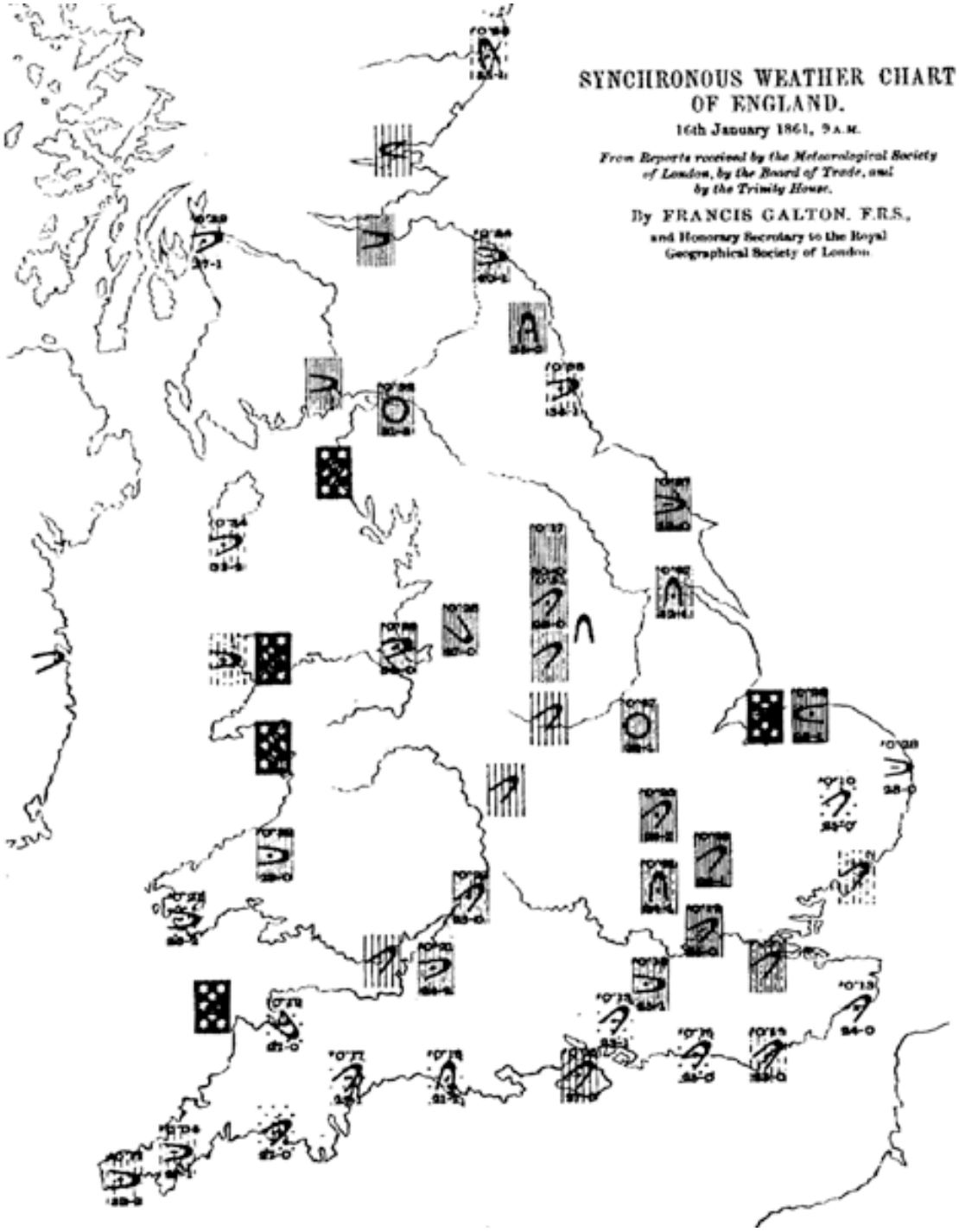
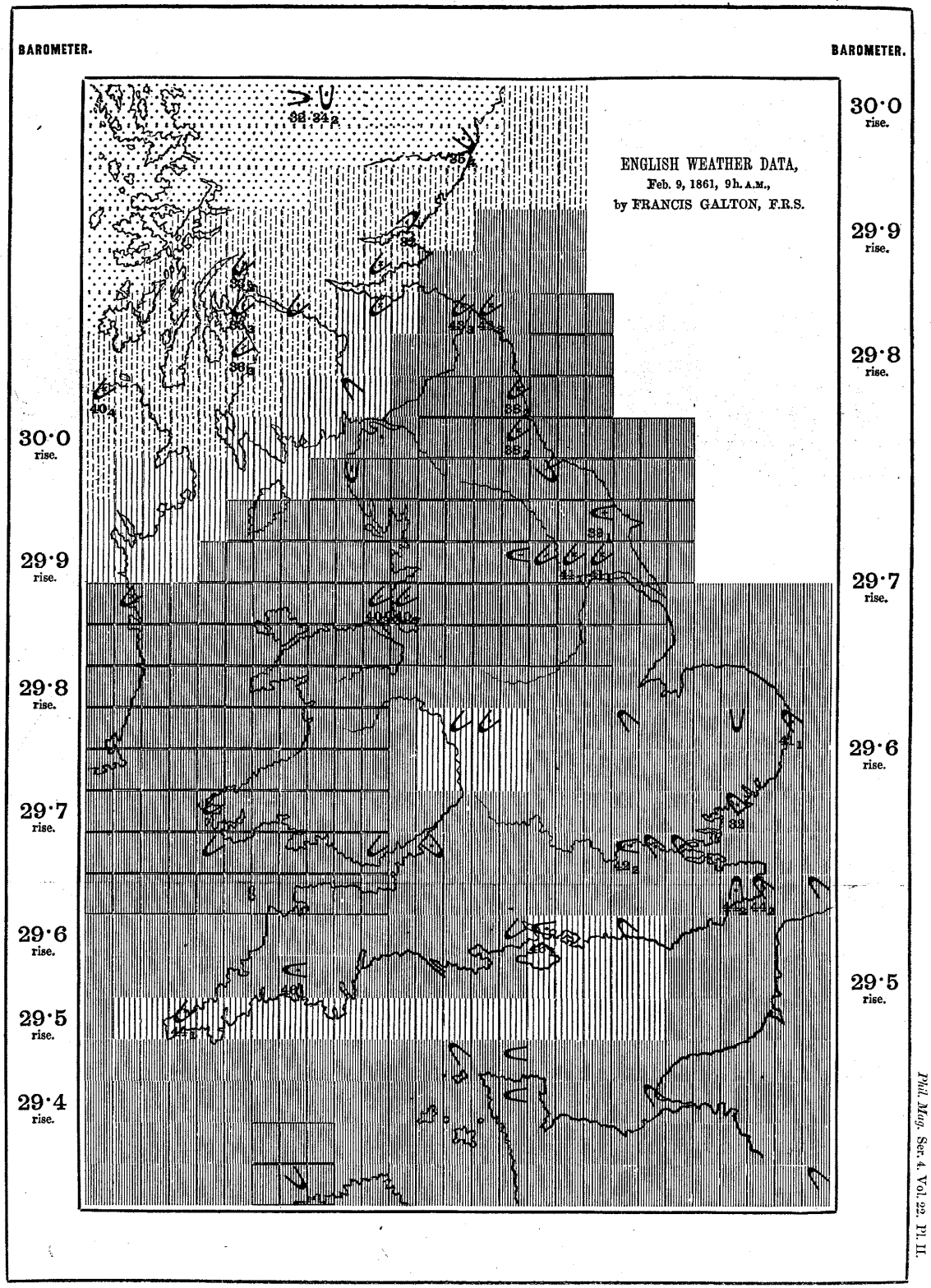


SYNCHRONOUS WEATHER CHART OF ENGLAND.

16th January 1861, 9 A.M.

From Reports received by the Meteorological Society of London, by the Board of Trade, and by the Trinity House.

By FRANCIS GALTON, F.R.S.,
and Honorary Secretary to the Royal Geographical Society of London



Phil. Mag. Ser. 4. Vol. 22. Pl. II.

METEOROGRAPHICA,

OR

METHODS OF MAPPING THE WEATHER;

ILLUSTRATED BY UPWARDS OF 600 PRINTED AND LITHOGRAPHED DIAGRAMS

REFERRING TO

THE WEATHER OF A LARGE PART OF EUROPE,

During the Month of December 1861.

By FRANCIS GALTON, F.R.S.

London and Cambridge:
MACMILLAN AND Co.

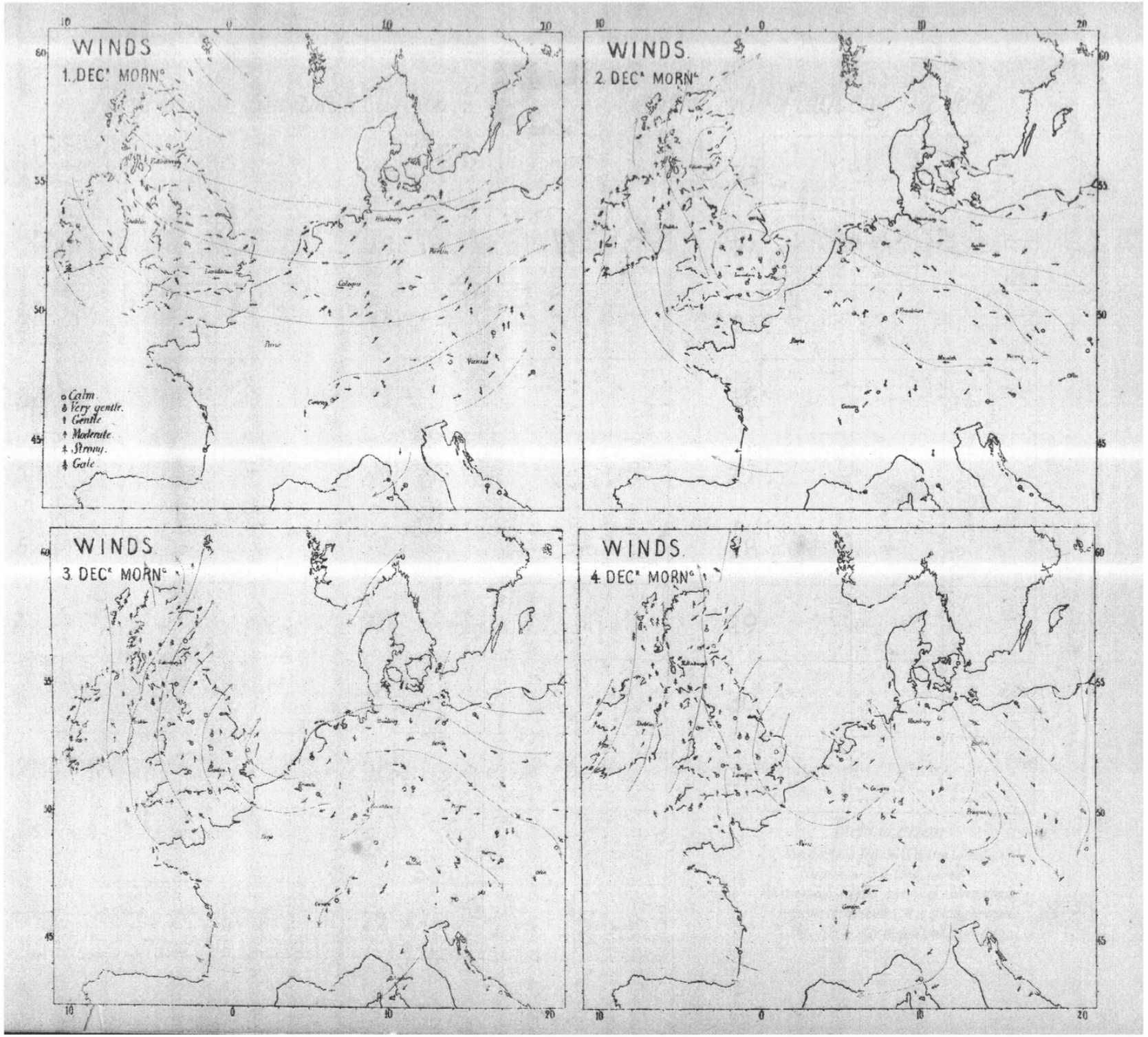
MDCCCLXIII.

$x°$ Fahrenheit $= \frac{5}{9}(x° - 32°)$ Centigrade $= \frac{4}{9}(x° - 32°)$ Reaumur.

Fahrenheit.	Centigrade.	Reaumur.	Fahrenheit.	Centigrade.	Reaumur.
100°	37°·8	30°·2	44°	6°·7	5°·3
99	37·2	29·8	43	6·1	4·9
98	36·7	29·3	42	5·6	4·4
97	36·1	28·9	41	5·0	4·0
96	35·6	28·4	40	4·4	3·6
95	35·0	28·0	39	3·9	3·1
94	34·4	27·6	38	3·3	2·7
93	33·9	27·1	37	2·8	2·2
92	33·3	26·7	36	2·2	1·8
91	32·8	26·2	35	1·7	1·3
90	32·2	25·8	34	1·1	0·9
89	31·7	25·3	33	0·6	0·4
88	31·1	24·9	32	0·0	0·0
87	30·6	24·4	31	− 0·6	− 0·4
86	30·0	24·0	30	− 1·1	− 0·9
85	29·4	23·6	29	− 1·7	− 1·3
84	28·9	23·1	28	− 2·2	− 1·8
83	28·3	22·7	27	− 2·8	− 2·2
82	27·8	22·2	26	− 3·3	− 2·7
81	27·2	21·8	25	− 3·9	− 3·1
80	26·7	21·3	24	− 4·4	− 3·6
79	26·1	20·9	23	− 5·0	− 4·0
78	25·6	20·4	22	− 5·6	− 4·4
77	25·0	20·0	21	− 6·1	− 4·9
76	24·4	19·6	20	− 6·7	− 5·3
75	23·9	19·1	19	− 7·2	− 5·8
74	23·3	18·7	18	− 7·8	− 6·2
73	22·8	18·2	17	− 8·3	− 6·7
72	22·2	17·8	16	− 8·9	− 7·1
71	21·7	17·3	15	− 9·4	− 7·6
70	21·1	16·9	14	−10·0	− 8·0
69	20·6	16·4	13	−10·6	− 8·4
68	20·0	16·0	12	−11·1	− 8·9
67	19·4	15·6	11	−11·7	− 9·3
66	18·9	15·1	10	−12·2	− 9·8
65	18·3	14·7	9	−12·8	−10·2
64	17·8	14·2	8	−13·3	−10·7
63	17·2	13·8	7	−13·9	−11·1
62	16·7	13·3	6	−14·4	−11·6
61	16·1	12·9	5	−15·0	−12·0
60	15·6	12·4	4	−15·6	−12·4
59	15·0	12·0	3	−16·1	−12·9
58	14·4	11·6	2	−16·7	−13·3
57	13·9	11·1	1	−17·2	−13·8
56	13·3	10·7	0	−17·8	−14·2
55	12·8	10·2	− 1	−18·3	−14·7
54	12·2	9·8	− 2	−18·9	−15·1
53	11·7	9·3	− 3	−19·4	−15·6
52	11·1	8·9	− 4	−20·0	−16·0
51	10·6	8·4	− 5	−20·6	−16·0
50	10·0	8·0	− 6	−21·1	−16·9
49	9·4	7·6	− 7	−21·7	−17·3
48	8·9	7·1	− 8	−23·2	−17·8
47	8·3	6·7	− 9	−22·8	−18·2
46	7·8	6·2	−10	−23·3	−18·7
45	7·2	5·8			

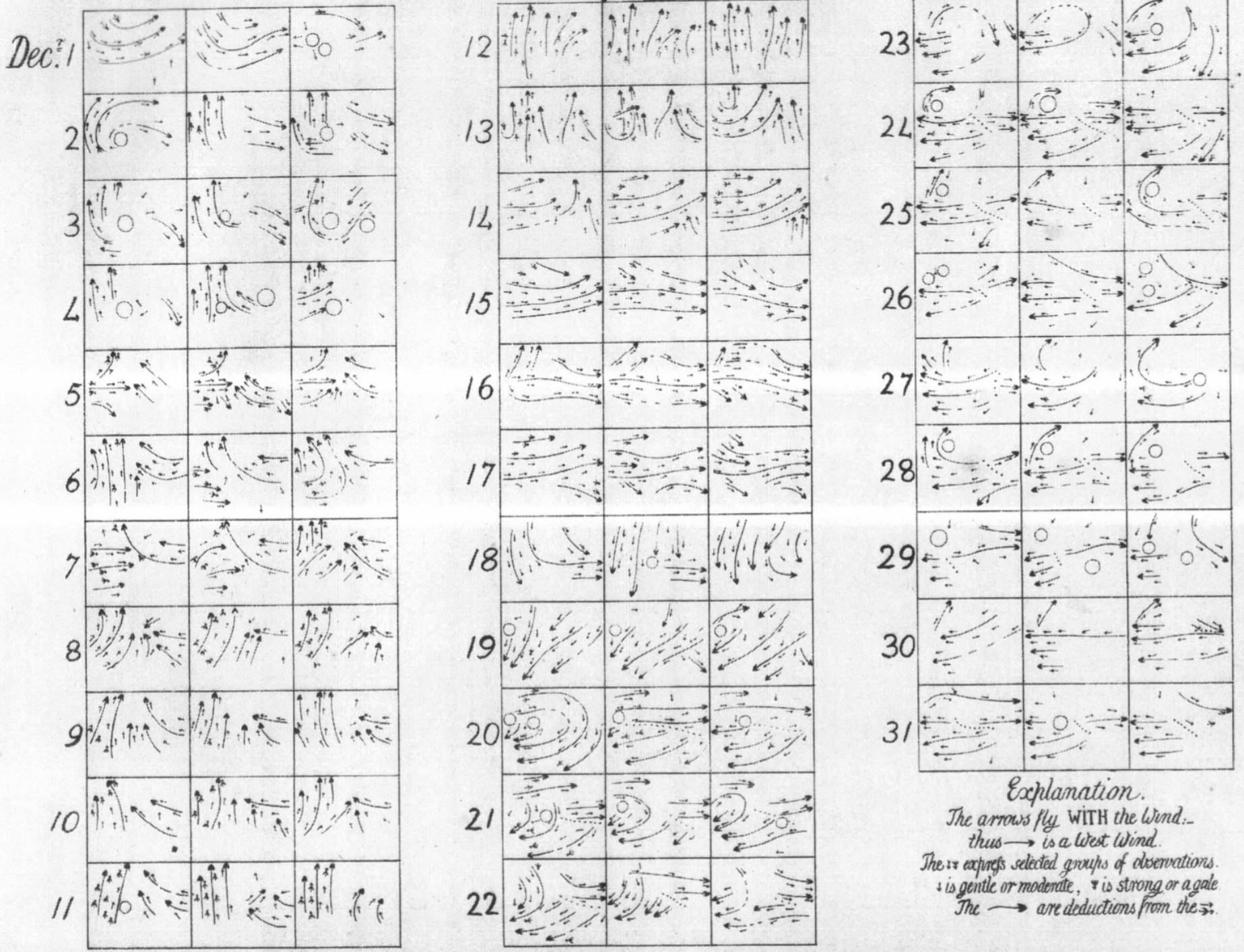
1 Inch = 25·3995 Millimètres
= 11·2595 Paris Lines
= 20 Russian Half Lines.

Inches and Tenths.	Millimètres.	Paris Lines.	Russian Half Lines.
28·0	711	315	560
1	714	316	562
2	716	318	564
3	719	319	566
4	721	320	568
5	724	321	570
6	726	322	572
7	729	323	574
8	732	324	576
9	734	325	578
29·0	737	327	580
1	739	328	582
2	742	329	584
3	744	330	586
4	747	331	588
5	749	332	590
6	752	333	592
7	754	334	594
8	757	336	596
9	759	337	598
30·0	762	338	600
1	765	339	602
2	767	340	604
3	770	341	606
4	772	342	608
5	775	343	610
6	777	345	612
7	780	346	614
8	782	347	616
9	785	348	618


WINDS
1. DEC^r MORN^g
Calm.
Very gentle.
Gentle
Moderate
Strong.
Gale
WINDS.
2. DEC^r MORN^g
WINDS
3. DEC^r MORN^g
WINDS.
4. DEC^r MORN^g

The Winds of Britain & Europe,

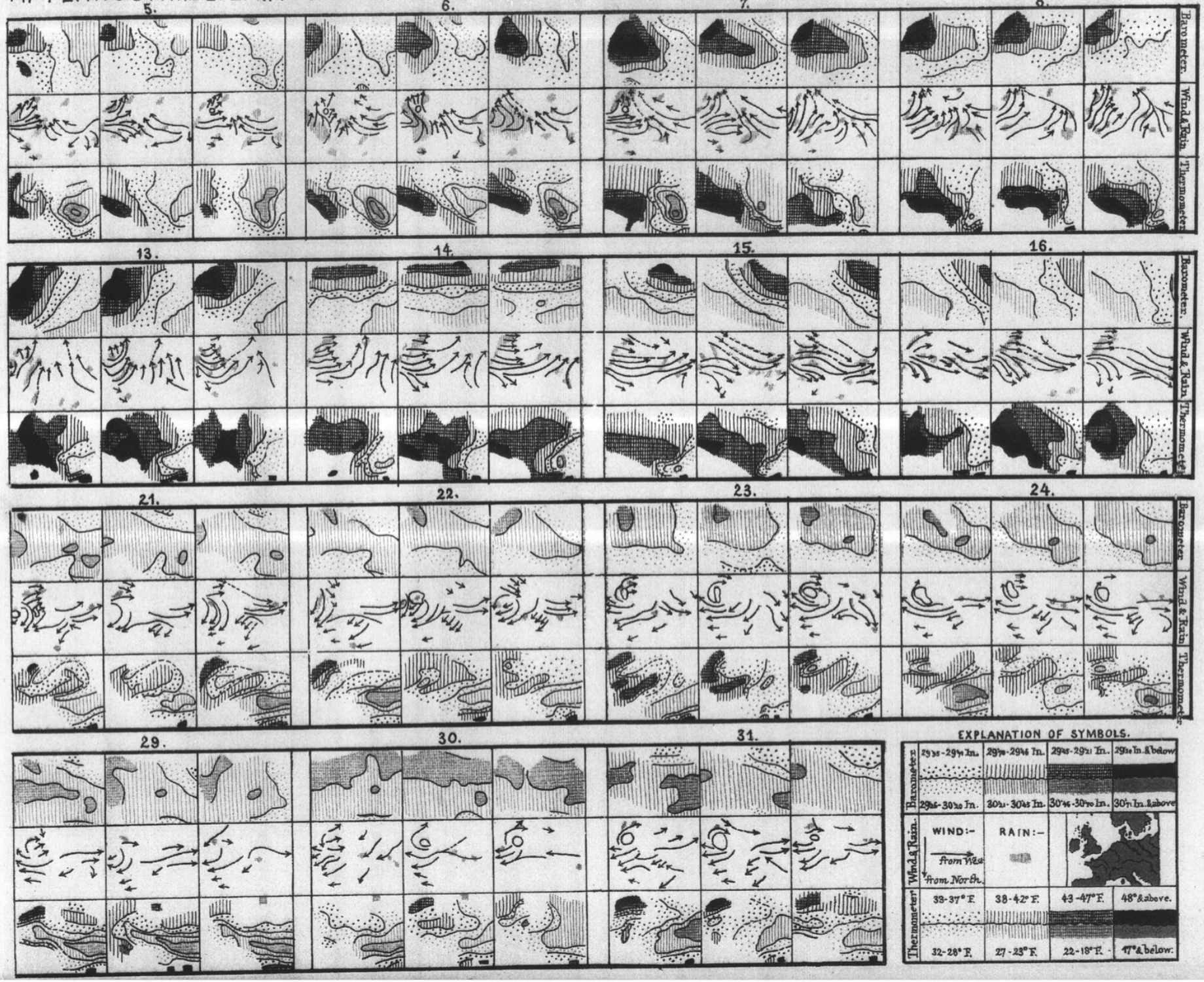
Morn.g, Aft.n & Ev.g of each day Dec.r 1861.



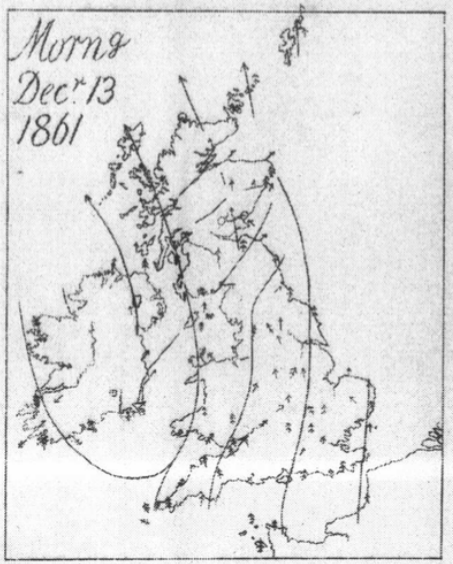
Explanation.

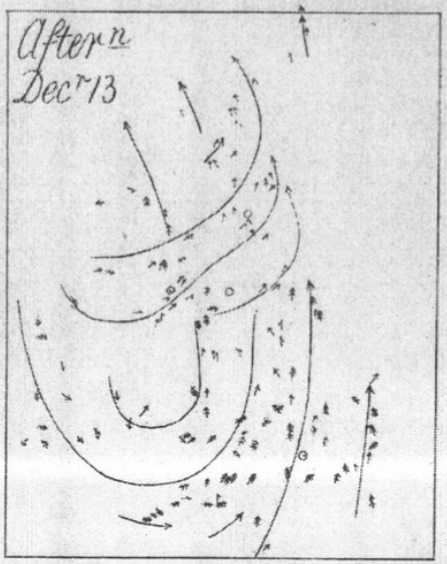
The arrows fly WITH the Wind:-
thus ⟶ is a West Wind.
The ⁝ expreſs selected groups of observations.
↓ is gentle or moderate, ▼ is strong or a gale
The ⟶ are deductions from the ⁝.

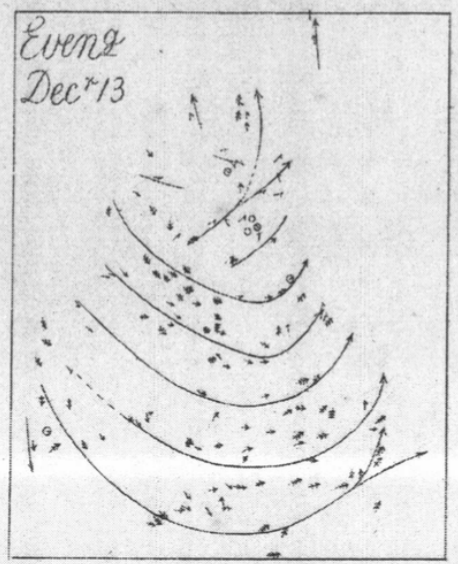
AFTERNOON AND EVENING ON EACH DAY DURING DECEMBER, 1861.

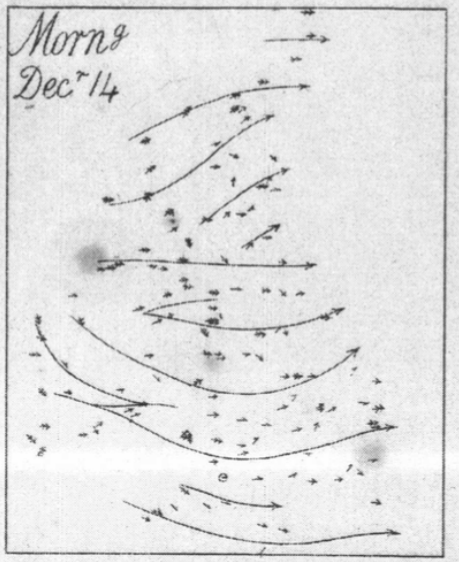


Change from a South to a West Gale.

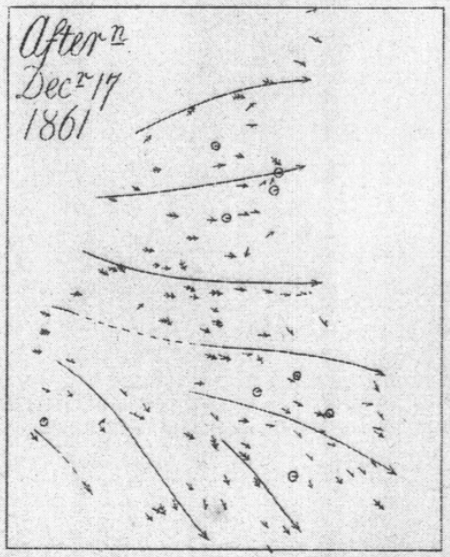


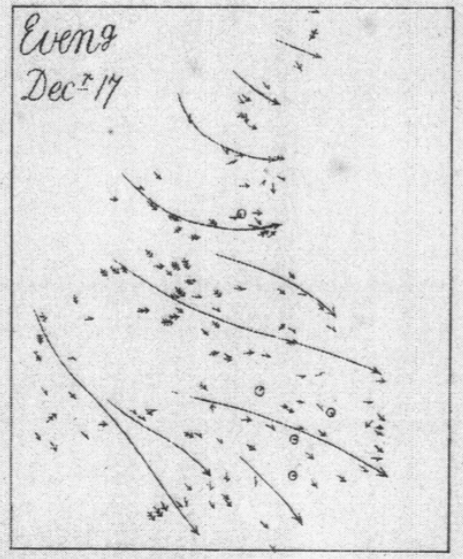


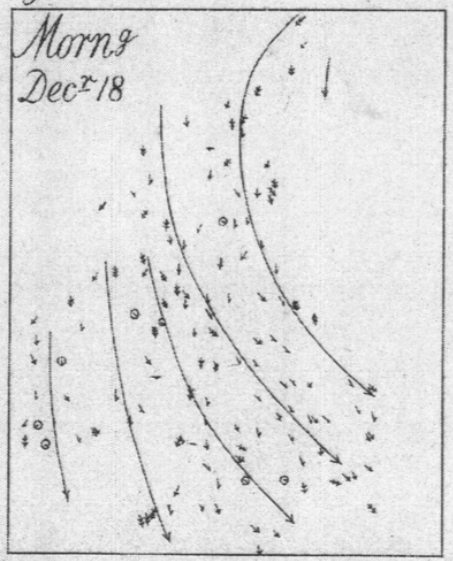




Change from a West to a North Gale







SYMBOLS FOR FORCE OF WIND

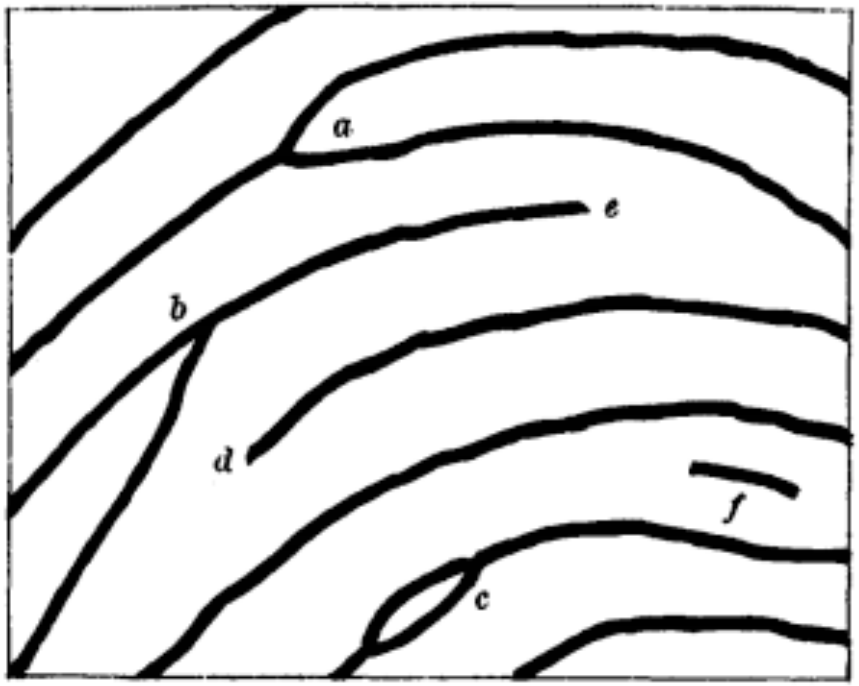
very gentle ①
gentle ↾
moderate ↓
strong ↡
gale ⇊

Fingerprints

The use of fingerprints “as a device for personal identification” were not in wide use before the late 1800s

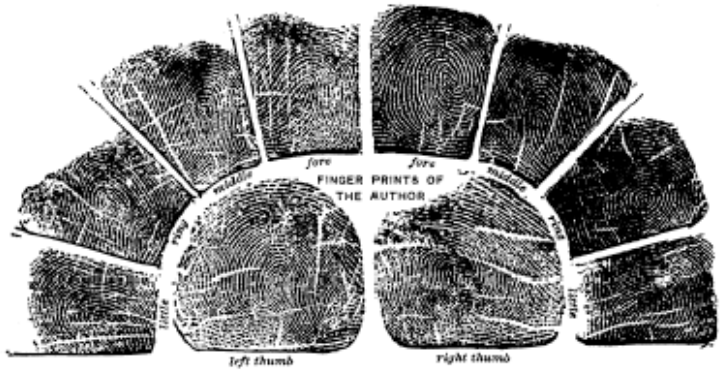
In 1890-95 Francis Galton (applied a scientific approach to assessing the reliability of fingerprints as a method for “criminal identification”

Galton emphasized that it wasn’t the overall features that were useful for classification, but rather “the minutia of the prints (tiny islets and forks in the ridges)”



“Galton and identification by fingerprints”, by Stephen Stigler in Genetics July 1, 1995 vol. 140 no. 3857-860

FINGER PRINTS



BY

FRANCIS GALTON, F.R.S., ETC.

London
MACMILLAN AND CO.
AND NEW YORK
1892

Galton and regression

Galton was also half-cousins with Charles Darwin (sharing the same grandfather) and took a strong interest in how physical and mental characteristics move from generation to generation — **Heredity**

His work on regression started with a book entitled Heredity Genius from 1869 in which **he studied the way "talent" ran in families** — The book has lists of famous people and their famous relatives (great scientists and their families, for example)

He noted that there was a rather dramatic reduction in awesomeness as you moved up or down a family tree from the great person in the family (the Bachs or the Bernoullis say) — And thought of this as a kind of **regression toward mediocrity**

NATURAL INHERITANCE

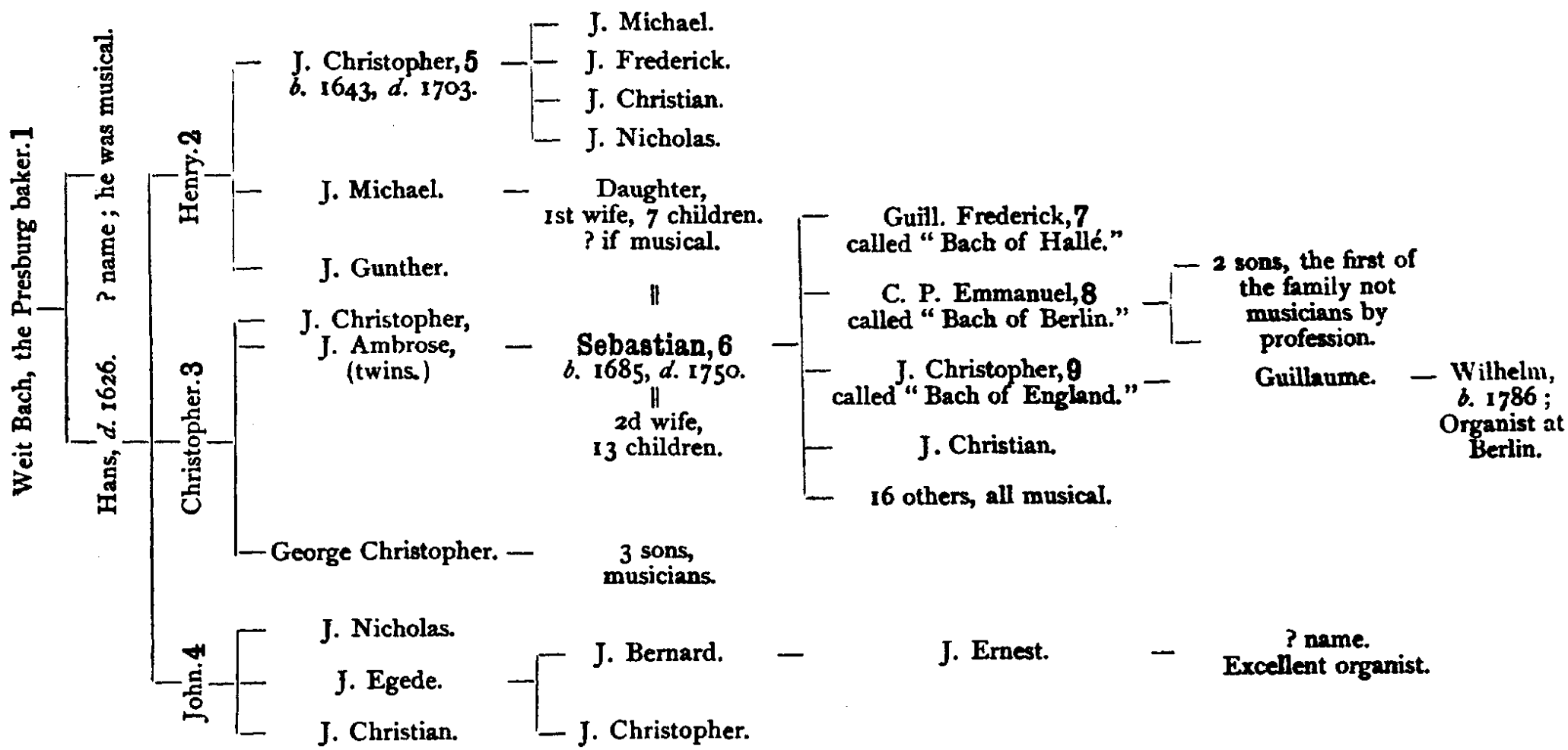
BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

"HEREDITARY GENIUS," "INQUIRIES INTO HUMAN FACULTY," ETC.

Pedigree of the Bachs.



Galton and regression

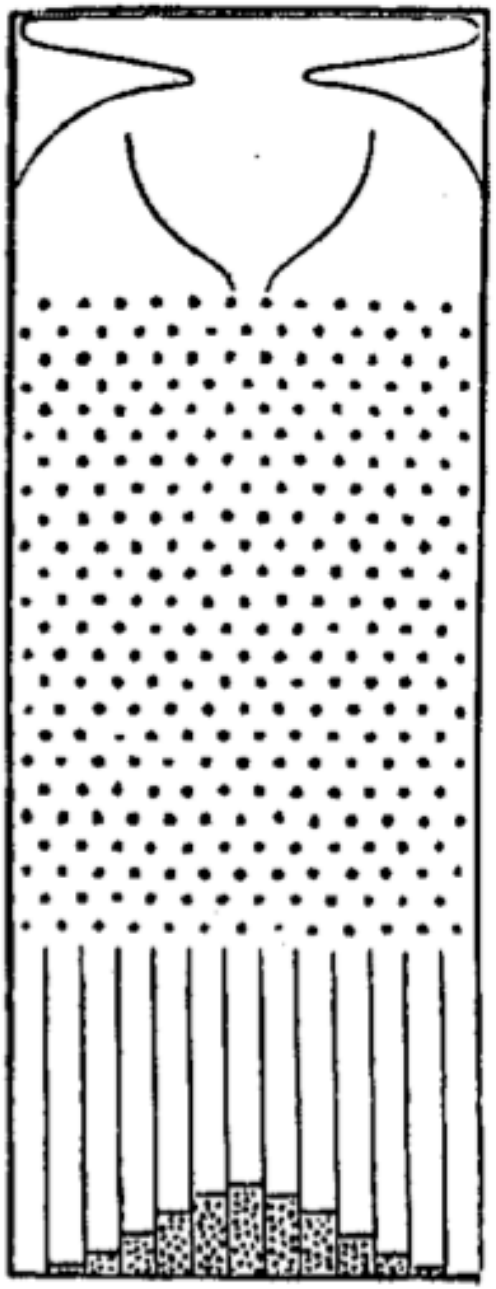
In some sense, his work builds on that of Adolphe Quetelet — Quetelet saw normal distributions in various aggregate statistics on human populations

Galton writes “Order in Apparent Chaos — I know of scarcely anything so apt to impress the imagination as the wonderful cosmic order expressed by the law of frequency of error. **The law would have been personified by the Greeks and deified, if they had known of it.**”

Galton and regression

In 1873, Galton had a machine built which he christened the Quincunx — The name comes from the similarity of the pin pattern to the arrangement of fruit trees in English agriculture (quincunxial because it was based on a square of four trees with a fifth in the center)

The machine was originally devised to illustrate the central limit theorem and to show how a number of independent events might add up to produce a normal distribution — Lead shot were dropped at the top of the machine and piled up according to the binomial coefficients at the bottom



Galton and regression

Relating the normal curve (and the associated central limit theorem) to heredity, however, proved difficult for Galton — He could not connect the curve to the transmission of ability or physical characteristics from one generation to the next writing

> "If the normal curve arose in each generation as the aggregate of large number of factors operating independently, no one of them of overriding or even significant importance, **what opportunity was there for a single factor such as parent to have a measurable impact?**"

So at first glance the normal curve that Galton was so fond of in Quetelet's work was at odds with the possibility of "inheritance" — Galton's solution to the problem would be the formulation of regression and its link to the bivariate normal distribution

Some history

Galton collected data from 928 children, recording, among other things, their heights and the heights of their parents (Quetelet, ironically had data of the same sort with both the heights and chest measurements of the Scottish soldiers)

He “transmuted” the heights of the girls and women in his data set, multiplying these heights by 1.08 and then forms a table of the heights of children versus the heights of their mid-parents (the average height of the father and transmuted mother)

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

2 FATHER ...

1. Date of birth. August 7th 1838. Birthplace. Neath, Glamorganshire.

2. Occupation. Clerk in Holy Orders. Residences.

3. Age at marriage.	The place for this entry is at 4 in next page.	23.
4. do. of wife	The place for this entry is at 3 in next page	23.

5. Mode of life so far as affecting growth or health.

6. Was early life laborious? why and how? No.

7. Adult height. 5ft. 6in. Colour of hair when adult. Dark brown. Colour of eyes. Blue.

8. General appearance. Slender.

9. Bodily strength and energy, if much above or below the average. During 22 years from Ordination, have preached 3600 times. Yearly average unable to preach (from temporary indisposition) only 4 Sundays during these 22 years.

10. Keenness or imperfection of sight or other senses. Have always possessed good sight, both for near & distant objects. No failure as yet. (age 46).

11. Mental powers and energy, if much above or below the average. Rapid reader.

12. Character and temperament. Cool, cautious, methodical.

FATHER ...

1. Date of birth. *August 7th 1838.* Birthplace. *South Glamorganshire.*
2. Occupation. *Clerk in Holy Orders.* Residences.
3. Age at marriage. The place for this entry is at 4 in next page. *23.*
4. do. of wife. The place for this entry is at 3 in next page. *23.*
5. Mode of life so far as affecting growth or health.
6. Was early life laborious? why and how? *No.*
7. Adult height. *5 ft. 6 in.* Colour of hair when adult. *Dark brown.* Colour of eyes. *Blue.*
8. General appearance. *Slender.*
9. Bodily strength and energy, if much above or below the average. *During 25 years from Ordination, have preached three times [illegible] unable to preach (from temporary indisposition) only [illegible] Sundays during these 25 years.*
10. Keenness or imperfection of sight or other senses. *Have always possessed good sight, both for near & distant objects. No failure as yet. (age 46).*
11. Mental powers and energy, if much above or below the average. *Rapid reader.*
12. Character and temperament. *Cool, cautious, methodical.*
13. Favorite pursuits and interests. Artistic aptitudes. *From earliest childhood a decided inclination to the study of music. [illegible] musical composition (not very difficult) at first sight, [illegible]*
14. Minor ailments to which there was special liability { in youth. *Colds in the head. Also headaches.* { in middle age. *Very rarely suffer now from these attacks.*
15. Graver illnesses { in youth. *None, excepting measles & whooping cough, but mild.* { in middle age. *None, excepting lumbago occasionally.*
16. Cause and date of death, and age at death. *Still living.*
17. General remarks.

MOTHER ...

1. Date of birth. *Dec. 18, 1838.* Birthplace. *Lindley, Huddersfield.*
2. Occupation. Residences. *Lindley, Scarborough, Skipton, [illegible]*
3. Age at marriage. *23* Total No. of { sons *1* { daus. *1* No. of sons deceased. *None.* ages
4. ditto of husband. *23* No. of daus. deceased. *None.* ages
5. Mode of life so far as affecting growth or health. *Educated at private schools.*
6. Was early life laborious? why and how? *No.*
7. Adult height. *5 ft. 1½ in.* Colour of hair when adult. *Dark brown.* Colour of eyes. *Brown.*
8. General appearance. *Slender frame. Dark complexion.*
9. Bodily strength and energy, if much above or below the average.
10. Keenness or imperfection of sight or other senses. *Sight & other senses good.*
11. Mental powers and energy, if much above or below the average.
12. Character and temperament. *Nervous & anxious.*
13. Favorite pursuits and interests. Artistic aptitudes.
14. Minor ailments to which there was special liability { in youth. *Bilious attacks, with vomiting & headache.* { in middle age. *Indigestion.*
15. Graver illnesses { in youth. { in middle age. *Bronchitis.*
16. Cause and date of death, and age at death. *Still living.*
17. General remarks.

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

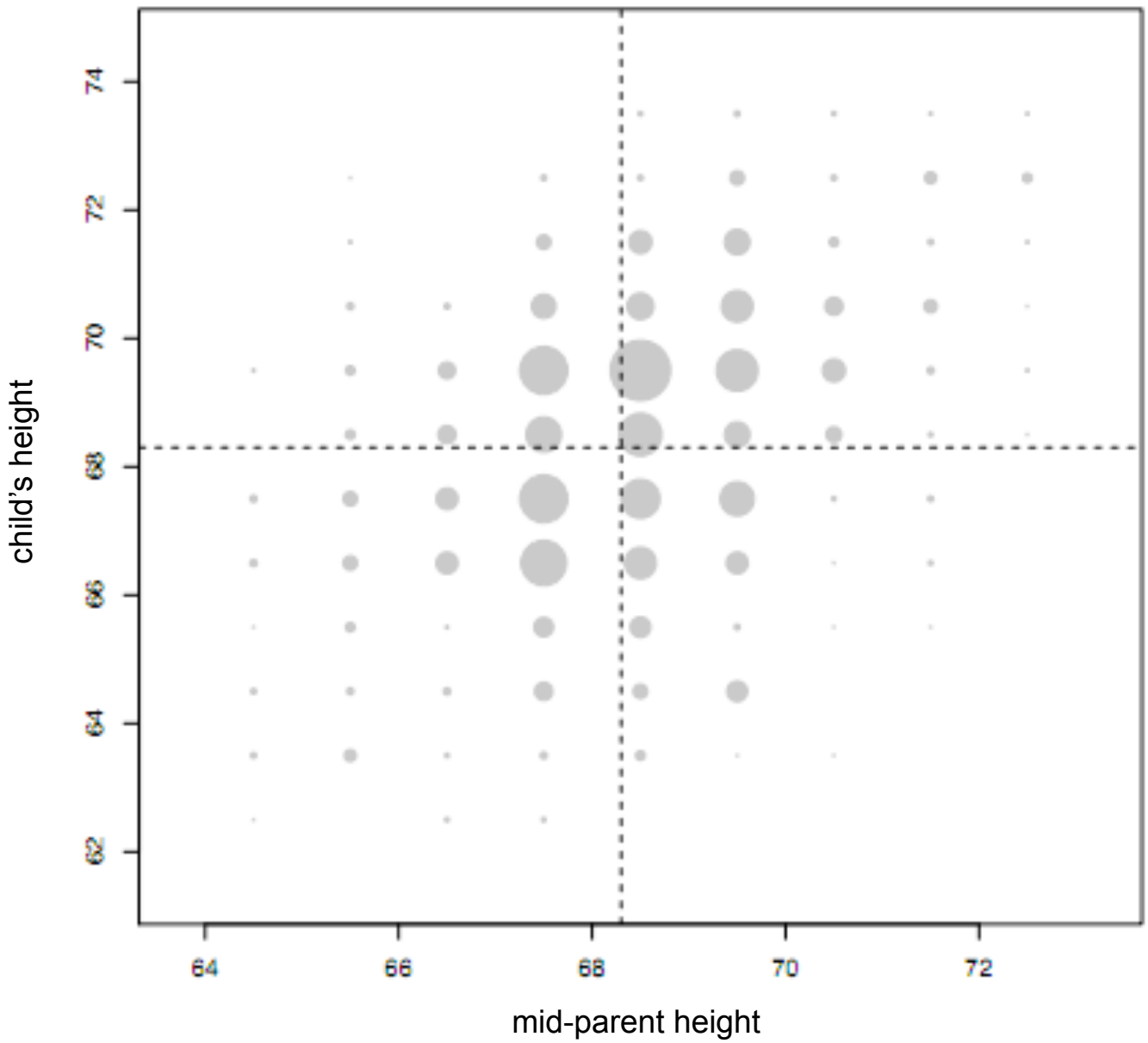
Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Heights

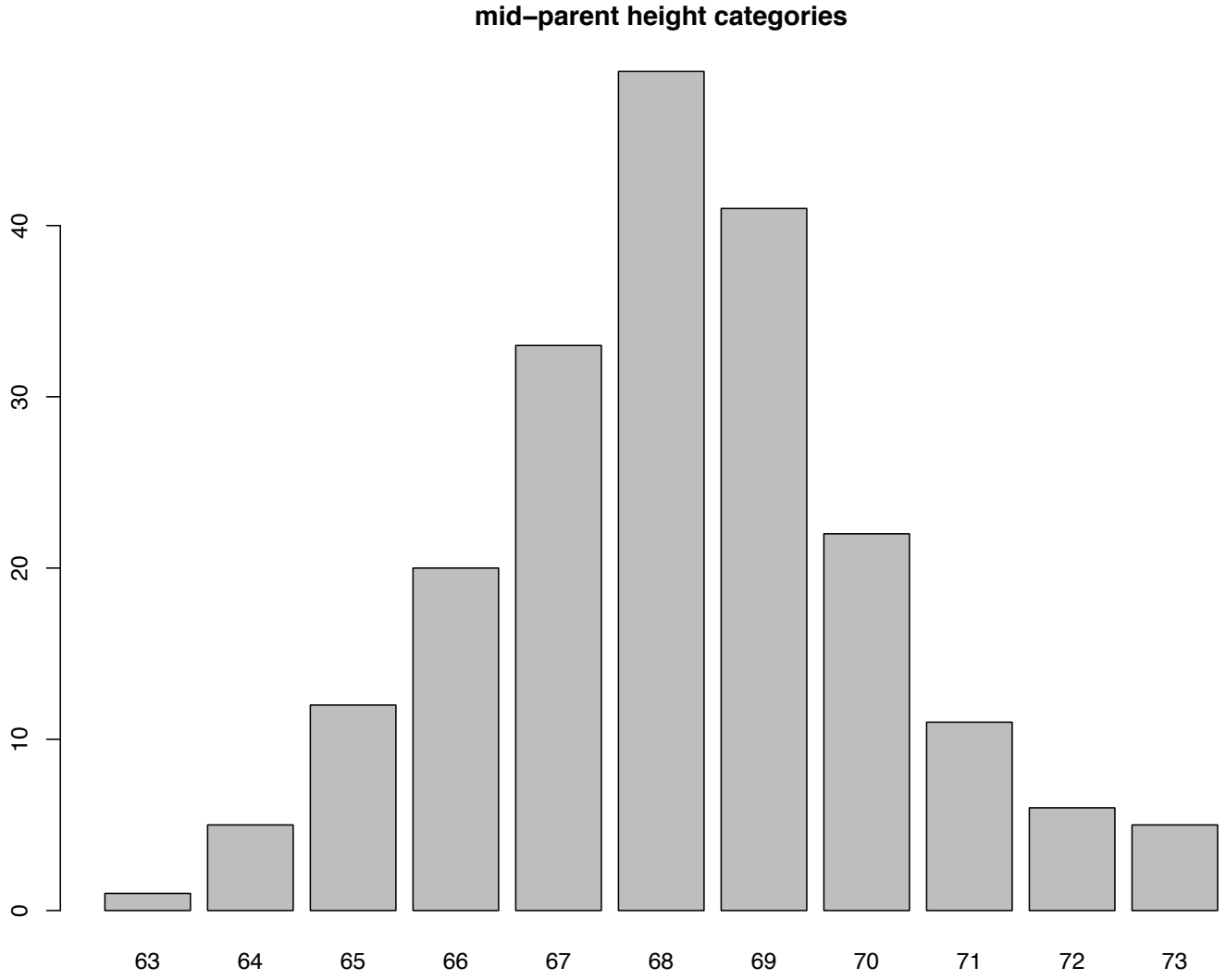
On the next slide we present another view of this table -- Here the different cells in the table are represented by circles that are sized according to the counts in each cell

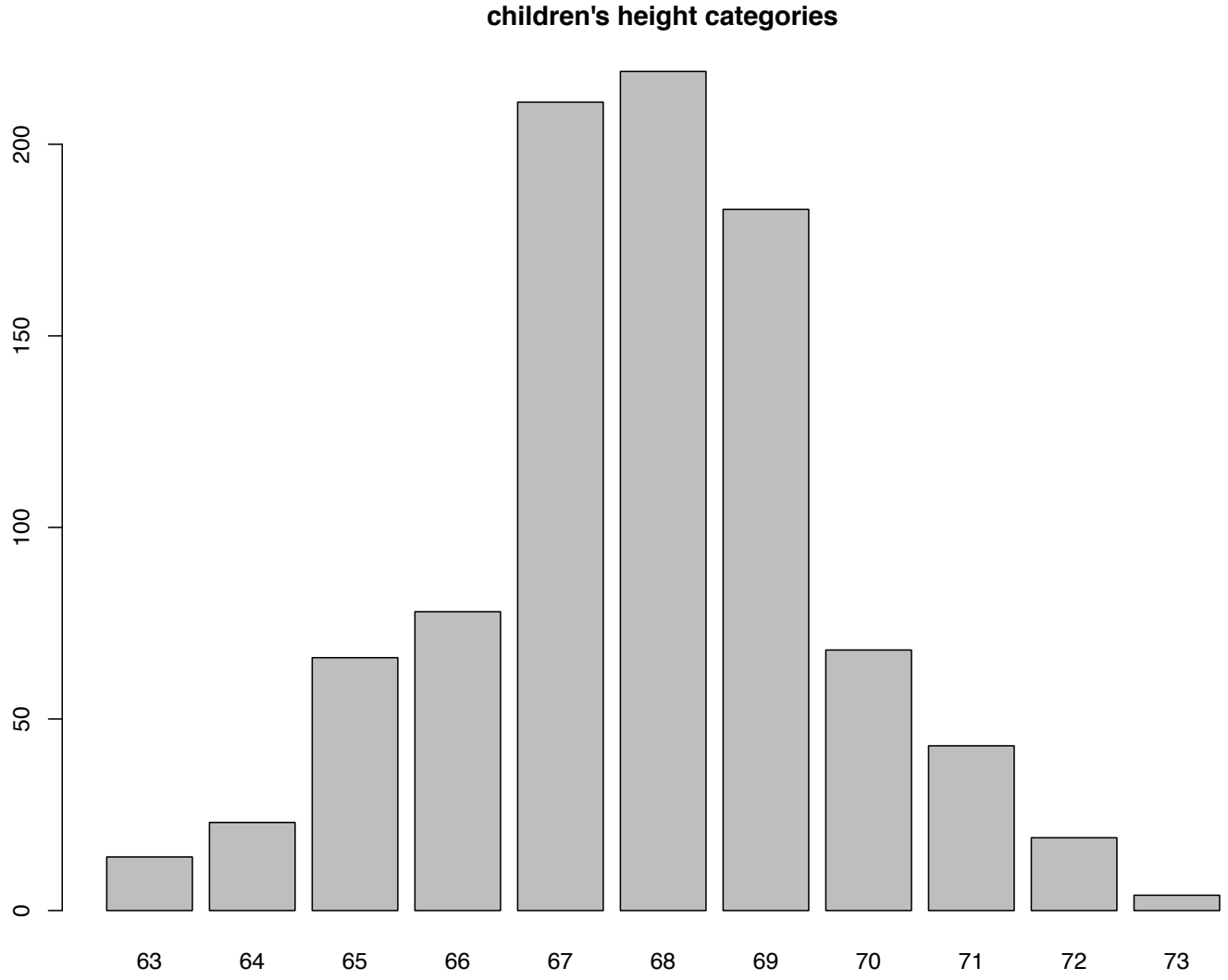
The dashed lines mark the means of the mid-parents' and (transmuted) children's heights (both about 68.3 inches) -- What does this display and the table suggest about the "data," the paired values of child and mid-parent heights?


child's height
mid-parent height
62
64
66
68
70
72
74
64
66
68
70
72

Heights

In addition to the elliptical look of the data distribution, the “marginal” height distributions (say, the distribution of mid-parent heights in the study considered on their own) also look normal...

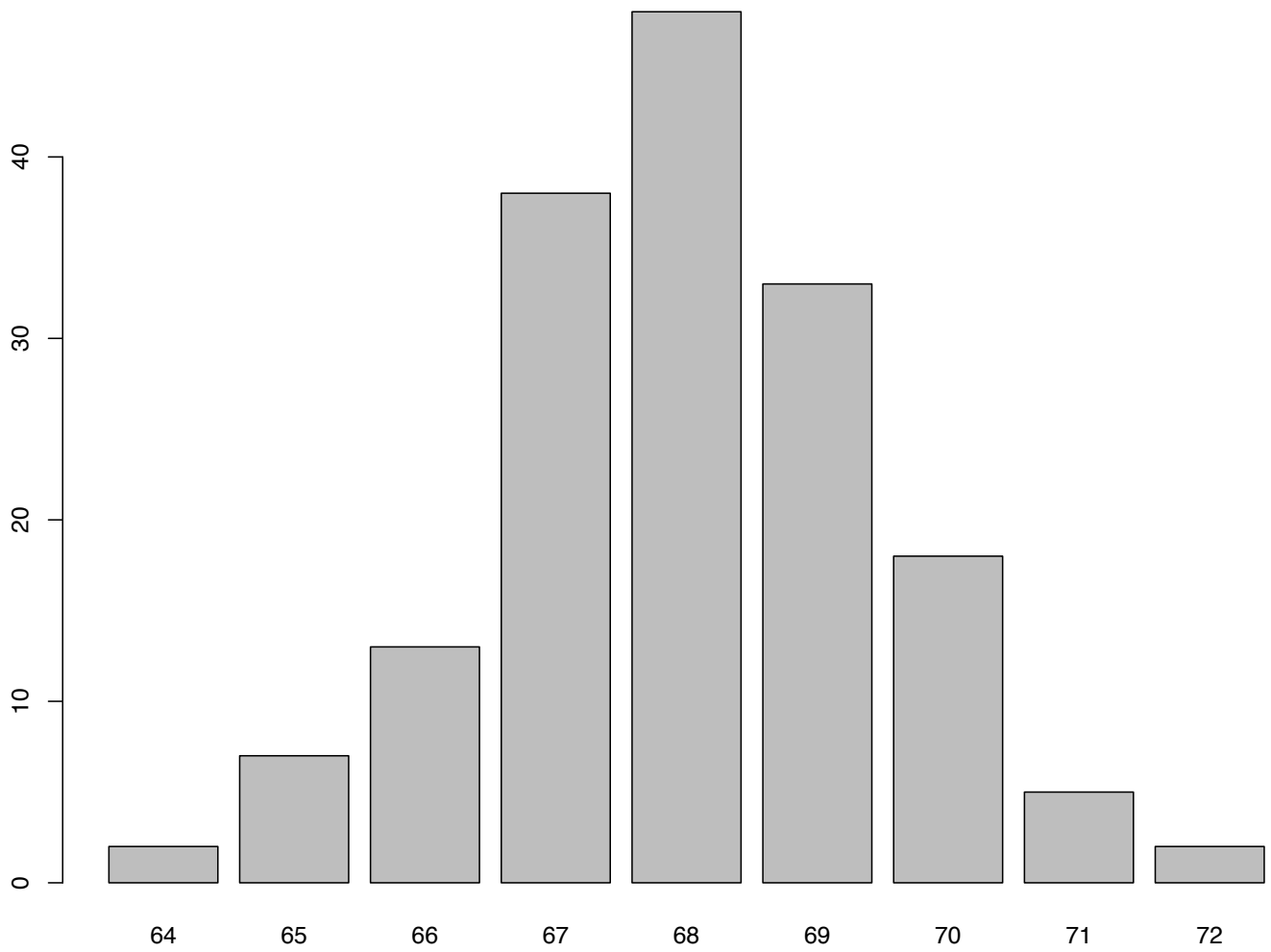

mid−parent height categories
0
10
20
30
40
63
64
65
66
67
68
69
70
71
72
73


children's height categories
0
50
100
150
200
63
64
65
66
67
68
69
70
71
72
73

Heights

He went on to further notice that in each of his separate categories of mid-parent heights, the (transmuted) children's heights also had normal distributions — That is, if you look at just one of the columns of his table, you see again a normal distribution

Here's one example...


children's heights, parents between 68.2 and 69.2 inches
0
10
20
30
40
64
65
66
67
68
69
70
71
72

Galton and regression

In 1873, Galton had a machine built which he christened **the Quincunx** -- The name comes from the similarity of the pin pattern to the arrangement of fruit trees in English agriculture (quincunxial because it was based on a square of four trees with a fifth in the center)

The machine was originally devised to **illustrate the central limit theorem** and how a number of independent events might add up to produce a normal distribution -- Lead shot were dropped at the top of the machine and piled up according to the binomial coefficients at the bottom

The other panels in the previous slide illustrate a thought experiment by Galton (it's not clear the other devices were ever made) -- The middle region (between the A's) in the central machine, could be closed, **preventing the shot from working their way down the machine**

FIG .7.

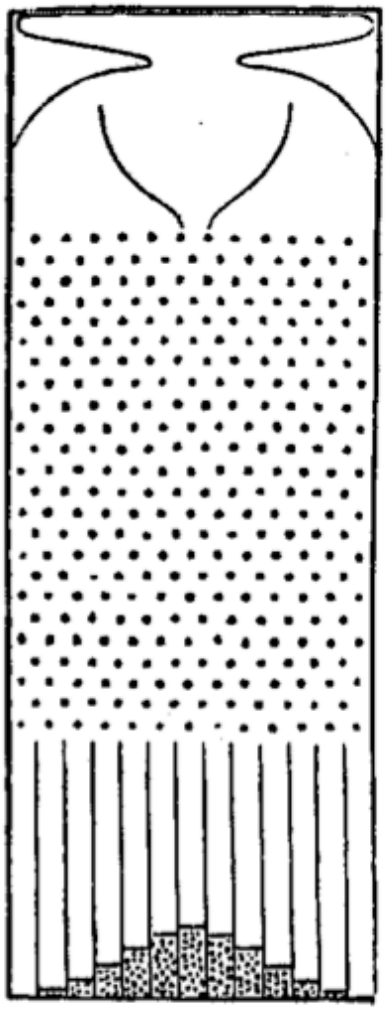


FIG .8.

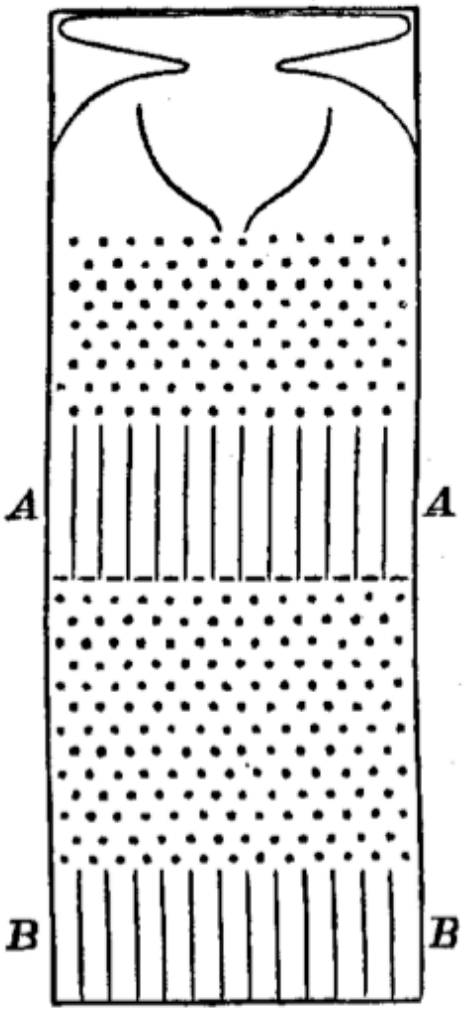
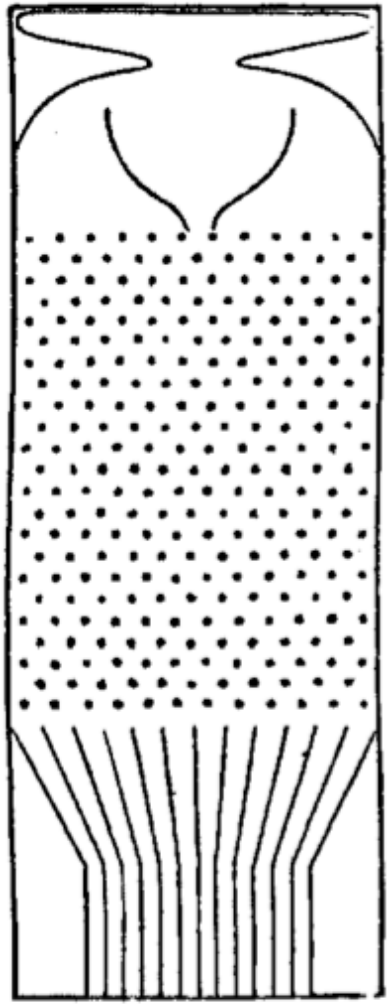


FIG .9.



Galton and regression

By imagining holding back a portion of the shots, Galton expected to still see a normal distribution at the bottom of the machine, but one with less variation -- As he opened each barrier, t**he shot would deposit themselves according to small normal curves**, adding to the pattern already established

Once all the barriers had been opened, you'd be left with the original normal distribution at the bottom -- Galton, in effect, showed how the normal curve **could be dissected into components** which could be traced back to the location of the shot at A-A level of the device

In effect, he established that a normal mixture of normals is itself normal -- But with this idea in hand, **we see his tables of human measurements in a different light**...

FIG .7.

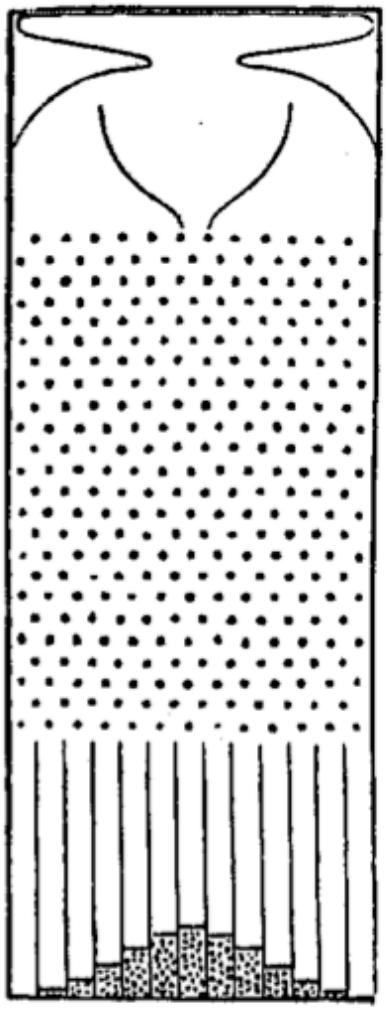


FIG .8.

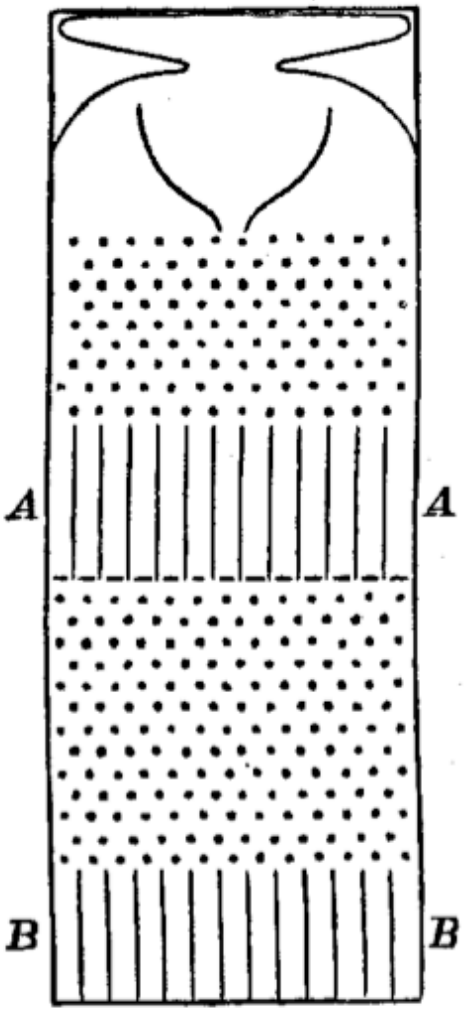


FIG .9.

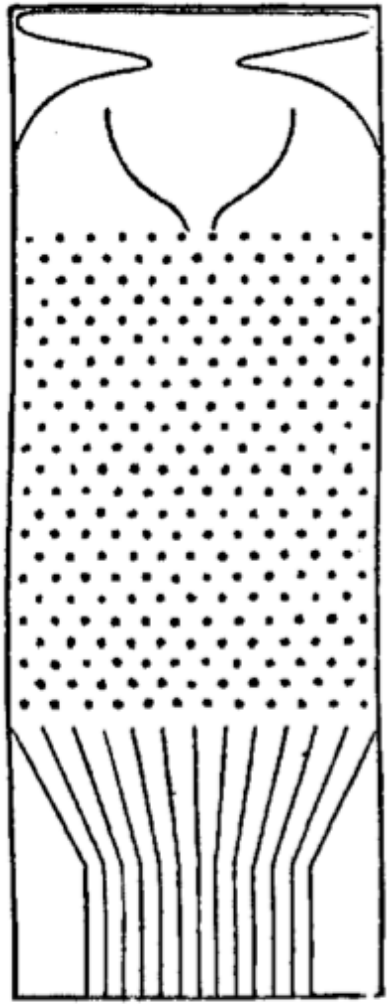


TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

TABLE 13 (Special Data).

RELATIVE NUMBER OF BROTHERS OF VARIOUS HEIGHTS TO MEN OF VARIOUS HEIGHTS, FAMILIES OF FIVE BROTHERS AND UPWARDS BEING EXCLUDED.

Heights of the men in inches.	Heights of their brothers in inches.													Total cases.	Medians.
	Below 63	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	Above 74		
74 and above	1	1	...	...	...	...	...	1	1	...	5	3	12	24	
73·5	...	...	...	...	...	1	3	4	8	3	3	2	3	27	
72·5	...	...	...	...	1	1	6	5	9	9	8	3	5	47	71·1
71·5	...	1	...	1	2	8	11	18	14	20	9	4	...	88	70·2
70·5	...	...	1	1	7	19	30	45	36	14	9	8	1	171	69·6
69·5	...	1	2	1	11	20	36	55	44	17	5	4	2	198	69·5
68·5	...	1	5	9	18	38	46	36	30	11	6	3	...	203	68·7
67·5	2	4	8	26	35	38	38	20	18	8	1	1	...	199	67·7
66·5	4	3	10	33	28	35	20	12	7	2	1	...	...	155	67·0
65·5	3	3	15	18	33	36	8	2	1	1	...	...	...	110	66·5
64·5	3	8	12	15	10	8	5	2	1	...	...	...	...	64	65·6
63·5	5	2	8	3	3	4	1	1	...	1	...	...	1	20	
Below 63.....	5	5	3	3	4	2	...	...	...	...	...	...	1	23	
Totals.....	23	29	64	110	152	200	204	201	169	86	47	28	25	1329	

Galton and regression

Looking at these tables, we see the Quincunx at work -- The righthand column labeled “Total number of Adult Children” being the **counts of shot at the A-A level**, while the row marked “Totals” can be thought of as **the distribution one would see at the bottom of the device** when all the barriers are opened and the **individual counts in each row as the corresponding normal curves**

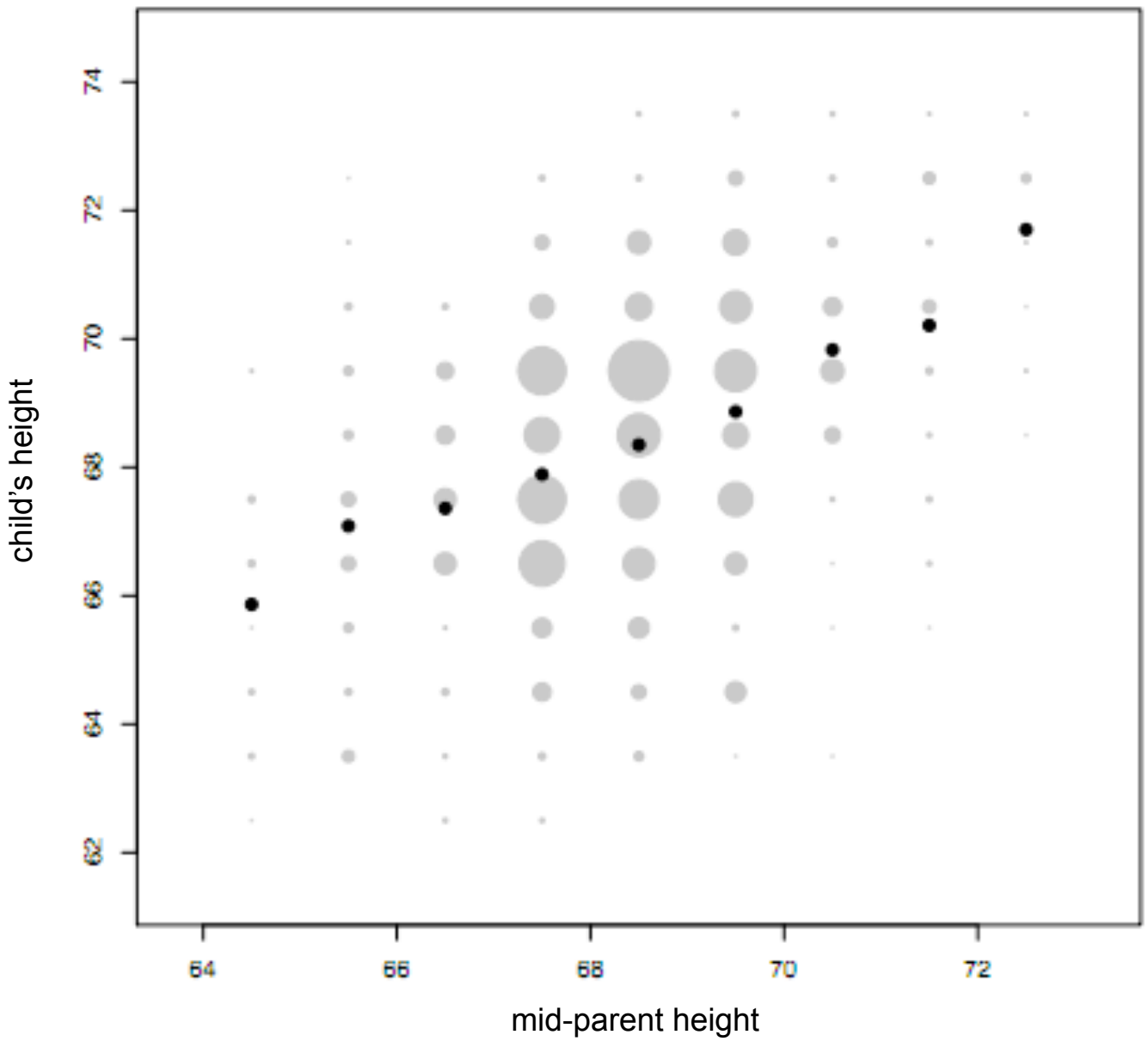
By 1877, Galton was starting to examine these ideas mathematically -- He essentially **discovered the important properties of the bivariate normal distribution** (the bivariate normal had been derived by theorists unknown to Galton, but they did not develop the idea of regression, nor did they attempt to fit it from data as Galton did)

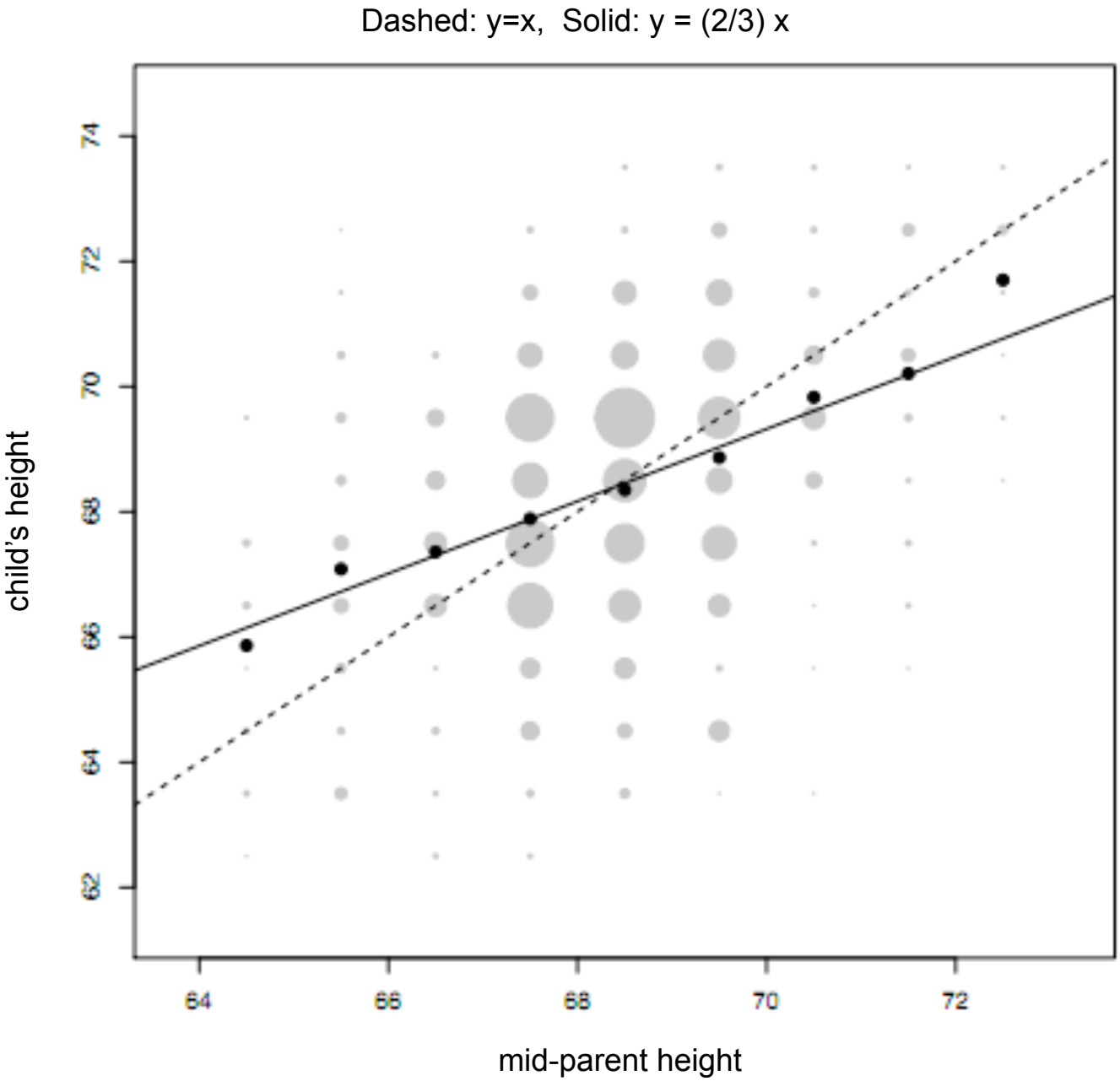
Galton and regression

In his text Natural Inheritance, he approached a table like this by first examining the **heights of the mid-parents** and noted that it appeared to be normal -- He then looked at the **marginal distribution of child heights** and found them to also be normally distributed

He then considered the heights of the children associated with different columns in his table, plotting median values against mid-parental height and finding a straight line (which he fit by eye)

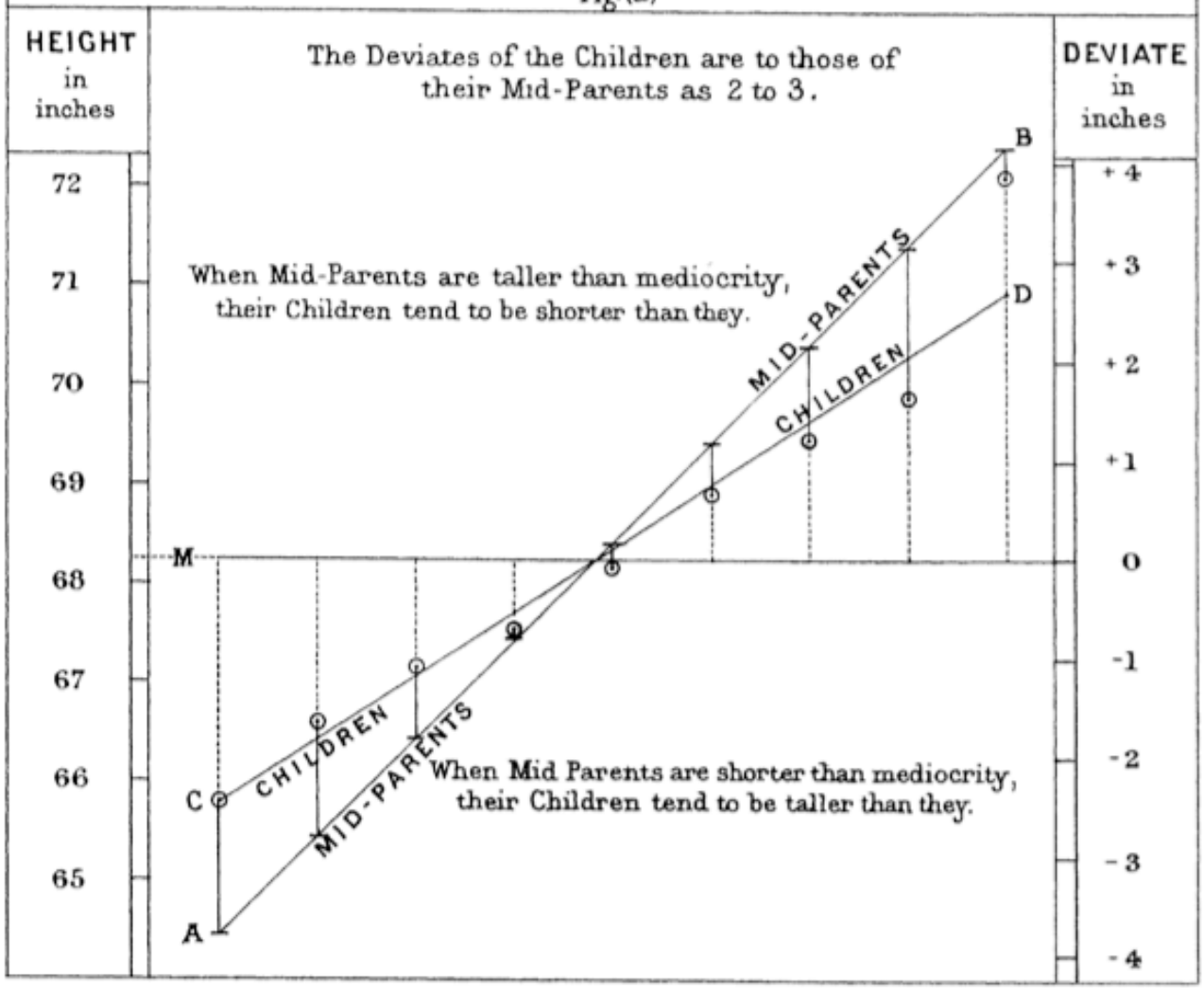
He found that the slope was about 2/3 -- If children were on average as tall as their parents, he'd expect a slope of 1, leading him to coin the phrase "regression toward mediocrity"


child's height
62
64
66
68
70
72
74
64
66
68
70
72
mid-parent height


Dashed: y=x, Solid: y = (2/3) x
child's height
mid-parent height
62
64
66
68
70
72
74
64
66
68
70
72

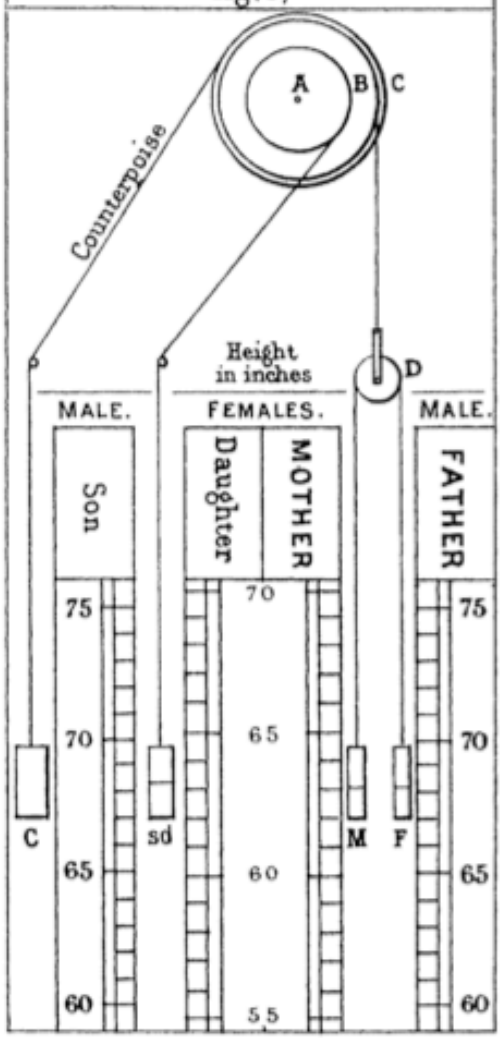
RATE OF REGRESSION IN HEREDITARY STATURE.

Fig. (a)



FORECASTER OF STATURE

Fig (b)



J P & W R Emslie, lith

Journ. Anthropolog. Inst., Vol. XV, Pl. IX

Galton and regression

What Galton found through essentially geometric means was the following relationship

$$\frac{y - \bar{y}}{sd(y)} = r \frac{x - \bar{x}}{sd(x)}$$

where we might take x to be the heights of mid-parents and y to be the heights of their adult offspring -- The quantity r is the correlation coefficient between x and y (another Galton innovation)

This gives a precise meaning to his phrase “regression to the mean”

Galton and regression

The r here is the so-called correlation coefficient -- If we are given data in n pairs, say, mid-parent heights and children's heights, $(X_1, Y_1), \ldots, (X_n, Y_n)$, then it is computed as

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\widehat{\sigma}_X}\right)\left(\frac{Y_i - \overline{Y}}{\widehat{\sigma}_Y}\right)$$

where $\overline{X}$ and $\widehat{\sigma}_X$ are the mean and sample standard deviation of $X_1, \ldots, X_n$ and $\overline{Y}$ and $\widehat{\sigma}_Y$ are the mean and sample standard deviation of $Y_1, \ldots, Y_n$

We will see what this is estimating shortly -- For now, notice that it seems to be measuring the degree of agreement (the co-relation) between X and Y

FIG. 10.

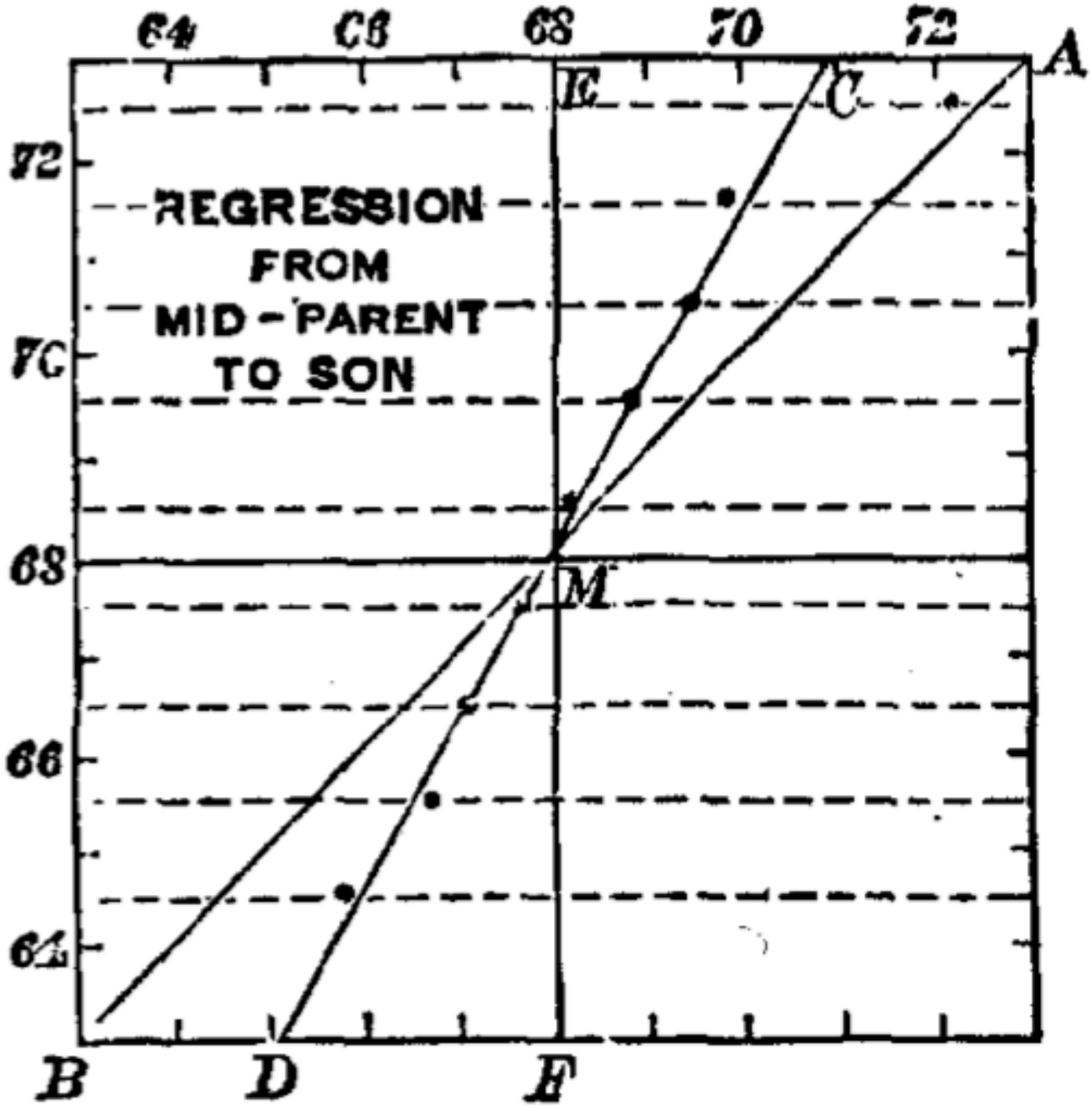
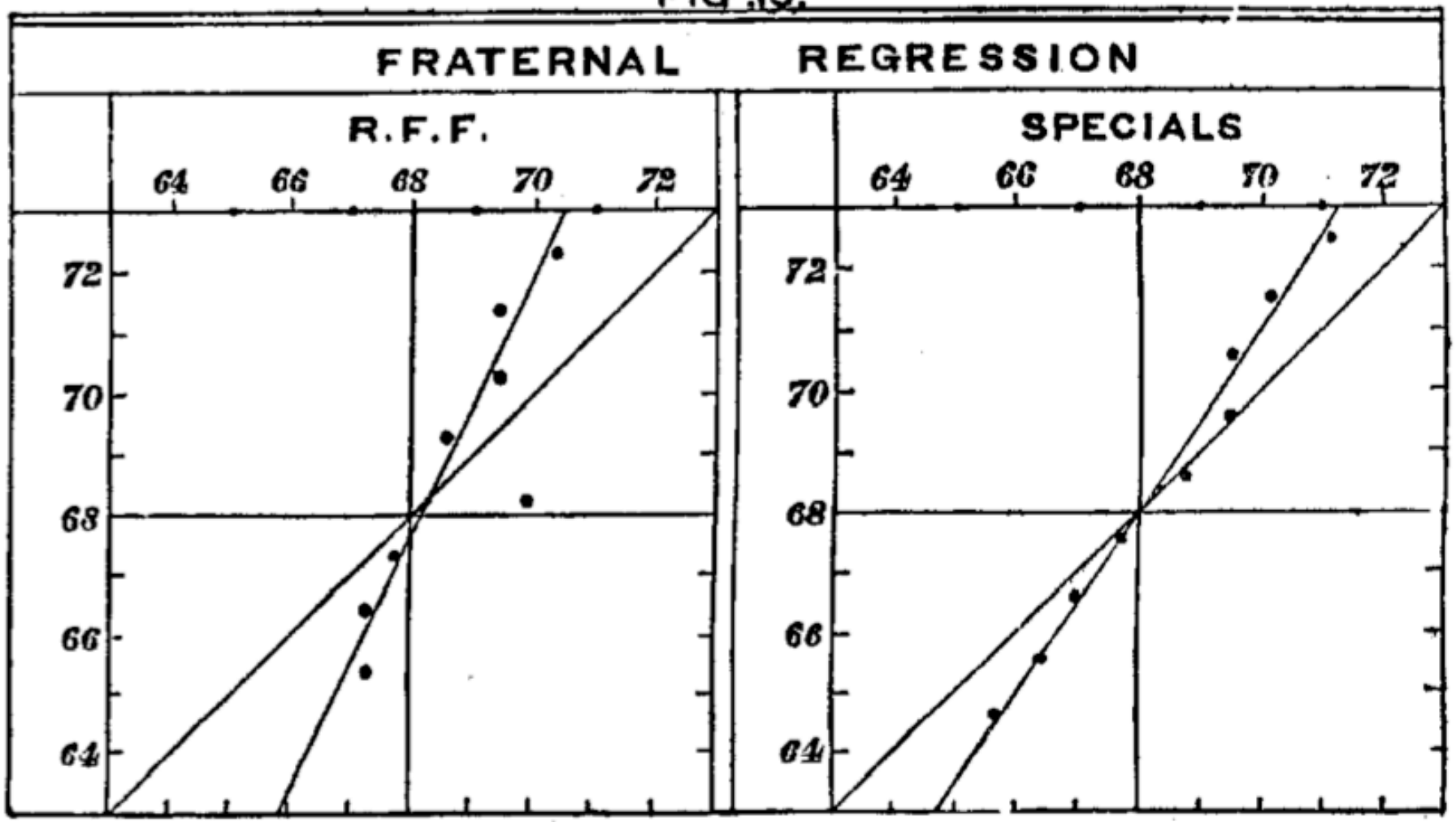

64
66
68
70
72
A
E
C
72
REGRESSION
FROM
MID-PARENT
TO SON
70
68
M
66
64
B
D
F

FIG .13.


FRATERNAL REGRESSION
R.F.F.
SPECIALS
64
66
68
70
72
72
70
68
66
64

Galton and regression

Galton also noticed, however, that **a similar kind of regression happened in reverse** -- That is, that if you transposed the table, you'd find a slope of 1/3 relating the average mid-parent's height to that of their children

He surmised that the regression effect was more a fact about **the bivariate normal distribution** than anything else -- This is a lesson that many researchers have failed to appreciate even now

I found it hard at first to catch the full significance of the entries in the table, which had curious relations that were very interesting to investigate. They came out distinctly when I "smoothed" the entries by writing at each intersection of a horizontal column with a vertical one, the sum of the entries in the four adjacent squares, and using these to work upon. I then noticed (see [fig. 6.6]) that lines drawn through entries of the same value formed a series of concentric and similar ellipses. Their common centre lay at the intersection of the vertical and horizontal lines, that corresponded to 68.25 inches. Their axes were similarly inclined. The points where each ellipse in succession was touched by a horizontal tangent, lay in a straight line inclined to the vertical in the ratio of 2/3; those where they were touched by a vertical tangent lay in a straight line inclined to the horizontal in the ratio of 1/3. These ratios confirm the values of average regression already obtained by a different method, of 2/3 from mid-parent to offspring, and of 1/3 from offspring to mid-parent, because it will be obvious on studying [fig. 6.6] that the point where each horizontal line in succession is touched by an ellipse, the greatest value in that line must appear at the point of contact. The same is true in respect to the vertical lines. These and other relations were evidently a subject for mathematical analysis and verification. (Galton 1885c, 254–255)

Galton and regression

To complete this story, Galton enlisted the help of a mathematician, Hamilton Dickson -- The problem he wanted solved was the following

Suppose x and y are expressed as deviations from the mean and that x is normal with mean zero and standard deviation Q_x

Also suppose that conditional on a fixed value of x, y is also normal with mean $\beta_{y|x}$ and standard deviation $Q_{y|x}$

What is the joint density of x and y and, in particular, are the contours of equal probability elliptical?

What is the conditional distribution of x given y, and in particular, what is the relation between the two regression coefficients?

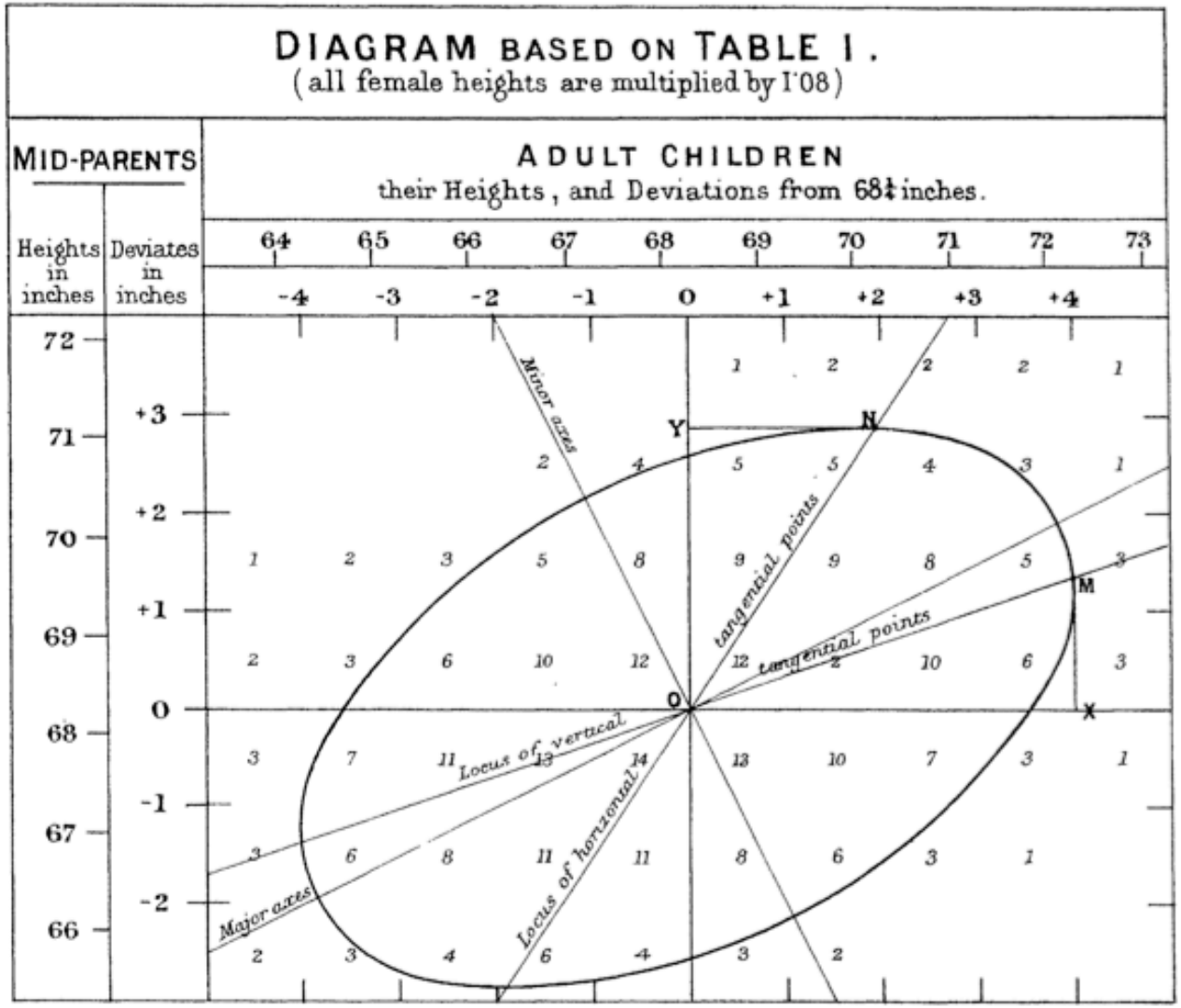
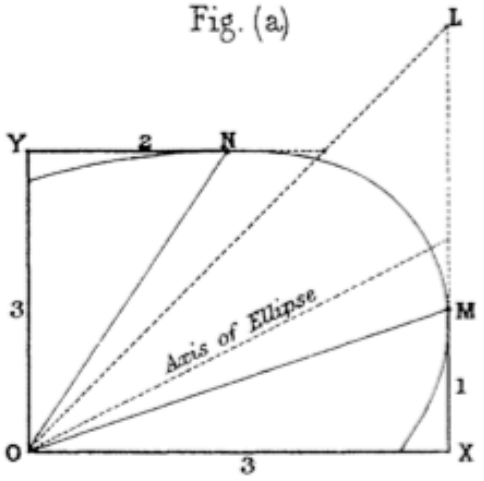


Fig. (a)



J.P. & W.R. Emslie, lith.

The bivariate normal

In his response, Dickson derived the bivariate normal distribution and the associated marginals and conditionals -- Suppose Y_1 is normal with mean μ_1 and variance σ_1^2 and that Y_2 is normal with mean μ_2 and variance σ_2^2

The pair Y_1 and Y_2 are said to have a bivariate normal distribution if their density is given by

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(y_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right\}\right]$$

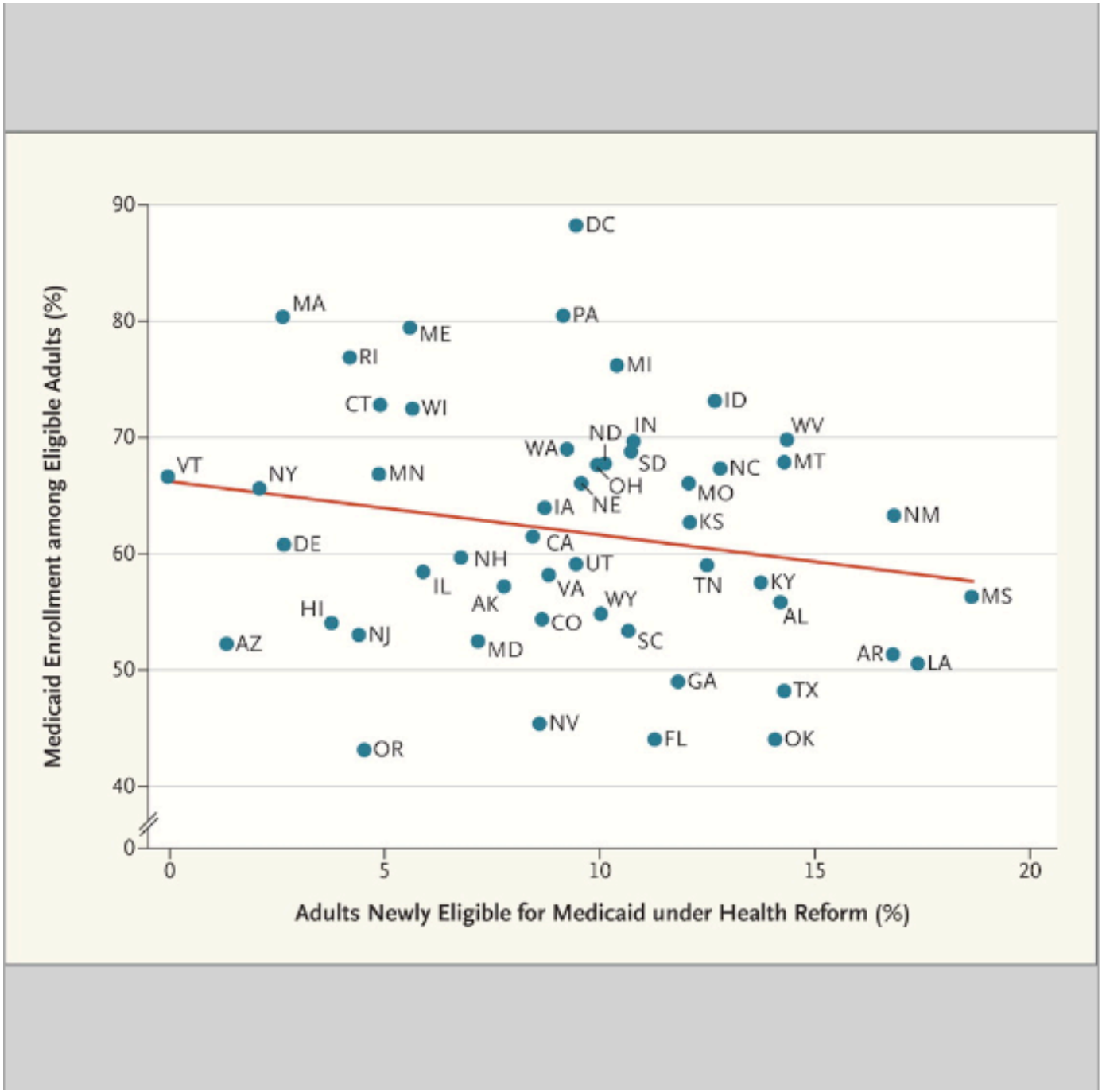
where ρ is the so-called correlation between Y_1 and Y_2 -- This is again a parametric family with 5 parameters, μ_1, μ_2, σ_1, σ_2, and ρ

Regression today

Of course univariate (simple linear) regression or curve and surface fitting is only the start of the story

Regression has become a powerful tool in a number of quantitative disciplines -- In many cases, regression models act as a kind of **social probe**, providing researchers with a glimpse into the workings of some larger phenomenon

How we reason with these models is the subject of a big part of today's lecture -- Let's start by looking at a few examples of modern regression analysis



Medicaid Enrollment among Currently Eligible Adults (2007 through 2009) and Percentage of Adults Who Will Become Eligible in 2014 under Health Care Reform, by State.

The population sample was restricted to eligible adults with no other form of health insurance; noncitizens were excluded from the analysis. Results are based on an analysis of data from the Current Population Survey of 2007 through 2009. The red line shows the regression equation: Enrollment=0.660.46×Newly Eligible (P=0.17).

Table 2. Percent Approving Same-Sex Marriage Ban; OLS Estimates, U.S. Counties

Independent Variable	1	2	3	4
Percent Women Not Working in Labor Force	.152**	.170**	.075*	.090*
Occupational Sex Segregation	16.310***	18.048***	8.216***	9.559***
Percent Households Married with Children	.219*	.322***	.156**	.217***
Percent Same-Sex Households	−2.038	−2.237	−5.428***	−5.956***
Percent Unmarried Opposite-Sex Households	−1.682***	−1.720***	−1.314***	−1.210***
Residential Instability	−.033	−.024	−.083**	−.066*
Percent Homes Not Owner Occupied	−.076*	−.084*	−.079***	−.109***
Crime Rate		.099***		.022*
Percent Production or Construction Occupations	.194***	.234***	.067**	.063*
Percent Professional Occupations	−.178*	−.100	−.118*	−.119*
Percent Self-Employed	−.100	−.093	−.198***	−.234***
Median Family Income ($1,000s)	−.138***	−.045	−.166***	−.148***
Percent Receiving Public Assistance	−.054	.037	.125	.207
Mean Years of Education	−1.328*	−2.611***	−2.040***	−2.606***
Percent Enrolled in College	−.060	.020	−.185**	−.139
Population Density (logged)	−.111	−.332	.001	.040
Percent Urban	.029**	.011	.015**	.011
Republican Voting (percent Bush 2000)	.350***	.381***	.287***	.317***
Percent Evangelical	.180***	.179***	.031***	.020
Percent Catholic	.063***	.064***	−.025**	−.035**
Median Age	−.155	−.022	−.314***	−.247**
Percent African American	.092***	.085***	−.006	.009
Percent Latino	.043*	.024	−.108***	−.118***
LGBT Organizations	−.114	.974	−2.096***	−1.576**
Civil Rights Organizations	−.007	−.004	.073	.113
Antidiscrimination Legislation	−3.283***	−2.797***	−2.656***	−1.918*
Alabama			−6.679***	−6.132***
Arizona			−23.814***	−23.878***
Arkansas			−9.454***	−9.085***
California			−2.079*	−1.867*
Colorado			−13.112***	−12.478***
Florida			−7.906***	−8.263***
Georgia			−2.787***	−3.307***
Idaho			−17.638***	−17.591***
Kansas			−5.857***	−6.051***
Kentucky			−7.623***	−9.111***
Louisiana			−4.771***	−5.492***
Michigan			−15.107***	−15.211***
Mississippi			−2.865***	−1.619
Missouri			−5.502***	−5.583***
Montana			−8.701***	−8.893***
Nebraska			−5.916***	−5.754***
Nevada			−3.886***	−3.964***

Table 3. Regression Equations Predicting Sales, Number of Customers, Market Share, and Relative Profitability with Racial and Gender Diversity and Other Characteristics of Establishments

	Model 1	Model 2	Model 3	Model 4
Independent Variables	Sales	Customers	Market Share	Profitability
Constant	4.998***	61545.4	3.403***	3.363***
Racial diversity	.093***	433.86***	.007**	.006*
Gender diversity	.028**	195.642**	.001	.005**
Proprietorship	−.821	−370.78	−.232*	−.161
Partnership	.663	−6454.6	−.017	.256
Public corporation	−.109	7376.29	.214*	.202
Private corporation	−1.484**	−8748.7*	.008	.019
Company size	.00001*	.352*	.000**	.000**
Establishment size	.00001**	.119	.000	.000
Organization age	.013**	44.813	.001	.001
Agriculture	−1.942	4188.66	−.206	−.033
Mining	.739	−28856*	−.168	−.264
Construction	−.967	−875.7	−.152	−.036
Transportation/communications	−.052	1498.75	.119	.226
Wholesale trade	.008	−16383**	.136	−.064
Retail trade	−1.183**	7209.83*	.08	−.087
F. I. R. E.	−.683	−8335.4	−.212	.085
Business services	−1.49*	1552.28	.204*	.112
Personal services	−1.566*	1480.03**	.423**	−.001
Entertainment	−4.708**	−1504.5	−.191	−.076
Professional services	−.615	−13539**	.138	−.023
North	2.196***	23143.9***	−.06	−.039
Midwest	2.616***	14968.3***	−.023	−.073
South	1.82***	21152.8***	−.055	.059
R^2	.165***	.155***	.075**	.064**
N	506	506	469	484

Notes: Coefficients are unstandardized. For the dummy (binary) variable coefficients, significance levels refer to the difference between the omitted dummy variable category and the coefficient for the given category.
* $p < .1$; ** $p < .05$; *** $p < .01$.

Table 2. OLS Estimates of Effect of Selected Measures of Residential Segregation on Log of Total Foreclosures

	Dissimilarity Index		Isolation Index	
Variables	*B*	*SE*	*B*	*SE*
Index of Segregation				
African Americans	3.718**	.725	2.122**	.619
Hispanics	−.773	.596	.080	.656
Asians	−2.080*	.920	−2.161	1.636
Control Variables				
Housing Starts Ratio	2.980**	.960	3.067**	1.077
Wharton Land Use Index	.250**	.082	.272**	.096
Change in Housing Price Index	.082**	.024	.092**	.029
CRA-Covered Lending Share	−1.295	.912	−.810	1.061
Subprime Loan Share	3.022*	1.353	4.310**	1.581
MSA Credit Score Index	−.015*	.007	−.016*	.007
Log of Population	1.008**	.089	1.013**	.093
Percent with College Degree	−1.341	1.315	−.997	1.459
Log Median Household Income	.253	.509	.340	.515
Percent with Second Mortgage	.751	3.687	.225	4.350
Percent Workforce Unionized	−.025**	.011	−.022*	.011
Unemployment Rate	−.010	.064	.012	.071
Change in Unemployment Rate	.245**	.052	.213**	.063
Age of Housing Stock	.004	.012	.014	.013
Region				
Midwest	.434*	.200	.631**	.200
South	.042	.257	.081	.296
West	.463	.384	.679	.436
Coastal MSA	−.053	.123	.070	.133
Borders Rio Grande	−1.030**	.370	−1.054**	.380
Constant	1.960	7.557	.979	8.150
R^2	.91		.90	
Joint *F*-Test for Region	3.35*		7.97**	
Joint *F*-Test for Segregation	10.48**		6.28**	

Note: $N = 99$. Robust standard errors. Model also includes percent black, percent Hispanic, and percent Asian.

*$p < .05$; **$p < .01$ (two-tailed tests).

Regression today

In each of these cases, the model being examined is of the form

$$Y = b_0 x_1 + \cdots + b_p x_p + \epsilon$$

where each response Y is taken to be some linear function of a set of input variables $x_1, \ldots, x_p$ and the error ϵ is assumed to be independent of the inputs and to have mean zero

The applications on the previous page take Y to be everything from segregation measures to indices of profitability and sales

Regression today

But researchers are rarely satisfied with simply producing coefficient estimates and instead they want to judge the size of the coefficients both in terms of **statistical as well as practical significance**

The coefficient b in a model of the form

$$Y = b_0 x_1 + \cdots + b_p x_p + \epsilon$$

tells us how a unit change in a predictor (unemployment rate in a county or the age of its housing stock, say) affects the response (some measure of racial segregation, say), holding all other conditions the same -- **The absolute size of the coefficient, then, provides us with a sense of the practical importance of the predictor**

Regression today

In each of these cases, the model being examined is of the form

$$Y = b_0 x_1 + \cdots + b_p x_p + \epsilon$$

where each response Y is taken to be some linear function of a set of input variables $x_1, \ldots, x_p$ and the error ϵ is assumed to be independent of the inputs and to have mean zero

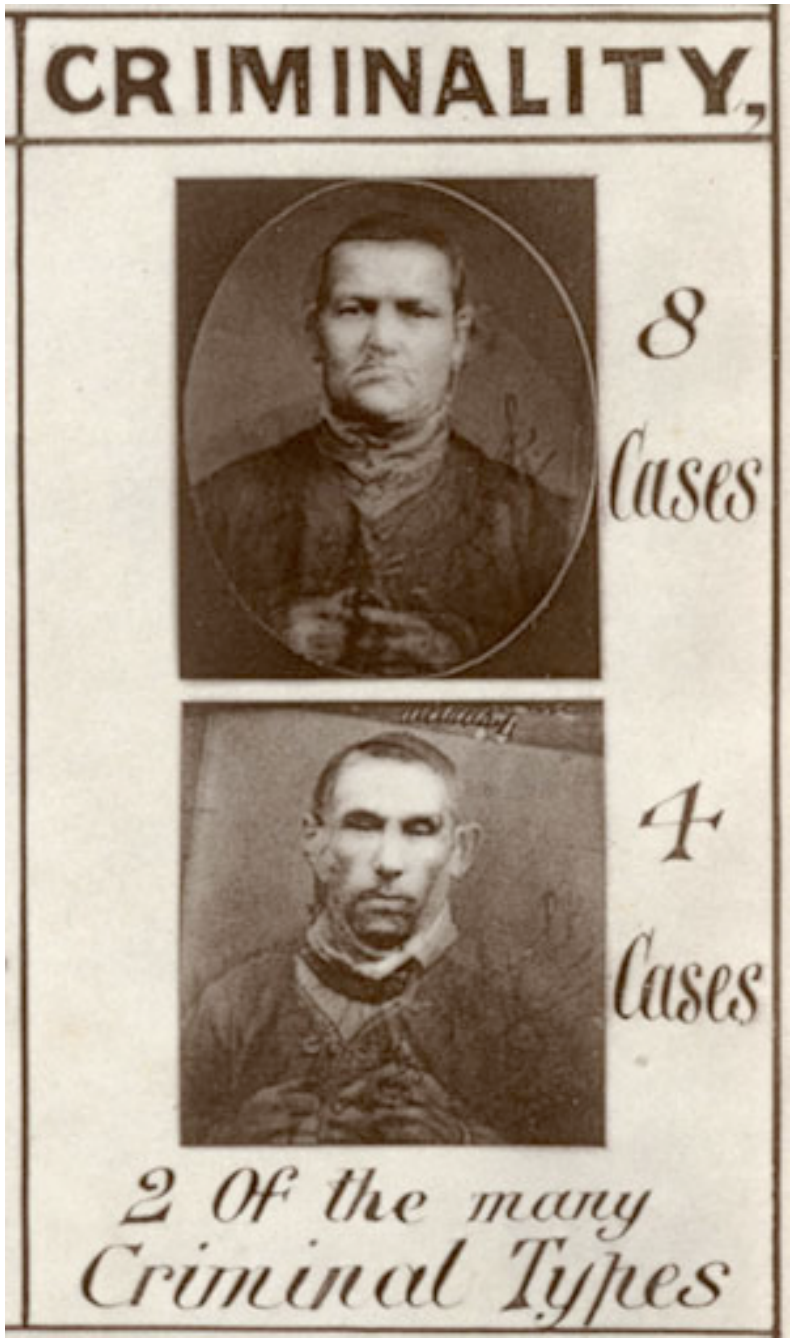
The applications on the previous page take Y to be everything from segregation measures to indices of profitability and sales

Galton

As we've seen, Galton was deeply committed to the idea of **the normal curve as an important force in nature** and (as with Quetelet) thought the mean value had particular importance as **an indicator of "type"**

Quetelet was more extreme than Galton, however, in that he believed deviations from the mean were more like small errors, and **regarded the mean as something perfect or ideal**

For Galton, these types were stable from generation to generation -- You can see this in his work on fingerprints or even in his **composite photography**



Lantern slide of composite photograph. Comprises faces of private soldiers collected by Chatham. Leonard Darwin. son of Charles Darwin. C. L. Darwin was in the Royal Engineers and was later President of the Eugenics Society.

It might be expected that when many different portraits are fused into a single one, the result would be a mere smudge. Such, however, is by no means the case... There are then so many traits in common, to combine and to reinforce one another that they prevail to the exclusion of the rest. All that is common remains, all that is individual tends to disappear.

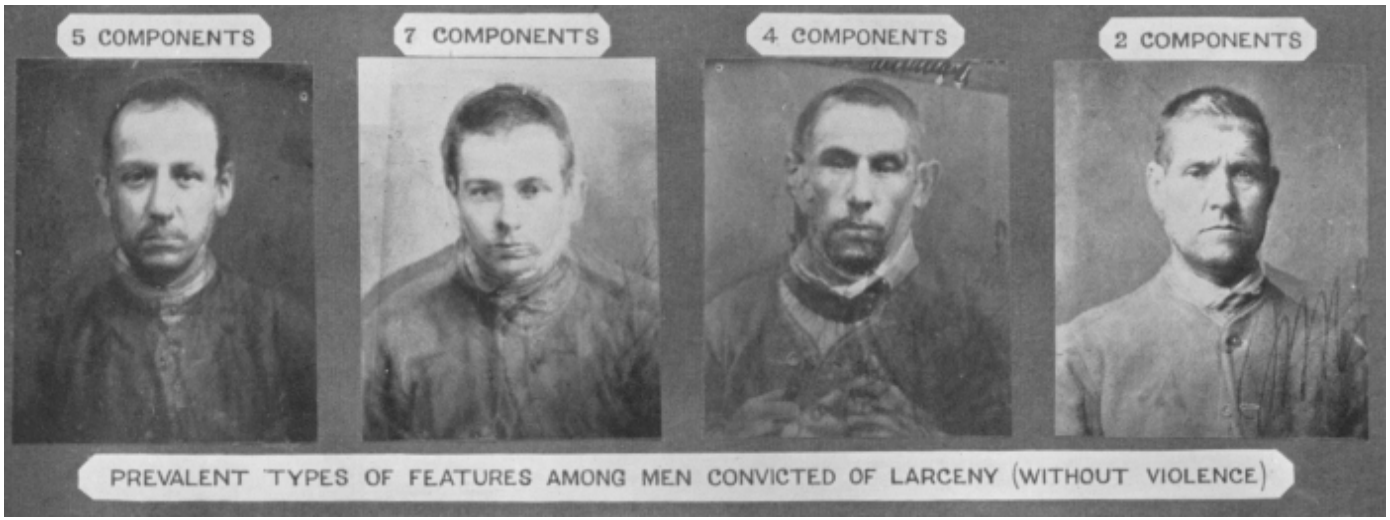
Composite pictures are... the equivalents of those large statistical tables whose totals divided by the number of cases and entered in the bottom line are the averages. They are real generalizations because they include the whole of the material under consideration.

Portraits of Napoleon prepared for making composites exhibited by Francis Galton at the Universal Exhibition, Paris, in 1889. Comprises 5 small photos of different likeness of Napoleon, as he appeared on various coins \medals. The 6th photo is a composite of the other 5.

Lantern slide of family composite photograph. Comprises photos of a father, mother, 2 sons & 2 daughters.

Lantern slide of composite photograph. Comprises 2 horses' heads - Raconteur & St Marnock. 1898

Portraits of three Sisters, full face and profile, with the corresponding Composites.


5 COMPONENTS
7 COMPONENTS
4 COMPONENTS
2 COMPONENTS
PREVALENT TYPES OF FEATURES AMONG MEN CONVICTED OF LARCENY (WITHOUT VIOLENCE)

SPECIMENS OF COMPOSITE PORTRAITURE

PERSONAL AND FAMILY.

Alexander the Great From 6 Different Medals.

Two Sisters.

From 6 Members of same Family Male & Female.

HEALTH.

23 Cases. Royal Engineers. 12 Officers. 11 Privates

DISEASE.

6 Cases

9 Cases

Tubercular Disease

CRIMINALITY.

8 Cases

4 Cases

2 Of the many Criminal Types

CONSUMPTION AND OTHER MALADIES

I

20 Cases

II

36 Cases

56 Cases

Co-composite of I & II

Consumptive Cases.

100 Cases

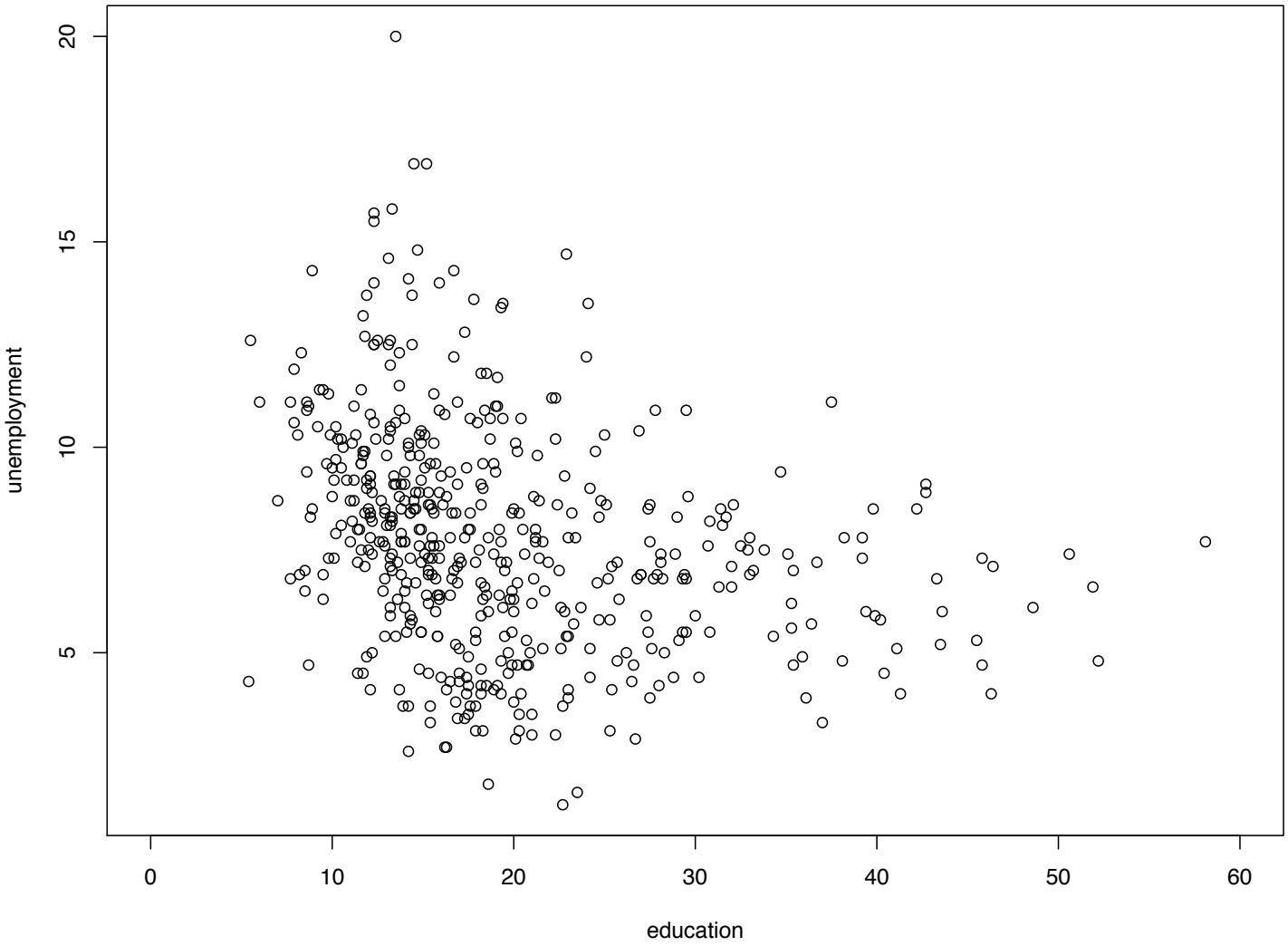
50 Cases

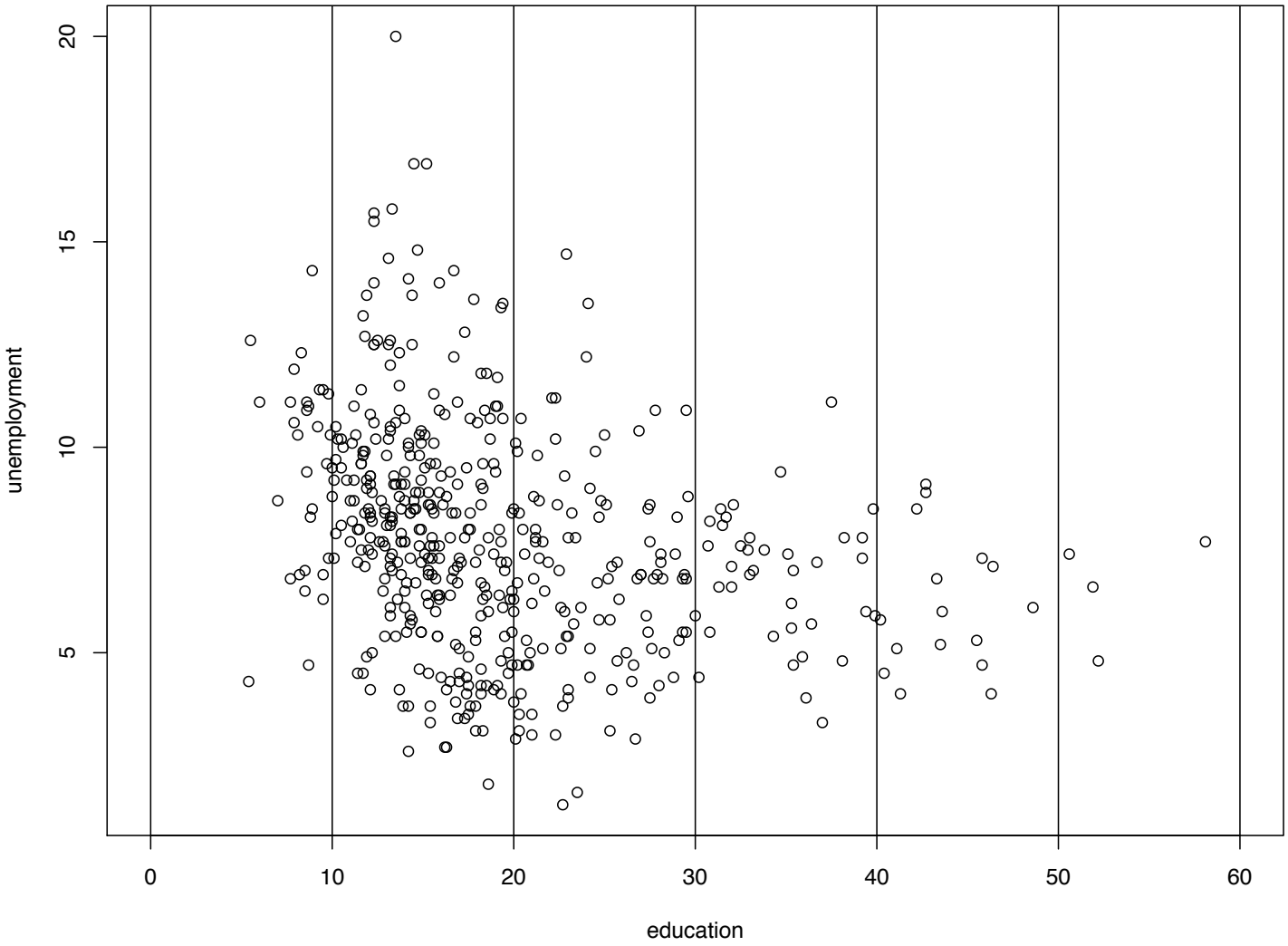
Not Consumptive.

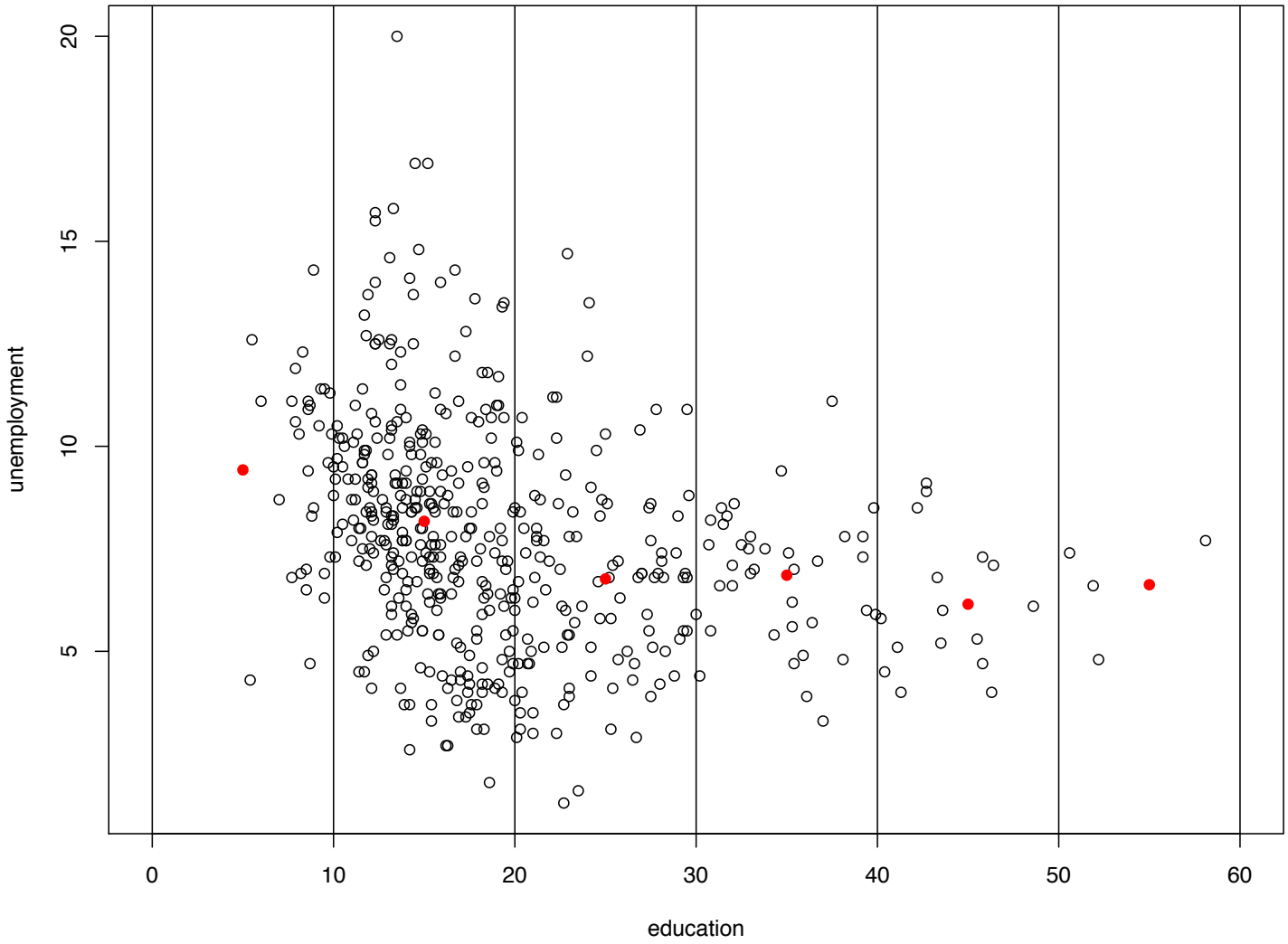
Conditioning

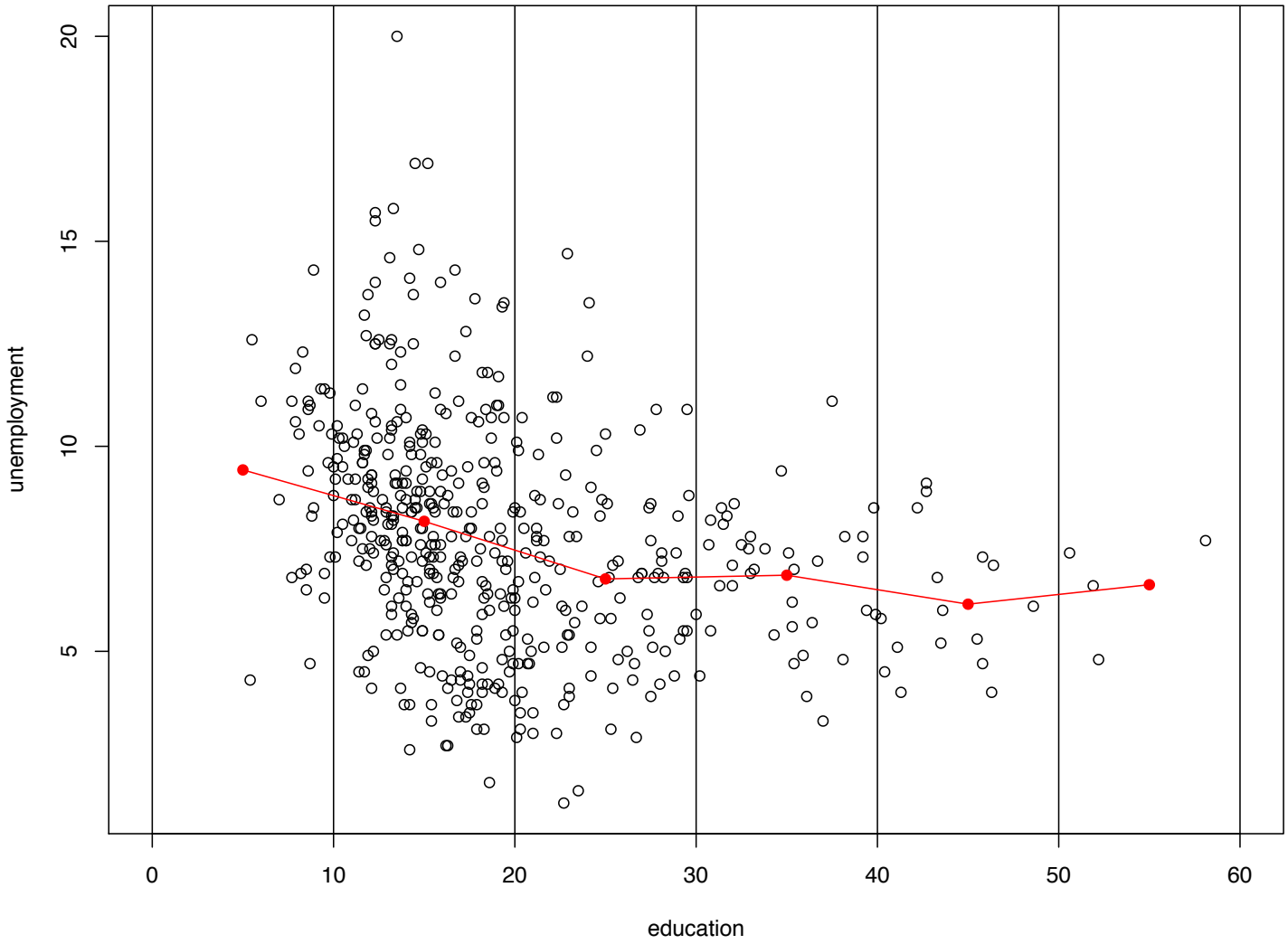
Galton's idea about conditioning (dividing the data into groups and then averaging) can be used in a wide range of settings -- Suppose we go back to our "hardest places to live" data and consider the relationship between education and unemployment

On the next few slides, we carve up the input space into intervals and form averages to give us a sense of the shape of the dependence...


unemployment
5
10
15
20
0
10
20
30
40
50
60
education


unemployment
5
10
15
20
0
10
20
30
40
50
60
education

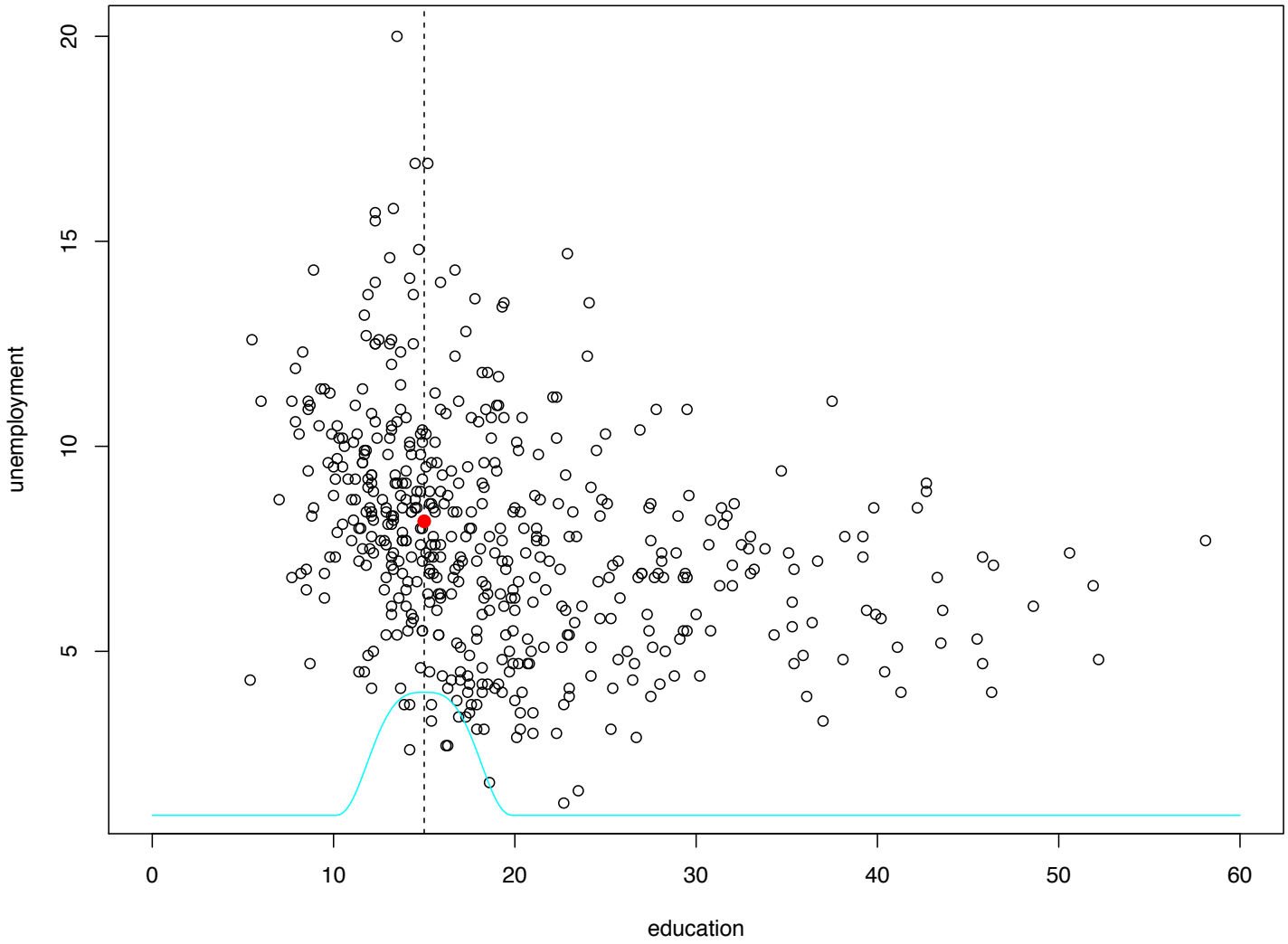

unemployment
5
10
15
20
0
10
20
30
40
50
60
education


unemployment
education
0
10
20
30
40
50
60
5
10
15
20

Smoothing

Rather than make hard decisions about intervals, we can improve this by making predictions using weighted averages instead, where the weights drop off as we get farther from a point we'd like to predict at

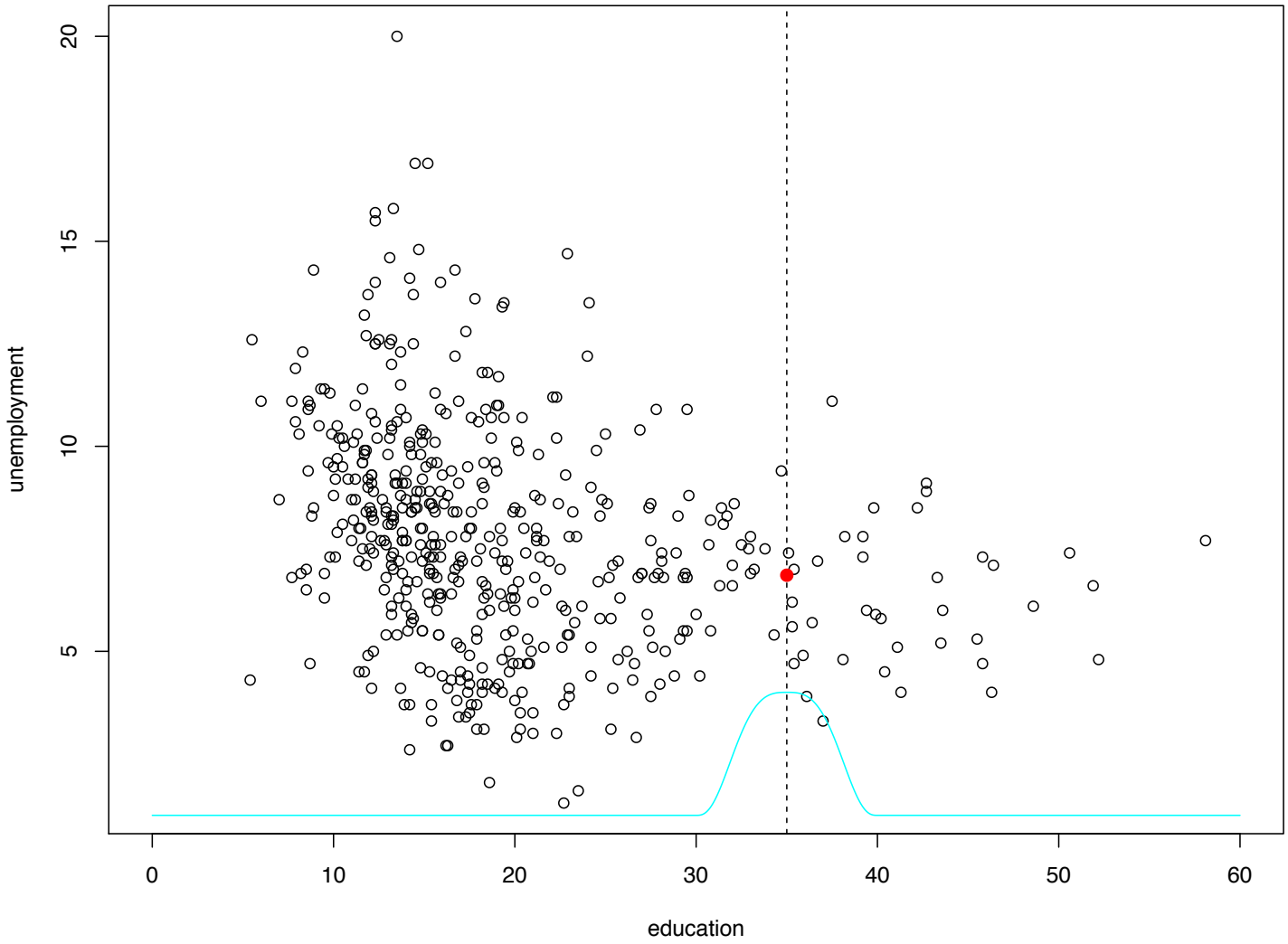
Here is a simple example where we are interested in the vulnerability (the death risk) associated with an graduation rate of 15…


unemployment
20
15
10
5
0
10
20
30
40
50
60
education

Smoothing

What we are doing here is a pretty direct implementation of what Galton was doing with his table of family's heights -- The kernel (the small bump at the bottom of the plot) is used as weights when forming our averages

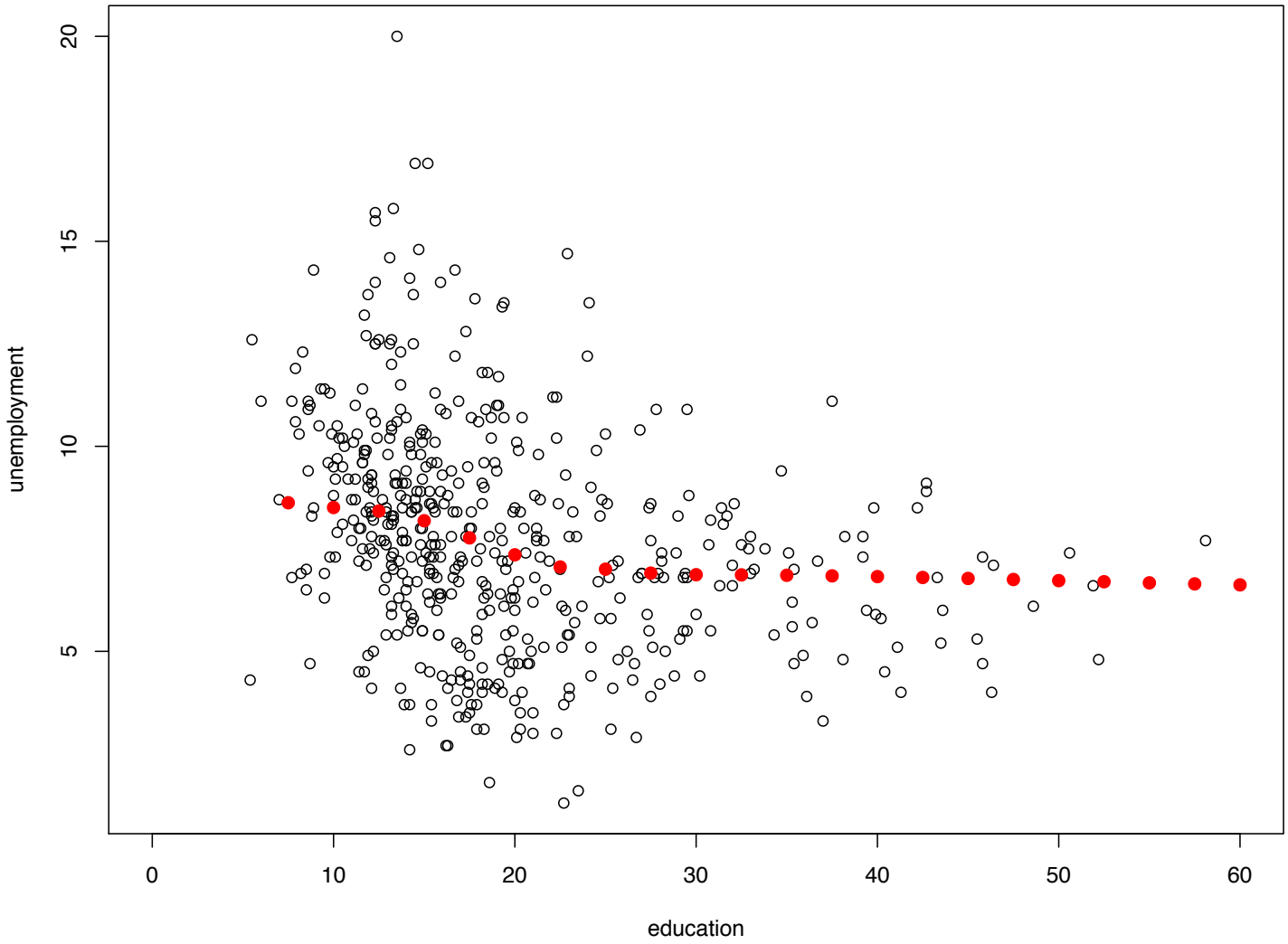
Here is a simple example where we are interested in the vulnerability (the death risk) associated with an education value of 35...

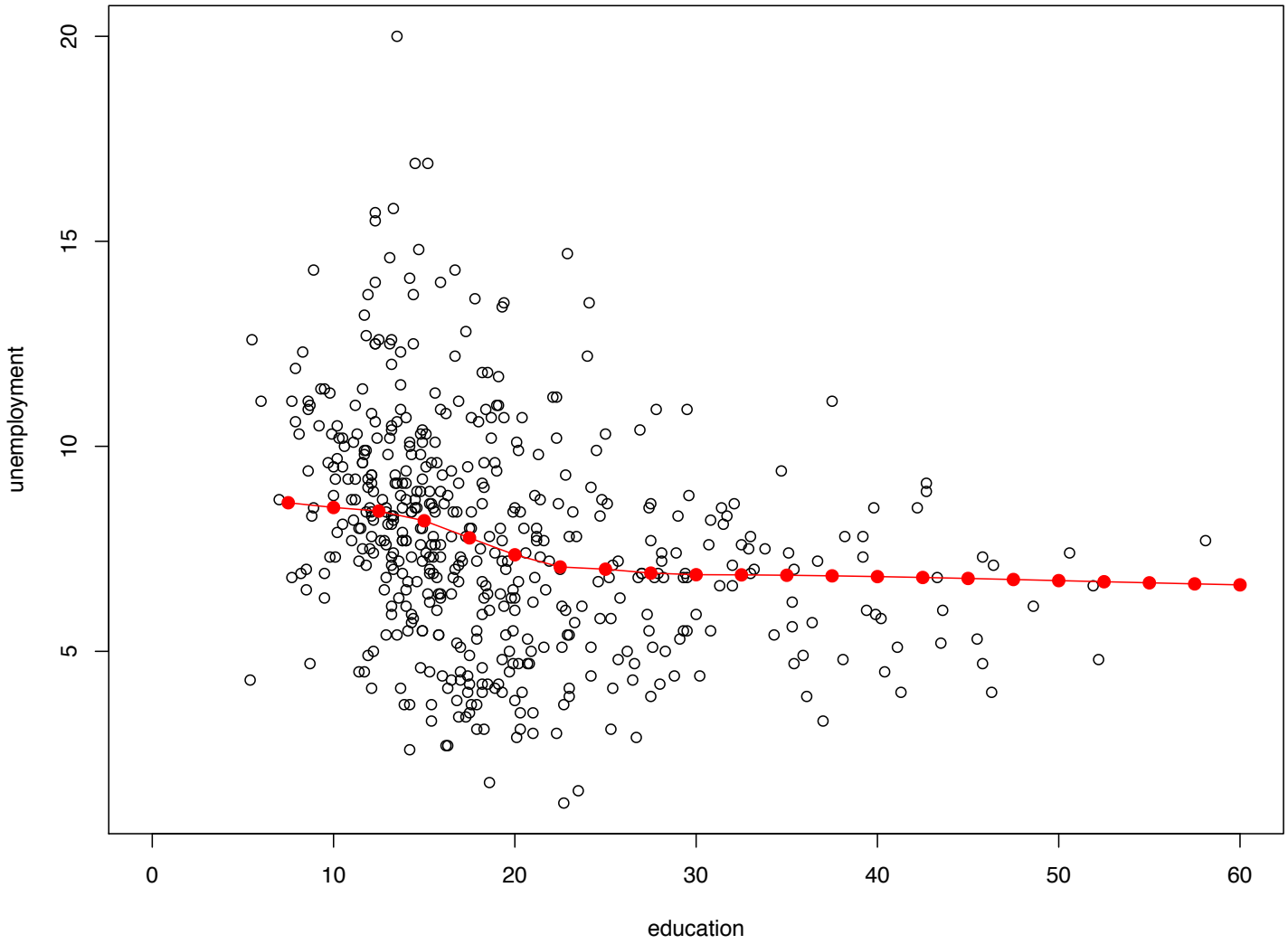

unemployment
5
10
15
20
0
10
20
30
40
50
60
education

Smoothing

If we do this at a number of points, we can start to see a “curve” emerge -- We can then simply connect the dots to come up with a simple model for how unemployment changes with education

Notice that unlike Galton’s case, the solution here is not a line -- It is a curve

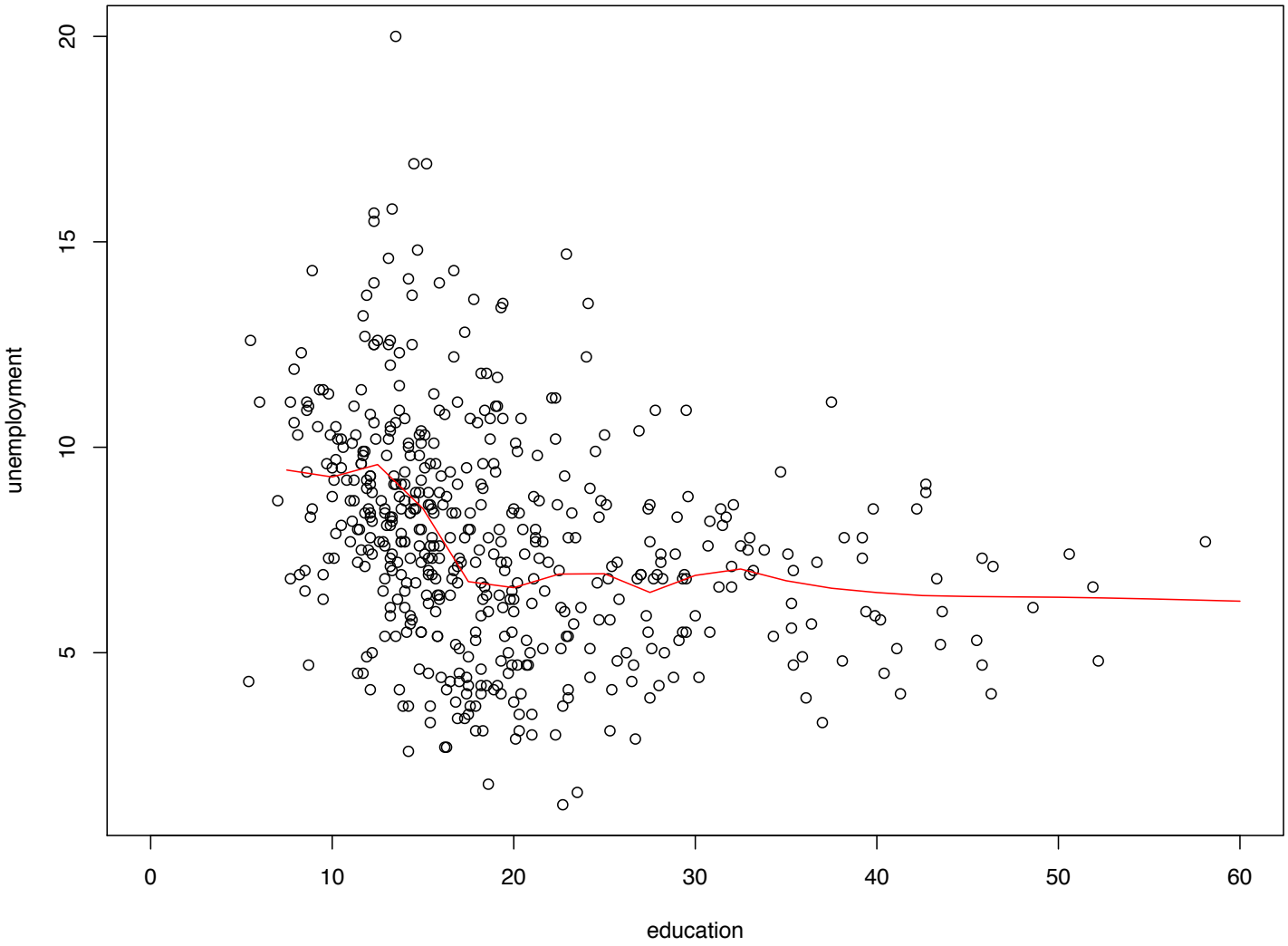

unemployment
5
10
15
20
0
10
20
30
40
50
60
education

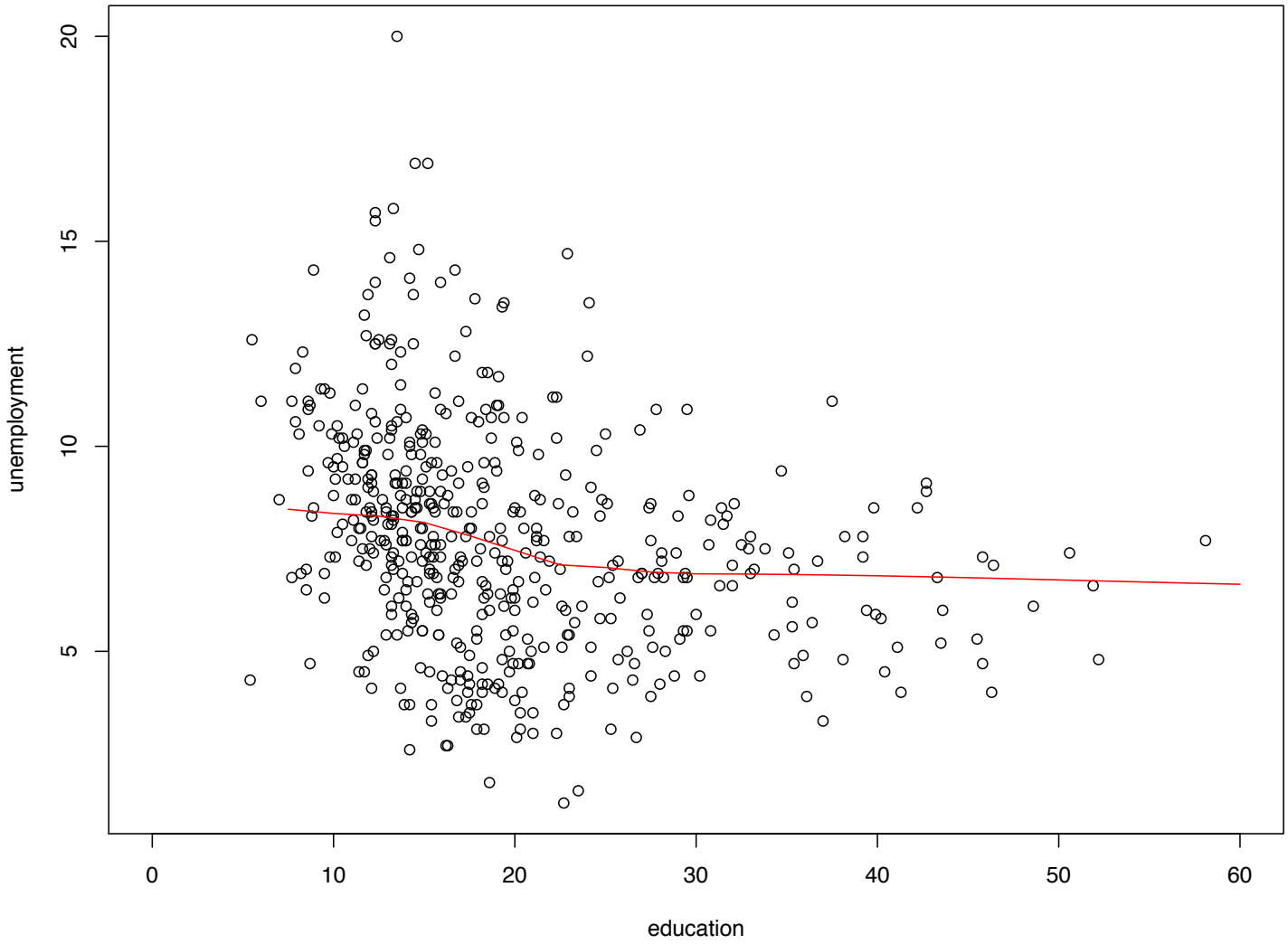

unemployment
education
5
10
15
20
0
10
20
30
40
50
60

Smoothing

The final ingredient here is how wide to make our kernel function —The wider the cyan bump at the bottom of the plots, the more points are included in the average and the flatter our curve becomes (eventually just becoming a constant, the overall mean

If we make it smaller, the curve becomes more wiggly...


unemployment
5
10
15
20
0
10
20
30
40
50
60
education


unemployment
education
5
10
15
20
0
10
20
30
40
50
60

Smoothing

This brings us to one of the essential tradeoffs in modeling -- The balance between **complexity of the object we are creating and the noise that is**

There are many factors that affect the death risk computed for a country and all of those unmeasured variables enter in as “noise” to the system -- **There is a lot of fuzz around the curve we fit**

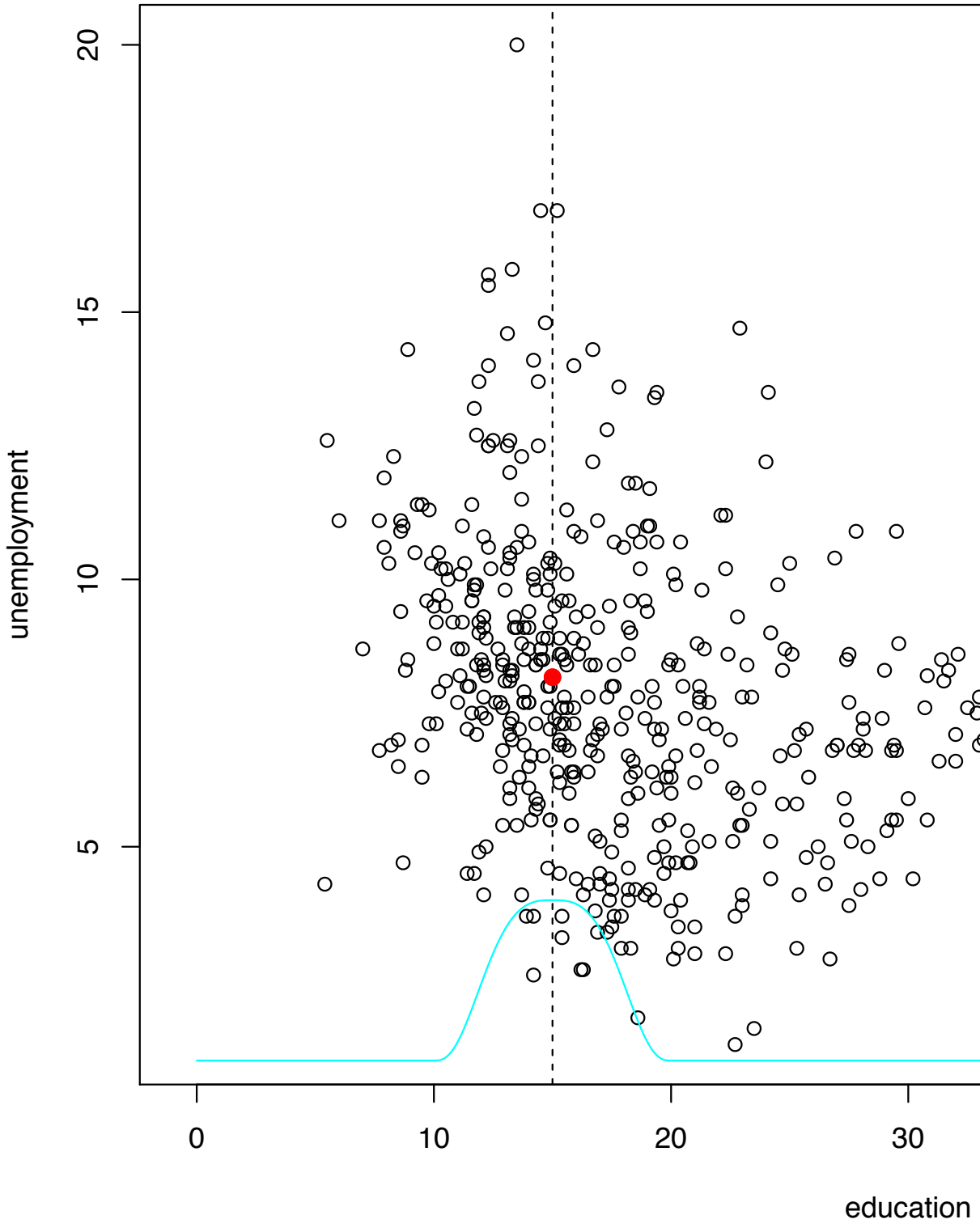
If we make the fit too wiggly, ,we are starting to fit the patterns in the noise -- Not wiggly enough and we might miss structure, like the fact that the curve bends

Smoothing

This kernel idea can be made rigorous if we think about a statistical model of the form

$$Y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

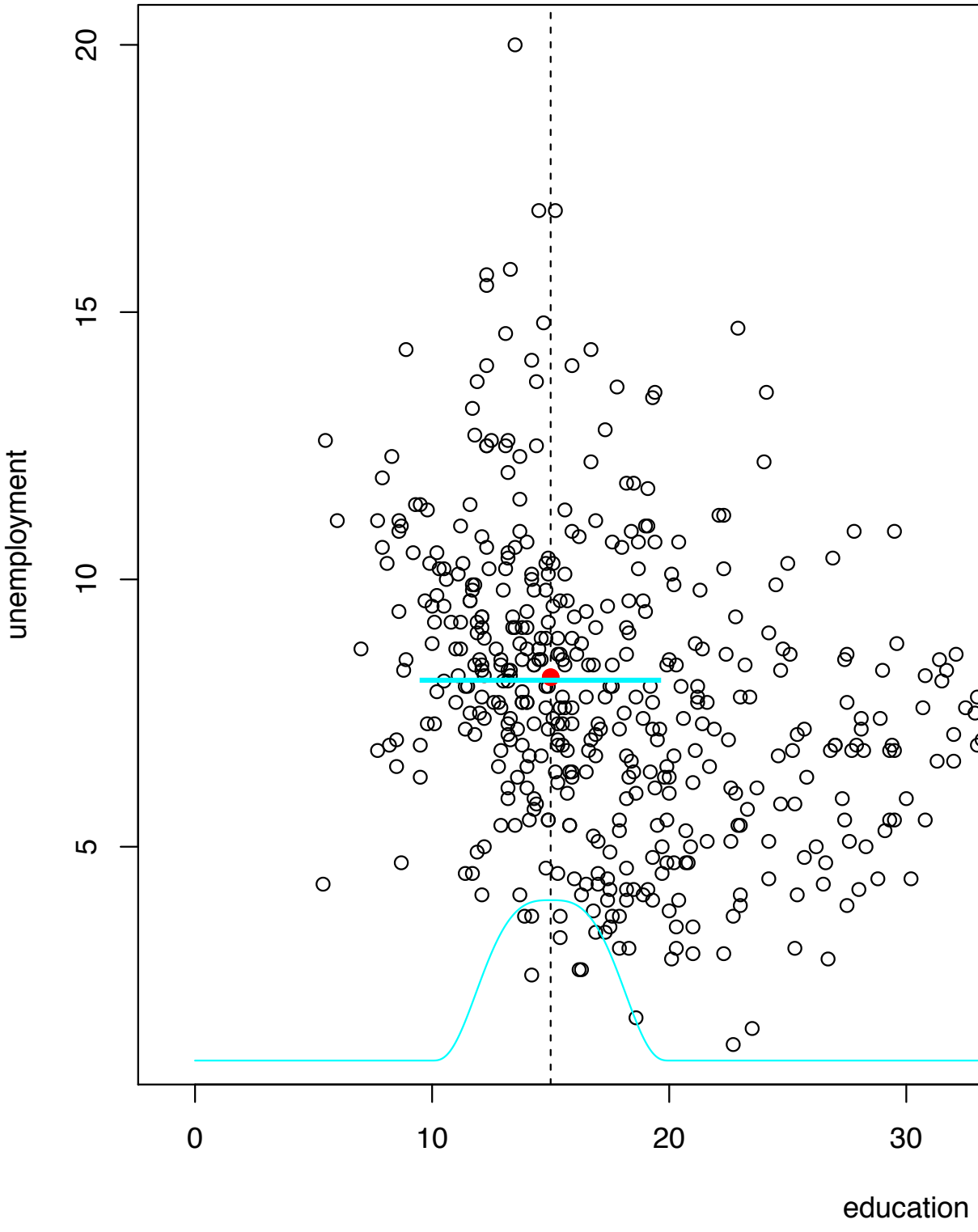
That is, rather than estimate a line like Galton, we will use a (smooth) function to describe our conditional means



Smoothing

For those of you who have had calculus, you might recall that in small neighborhoods, you can describe smooth functions “locally”

Here we are implicitly using a constant to describe the function -- What else might we do?

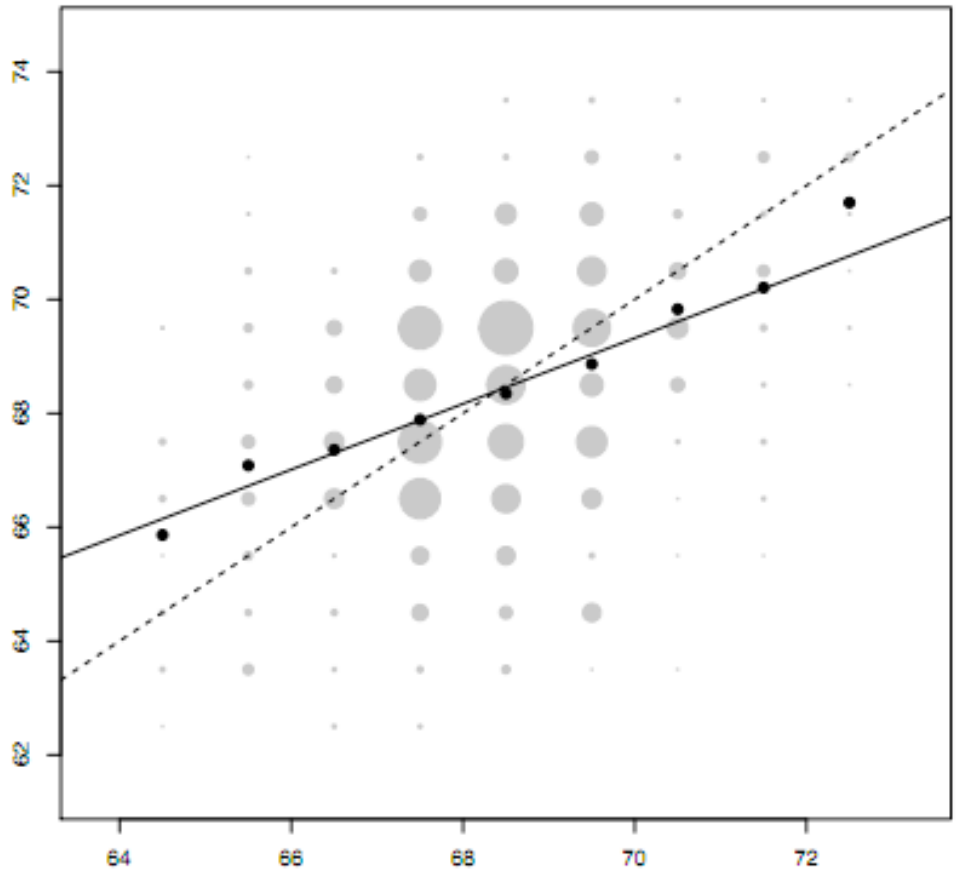


Another view

Galton noticed that the conditional means of children's heights followed (essentially) a straight line

Rather than eyeball this line, we can also apply a rather simple idea to pick the "best" line through the data

In steps least squares...



Ordinary least squares

Given a set of n data points $y_1, \ldots, y_n$ recall that the mean $\bar{y}$ is the unique point that minimizes the squared deviations

$$\sum_{i=1}^{n} (y_i - b)^2$$

Now, by analogy, if our data are pairs $(x_1, y_1), \ldots, (x_n, y_n)$ (say parents' heights x and children's heights y), we can specify any line through the data as $b_0 + b_1 x$ and choose the best by selecting b_0 and b_1 to minimize

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$