Example. Students in a certain kindergarten class are given an IQ test in the fall and then again in the Spring. Researchers want to know if the academic program in this kindergarten helps boost the children's IQ.

(*) The average on both tests is about 100 and both SD s are about 15, so at first glance it seems that a year of kindergarten had no overall effect.

(*) A closer look at the data finds shows that students with high scores on the first test, tended to have lower scores on the second test, on average. Also, students with lower scores on the first test did better, on average, on the second test.

(*) Why?

(*) Because of the *regression effect.*

(*) The data in a scatter plot is (more or less) symmetric around the SD line.

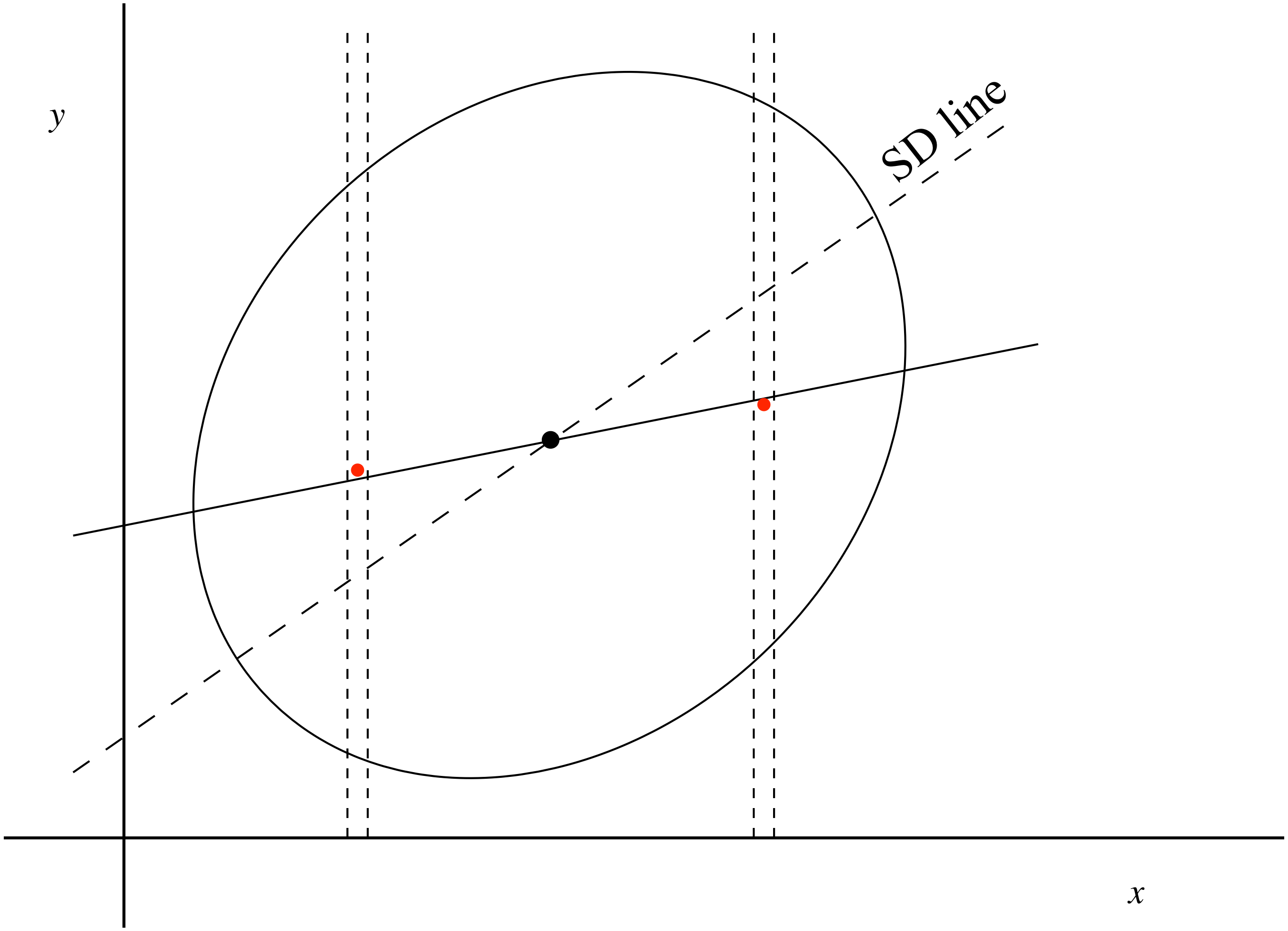
$\Rightarrow$ the SD line increases (or decreases) at the rate of 1 SD_y for every 1 SD_x.

(*) Vertical strips are generally *not* symmetric around the SD line. They are (more or less) symmetric around the regression line.

$\Rightarrow$ the regression line increases (or decreases) at the rate of $r_{xy} \times SD_y$ for every 1 SD_x.

(*) So, the mean score on the second test of students who scored above average on the first test will not be as high as their score on the first test (but still above than average)...

(*) ... and the mean score on the second test of students who scored below average on the first test will be higher than their score on the first test (but still below average).


y
SD line
x

The regression effect – a famous example.

Example: Heights of sons on heights of fathers.

$$\text{average height of fathers} \approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches}$$

$$\text{average height of sons} \approx 69 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches} \quad r \approx 0.5$$

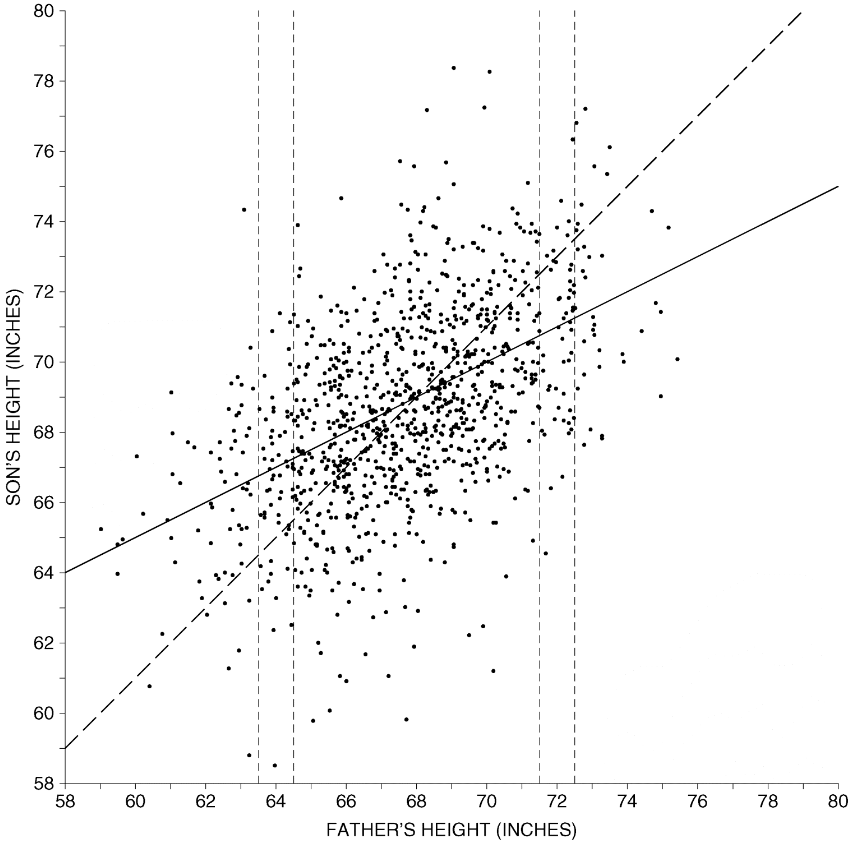


Figure 5., p.171 in FPP, sons and fathers' heights, with SD line and regression line.

- The average heights of the sons for each height class of the fathers follow the regression line, not the SD line.
- The average height of the sons grows more slowly than the height of their fathers.
- Fathers that are much taller than 70 inches, will have sons that are, on average, shorter than them.
- Fathers that are shorter than 70 inches will have sons that are, on average, taller than them.
- The same geometric-logic applies as in the test-retest scenarios: higher than average scores on the first test will be followed by somewhat lower scores on the second test, on average. Likewise, lower than average scores on the first test will be followed by somewhat better scores on the second test, on average.
- The belief that the regression effect is anything more than a statistical fact of life is the ***regression fallacy***.

Where does the regression line come from?

Given a set of paired data, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to find a straight line that predict a y-value as accurately as possible from a known x-value.

(*) Want the observed y-value(s) to be as close as possible to the y-values predicted by the line... *on average*.

(*) If the equation of a line is $\tilde{y} = ax + b$ we want to find the specific values of a and b that make the expression

$$\sqrt{\frac{1}{n}\sum(y_j - \tilde{y}_j)^2} = \text{R.M.S. error of the line}$$

as small as possible.

(*) This problem can be solved using calculus or linear algebra, and it turns out that the line with the smallest R.M.S. error is precisely the regression line.

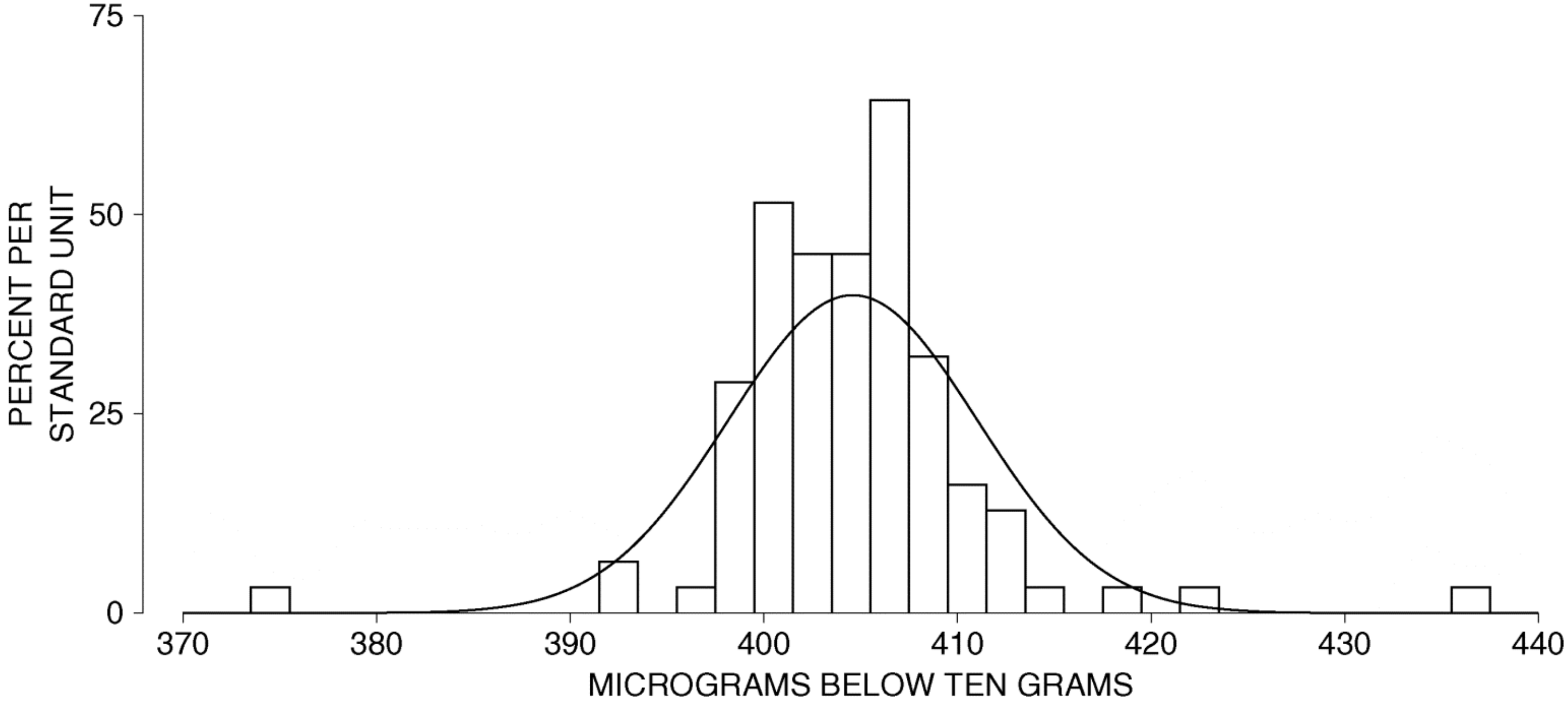
(*) The regression line is also called the *least-squares* line

Measurement error

Table 1. One hundred measurements on NB 10. Almer and Jones, National Bureau of Standards. Units are micrograms below 10 grams.

No.	*Result*	*No.*	*Result*	*No.*	*Result*	*No.*	*Result*
1	409	26	397	51	404	76	404
2	400	27	407	52	406	77	401
3	406	28	401	53	407	78	404
4	399	29	399	54	405	79	408
5	402	30	401	55	411	80	406
6	406	31	403	56	410	81	408
7	401	32	400	57	410	82	406
8	403	33	410	58	410	83	401
9	401	34	401	59	401	84	412
10	403	35	407	60	402	85	393
11	398	36	423	61	404	86	437
12	403	37	406	62	405	87	418
13	407	38	406	63	392	88	415
14	402	39	402	64	407	89	404
15	401	40	405	65	406	90	401
16	399	41	405	66	404	91	401
17	400	42	409	67	403	92	407
18	401	43	399	68	408	93	412
19	405	44	402	69	404	94	375
20	402	45	407	70	407	95	409
21	408	46	406	71	412	96	406
22	399	47	413	72	406	97	398
23	399	48	409	73	409	98	406
24	402	49	404	74	400	99	403
25	399	50	402	75	408	100	404

Histogram of the data in table.



Model for measurement error

$$\boxed{\textbf{Individual measurement} = \textbf{true value} + \textbf{bias} + \textbf{chance error}}$$

- ***Bias*** pushes the results in one direction. I.e., for a given set of measurements, bias (if present) is either always positive or always negative.
- ***Chance error*** is just as likely to be positive as negative.
- The ***true value*** is in many cases *unknown*, and perhaps even unknowable.

  One of the central applications of statistical analysis is to estimate the true value of a given quantity based on a sequence of repeated measurements (or repeated experiments).

  Thus, a key element in the design of both experiments and observational studies is to ***minimize the bias*** *as much as possible.*

Bias in sample surveys:

1. The Literary Digest Poll of 1936

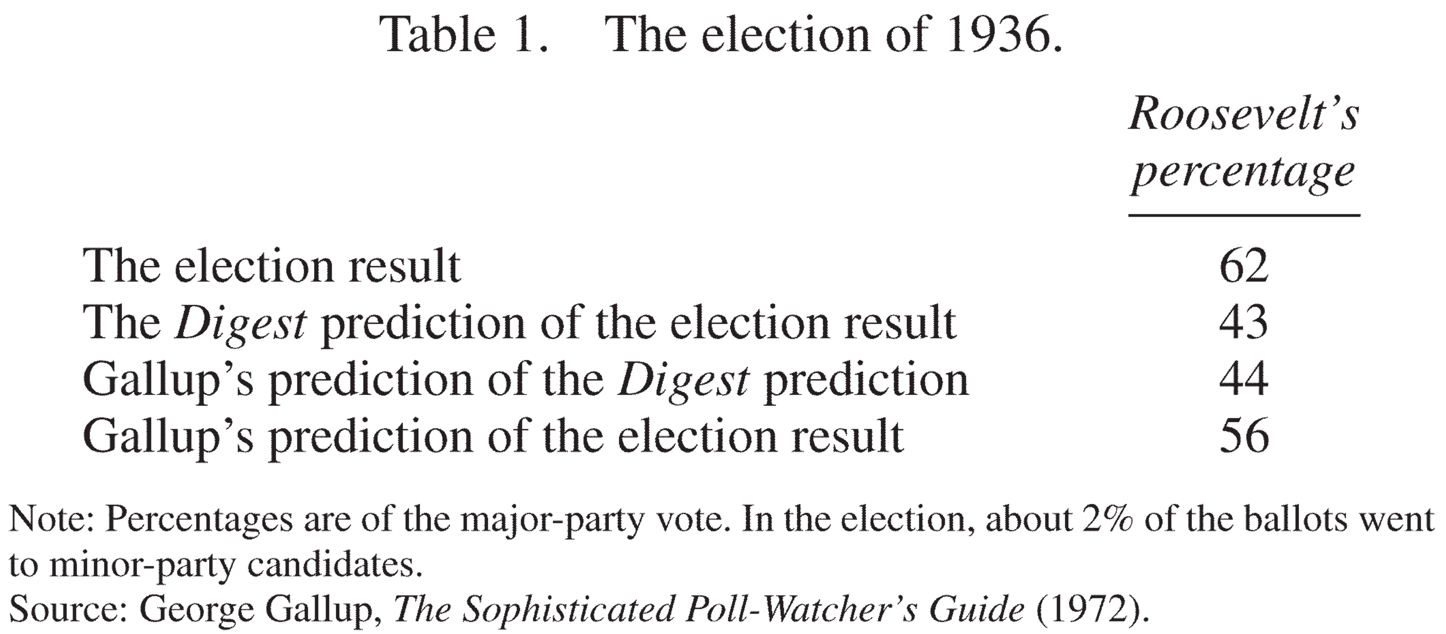


Table 1. The election of 1936.

	Roosevelt's percentage
The election result	62
The *Digest* prediction of the election result	43
Gallup's prediction of the *Digest* prediction	44
Gallup's prediction of the election result	56

Note: Percentages are of the major-party vote. In the election, about 2% of the ballots went to minor-party candidates.
Source: George Gallup, *The Sophisticated Poll-Watcher's Guide* (1972).