UNIVERSITY OF LINCOLN

SCHOOL OF MATHEMATICS
AND PHYSICS

MTH1005
PROBABILITY AND STATISTICS

Semester B
Lecture 8 (12/3/2024)

Danilo Roccatano

Office: INB 3323
Email: droccatano@lincoln.ac.uk

SUMMARY LAST LESSON

We have

- defined the covariance of two random variables X and Y
- examined the meaning of the covariance
- shown that Cov(X, Y) = 0 for independent random variables.
- defined the correlation between two random variables X and Y.

The Simple Linear Regression Model

SIMPLE LINEAR REGRESSION MODEL

- The simplest deterministic mathematical relationship between two variables *x* and *y* is a linear relationship

$$y = a_0 + a_1 x$$

- The set of pairs (*x*, *y*) for which $y = a_0 + a_1 x$ determines a straight line with slope a_1 and *y*-intercept a_0. The objective of this section is to develop a linear probabilistic model.

- If the two variables are not deterministically related, then for a fixed value of *x*, there is uncertainty in the value of the second variable.

SIMPLE LINEAR REGRESSION MODEL

- More generally, the variable whose value is fixed by the experimenter will be denoted by *x* and will be called the **independent, predictor,** or **explanatory variable.**
- For fixed *x*, the second variable will be random; we denote this random variable and its observed value by *Y* and *y*, respectively, and refer to it as the **dependent** or **response variable.**
- Usually observations will be made for a number of settings of the independent variable.

SIMPLE LINEAR REGRESSION MODEL

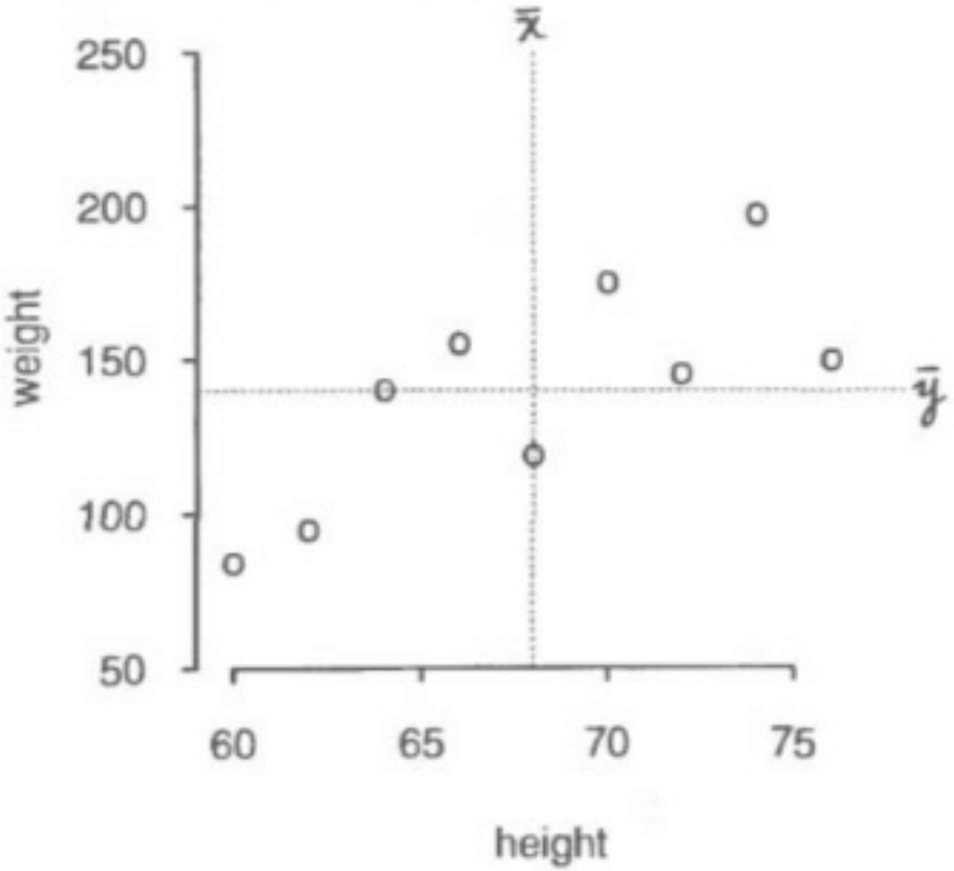
- Let $x_1, x_2, \ldots, x_n$ denote values of the independent variable for which observations are made, and let Y_i and y_i, respectively, denote the random variable and observed value associated with x_i. The available bivariate data then consists of the n pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

- A picture of this data called a **scatter plot** gives preliminary impressions about the nature of any relationship. In such a plot, each (x_i, y_i) is represented as a point plotted on a two-dimensional coordinate system.

SIMPLE LINEAR REGRESSION MODEL

We want to correlate the weight with the height of a sample of students

HEIGHT	WEIGHT
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



From: Gonick & Smith The cartoon guide to Statistics

SIMPLE LINEAR REGRESSION MODEL

Given our assumption of a straight-line

$$y_i = a_0 + a_1 x_i + e_i$$

the error at each point is given by

$$e_i = y_i - a_0 + a_1 x_i$$

So, in some sense the some of the total errors would be given by the sum of the errors. We will take the sum of the squares of the errors

$$S_r = \sum_{i=0}^{n-1} e_i^2 = \sum_{i=0}^{n-1} (y_i - a_0 - a_1 x_i)^2$$

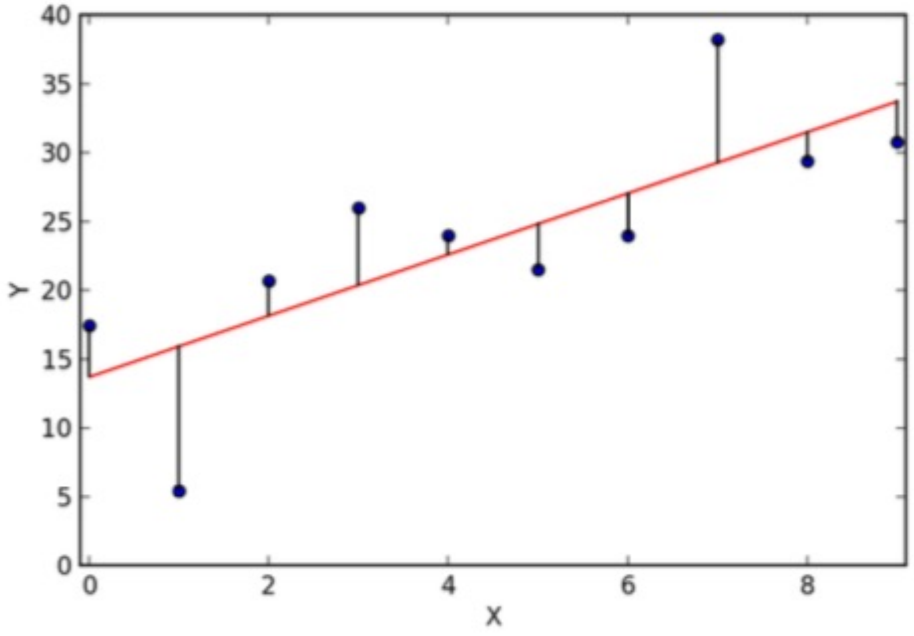
as our error criterion.

SIMPLE LINEAR REGRESSION MODEL

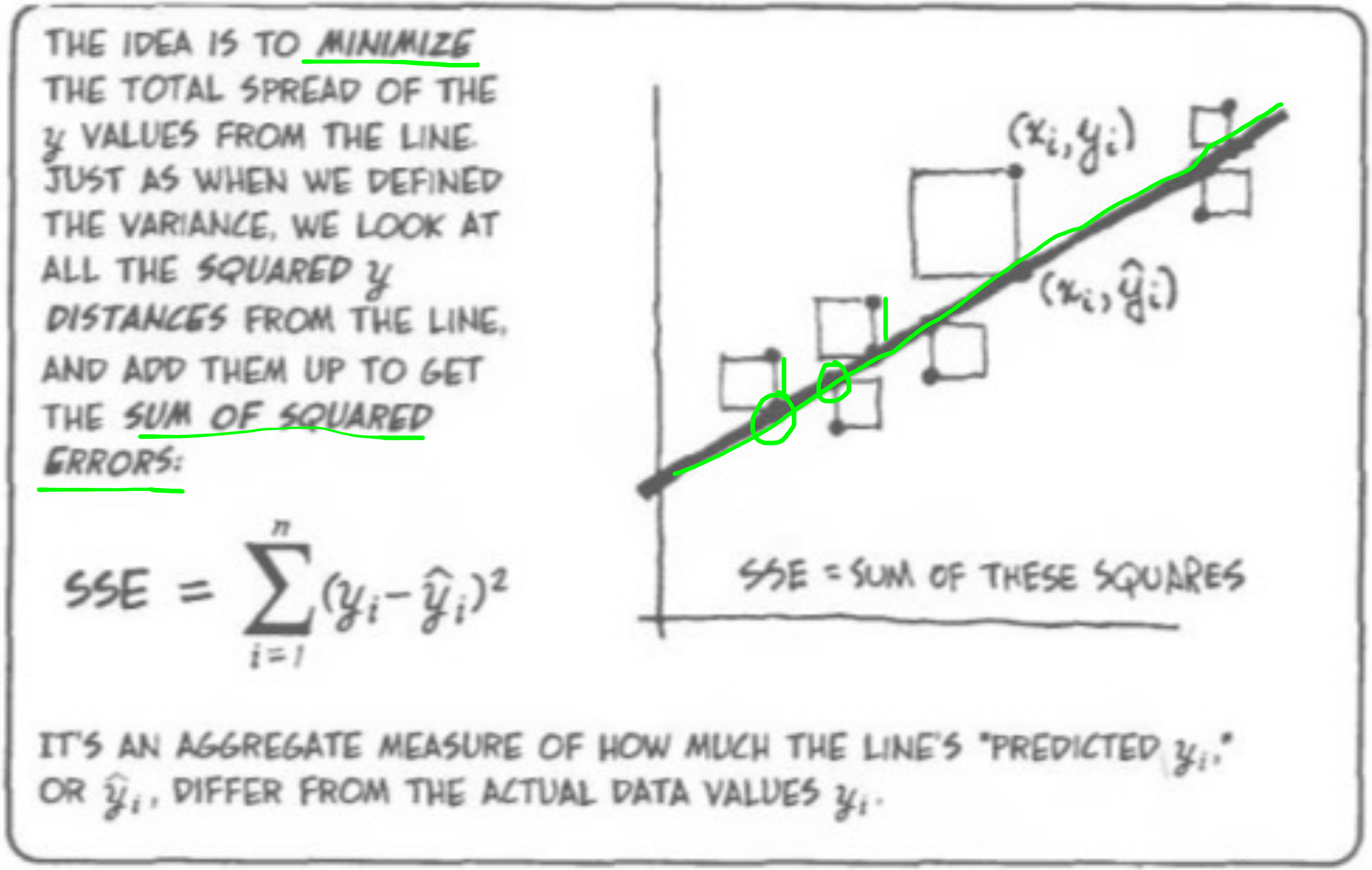
So in some sense the some of the total errors would be given by the sum of the   errors. We will take the sum of the squares of the errors

$$S_r = \sum_{i=0}^{n-1} e_i^2 = \sum_{i=0}^{n-1} (y_i - a_0 - a_1 x_i)^2$$

as our error criterion.



SIMPLE LINEAR REGRESSION MODEL



THE IDEA IS TO *MINIMIZE* THE TOTAL SPREAD OF THE y VALUES FROM THE LINE. JUST AS WHEN WE DEFINED THE VARIANCE, WE LOOK AT ALL THE *SQUARED* y *DISTANCES* FROM THE LINE, AND ADD THEM UP TO GET THE *SUM OF SQUARED ERRORS*:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

IT'S AN AGGREGATE MEASURE OF HOW MUCH THE LINE'S "PREDICTED y_i," OR $\hat{y}_i$, DIFFER FROM THE ACTUAL DATA VALUES y_i.

From: Gonick & Smith The cartoon guide to Statistics

DETERMINATION OF THE OPTIMAL PARAMETERS

If we look at our model

$$S_r = \sum_{i=0}^{n-1} e_i^2 = \sum_{i=0}^{n-1} (y_i - \underline{a_0} - \underline{a_1} x_i)^2$$

we see that there are 2 parameters, a_0 and a_1 that control the slope and intercept of our model.

It is a model, we are assuming that there is a linear relationship between x and y.

We want to minimize the value of S_r , so we differentiate with respect to our parameters

DETERMINATION OF THE OPTIMAL PARAMETERS

$$\frac{\partial S_r}{\partial a_0} = -2\sum(y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum[(y_i - a_0 - a_1 x_i)x_i]$$

where the summations go from 0 to n − 1 .

Note that the points (x_i , y_i) are not variables, they are things we have measured. What we can vary is the parameters of our model. So S_r is a function of the two parameters a_0 and a_1.

For more general models we will have more parameters and a more complex relations ship than the straight line assumed here.

DETERMINATION OF THE OPTIMAL PARAMETERS

We want to minimize the value of Sr, so we differentiate with respect to our parameters

$$\frac{\partial S_r}{\partial a_0} = -2\sum(y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum[(y_i - a_0 - a_1 x_i)x_i]$$

This gives us a pair of simultaneous linear equations, sometimes called the normal equations.

We can solve these for a_1 and a_0.

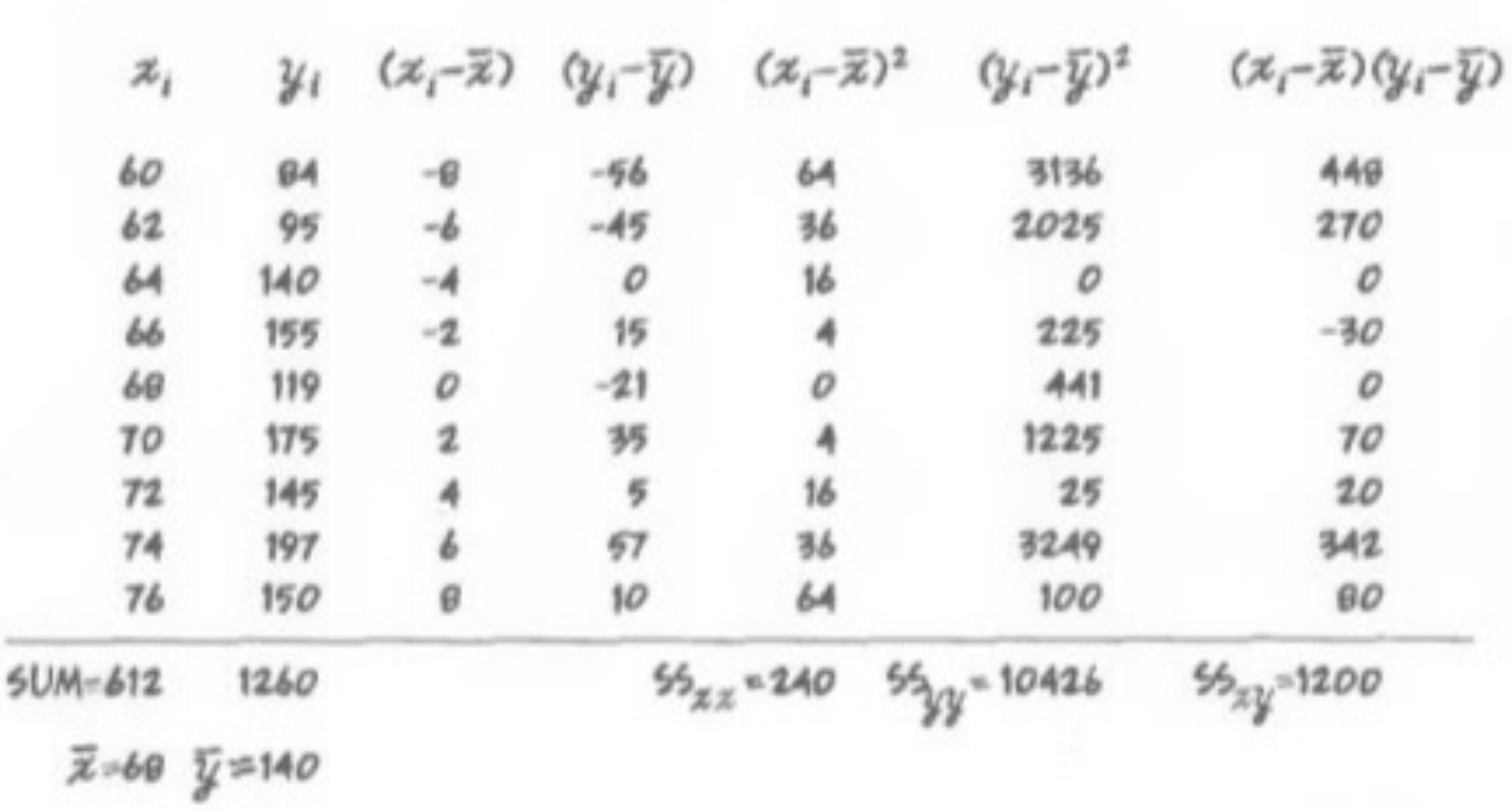
$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

and plug this into the first equation to get

$$a_0 = \frac{\sum y_i}{n} - a_1\frac{\sum x_i}{n} = \bar{y} - a_1\bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the means of the x and y values.

DET. OF THE OPTIMAL PARAMETERS: THE EXAMPLE



x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
SUM=612	1260			$SS_{xx}=240$	$SS_{yy}=10426$	$SS_{xy}=1200$
$\bar{x}=68$	$\bar{y}=140$					

OPTIMAL PARAMETERS: THE EXAMPLE

$$y = a + \underline{b}x$$

WHERE

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

AND

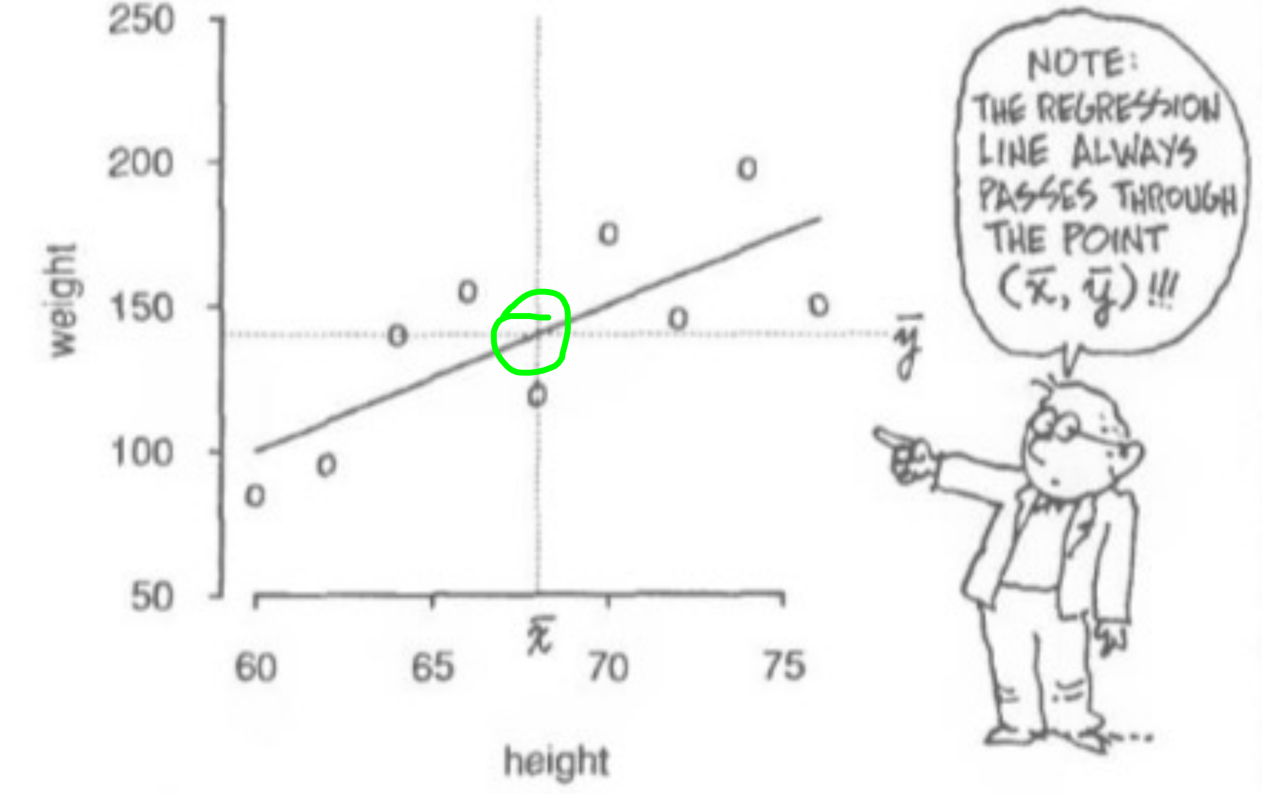
$$a = \bar{y} - b\bar{x}$$

(HERE $\bar{x}$ AND $\bar{y}$ ARE THE MEANS OF $\{x_i\}$ AND $\{y_i\}$ RESPECTIVELY.)

WHICH GIVES VALUES OF a AND b:

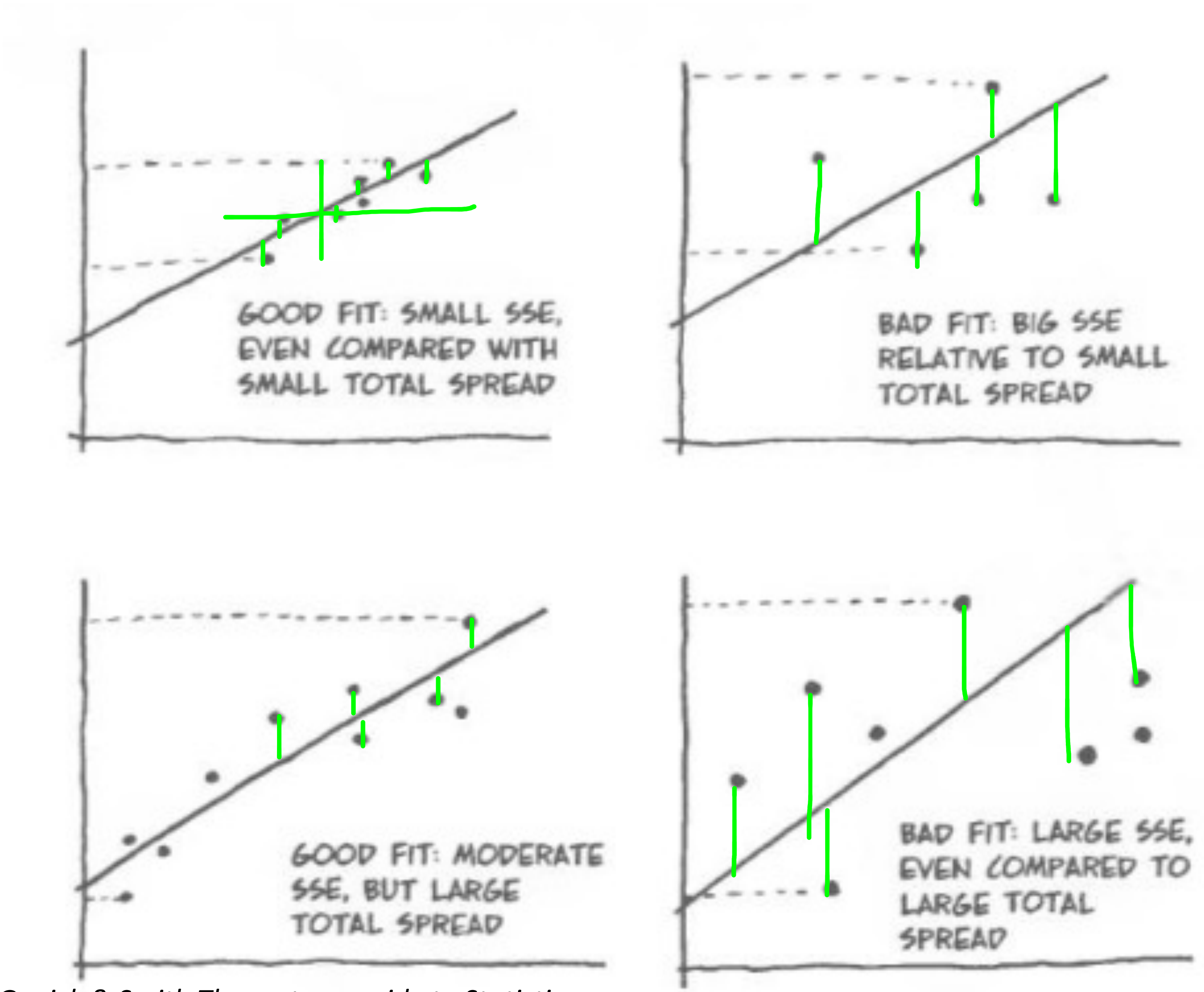
$$\underline{b = \frac{1200}{240} = 5} \qquad \underline{a = \bar{y} - b\bar{x} = 140 - 5(68)} = -200$$

SO $\boxed{y = -200 + 5x}$



From: Gonick & Smith The cartoon guide to Statistics

QUANTIFYING THE ERROR


GOOD FIT: SMALL SSE, EVEN COMPARED WITH SMALL TOTAL SPREAD
BAD FIT: BIG SSE RELATIVE TO SMALL TOTAL SPREAD
GOOD FIT: MODERATE SSE, BUT LARGE TOTAL SPREAD
BAD FIT: LARGE SSE, EVEN COMPARED TO LARGE TOTAL SPREAD

From: Gonick & Smith The cartoon guide to Statistics

QUANTIFYING THE ERROR

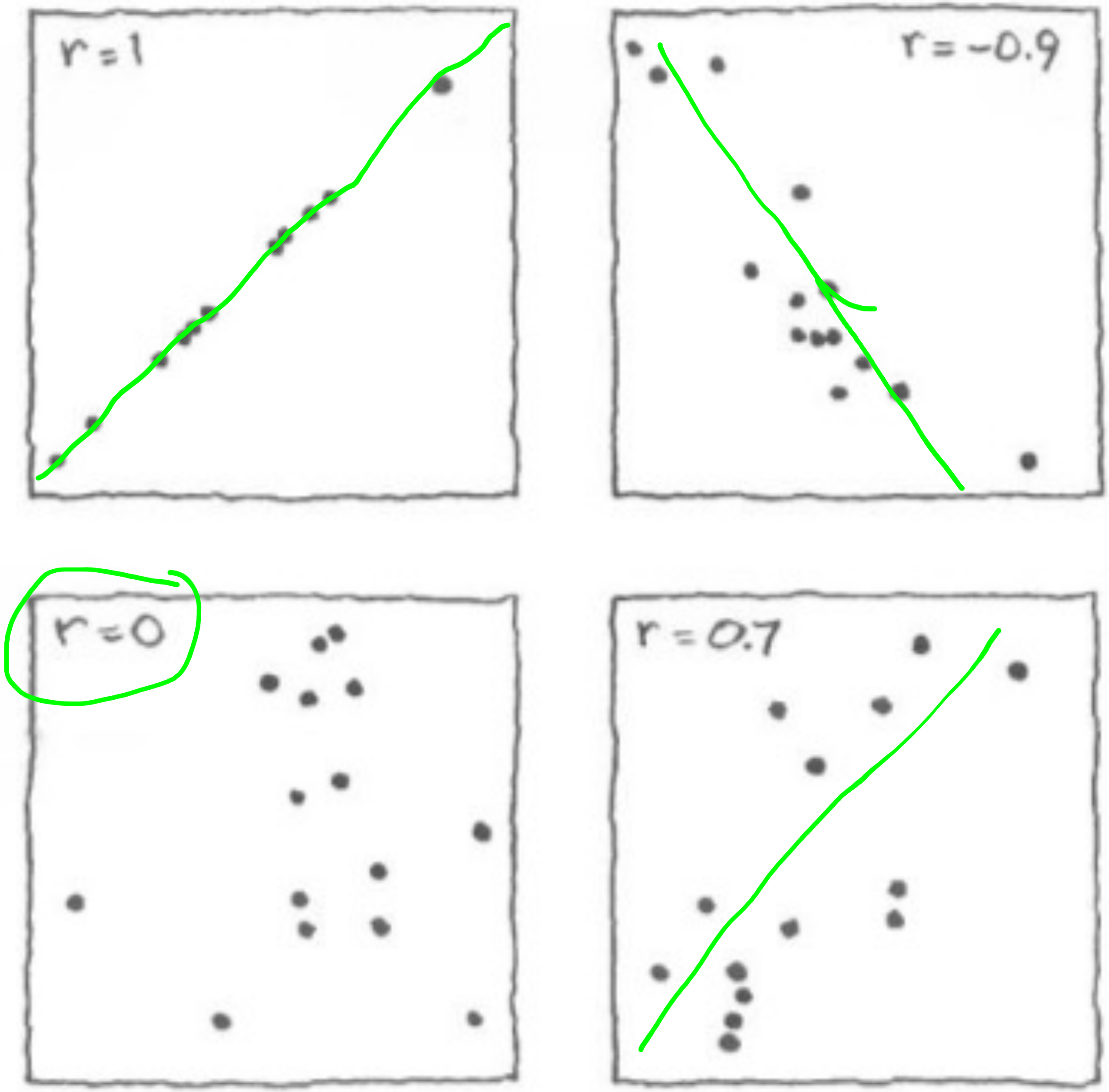
To quantify the error in our model we can look at the **correlation coefficient (r)** of our data, conventionally called **r.**

This is the covariance of the data (x, y) divided by the square roots of the variance of x and y.

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

The measures the difference between the scatter of the data from the mean and the scatter of the data from the ideal straight line.

QUANTIFYING THE ERROR



From: Gonick & Smith The cartoon guide to Statistics

MEASURES OF CENTRAL TENDENCY - REVIEW

Where is the middle of the distribution?

We have looked at the mean, which we saw is also termed the expectation value of a random variable.

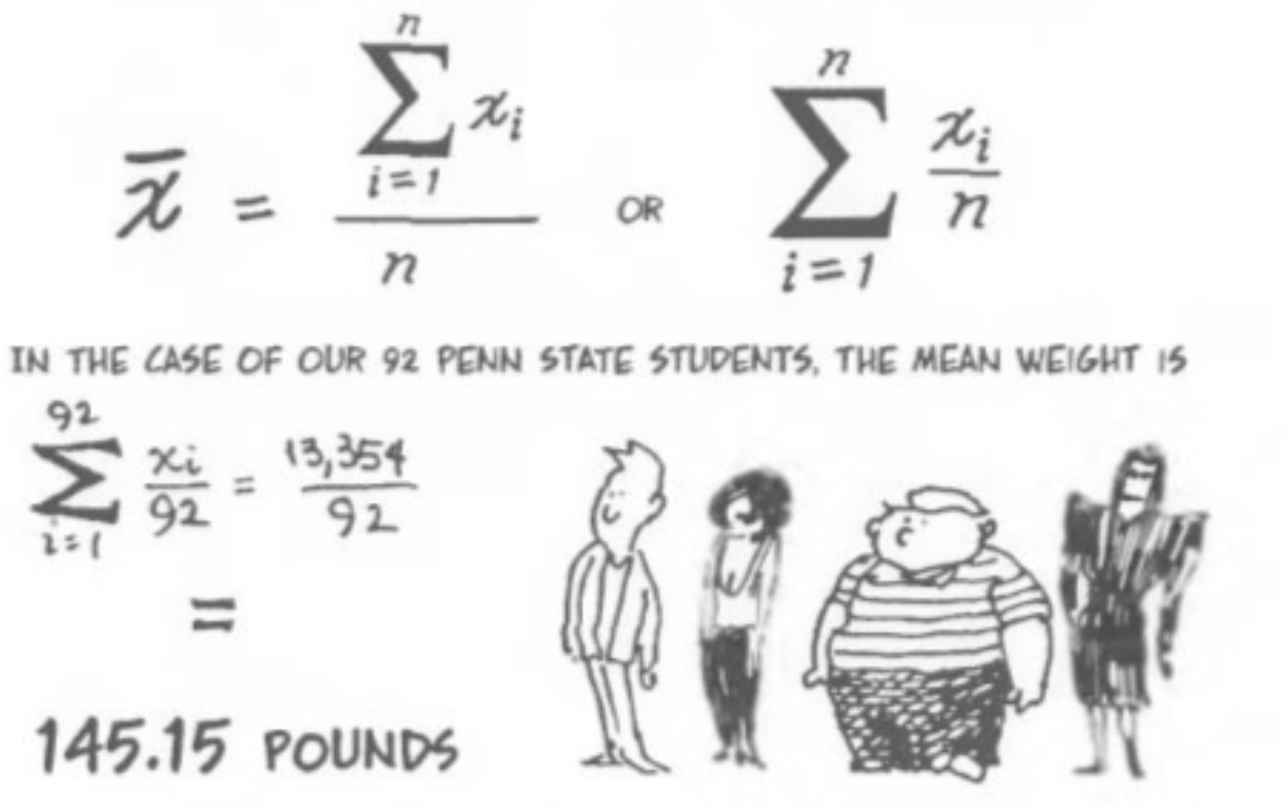
There are two other common measures of the centre of the distribution

- **Mean** *of data or expectation value of a random variable*
- **Mode**, *which is the most commonly occurring value in the distribution (may not be unique)*
- **Median**, *which is middle value of the distribution if values are listed in order, for an odd number of values. Or the mean of the two central values, for even numbers of data points.*

These measures are often applied to samples, which may or may not come from some idealised probability distribution.

RANDOM VARIABLES: Properties of the random variable

Mean or Expectation value


$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ OR $\sum_{i=1}^{n} \frac{x_i}{n}$
IN THE CASE OF OUR 92 PENN STATE STUDENTS, THE MEAN WEIGHT IS
$\sum_{i=1}^{92} \frac{x_i}{92} = \frac{13{,}354}{92}$
=
145.15 POUNDS

We will see that the means is also called **expectation value of the random variable** and it is indicated ad E[X]

From: Gonick & Smith The cartoon guide to Statistics

MEDIAN FOR DISCRETE DATA

For a discrete set of data the median is easily defined:

median is middle value of the distribution: if values are listed in order, for an odd number of values. Or the (arithmetic) mean of the two central values, for even numbers of data points.

(1, 2, 3, 4, 5)

has median 3, the central value.

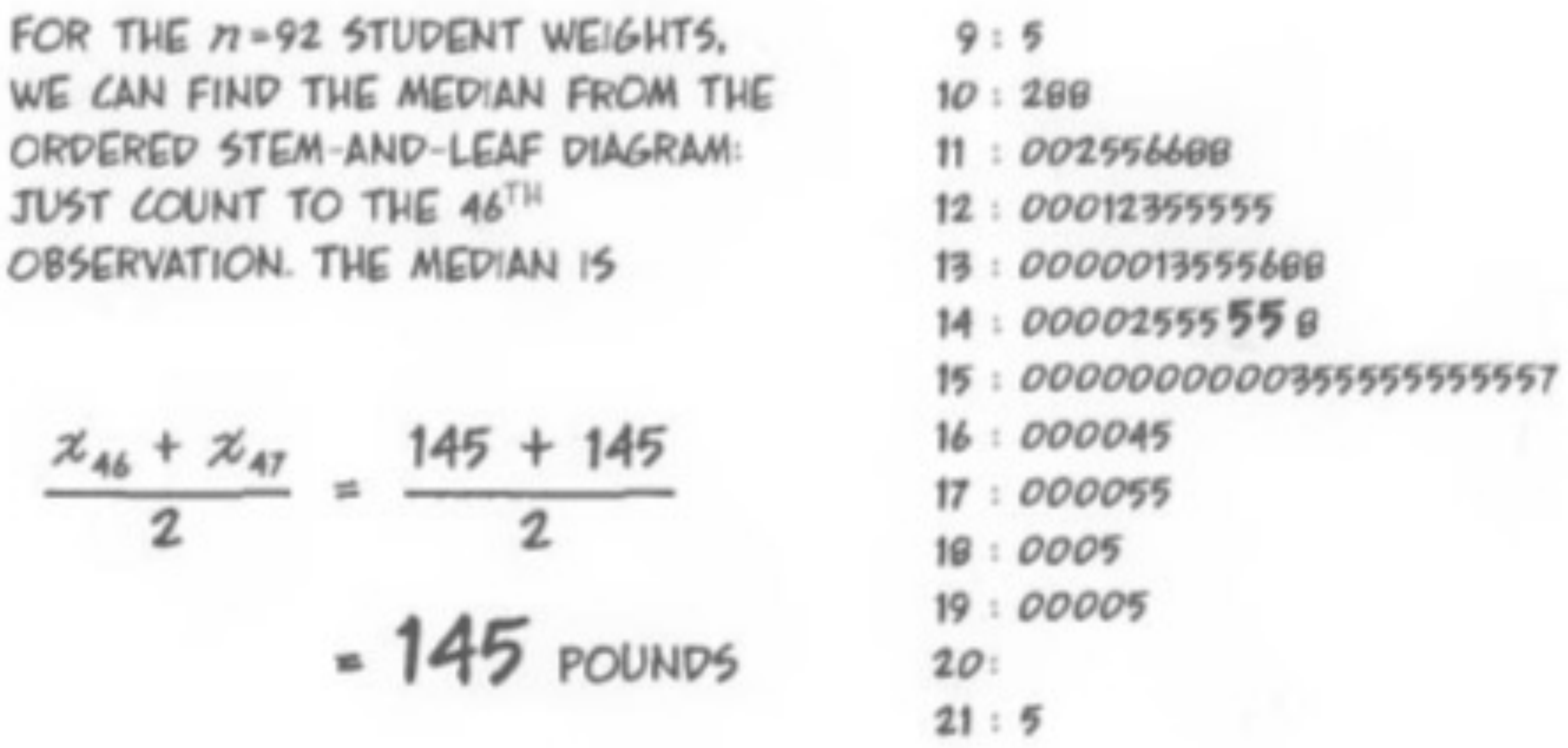
(1, 2, 3, 4, 5, 6)

has median 3.5 - the mean of 3 and 4.

Algorithmically, we sort the list of numbers, check whether there are an odd or even number, then select the central value, or the central two and average. We could also think of this as defining a partition that cuts the data in to two pieces, of equal weight i.e. with equal numbers of data in each partition.

RANDOM VARIABLES: Properties of the random variable

Median



FOR THE $n = 92$ STUDENT WEIGHTS, WE CAN FIND THE MEDIAN FROM THE ORDERED STEM-AND-LEAF DIAGRAM: JUST COUNT TO THE 46TH OBSERVATION. THE MEDIAN IS

$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$

$$= 145 \text{ POUNDS}$$

```
 9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 00002555558
15 : 0000000000035555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5
```

MEDIAN OF A DISCRETE RANDOM VARIABLE

We could also think of this as defining a partition that cuts the data in to two pieces, of equal weight i.e. with equal numbers of data in each partition.

The median is the value where it is likely that the variable would take a value greater than or equal, i.e. P $(Z \geq z_{median}) \geq 0.5$, or less than or equal to , i.e. P $(Z \leq z_{median}) \geq 0.5$.

Remember that the cumulative distribution function of a discrete random variable Z is given by

$$F_Z(a) = P(Z \leq a) = \sum_{z \leq a} p_Z(z)$$

If we look back at the list (1, 2, 3, 4, 5) and convert it to a random variable - say that each value is equally likely, we get

$$p_Z = \frac{1}{5}, \qquad z = 1,2,3,4,5$$

MEDIAN OF A DISCRETE RANDOM VARIABLE

Now

$$F_Z(z) = \begin{cases} 0, & z < 1 \\ \frac{1}{5}, & 1 \leq z < 2 \\ \frac{2}{5}, & 2 \leq z < 3 \\ \frac{3}{5}, & \boxed{3} \leq z < 4 \\ \frac{4}{5}, & 4 \leq z < 5 \\ 1, & z \geq 5 \end{cases}$$

We can deduce that the median must be 3:

$$P(Z \leq 3) = F(3) = \tfrac{3}{5}.$$

The other equality is a bit more tricky

$$P(Z \geq 3) = \sum_{z \geq 3} p_Z(z) = p_Z(3) + p_Z(4) + p_Z(5) = \frac{3}{5}$$

MEDIAN FOR CONTINUOUS DATA

The computation of the **median for a continuous random variable** is more interesting.
Recall cumulative distribution function of a random variable gives the probability that the random variable will take on a value equal to, or smaller than, the argument of the function. For a continuous random variable X

$$F_X(a) = \int_{-\infty}^{a} f(x)dx = P(X \leq a)$$

If we take a value a such that $F_X(a)$ this means that at that value the probability

$$P(X \leq a) = P(X > a) = \frac{1}{2}$$

If we think about it this would also apply to the discrete case. If we take a random value from our list, half the time it will be smaller than our median, and half the time it will be larger.

From this argument, we can say that a is the median of the random variable X .

SUMMARY OF CALCULATING THE MEDIAN

Median of a discrete set of data is the middle value of the distribution if values are listed in order, for an odd number of values. Or the mean of the two central values, for even numbers of data points.

The median of a discrete random variable Z is the value z such that

$$P(Z \leq z) \geq 0.5$$

The median of a continuous random variable X is the value x such that

$$F(x) = \frac{1}{2}.$$

MEDIAN, QUANTILES AND STATISTICAL TABLES

We start to examine in more detail the properties of a probability distribution, and hence the distribution of the values of a random variable. Then we looked at

- Measures of Central tendency
  - Mean, Mode
  - Medians

We now look to other statistics descriptors to measure the variation

  - Quantiles and Percentiles

as well as a remarkable result about probability distributions called

=> **Chebyshev's Inequality.**

which leads to

=> The Weak Law of Large Numbers

MEASURES OF VARIATION

We have already looked at how to calculate the variance and standard deviation of a random variable.
They measure how much the different values of a random variable differ from the mean of the variable, and how likely the given value is to occur.

- **Variance and standard deviation**
- **Range:** *largest value minus the smallest value.*
- **inter-quartile distance:** *the 'distance' between the first quartile and the third quartile*

These measures are often applied to samples, which may or may not come from some idealised probability distribution.
The first two we have covered, or are obvious. What are quartiles?

QUANTILES

Quantiles are cut points dividing a set of observations into equal sized groups.

Actually, we've already met one - the median. This cuts the set of observations/values of the distribution in 2.

For a continuous distribution life is actually easier, the median is just the value at which the cumulative distribution function of a variable reaches $\frac{1}{2}$.

- median of a discrete set of data is the middle value of the distribution if values are listed in order, for an odd number of values. Or the mean of the two central values, for even numbers of data points.

- the median of a discrete random variable Z is the value z such that $P\ (Z \leq z) \geq 0.5$ and $P\ (Z \geq z) \geq 0.5$.

- median of a continuous random variable X is the value x such that $F(x) = \frac{1}{2}$.

QUANTILES

For a general quantile we just replace

$$\frac{1}{2} \text{ with} \frac{a}{n}$$

where $(n-1)$ is number of cuts we will make in the data, gives a^{th} n-tile.

Some of them have special names.

QUARTILES

If we take n= 4, then we would make 3 cuts in the data - we've quartered it.

the values at which we make the three cuts are called quartiles:

- **first quartile**, Q_1 , is defined by $F(Q_1) = \frac{1}{4}$
- **second quartile**, Q_2, is defined by $F(Q_2) = \frac{2}{4} = \frac{1}{2}$
- **third quartile**, Q_3, is defined by $F(Q_3) = \frac{3}{4}$

Note again that Q_2 is the median.

EXAMPLE

Consider a probability density function for the random variable Y

$$f_Y(y) = \begin{cases} 2e^{-2y} & \text{y} > 0\ , \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F_Y(a) = \int_0^a f(y)\mathrm{d}y = 1 - e^{-2a}$$

we can then calculate any quartile as $F_Y(Q_n) = \frac{n}{4}$

This implies, making $F_Y(Q_n)$ explicit, that

$$1 - e^{-2Q_n} = \frac{n}{4}$$

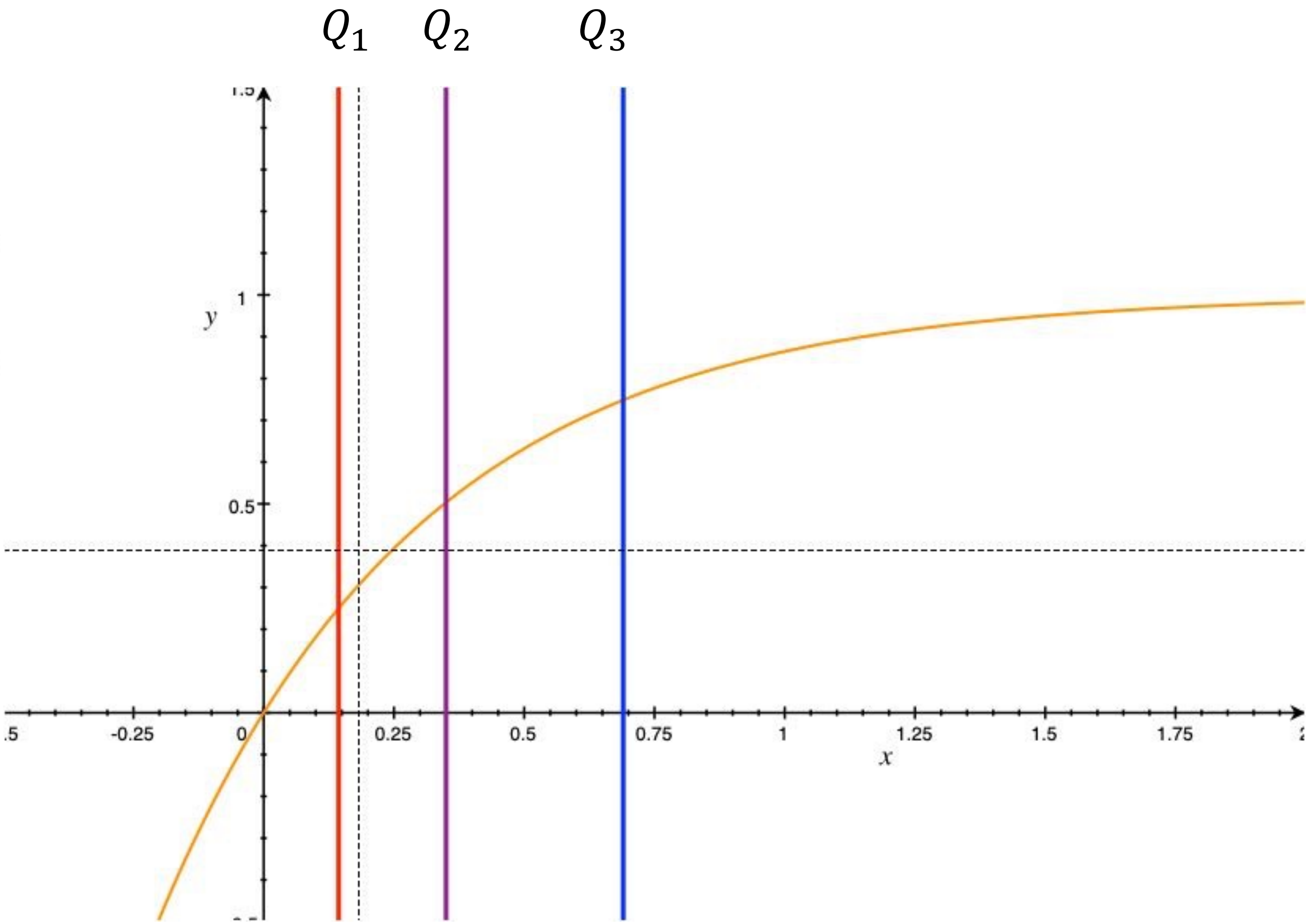
and rearranging we get

$$Q_n = -\frac{\ln(1 - \frac{n}{4})}{2}$$

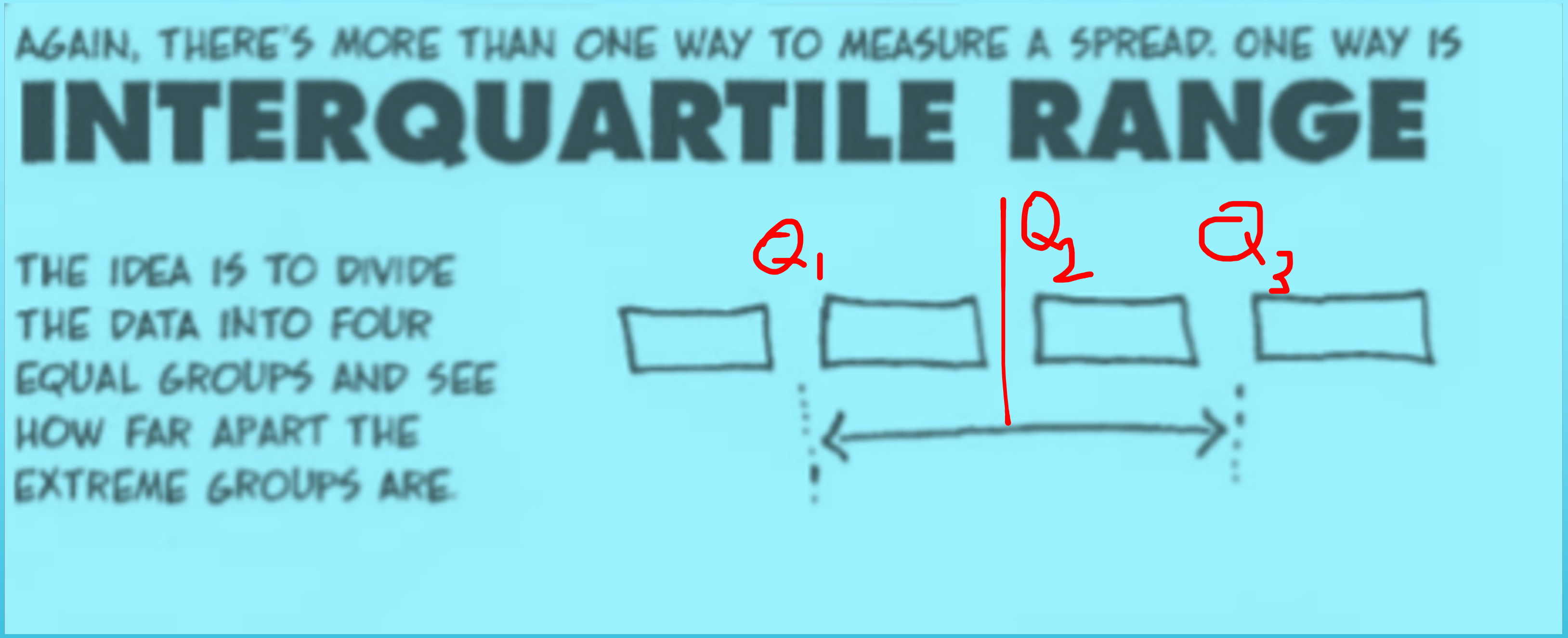
EXAMPLE

So we have

$$Q_1 = -\frac{\ln(1-\frac{1}{4})}{2} \approx 0.1438$$
$$Q_2 = -\frac{\ln(1-\frac{2}{4})}{2} \approx 0.3466$$
$$Q_3 = -\frac{\ln(1-\frac{3}{4})}{2} \approx 0.6931$$



USE OF THE QUARTILES



AGAIN, THERE'S MORE THAN ONE WAY TO MEASURE A SPREAD. ONE WAY IS

INTERQUARTILE RANGE

THE IDEA IS TO DIVIDE THE DATA INTO FOUR EQUAL GROUPS AND SEE HOW FAR APART THE EXTREME GROUPS ARE

From: Gonick & Smith The cartoon guide to Statistics

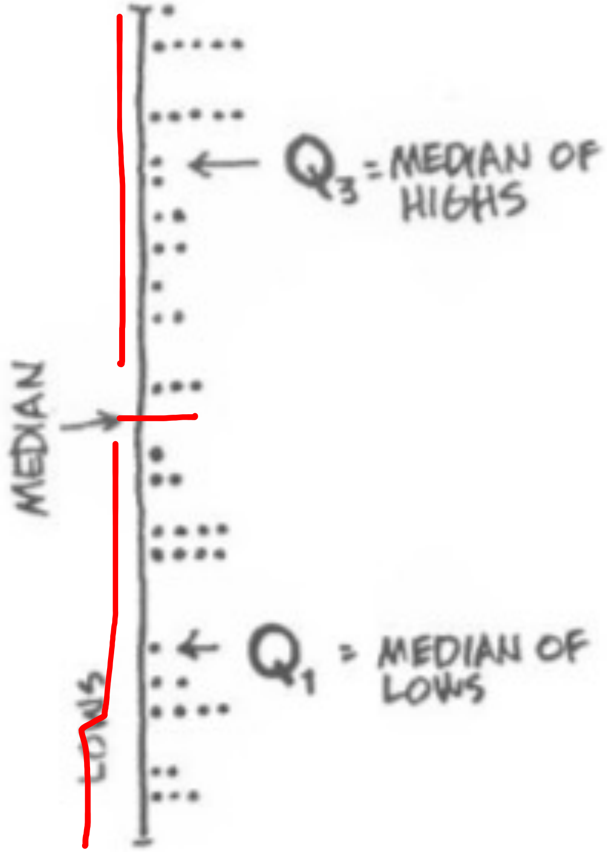
FIND QUARTILES IN A SET OF DATA

1) PUT THE DATA IN NUMERICAL ORDER.

2) DIVIDE THE DATA INTO TWO EQUAL HIGH AND LOW GROUPS AT THE MEDIAN. (IF THE MEDIAN IS A DATA POINT, INCLUDE IT IN BOTH THE HIGH AND LOW GROUPS.)

3) FIND THE MEDIAN OF THE LOW GROUP. THIS IS CALLED THE FIRST QUARTILE, OR Q_1.

4) THE MEDIAN OF THE HIGH GROUP IS THE THIRD QUARTILE, OR Q_3.

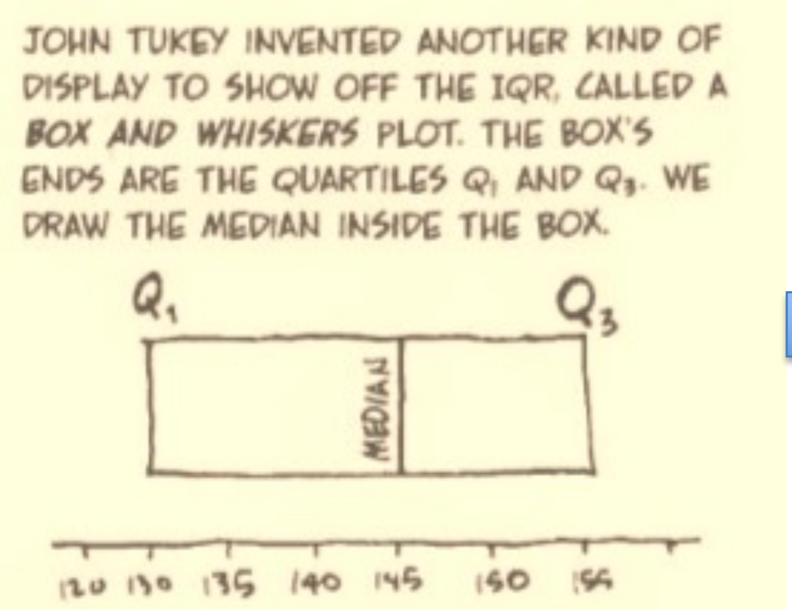


NOW THE INTERQUARTILE RANGE (IQR) IS THE DISTANCE (OR DIFFERENCE) BETWEEN THEM:

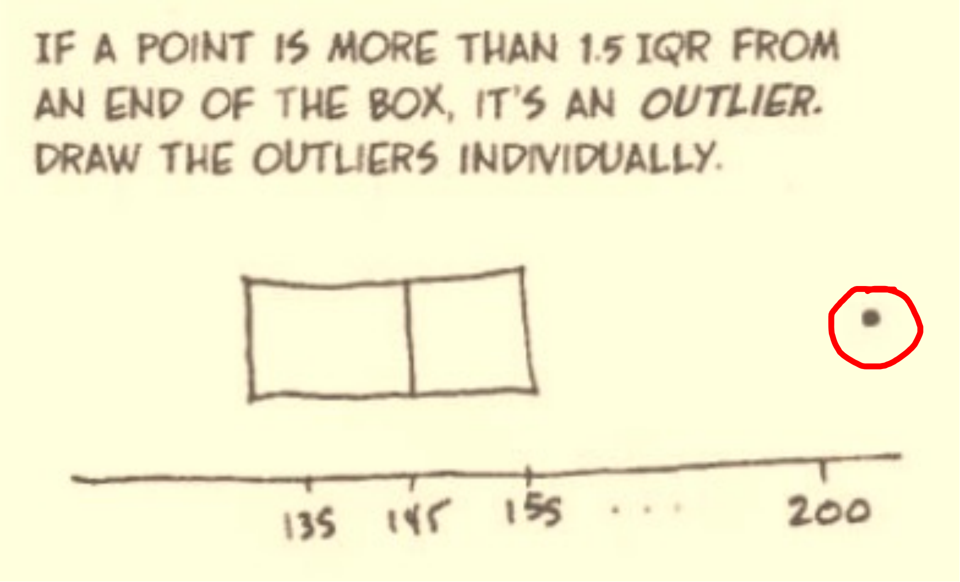
$$IQR = Q_3 - Q_1$$

From: Gonick & Smith The cartoon guide to Statistics

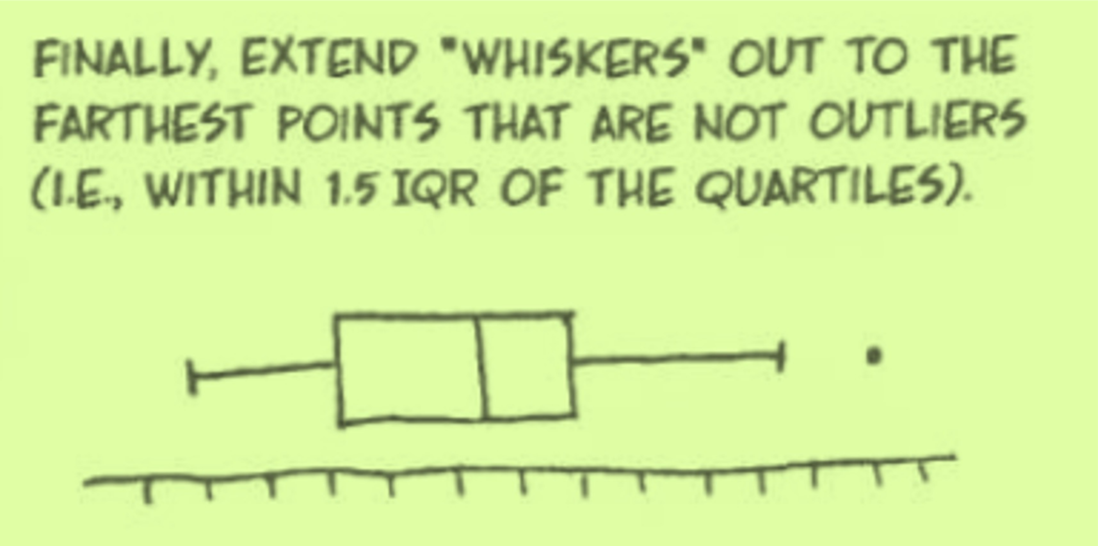
BOX AND WHISKERS REPRESENTATION OF THE IQR



JOHN TUKEY INVENTED ANOTHER KIND OF DISPLAY TO SHOW OFF THE IQR, CALLED A ***BOX AND WHISKERS*** PLOT. THE BOX'S ENDS ARE THE QUARTILES Q_1 AND Q_3. WE DRAW THE MEDIAN INSIDE THE BOX.



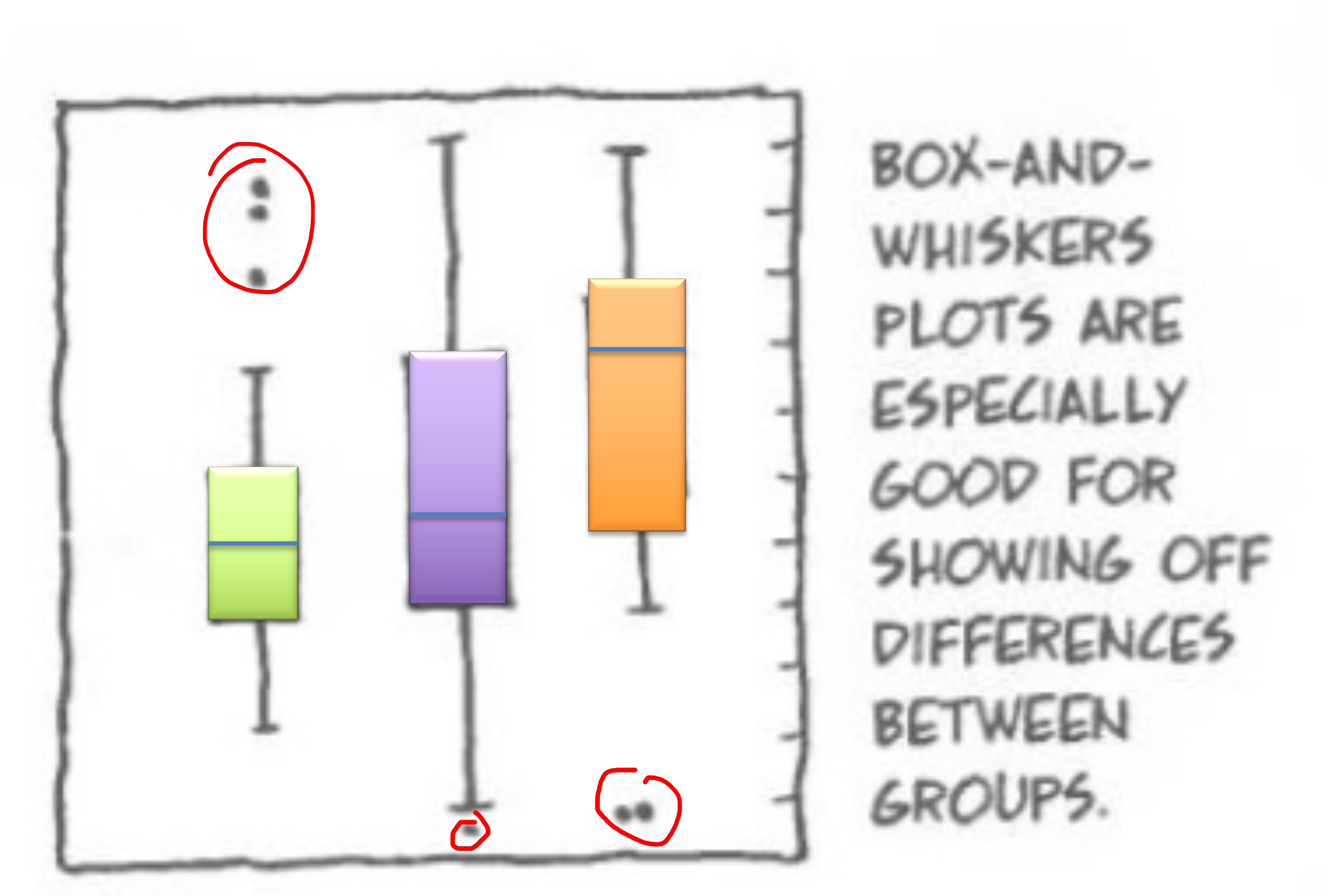
IF A POINT IS MORE THAN 1.5 IQR FROM AN END OF THE BOX, IT'S AN ***OUTLIER.*** DRAW THE OUTLIERS INDIVIDUALLY.



FINALLY, EXTEND "WHISKERS" OUT TO THE FARTHEST POINTS THAT ARE NOT OUTLIERS (I.E., WITHIN 1.5 IQR OF THE QUARTILES).

From: Gonick & Smith The cartoon guide to Statistics

GRAPHICAL REPRESENTATION OF THE IQR


BOX-AND-WHISKERS PLOTS ARE ESPECIALLY GOOD FOR SHOWING OFF DIFFERENCES BETWEEN GROUPS.

From: Gonick & Smith The cartoon guide to Statistics

PERCENTILES

Of course, there is no reason to restrict ourselves to quartiles.

We can also divide the interval in 100 parts and calculate
The so-called **percentiles**

We just need to set the cumulative distribution function equal to

$$F(x) = \frac{a}{100}$$

And chose the a^{th} percentile

Example

Consider a probability density function

$$f_Y(y) = \begin{cases} 2e^{-2y}, & y > 0, \\ 0, & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F_Y(a) = \begin{cases} 0, & a < 0 \\ \int_0^a f_Y(y)\mathrm{dy} = 1 - \mathrm{e}^{-2\mathrm{a}}, & a \geq 0 \end{cases}$$

Example

We can then calculate the n^{th} percentile, p_n , as

$$F_Y(p_n) = 1 - e^{-2p_n} = \frac{n}{100}$$

rearranging we get

$$p_n = -\frac{\ln(1 - \frac{n}{100})}{2}$$

So we have for instance

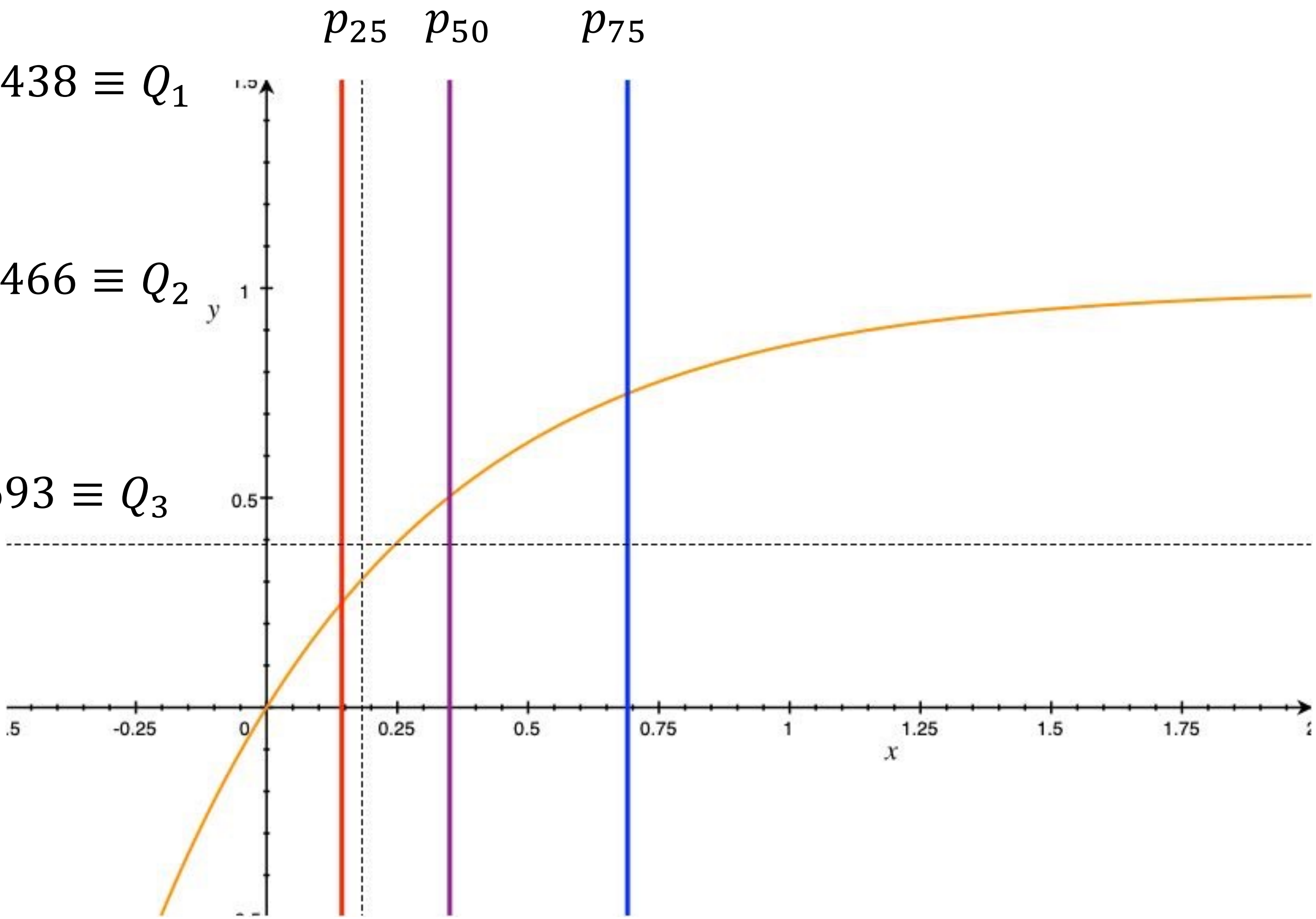
EXAMPLE

$$p_5 = -\frac{\ln\left(1-\frac{5}{100}\right)}{2} \approx 0.026$$

$$p_{\underset{25}{20}} = -\frac{\ln\left(1-\frac{25}{100}\right)}{2} \approx 0.1438 \equiv Q_1$$

$$p_{50} = -\frac{\ln\left(1-\frac{50}{100}\right)}{2} \approx 0.3466 \equiv Q_2$$

$$p_{75} = -\frac{\ln\left(1-\frac{75}{100}\right)}{2} \approx 0.693 \equiv Q_3$$



STATISTICAL TABLES

These give values of these percentiles for various distributions, that is all. Nowadays it is easier to get a computer to calculate them for you.

Particularly, in the IT sessions, we will use calls to the python **scipy library** to generate them.

SUMMARY: MEASURES OF VARIATION

We have already looked at how to calculate the variance and standard deviation of a random variable.
They measure how much the different values of a random variable differ from the mean of the variable, and how likely the given value is to occur.

- **Variance and standard deviation**
- **Range:** *largest value minus the smallest value.*
- **inter-quartile distance:** *the 'distance' between the first quartile and the third quartile*

These measures are often applied to samples, which may or may not come from some idealised probability distribution.

MARKOVS'S INEQUALITY

Lets actually use all this machinery that we've built up to see if we can say anything general about a distribution.

Let *X* be a continuous random variable with probability density function f(x).

We can show that if *X* only takes on positive values that

$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$$

MARKOVS'S INEQUALITY

Now

$$E[X] = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

$$= \int_{-\infty}^{a} x f(x) \mathrm{d}x + \int_{a}^{\infty} x f(x) \mathrm{d}x$$

Now we use the condition that X is non-negative to set up an inequality

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{a} x f(x) \mathrm{d}x + \int_{a}^{\infty} x f(x) \mathrm{d}x \\ &\geq \int_{a}^{\infty} x f(x) \mathrm{d}x \\ &\geq \int_{a}^{\infty} a f(x) \mathrm{d}x \\ &= a \int_{a}^{\infty} f(x) \mathrm{d}x \\ &= a P(X \geq a) \end{aligned}$$

Which proves our initial statement.

CHEBYSHEV'S INEQUALITY

We've proved that

$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$$

Take the particular case of $a = k^2$. And consider the expectation value of $(X\text{-}E[X])^2$

$$P((X - \mathrm{E}[X])^2 \geq k^2) \leq \frac{\mathrm{E}[(X - \mathrm{E}[X])^2]}{k^2} = \frac{\sigma_X^2}{k^2}$$

As long as the mean of X , μ , and its variance, σ^2 , are finite then we can equivalently say

$$P\Big(|X - \mathrm{E}[X]| \geq k\Big) \leq \frac{\sigma^2}{k^2}$$

CHEBYSHEV'S INEQUALITY

or if we define a new k′ in units of σ_X, the standard deviation, with $k = k'\sigma_X$), we get

$$P\left(|X - \mathrm{E}[X]| \geq k'\sigma_X\right) \leq \frac{1}{k'^2}$$

→ CHEBYSHEV'S INEQUALITY

Pafnuty Lvovich Chebyshev (1821 –1894) was a Russian mathematician.

His name can be alternatively transliterated as Chebychev, Chebysheff, Chebyshov; or Tchebychev, Tchebycheff (French transcriptions); or Tschebyschev, Tschebyschef, Tschebyscheff (German transcriptions)!

IMPLICATION OF CHEBYSHEV'S INEQUALITY

What does

$$P\left(|X - \mathrm{E}[X]| \geq k'\sigma_X\right) \leq \frac{1}{k'^2}$$

tell us?

Lets take k' = 2 to be concrete. Then

$$P\left(|X - \mathrm{E}[X]| \geq 2\sigma_X\right) \leq \frac{1}{4}$$

or put another way, *75% of the probability distribution is within 2 standard deviations of the mean.*

OR, also that only at most $\frac{1}{3^2} = 0.111 => \ 11.1\%$ of the data will be further than k’=3 standard deviations from the mean.

NOTE: We have made no assumptions at all about the distribution - it applies to any random variable! The Chebyshev’s inequality is generally a rather loose bound, if we know something about the real distribution we can do (much) better.

EXAMPLE

Suppose it is known that the number of items produced in a factory during a week is a random variable, X , with the mean value of 50.

What can be said about the probability that this week's production exceeds 75?

EXAMPLE

we can use Markov's inequality

$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$$

here we know that E[X] is 50, and the production is a random variable that does not take on negative values.

Plugging in the numbers we have

$$P(X \geq 75) \leq \frac{50}{75} = \frac{2}{3}$$

So not very useful in this case.

EXAMPLE

Q: If the production variance equals 25, then what can be said about the probability that the production this week will be between 40 and 60 items?

By Chebyshevs inequality in the form

$$P\left(|X - \mathrm{E}[X]| \geq k\right) \leq \frac{\sigma^2}{k^2}$$

we can plug in the numbers to get

$$P(|X - 50| \geq 10) \leq \frac{25}{10^2} = \frac{1}{4}$$

so the probability that the production is between 40 and 60 is at least $1 - \frac{1}{4} = \frac{3}{4}$

EQUIVALENT PROBABILITIES

In the last expression we manipulated equivalent probabilities

$$\begin{aligned}
P(\{|X-50|\geq 10\}) &= P(\{X-50>10\cup X-50<-10\}) \\
&= P(\{X>60\cup X<40\}) \\
&= P(\{40<X<60\}^{C}) \\
&= 1-P(40<X<60)
\end{aligned}$$

If we remember the things in the probabilities are really sets, we can rearrange them to get them into a useful form, and get different statements that define the same sets and hence the same probabilities.

WEAK LAW OF LARGE NUMBERS

We're getting a bit ahead of ourselves but let us go ahead and follow through a remarkable implication of the inequality we've just derived.

Let X_1 , X_2 , … , X_n be a sequence of independent and identically distributed random variables, each having E[X] = μ.

Then for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) \to 0 \text{ as } n \to \infty$$

WEAK LAW OF LARGE NUMBERS: EXAMPLE

If the X_i correspond to consecutive tosses of a coin, and we give them a value 0 if they come up a tail and 1 a head then the weak law of large numbers will say that

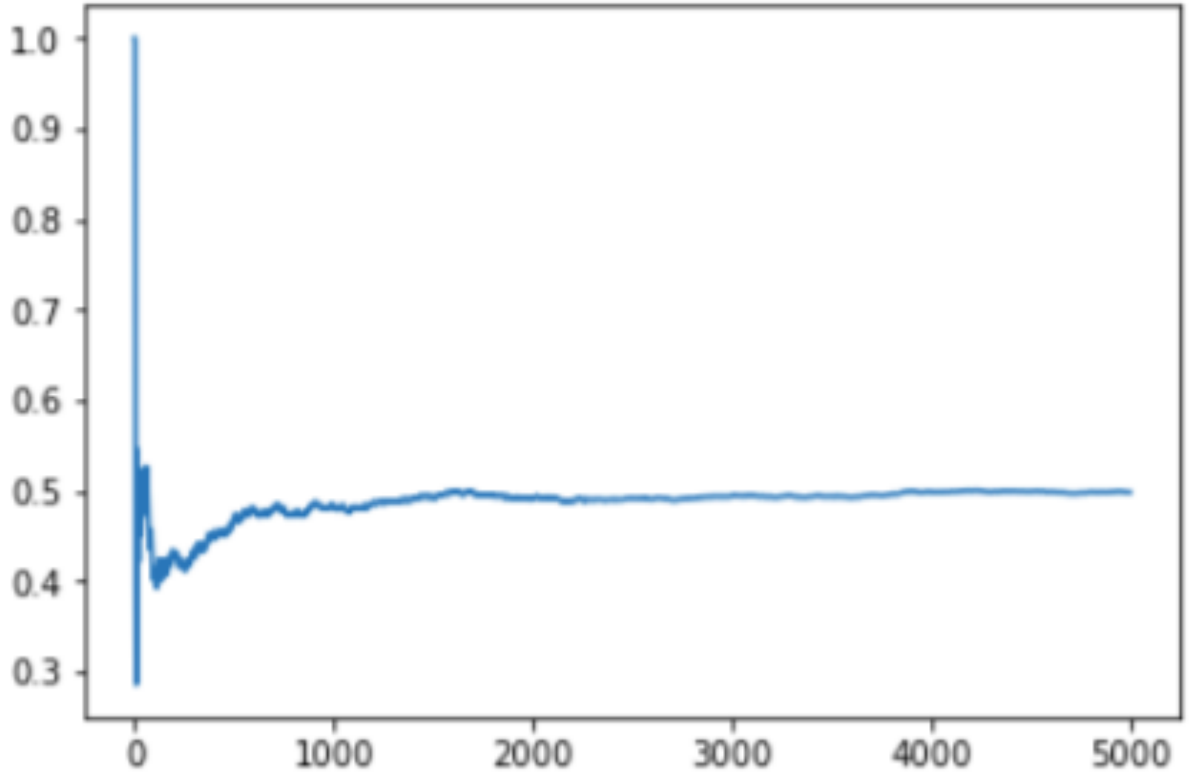
$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{1}{2}\right| > \epsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty$$

In this case, as $E[X_i]$ = P ({heads}) we see that the average proportion of the first n flips that differ from the idealised probability P (E) by more than

ϵ → 0 as *N*-> ∞ .

WEAK LAW OF LARGE NUMBERS: EXAMPLE

Expectation value of our variable, which equals P ({heads}) , as we extend the length of a simulation.



SUMMARY

- Measures of Central tendency
- Medians, Quantiles and Percentiles
- Measures of Variation

as well as a remarkable result about probability distributions called

- Chebyshev's Inequality.

Which leads to

The Weak Law of Large Numbers