Lecture 3: Middle

Last time

We examined some basic graphical and numerical summaries, essentially covering and extending Chapter 6 of your text -- We saw a series of plots, some familiar, some new, but all aimed at helping you read a data set

Some plots were designed to illustrate the “structure” of a single variable (symmetry, uni- or multi-modality, skew) while others helped us assess association between two variables

Today

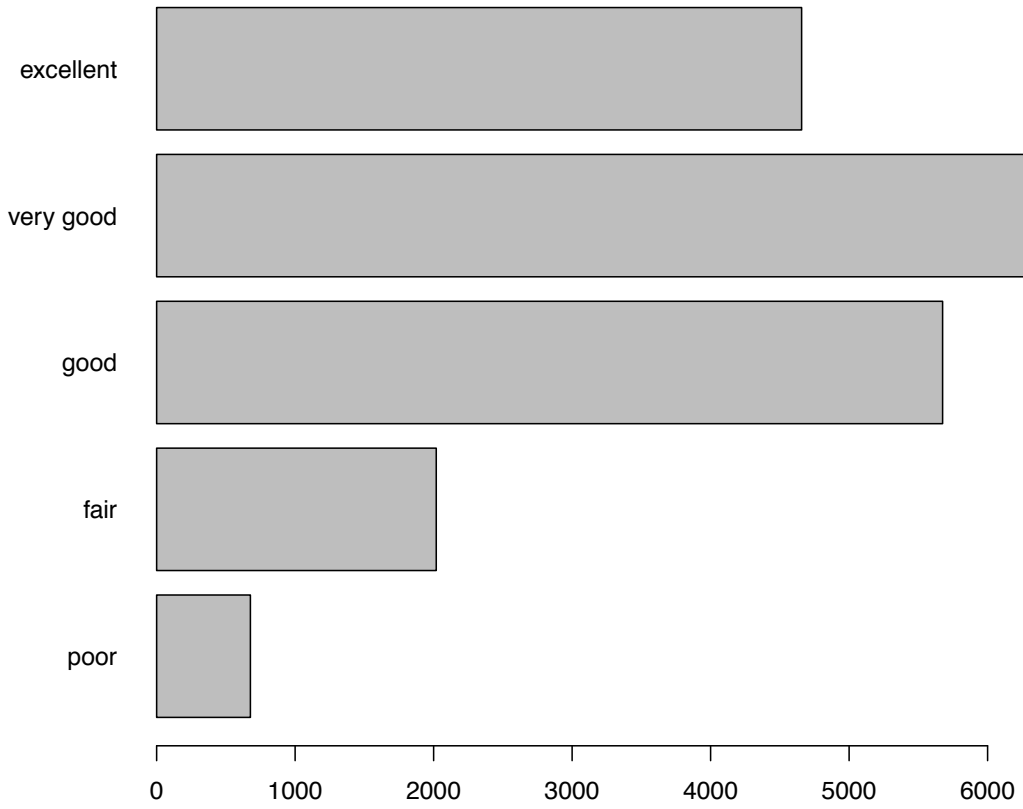
We will start by taking up our discussion of developing a boxplot for more than a single variable, a graphic to summarize the shape of a 2-dimensional point cloud

We will then examine tools for viewing (continuous) data in 2 or more dimensions, spending some time with projections and linked displays

We'll end with some material for your (first) homework assignment -- The subject of graphics will not end here, however, in that we'll also examine spatial (map-based) data as well as text as data later in the term

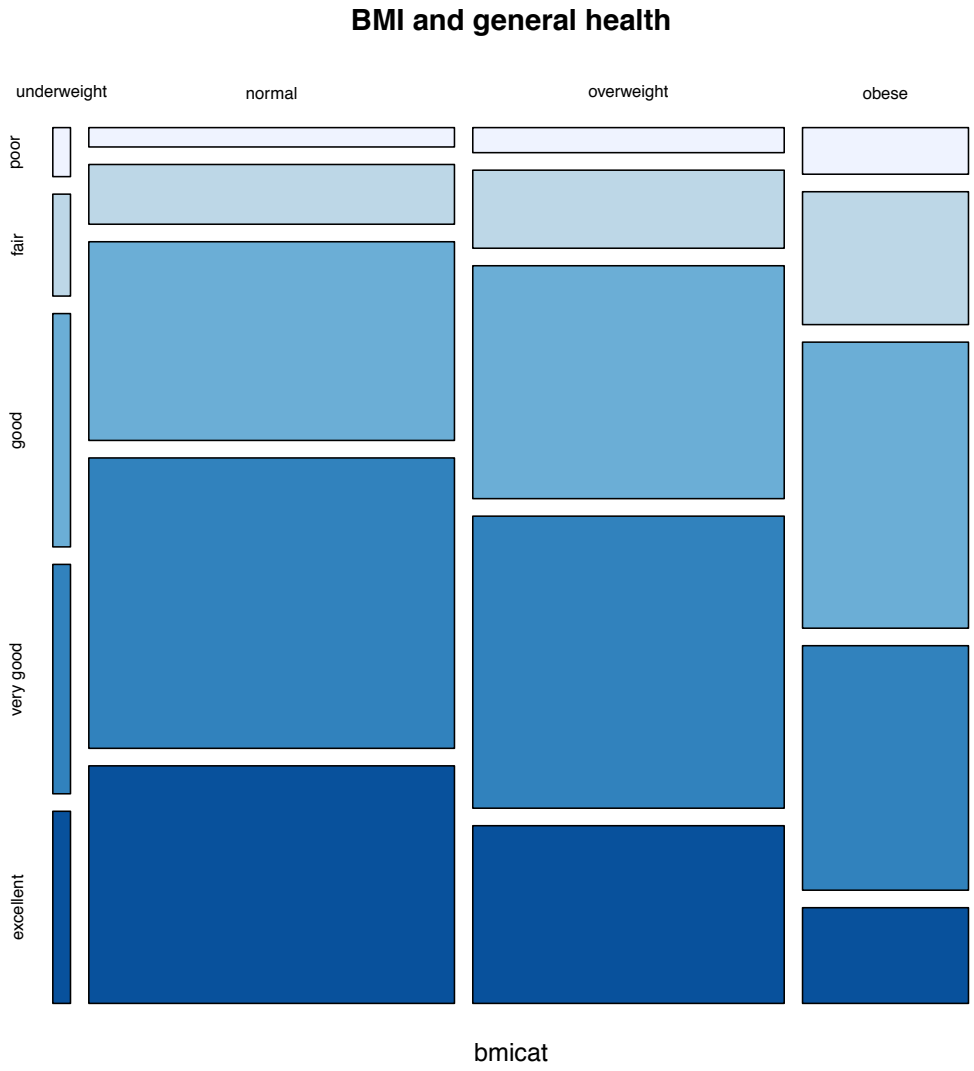
Frequency displays

We began examining some simple graphical devices to display the counts per category for a qualitative variable



Frequency displays

And we introduced a new display, a mosaic plot, to exhibit association between two qualitative variables



Frequency displays

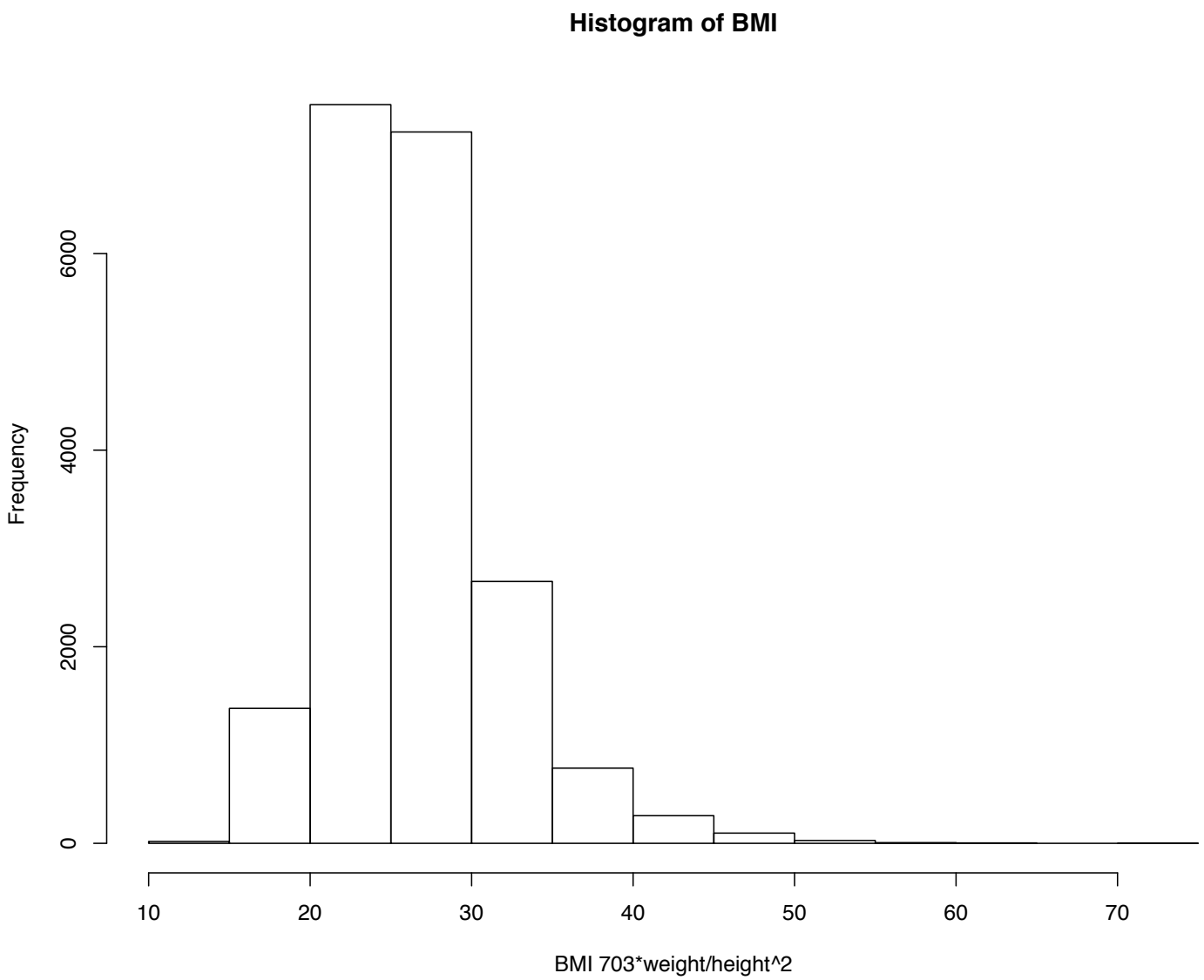
In the previous case, we **discretized a quantitative variable**, respondents' BMI values, into categories and then used a mosaic plot to exhibit possible associations with variables like the respondents' general health

If we didn't know about the CDC's categories (obese, overweight, normal, etc.), could we still employ this technique? **How would we divide the continuous BMI measure into categories?**

Continuous variables

Frequency displays for continuous variables help us examine **the "shape" of a data set** -- Histograms, for example, function in the same way as barplots after binning the data into (for our purposes, equally sized) intervals

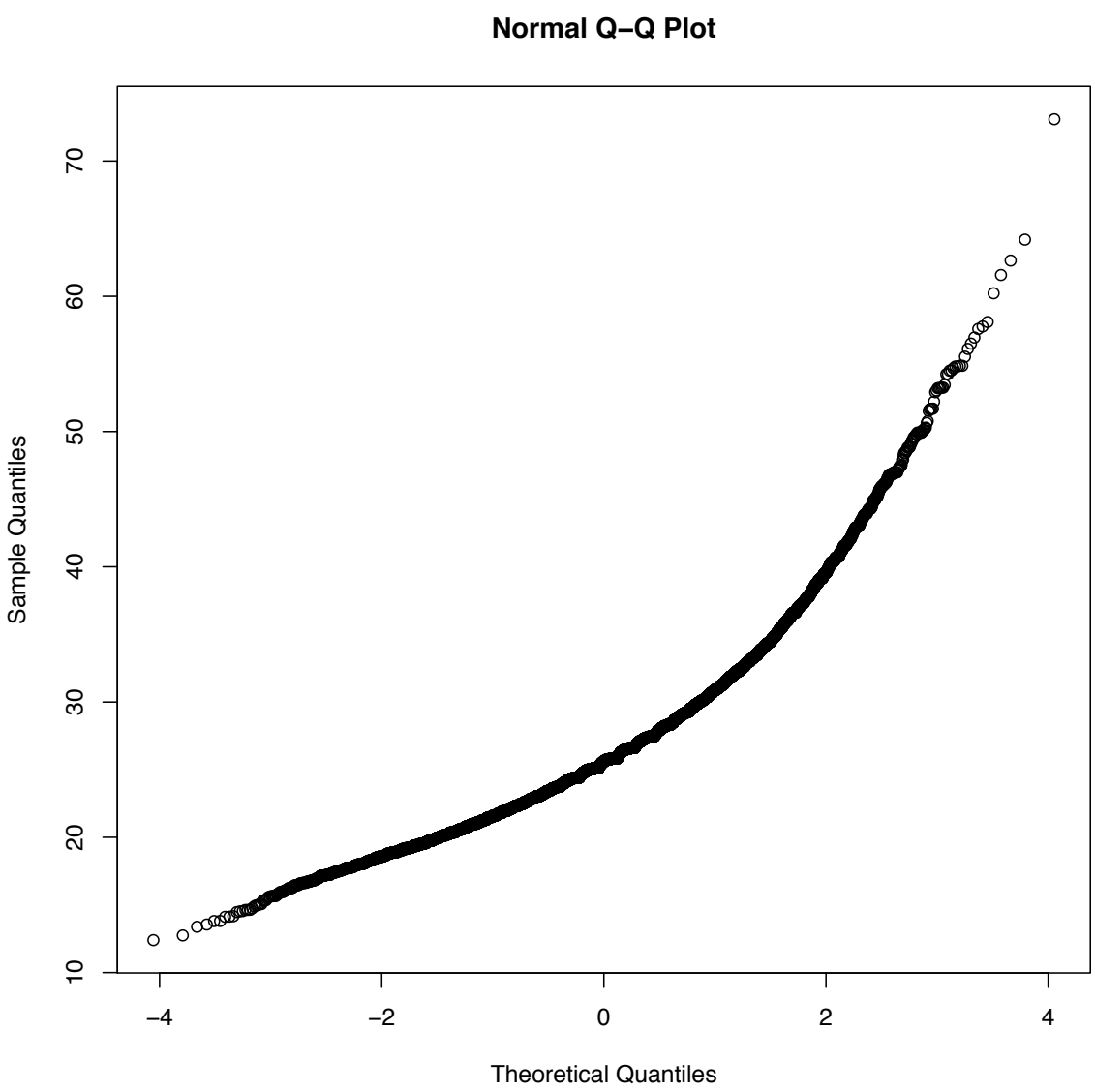
We describe these plots with terms like symmetric or skewed, uni- or multi-modal -- The shape is a story, and opens up questions for further study


Histogram of BMI
Frequency
0
2000
4000
6000
10
20
30
40
50
60
70
BMI 703*weight/height^2

Continuous variables

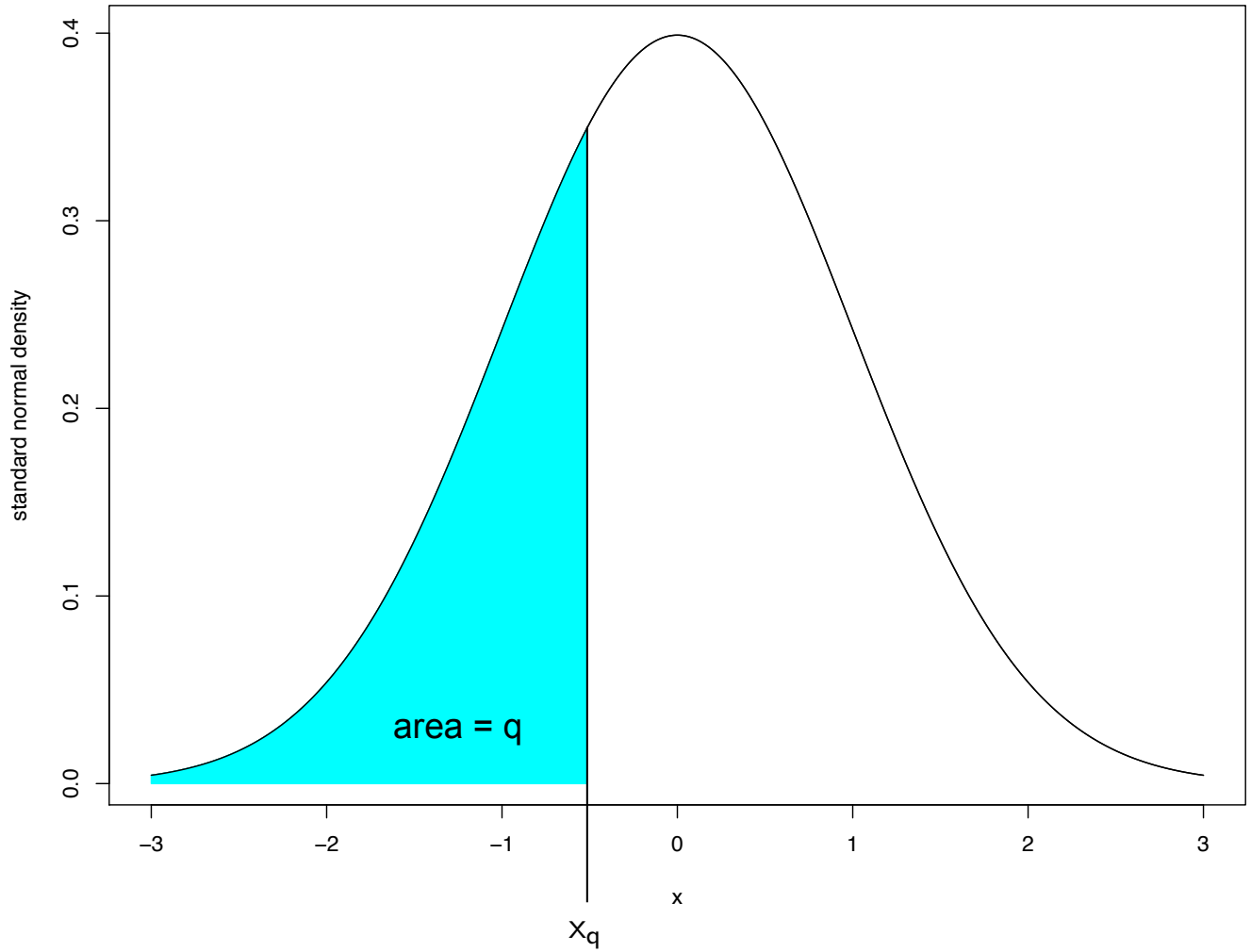
Last time we informally presented the construction of another kind of graphic, a **quantile-quantile plot** that allows us to compare a known distributional shape to that of our data

In the last lecture, we used the normal distribution as a kind of ruler...


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
10
20
30
40
50
60
70
−4
−2
0
2
4

Quantiles

Last time we made a passing reference to quantiles -- The qth quantile of a probability distribution is the point x_q such that

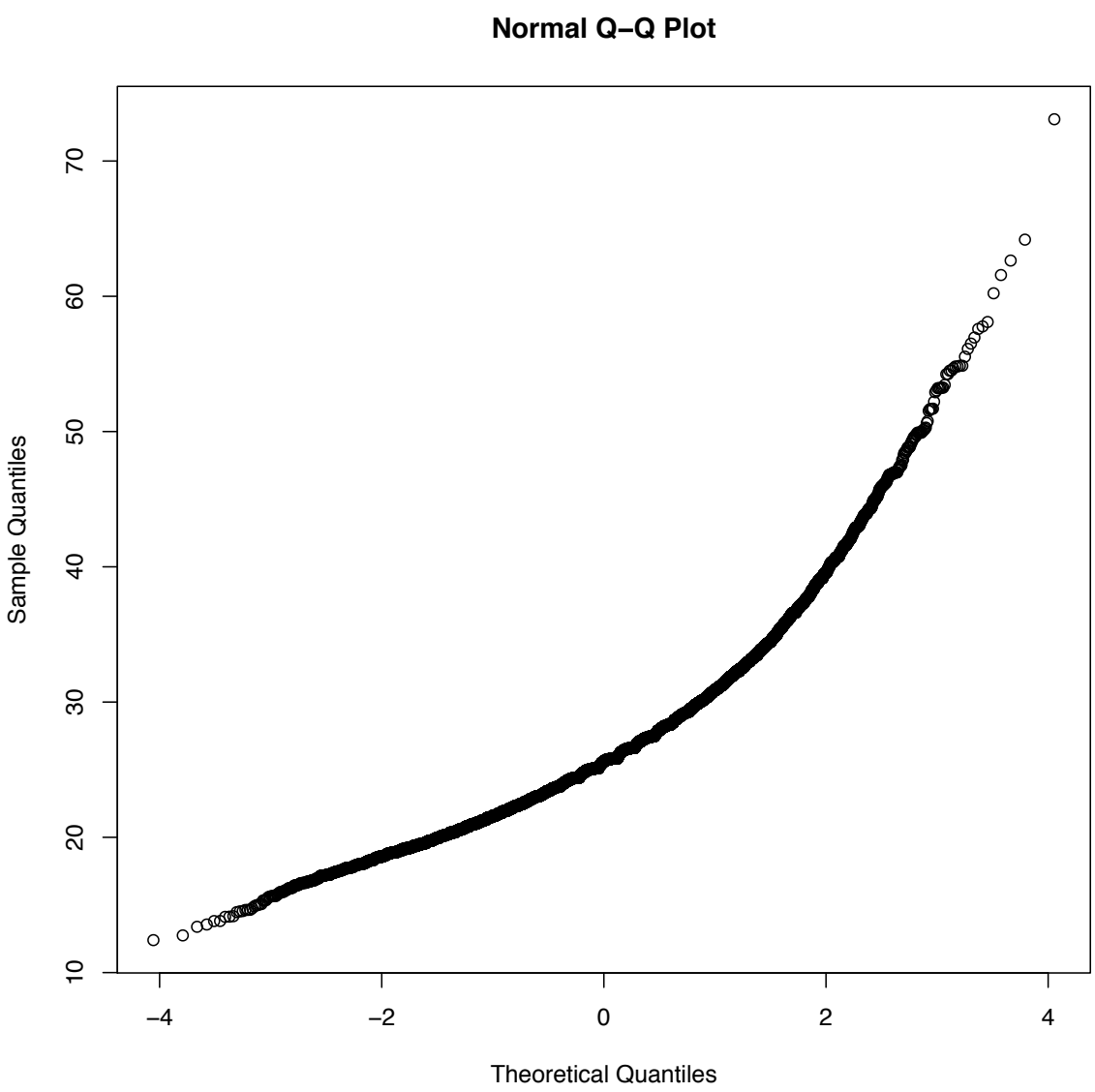


Sample quantiles

Similarly, given a data set with n points, $x_1, \ldots, x_n$, we can represent the sorted data as $x_{(1)}, \ldots, x_{(n)}$ where $x_{(1)}$ is the smallest point and $x_{(n)}$ is the largest in our sample, for example

Assuming we have no repeat values in $x_1, \ldots, x_n$, it should be clear then that j/n points of our sample are less than or equal to $x_{(j)}$-- These are the sample quantiles and you can, with various extension strategies, define the sample quantiles for any q between 0 and 1

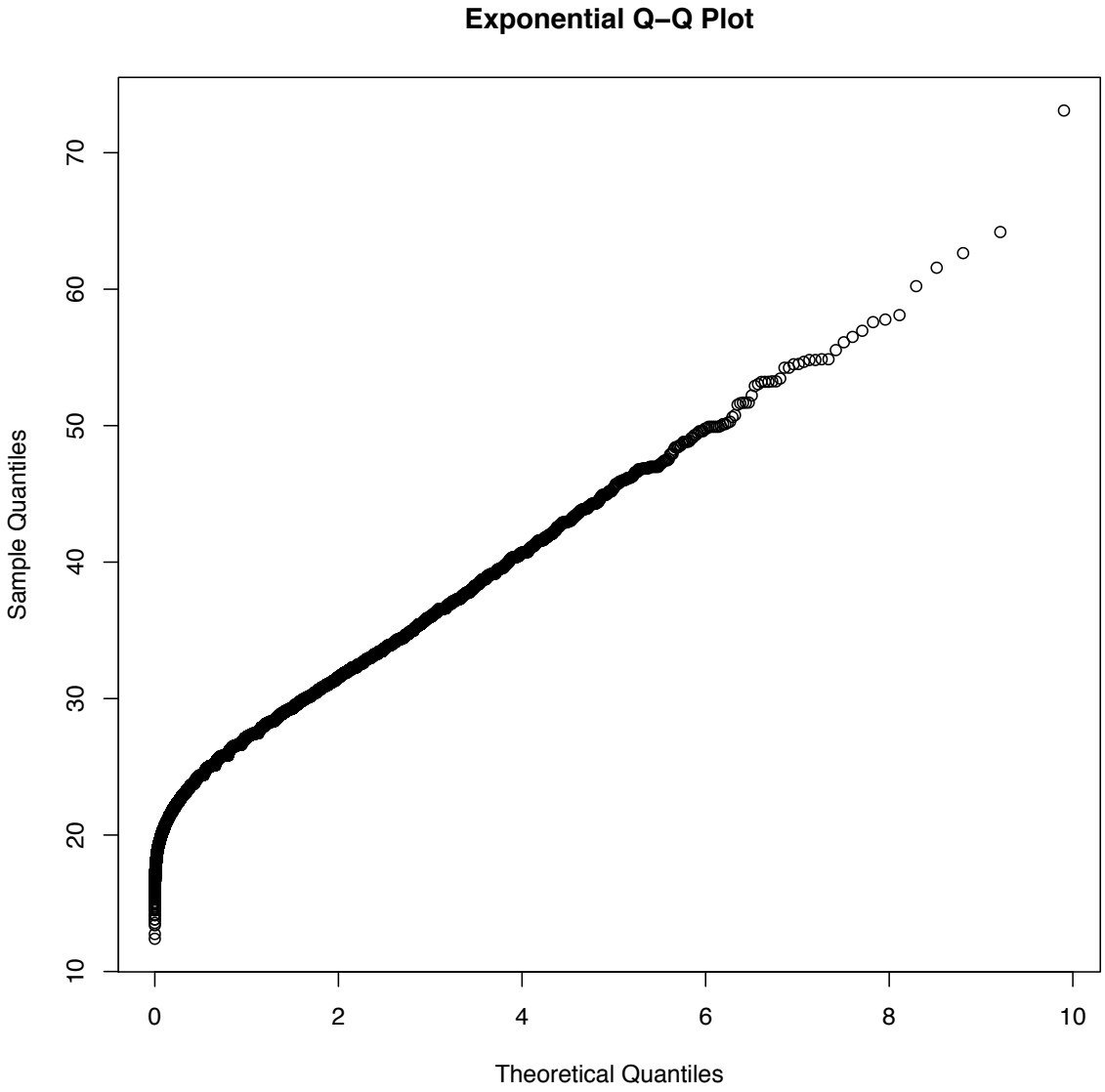
The Q-Q plot then plots the theoretical j/n quantile from the normal distribution against the sorted data $x_{(1)}, \ldots, x_{(n)}$


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
10
20
30
40
50
60
70
−4
−2
0
2
4

Shape

Of course, the normal distribution is not our only “ruler” and we often want to compare the data to some other known distribution -- On the next slide, we compare the **BMI values to the quantiles of the exponential distribution**

We can also compare two data sets in this way, plotting sorted data sets against each other

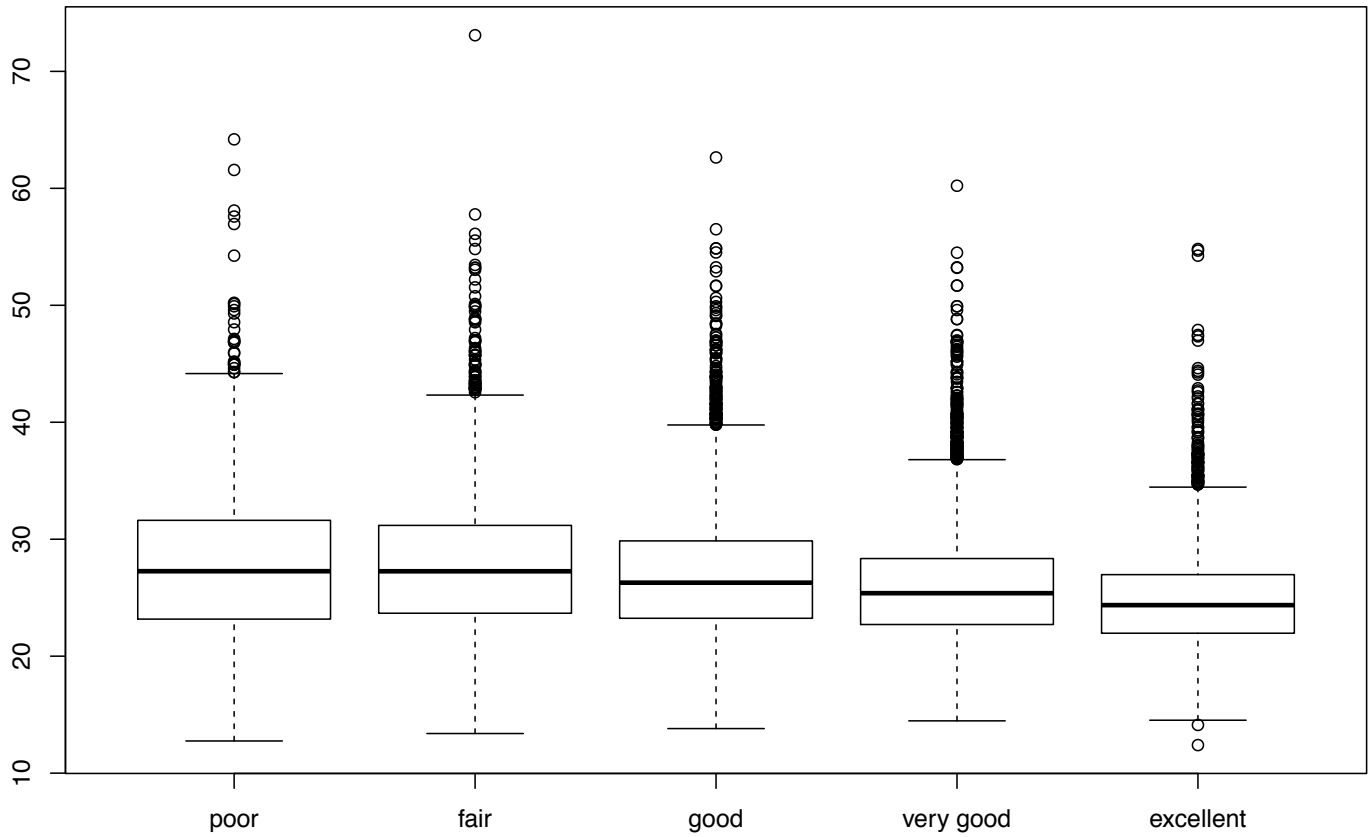

Exponential Q–Q Plot
Sample Quantiles
Theoretical Quantiles
10
20
30
40
50
60
70
0
2
4
6
8
10

Box plots

We then considered a **cartoon or thumbnail of a distribution to compare data that fall into different groups** -- The box covers the central 50% of the data (the region between the 25th and 75th percentiles (quantiles with q=0.25 and q=0.75)

We also discussed a technique to highlight outlying or “outside” points that are possibly too large (or too small) and might warrant further investigation

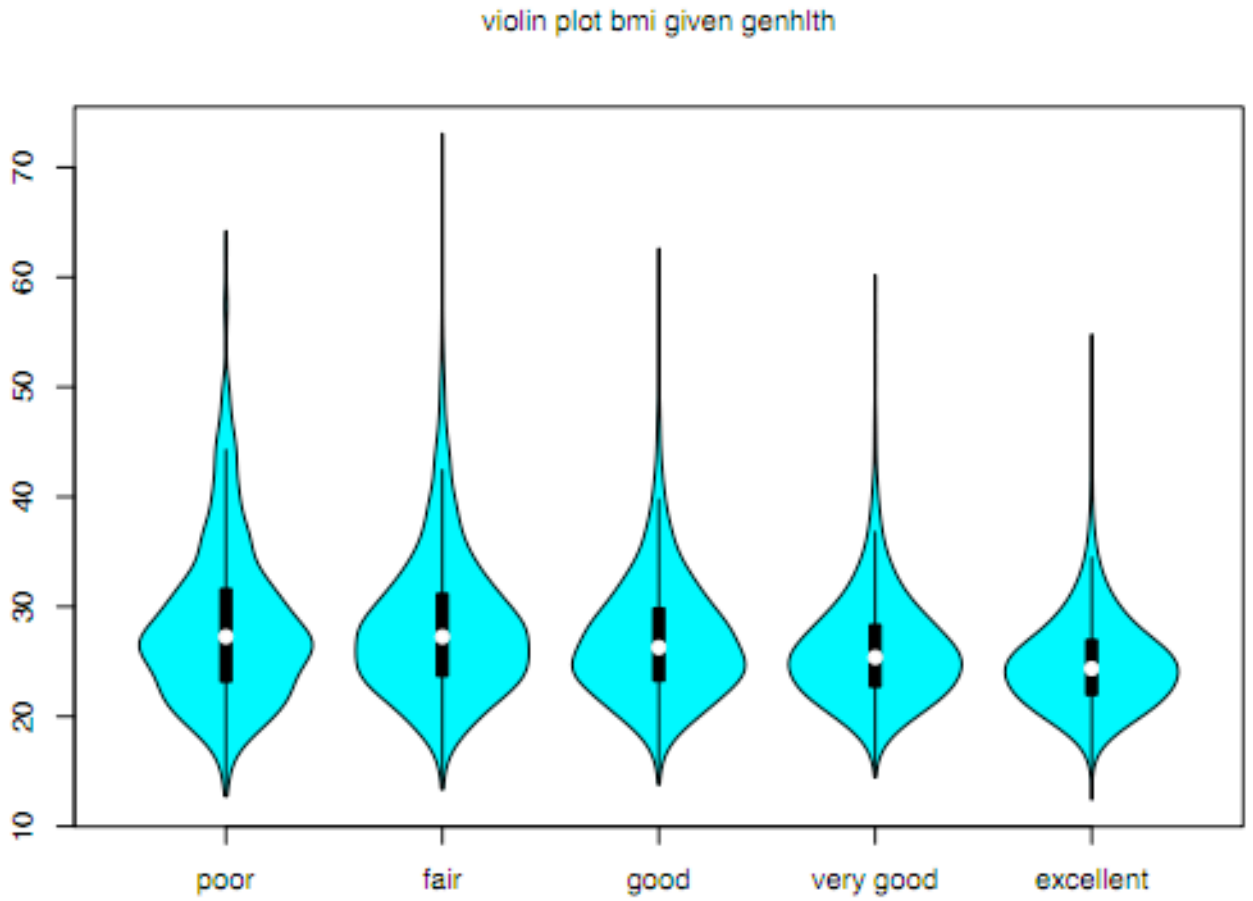
On the next slide, we present the BMI data again, this time broken down by respondent’s general health -- You might compare this display to what we saw a few slides back using mosaic plots


10
20
30
40
50
60
70
poor
fair
good
very good
excellent

Violin plots

The so-called violin plot might be more artistry than data analysis; but it uses the smoothed histogram tipped on its side and mirrored left and right in place of a box

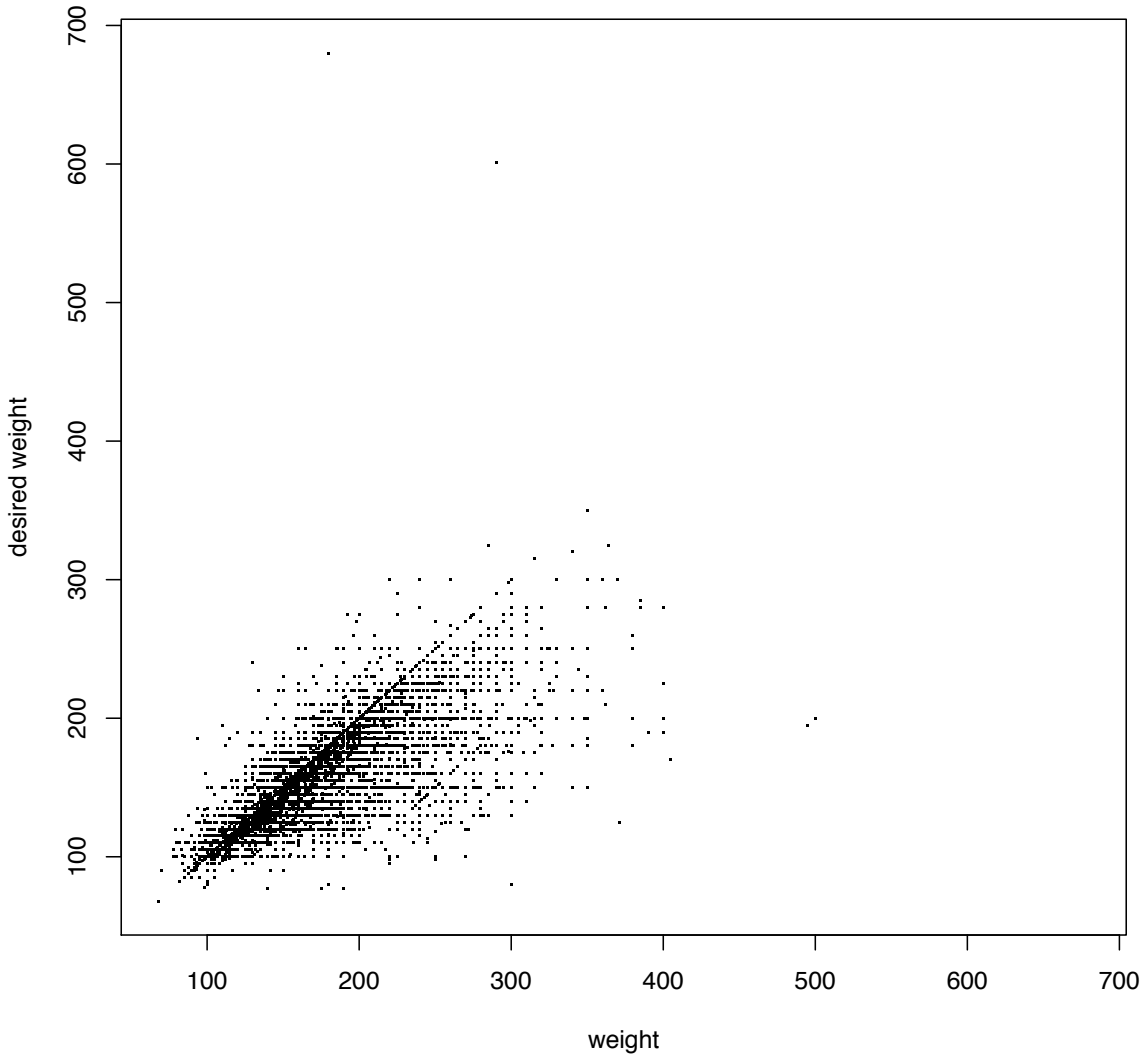
Compare this plot to the boxplot three slides back; what do you think?



Extensions

Last time, we considered an extension of the box plot to more than one variable -- That is, create a display that can create a cartoon “spatial” distribution of points

What concepts do we have to generalize to do this?


desired weight
weight
100
200
300
400
500
600
700

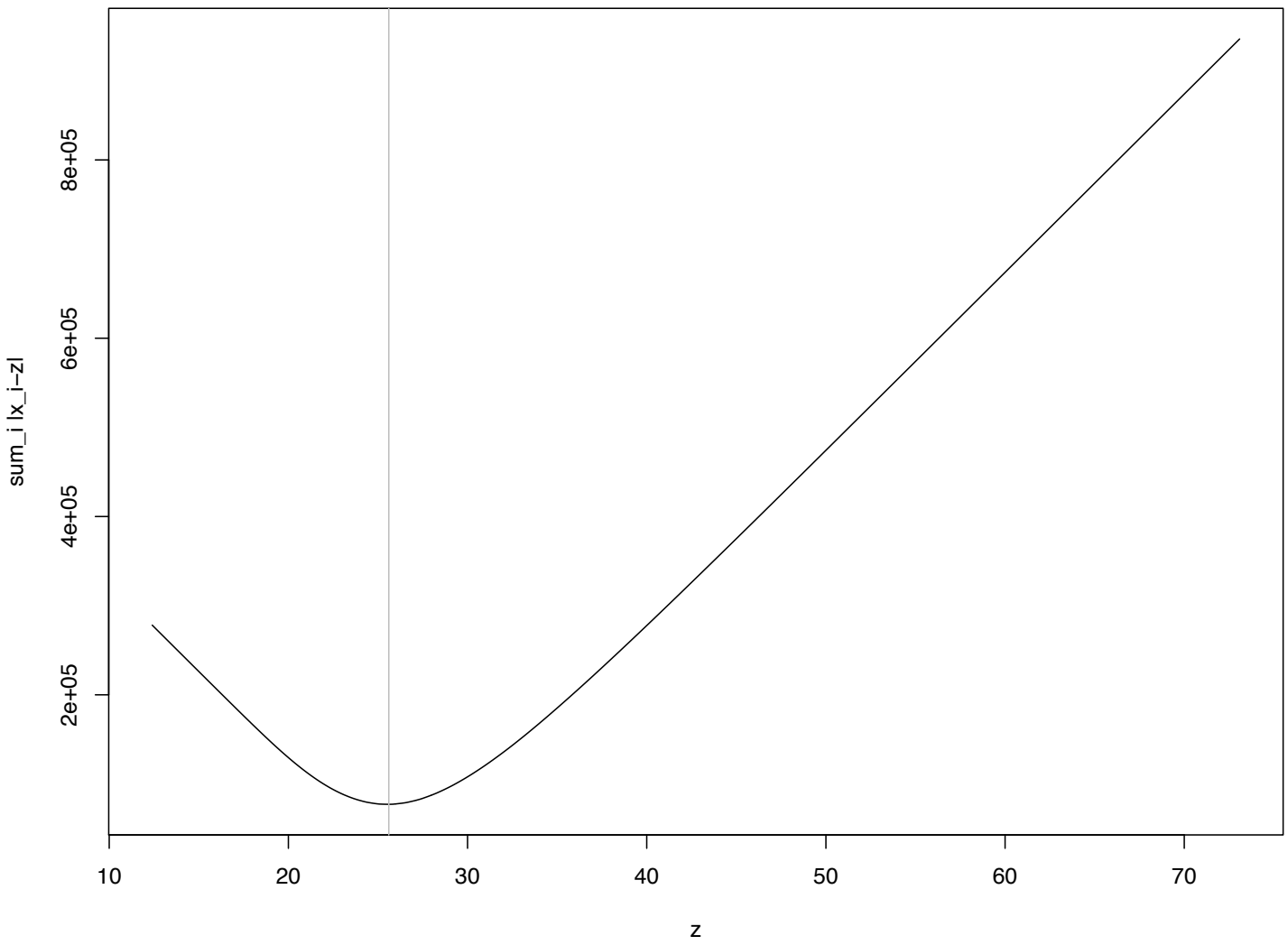
Extensions: Optimization

Last time, one of you mentioned capturing the “center” through **some notion of distance** -- Let’s make that a bit concrete

Suppose we have a set of n values $x_1, \ldots, x_n$ for a single variable and consider the expression

$$\sum_{i=1}^{n} |z - x_i|$$

The value of z that minimizes this quantity turns out to be the median!


sum_i |x_i−z|
2e+05
4e+05
6e+05
8e+05
10
20
30
40
50
60
70
z

Extensions: Optimization

Assuming $|z - x_i|$ is not zero, its derivative (as a function of z) is simply the sign of $z - x_i$ -- If it is zero, then its derivative from the left is -1 and its derivative from the right is +1

Using this fact, you can show that the the expression

$$\sum_{i=1}^{n} |z - x_i|$$

must be convex -- If we have an even number of data points (n even) with no repetitions among the $x_1, \ldots, x_n$, then any point between $x_{n/2}$ and $x_{(n/2+1)}$ has a zero derivative

Extensions: Optimization

We have already seen the interquartile range as a notion of spread in the data -- The width of the interval (the height of a box in a box plot) that covers the central 50% of the data

There are **competing notions for the spread** that are based on absolute deviations -- For example, the MAD or median absolute deviation is defined to be

$$\text{median } \{\, |x_1 - x_{\text{med}}|, |x_2 - x_{\text{med}}|, \ldots, |x_n - x_{\text{med}}| \,\}$$

We say that the median and either the IQR or the MAD are “robust” to outlying data -- We’ll make that precise in a moment...

Aside: Mean and variance

The arithmetic mean is another notion of the center of a distribution -- We recall that the mean of n values $x_1, \ldots, x_n$ of a variable is simply

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

The associated measure of spread is the sample variance -- It is the sum of squared deviations from the mean

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

Aside: Mean and variance

If we consider the sum of squared deviations as a function of any point z, then we can show (this time taking derivatives is easy!) that the minimizer of

$$\sum_{i=1}^{n} (x_i - z)^2$$

is the sample mean, $z = \bar{x}$! Go ahead, try it!

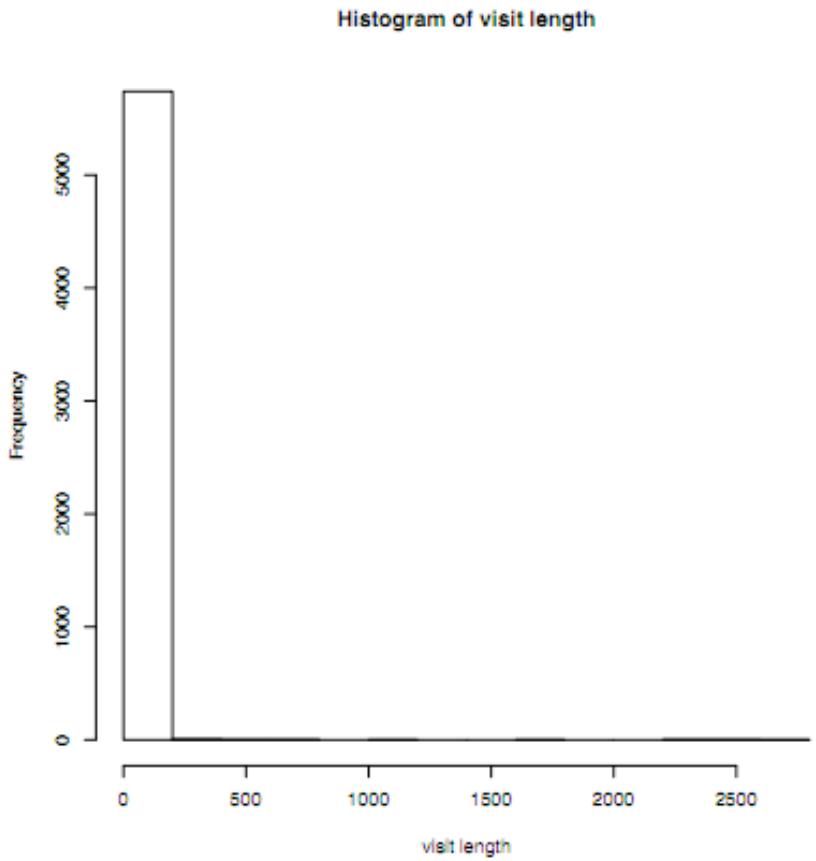
Aside: Mean and variance

While the median and the interquartile range are very direct 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”**

Why would we expect this? Does the median have the same problem?

Aside: Web visits example

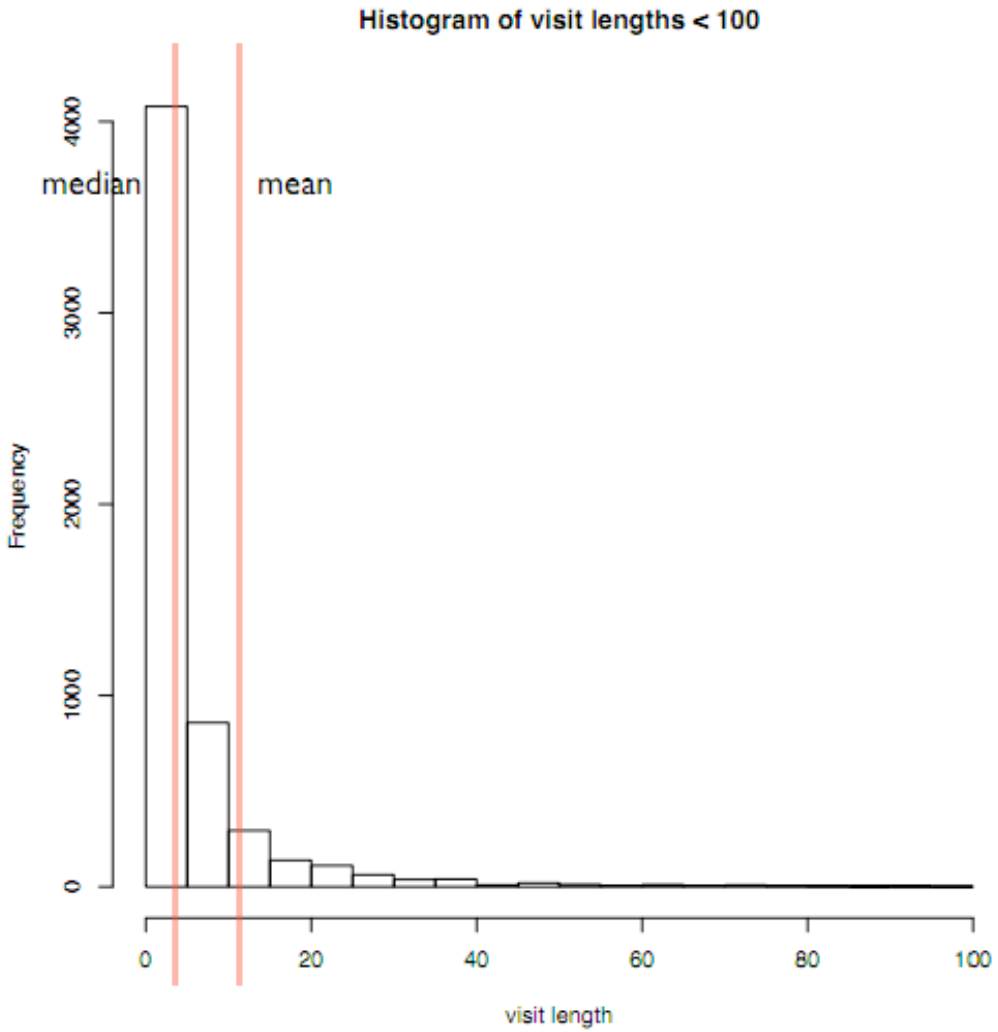
Here again is our poor histogram of web visits; remember that these data are heavily skewed to the right



Aside: 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?

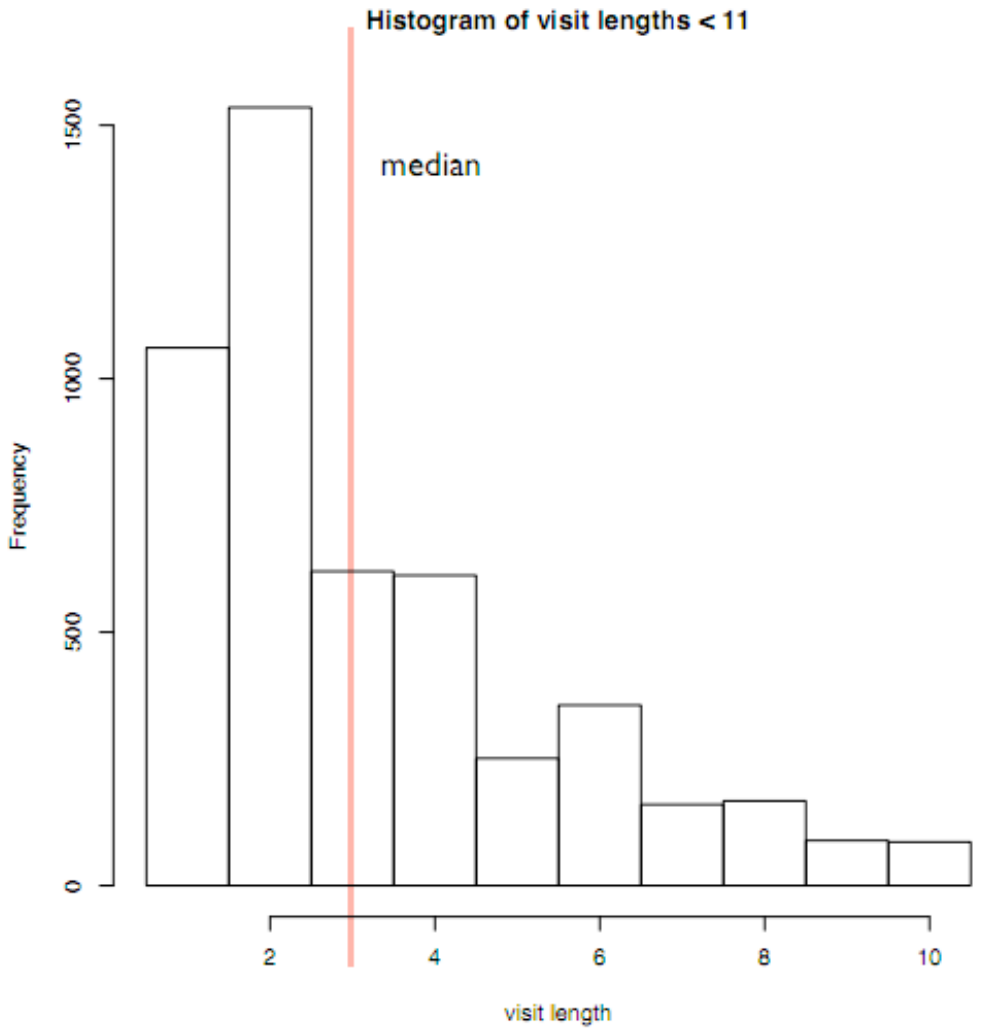


Aside: Web visits example

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?



Aside: Web visits 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

Aside: The normal distribution

As we will see next time, the sample **mean and variance are tied up with estimating the parameters of a normal distribution** (the population mean and variance) -- In a modeling context, you certainly wouldn't propose a normal distribution for the Web visits data!

John Tukey was one of the first statisticians to call attention to the fact that departures from a normal distribution could hurt the mean and variance -- His test case was a "mixture" of two normals

Aside: Robustness

Imagine tossing a coin such that 99.2% of the time, you sampled an observation from the normal distribution with mean μ and variance σ^2; and 0.8% of the time you selected an observation from a normal with mean μ and variance $9\sigma^2$

This was Tukey's idea of **"contamination"** -- That there could be some error process that was introducing wild observations at a very low rate, but even with this low rate bad things happened

His work from the early 60s (and before, actually) led to a subfield of statistics concerned with the robustness of estimators to departures from assumptions (like the data come from a single normal population)

Aside: Quantiles (again)

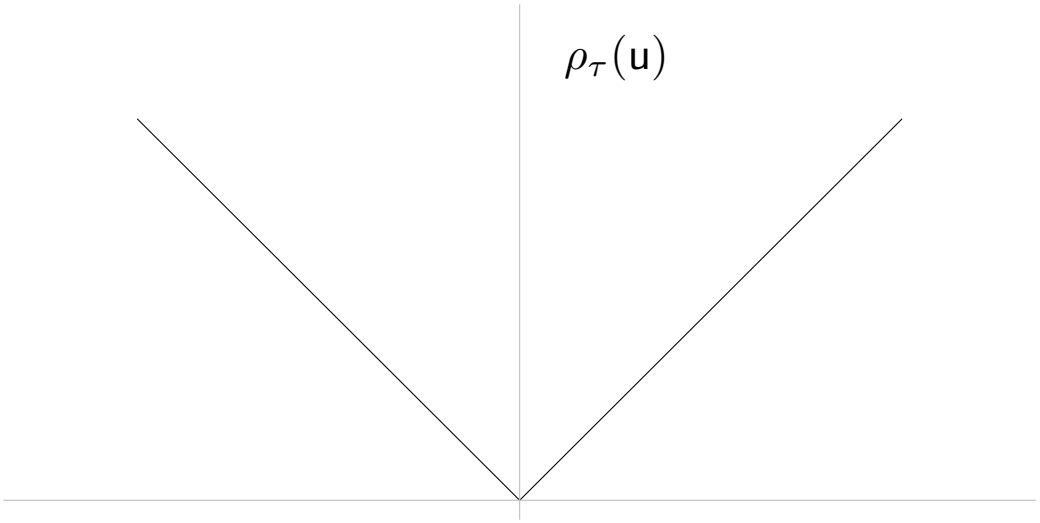
It turns out that we can define quantiles as a minimization problem as well -- Let's define a new function

$$\rho_\tau(u) = |u| + \tau u$$

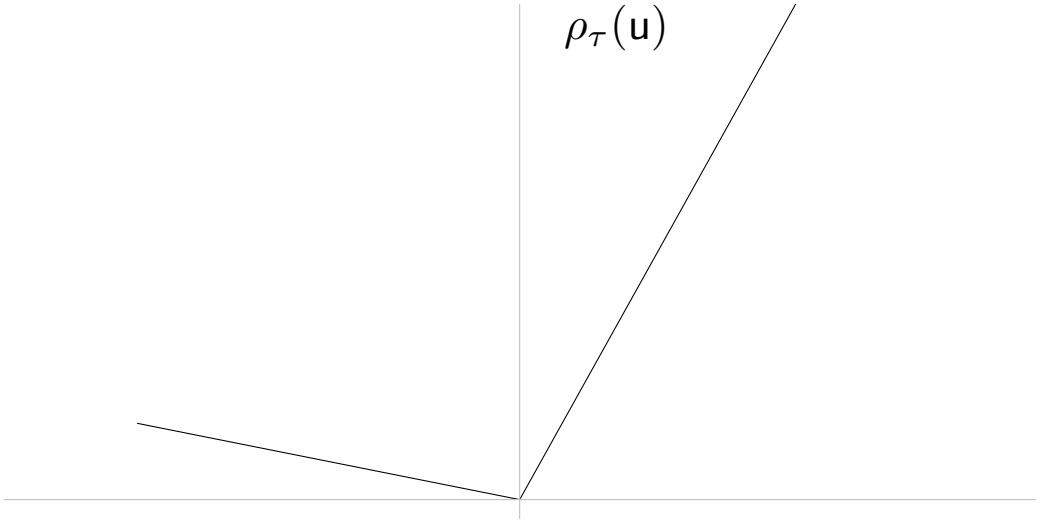
For a given level 0 < q < 1, we can define the set $\tau = 2q - 1$ and then define the qth quantile to be the value of z that minimizes

$$\sum_{i=1}^{n} \rho_\tau(z - x_i)$$

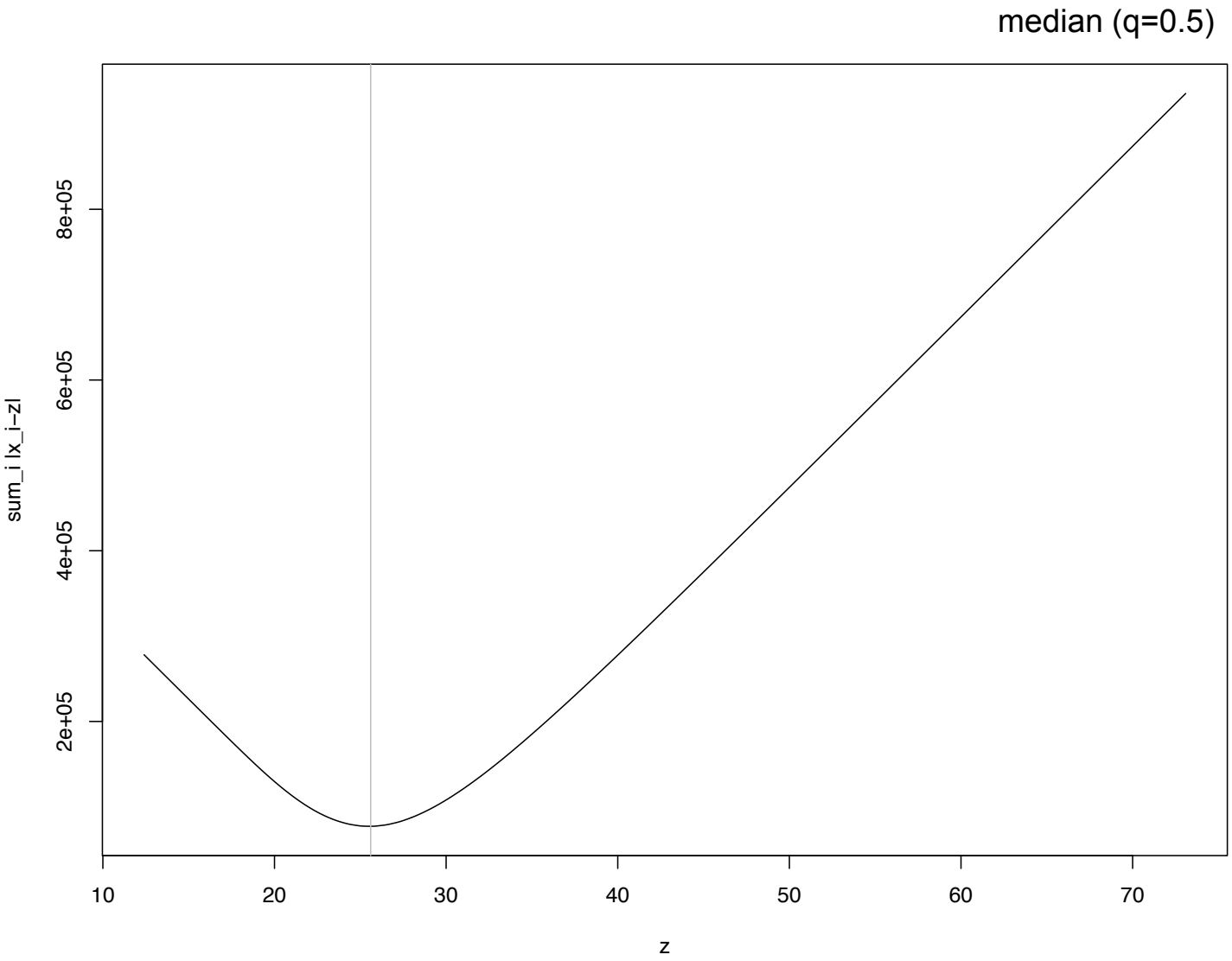
At the right we have an example of the function $\rho_\tau(u)$ for q=0.5 (the median) and q=0.9 the 90th percentile

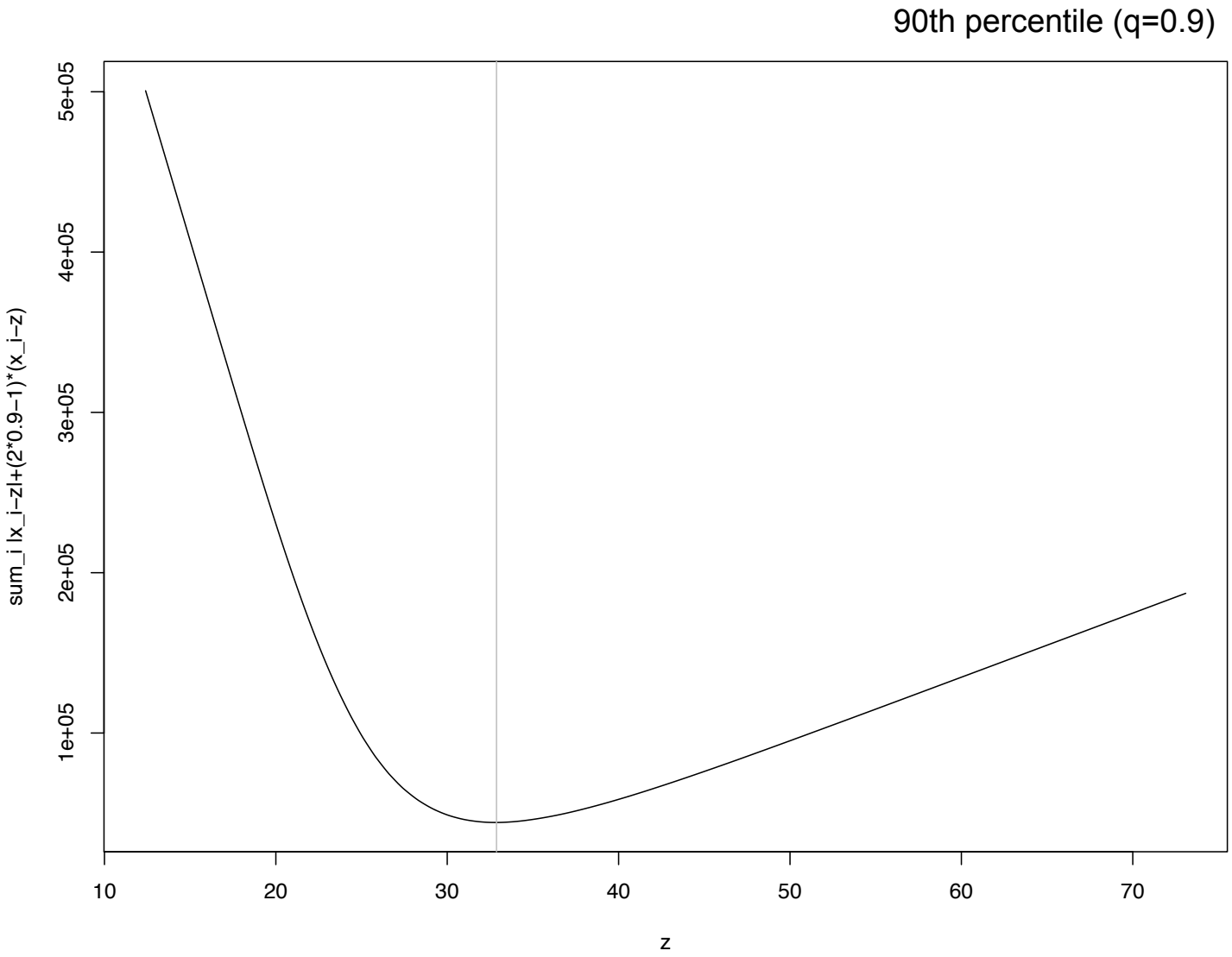


median (q=0.5)



90th percentile (q=0.9)


median (q=0.5)
sum_i |x_i−z|
2e+05
4e+05
6e+05
8e+05
10
20
30
40
50
60
70
z


90th percentile (q=0.9)
sum_i |x_i−z|+(2*0.9−1)*(x_i−z)
1e+05
2e+05
3e+05
4e+05
5e+05
10
20
30
40
50
60
70
z

Back to the center of a points in space...

Extensions: Optimization

One way to think about the center of a 2-variable data cloud would be to extend this optimization approach -- This was done, for example, in the late 1800s and early 1900s by the U.S. Census

One definition of the “center” of the U.S. population involved a hypothetical assembly of all the people in the country -- Simply, the median of this spatial distribution of people is the point that minimizes the total distance the population would have to travel to assemble there

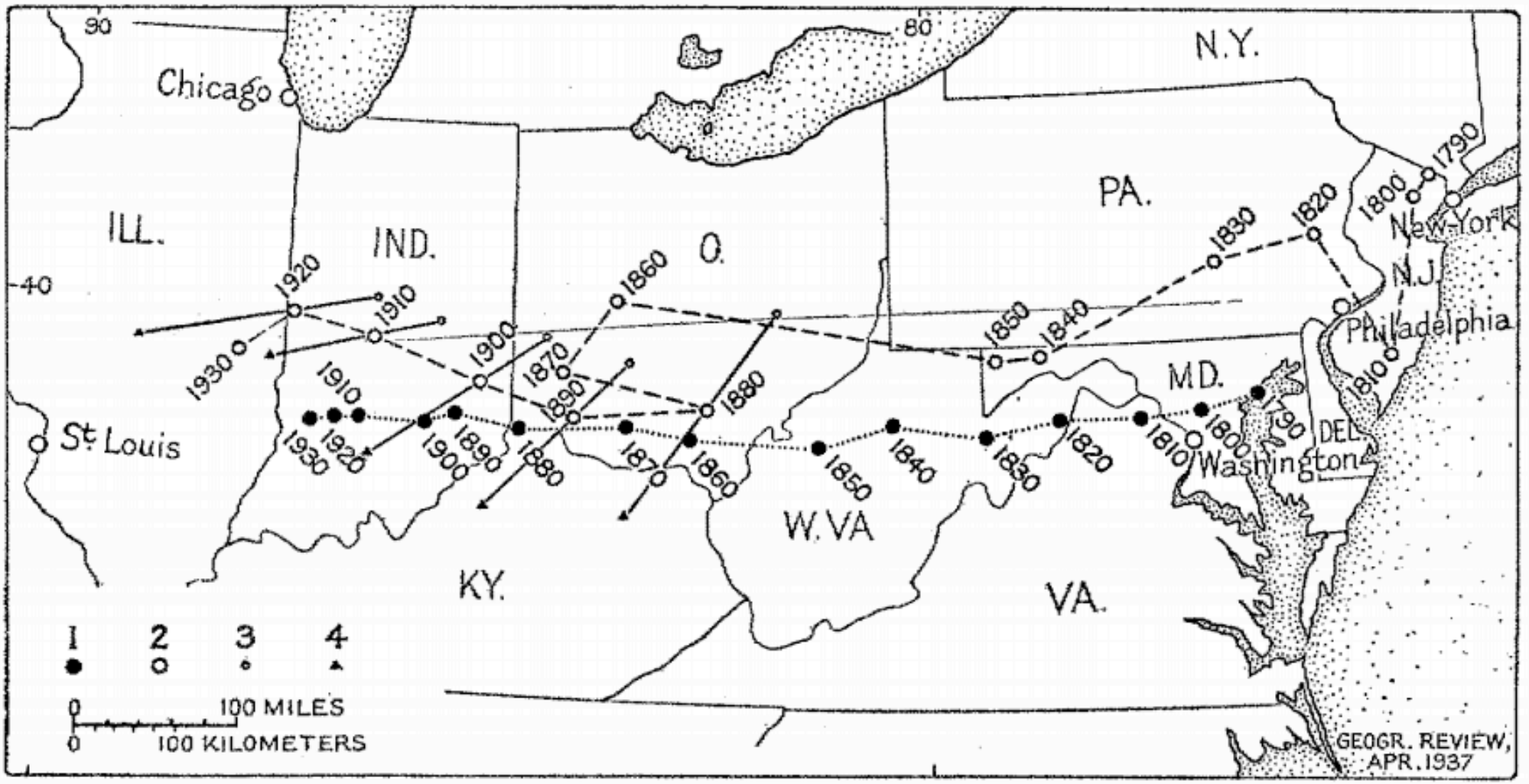
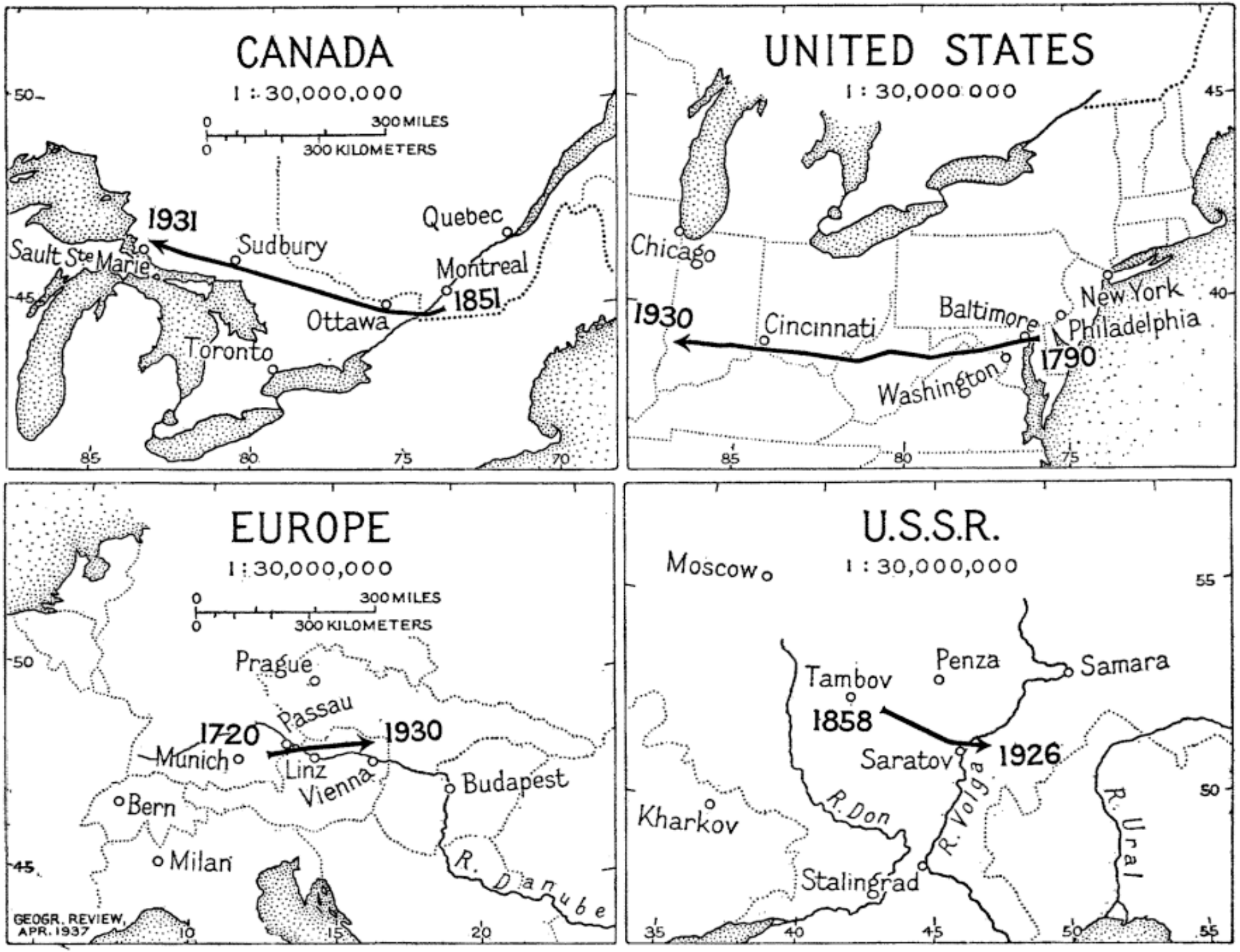


FIG. 5—Centrogram showing movements of the centers of population and higher education in the United States of America from 1790 to 1930. Key: centers of 1, general population; 2, higher educational population (universities and colleges); 3, higher educational population, men; 4, higher educational population, women. (From a study made by Walter C. Eells for publication in the forthcoming Mendeleev Memorial Volume of the Centrographical Laboratory in Leningrad.)


CANADA
1 : 30,000,000
0 300 MILES
0 300 KILOMETERS
1931
Sault Ste Marie
Sudbury
Quebec
Montreal
1851
Ottawa
Toronto
50
45
85
80
75
70
UNITED STATES
1 : 30,000 000
Chicago
1930
Cincinnati
Baltimore
New York
Philadelphia
1790
Washington
45
40
85
80
75
EUROPE
1 : 30,000,000
0 300 MILES
0 300 KILOMETERS
Prague
Passau
1720
1930
Munich
Linz
Vienna
Budapest
Bern
Milan
R. Danube
50
45
10
15
20
GEOGR. REVIEW, APR. 1937
U.S.S.R.
Moscow
1 : 30,000,000
55
Tambov
Penza
Samara
1858
Saratov
1926
Kharkov
R. Don
R. Volga
R. Ural
Stalingrad
50
35
40
45
50
55

Extensions: Optimization

In symbols, if the distance between a point on the map z and the spatial coordinate where a person lived (the latitude and longitude, say) is the Euclidean distance $\|x_i - z\|$, then the median is defined as the value of z minimizing

$$\sum_{i=1}^{n} \|x_i - z\|$$

Notice that when we have univariate data again, the expression above reduces to the sum of absolute deviations and we get back our univariate median!

Extensions: Optimization

If are data are in the plane (consisting of two measurements per point), $x_i = (x_{i1}, x_{i2})$, then the distance is just

$$\|x_i - z\| = \sqrt{(x_{i1} - z_{i1})^2 + (x_{i2} - z_{i2})^2}$$

Extensions: Optimization

If we have a single measurement per point so that x_i and z are scalars, then

$$\|x_i - z\| = \sqrt{(x_i - z_i)^2} = |x_i - z_i|$$

so that the quantity we are minimizing (with respect to z) is again

$$\sum_{i=1}^{n} |x_i - z|$$

Extensions: Optimization

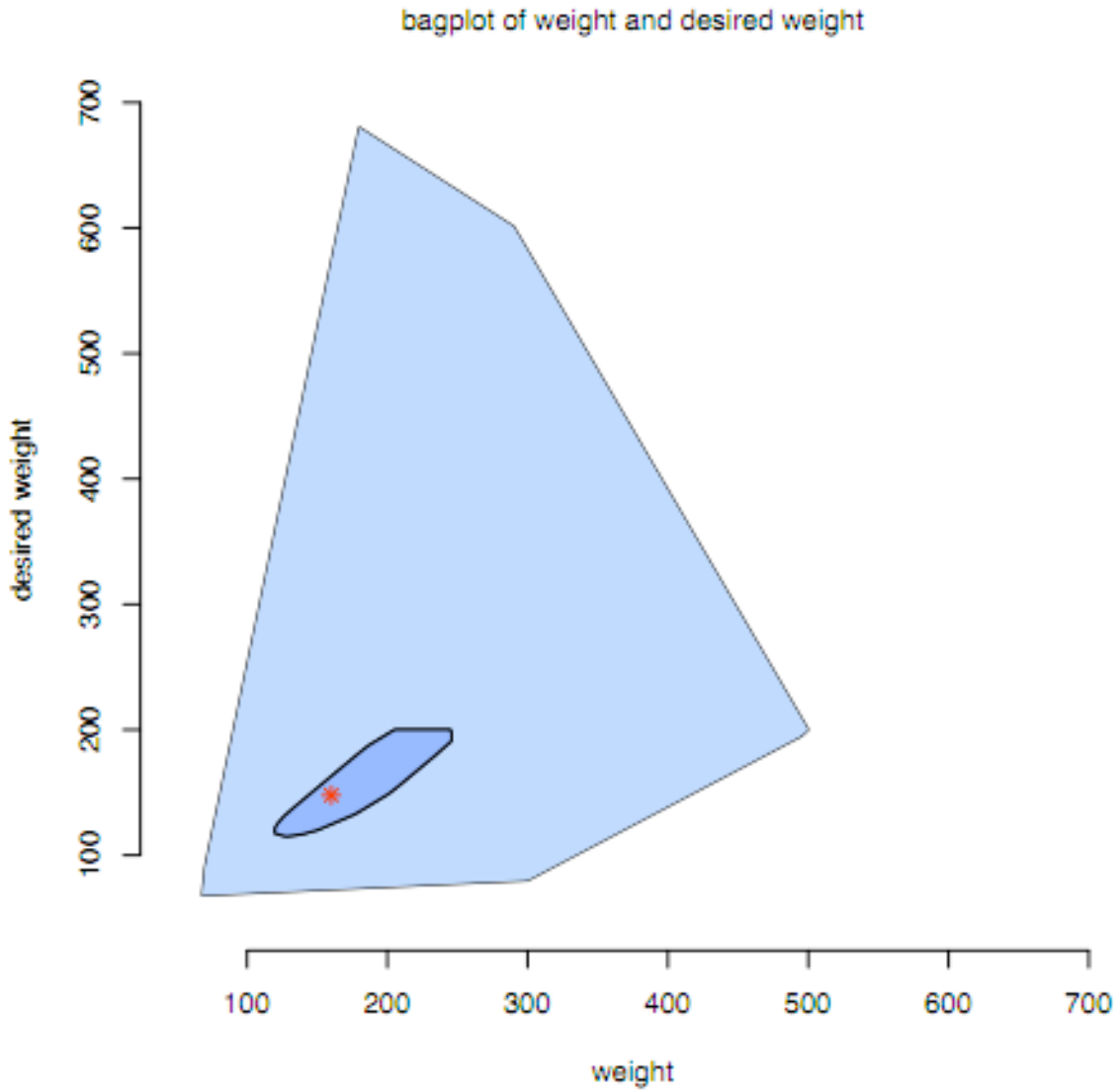
While we won't prove it, defined in this way, **we can rotate our points in the plane and still come up with the same center or median** -- This would not be true if we took our center to be the vector consisting of the median of $x_{11}, \ldots, x_{n1}$ (the values of the first coordinate), and the median of $x_{12}, \ldots, x_{n2}$, the values of the second

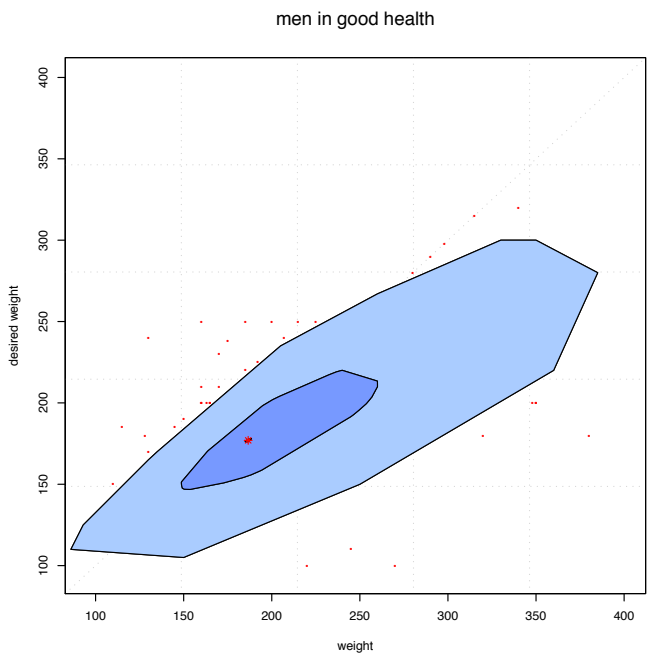
Why might this kind of invariance be important?

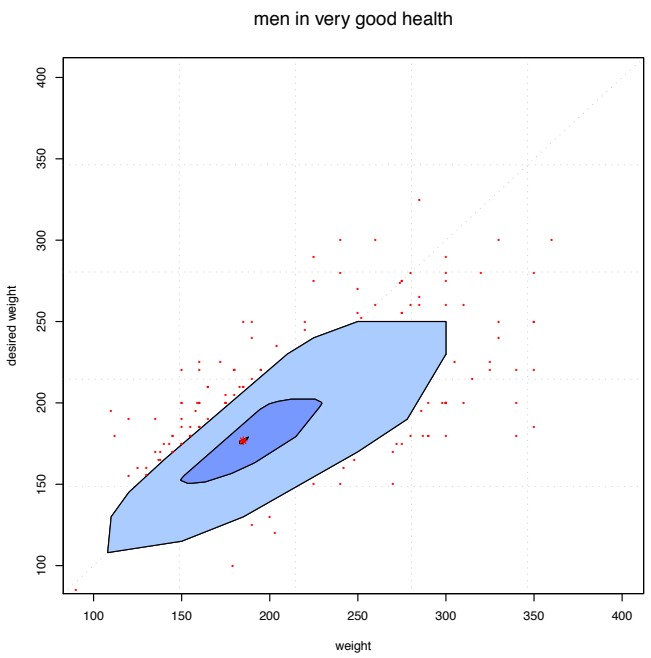
Other metaphors

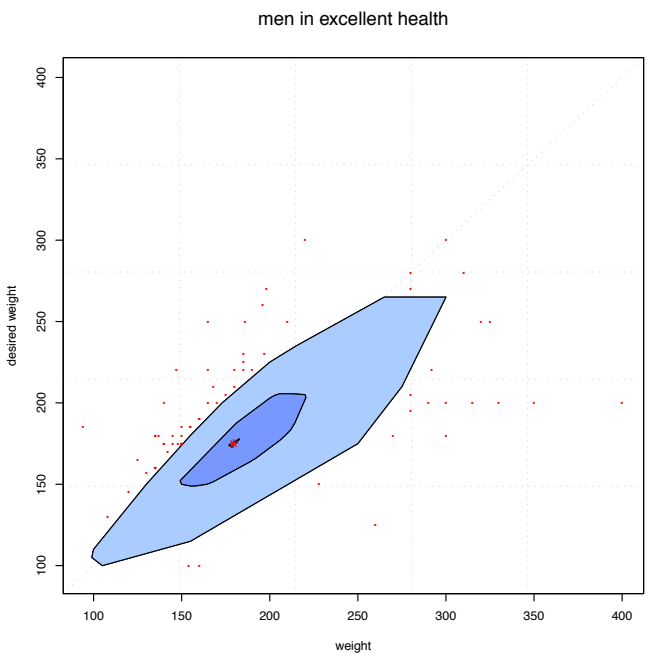
The distance approach is just one way to generalize the median or center of the distribution, and, as you might expect, **statisticians have had a lot of time to think this concept over**

We'll talk about just one other approach because it gives some insights into the structure of data in two (and higher) dimensions -- On the next slide we present the associated **"bagplot", an extension of the boxplot**


bagplot of weight and desired weight
desired weight
100
200
300
400
500
600
700
weight
100
200
300
400
500
600
700


men in good health
desired weight
weight
100
150
200
250
300
350
400

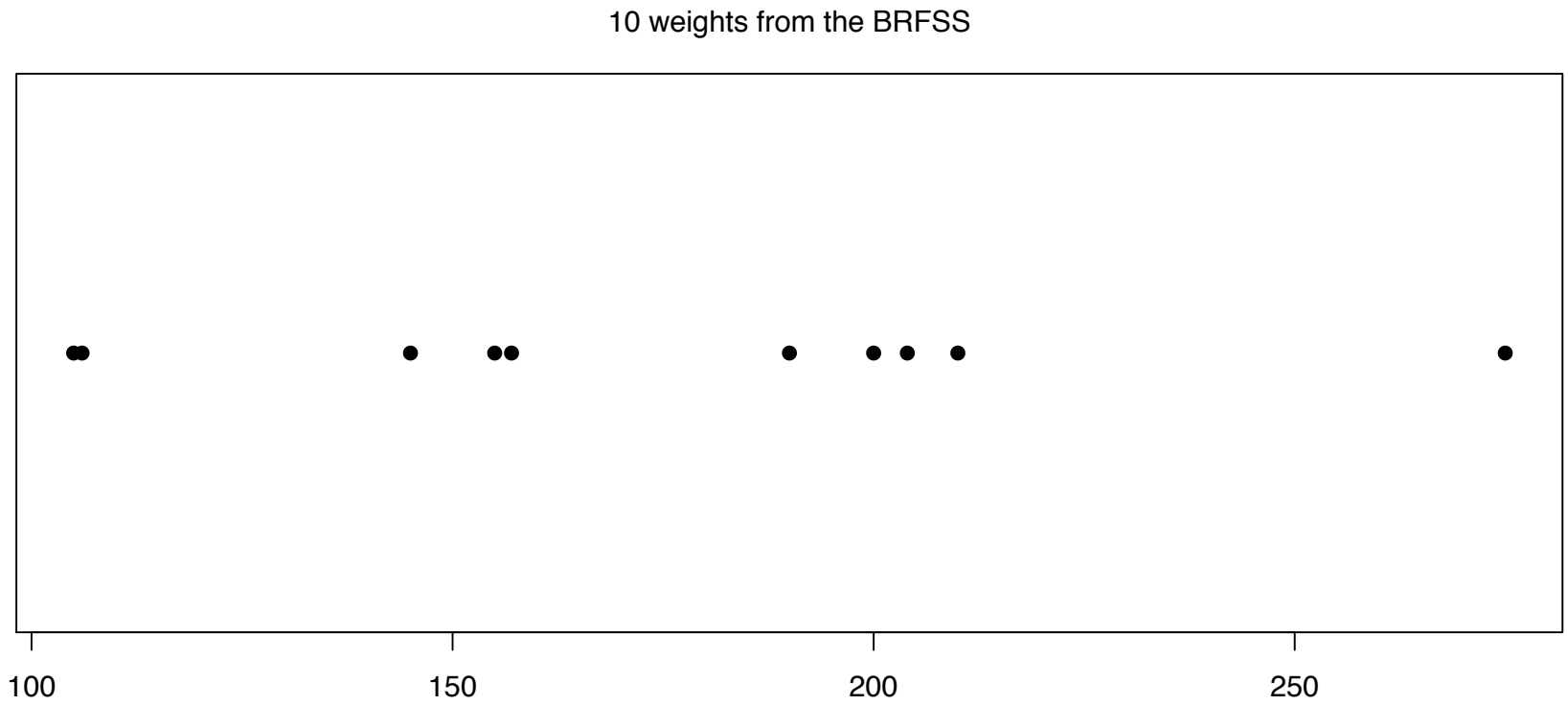

men in very good health
desired weight
weight
100
150
200
250
300
350
400

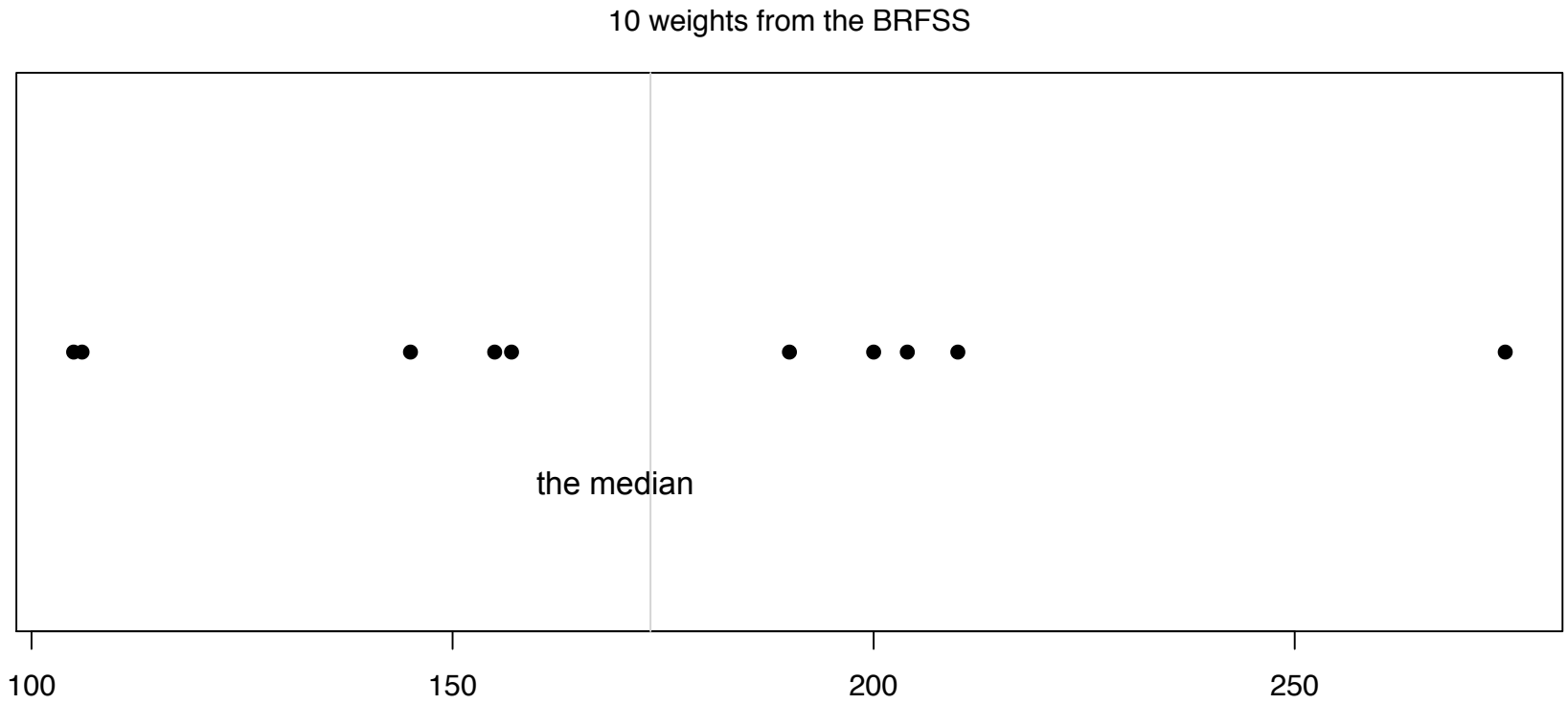

men in excellent health
desired weight
weight
100
150
200
250
300
350
400

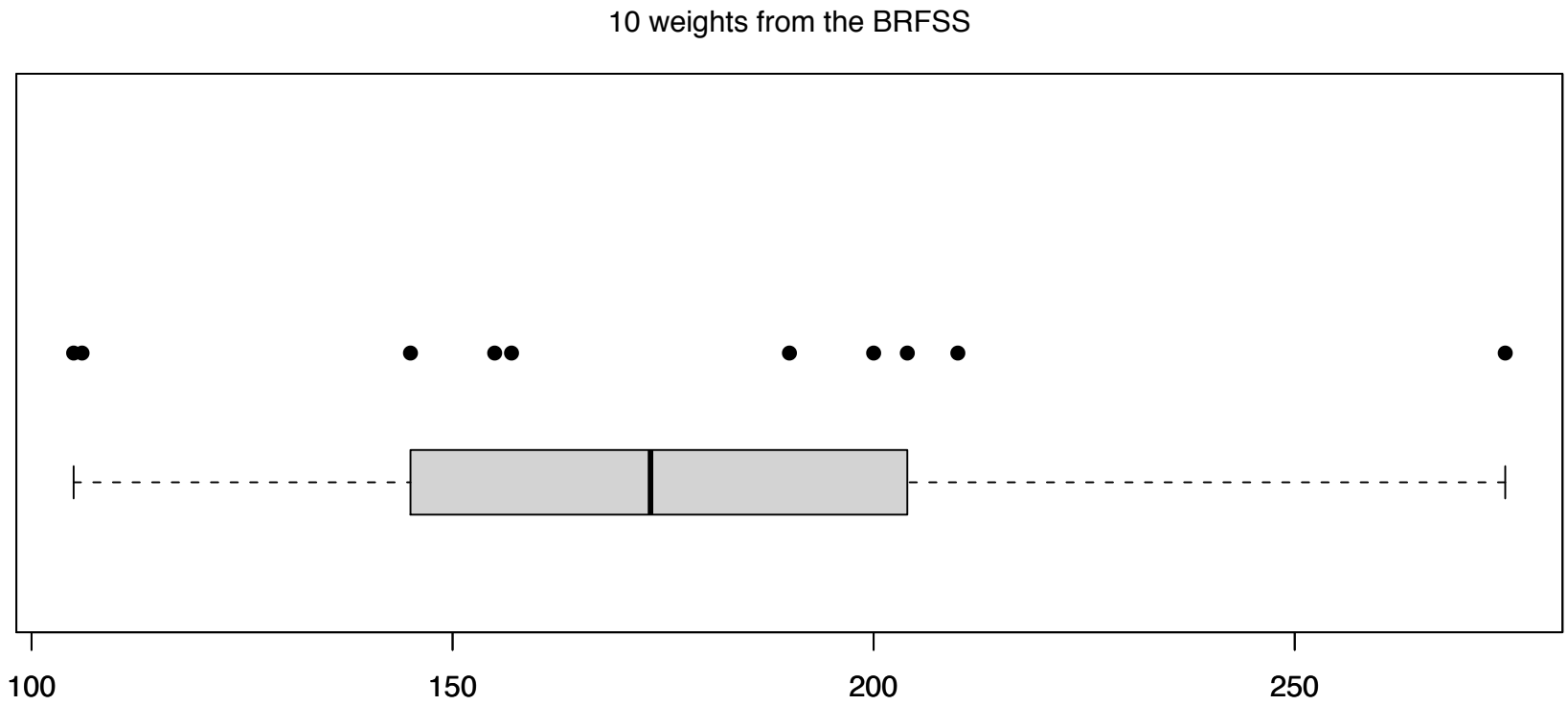
Depth

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
100
150
200
250

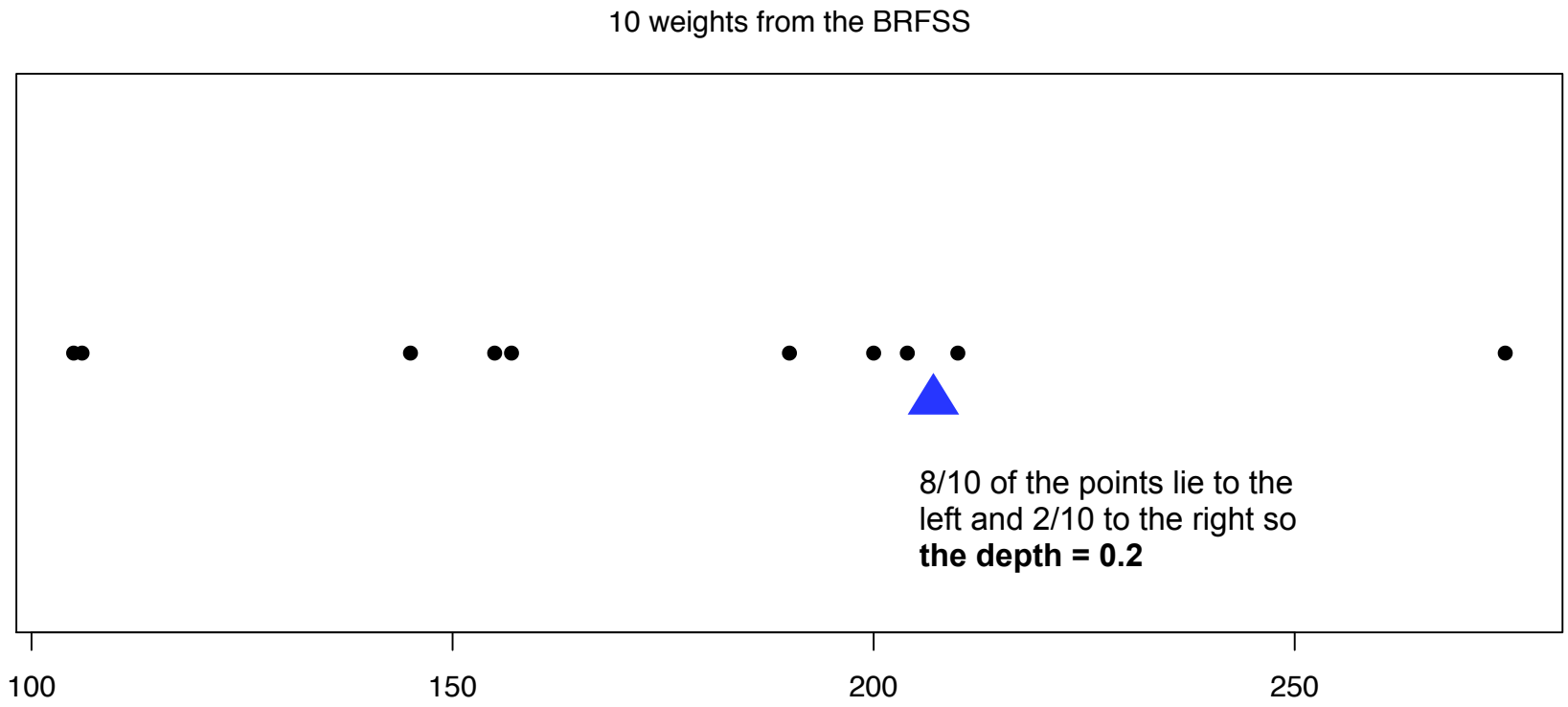

10 weights from the BRFSS
the median
100
150
200
250

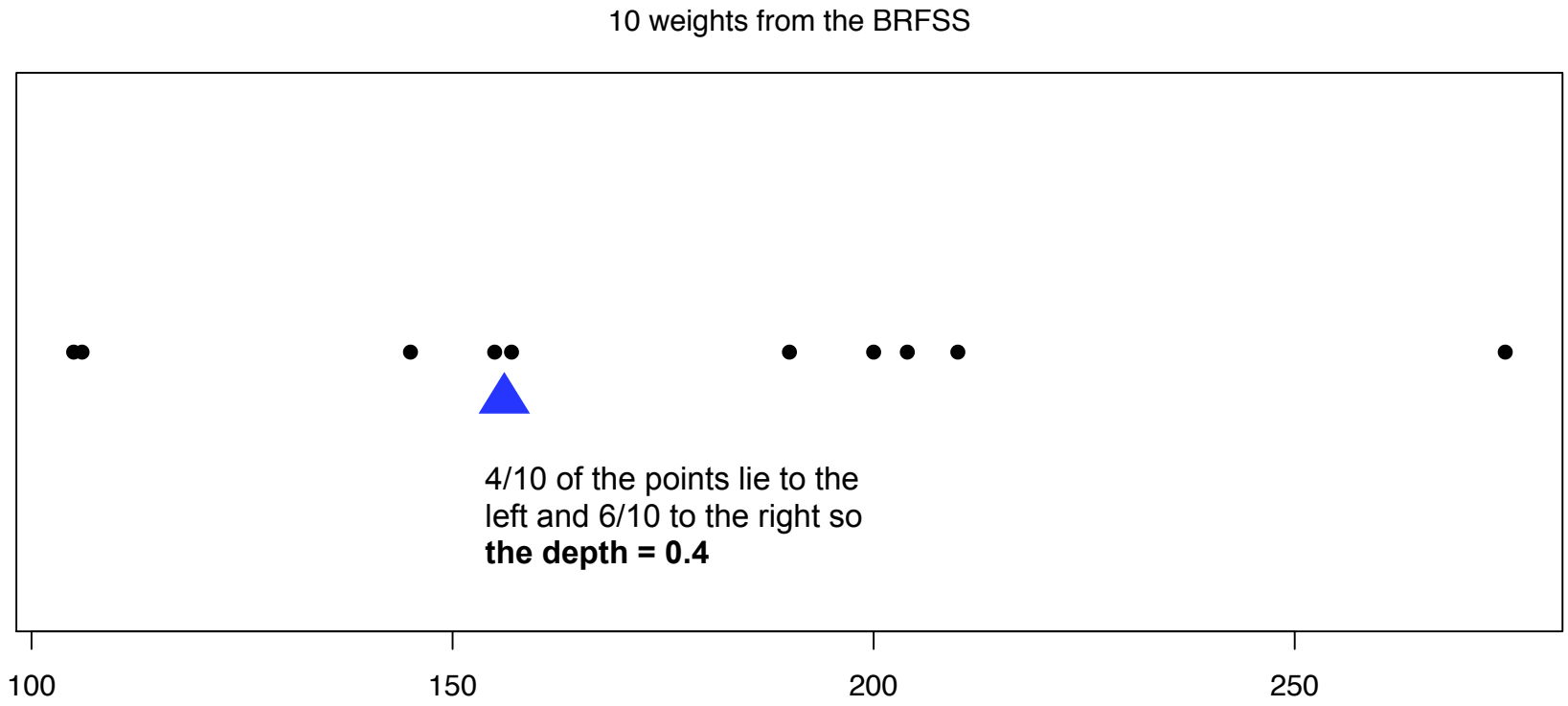

10 weights from the BRFSS
100
150
200
250

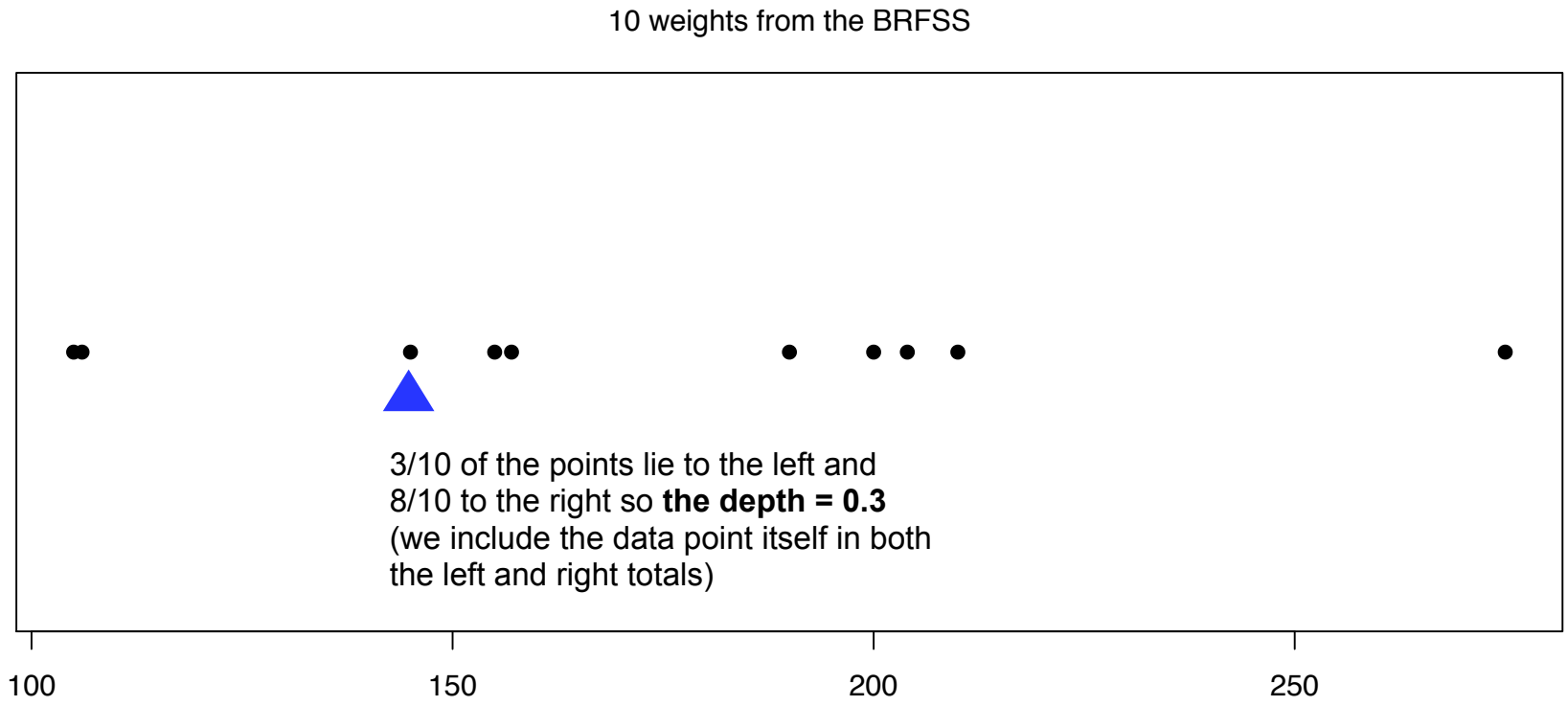
Depth

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
8/10 of the points lie to the
left and 2/10 to the right so
the depth = 0.2
100
150
200
250

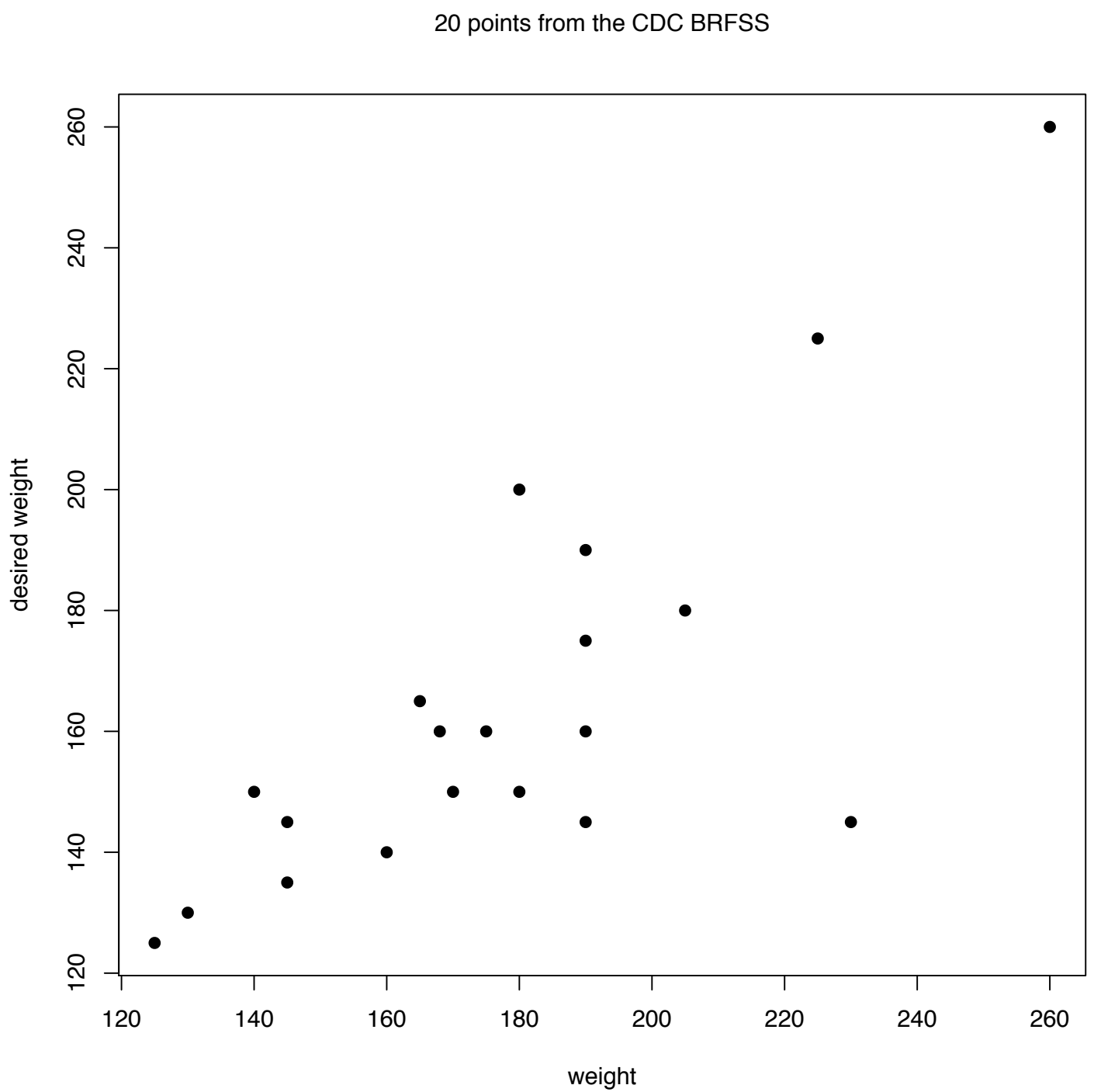

10 weights from the BRFSS
4/10 of the points lie to the
left and 6/10 to the right so
the depth = 0.4
100
150
200
250

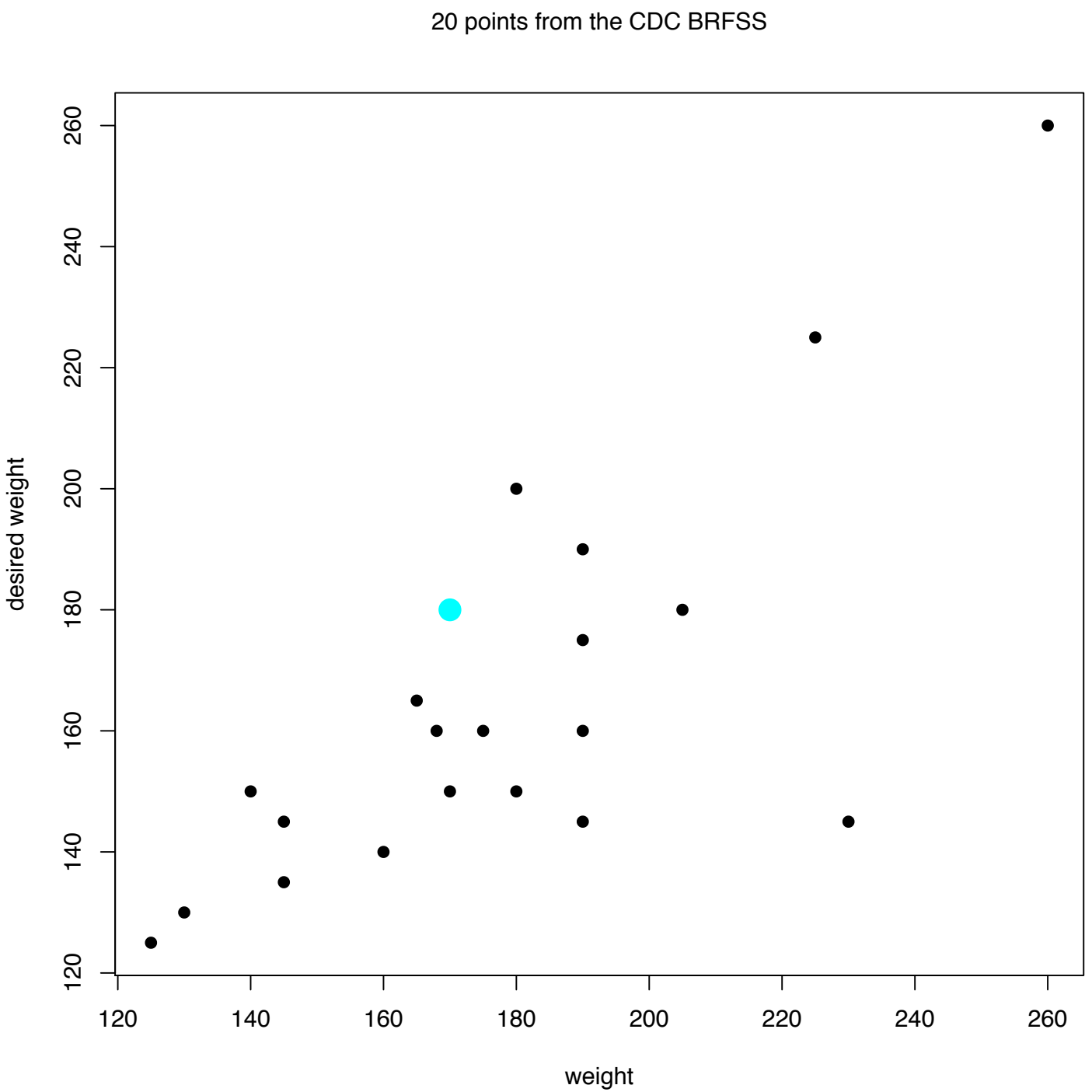

10 weights from the BRFSS
3/10 of the points lie to the left and
8/10 to the right so **the depth = 0.3**
(we include the data point itself in both
the left and right totals)
100
150
200
250

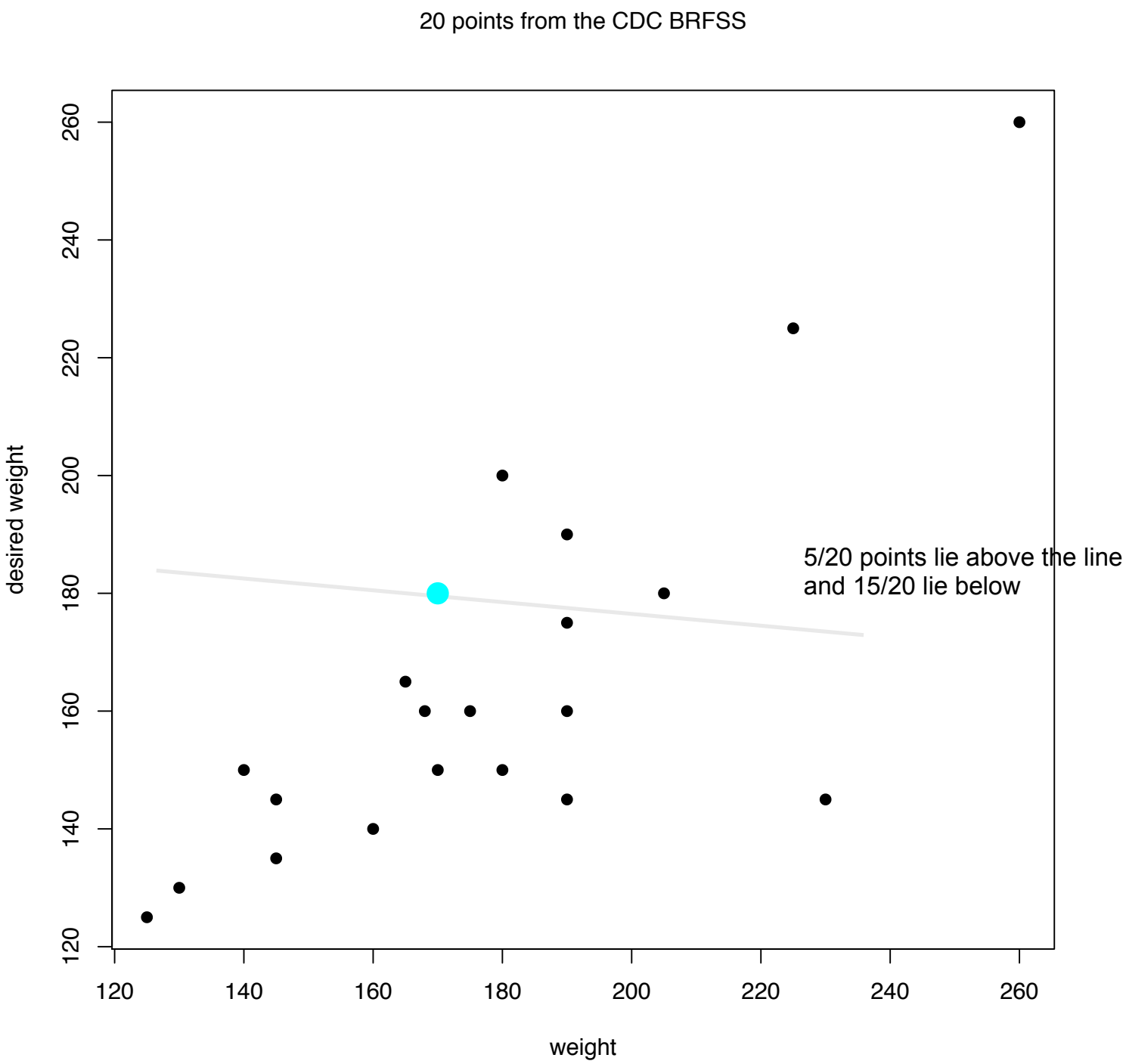
Depth

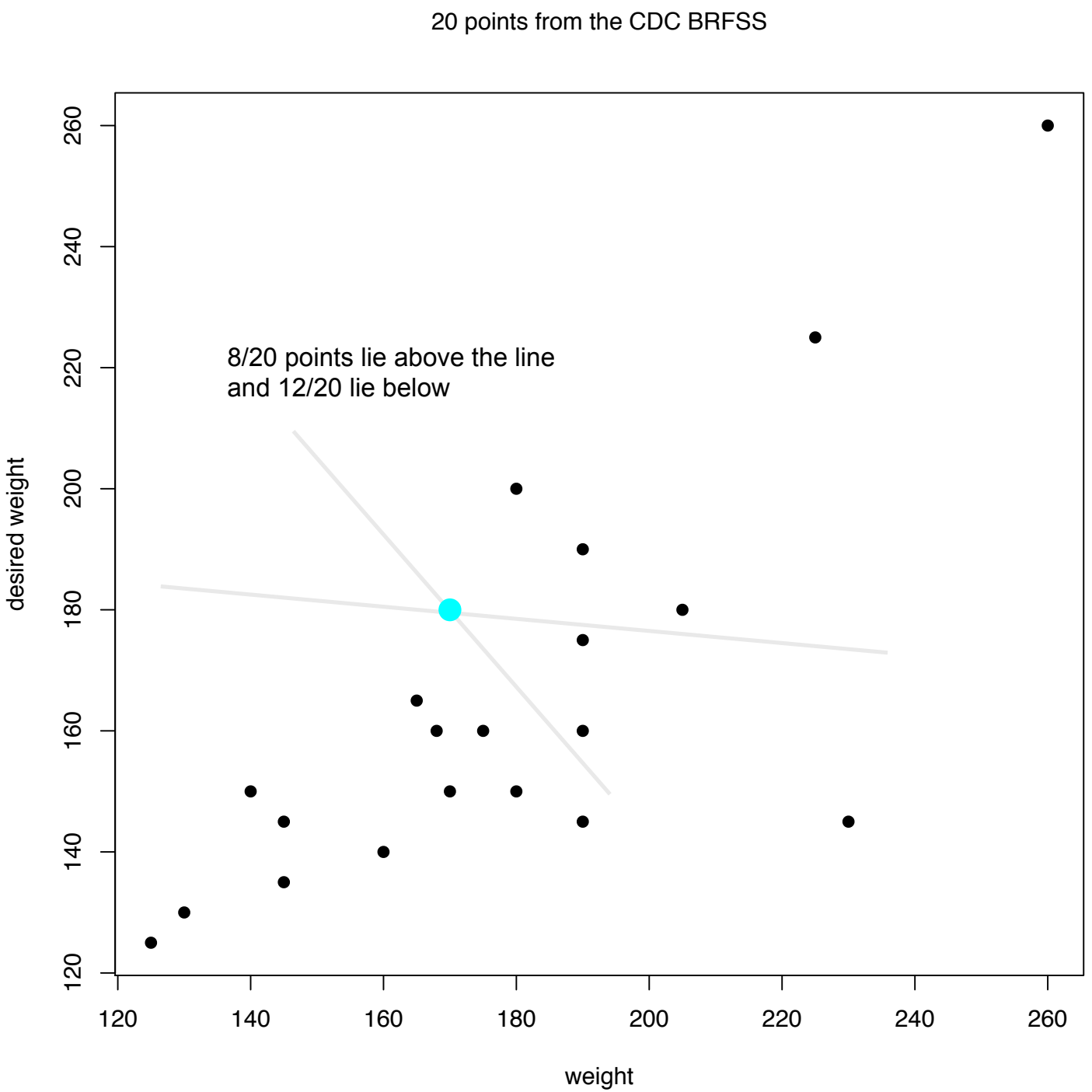
The median, then, is **the location on the real line having the greatest depth** -- If an “interval” of locations have greatest depth (as is the case on the previous slides, where any location between and including the 5th and 6th data points all have depth 1/2) we take the midpoint of the interval as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

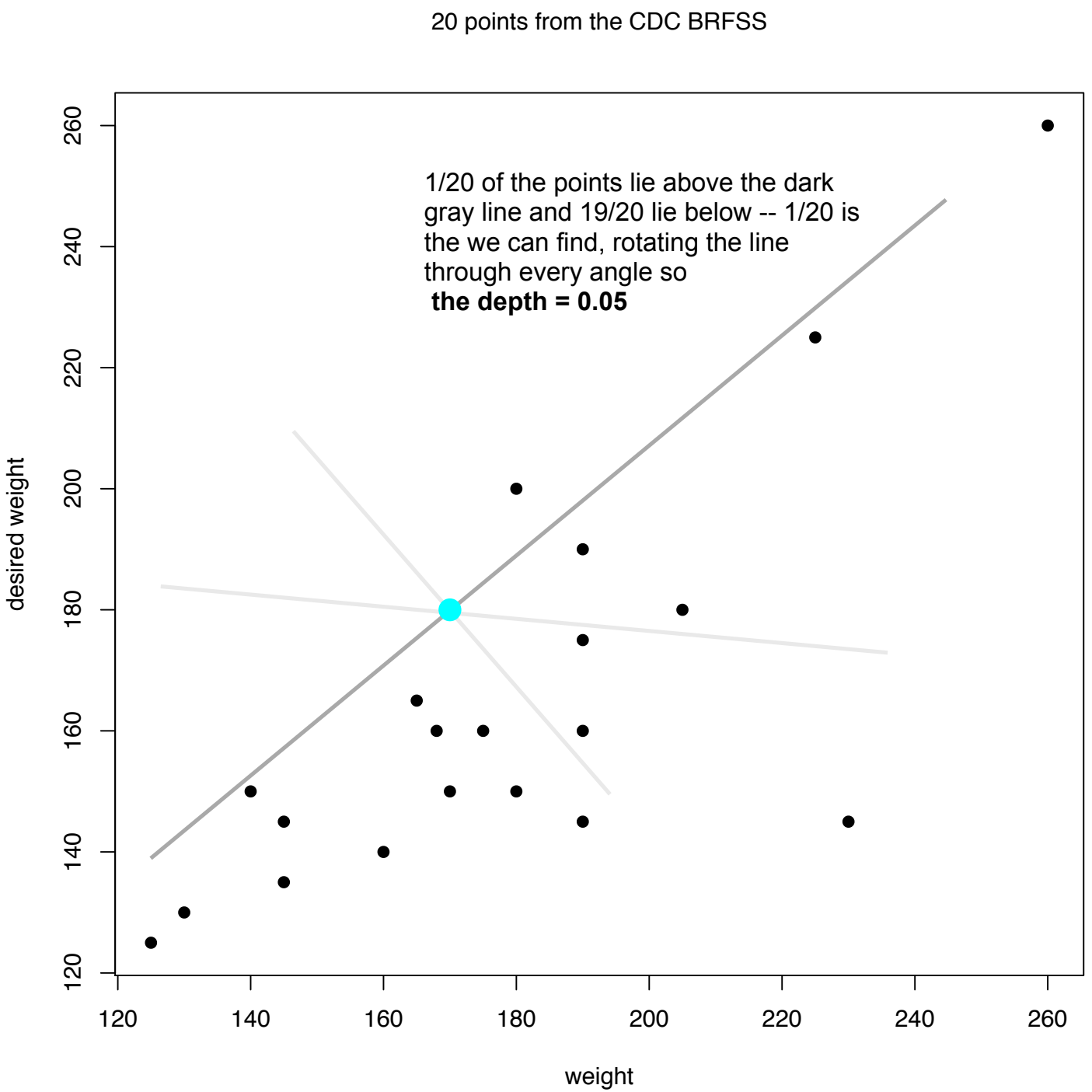
Now, how do we generalize this to two dimensions? How do we generalize the notion of left and right?

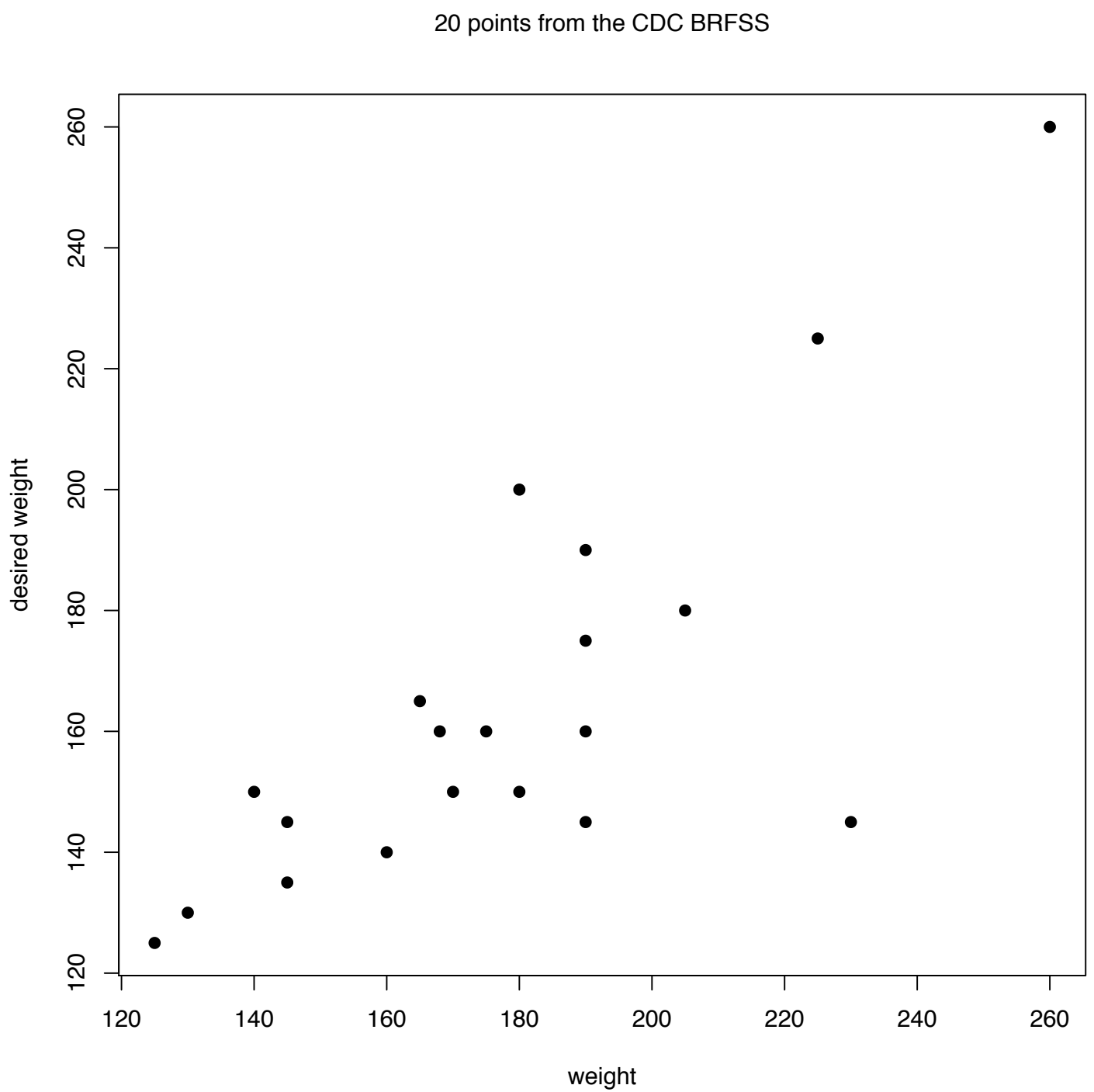

20 points from the CDC BRFSS
desired weight
weight
120
140
160
180
200
220
240
260

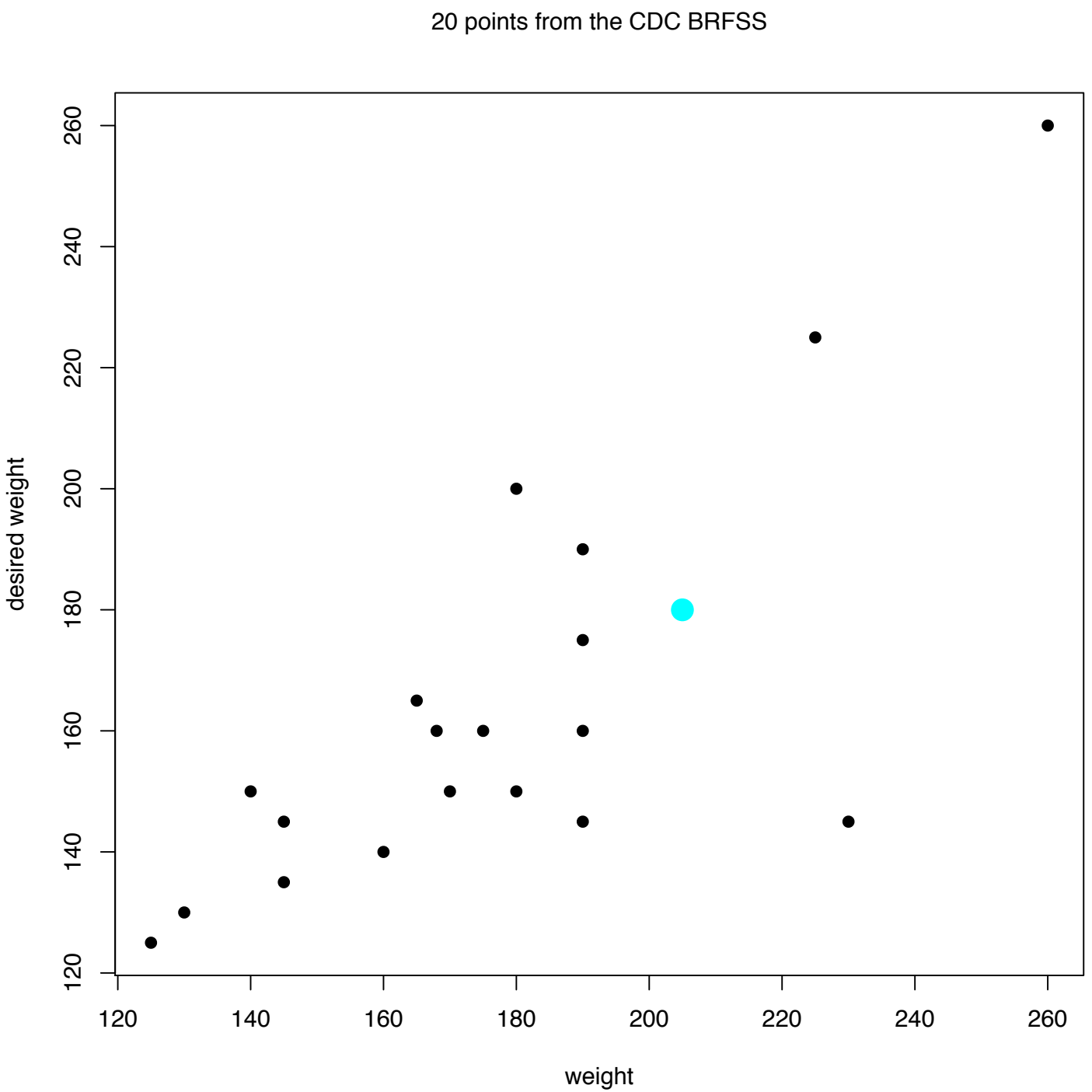

20 points from the CDC BRFSS
desired weight
weight
120
140
160
180
200
220
240
260

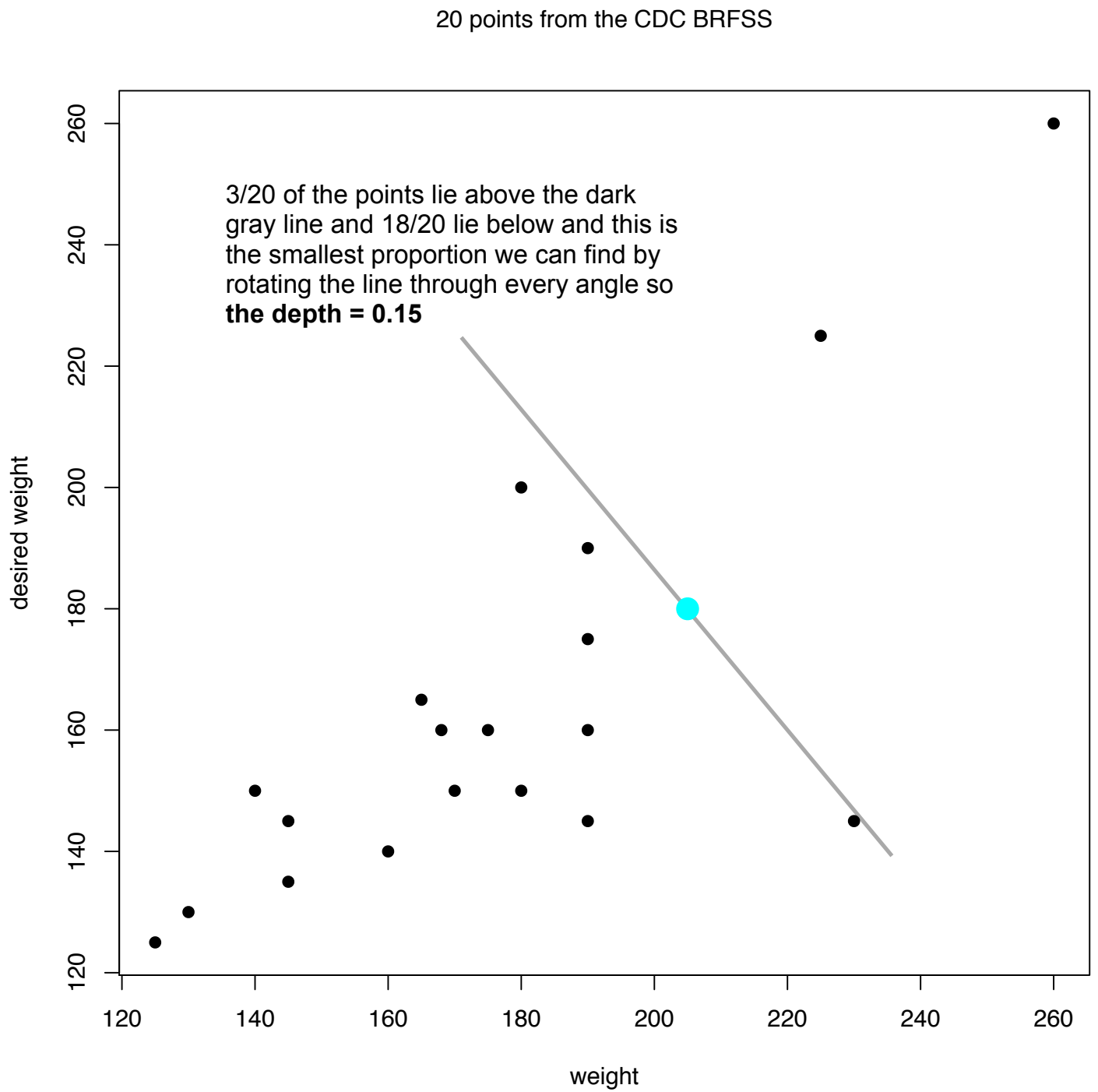

20 points from the CDC BRFSS
desired weight
weight
120
140
160
180
200
220
240
260
5/20 points lie above the line
and 15/20 lie below

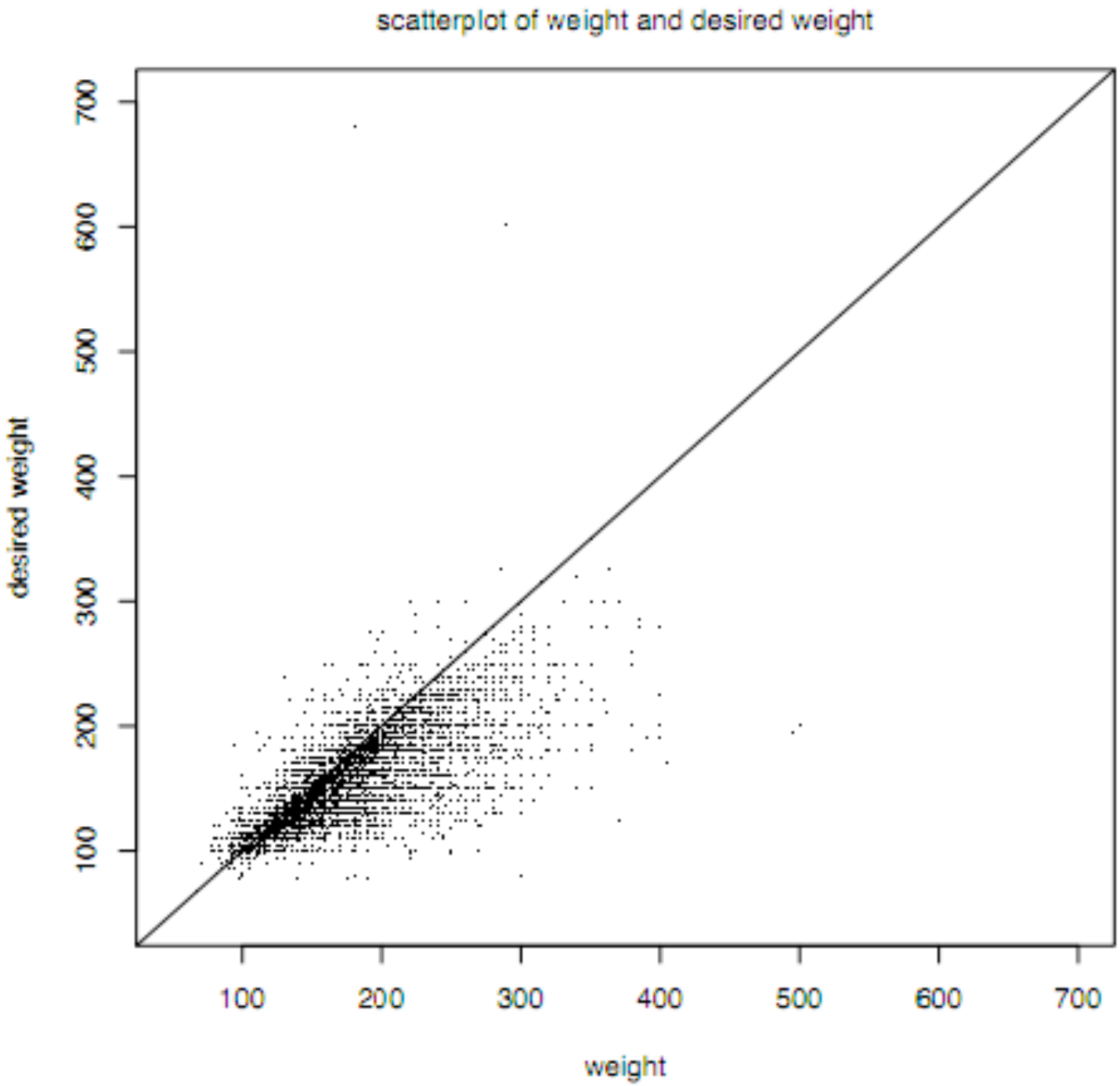

20 points from the CDC BRFSS
8/20 points lie above the line
and 12/20 lie below
desired weight
weight
120
140
160
180
200
220
240
260

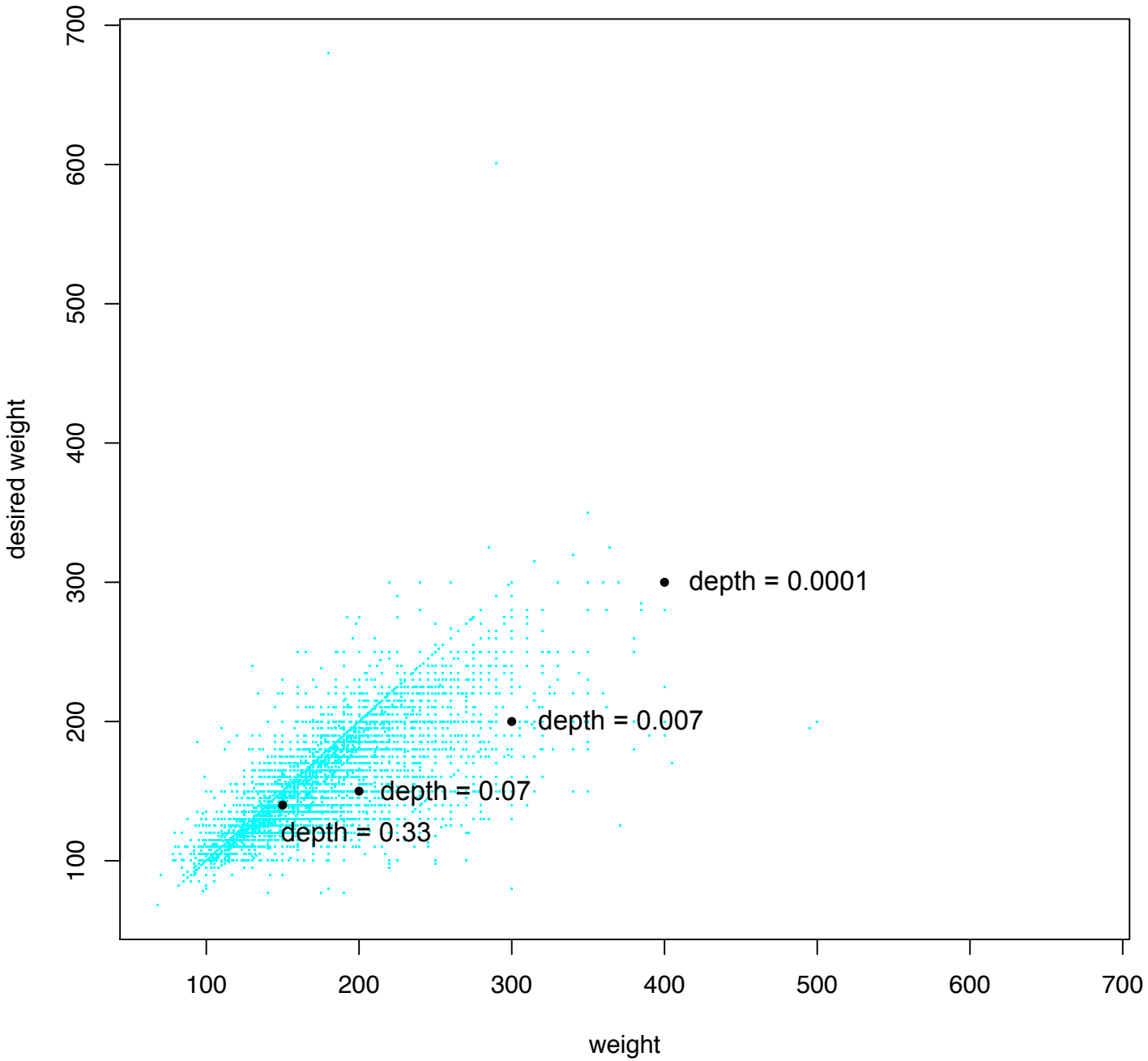

20 points from the CDC BRFSS
1/20 of the points lie above the dark gray line and 19/20 lie below -- 1/20 is the we can find, rotating the line through every angle so
the depth = 0.05
desired weight
weight
120
140
160
180
200
220
240
260


20 points from the CDC BRFSS
desired weight
weight
120
140
160
180
200
220
240
260


20 points from the CDC BRFSS
desired weight
weight
120
140
160
180
200
220
240
260


20 points from the CDC BRFSS
3/20 of the points lie above the dark gray line and 18/20 lie below and this is the smallest proportion we can find by rotating the line through every angle so
the depth = 0.15
desired weight
weight
120
140
160
180
200
220
240
260


scatterplot of weight and desired weight
desired weight
weight
100
200
300
400
500
600
700


desired weight
weight
100
200
300
400
500
600
700
depth = 0.0001
depth = 0.007
depth = 0.07
depth = 0.33

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:

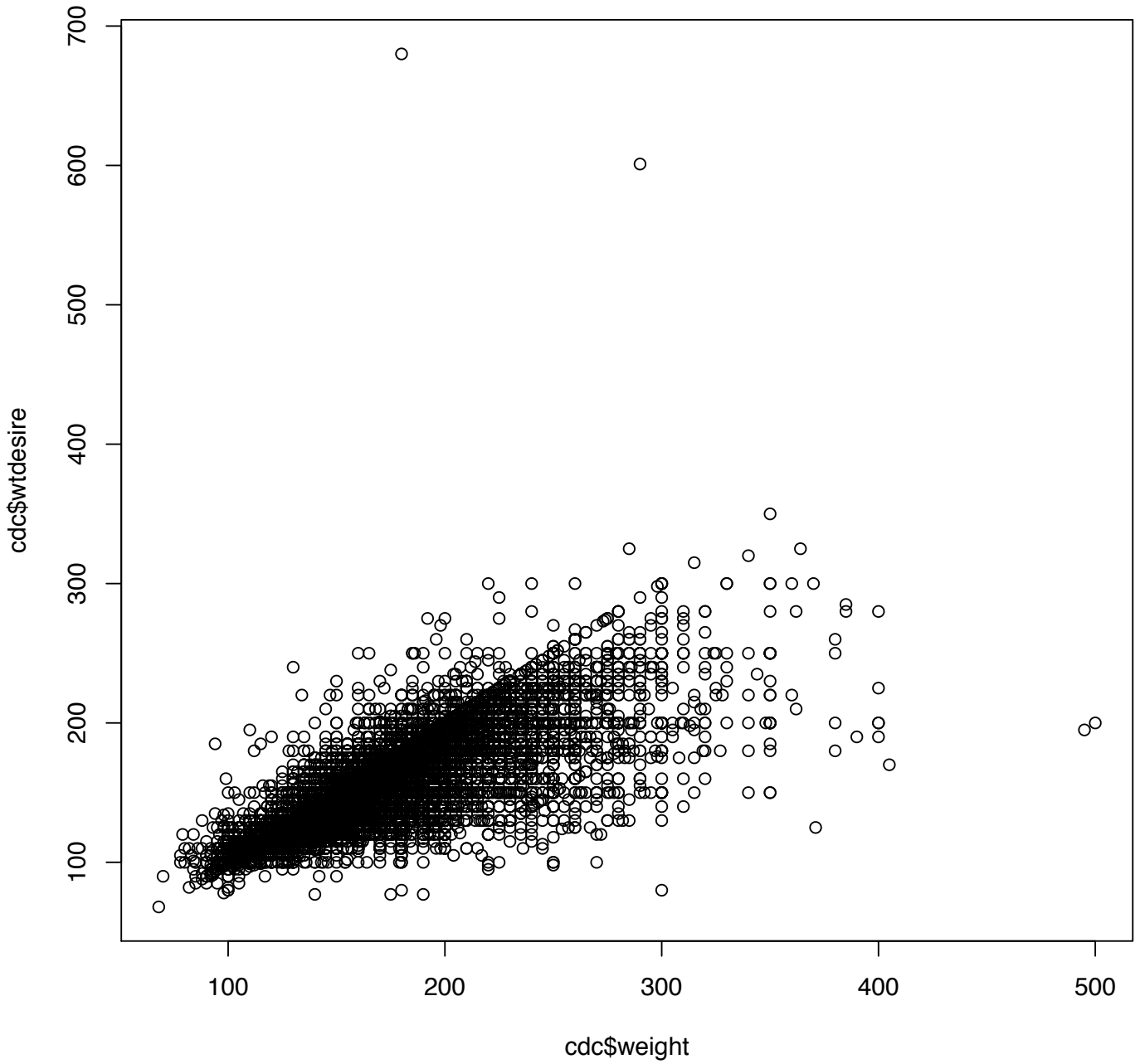
> *Like the univariate boxplot, the bagplot also visualizes several characteristics of the data:* **its location** *(the depth median),* **spread** *(the size of the bag),* **correlation** *(the orientation of the bag),* **skewness** *(the shape of the bag and the loop), and* **tails** *(the points near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

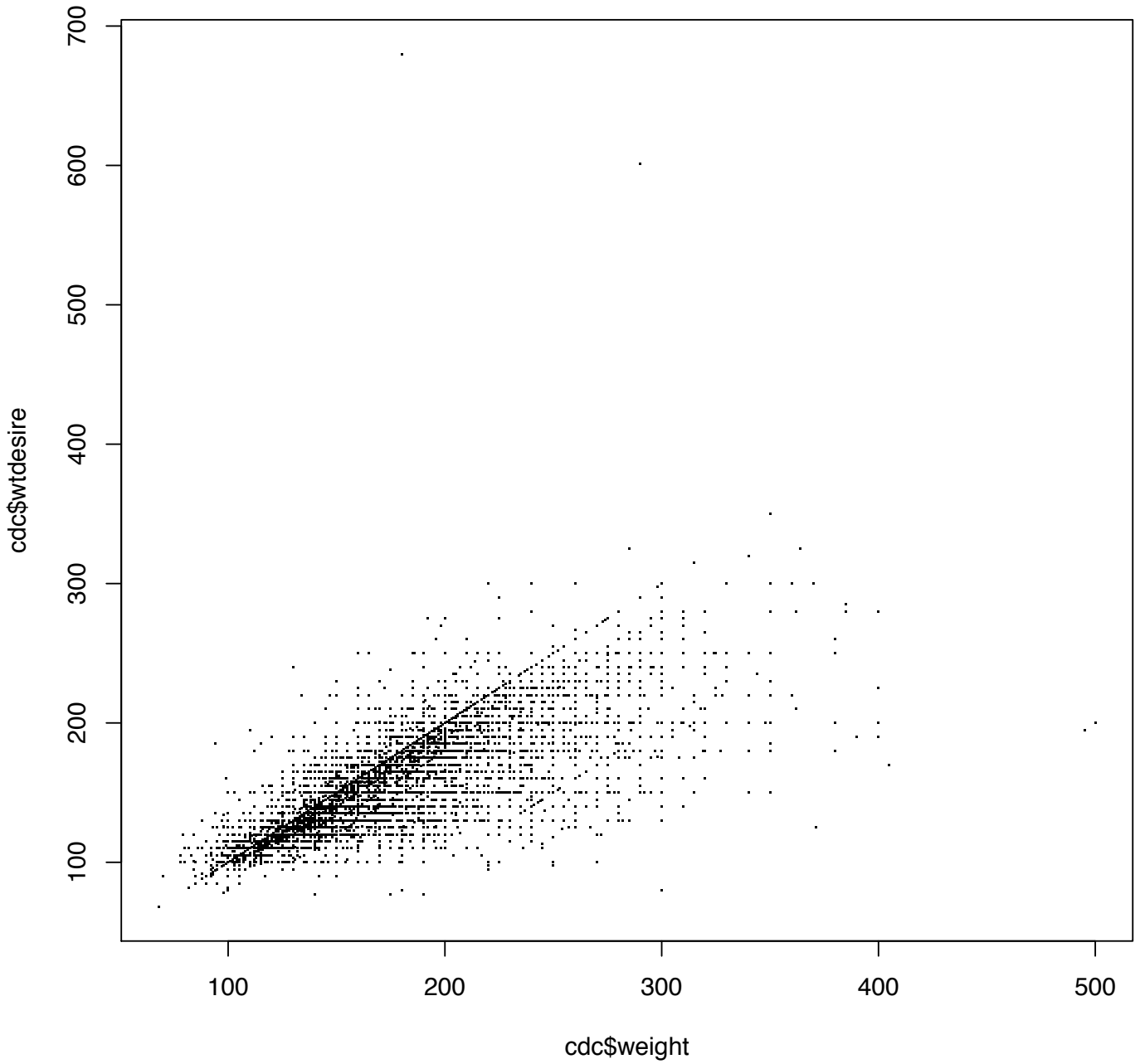
Scatterplots

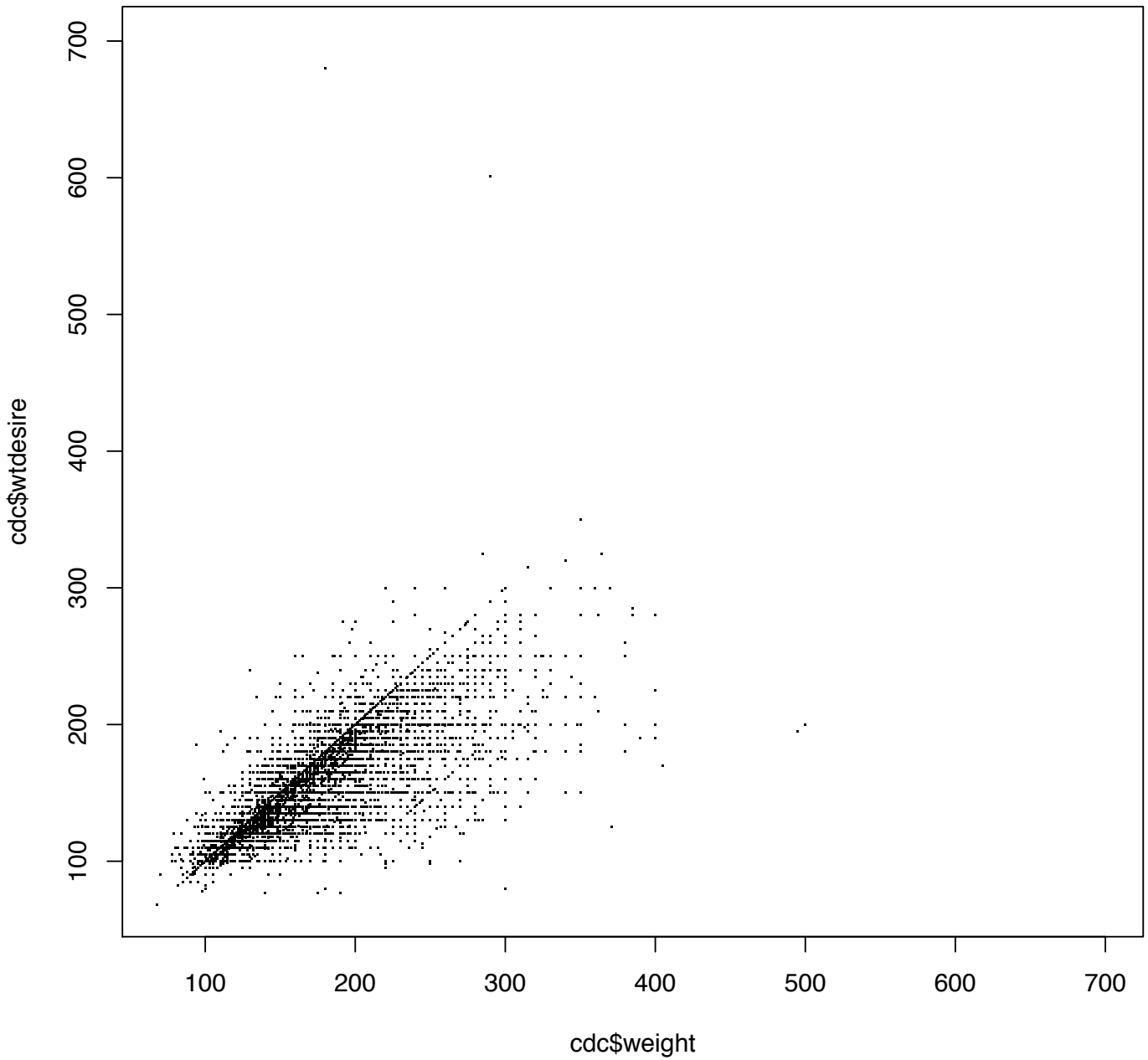
For two continuous variables, we have already introduced a scatterplot as a tool for assessing associations -- R provides considerable control over what a plot like this looks like

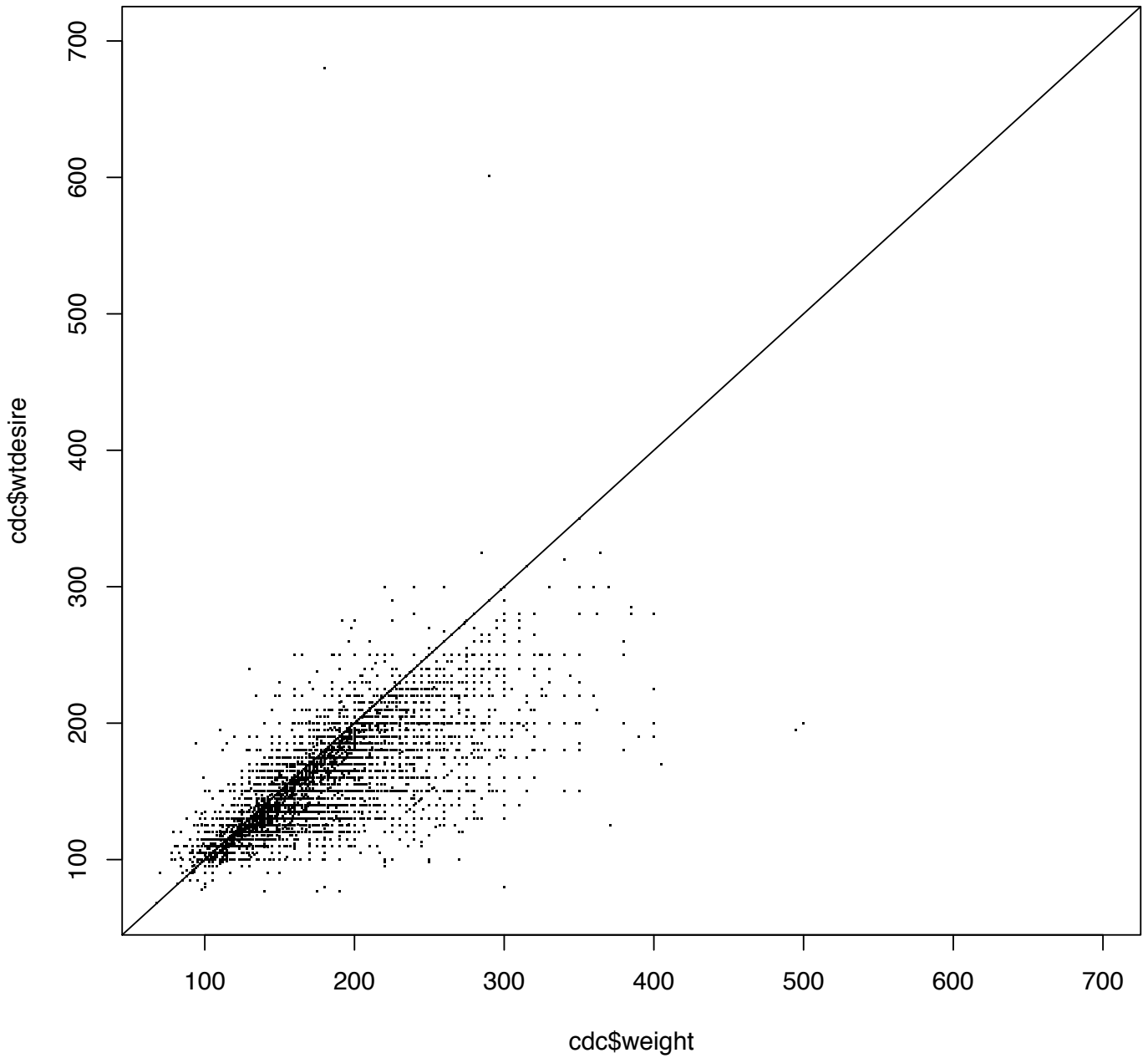
On the next few slides, we change plotting characters as well as the range of the x- and y-axes on the plot of respondents' desired weight versus their current weight -- We also add a reference line with mean 0 and slope 1 (Why?)

```
# first plot
plot(cdc$weight,cdc$wtdesire)

# second plot, changing plotting character
plot(cdc$weight,cdc$wtdesire,pch=".")

# third plot, changing range of x- and y-axes
plot(cdc$weight,cdc$wtdesire,pch=".",xlim=c(70,700),ylim=c(70,700))

# fourth plot, adding a reference line with slope 1 and intercept 0
plot(cdc$weight,cdc$wtdesire,pch=".",xlim=c(70,700),ylim=c(70,700))
abline(0,1)
```


cdc$wtdesire
cdc$weight
100
200
300
400
500
600
700


cdc$wtdesire
700
600
500
400
300
200
100
100
200
300
400
500
cdc$weight


cdc$wtdesire
100
200
300
400
500
600
700
cdc$weight
100
200
300
400
500
600
700

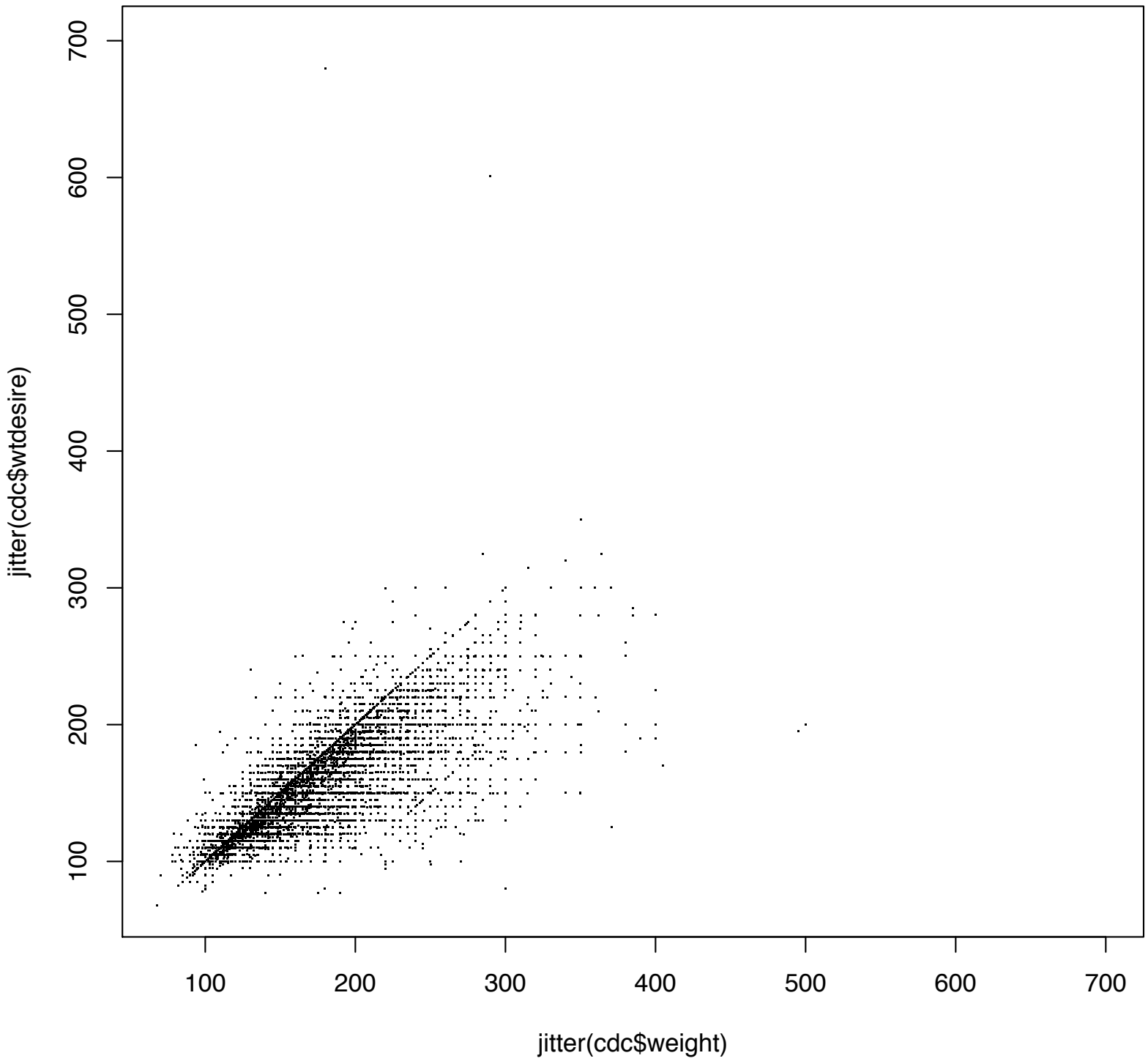

cdc$wtdesire
100
200
300
400
500
600
700
cdc$weight
100
200
300
400
500
600
700

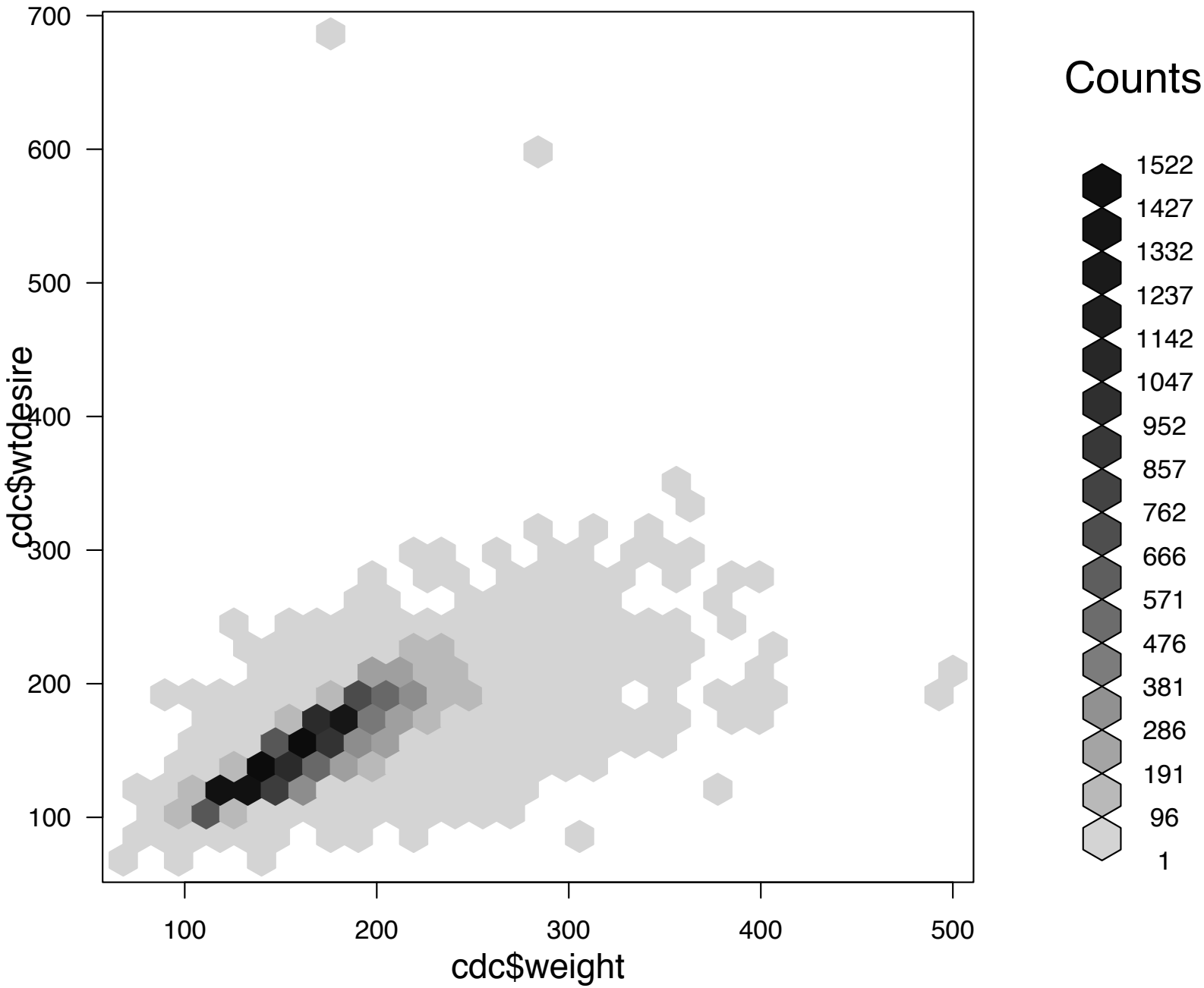
Alterations

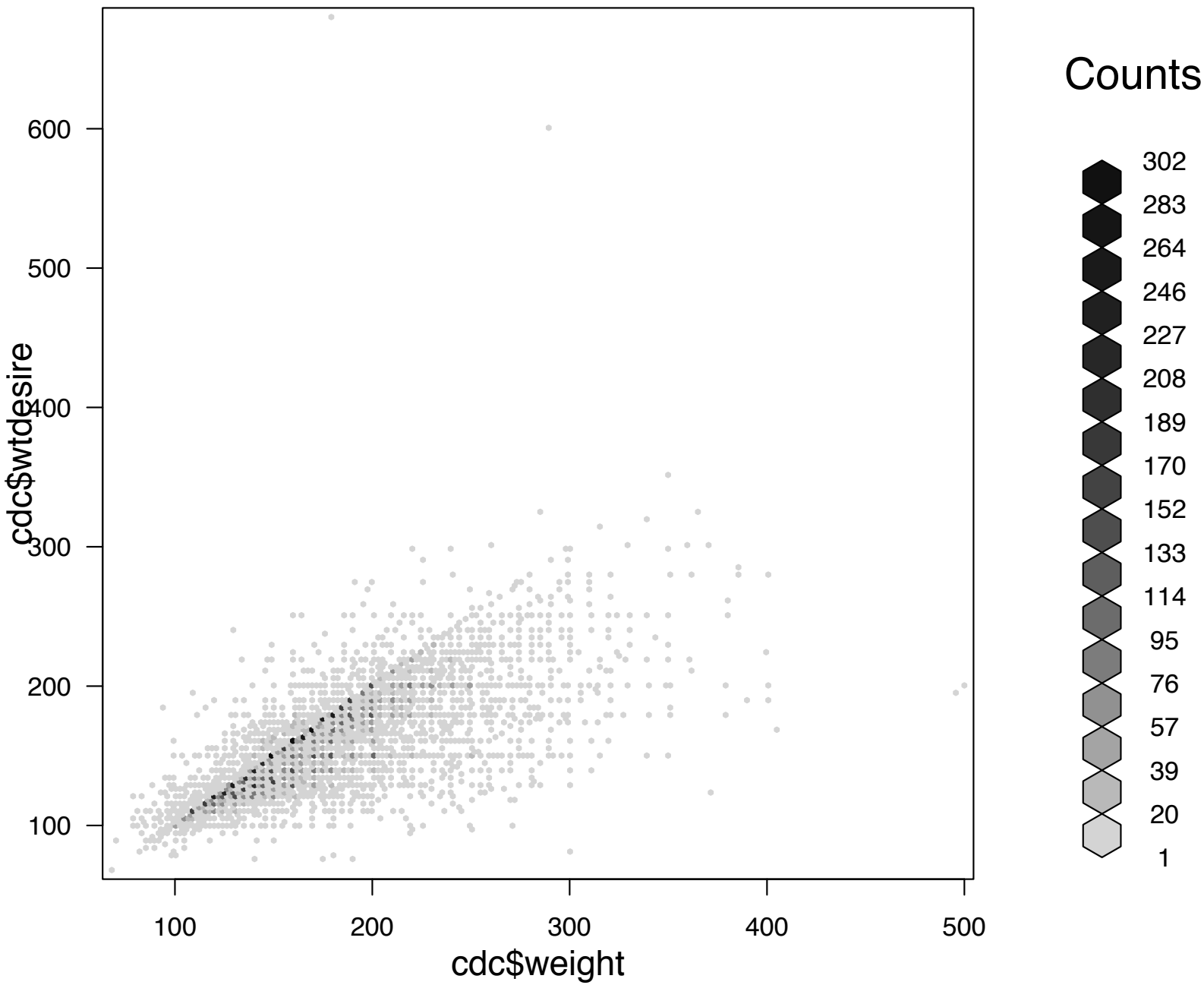
One problem that we've ignored with this plot is that it represents **20,000 pairs of points** -- It's hard to see that because there is considerable **over-plotting**

That is, we have seen already that weight tends to be reported in 5-pound increments, and the same is true for respondents' desired weights -- Inevitably that means that some pairs of points are represented multiple times

What can we do to bring these points out?


jitter(cdc$wtdesire)
100
200
300
400
500
600
700
jitter(cdc$weight)
100
200
300
400
500
600
700


700
600
500
400
300
200
100
cdc$wtdesire
100
200
300
400
500
cdc$weight
Counts
1522
1427
1332
1237
1142
1047
952
857
762
666
571
476
381
286
191
96
1


600
500
400
300
200
100
cdc$wtdesire
100
200
300
400
500
cdc$weight
Counts
302
283
264
246
227
208
189
170
152
133
114
95
76
57
39
20
1

```
# first plot
plot(cdc$weight,cdc$wtdesire)

# second plot, changing plotting character
plot(cdc$weight,cdc$wtdesire,pch=".")

# third plot, changing range of x- and y-axes
plot(cdc$weight,cdc$wtdesire,pch=".",xlim=c(70,700),ylim=c(70,700))

# fourth plot, adding a reference line with slope 1 and intercept 0
plot(cdc$weight,cdc$wtdesire,pch=".",xlim=c(70,700),ylim=c(70,700))
abline(0,1)

# final plots, jittering...
plot(jitter(cdc$weight),jitter(cdc$wtdesire),pch=".",xlim=c(70,700),ylim=c(70,700))

# then binning the data using hexbin...
h <- hexbin(cdc$weight,cdc$wtdesire)
plot(h)

# ... or in one go (and changing the number of bins to 200)
plot(hexbin(cdc$weight,cdc$wtdesire,200))
```

Scatterplots

The scatterplot is so ubiquitous that we barely recognize what we're doing -- It seems sensible and almost automatic and, hence, **practically invisible**

While this kind of plot is great for two dimensions, what do we do if we have three variables? Four variables? Twenty? What are our options now?

PNAS PNAS PNAS

Estimating least-developed countries' vulnerability to climate-related extreme events over the next 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwabena Asante[d], Marc Metzger[e,f], Jose Rafael[g], Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]Climate Systems Analysis Group, University of Cape Town, Rondebosch 7701, South Africa; [c]Institute of Environmental Science and Technology, Autonomous University of Barcelona, 08193 Bellaterra, Spain; [d]Climatus LLC, Mountain View, CA 94041; [e]Centre for the Study of Environmental Change and Sustainability, University of Edinburgh, EH8 9XP, Scotland; [f]Alterra, Wageningen University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Department of Geography, University of Eduardo Mondlane, Maputo, Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences, 1040 Vienna, Austria; and [i]Department of Oceanography, University of Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approved December 4, 2009 (received for review September 10, 2009)

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may

sensitivity to those stressors, which in turn is determined by a complex set of social, economic, and institutional factors collectively described as determining its adaptive capacity (5, 6). As the UNFCCC secretariat suggested in its needs assessment, "one of the key limitations in estimating the costs of adaptation is the uncertainty about adaptive capacity. Adaptive capacity is essentially the ability to adapt to stresses such as climate change. It does not predict what adaptations will happen, but gives an indication of differing capacities of societies to adapt *on their own* to climate change or other stresses" (1, p. 97).

Human losses to extreme weather events can serve as a reliable indicator for this vulnerability, and with it the need for financial assistance, for two reasons. First, measures to reduce vulnerability to extreme weather events account for a particularly large share of estimated adaptation financial needs (1). Second, in the context of efforts to achieve a wide range of development goals, it is only

Vulnerability

The underlying question here is interesting and relevant (they usually are, for what it's worth) -- Here we are interested in understanding how climate change (and the accompanying increase in extreme weather events) will affect different parts of the world

Specifically, the researchers produce a model that relates variables capturing some notion of vulnerability to the impacts that weather-related natural disasters have had, country by country

Estimating least-developed to climate-related extreme 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwa
Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]
South Africa; [c]Institute of Environmental Science and Technology, Autonomous
View, CA 94041; [d]Centre for the Study of Environmental Change and Sustainab
University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Depa
Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences
Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approv

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may begin to offset rising climate exposure in the second quarter of the century, and that it is in the period between now and then that vulnerability will rise most quickly. This implies an urgency to the need for international assistance to finance adaptation.

vulnerability | adaptive capacity | development | natural disasters | natural hazards

Results

The first stage of our analysis was to estimate statistical models of losses from climate-related disasters, based on a set of climatic and socio-economic variables that will likely change over time, which appear in Table 1. The dependent variables are logged values of the number of people per million of national population killed or affected, respectively, by droughts, floods, or storms over the period 1990–2007. **The variable number of disasters is the logged value of numbers reported by each country over the same period, and accounts for climate exposure**; estimated coefficient values greater than 1 in both models indicate that average losses per disaster are higher in more disaster-prone countries. We expected that larger countries are likely to experience disasters over a smaller proportion of their territory or population, and also benefit from potential economies of scale in their disaster management infrastructure, both resulting in lower average per capita losses; the negative coefficient estimates for the variable national population in both models are consistent with this expectation. **The variable HDI represents the Human Development Index, a United Nations (UN) indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth.** Recent studies of disaster losses —not limited to climate-related events—have shown that countries with medium HDI values experience the highest average losses, whereas countries with high HDI values experience the lowest (14, 15). We therefore included the logged HDI values in quadratic form. Negative coefficient estimates for both HDI and HDI^2 in both models are thus consistent with these expectations, given that logged HDI values are always negative, and the square of the logged values are in turn positive. Finally, we considered several additional socio-economic variables not directly captured by HDI, and found only two that improved model fit. **For the model of the number of people killed, the positive coefficient estimate for female fertility indicates that countries with higher birth rates experience greater average numbers of deaths. We do not take this to mean that there is a direct connection between fertility and natural hazard deaths, but rather that higher birth rates are associated with lower female empowerment, and lower female empowerment is associated with higher disaster vulnerability**, as has been shown previously (16, 17). For the model of the number of people affected, the negative coefficient estimate for the proportion urban population is consistent with urban residents being less likely to require postdisaster assistance than rural residents, also observed previously (18, 19). Both models yield an R^2statistic slightly greater than 0.5, indicating that variance in the independent variables explains just over half ofthe variance in the numbers killed and affected. This is consistent with results from past analyses based on similar data and methods (8–10)

Vulnerability

In the end, a great deal of attention is paid to a regression table (below), the form of which we should be fairly familiar with

In each row they present the regression of the logarithm of the number of people killed by weather-related natural disasters from 1990 to 2007 as a function of several predictors, one of which is slightly special...

Table 1. Ordinary least-squares regression results

Independent variables	Killed	Affected
Number of disasters	1.36* (0.15)	1.88* (0.19)
National population	−0.56* (0.09)	−0.79* (0.11)
HDI	−5.97* (1.95)	−13.55* (2.16)
HDI^2	−6.26* (1.52)	−9.82* (1.86)
Female fertility	1.45* (0.43)	
Proportion urban population		−0.41 (0.37)
Constant	−3.86* (0.49)	5.33* (1.71)
Number of observations	150	154
R^2	0.52	0.55

The dependent variable in the Killed model is the logged value of the number of people reported by CRED as killed by the three types of disasters considered (droughts, floods, and storms) divided by population. The dependent variable in the *Affected* model is the same for the number of people reported affected, but not killed, by the same disasters. All independent variables are logged values. Because HDI occupies the range of 0–1, all logged HDI values were negative, whereas the squares of these values were positive. *Values significant (*two-tailed student's t* test) at the 99% confidence level. Values in parentheses are SEs.

Looking at the data

The data we were given consist of measurements associated with 144 different countries -- For each we have the following variables

`country_name` the name of the country

`ln_events` the natural logarithm of the number of droughts, floods and storms occurring in the country from 1990-2007

`ln_pop` the natural logarithm of the country's population

`ln_fert` the natural logarithm of an estimate of the country's female fertility

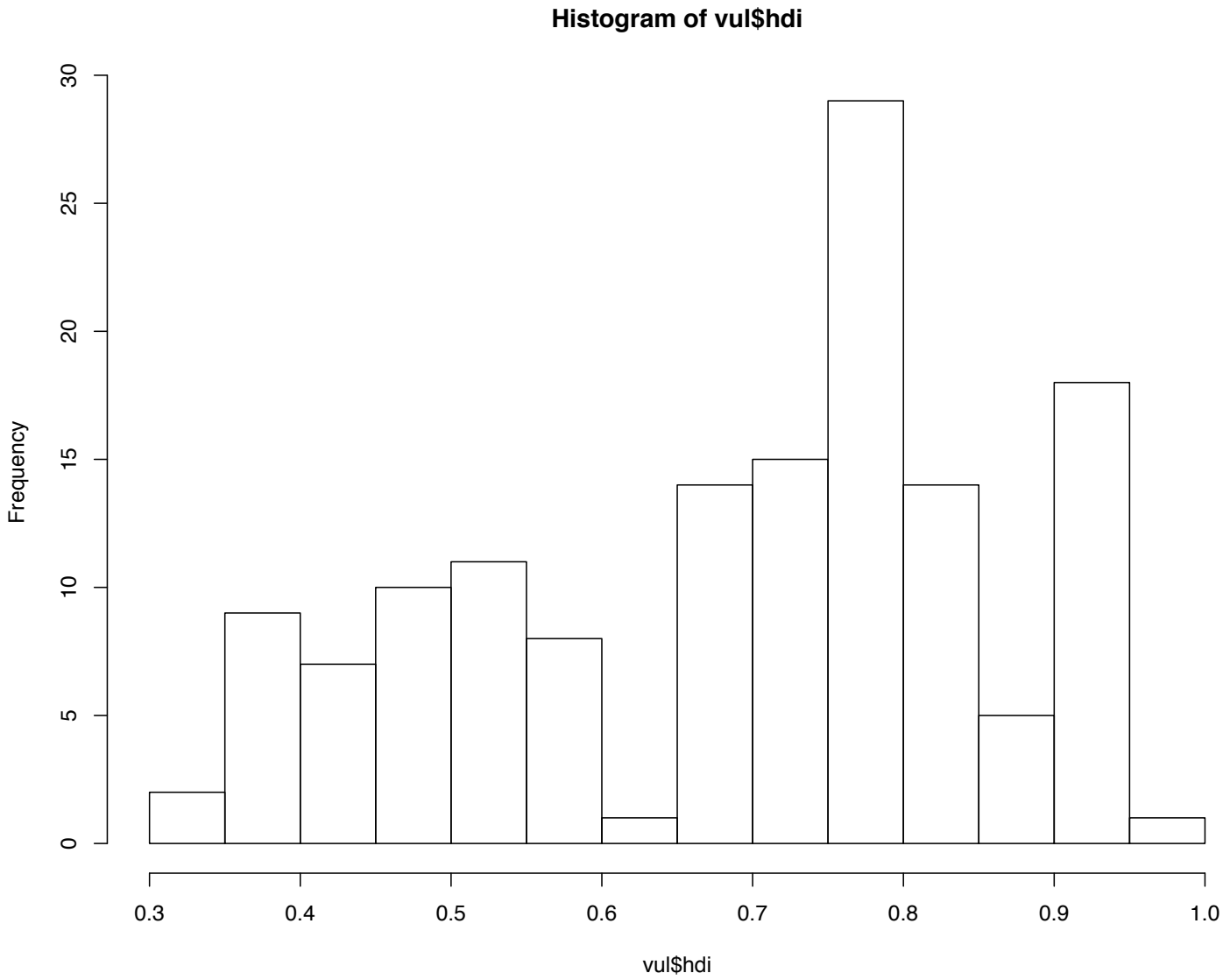
`hdi` the Human Development Index for the country

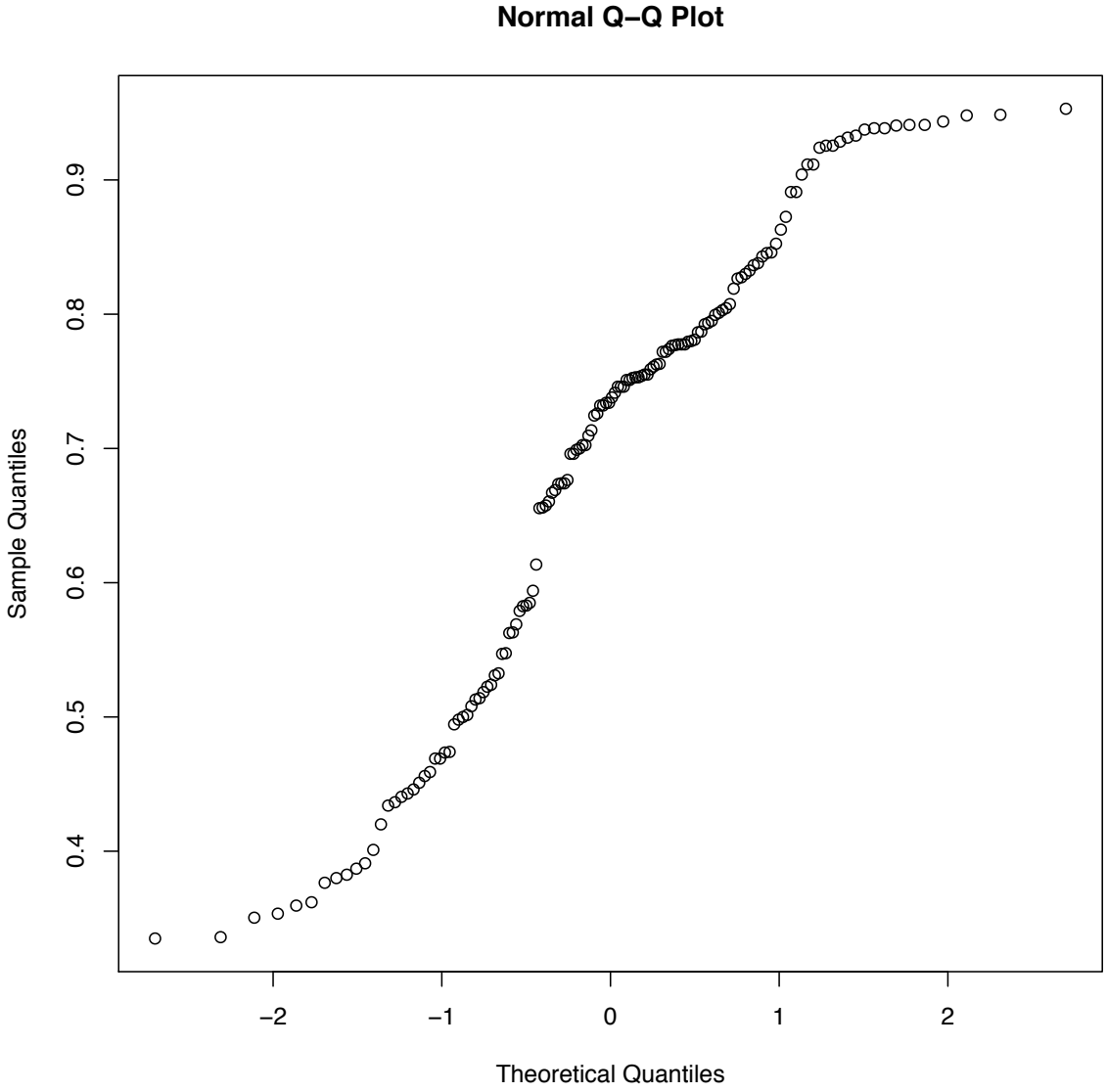
`death_risk` the proportion of people out of 1M in population killed in droughts, floods and storms

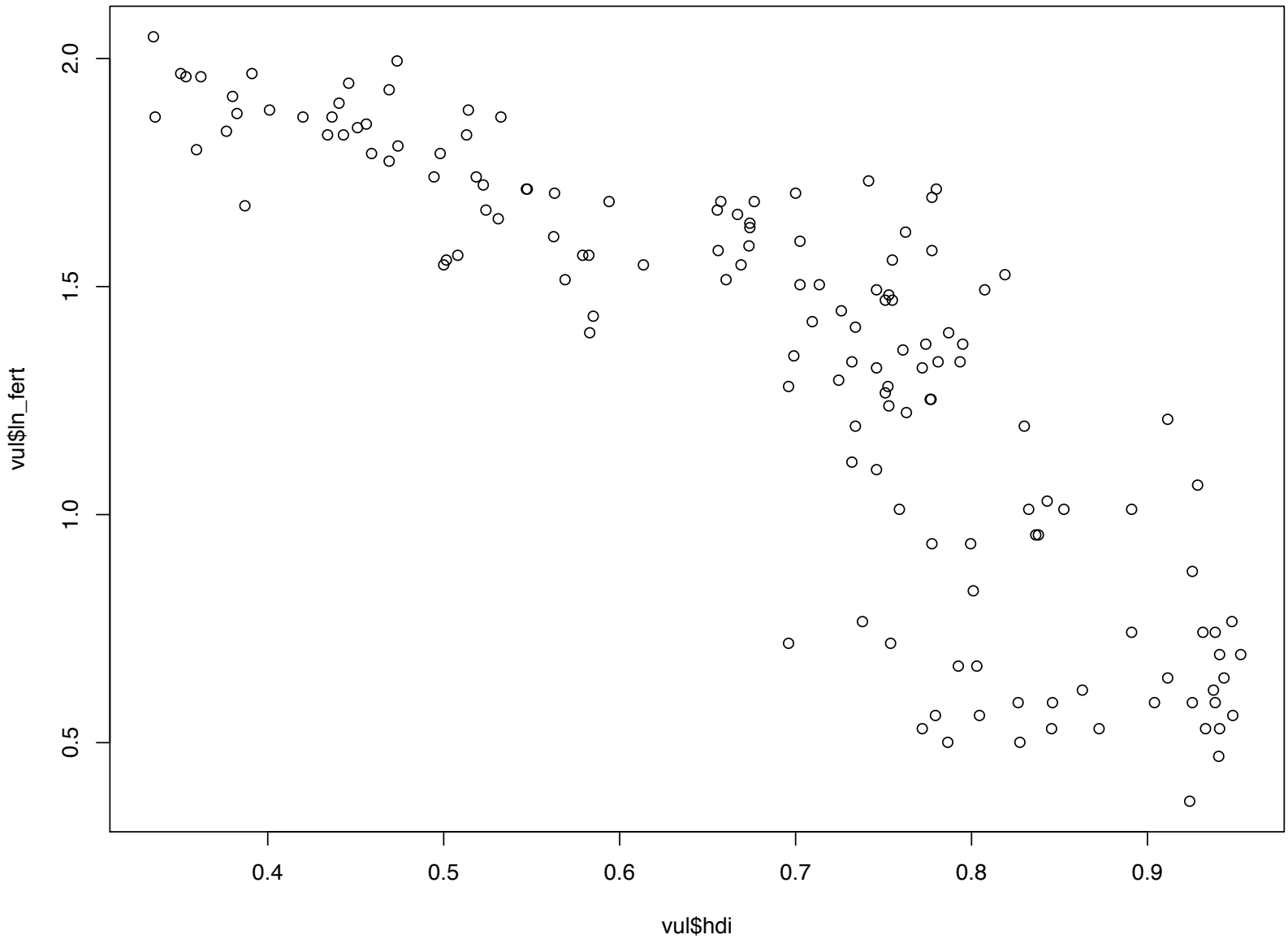
There are four predictor variables (if you count HDI and its square as one) which, while not big by any stretch of the imagination, is complex enough to keep us from "seeing" the whole data set

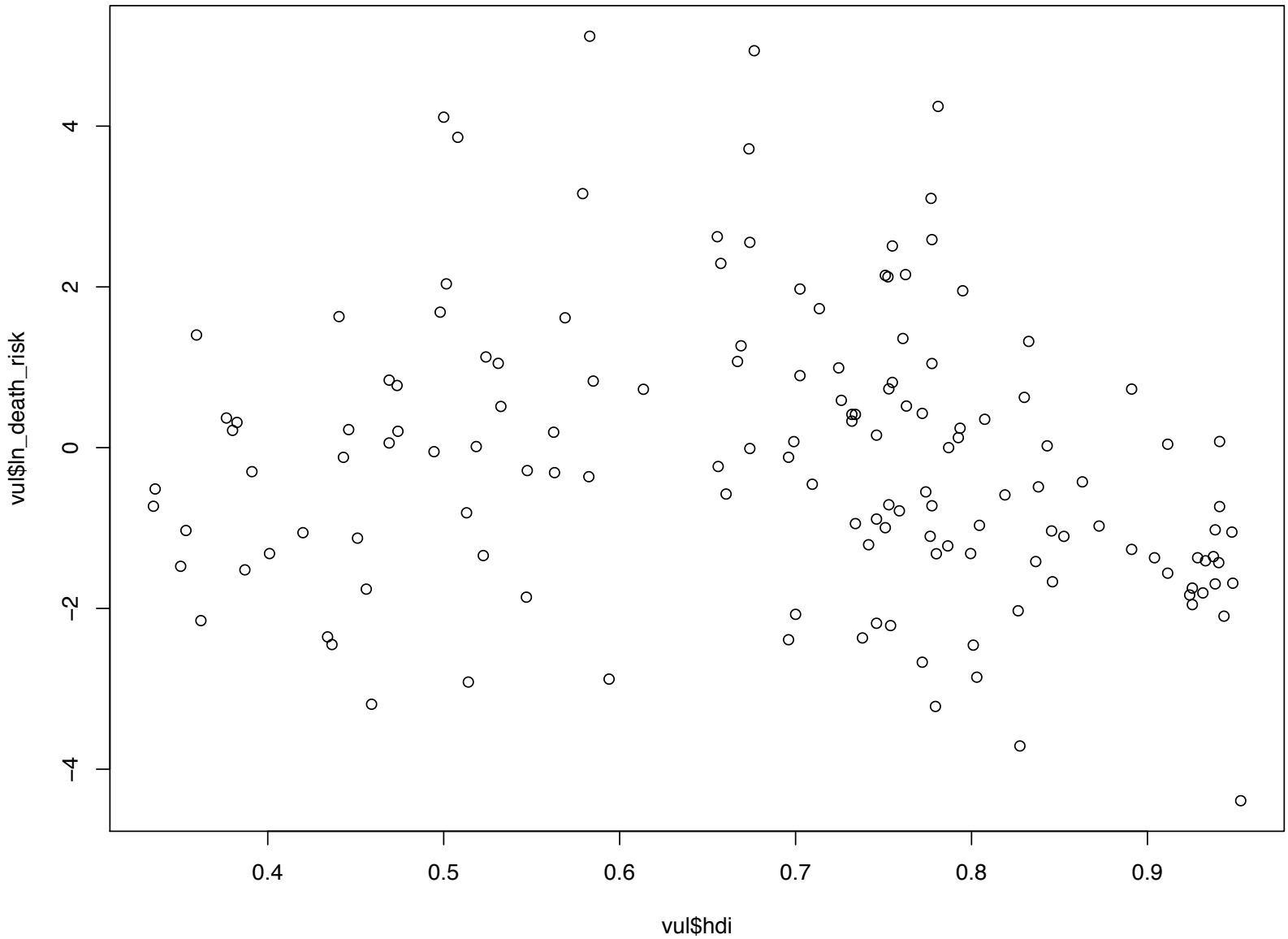
Instead, we might opt for partial views...

```
# the first few countries...

> head(vul)
  country_name ln_events   ln_fert       hdi   ln_pop ln_death_risk
1      Albania  2.302585 1.2383740 0.7530000 4.006120    -0.7102835
2      Algeria  3.496508 1.5993880 0.7025000 6.283885     0.8961845
3       Angola  3.044523 1.9459100 0.4460000 5.556056     0.2246880
4    Argentina  3.637586 1.0116010 0.8525001 6.483515    -1.1036180
5      Armenia  1.386294 0.7654679 0.7380000 3.976562    -2.3671240
6    Australia  4.394449 0.7654679 0.9480000 5.837925    -1.0504330

# ... and the last few countries

> tail(vul)
           country_name ln_events  ln_fert    hdi   ln_pop ln_death_risk
139           Venezuela  2.995732 1.335001 0.7810 6.072891     4.2457150
140            Viet Nam  4.595120 1.504077 0.7025 7.248978     1.9736860
141               Yemen  3.091043 1.994700 0.4735 5.808543     0.7734824
142 Zaire/Congo Dem Rep  2.944439 1.887070 0.4010 6.846872    -1.3174430
143              Zambia  2.564949 1.871802 0.4365 5.222156    -2.4495670
144            Zimbabwe  2.484907 1.704748 0.5630 5.360353    -0.3104967

# order according to death_risk

> tail(vul[order(vul$ln_death_risk),])
    country_name ln_events  ln_fert    hdi   ln_pop ln_death_risk
93     Nicaragua  3.135494 1.589235 0.6735 4.499810      3.717359
55         Haiti  3.784190 1.568616 0.5080 5.033049      3.860935
10    Bangladesh  4.836282 1.547562 0.5000 7.828728      4.111300
139    Venezuela  2.995732 1.335001 0.7810 6.072891      4.245715
56      Honduras  3.496508 1.686399 0.6765 4.701086      4.938241
88       Myanmar  2.772589 1.398717 0.5830 6.688770      5.118413

> max(vul$ln_events)
[1] 5.948035

> exp(max(vul$ln_events))
[1] 383
```


Histogram of vul$hdi
Frequency
0
5
10
15
20
25
30
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
vul$hdi


Normal Q–Q Plot
Sample Quantiles
0.4
0.5
0.6
0.7
0.8
0.9
−2
−1
0
1
2
Theoretical Quantiles

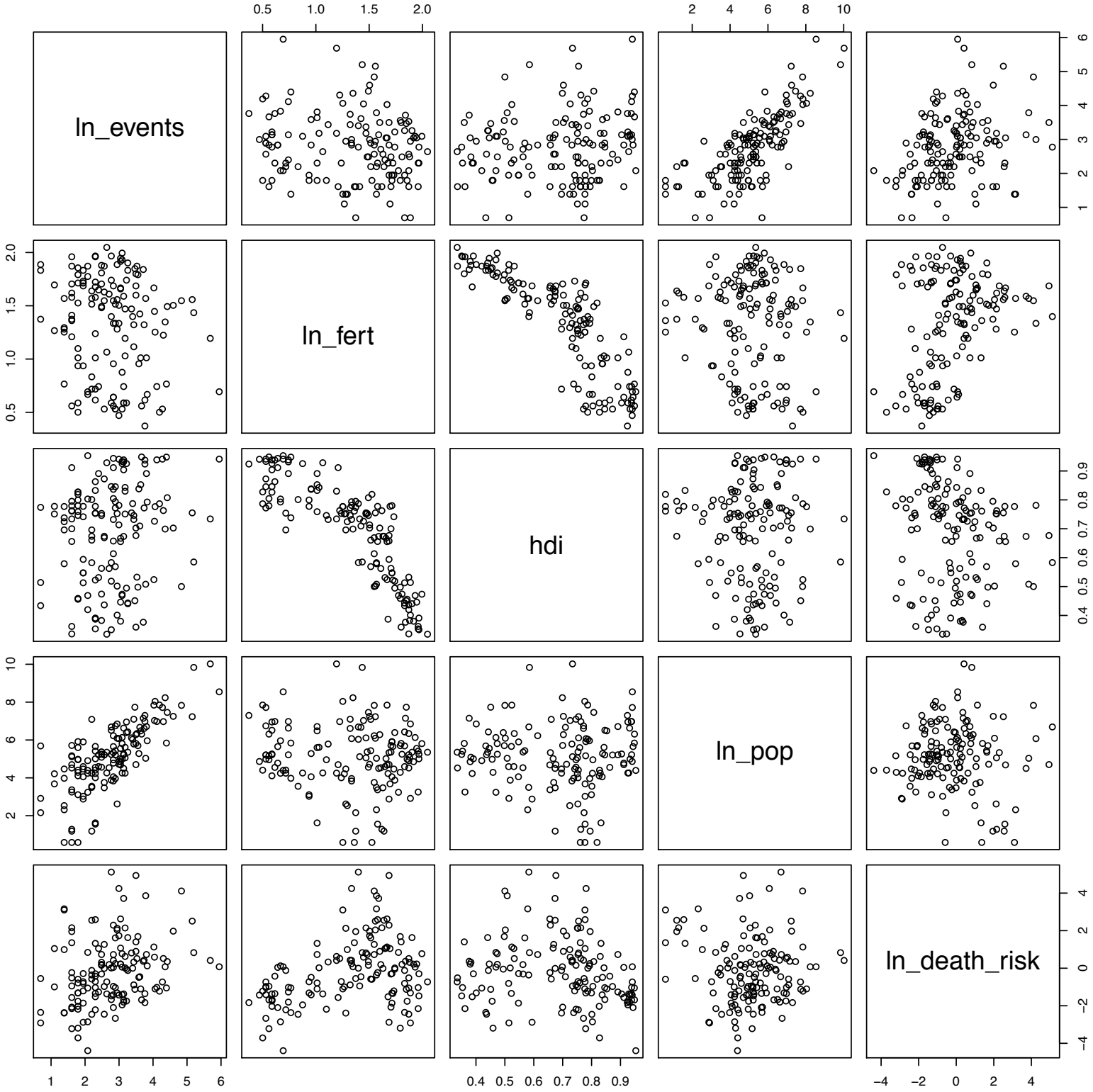

vul$ln_fert
2.0
1.5
1.0
0.5
0.4
0.5
0.6
0.7
0.8
0.9
vul$hdi


vul$ln_death_risk
4
2
0
−2
−4
0.4
0.5
0.6
0.7
0.8
0.9
vul$hdi

Scatterplot matrix

A scatterplot lets us examine pairs of variables -- We can stack the plots for all possible pairs of data points in the form of a matrix, a scatterplot matrix

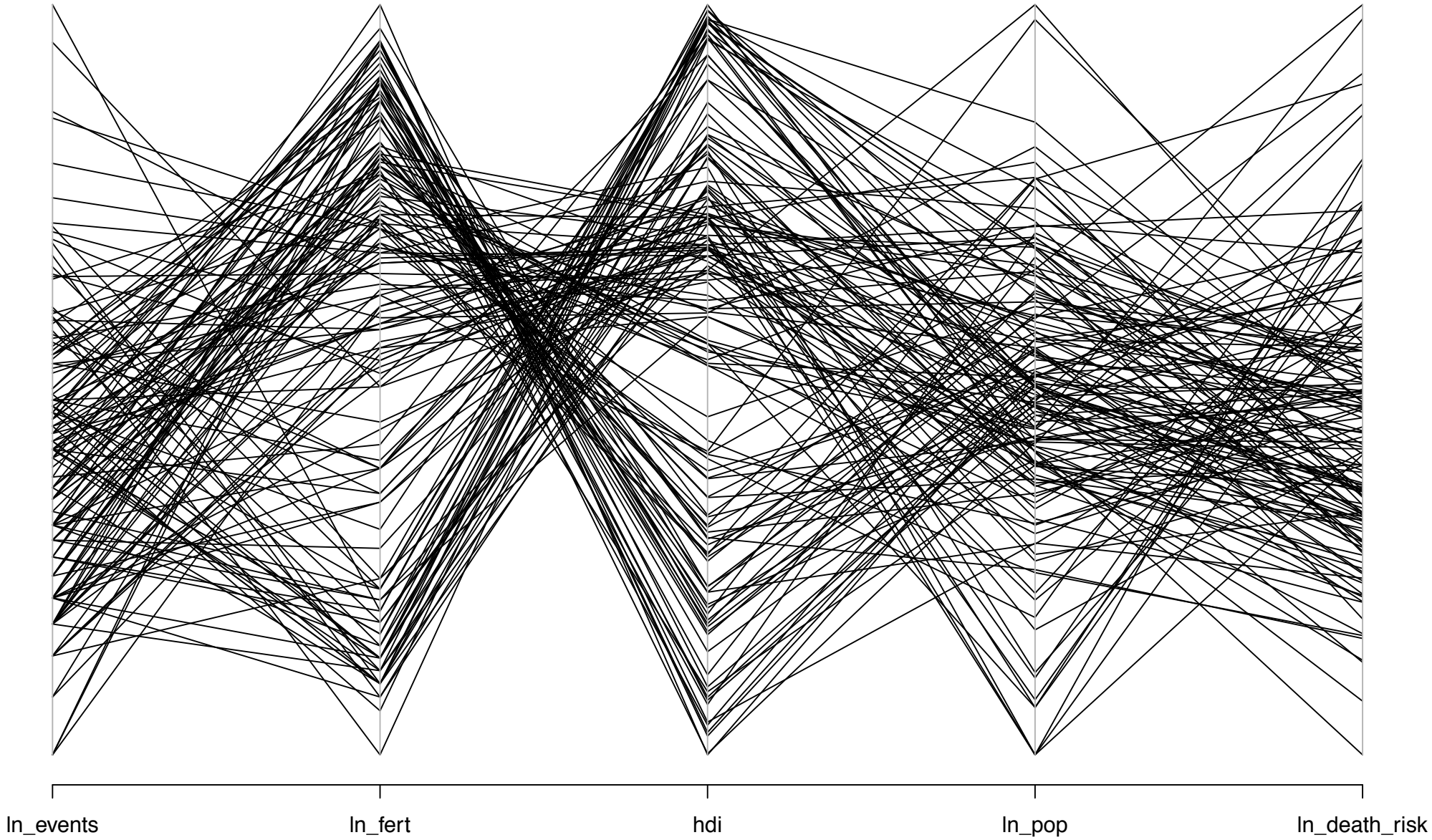
On the next slide, we plot all but the country names -- Tell me something about the associations you see...

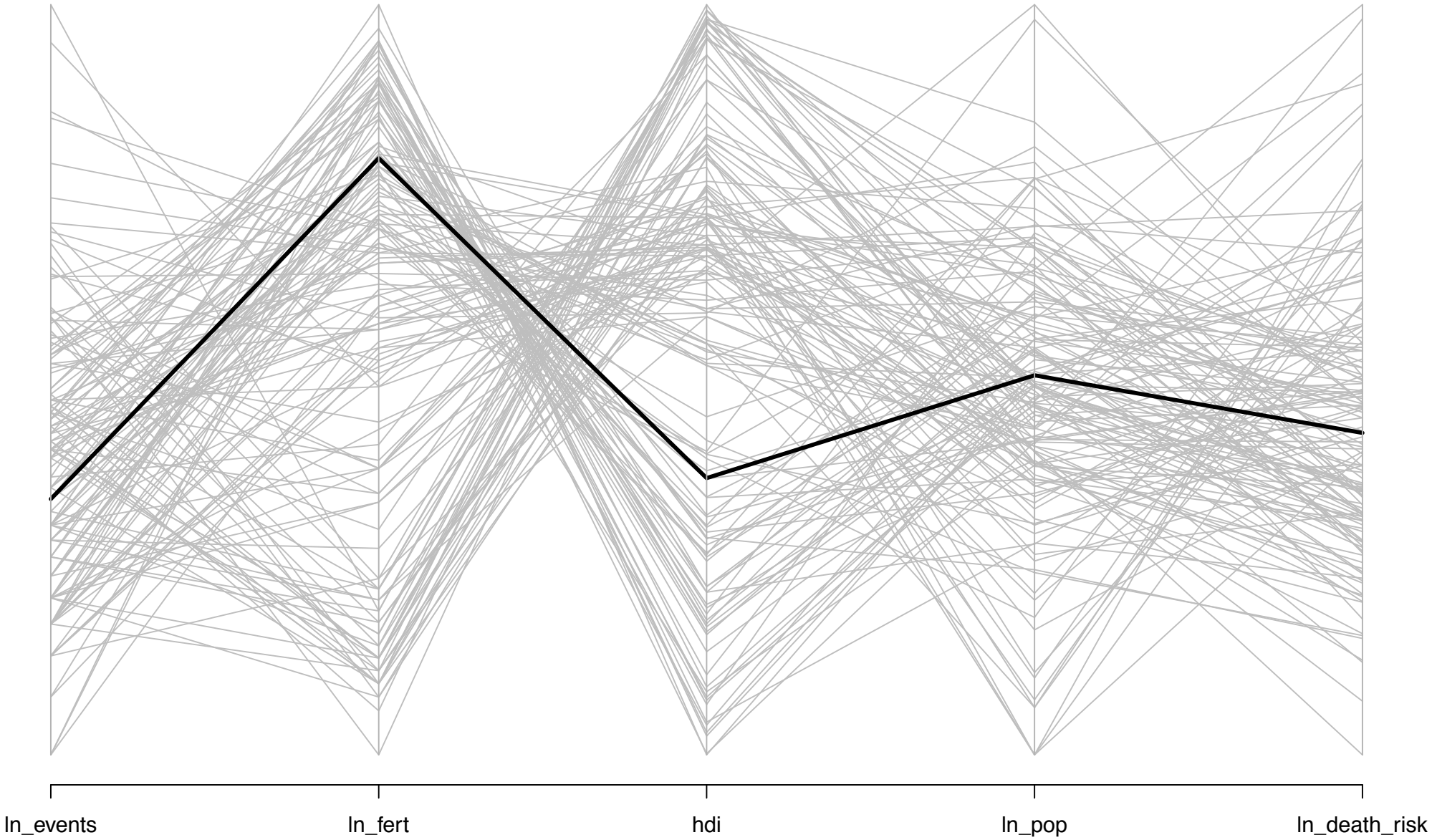

ln_events
ln_fert
hdi
ln_pop
ln_death_risk

Aside: Parallel coordinates plot

As with the multivariable median, statisticians have had decades to think about visualizing many variables at one time -- Here, however, we haven't been as successful in coming up with definitive tools

The next plot views each variable as a vertical axis and then represents the values from a single country as a broken line running between these axes


ln_events
ln_fert
hdi
ln_pop
ln_death_risk

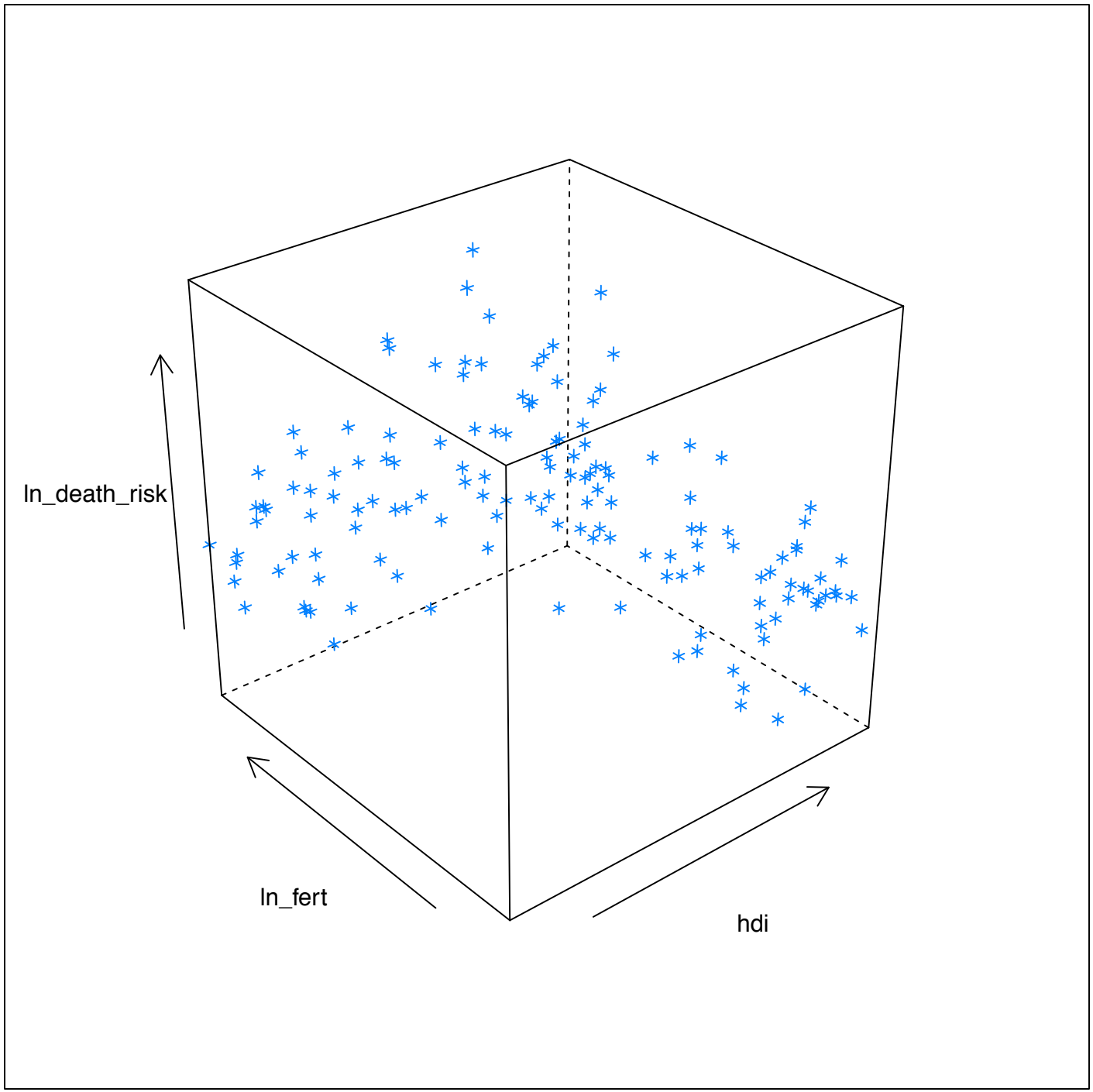

ln_events
ln_fert
hdi
ln_pop
ln_death_risk

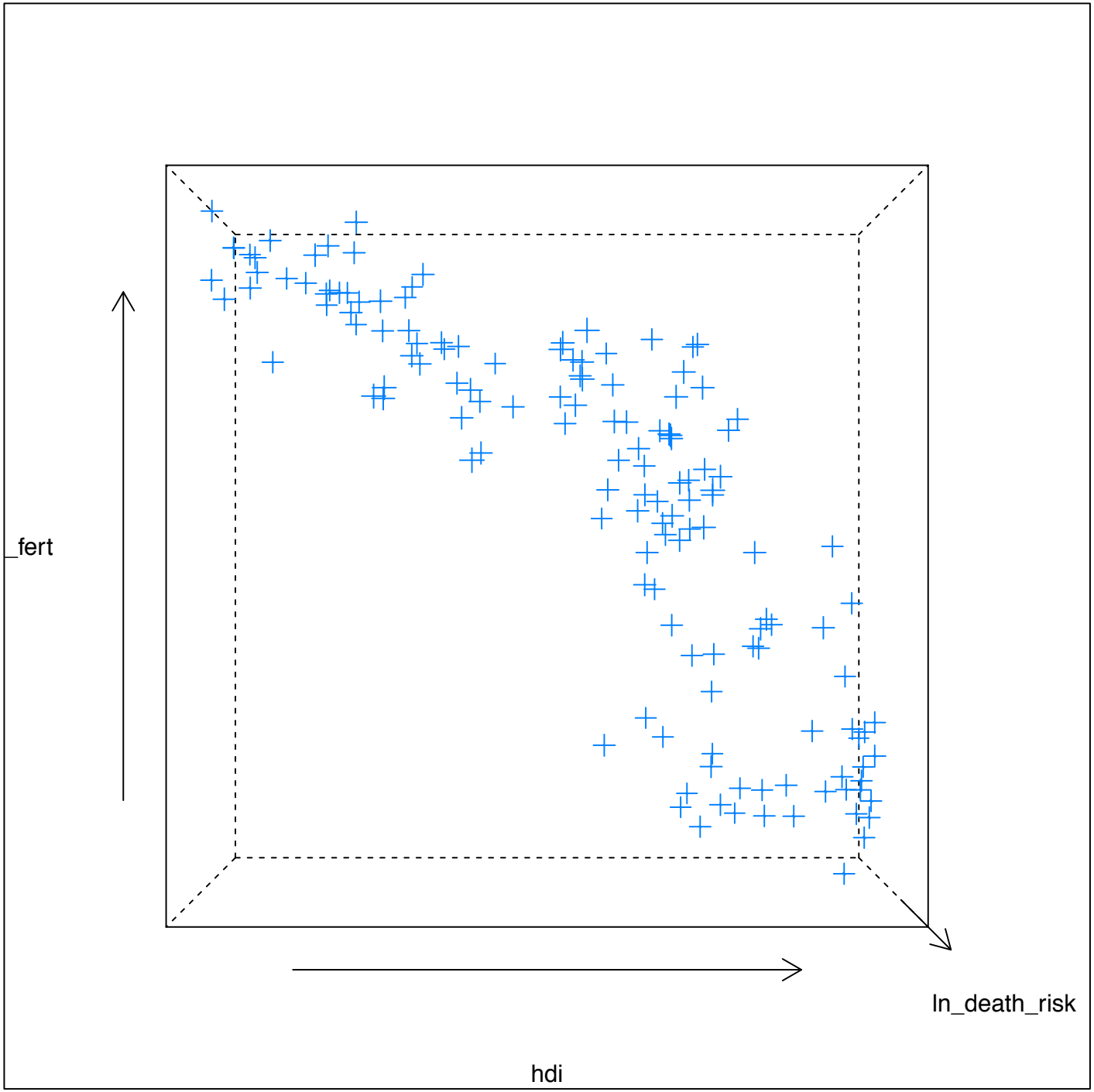
Multidimensional data

Although I'm not sure exactly the first person to do this and in what context, we can make a mapping between the rows in a data table and a point in Euclidean space -- If you think about it, this is a big conceptual leap

So in the case of the vulnerability data, we have 5 quantities recorded for each country (well, 6 if you consider its name, but we'll leave that out for now, viewing it more as an index than a measured quantity)

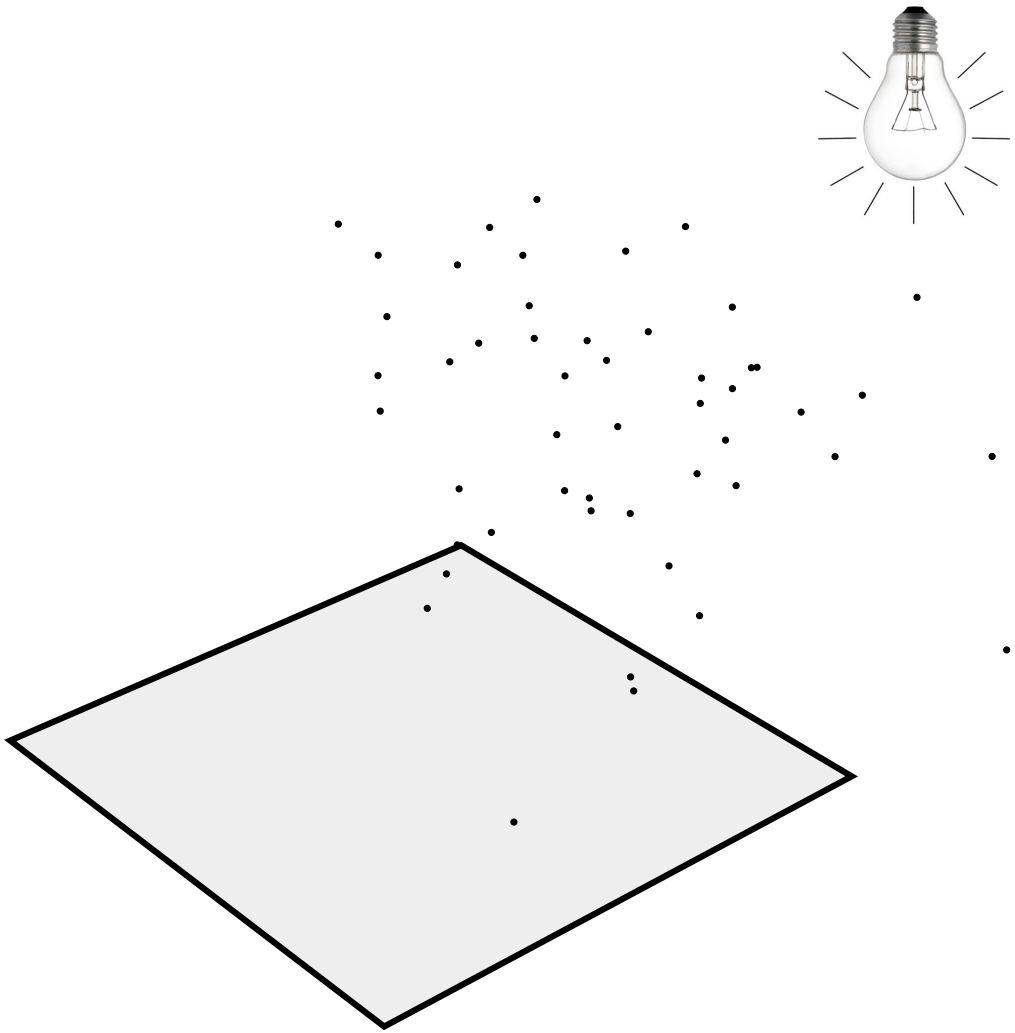
That means each row of data can be thought of as a point in 5-dimensional Euclidean space -- The scatterplot matrix representing a series of projections of the data into two dimensions


ln_death_risk
ln_fert
hdi


_fert
ln_death_risk
hdi

Multidimensional data

We can carry this idea farther and examine two-dimensional projections of our data set that are not “axis-aligned” as in the scatterplot matrix -- We can consider casting shadows of the data when viewed from different angles



Multidimensional data

For those of you who have had linear algebra (or for whom this material still feels familiar), we can make this projection idea rigorous -- As a graphical tool, projections let us turn the data over and examine it for structures that might not be obvious from axis-aligned projections

One approach to this idea is the so-called grand tour -- We select a series of random directions from which to view the data and then move smoothly between them, interpolating the motion

There is an excellent tool for doing this...

GGobi data visualization syst

www.ggobi.org

Overview | Learn | Blog | Foundation | Packages | Publications | Download | Support

GGobi

Good pictures force the unexpected upon us

News: **Hack-at-it 2010**

Download GGobi for Windows, Mac and Linux

Introduction

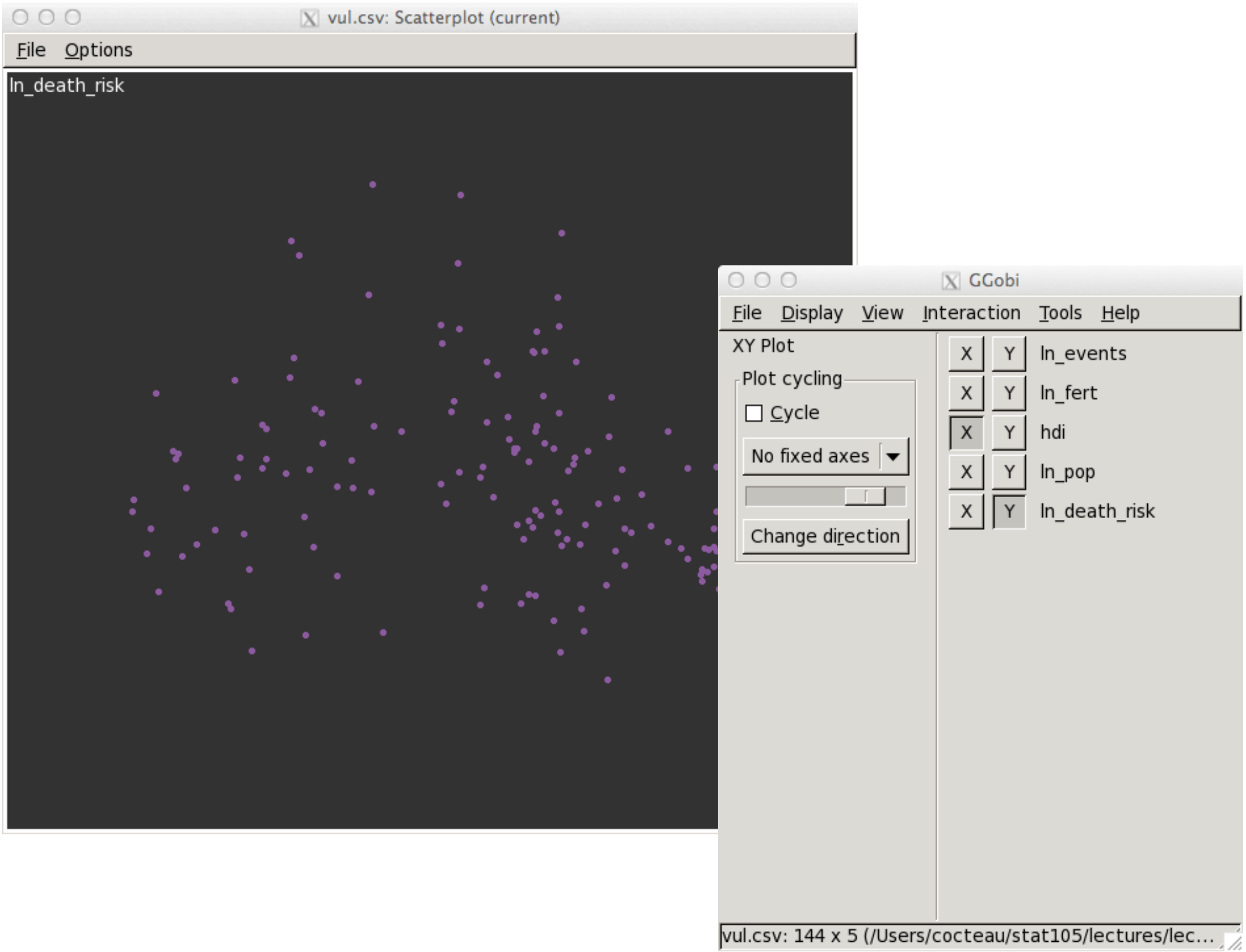
GGobi is an open source visualization program for exploring high-dimensional data. It provides highly dynamic and interactive graphics such as tours, as well as familiar graphics such as the scatterplot, barchart and parallel coordinates plots. Plots are interactive and linked with brushing and identification.

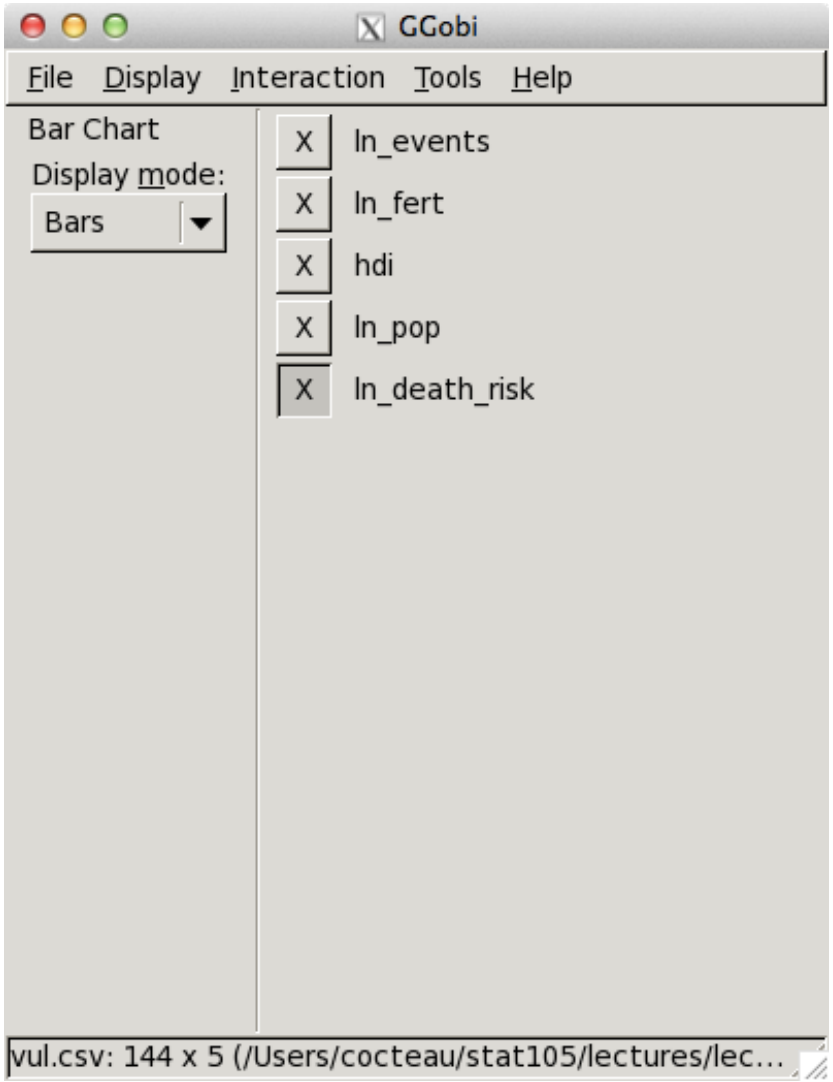
GGobi is fully documented in the GGobi book: "Interactive and Dynamic Graphics for Data Analysis".

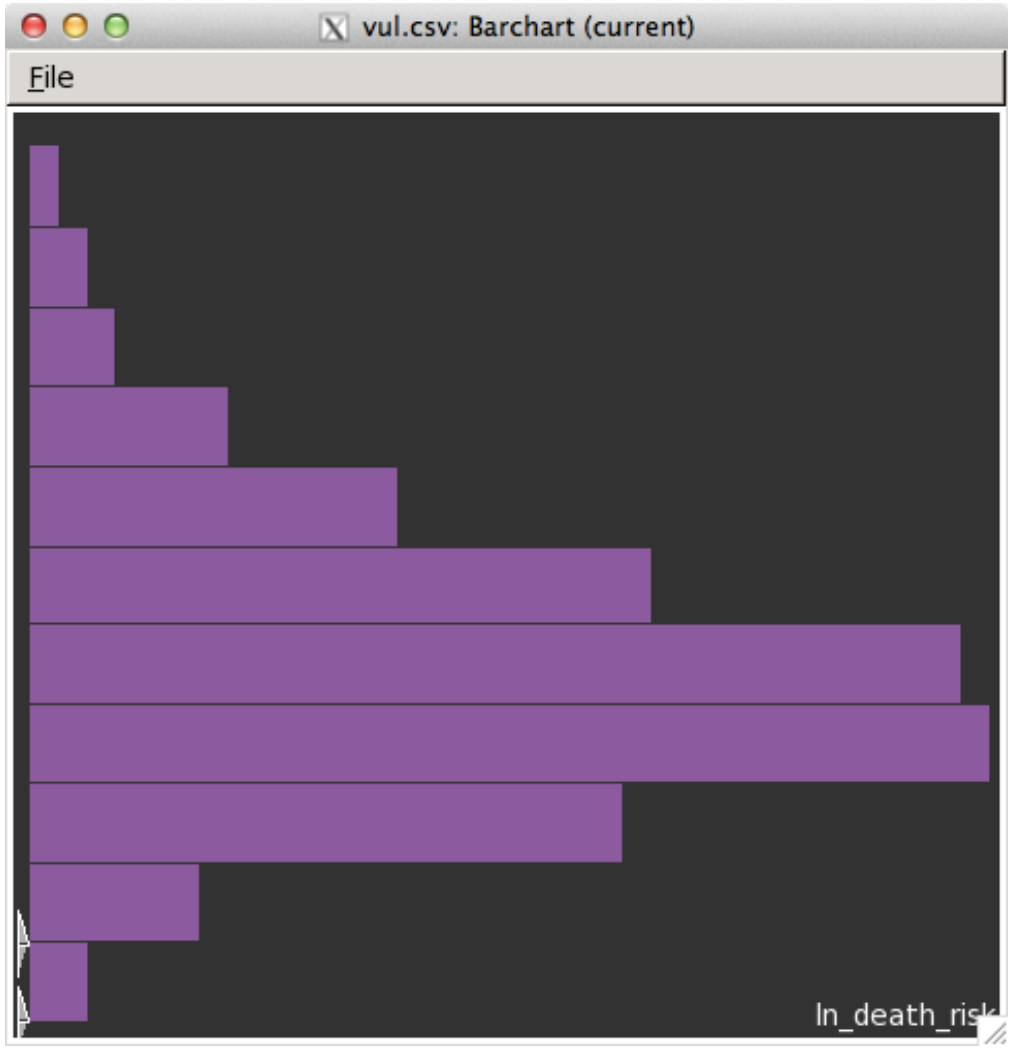
If you are interested in how GGobi came to be, you can read more about it on our history page.

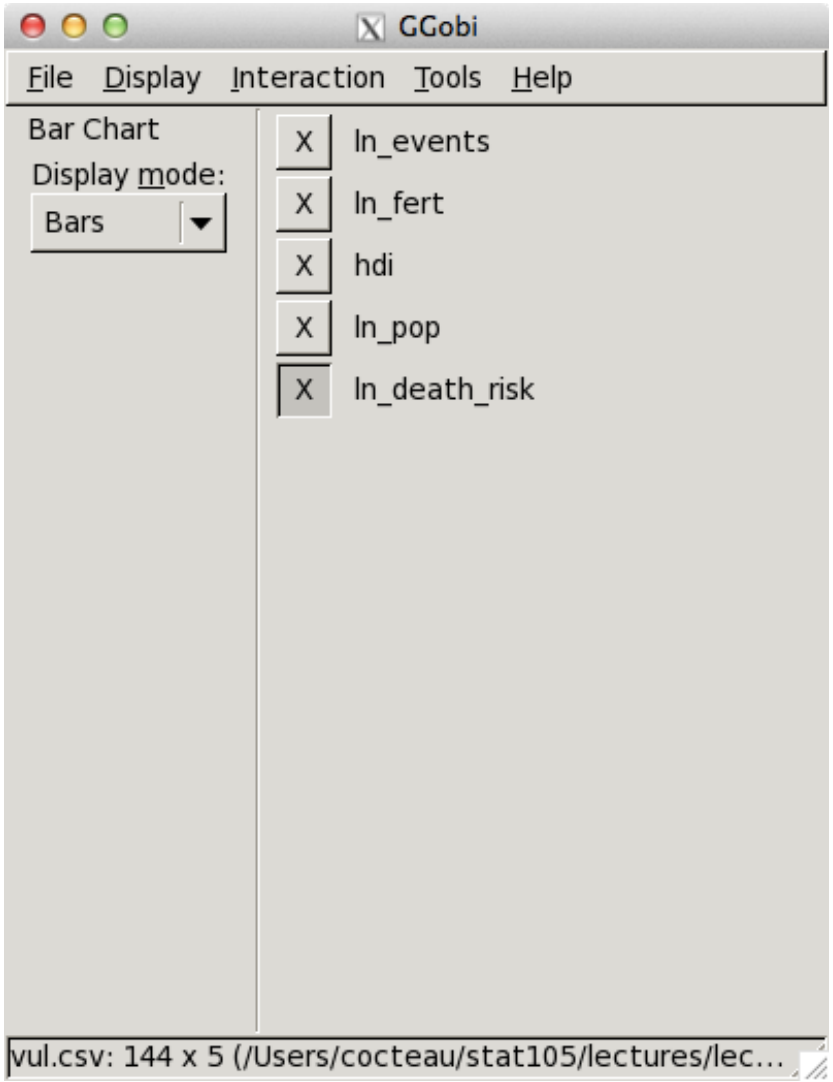
Features

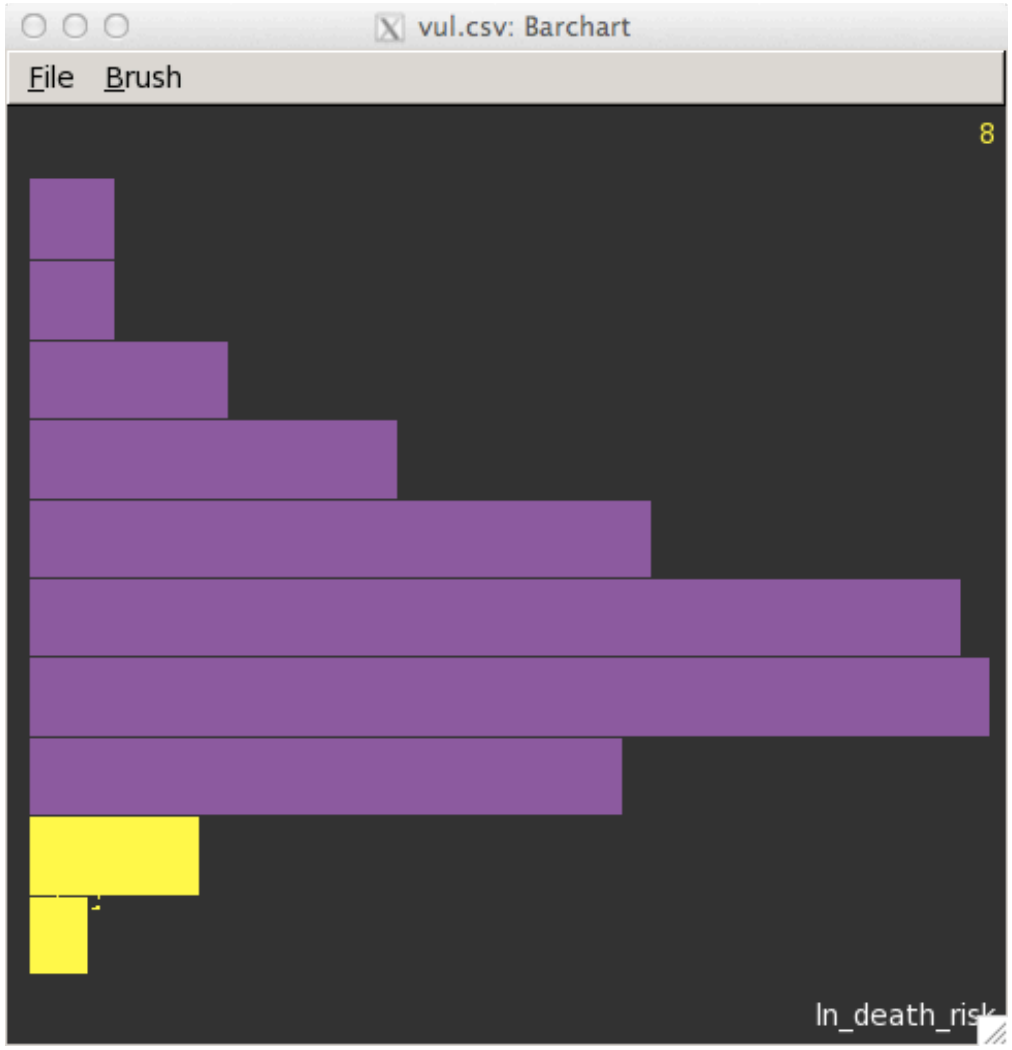
- Need to look up cases with low or high values on some variables (price, weight,...) and show how they behave in terms of other variables? → brush in linked plots.


vul.csv: Scatterplot (current)
File Options
ln_death_risk
GGobi
File Display View Interaction Tools Help
XY Plot
Plot cycling
Cycle
No fixed axes
Change direction
X Y ln_events
X Y ln_fert
X Y hdi
X Y ln_pop
X Y ln_death_risk
vul.csv: 144 x 5 (/Users/cocteau/stat105/lectures/lec…


GGobi
File Display Interaction Tools Help
Bar Chart
Display mode:
Bars
X ln_events
X ln_fert
X hdi
X ln_pop
X ln_death_risk
vul.csv: 144 x 5 (/Users/cocteau/stat105/lectures/lec...


vul.csv: Barchart (current)
File
ln_death_risk

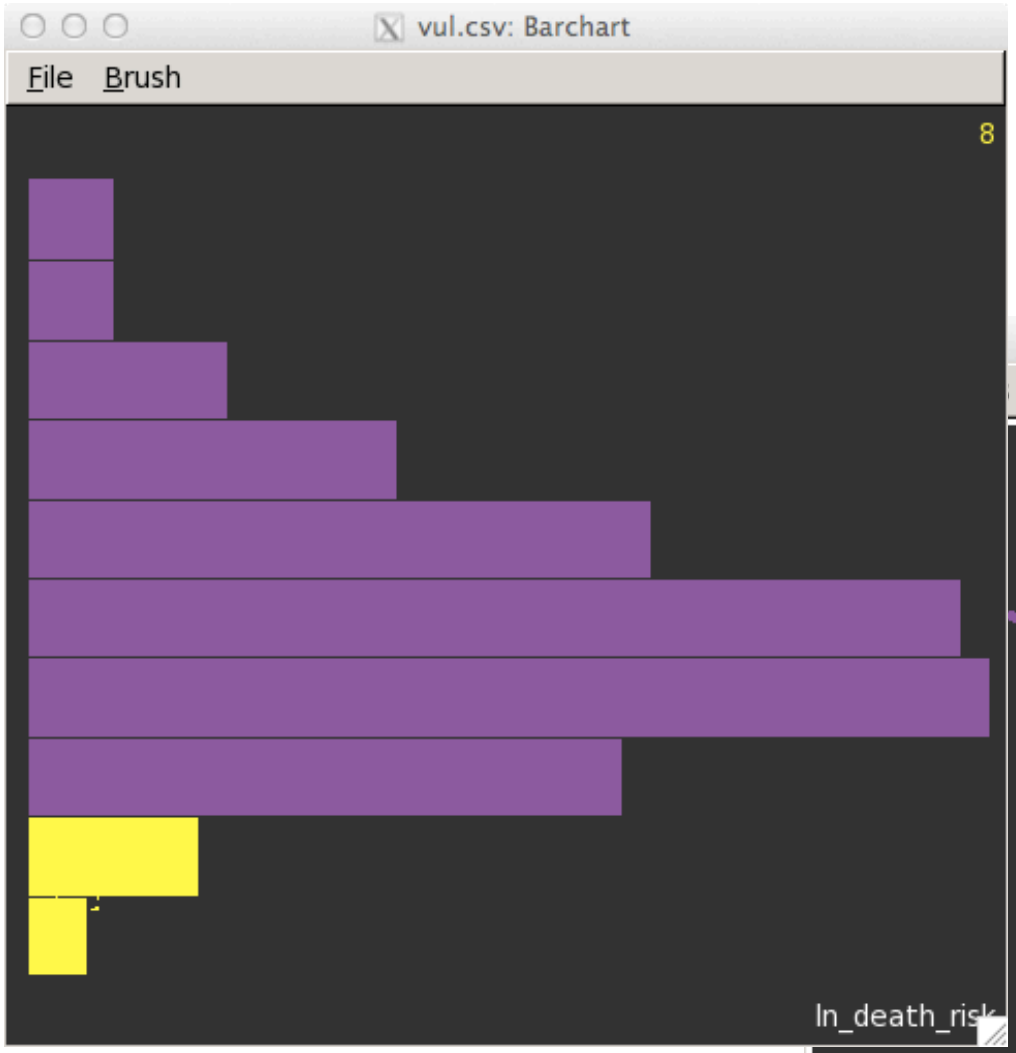

GGobi
File Display Interaction Tools Help
Bar Chart
Display mode:
Bars
X ln_events
X ln_fert
X hdi
X ln_pop
X ln_death_risk
vul.csv: 144 x 5 (/Users/cocteau/stat105/lectures/lec...

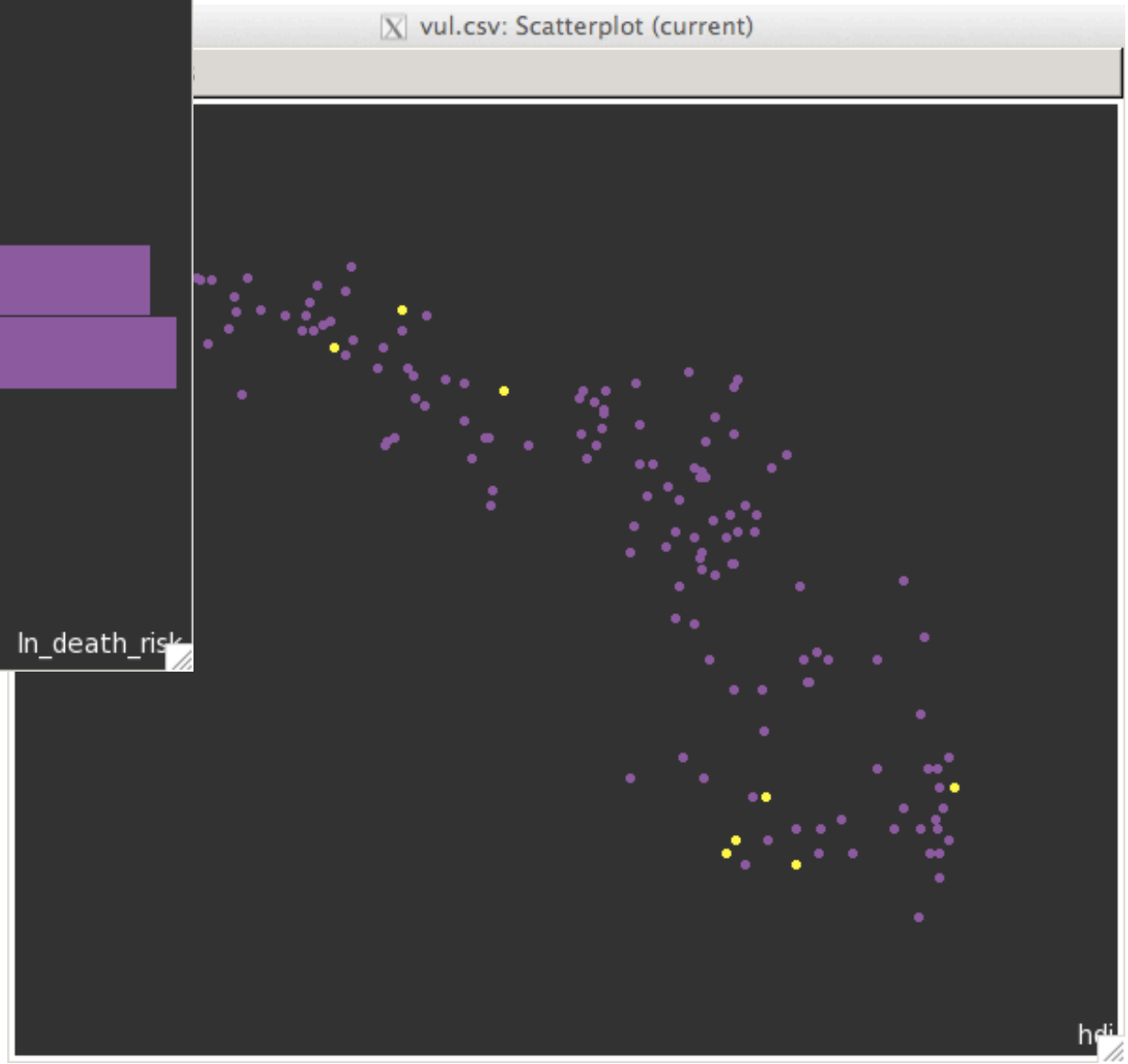

vul.csv: Barchart
File Brush
8
ln_death_risk

Linked displays

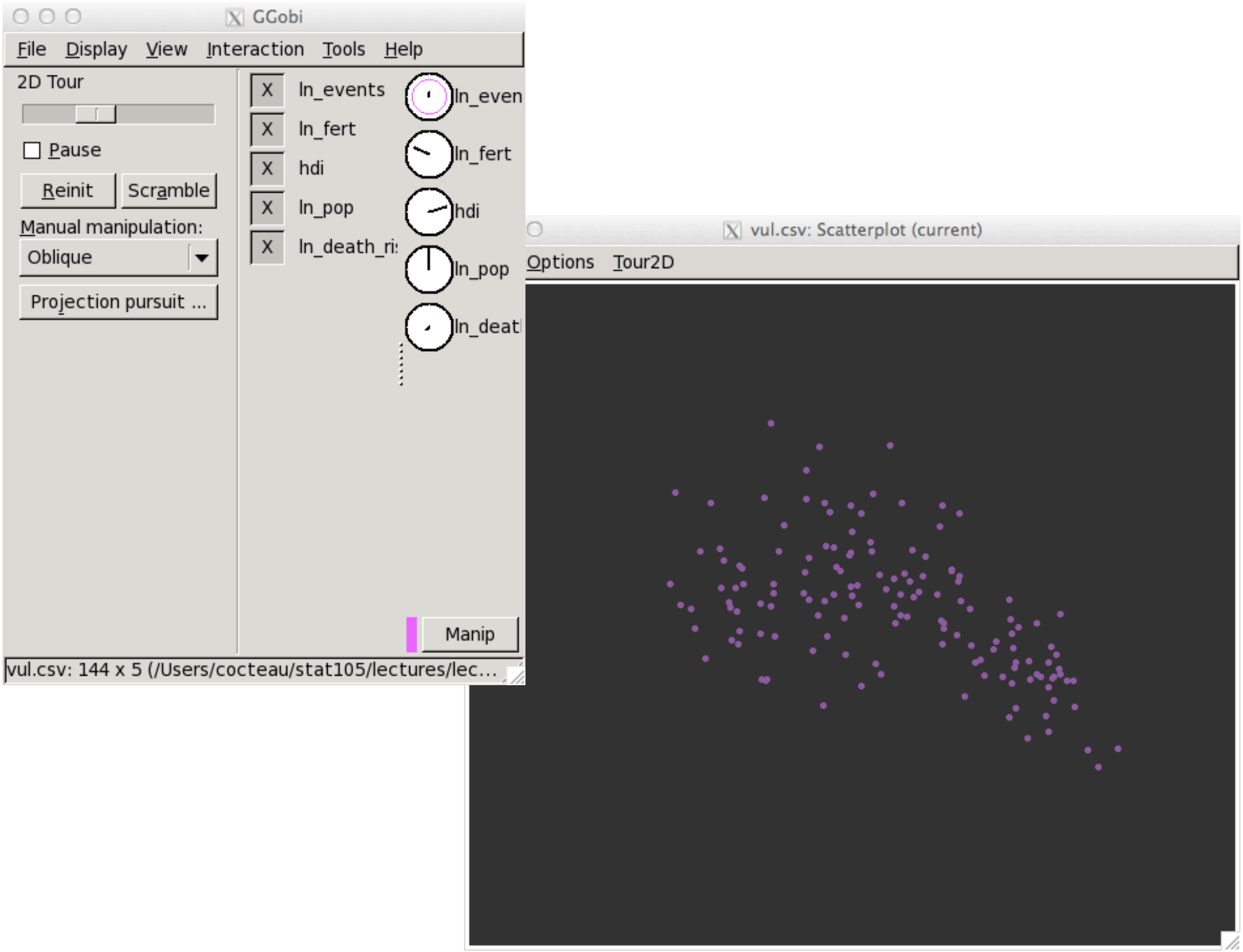
GGobi also implements linking between the displays, allowing you to highlight data in one window and examine where it appears in other plots -- On the next page we “bush” a histogram and see where the points fall on a scatterplot

Linking displays in this way is a powerful way to examine the dependence structure in your data and to examine outliers...


vul.csv: Barchart
File Brush
8
ln_death_risk


vul.csv: Scatterplot (current)
hdi

And GGobi implements the grand tour...


GGobi
File Display View Interaction Tools Help
2D Tour
Pause
Reinit
Scramble
Manual manipulation:
Oblique
Projection pursuit ...
X ln_events
X ln_fert
X hdi
X ln_pop
X ln_death_ri
ln_even
ln_fert
hdi
ln_pop
ln_deat
Manip
vul.csv: 144 x 5 (/Users/cocteau/stat105/lectures/lec...
vul.csv: Scatterplot (current)
Options Tour2D

Projections

As you watch the projections dance across the screen, we are scanning for directions that are “interesting”, providing us with a view into the clustering or grouping of data that might not be immediately evident otherwise

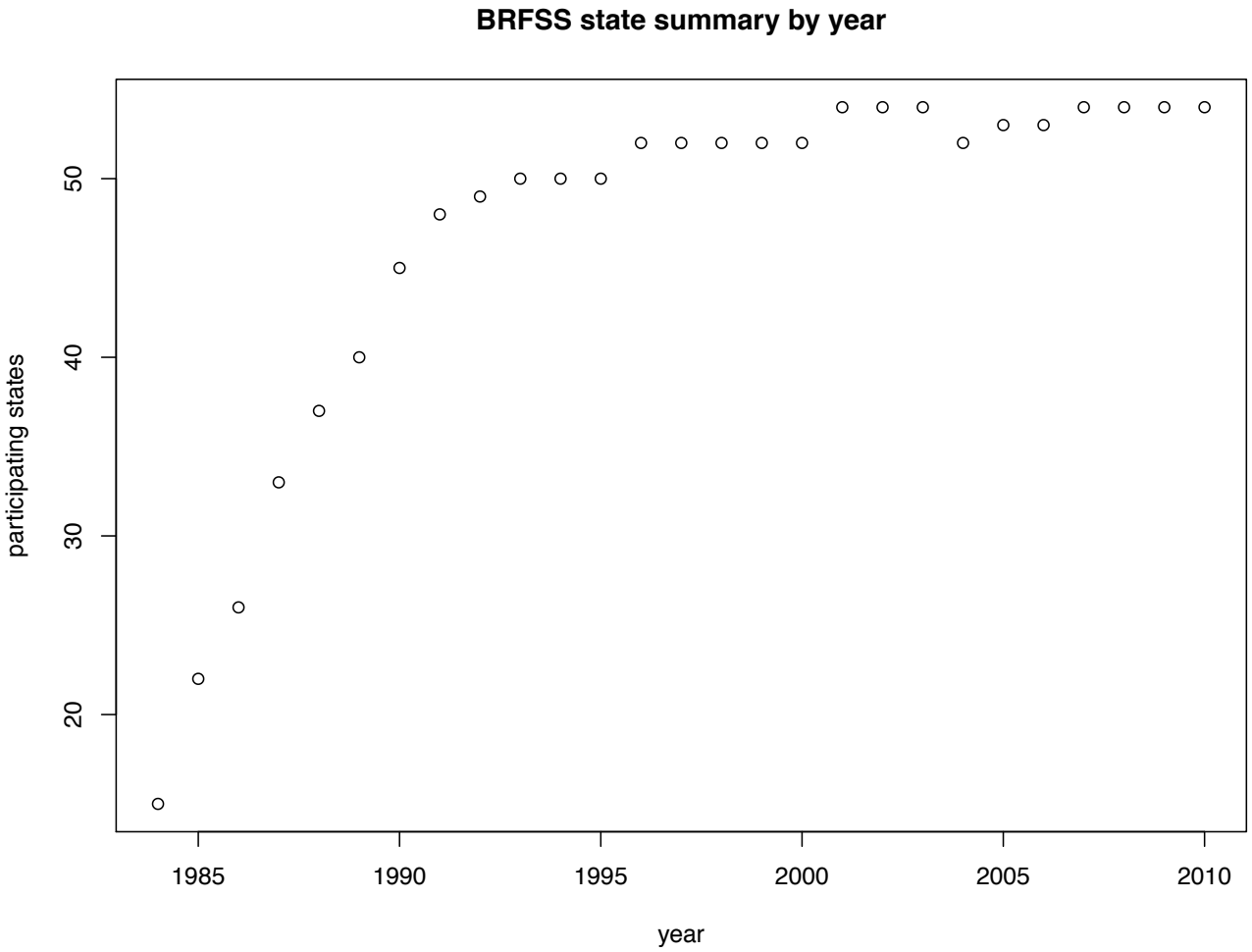
It turns out (a consequence of the Central Limit Theorem) that these projected views of the data will be “uninteresting” in that they will look like a bivariate normal distribution

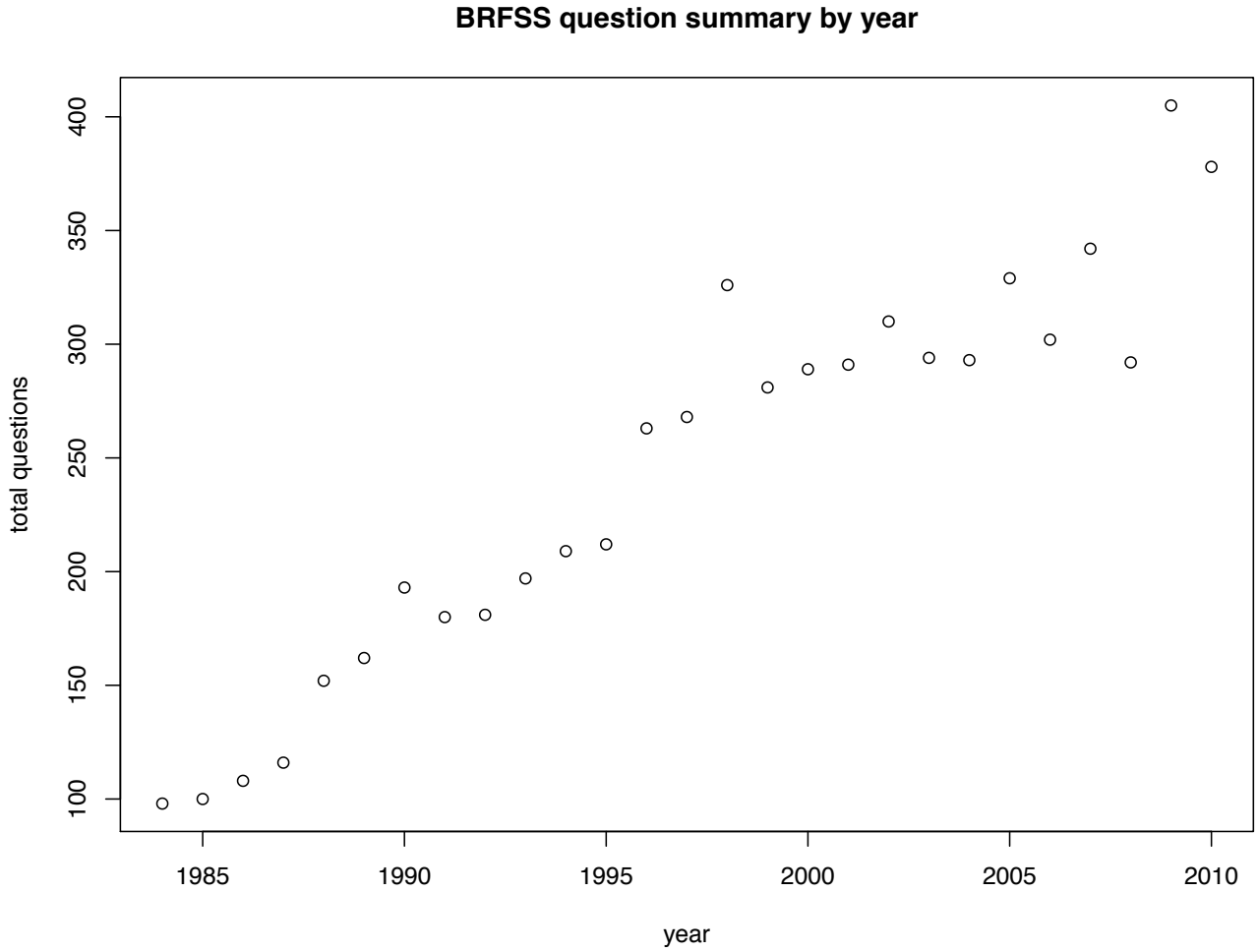
This, then, becomes one possible definition of “uninteresting” and we can score views by how dissimilar they are from this distribution -- In the late 1970s and early 1980s, this led to a statistical technique known as projection pursuit

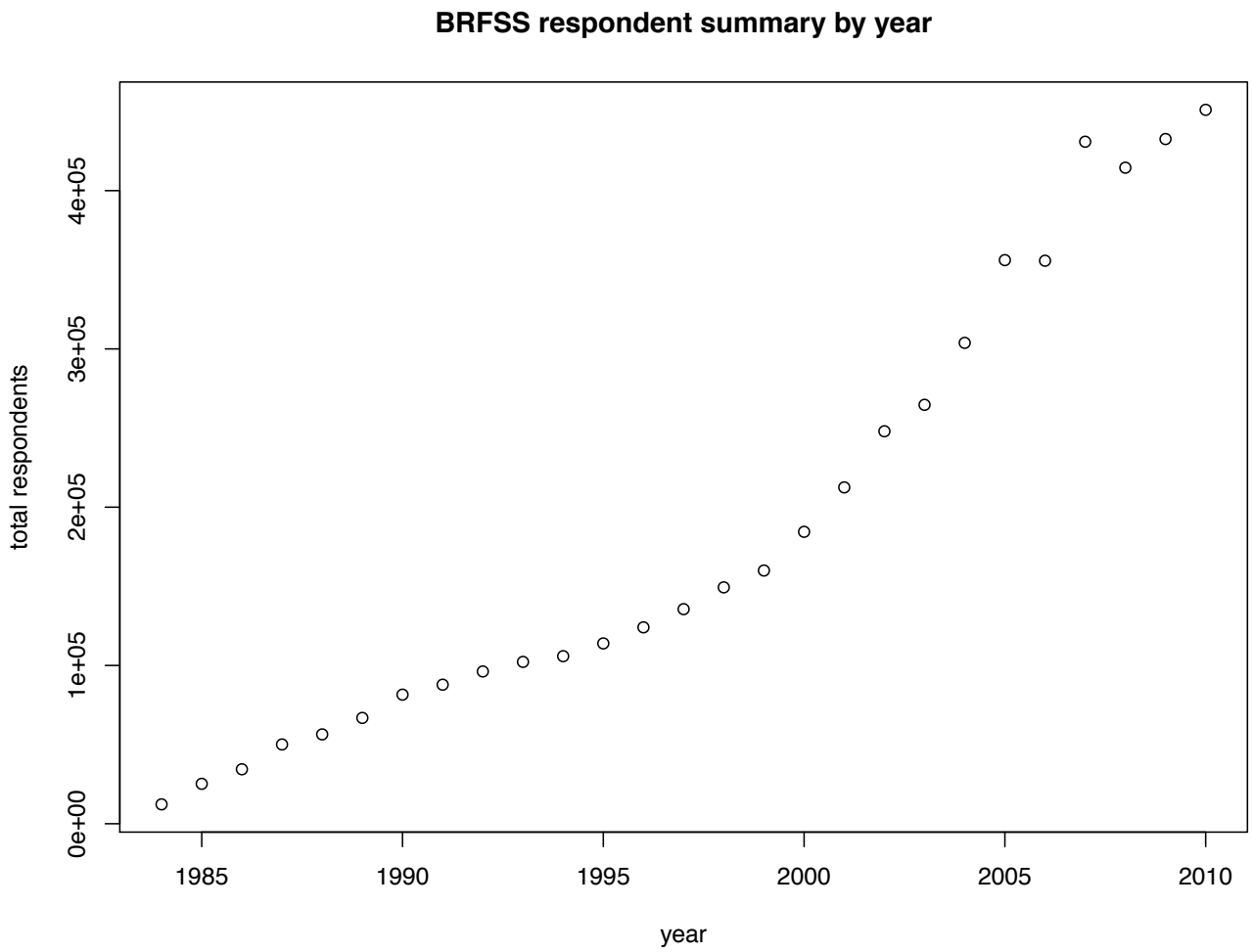
Time series

Finally, your book presents time series plots -- That is, plots of a single variable as a function of time (in quality control or process monitoring applications, these kinds of plots are exceedingly common)

The next few pages are plots of some basic statistics about the BRFSS as it has evolved in time -- These are mostly token figures and we'll be making much more extensive use of time series plots later in the quarter


BRFSS state summary by year
participating states
20
30
40
50
1985
1990
1995
2000
2005
2010
year


BRFSS question summary by year
total questions
100
150
200
250
300
350
400
1985
1990
1995
2000
2005
2010
year


BRFSS respondent summary by year
total respondents
0e+00
1e+05
2e+05
3e+05
4e+05
1985
1990
1995
2000
2005
2010
year

Finally, a prelude to your homework...

The registrar

The registrar maintains **a record of every class you take**; in addition to what class, it publishes a catalog of when classes meet and how many people were enrolled

On the next page, we present a few lines from a data file we will eventually consider in lab; it is was provided by the registrar (at a cost of $85) and contains the schedules for every student on campus last quarter*

In all, we have 162380 separate rows in this table, each corresponding to a different student and a single class with 31981 total students; What can we learn from these data? And, more importantly, how?

*Note that the identification number in this table is not your student ID, or even part of it, but a random number generated to replace your real ID

	id	subj_area_cd	cat_sort	MEET_BLDG_CD	meet_rm_sort	strt_time	end_time	DAYS_OF_WK_CD	career_cd
1	816640632	ANTHRO	0009	HAINES	00314	10:00:00	10:50:00	M	U
2	816640632	ANTHRO	0009	FOWLER	A00103B	11:00:00	12:15:00	TR	U
3	816640632	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U
4	816640632	ENGCOMP	0003	HUMANTS	A00046	09:30:00	10:45:00	TR	U
5	816640632	GEOG	0005	BUNCHE	A00170	11:00:00	12:50:00	M	U
6	816643648	MGMT	0403	GOLD	B00313	09:30:00	12:45:00	S	G
7	816643648	MGMT	0405	GOLD	B00313	14:00:00	17:15:00	S	G
8	816577472	COMM ST	0187	PUB AFF	01222	09:30:00	10:45:00	TR	U
9	816577472	COMM ST	0168	ROYCE	00362	17:00:00	19:50:00	M	U
10	816577472	COMM ST	0133	DODD	00175	10:00:00	10:50:00	MWF	U
12	806029941	EDUC	0491	KAUFMAN	00153	17:00:00	19:50:00	W	G
13	806029941	EDUC	0330D	FIELD		08:00:00	14:50:00	MTWRF	G
14	821748664	ANTHRO	0007	HAINES	00039	09:00:00	09:50:00	MWF	U
15	821748664	SPAN	0120	FOWLER	A00139	15:30:00	16:50:00	MW	U
16	821748664	SPAN	0120	HUMANTS	A00046	11:00:00	11:50:00	R	U
17	821748664	WOM STD	0107C M	HAINES	A00025	14:00:00	15:50:00	TR	U
18	821748664	ANTHRO	0007	HAINES	00350	12:00:00	12:50:00	R	U
19	820969784	ENGR	0180	BOELTER	02444	18:00:00	18:50:00	M	U
20	820969784	EL ENGR	0115AL	ENGR IV	18132	12:00:00	15:50:00	T	U
21	820969784	EL ENGR	0115A	ROLFE	01200	08:00:00	09:50:00	MW	U
22	820969784	EL ENGR	0115A	BOELTER	05280	09:00:00	09:50:00	F	U
23	820969784	STATS	0105	PAB	02434	15:00:00	15:50:00	R	U
24	820969784	STATS	0105	FRANZ	02258A	12:00:00	12:50:00	MWF	U
25	820969784	ENGR	0180	BOELTER	02444	16:00:00	17:50:00	MW	U
26	821030697	GEOG	0005	HAINES	00039	13:00:00	14:15:00	MW	U

The registrar

At the end of your career here, you receive a transcript, an aggregate of all the data the registrar has on you (but printed in some reduced format) -- I'd like you to consider these data and what we might learn from it when it is aggregated across all the students at UCLA

"Learning" here might involve recoding variables, sometimes reshaping the data set entirely, to give us a different view (changing, say the unit of observation which here is a "student in a class" or an "enrollment event")

This is part of your first homework assignment and is due by Friday -- I will create data sets based on your suggestions and you are going to apply your data summary and visualization skills to tell me something!