

Lecture 4: (Re)shuffle

Last time

We started to work with a new data set, the Behavioral Risk Surveillance Survey from the CDC -- It is a larger telephone survey of the adult population in the United States

We used the new data as an opportunity to consider again our basic plotting tools (barplots, mosaic plots, histograms) and introduced a new category of frequency display, **the boxplot**

Boxplots are meant to be **a slightly cartoonish simplification of a set of data**, providing you with visual cues about the “center” of the distribution and how the data “spread” out around the center

We considered both a vanilla version of the displays that rely on the five number summary (the minimum, the lower quartile, the median, the upper quartile and the maximum) and a modified version...

Last time

The modified version is **an attempt to highlight data points that might be “outside” our expectations** in some way, where our expectations are set based on a bell-shaped distribution (we will make this concept more formal in a lecture or two)

The modified version exposes points that are “outside” our expectations and might be the subject of discussion -- **Are these points different from the rest in some way?**

We ended by extending the humble boxplot, first by augmenting it with a smoothed histogram (violin and bean plots) and then took the idea two data with two or more “dimensions” or variables

Today

We are going to **finish our extensions to the boxplot**, looking at how they can be made to create a cartoon view of a bivariate (two variables) scatter of points -- We'll entertain a new notion called "depth"

Then we'll **start our first rigorous treatment of statistical inference** -- We will examine so-called significance tests in the context of a **clinical trial**, essentially deciding if there is a "significant" difference between two treatments (or rather between a treatment and a "control")

First, however, something from the Wall Street Journal yesterday...

Joining data

Last time we took a “step too far” perhaps and joined demographic data about students with the table of enrollment events -- This seemed to be a little too close for comfort for many of you

A news story reported by the Wall Street Journal yesterday brings some of this even closer -- The story involved **applications that run on smart phones** (your iPhone or Android device)

How many of you have read **the privacy policies associated with your favorite web sites or apps?**

Mobile-App Makers Face U.S. Privacy Investigation

Article

Video

Stock Quotes

Comments (22)



Email



Print

Save This



Like

467



+ More



Text



By [AMIR EFRATI](#), [SCOTT THURM](#) and [DIONNE SEARCEY](#)

Federal prosecutors in New Jersey are investigating whether numerous smartphone applications illegally obtained or transmitted information about their users without proper disclosures, according to a person familiar with the matter.



Online-music streaming service Pandora, which plans an initial public offering, says in an SEC filing that it has been subpoenaed in an investigation probing

The criminal investigation is examining whether the app makers fully described to users the types of data they collected and why they needed the information—such as a user's location or a unique identifier for the phone—the person familiar with the matter said. Collecting information about a user without proper notice or authorization could violate a federal computer-fraud law.

Online music service Pandora Media Inc.

Mobile-App Makers Face U.S. Privacy Investigation

Article

Video

Stock Quotes

Comments (22)

 Email

 Print


Save This


 Like

467



+ More

 Text

By AMIR EFRATI, SCOTT THURM and DIONNE SEARCEY

Federal prosecutors in New Jersey are investigating whether numerous smartphone applications illegally obtained or transmitted information about their users without proper disclosures, according to a person familiar with the matter.



Online-music streaming service Pandora, which plans an initial public offering, says in an SEC filing that it has been subpoenaed in an investigation probing information-sharing by mobile applications. John Letzing and Stacey Delo discuss.

The criminal investigation is examining whether the app makers fully described to users the types of data they collected and why they needed the information—such as a user's location or a unique identifier for the phone—the person familiar with the matter said. Collecting information about a user without proper notice or authorization could violate a federal computer-fraud law.

Online music service Pandora Media Inc. said Monday it received a subpoena related to a federal grand-jury investigation of

information-sharing practices by smartphone applications.

Pandora disclosed the subpoena, issued "in early 2011," in a Securities and Exchange Commission filing. The Oakland, Calif., company said it had been informed it is "not a specific target of the investigation." Pandora said it believed similar subpoenas had been issued "on an industry-wide basis to the publishers of numerous other smartphone applications."

A Pandora spokeswoman declined to comment.

The Wall Street Journal reported in December that popular applications on the iPhone and Android mobile phones, including Pandora, transmit information about the phones, their users and their locations to outsiders, including advertising networks.



View Full Image

Bloomberg News

Smartphone apps—of which there are thousands—are software programs that allow, say, a user to read an e-book, play a game, get sports scores or search for a restaurant.

The Journal tested 101 apps and found that 56 transmitted the phone's unique device identifier to other companies without users' awareness or consent. Forty-seven apps transmitted the phone's location in some

*The Wall Street Journal reported in December that popular applications on the iPhone and Android mobile phones, including Pandora, **transmit information about the phones, their users and their locations to outsiders, including advertising networks.***

Smartphone apps—of which there are thousands—are software programs that allow, say, a user to read an e-book, play a game, get sports scores or search for a restaurant.

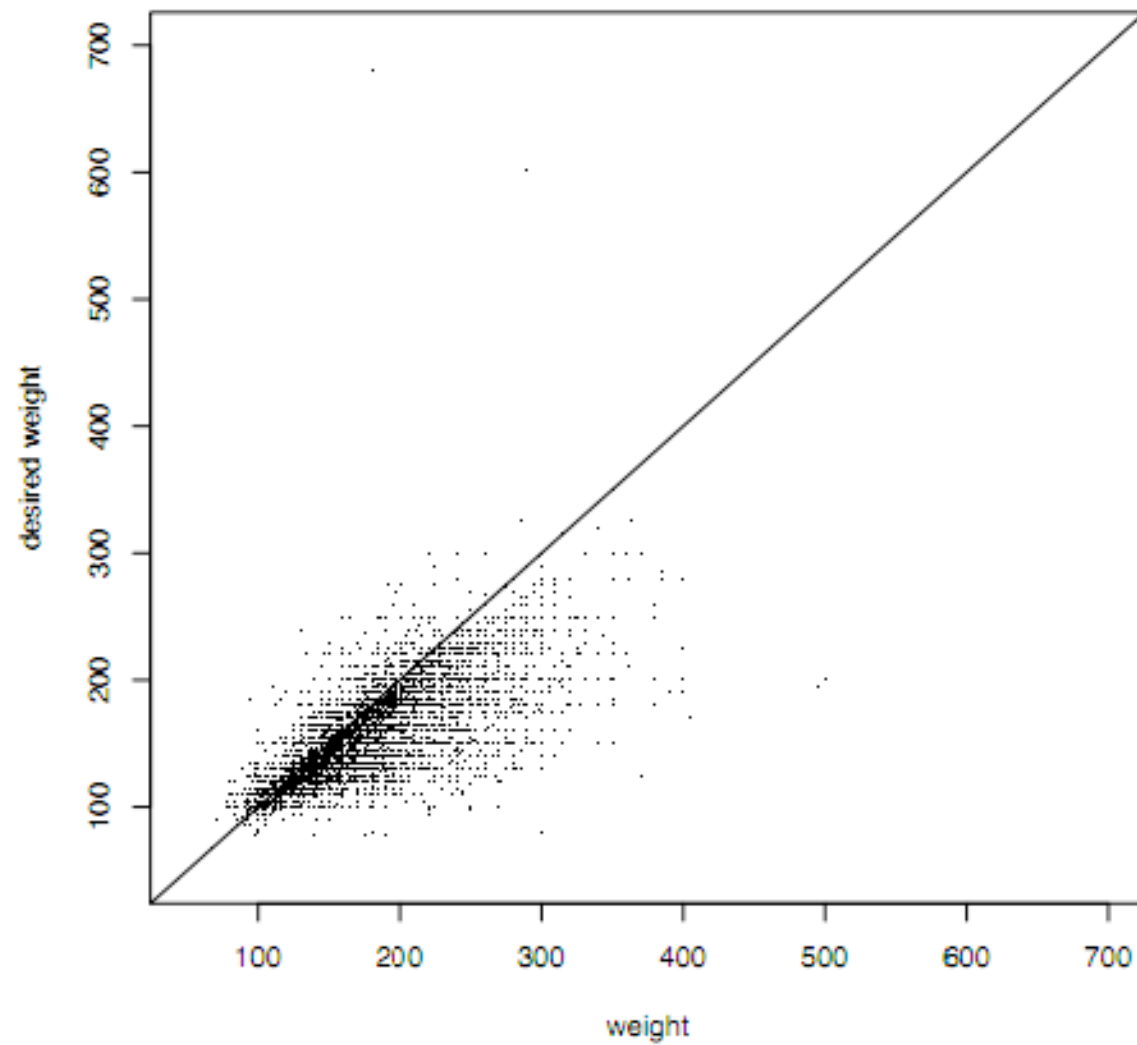
*The Journal tested 101 apps and found that 56 transmitted the phone's unique device identifier to other companies without users' awareness or consent. Forty-seven apps transmitted the phone's location in some way. **Five sent a user's age, gender and other personal details to outsiders. At the time they were tested, 45 apps didn't provide privacy policies on their websites or inside the apps.***

Extensions

Last time we considered how we might **extend the boxplot to two dimensions** (two variables) -- We saw a scatterplot of BRFSS respondents' desired weight against their weight

Did you have any thoughts about what we could do?

scatterplot of weight and desired weight

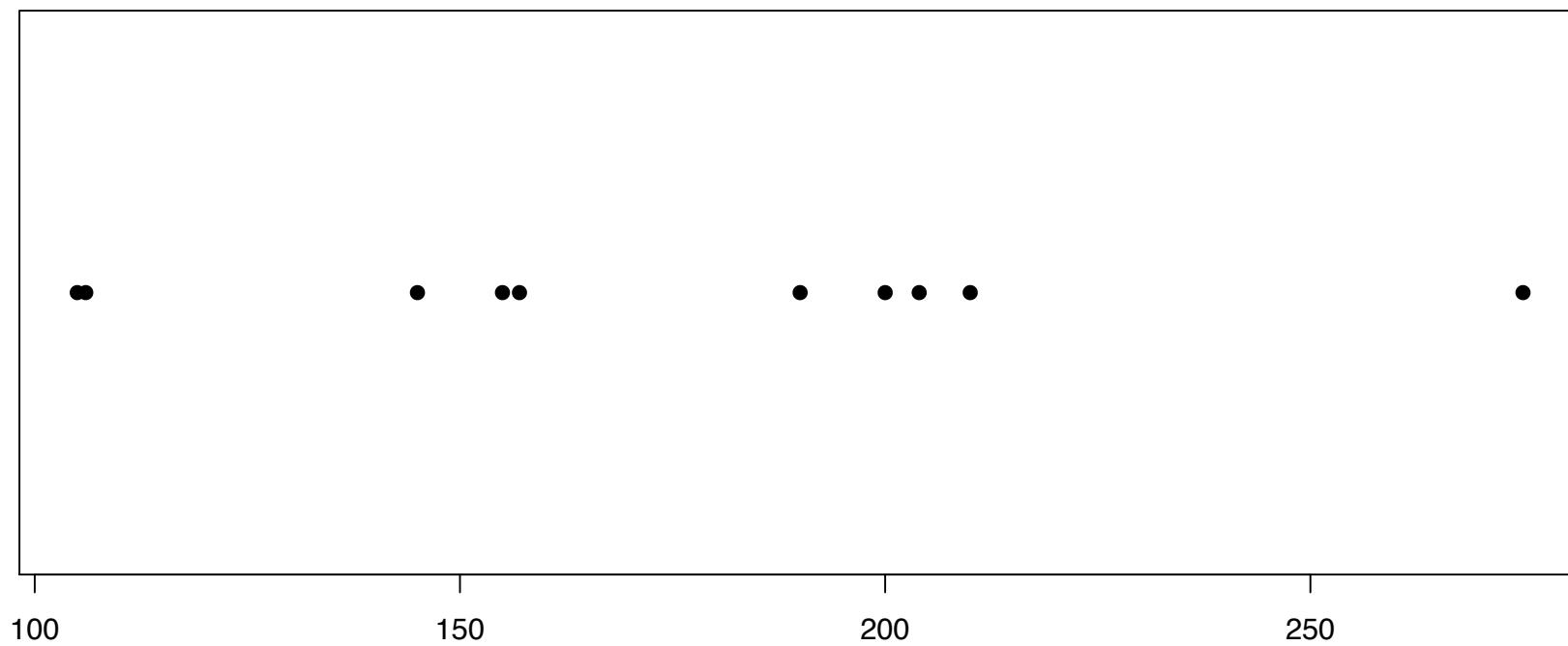


Depth

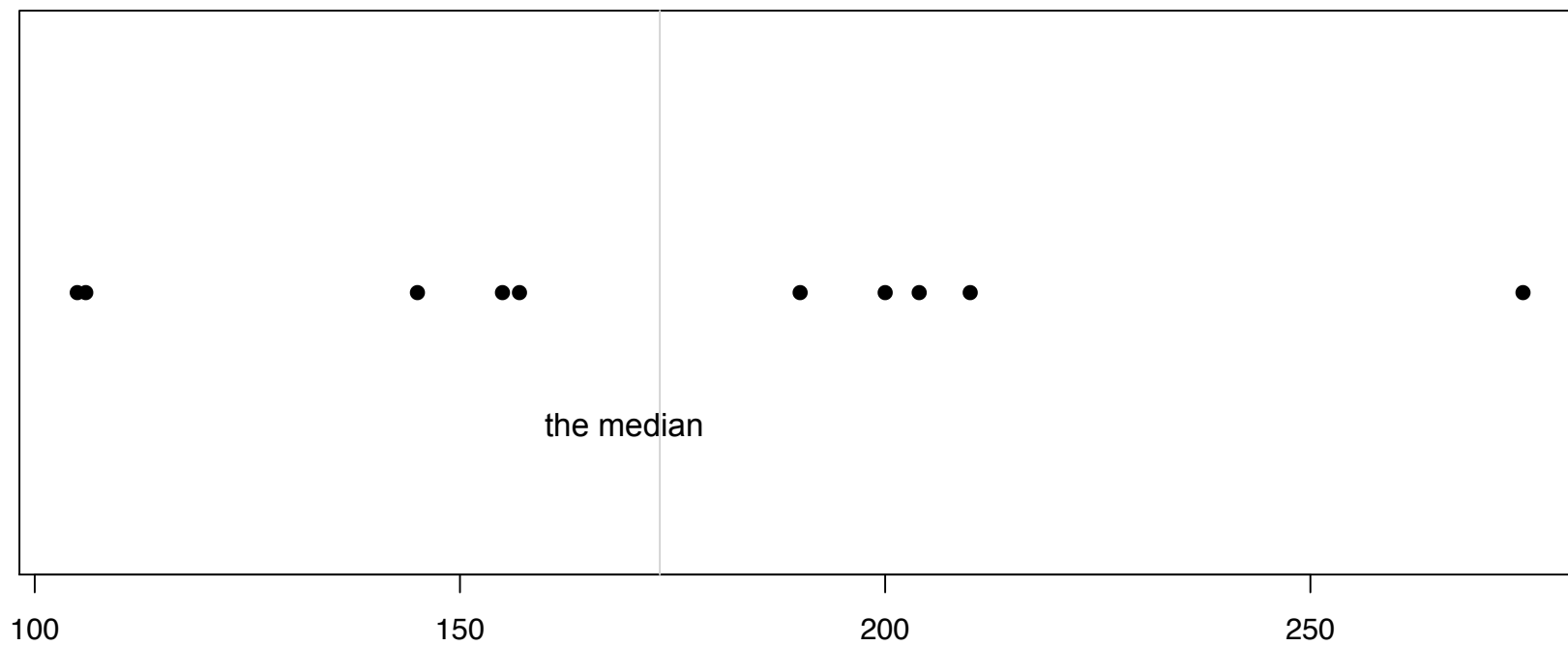
Let's start by thinking about what we're doing when we define the median for a data from a single variable (say just the BRFSS reported weights) -- Last time we took it to be **the point that divides our data into two pieces** (plus or minus some extra details when we have an even number of points)

We will first consider the median as **the “deepest” location relative to our data set** and then consider how to generalize that notion -- Again, we do this because it gives us insight into concepts like the median and displays like boxplots

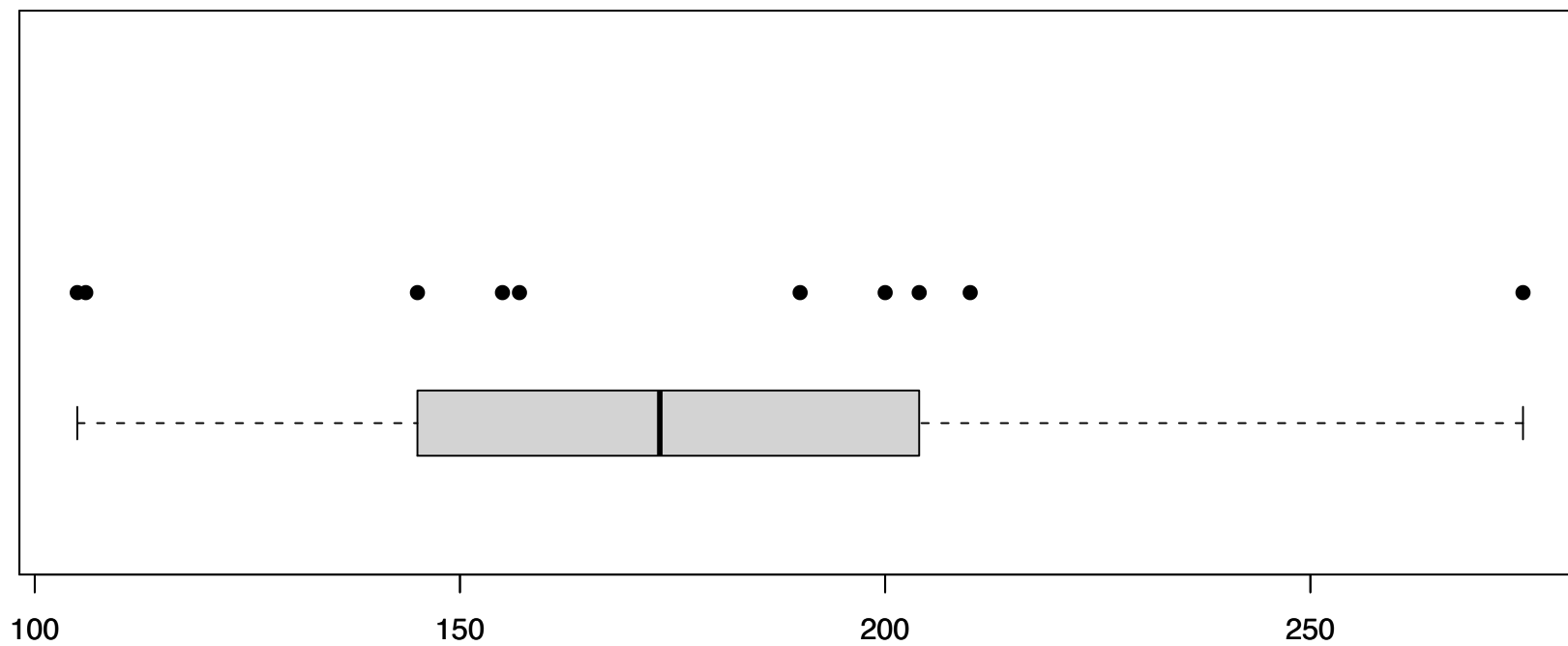
10 weights from the BRFSS



10 weights from the BRFSS



10 weights from the BRFSS

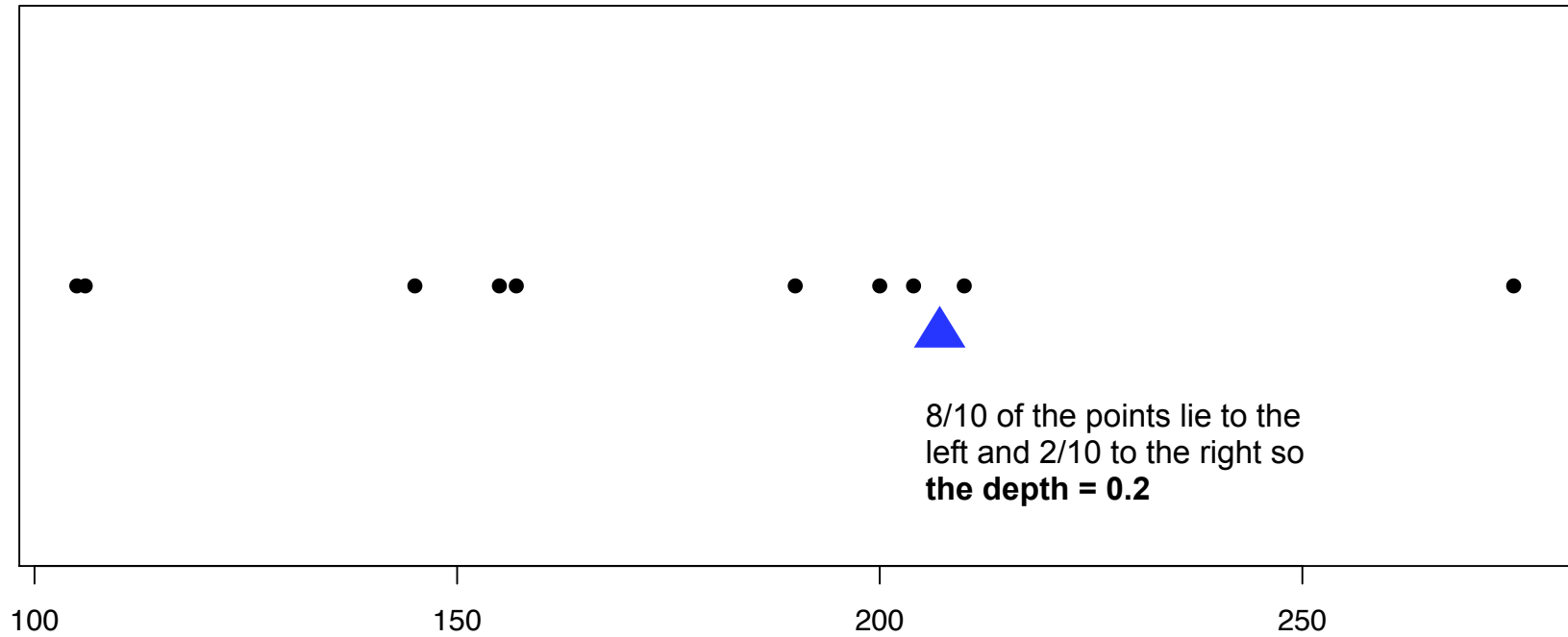


Depth

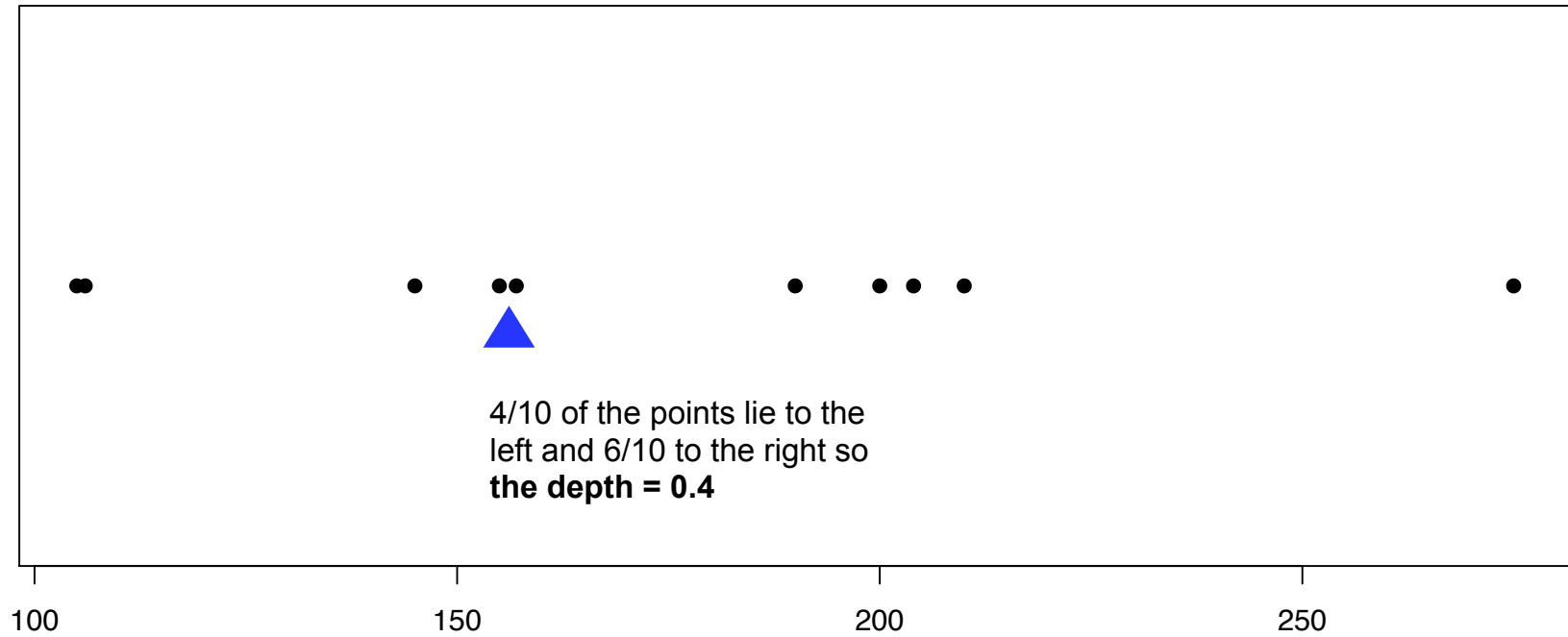
To define the depth of any location on the real line relative to this data set, we count **the proportion of points to the left and to the right** and define its depth to be **the smaller of the two**

Here are a couple of examples...

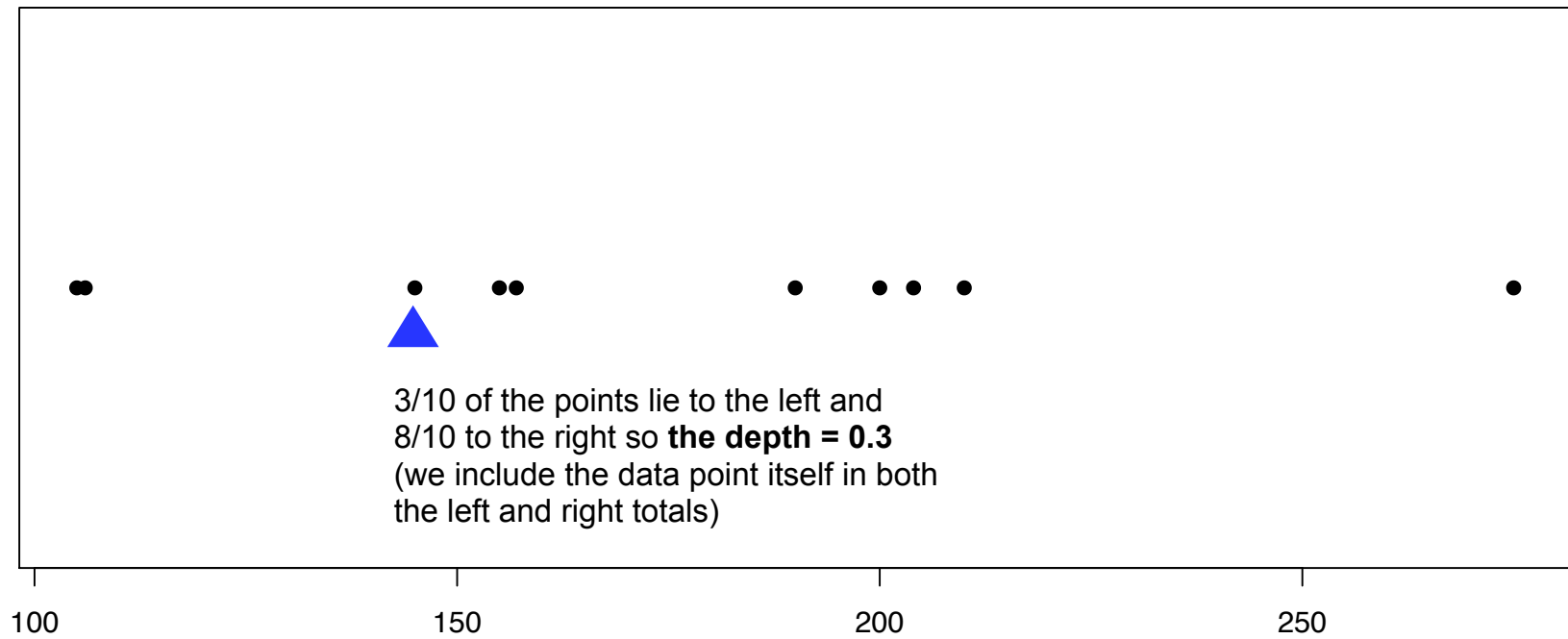
10 weights from the BRFSS



10 weights from the BRFSS



10 weights from the BRFSS

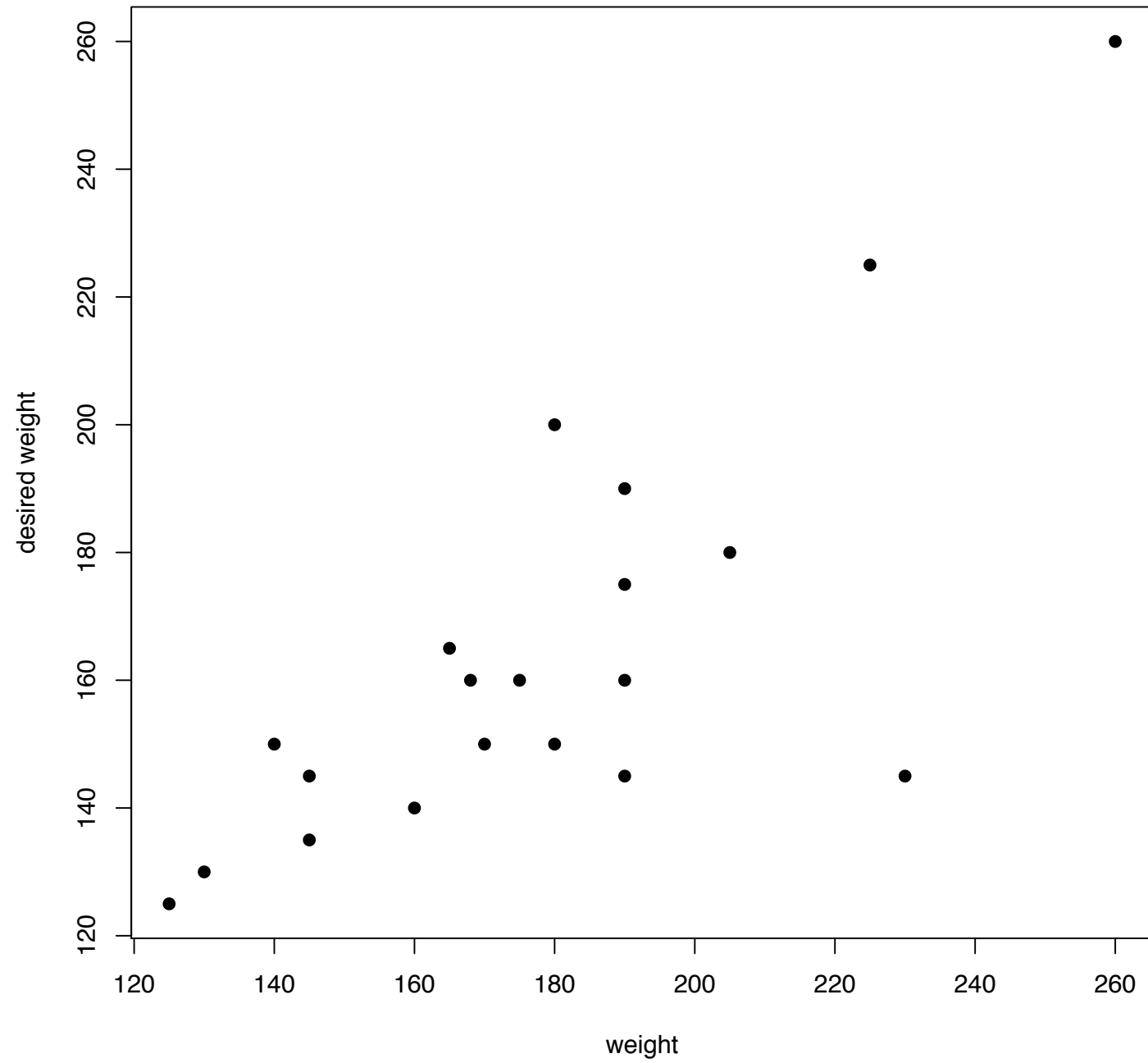


Depth

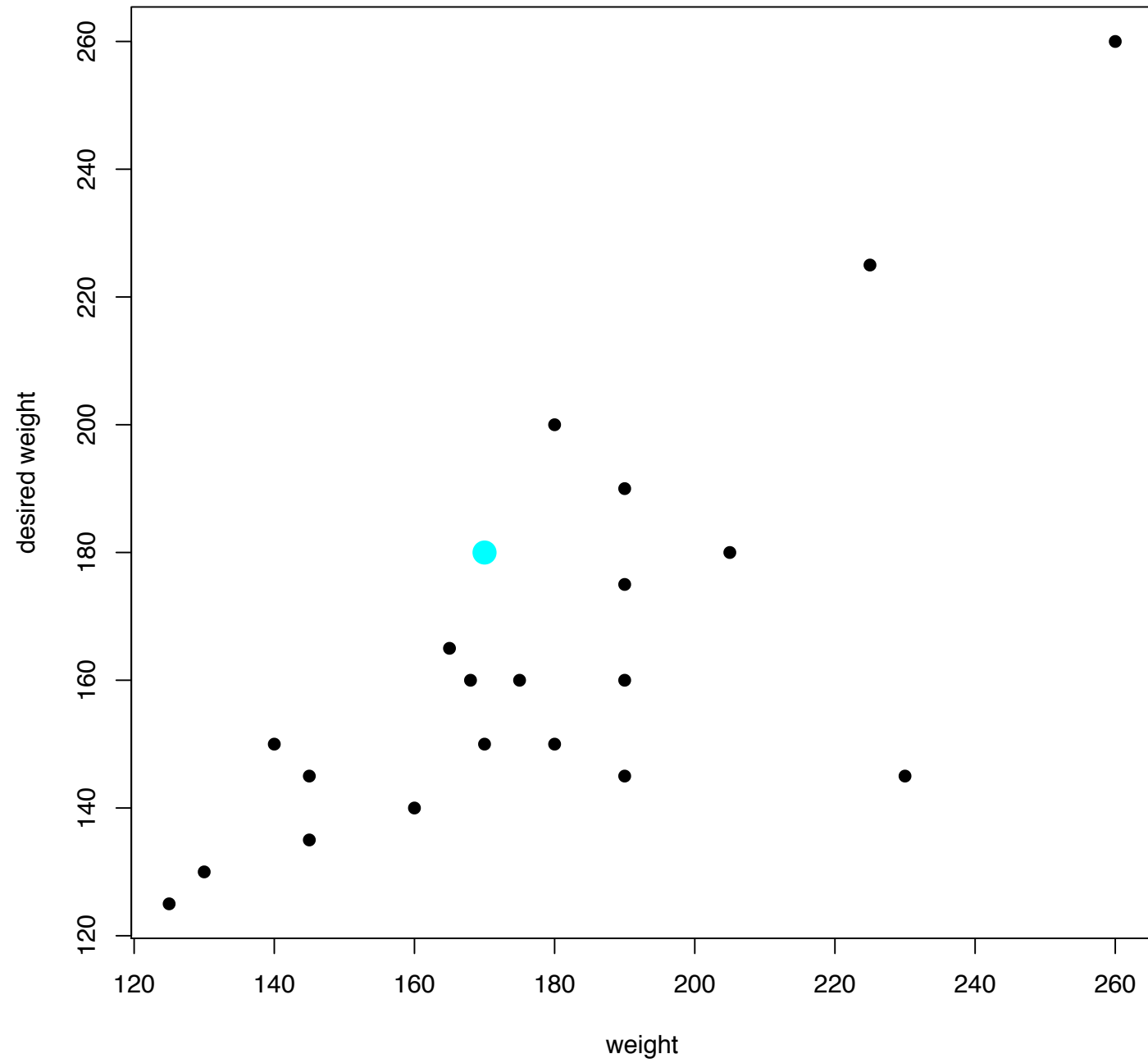
The median, then, is **the location on the real line having the greatest depth** -- If an “interval” of locations have greatest depth (as is the case on the previous slides, where any location between and including the 5th and 6th data points all have depth $1/2$) we take the midpoint of the interval as the median

Now, how do we generalize this to two dimensions? How do we generalize the notion of left and right?

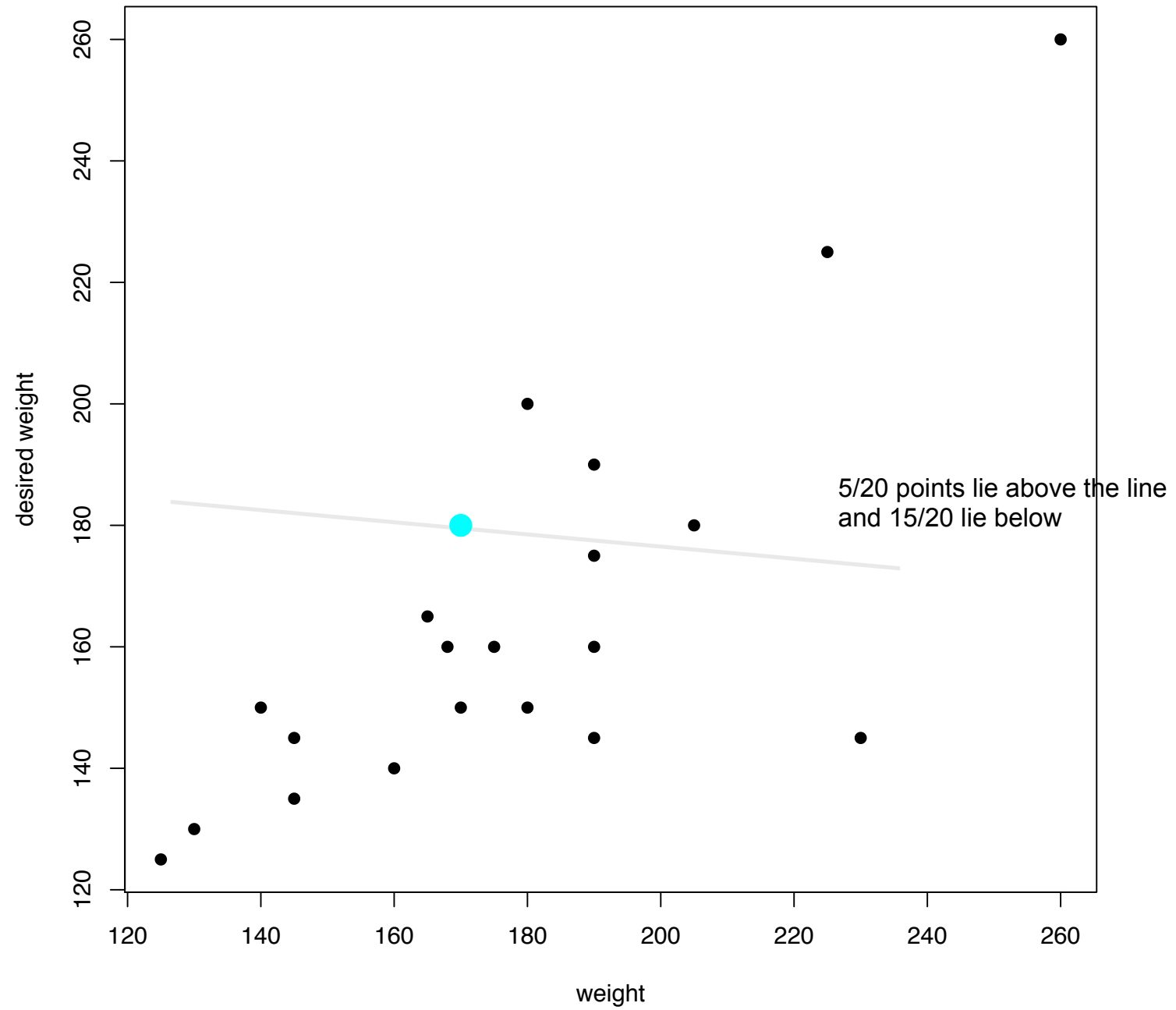
20 points from the CDC BRFSS



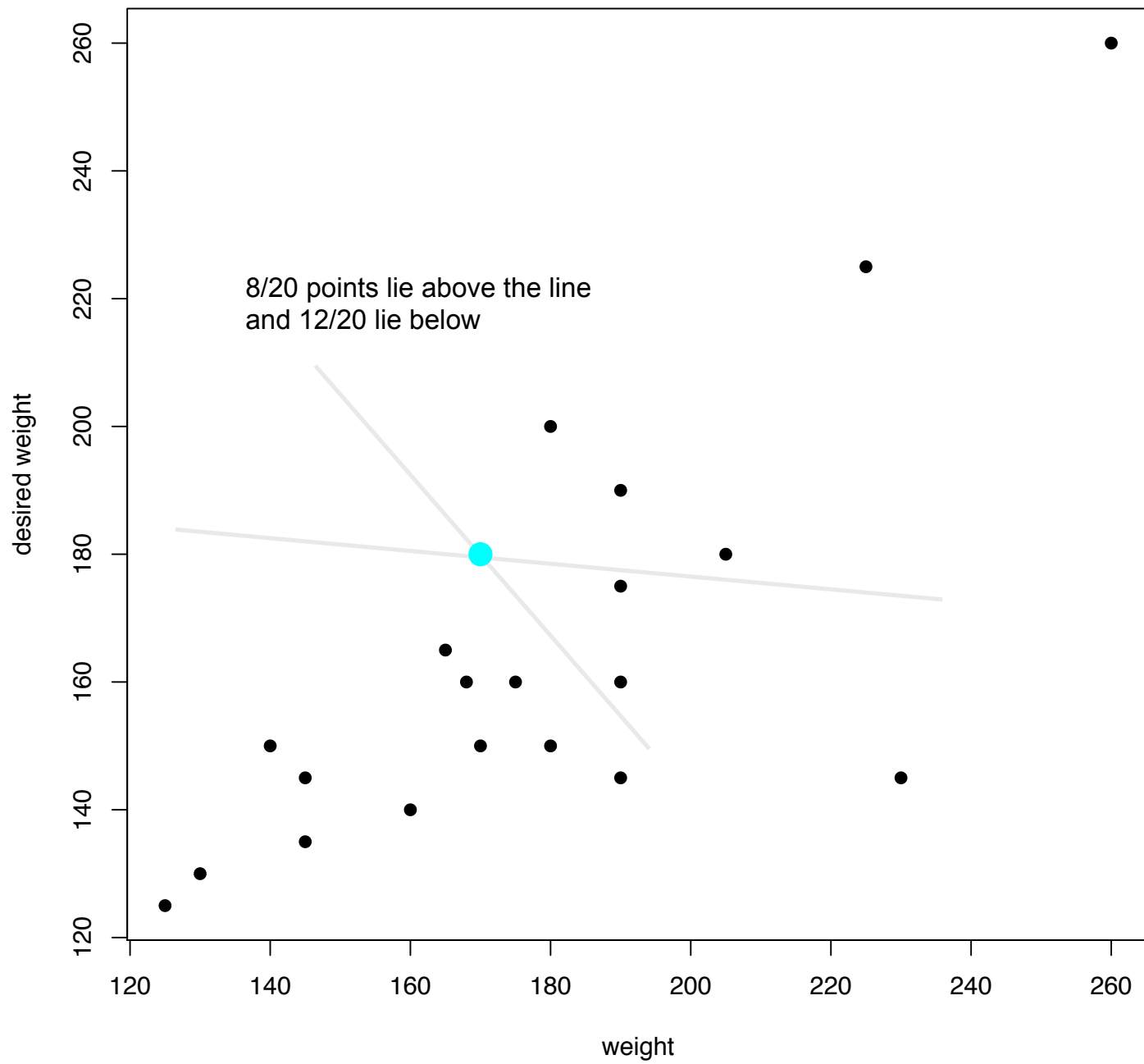
20 points from the CDC BRFSS



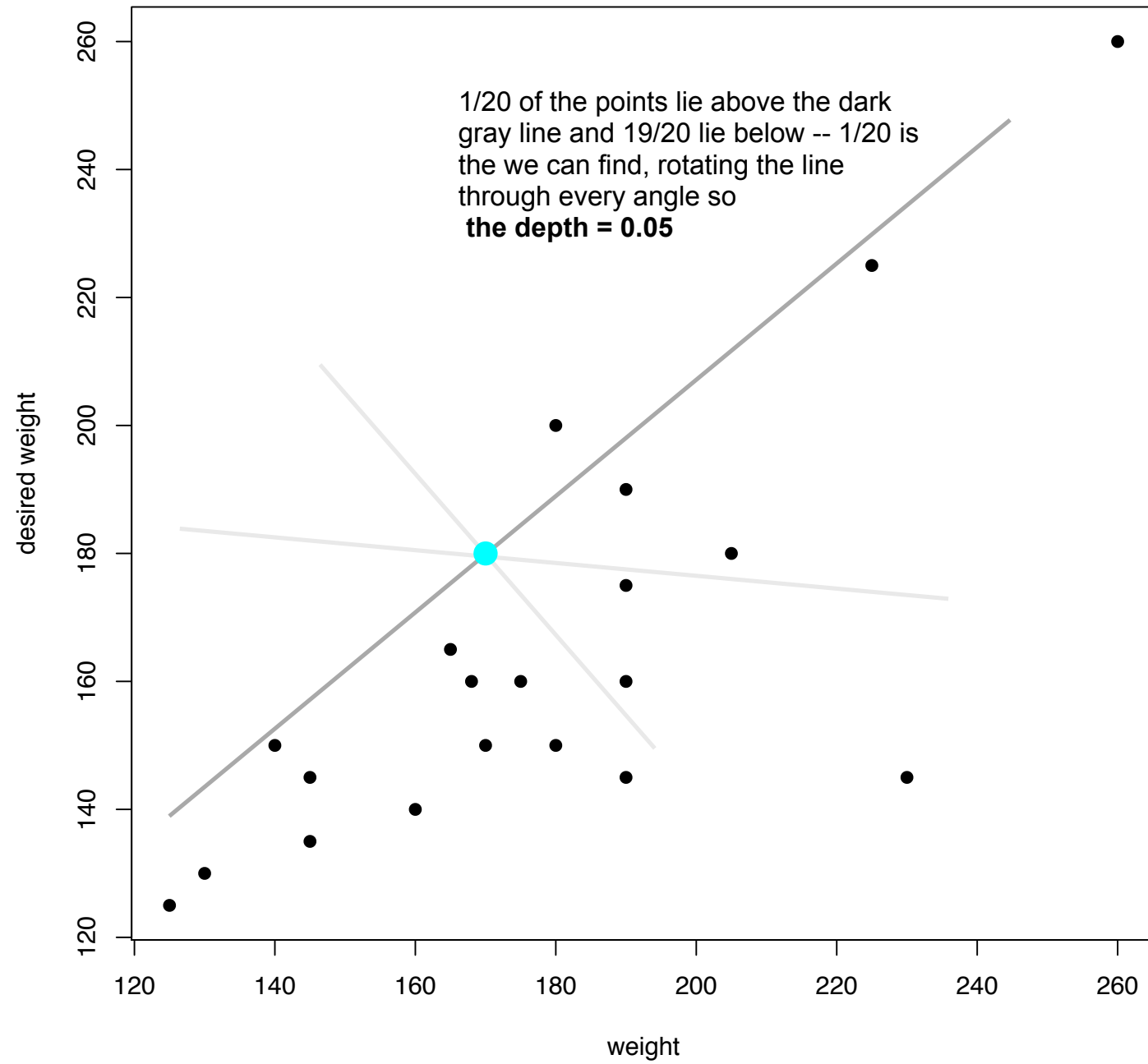
20 points from the CDC BRFSS



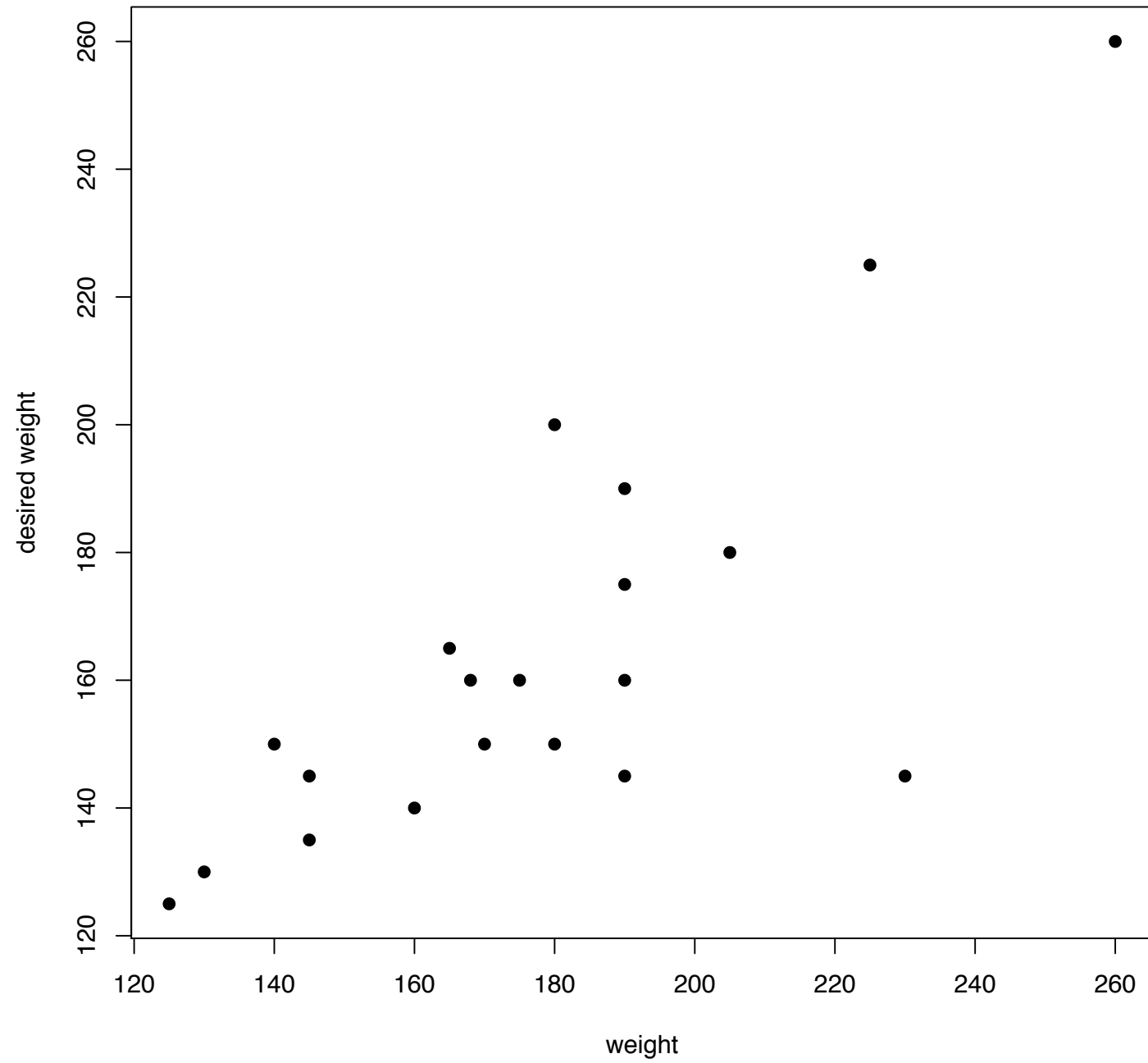
20 points from the CDC BRFSS



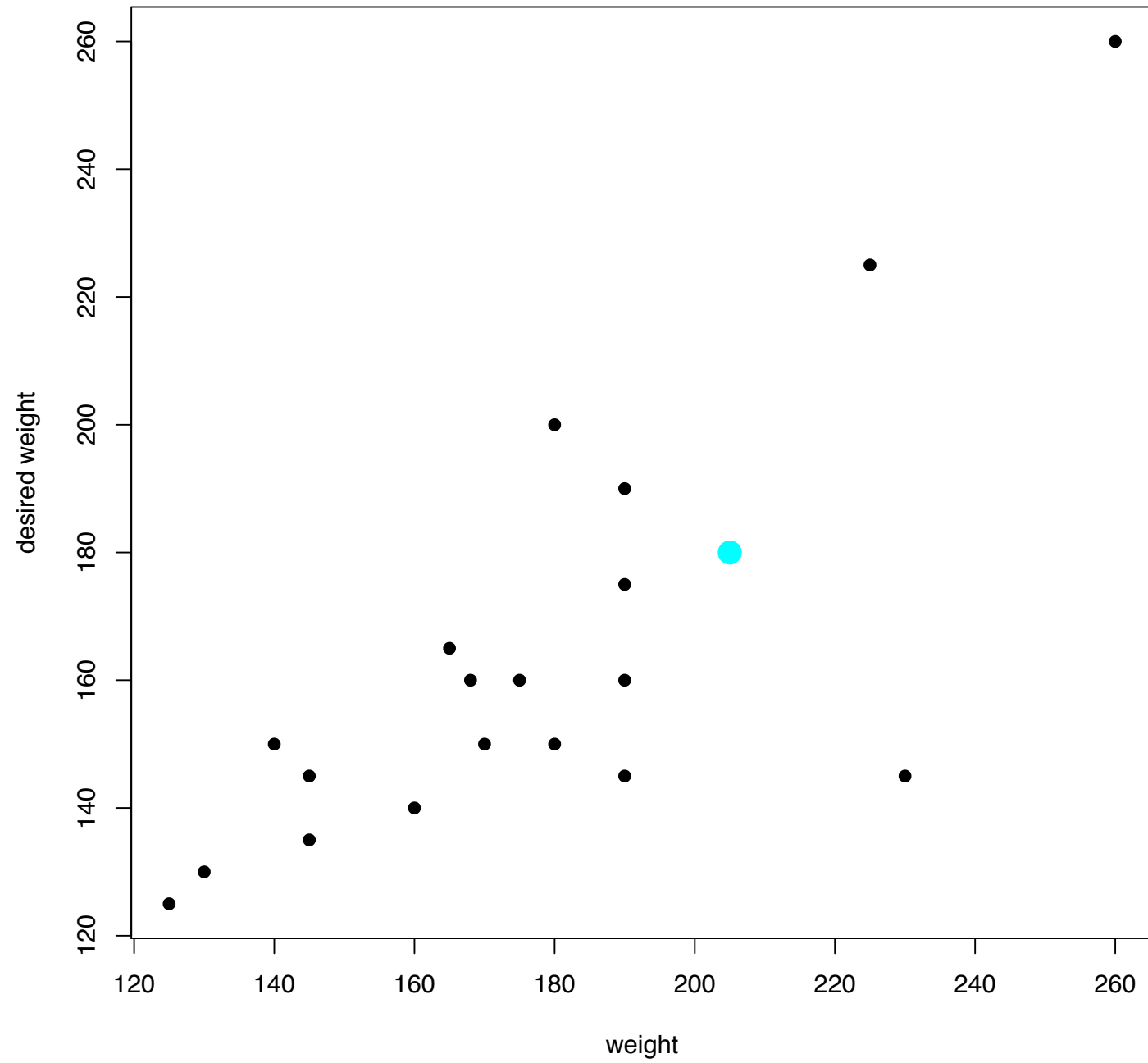
20 points from the CDC BRFSS



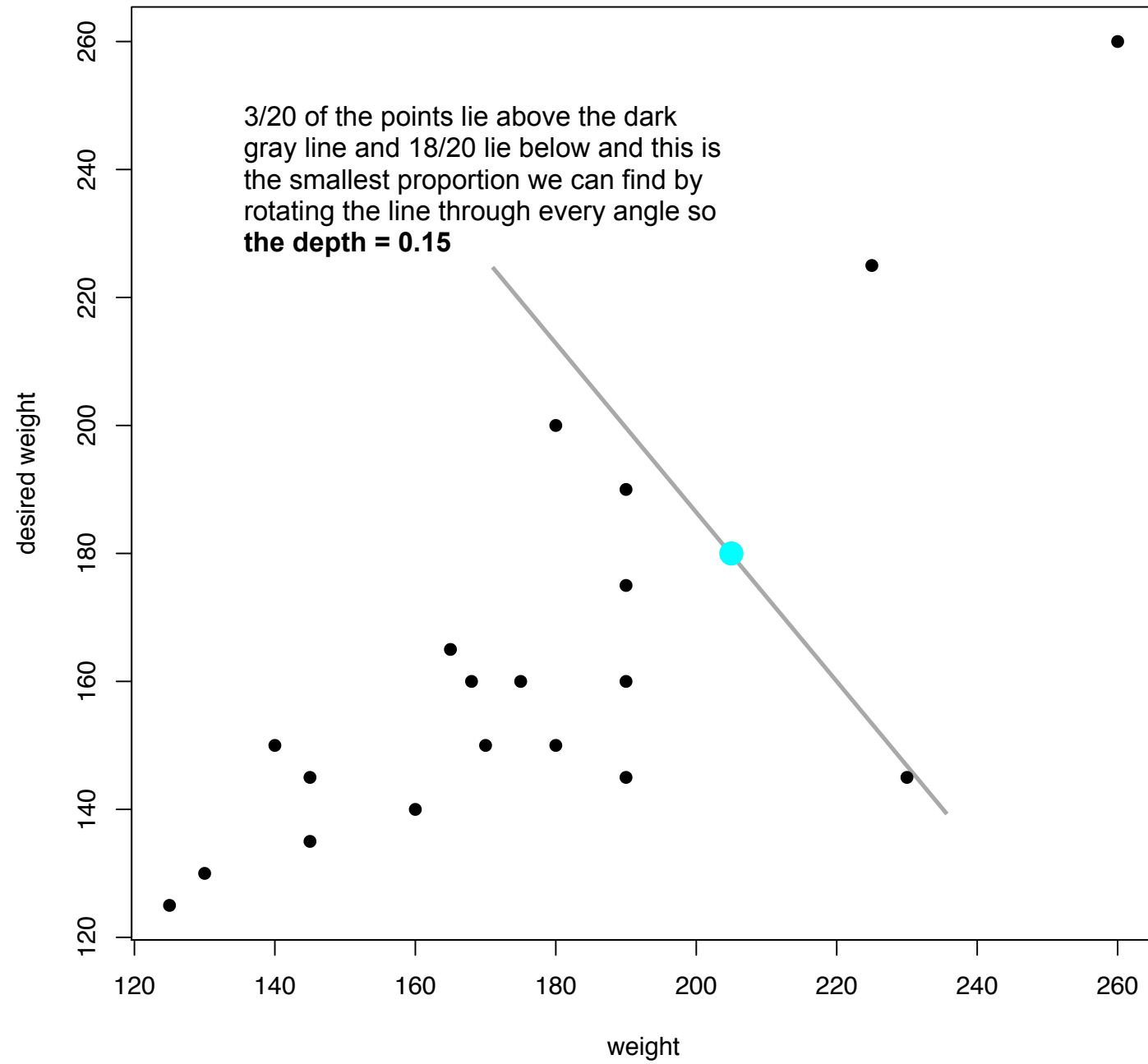
20 points from the CDC BRFSS



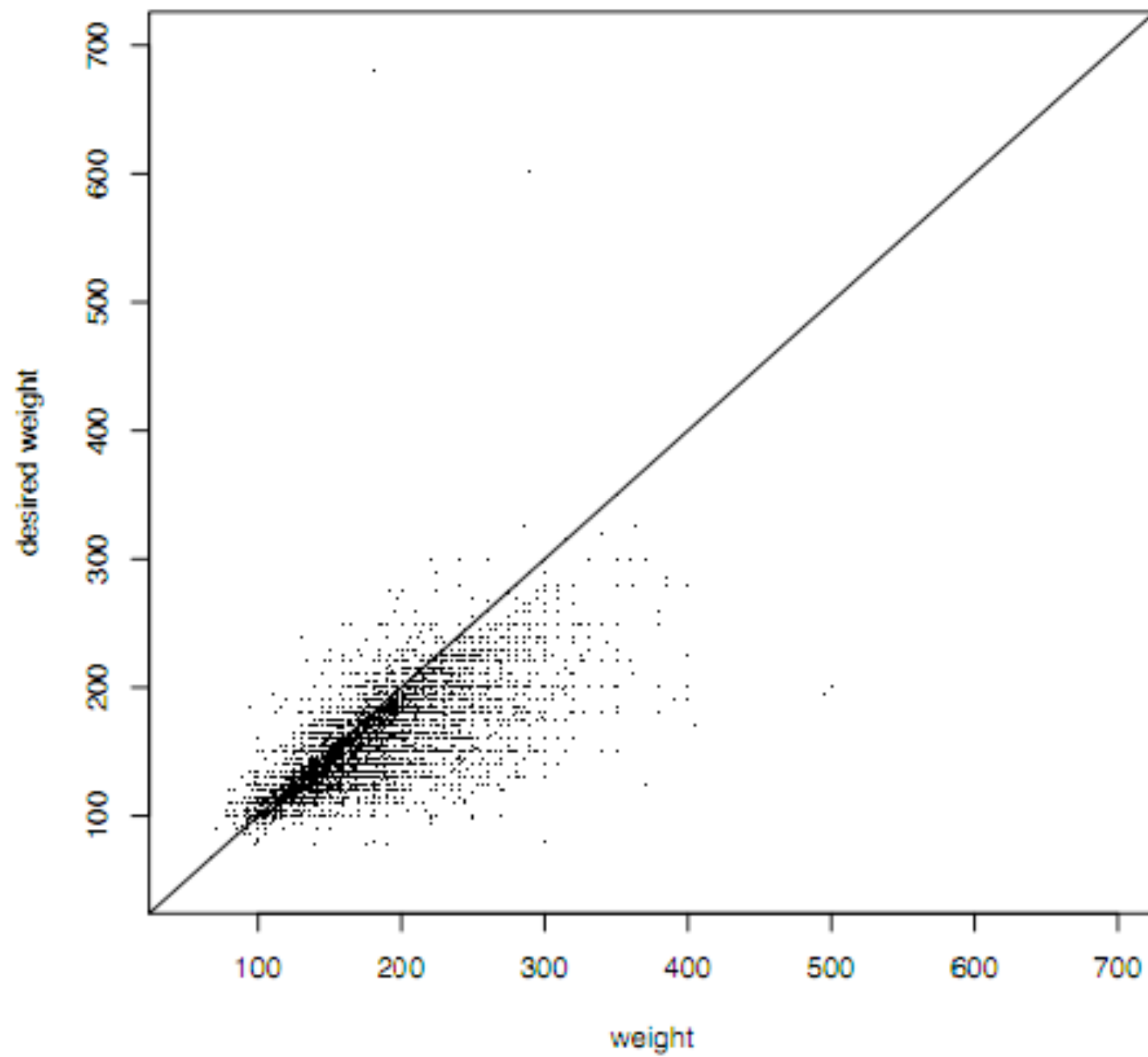
20 points from the CDC BRFSS

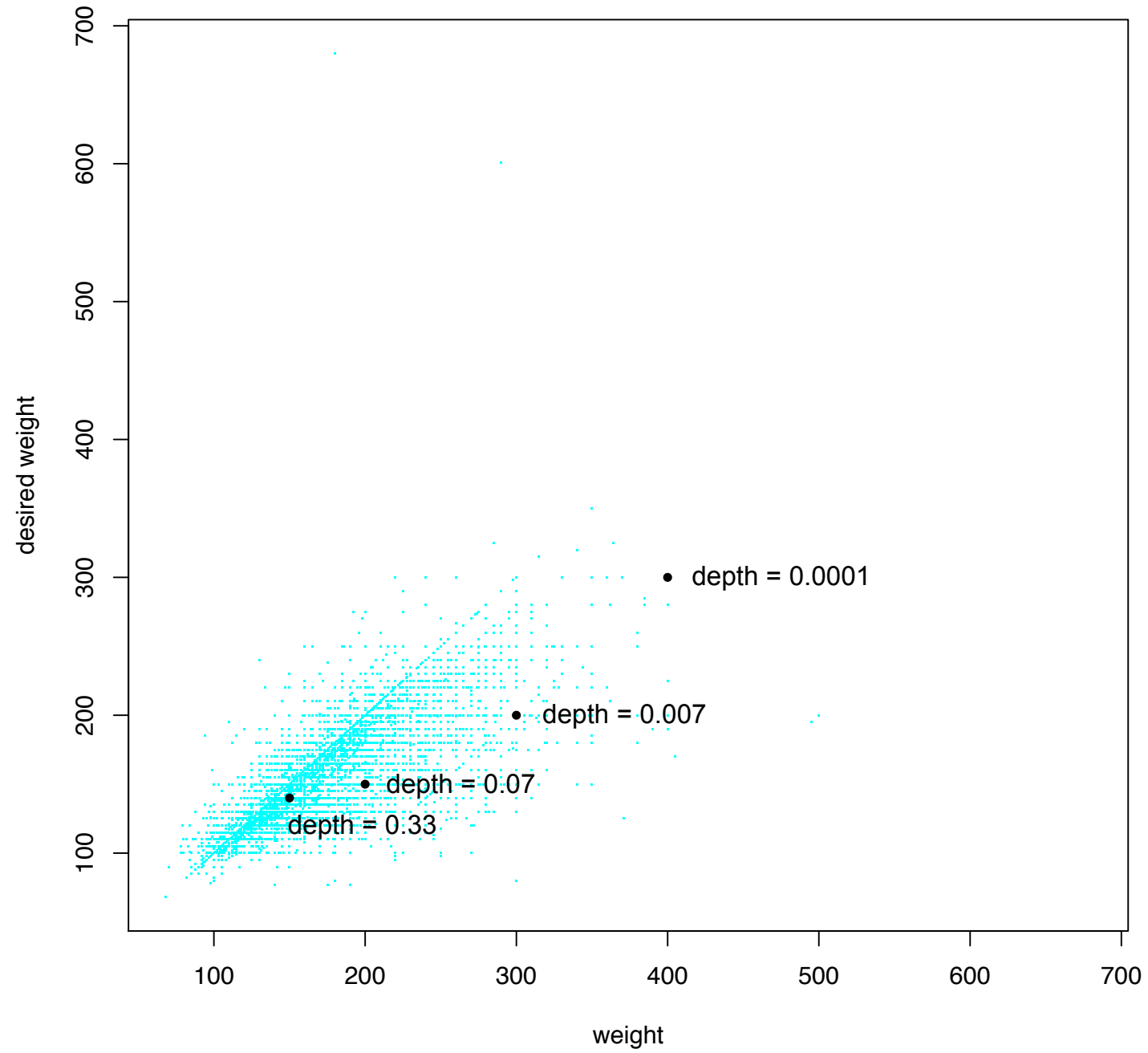


20 points from the CDC BRFSS



scatterplot of weight and desired weight





A generalization

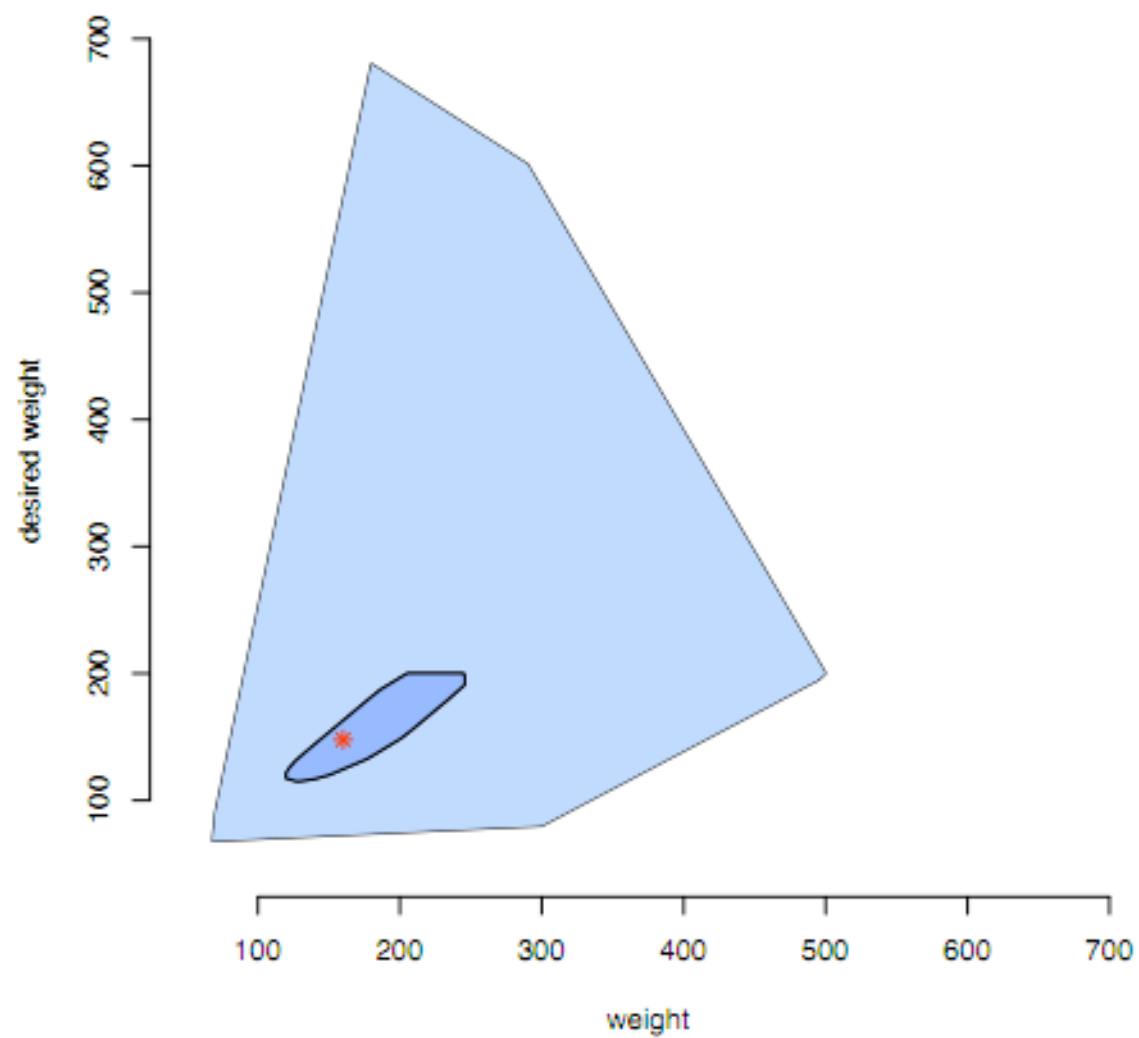
We can then **define the “depth median” as the deepest location** (if it’s unique) or the “center of gravity” of the set if there are more (it’s guaranteed to be a closed, bounded and convex set if any of those words speak to you -- and it’s not important if not)

Similarly, we can **use depth to define the deepest 50% of the data** (essentially), creating a generalization of the box part of the box plot -- The “whiskers” or in this case an outer “loop” is defined by inflating the middle 50% (default is a factor of 3, again based on simulations) and settling back on the data

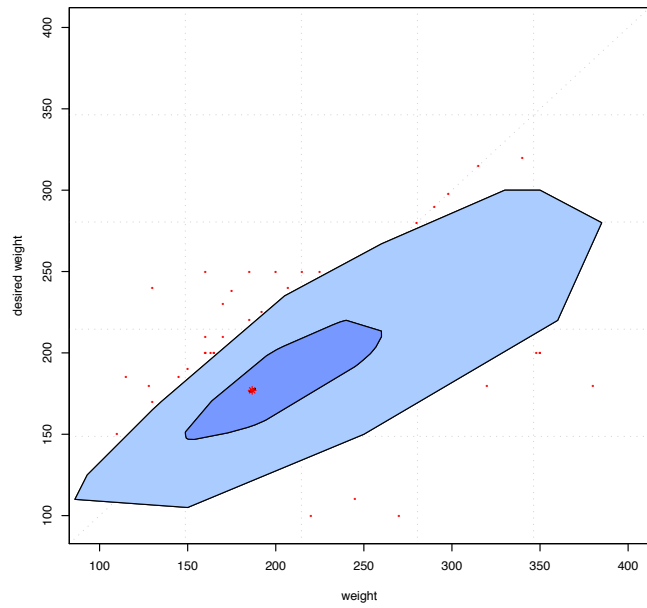
The authors of the graphic say:

*Like the univariate boxplot, the bagplot also visualizes several characteristics of the data: **its location** (the depth median), **spread** (the size of the bag), **correlation** (the orientation of the bag), **skewness** (the shape of the bag and the loop), and **tails** (the points near the boundary of the loop and the outliers)*

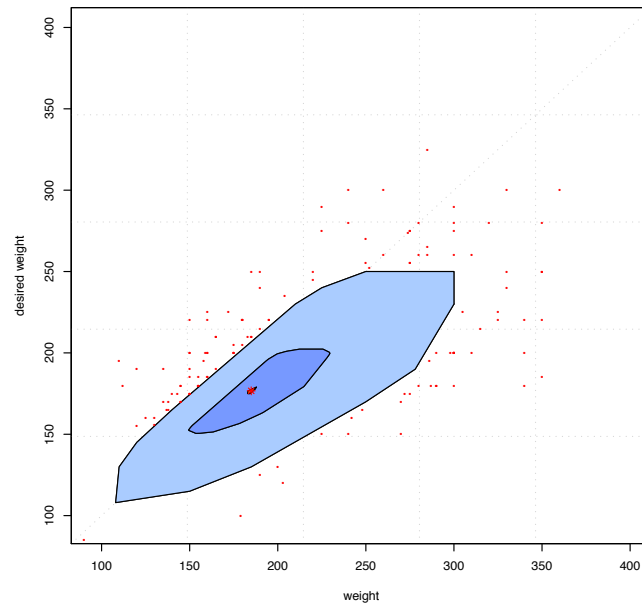
bagplot of weight and desired weight



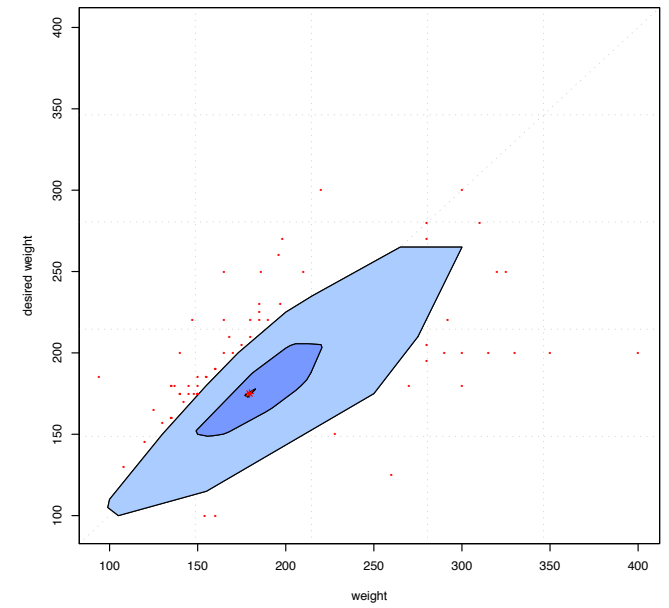
men in good health



men in very good health



men in excellent health



Other measures of center and spread

So far, we have used techniques that are more about sorting than anything else; we found the “center” of a distribution by selecting a point in the middle of a sample, the median

Our notion of “spread” came from the IQR, the interquartile range; this is the central region containing 50% of the data

Over the next few slides, we illustrate two other measures, **the mean and variance**; I’d like you to read this, together with the sections in your book, but we won’t talk about it much in class for another week when we start using these numerical summaries for inference

Mean and variance

A perhaps more primary notion of center for a sample of data is the **arithmetic mean or average**; given observations x_1, x_2, \dots, x_n , this is just

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

In a similar spirit, **the variance** of a sample of data is the sum of squared deviations from the mean; that is, for each point x_1 , the deviation is $(x_1 - \bar{x})$ and the variance s^2 is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

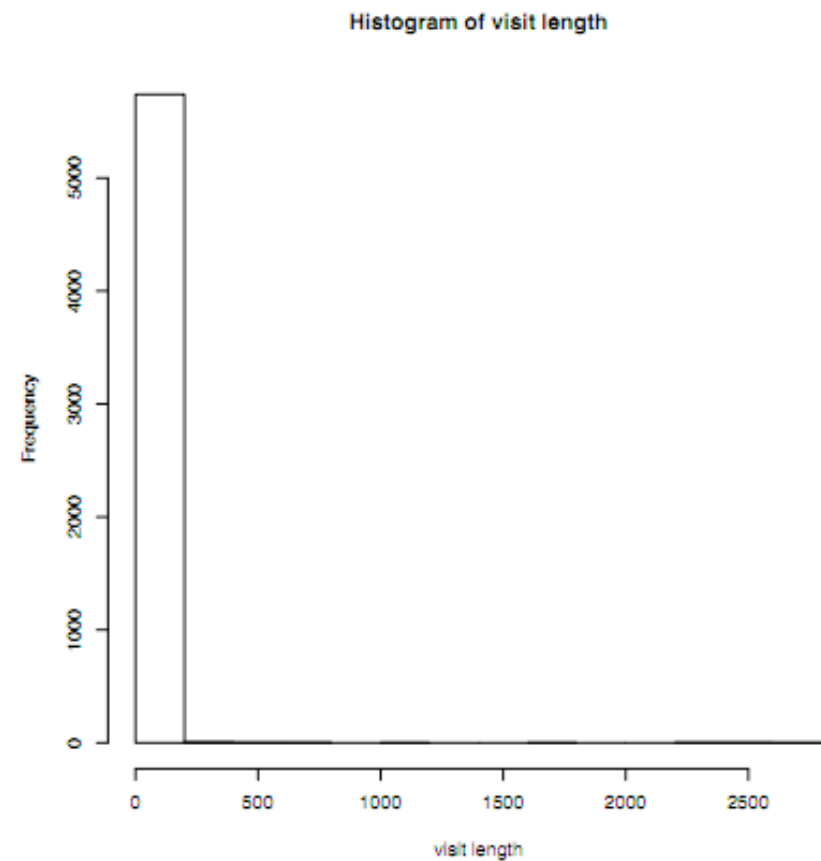
Mean and variance

While the median and the IQR were very direction notions of center and spread, the mean and standard deviation are slightly more delicate; for example, the mean is very much influenced by one or more “extreme” points (in the sense of extreme that we discussed earlier)

Why would we expect that? Is the median similarly vexed? Why or why not?

Web visits example

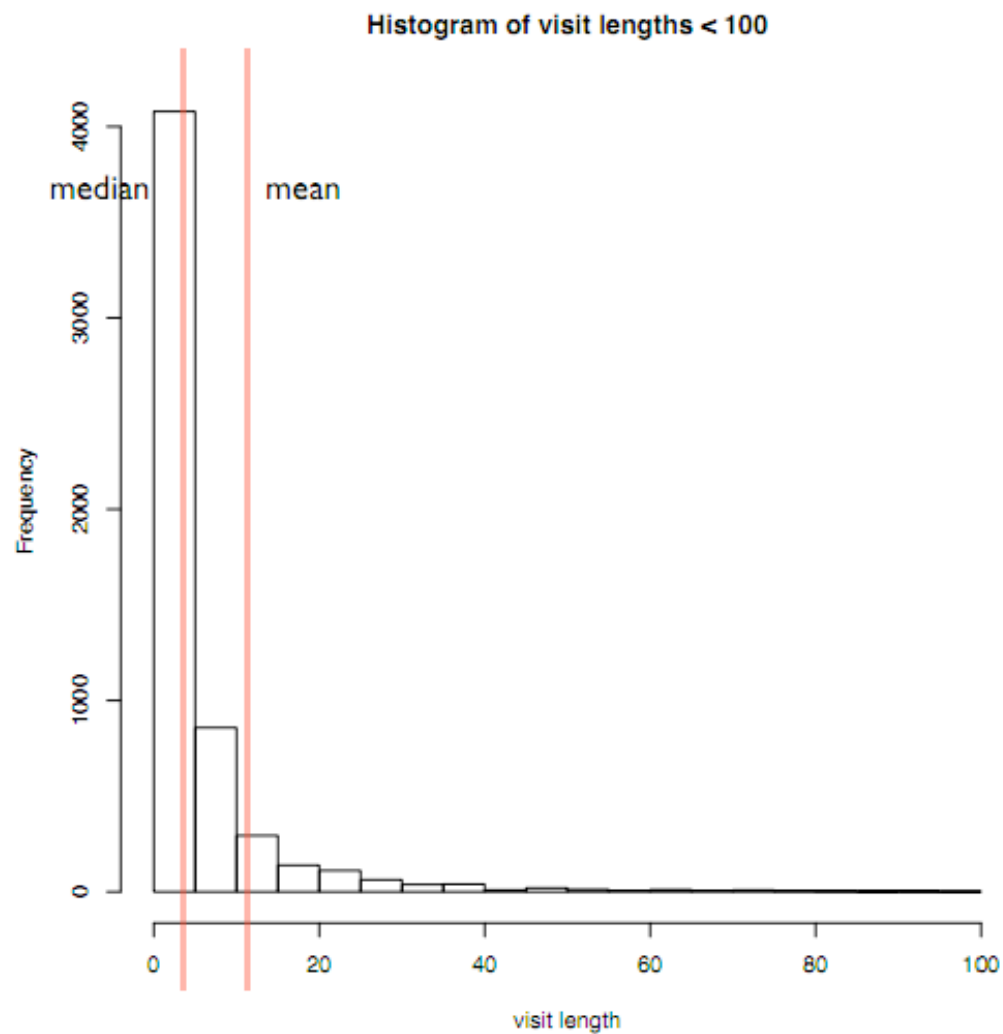
Here again is our poor histogram of web visits; remember that these data are heavily skewed to the right



Web visits example

Here we is the histogram
restricted to visits of length 100

The two red lines mark the
median and the mean; what do
you notice?

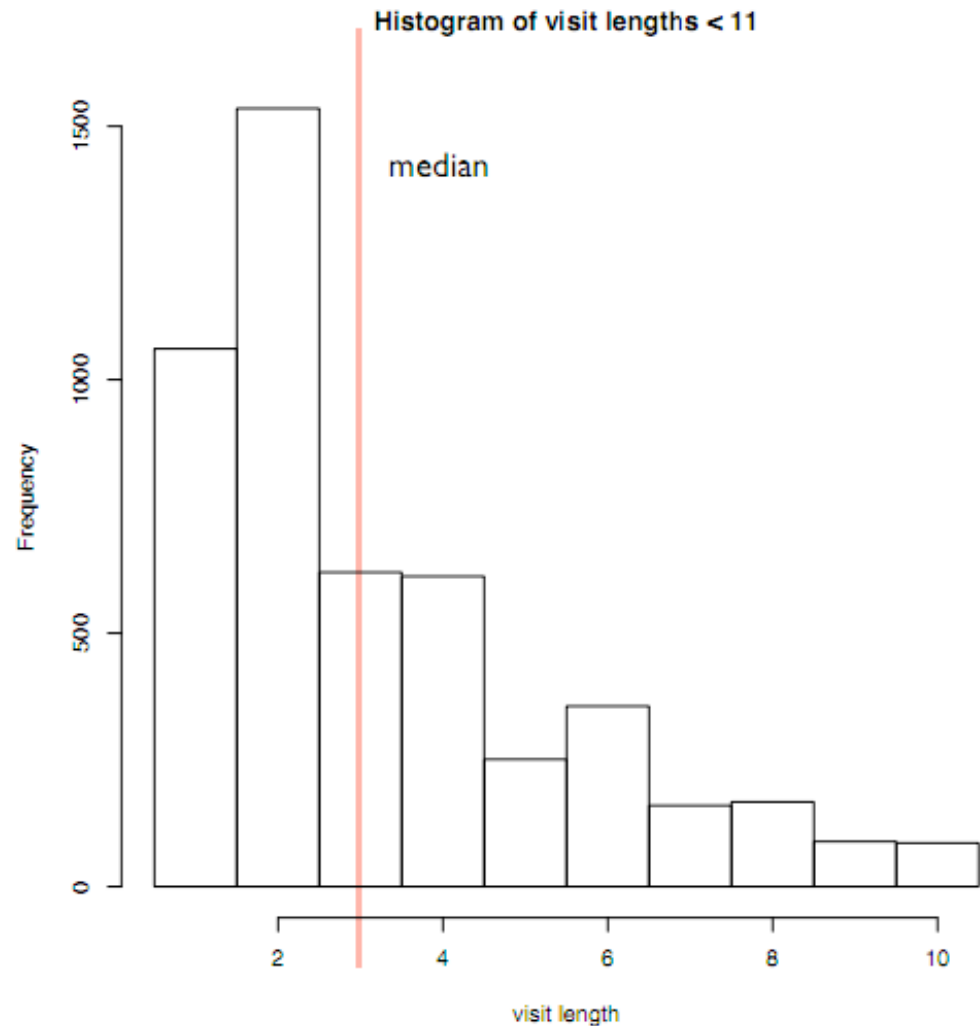


Web visits example

Only 824 of the 5,761 visits are longer than 10 hits

Here we present an even more severe restriction on the x-axis; is this a better way to summarize things?

The red lines again mark the median and mean; um, what's missing?



Example

Interestingly, the mean (11.12) doesn't even appear on a plot that contains $(5761-824)/5761 = 86\%$ of the data!

What notion of "center" is this providing, then?

Since a lot of statistical work involves using means, it is often suggested that we **transform the data** to give us something that is "better behaved" or has fewer "extreme" points

Inference

With the CDC BRFSS data we saw our first example of a survey -- **A random sample of the adult U.S. population** was asked a series of questions about their habits (um, health-related habits)

As we mentioned, we are often not just interested with the people who were surveyed, but instead what we might learn about the larger population -- **This is the process of statistical inference**

Today, we're going to look at one particular kind of study design and see how we might use the resulting data to draw conclusions

Types of studies

In the health and life sciences, we are faced with two kinds of studies that differ in terms the conditions under which data are collected

- In an **experimental study**, we impose some **change or treatment** and measure the result or response
- In an **observational study**, we simply **observe and measure something that has taken place or is taking place** (while trying not to cause any changes by our presence)

The kinds of inference you can make will depend on the type of study you conduct, **as well as its overall design or program for how data are to be collected** -- These kinds of considerations will lead to a range of (admittedly more technical) questions you should ask of a data set

In the last two lectures, we've talked about two data sets, the CDC Behavioral Risk Surveillance System and a list of courses from the registrar -- What kinds of "studies" do these represent

Clinical trials

A clinical trial is simply an experimental study in which two or more treatments are assigned to human subjects -- Experimental studies in all areas of biology have been greatly informed by procedures used in clinical trials

The clinical trial, however, has evolved considerably -- It was not always the “gold standard” of experimental designs -- Richard Doll (a well known epidemiologist who studied lung cancer) noted that before 1946

*... new treatments were almost always introduced on the grounds that in the hands of professor A or in the hands of a consultant at one of the leading teaching hospitals, the results in a small series of patients (seldom more than 50) had been superior to those recorded by professor B (or some other consultant) or by the same investigator previously. **Under these conditions variability of outcome, chance, and the unconscious (leave alone the conscious) in the selection of patients brought about apparently important differences in the results obtained**; consequently, there were many competing new treatments*

Clinical trials

In an attempt to improve the evaluation of different treatments, **Austin Bradford Hill began advocating a more systematic approach to designing clinical trials**; like Doll, he was frustrated with the quality of research at the time, going so far as to question the ethics of the existing system



Hill was the son of a distinguished physiologist; his hope of a medical career was thwarted by the onset of tuberculosis in 1917, and instead, while an invalid, he completed a degree in economics by correspondence

In 1927 Hill moved to the London School of Hygiene and Tropical Medicine and during the 1930s he researched mainly in occupational epidemiology; his renown in medical statistics started in 1937 with the publication of his textbook, *Principles of Medical Statistics*, based on a series of articles in the *Lancet*

Clinical Trials

Hill's work emphasizes the **practical snags and difficulties of applying statistics** in a clinical setting rather than theoretical minutiae -- It seems that his advice, while often statistically sound, was motivated by practical concerns

In terms of clinical trials, Hill argued for **well-specified study aims or outcomes**, and the consistent use of controls -- Patients were to be divided into two groups: **the “treatment” group would receive a new drug or procedure**, while **the “control” group would be prescribed the standard therapy**

Upon completion of the trial, researchers would examine the differences between the two groups, measuring outcomes, and determine if the proposed treatment is superior to the existing therapy

With his very practical approach to clinical work, Hill took a special interest in how patients were divided into the treatment and control groups -- **Left solely to physicians, he felt there could be a problem**

What was he worried about?

Clinical Trials

To remove the subjective bias of physicians in making assignments, some clinicians (including Hill, initially) had recommended the so-called **alternation method** -- That is, as patients appear at a clinic or study center, researchers alternately assign them to treatment or control

Other similar schemes include the assignment of a patient based on his or her initials or even their birthdate -- Taking Hill's very practical stance, do these methods completely remove potential bias?

Clinical trials

In 1948, Hill published a groundbreaking study on the effectiveness of streptomycin (an antibiotic) in treating pulmonary tuberculosis; here is how he assigned patients to the treatment and control groups



*Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on **random sampling numbers drawn up for each sex** at each centre by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-ordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and number. After acceptance of a patient by the panel, and before admission to the streptomycin centre, **the appropriate numbered envelope was opened at the central office: the card inside told if the patient was to be an S or C case**, and this information was then given to the medical officer of the centre. Patients were not told before admission that they were to get special treatment; C patients did not know throughout their stay in hospital that they were control patients in a special study; they were in fact treated as they would have been in the past, the sole difference being that they had been admitted to the centre more rapidly than was normal. Usually they were not in the same wards as S patients, but the same regimen was maintained."*

An aside: Some history

Following the **immense success of penicillin**, there was a great deal of research activity around detecting other potential antibiotics

Also, tuberculosis was the “**most important cause of death**” of young adults in Europe and North America at the time

Considerable laboratory work and some early experiments on patients suggested that Streptomycin would be an effective treatment for pulmonary tuberculosis

The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line, J. Crofton
<http://jrm.rsmjournals.com/cgi/reprint/99/10/531>

Clinical Trials

The tuberculosis study was the first time **randomization of treatments** was used in a clinical trial; after its publication, Hill wrote a series of articles describing its use

*In these articles, I had set out the need for controlled experiments in clinical medicine with groups chosen at random. At the outset, I think I pleaded that trials should be made using alternate cases. I suspect if (and its a very large IF) if that, in fact, were done strictly they would be random. I **deliberately left out the words "randomization" and "random sampling numbers" at that time, because I was trying to persuade the doctors to come into controlled trials in the very simplest form and I might have scared them off.** I think the concepts of "randomization" and "random sampling numbers" are slightly odd to the layman, or, for that matter, to the lay doctor, when it comes to statistics. I thought it would be better to get doctors to walk first, before I tried to get them to run.*

Memories of the British streptomycin trial in tuberculosis: The first randomized clinical trial, Sir Austin Bradford Hill

Clinical Trials

Through randomization (and the blinding of the physicians), Hill achieved his goal of reducing bias by allocating “the patients to the ‘treatment’ and ‘control’ groups in such a way that the two groups are initially equivalent in all respects relevant to the inquiry” -- He writes

It ensures that neither our personal idiosyncrasies (our likes or dislikes consciously or unwittingly applied) nor our lack of balanced judgement has entered into the construction of the different treatment groups—the allocation has been outside our control and the groups are therefore unbiased;

... it removes the danger, inherent in an allocation based on personal judgement, that believing we may be biased in our judgements we endeavour to allow for that bias, to exclude it, and that in doing so we may overcompensate and by thus ‘leaning over backward’ introduce a lack of balance from the other direction;

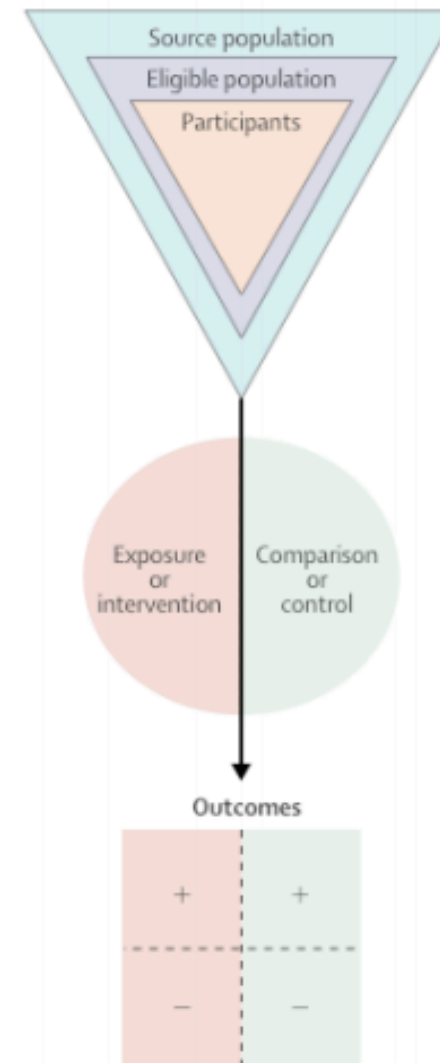
... and, having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity.

Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial. *Control Clin Trials* 1990;11:77–7

Randomized controlled trials

As an experiment then, the design is straightforward: participants are assigned randomly to either receive a treatment under study or a “control,” perhaps a placebo or a standard therapy

At the end of the study, an outcome is recorded for each participant; in some cases, scientists are evaluating whether a drug helps with a particular condition, say



The use of randomization

This kind of trial is not unique to the medical sciences; in fact, a number of “intervention” studies took place in the 1970s and 80s, all based on this idea of randomization

On the next slide, we present a few examples, and then a piece of a large list of such studies (245 in all) published in the late 80s

Examples of controlled trials of social programmes carried out in the United States during 1968-91

Trial	Years	Aim	Design	Outcomes assessed
Income maintenance experiments				
New Jersey-Pennsylvania negative income tax experiment (see Ferber and Hirsch, ¹⁹ Rossi and Lyall ²⁰)	1968-72	To study effects on work incentives of negative income tax	Random allocation of 1216 low income families to 8 intervention and 1 control groups	Participation in labour force; consumption expenditure; health and family behaviour; school attendance
Rural negative income tax experiment (see Ferber and Hirsch, ¹⁹ Maynard ²¹)	1970-2	To replicate above experiment in poor rural areas with non-intact families with female or male heads	Stratified random allocation of 809 low income families to 5 intervention and 1 control groups	Participation in labour force; consumption expenditure; health and family behaviour; school attendance
Gary income maintenance experiment (see Ferber and Hirsch, ¹⁹ Kehrer, ²² Kehrer and Wolin ²³)	1971-4	To study effects on participation in labour force and other family behaviours of different levels and forms of income maintenance, day care subsidies, and information and referral services	Stratified random allocation of 1799 low income, single parent families	Participation in labour force; consumption expenditure; health and family behaviour; school attendance; social and psychological attitudes
Denver-Seattle income maintenance experiments (see Ferber and Hirsch, ¹⁹ Rossi and Lyall ²⁰)	1970-91	To study effects on participation in labour force and other household behaviours of different levels and forms of income maintenance, job counselling, and training subsidies	Stratified random allocation of 2042 families allocated to 84 experimental "cells" with different combinations of support levels, tax rates, etc, and 1 control group	Participation in labour force; consumption expenditure; health and family behaviour; school attendance
Housing allowances				
Experimental housing allowance program (demand experiment) (see Bradbury and Downs, ²⁴ Friedman and Weinberg ²⁵)	1978-80	To study effects on households' housing behaviour of different forms of housing allowances and estimate of cost effectiveness	Stratified random allocation of 2241 low income households to 17 intervention groups with different housing allowance formulae and 2 control groups	Quality of housing; housing consumption behaviour; mobility
Supporting workers programmes				
Supported work program (see Ferber and Hirsch ¹⁹)	1975-8	To study effects and costs of supported work environment for disadvantaged workers	Random allocation of 6616 disadvantaged workers to 1 intervention and 1 control group	Participation in labour force; hours worked; total earnings
Texas worker adjustment program (see Bloom ²⁶)	1984-5	To study effects and costs of combination of job search assistance and occupational skills training for displaced workers	Random allocation of 2259 hard to employ individuals by random numbers table to 2 intervention and 1 control groups on 1 site, and 1 intervention and 1 control groups on 2 sites	Earnings; unemployment; unemployment benefits
Penal experiments				
Living insurance for ex-prisoners (LIFE) (see Rossi et al ²⁷)	1971-4	To study effects on re-arrests and participation in labour force of different levels of post-release payment and job assistance schemes	Stratified random allocation of 432 released prisoners to 3 intervention groups (payments only, counselling and placement only, both combined) and 1 control group	Arrests and convictions by type of offence; participation in labour force; health and living arrangements
Transitional aid research project (TARP) (see Rossi et al ²⁷)	1975-7	To study effects on re-arrests and participation in labour force of different levels of post-release payment and job assistance schemes	Stratified random allocation of 3982 released prisoners to 4 intervention groups with combinations of different payment periods and tax rates (3 groups) and job placement services (1 group) plus 2 control groups	Arrests and convictions by type of offence; participation in labour force; health and living arrangements

ORGANIZATION AND COMPOSITION OF THE BIBLIOGRAPHY

Experiments listed here are divided into 10 major categories which correspond to the type of program undergoing tests. They include:

- (1) Criminal and Civil Justice
- (2) Mental Health
- (3) Training and Education
- (4) Mass Communications
- (5) Information Collection, Transmission, and Retrieval
- (6) Research Utilization
- (7) Commerce, Industry, and Public Utilities
- (8) Social Welfare
- (9) Health Services and Medical Treatment
- (10) Fertility Control

Some categories are rather broad and so they are divided further. So for example, Criminal and Civil Justice has subsections for correctional treatments, judicial procedures, and traffic safety programs.

With each category, experiments are listed alphabetically, and only one reference is supplied for a given research project. That reference may, for example, refer to only one test in a long series of tests done by the same investigators. The single reference we have chosen to include is the one which we believe best characterizes the tests.

The categories we use are misleading to the extent that some experiments quite properly fall into more than one class. The experiments on Sesame Street, for example, might just as well be classified as research in mass communications as in education. In such case, we classify the experiment in accord with what we regard as the main purpose of research, so the tests of Sesame Street then are listed in Section 3 on Training and Education. For brevity's sake alone, we neither cross-reference entries nor place any entries into more than one category.

8. SOCIAL WELFARE: CASEWORK, HOUSING, INSURANCE AND OTHER SUPPORT PROGRAMS

- Behling, J. H. *An experimental study to measure the effectiveness of casework service*. Columbus, Ohio: Franklin County Welfare Department, 1961.
- Blenker, M., Jahn, J., & Wasser, E. *Serving the aging: An experiment in social work and public health nursing*. New York: Community Service Society of New York, 1964.
- Brown, G. E. (Ed.). *The multi-problem dilemma*. Metuchen, N.J.: Scarecrow Press, Inc., 1968.
- Eudey, E. *A move to home-ownership: Report on LIHD-2*. San Francisco, Calif.: San Francisco Development Fund, Inc., December 1970.
- Geismar, L. L., Lagay, B., Wolock, I., Gerhart, U., & Fink, H. *Early support of family life*. Metuchen, N.J.: Scarecrow, 1972.
- Hedrick, T., Oros, C., & Schmutte, G. Out of school youth employment demonstration project: 1977-78. Akron, Ohio: Akron Medine Employment Training Consortium, 1978 (Available from the authors, Kent State University, Kent, Ohio).
- Hill, D. B. & Veney, J. E. Kansas Blue Cross/Blue Shield Outpatient benefits experiment. *Medical Care*, 1970, 8, 143-158.
- Kuhl, P. H. *The Familycenter Project, and action research on socially deprived families*. Copenhagen, Denmark: The Danish National Institute of Social Research, 1969.
- Marin, R. C. *A comprehensive program for multiproblem families—Report on a four-year controlled experiment*. Rio Piedras, Puerto Rico: University of Puerto Rico, Institute of Caribbean Studies, 1969.
- McCabe, A. R., Selligman, A., Pyrke, M., Berkowitz, L., Kogan, L., & Pettiford, P. *The pursuit of promise: A study of the intellectually superior child in a socially deprived area*. New York: Community Service Society of New York, 1967.
- Medicus Systems Corporation. *Home care and day care services*. (Final Report) Chicago, Ill.: Medicus, 1978.

Fisher and randomization

It would be incorrect to suggest that the idea of randomization is due to Hill; Hill was working in the 1940's and 50's and became an advocate of randomization on fairly practical grounds (reducing bias)

In the 1920's and 1930's, R A Fisher (who we met in the first lecture, leaning thoughtfully over his calculator) was promoting the idea of randomization from a technical perspective; **to Fisher, randomization gave rise to valid statistical procedures**



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

Another aside: Fisher and Hill

There is, in fact, an interesting story that connects these two researchers; both were active in roughly the same time period and they were certainly aware of each other's work

They exchanged correspondence starting in 1929, “**Dear Sir**”; and then in 1931 “**Dear Fisher**” and “**Dear Bradford Hill**”; and then in 1940 “**My dear Fisher**” and “**My dear Bradford Hill**”; and then by 1952 “**My dear Ron**” and “**My dear Tony**” (Hill went by Tony)

But by 1958 they were back to “**Dear Fisher**” and “**Dear Bradford Hill**” as the two (Doll, a significant co-investigator with Hill) were on opposite sides in a dispute as to **whether or not smoking caused lung cancer**

From the point of view of our discussion, one of Fisher's main criticisms of the studies suggesting that smoking caused lung cancer was the fact that **they were entirely observational** -- He wanted a “properly randomized experiment” (which of course would be difficult as you can't force people to start smoking)

We will speak more about causation and what you can conclude from different types of studies over the next couple of lectures

Hill's tuberculosis trial

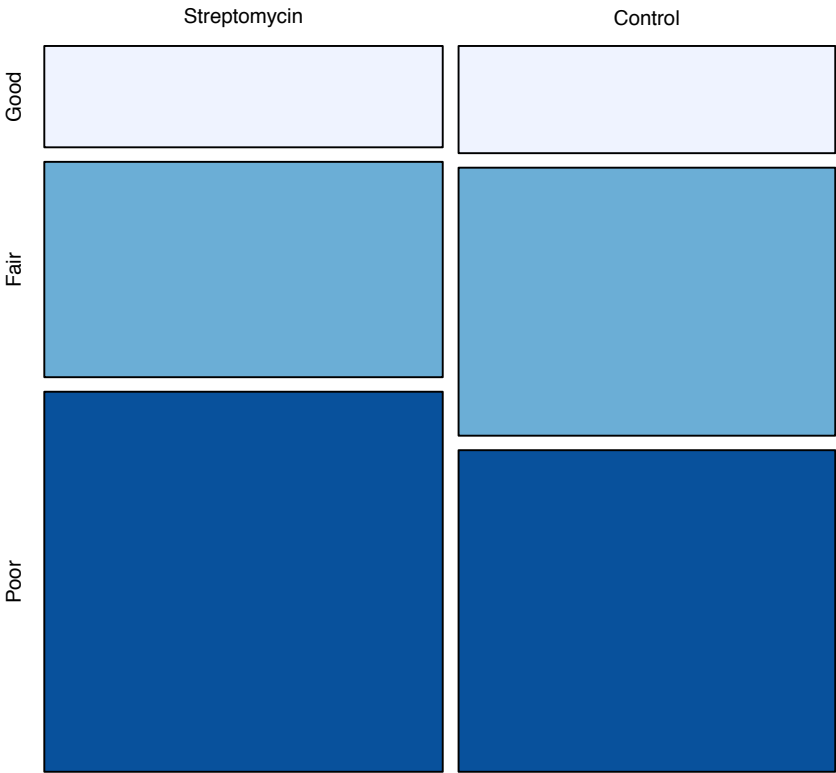
But back to the task at hand; here is a simple summary of the patients enrolled in Hill's tuberculosis trial -- as Hill hoped, the groups seem relatively well balanced in terms of their measured "condition"

TABLE I.—*Condition on Admission*

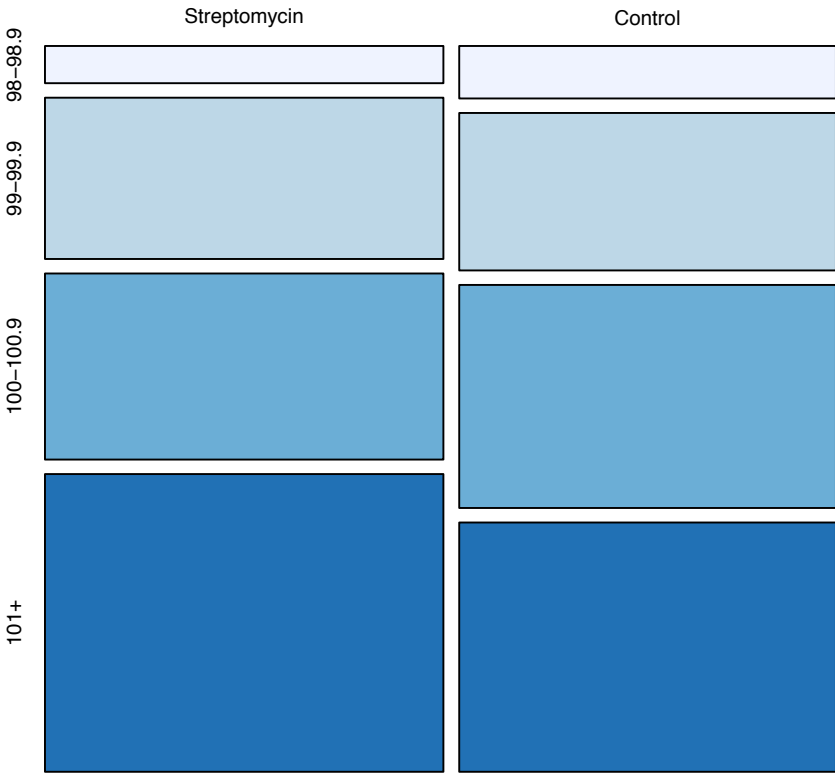
General Condi- tion	S Group	C Group	Max. Evening Temp. in First Week*	S Group	C Group	Sedimenta- tion Rate	S Group	C Group
Good .	8	8	98–98·9° F. (36·7–37·15° C.)	3	4	0–10	0	0
Fair ..	17	20	99–99·9° F. (37·2–37·75° C.)	13	12	11–20	3	2
Poor ..	30	24	100–100·9° F. (37·8–38·25° C.)	15	17	21–50	16	20
			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, after Hill



Max. Evening Temp., after Hill



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

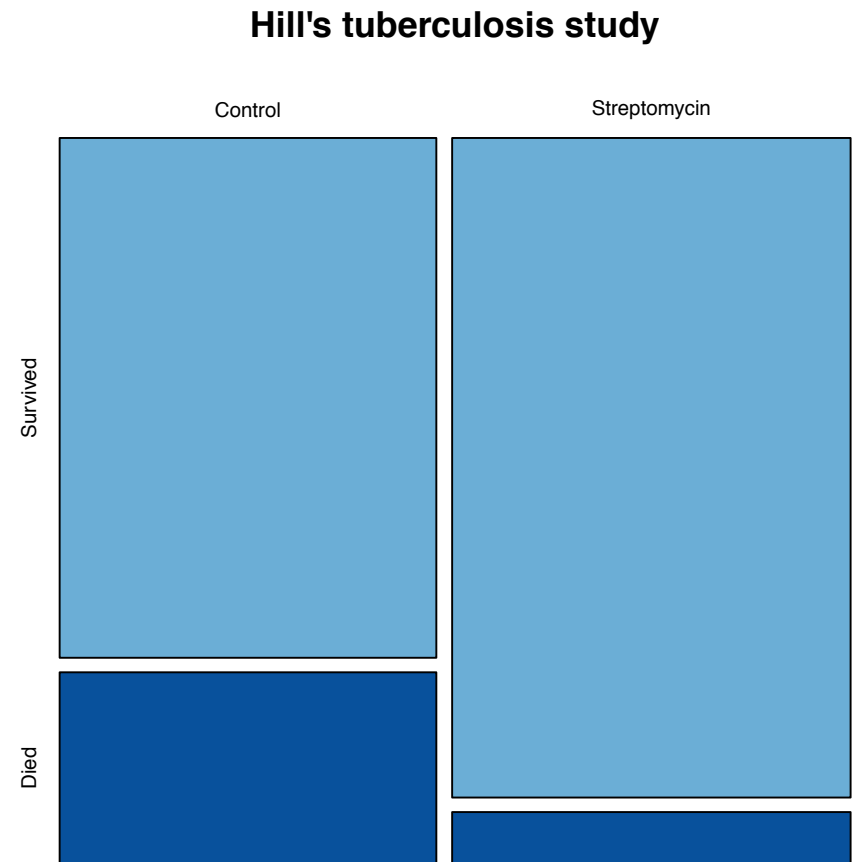
		Treatment		
		C	S	
Status	Survived	38	51	89
	Died	14	4	18
		52	55	107

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



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	89
	Died	14	4	18
		52	55	107

Some analysis with Hill's data

When you read about these kinds of trials in the medical literature, it is not uncommon **to work with a single figure of merit** -- Rather than look at the two conditional proportions, it is customary to look at their fraction

In this case, the ratio of the proportion of patients that died in the Streptomycin group (7.3%) to those that died in the Control group (27%) is 0.27 -- Streptomycin reduced the rate of mortality by nearly a quarter

This ratio is often called **the relative risk** -- The language comes from epidemiological studies where “treatment” is really exposure to some toxic substance and the outcome is not that you get better but that something horrible happens to you

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

Let's go back to the cartoon of a statistical inference problem that we started with in the first lecture

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%

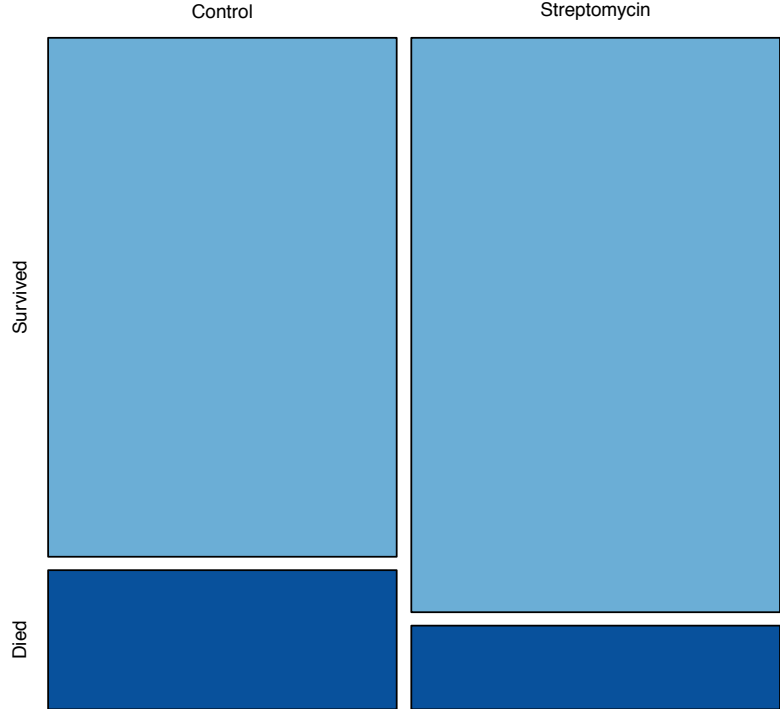
		Treatment		
		C	S	
Status	Survived	41	48	89
	Died	11	7	18
		52	55	107

Simulating random assignments

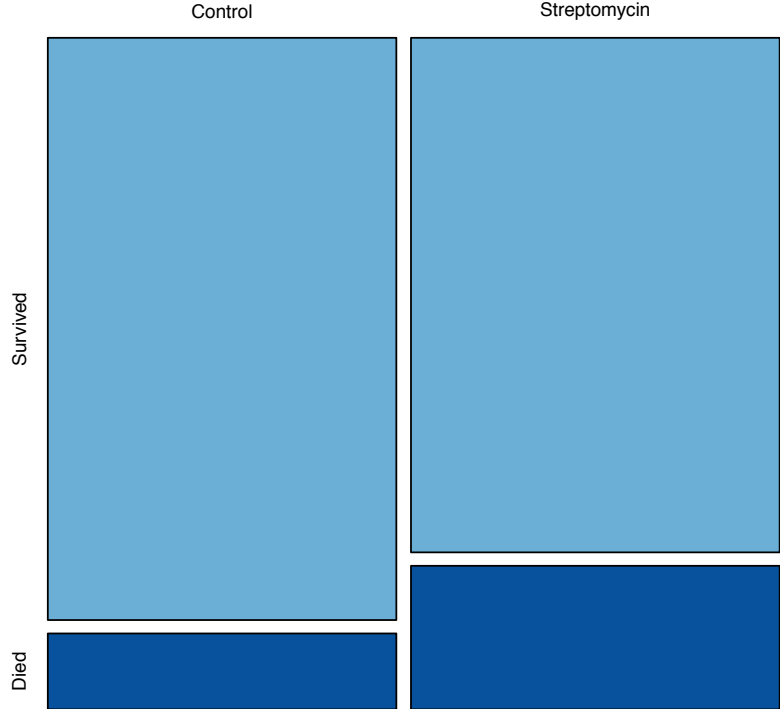
In this simulated table, we have the opposite, with 6/52 or 12% chance of dying under the control, and a 12/55 or 22% chance under Streptomycin; the treatment almost doubled the mortality rate among the participants

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

Simulated data



Simulated data



Simulating random assignments

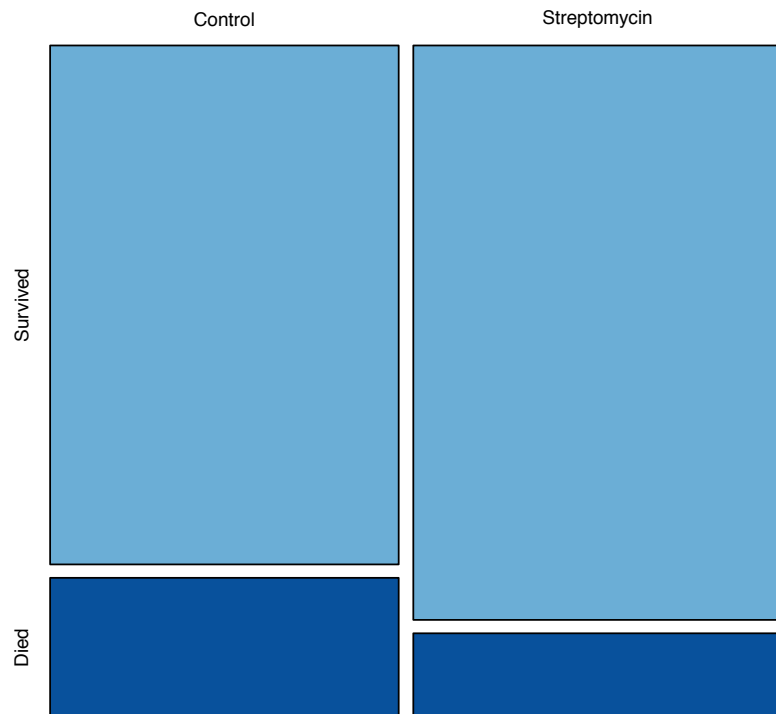
Notice that we only need to record **one piece of information for each trial, the number of deaths under Streptomycin** -- Knowing that we know all the other entries in the table

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?**

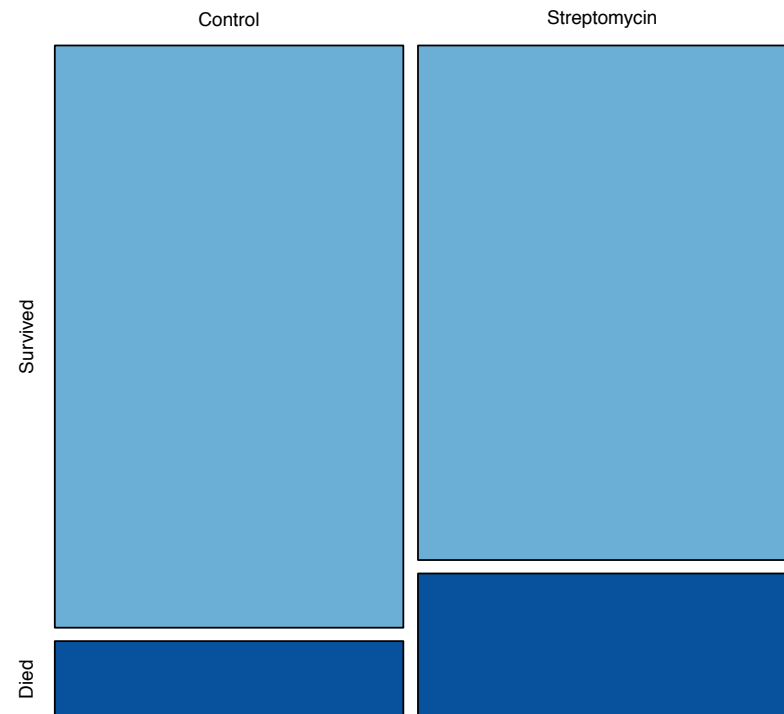
How would we figure this out?

Simulated data

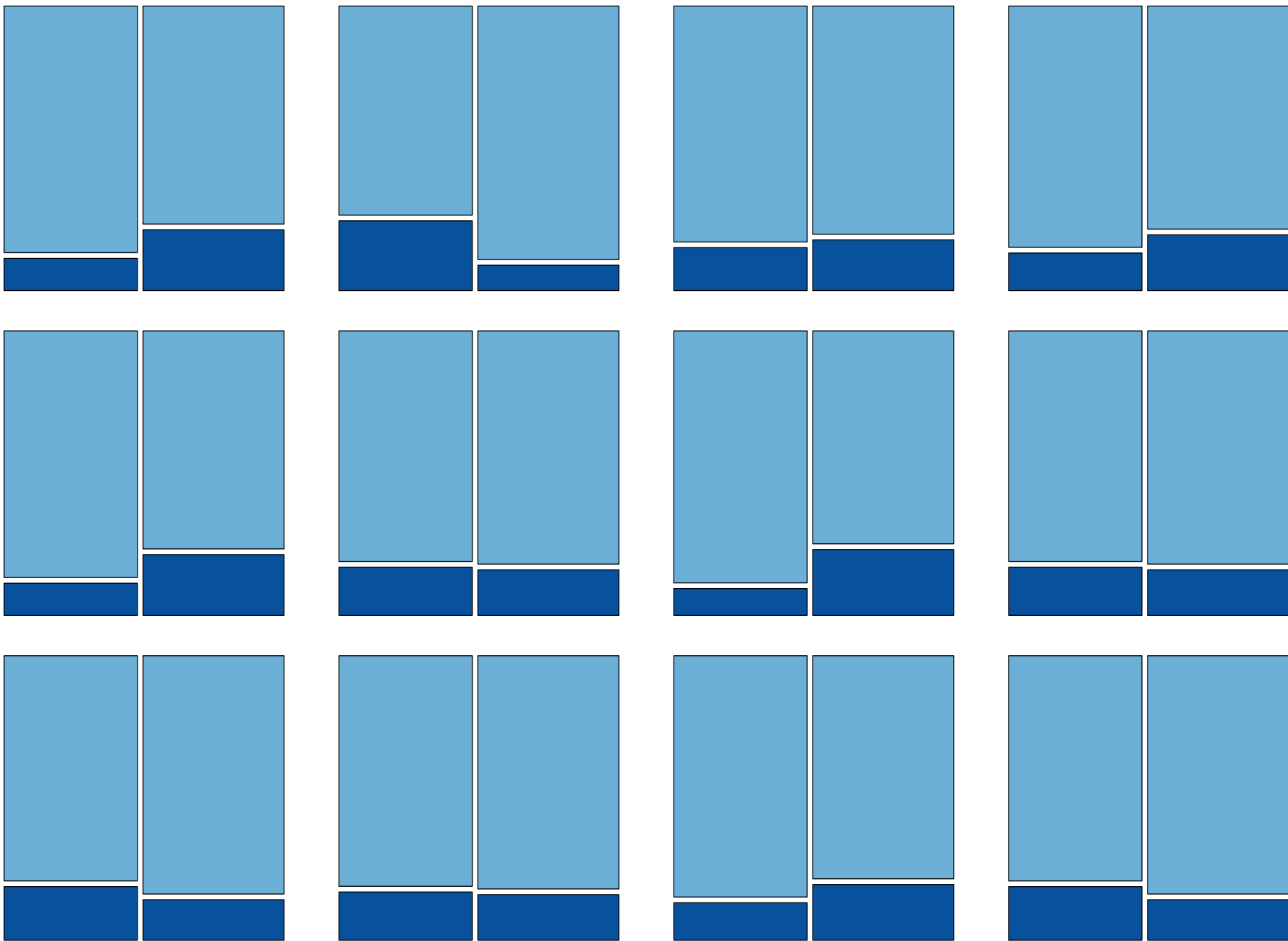


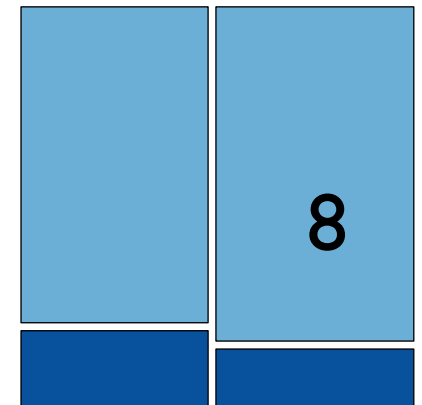
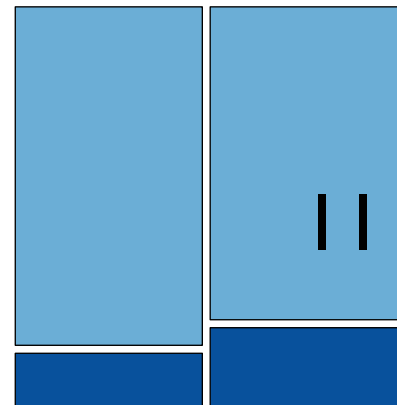
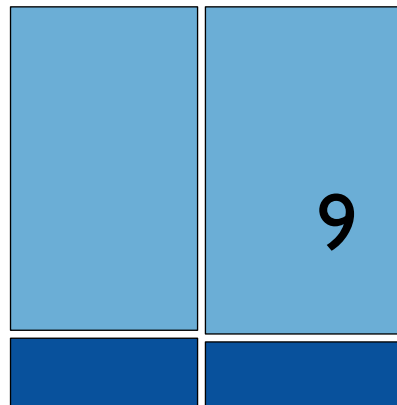
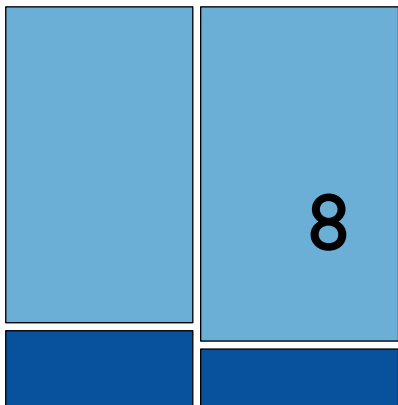
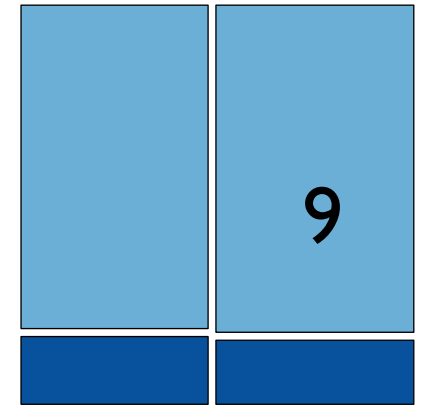
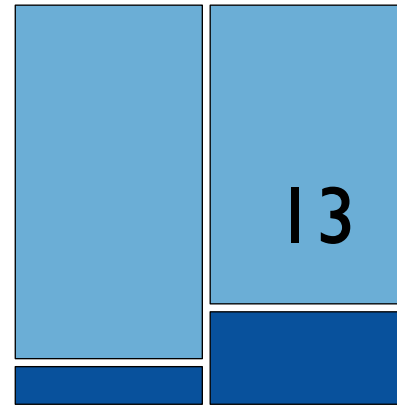
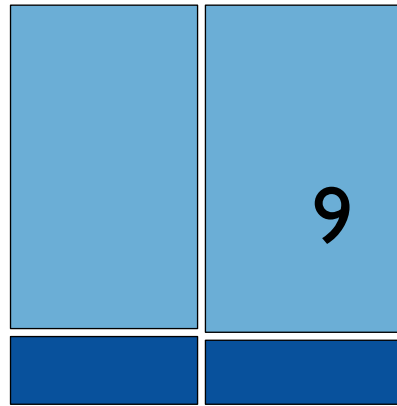
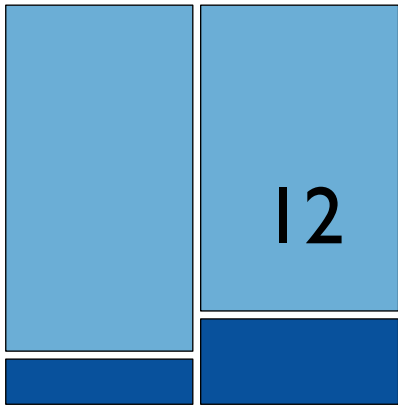
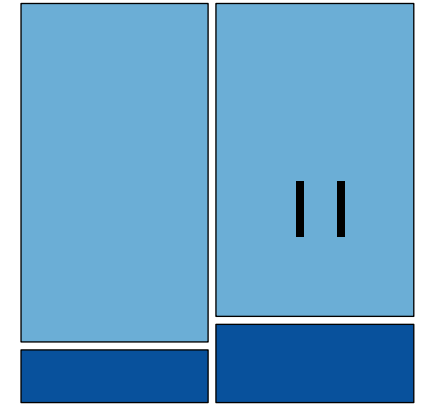
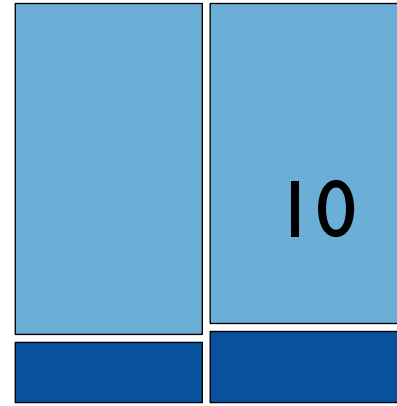
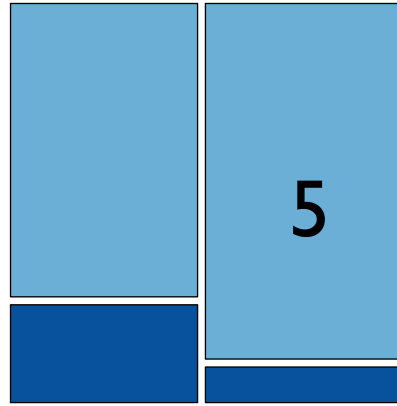
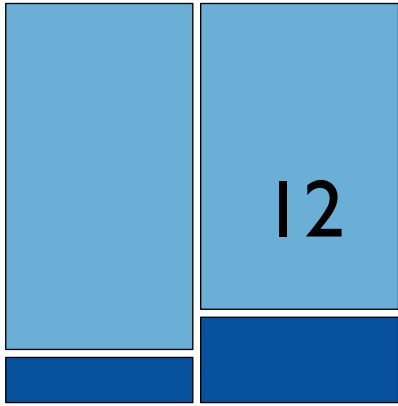
7 deaths in the Streptomycin group

Simulated data



12 deaths in the Streptomycin group





12

5

10

11

12

9

13

9

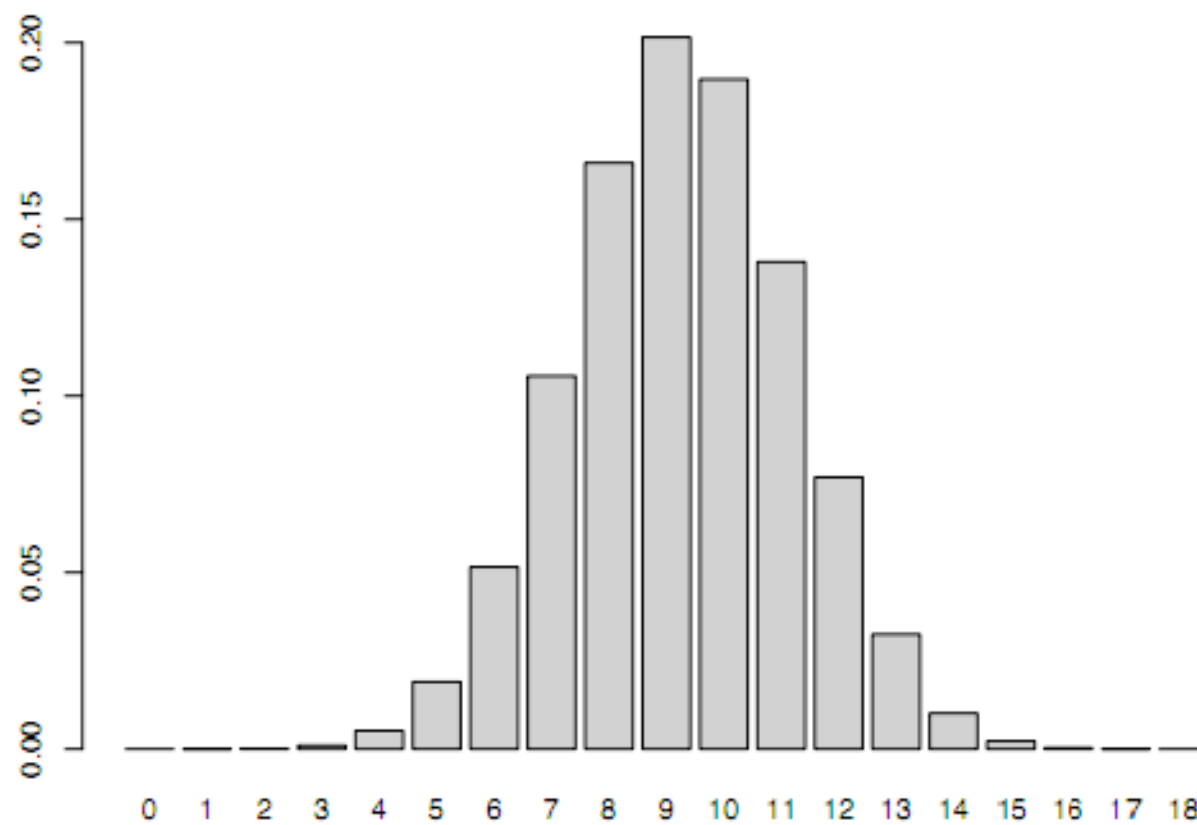
8

9

11

8

Proportion of simulated tables with n deaths under Streptomycin

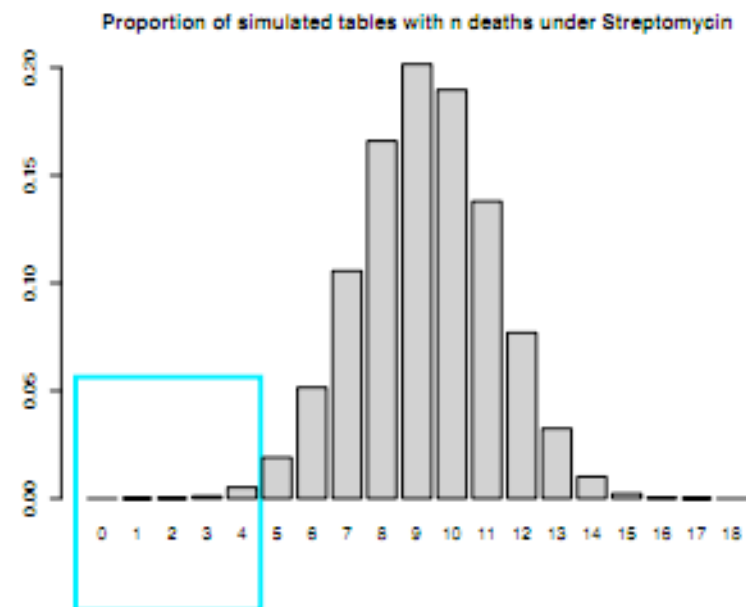


Simulating random assignments

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

Generalities

Notice that it was Hill's **random assignment that gives our analysis its validity** -- The way we collect data dictates the kinds of inferences we are allowed to make

However, **we have not said anything about what the effect of Streptomycin might be on patients outside the study** -- For that, we have to make more assumptions about how people were recruited into the trial (more on this next time)

This means that in addition to all the questions I had you ask about where data come from, you can now add a few technical ones -- **We are going to start paying attention to the role that randomness plays and, in particular, we will analyze as we randomized!**

Finally, the hypothesis testing framework is often referred to as **significance testing** in that we are attempting to establish the significance of some effect, the null hypothesis typically being that there is no effect

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).

Vioxx

Randomized controlled trials are common in medical research; let's have a look at a more recent case

Vioxx, an anti-inflammatory agent was introduced to the market in the late 1990s and was prescribed for the treatment of arthritis and acute pain

In 2000, the New England Journal of Medicine published the results from a large randomized controlled trial designed to examine whether patients receiving rofecoxib (Vioxx) would have fewer upper gastrointestinal "events" (perforations, ulcers, bleeding) than those taking naproxen (marketed as Aleve)

8,076 patients suffering from rheumatoid arthritis were randomized into two treatment groups: One received a twice daily dose of 50 mg of rofecoxib while the other received 500 mg of naproxen

COMPARISON OF UPPER GASTROIN AND NAPROXEN IN PATIENTS 1

CLAIRE BOMBARDIER, M.D., LOREN LAINE, M.D.,
RUBEN BURGOS-VARGAS, M.D., BARRY DAVIS, M.D., PH.D
CHRISTOPHER J. HAWKEY, M.D., MARC C
AND THOMAS J. SCHNITZER, M.D., F

ABSTRACT

Background Each year, clinical upper gastrointestinal events occur in 2 to 4 percent of patients who are taking nonselective nonsteroidal antiinflammatory drugs (NSAIDs). We assessed whether rofecoxib, a selective inhibitor of cyclooxygenase-2, would be associated with a lower incidence of clinically important upper gastrointestinal events than is the nonselective NSAID naproxen among patients with rheumatoid arthritis.

Methods We randomly assigned 8076 patients who were at least 50 years of age (or at least 40 years of age and receiving long-term glucocorticoid therapy) and who had rheumatoid arthritis to receive either 50 mg of rofecoxib daily or 500 mg of naproxen twice daily. The primary end point was confirmed clinical upper gastrointestinal events (gastroduodenal perforation or obstruction, upper gastrointestinal bleeding, and symptomatic gastroduodenal ulcers).

Results Rofecoxib and naproxen had similar efficacy against rheumatoid arthritis. During a median follow-up of 9.0 months, 2.1 confirmed gastrointestinal events per 100 patient-years occurred with rofecoxib, as compared with 4.5 per 100 patient-years with naproxen (relative risk, 0.5; 95 percent confidence interval, 0.3 to 0.6; $P < 0.001$). The respective rates of com-

Vioxx

In addition to GI problems, the researchers considered a variety of possible side-effects from taking rofecoxib (R) or naproxen (N); here we present a two-by-two table of patients who experienced cardiovascular adverse events (CE)

		Treatment		
		N	R	
Status	no CE	4010	4002	8012
	CE	19	45	64
		4029	4047	

Vioxx

Here we see that in the naproxen group, 19/4029 or 0.5% patients experienced cardiovascular adverse events, while under rofecoxib 45/4047 or 1.1% of patients had problems; the chance that a patient develop CE under rofecoxib is over twice as high

		Treatment		
		N	R	
Status	no CE	4010	4002	8012
	CE	19	45	64
		4029	4047	

Vioxx

As with Hill's data, this number seems convincing; and yet, we should ask whether or not these results could be produced by pure chance

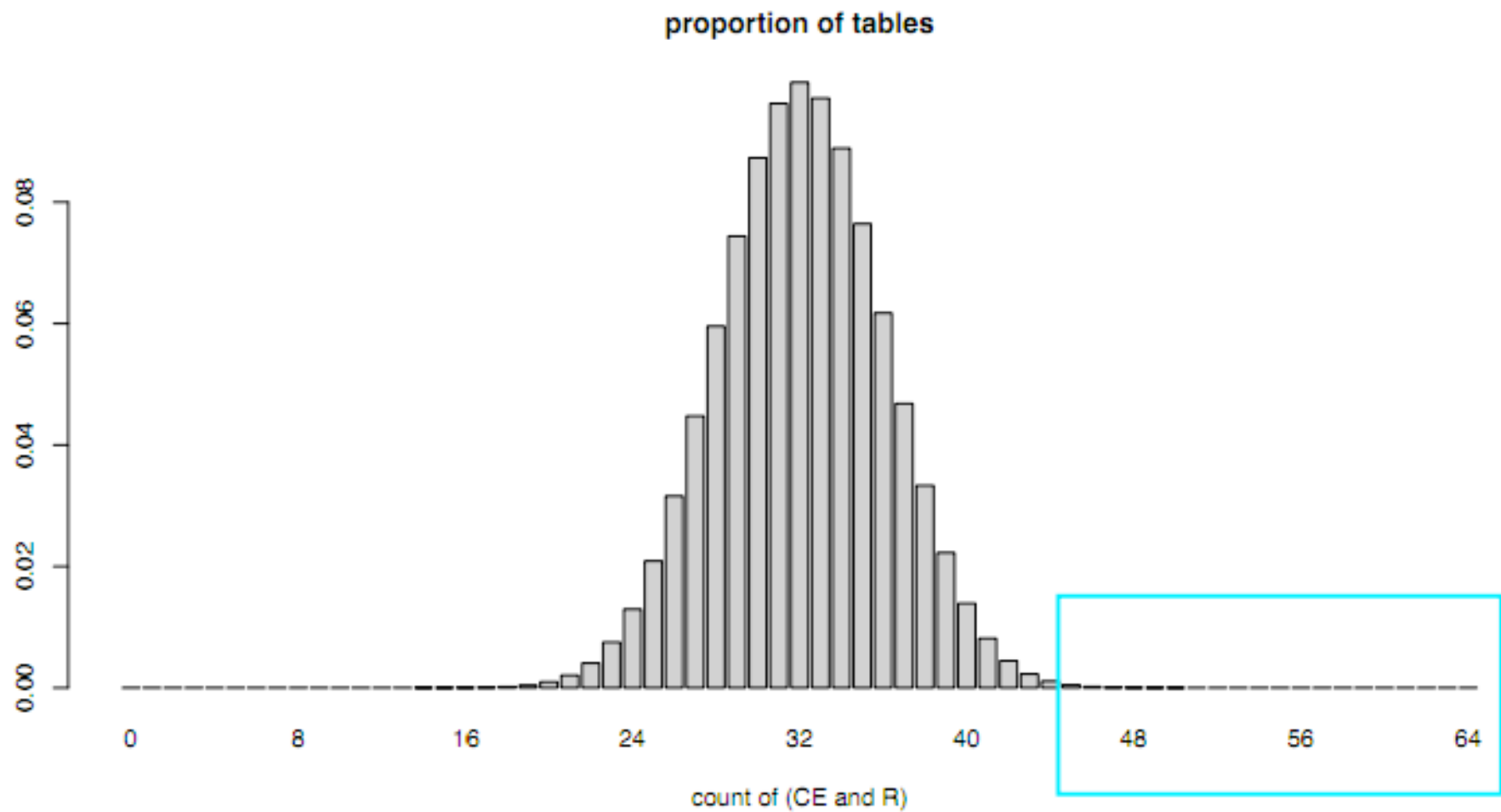
The same analysis framework holds up here; under the null hypothesis that patients are equally likely to have a CE under rofecoxib as naproxen, we can repeat the assignment of patients to treatment groups and examine the resulting tables

We're specifically interested in seeing how many tables yield results that are as strong or stronger than what we observed; in this case, under different randomizations of the patients, how often do we see tables with 45 or more deaths under Vioxx?

Well, we can simulate..

Vioxx

In this case, the data provide strong evidence that the proportions we observed were not due to the random assignment; the P-value here is 0.0000004



Vioxx

The evidence here is seems very strong; Merck, the manufacturer of Vioxx, however, concluded instead that **it wasn't that Vioxx was responsible for more adverse events, it was that naproxen helped reduce them**

As the research community debated the meaning of this particular trial, several ongoing trials began to exhibit similar problems, this time with placebos as the control treatment; eventually, **in 2004, Merck withdrew the drug from the market, citing increased risk of cardiovascular adverse events**

The debate then turned from whether Vioxx was harmful to what Merck scientists and senior management knew of these hazards, and **when they knew it**

	Protocol number	Submitted to FDA (year)	Treated disorder (number of patients)	Intervention (number of patients)		Duration (weeks)
				Rofecoxib	Control	
Ehrich et al (1999) ²³	010	1998	Osteoarthritis (n=145)	Rofecoxib 25 mg (n=73)	Placebo (n=72)	6
Laine et al (1999) ²⁴	044	1998	Osteoarthritis (n=742)	Rofecoxib 25 mg (n=195) Rofecoxib 50 mg (n=186)	Placebo (n=177) Ibuprofen 2400 mg (n=184)	24
Schnitzer et al (1999) ²⁴	068	2001	Rheumatoid arthritis (n=500)	Rofecoxib 25 mg (n=171) Rofecoxib 50 mg (n=161)	Placebo (n=168)	8
Extension of Schnitzer et al (1999) ²⁴	068-P2	2001	Rheumatoid arthritis (n=544)	Rofecoxib 25 mg (n=235) Rofecoxib 50 mg (n=223)	Naproxen 1000 mg (n=86)	44
Bombardier et al (2000) ⁴	088c	2000	Rheumatoid arthritis (n=8076)	Rofecoxib 50 mg (n=4047)	Naproxen 1000 mg (n=4029)	Up to 56
Cannon et al (2000) ¹⁴	035	1998	Osteoarthritis (n=784)	Rofecoxib 12.5 mg (n=259) Rofecoxib 25 mg (n=257)	Diclofenac 150 mg (n=268)	52
Day et al (2000) ²⁷	040	1998	Osteoarthritis (n=809)	Rofecoxib 12.5 mg (n=244) Rofecoxib 25 mg (n=242)	Placebo (n=74) Ibuprofen 2400 mg (n=249)	6
Hawkey et al (2000) ²⁵	045	1998	Osteoarthritis (n=775)	Rofecoxib 25 mg (n=195) Rofecoxib 50 mg (n=193)	Placebo (n=194) Ibuprofen 2400 mg (n=193)	24
Saag et al (2000) ¹⁸	033	1998	Osteoarthritis (n=736)	Rofecoxib 12.5 mg (n=219) Rofecoxib 25 mg (n=227)	Placebo (n=69) Ibuprofen 2400 mg (n=221)	6
Saag et al (2000 A) ¹⁸	034	1998	Osteoarthritis (n=693)	Rofecoxib 12.5 mg (n=231) Rofecoxib 25 mg (n=232)	Diclofenac 150 mg (n=230)	52
Ehrich et al (2001) ²⁸	029	1998	Osteoarthritis (n=523)	Rofecoxib 12.5 mg (n=144) Rofecoxib 25 mg (n=137) Rofecoxib 50 mg (n=97)	Placebo (n=145)	6
Unpublished extension of Ehrich et al (2001) ²⁸	029-10	1998	Osteoarthritis (n=438)	Rofecoxib 12.5 mg (n=102) Rofecoxib 25 mg (n=146) Rofecoxib 50 mg (n=100)	Diclofenac 150 mg (n=90)	26
Geba et al (2001) ²⁹	090	2000	Osteoarthritis (n=978)	Rofecoxib 12.5 mg (n=390)	Placebo (n=196) Nabumetone 1000 mg (n=392)	6
Truitt et al (2001) ³¹	058	1998	Osteoarthritis (n=341)	Rofecoxib 12.5 mg (n=118) Rofecoxib 25 mg (n=56)	Placebo (n=52) Nabumetone 1500 mg (n=115)	6
Truitt et al (2001 A) ³¹	096	2001	Rheumatoid arthritis (n=909)	Rofecoxib 12.5 mg (n=148) Rofecoxib 25 mg (n=311)	Placebo (n=301) Naproxen 1000 mg (n=149)	12
Unpublished extension of Truitt et al (2001 A) ³¹	096-P2	2001	Rheumatoid arthritis (n=673)	Rofecoxib 25 mg (n=335) Rofecoxib 50 mg (n=114)	Naproxen 1000 mg (n=224)	40
Geusens et al (2002) ²⁶	097	2001	Rheumatoid arthritis (n=1058)	Rofecoxib 25 mg (n=315) Rofecoxib 50 mg (n=297)	Placebo (n=299) Naproxen 1000 mg (n=147)	12
Unpublished extension of Geusens et al (2002) ²⁶	097-P2	2001	Rheumatoid arthritis (n=893)	Rofecoxib 25 mg (n=253) Rofecoxib 50 mg (n=392)	Naproxen 1000 mg (n=248)	40
Hawkey et al (2003) ²⁷	098/103	-	Rheumatoid arthritis (n=660)	Rofecoxib 50 mg (n=219)	Placebo (n=221) Naproxen 1000 mg (n=220)	12
Katz et al (2003) ³⁰	-	-	Chronic low back pain (n=690)	Rofecoxib 25 mg (n=233) Rofecoxib 50 mg (n=229)	Placebo (n=228)	4
Lisse et al (2003) ²¹	102	2000	Osteoarthritis (n=5586)	Rofecoxib 25 mg (n=2799)	Naproxen 1000 mg (n=2787)	12
Kivitz et al (2004) ²²	085	2000	Osteoarthritis (n=1042)	Rofecoxib 12.5 mg (n=424)	Placebo (n=208) Nabumetone 1000 mg (n=410)	6

A problem with thresholds

In 2003, before Merck pulled Vioxx from the shelves, Lisse et al published a report in the Annals of Internal Medicine; **with Merck funding, this group again compared rofecoxib to naproxen** for relief of osteoarthritis pain

The bulk of the paper was concerned with “gastrointestinal tolerability” of rofecoxib; but they did mention some cardiovascular problems; specifically, **five people died of myocardial infarction (heart attack) while on Vioxx, while only 1 did from the naproxen group**

In the study they say that “a Fisher exact test was used to compare incidence of confirmed... cardiovascular events” **but that the results were not significant** (evidently appealing to a 5% cutoff level for the P-value)

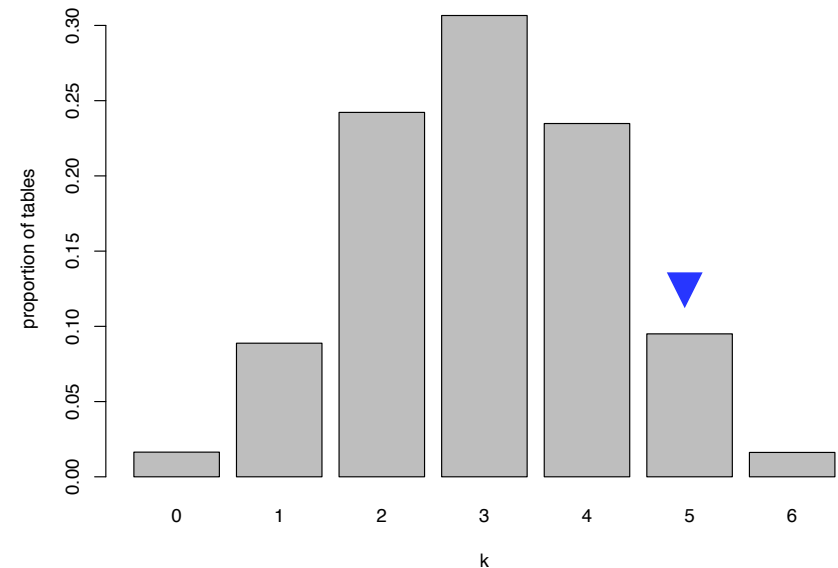
Lisse et al (2003)

Here are the results in tabular form -- While small, the conditional proportion of people having MI under Vioxx is five times that for naproxen

		Treatment		
		rofecoxib	naproxen	
Status	no MI	2780	2771	5551
	MI	5	1	6
		2785	2772	5557

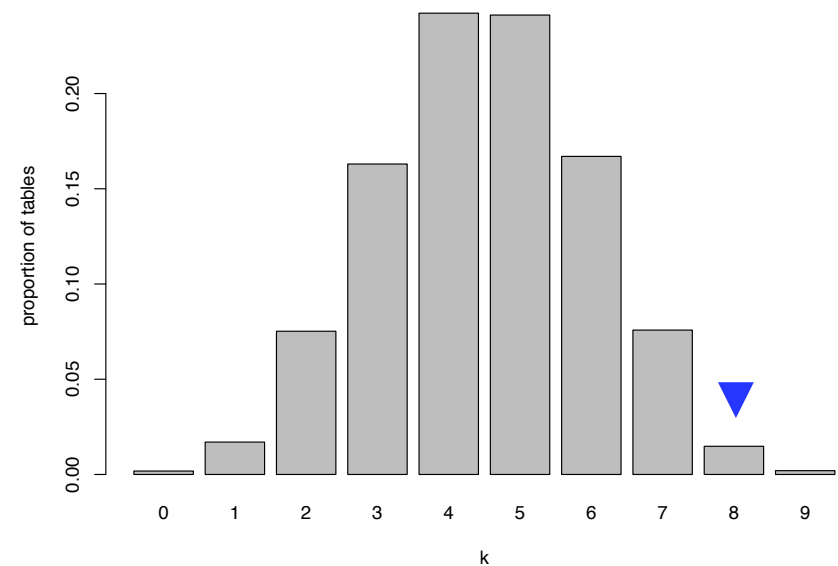
The problem with thresholds

At the right (top) we show the null distribution associated with Lisse's data; clearly the results are not very extreme given the randomness involved (a P-value of 0.11)



The problem, comes, however, as part of a Federal investigation into Vioxx, **Merck was forced to disclose three more deaths in the Vioxx group**; this changes the distribution to that shown in lower panel

In this case, the P-value is 0.02, smaller than the 0.05 cutoff and now significant



The problem with thresholds

We appeal to statistics because we want some kind of a simple procedure for telling truth from fiction -- This leads to cutoffs on P-values

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 a significance test

Thresholds

In the next lecture, we'll examine the use of thresholds in a systematic way -- We'll see how one pair of statisticians, Neyman and Pearson, formalized this testing framework into **one of making decisions**

*"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