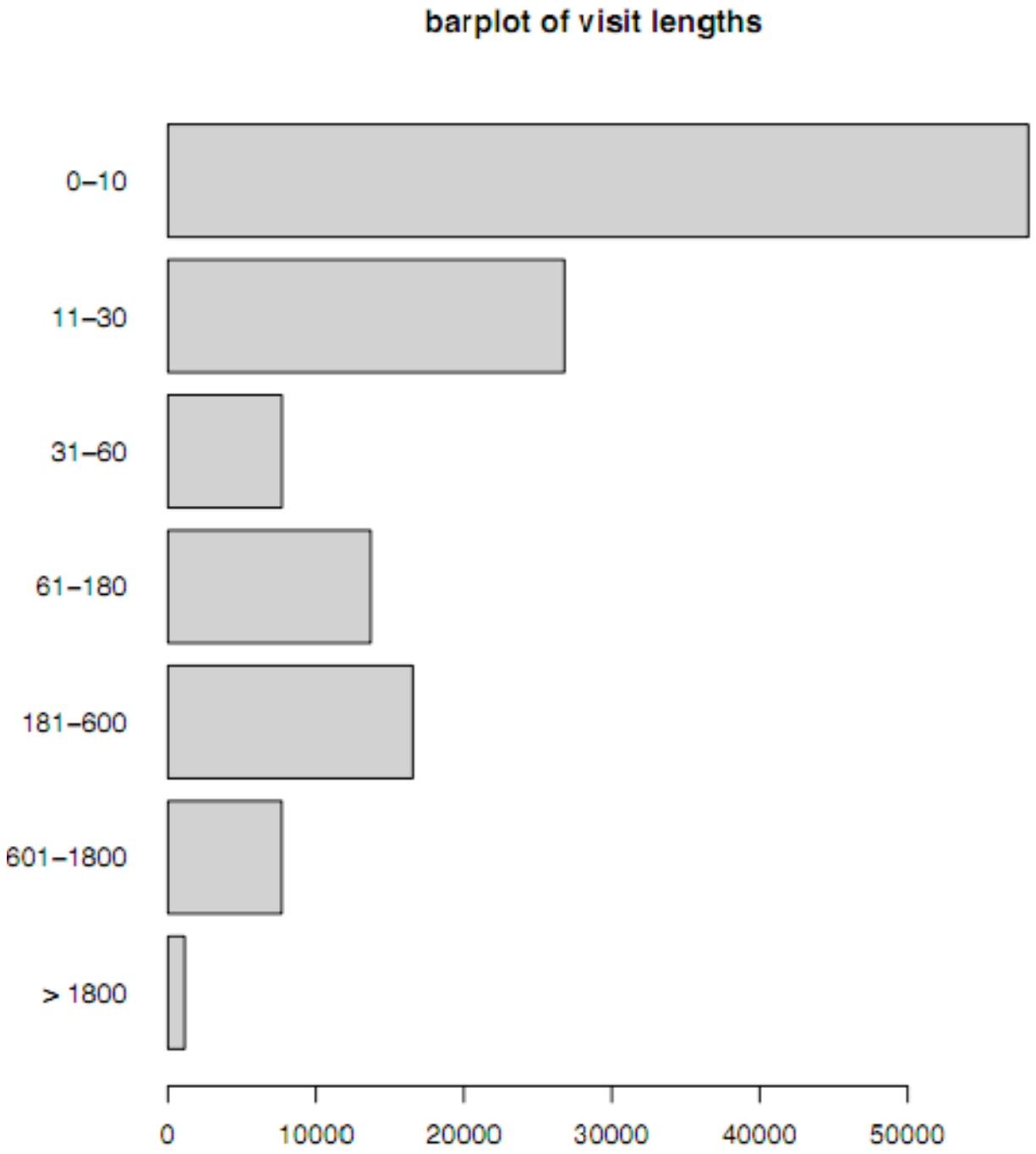
Alternatives

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

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

His view of Visit Length looked like this...


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


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

Alternatives

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

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

Transformations

With that out of the way...

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

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

Transformations

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

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

1 U.S. dollar = 5.78338895 Danish kroner

1 US gallon = 3.78541178 litres

1 short ton = 2000 pounds

Transformations

In fact, non-linear scaling is used across science

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

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

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

Transformations

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

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

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

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

* here log = natural log!

Question

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

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