STAT 513/413: Lecture 7
Randomness and distribution

(what is randomness, after all?)

The second part is rather STAT 413 matter - STAT 513 people are rather assumed to know it by now. Rizzo Section 5.1-5.5 will not be covered in the lectures, and their contents is irrelevant here, but they are assumed to be known (by everybody). Rizzo Sections 12.1 and 12.2 may turn out to be remotely useful - but still, the most useful source are these lectures. (The most relevant from the book would be Figure 3.2 of Rizzo, if there were more surrounding explanation.)

We would like to leave philosophy aside...

... but then - is it a surprise? - some may come out as totally clueless even in seemingly obvious technical issues!

So, what do we consider "random"?

People have written books to this effect; anything I may try here just in passing may be thus totally inappropriate; but hopefully at least somewhat helpful. Let me try:

"Random" is something beyond (our) complete reach

"Beyond reach" - something (we are) unable to influence completely (only partially maybe); includes something that is way too complex to be apprehended ("beyond grasp")

"Random" as demonstrating itself in the simplest sequences one is able to conceive

(A digression for geeks. Kolmogorov complexity theory: "random" mean "too complex to describe")

A bit of meditation upon the sequences:


```
> seq1
[1] 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
```



```
> seq2
[1] 0 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1
```



```
> seq3
[1] 1 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 0 1 1 1 0 1 0 1 0 1 0
> seq4
[1] 0 1 1 0 1 1 1 0 0 1 0 1 1 1 0 1 1 1 1 0 0 0 1 0 0 1 1 0 1 0 1 0
> seq5
[1] 1 1 0 0 1 0 0 1 0 0 0 0 1 1 1 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 1 0
```

"Random" may be also "haphazard"...

... but that will not be studied here (and is not in any similar course)

Of course, there is more to life than 0-1

Example. You feel thirsty: you desire a beer. You thus go out to a convenience store; you buy a can of beer there and return home.

That is completely "within reach": nothing "random" there

Now, on your way back home a truck deliberately leaves the road for the sidewalk and kills you (did happen in Toronto)

That is hardly within reach... It is random ("bad luck"), but...

Another example. You contract an influenza; most people do not die of it, but you do.

That is hardly within reach as well. At least, not for you; but for somebody else perhaps...

"The death of one person is a tragedy. The death of millions is just statistics."

Note: the "definition" of random says *complete* reach

Our "random" will be within at least some reach

This is quite impossible to define, but: for instance, anything that can happen *repetitively*, is within some reach

(Digression for geeks. We do not completely dismiss probabilistic analysis of non-repetitive events, as such may be important for decisions. We only say that repetitive phenomena are easier to handle - and in this lectures, they are the only ones we handle.)

Repetitive outcomes with only finitely many possibilities, repeated numbers, points in the plane, ...

In the repetitive setting, it is possible to work with relative frequency and to introduce a notion of probability based on a belief that it something like a "frequency in a long run"

"Chance is like fire: a good servant, but bad master."

What is "long run"

What "long run"? Mathematics, as we know it nowadays, does not know better than to resort to infinity - which in real life does not exist

But in real life there are "long runs": those are "hypothetically possible but practically way too long"

Example. What is the probability that a randomly drawn Canadian is female?

Simple:

$$\frac{\text{number of Canadian females}}{\text{total number of Canadians}}$$

We believe there is a such a thing - although changing perhaps every minute or second due to births and deaths. We can, however, calculate with it (call the result of the above ratio p: what is the probability that among 10 randomly chosen Canadians, at least 5 are women?), we can estimate it - it is hard to figure that out for all Canadians, but we can do it just in our neighborhood, for instance; and the larger the neighborhood(s) and more spread around the country, the more precise we believe our estimate is

Also, we can use physical symmetry

Everybody believes that coin lands on both sides with equal chance (and never stays in the air or even in the vertical position). This is a blessing in the classroom setting, because then we can fairly easily solve problems like the following one

Problem: what is the probability of obtaining 2 or less heads in 3 coin tosses. Closed-form solution: because this is very easy problem, we can use the equiprobability argument. There are 8 possibilities


```
[1,]    0    0    0
[2,]    0    0    1
[3,]    0    1    0
[4,]    0    1    1
[5,]    1    0    0
[6,]    1    0    1
[7,]    1    1    0
[8,]    1    1    1
```


Each of them has the same probability - that is, 1/8. The event "2 or less" includes 7 of them; hence the probability is $7/8 = 0.875$.

Personal numbers

Two people pick up a number between 100 and 999. What is the probability they pick up the same one?

Answer. There are 900 such numbers. *If we consider them all equally likely* and if we believe that the said two people do they picking *independently of each other*, then the result is

```
> (1/900)*(1/900)
1] 1.234568e-06
```


That is, 0.000001234568 - almost 1 in milion

While producing an answer like this is no big deal, the *real* answer may be tricky: see all the assumptions above. Even when there are no obvious violations, one never knows... They once wanted me to speak on TV about what is the probability that somebody is born on the same day of year as his father and grandfather. (I declined. Neither clerics explain their religion in five minutes.)

(What is the probability that three my Facebook friends have birthday on the same day, and another two too, albeit on a different one?)

Back to the coin: but "biased" now

"Biased coin": an invention of teachers when they want to assign a more interesting and realistic problem. It does exist in reality: a thumbtack

What is the probability that it falls in the "dangerous" way? Do you believe it is 1/2? If yes, could we bet? If we agree that it is p: what is the probability that in 10 tosses, it falls 2 or less times in the "dangerous" way?

(It happened at Tour de France: somebody threw a handful of them to harm cyclists. But long ago.)

It would not happen now

Something operatively useful now: distributions

So far, we looked at the repetitive random events with finite number of outcomes. These outcomes have a probability which can sometimes be calculated out of symmetry considerations in closed form. Nonetheless, even if this is not that feasible, we can always *estimate* such a probability as a relative frequency - and make it also visual

(Before going into that, still a digression to the character of randomness. For a finite number of outcomes, "completely at random" is usually understood as all of them having the same chance. In such a case, there are maximally "out of reach" - while in other cases, they are somewhat less, ending up with the case when the probability of one outcome is one and that of the others is zero - which is not considered random at all. Somewhat pradoxically, but still: probability zero events *do not happen* in real life, only in its mathematical - probabilistic - modeling.)

A useful function: `table()`

It performs *cross-tabulation*

```
> table(seq1)
seq1
 0  1
16 16
> table(seq2)
seq2
 0  1
16 16
> table(seq3)
seq3
 0  1
19 13
> table(seq4)
seq4
 0  1
14 18
> table(seq5)
seq5
 0  1
16 16
```

Probability estimates are only one step ahead

```
> table(seq3)/sum(table(seq3))
seq3
      0       1
0.59375 0.40625
```

Something more interesting

Thrown 10 thumbtacks 100 times (did I?)

```
> table(seq6)
seq6
 0  1  2  3  4  5  6  7  8 10
 1  3 10 22 30 15 12  5  1  1
> table(seq6)/sum(table(seq6))
seq6
   0    1    2    3    4    5    6    7    8   10
0.01 0.03 0.10 0.22 0.30 0.15 0.12 0.05 0.01 0.01
```


(Wow! When I did it first time, there were no results under 9 or 10; I had to try 1000 times, I obtained result under 9, but still no 10, even when tried 10000 times; only 100000 times worked! Now 10000 was enough.)

```
> table(seq66)
seq66
   0    1    2    3    4    5    6    7    8    9   10
  72  390 1168 2229 2505 1951 1135  417  119   12    2
> table(seq66)/sum(table(seq66))
seq66
     0      1      2      3      4      5      6      7      8      9     10
0.0072 0.0390 0.1168 0.2229 0.2505 0.1951 0.1135 0.0417 0.0119 0.0012 0.0002
```

We can plot it too

```
> plot(table(seq66)/sum(table(seq66)),lwd=10)
```

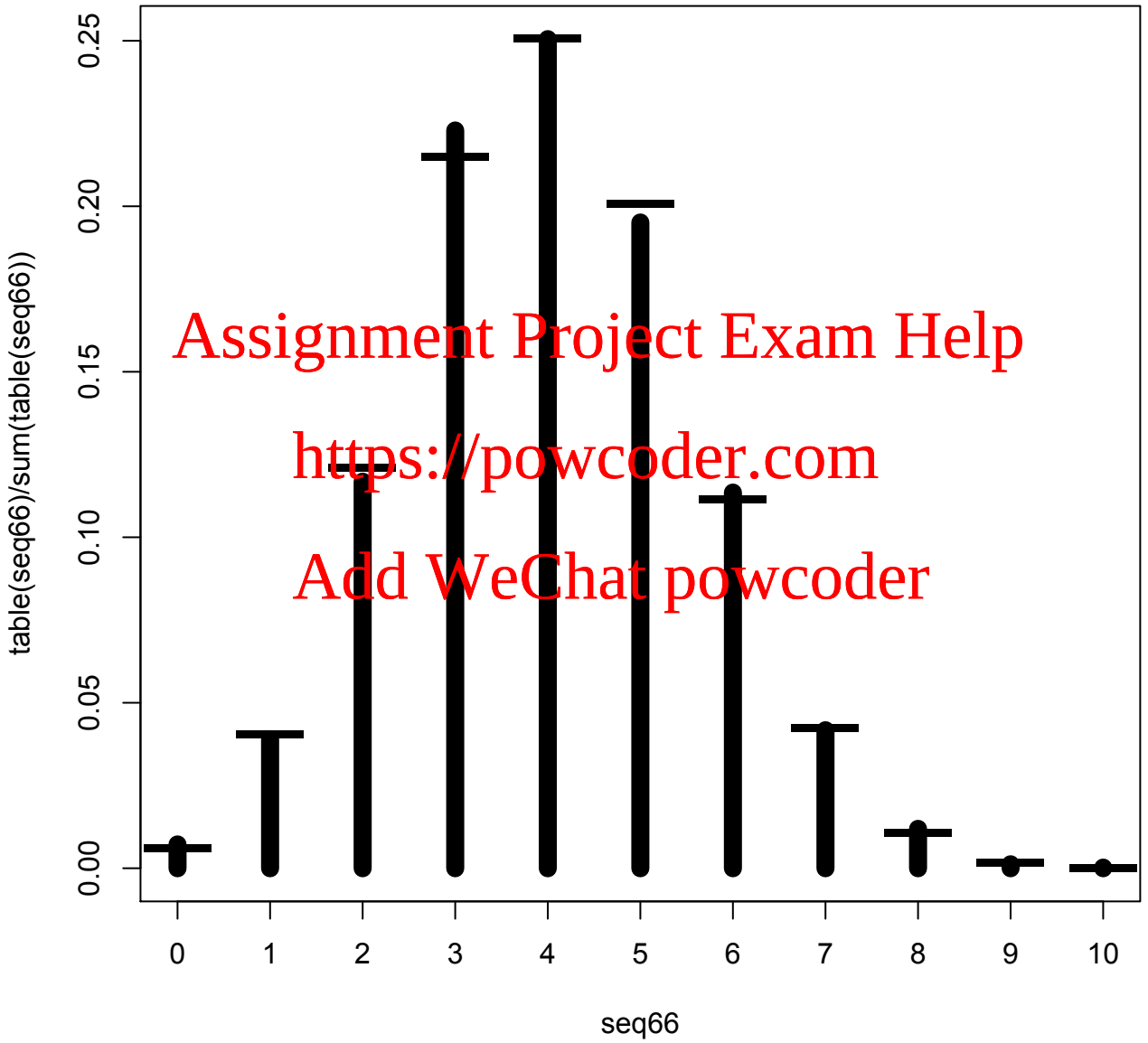
Now, what is the probability of the thumbtack falling "1"?

```
> table(seq66)
seq66
   0    1    2    3    4    5    6    7    8    9   10
  72  390 1168 2229 2505 1951 1135  417  119   12    2
> table(seq66)*0:10
seq66
    0     1     2     3     4     5     6     7     8     9    10
    0   390  2336  6687 10020  9755  6810  2919   952   108    20
> ptack=sum(table(seq66)*0:10)/(10000*10)
> ptack
[1] 0.39997
```

When we know that probability: is it not binomial distribution? Let us compare!

```
> points(0:10,dbinom(0:10,10,ptack),pch="_",cex=4)
```

The result



Also

For the sake of comparison, we can also do this, but we will seldom need to go that far

```
> chisq.test(table(seq66),p=dbinom(0:10,10,ptack))

Chi-squared test for given probabilities

data:  table(seq66)
X-squared = 12.296, df = 10, p-value = 0.2658
```


(This is just for those who know what is it about. A geek digression sort of, too. Otherwise, there is no time here to go that deep.)

Now, the continuous case

We have now outcomes that are numbers. Like this sequence with 1000 numbers

```
> seq7
   [1] 6.375508 6.917807 7.109299 6.749246 6.200092 5.766844
   [7] 4.986468 4.736126 4.310287 5.389913 7.874240 3.922325
  [13] 5.467302 5.326301 6.355067 6.038211 6.167680 5.696233
  [19] 3.828023 5.441739 4.771107 6.182582 7.419830 7.759132
...
```


Cross-tabulation is not very helpful here


```
> table(seq7)
seq7
2.2037467292975 2.6592409301083 2.7604466639459 3.4369396034162
              1               1               1               1
3.4706228903960 3.5918393908068 3.6490797747392 3.6644867989234
              1               1               1               1
 3.666546334279 3.7674304551910 3.7760527541395  3.800437211524
              1               1               1               1
...
```

Density estimate

If we are completely clueless, we can estimate the probability density underlying the sequence and plot it

```
> plot(density(seq7),lwd=3)
```

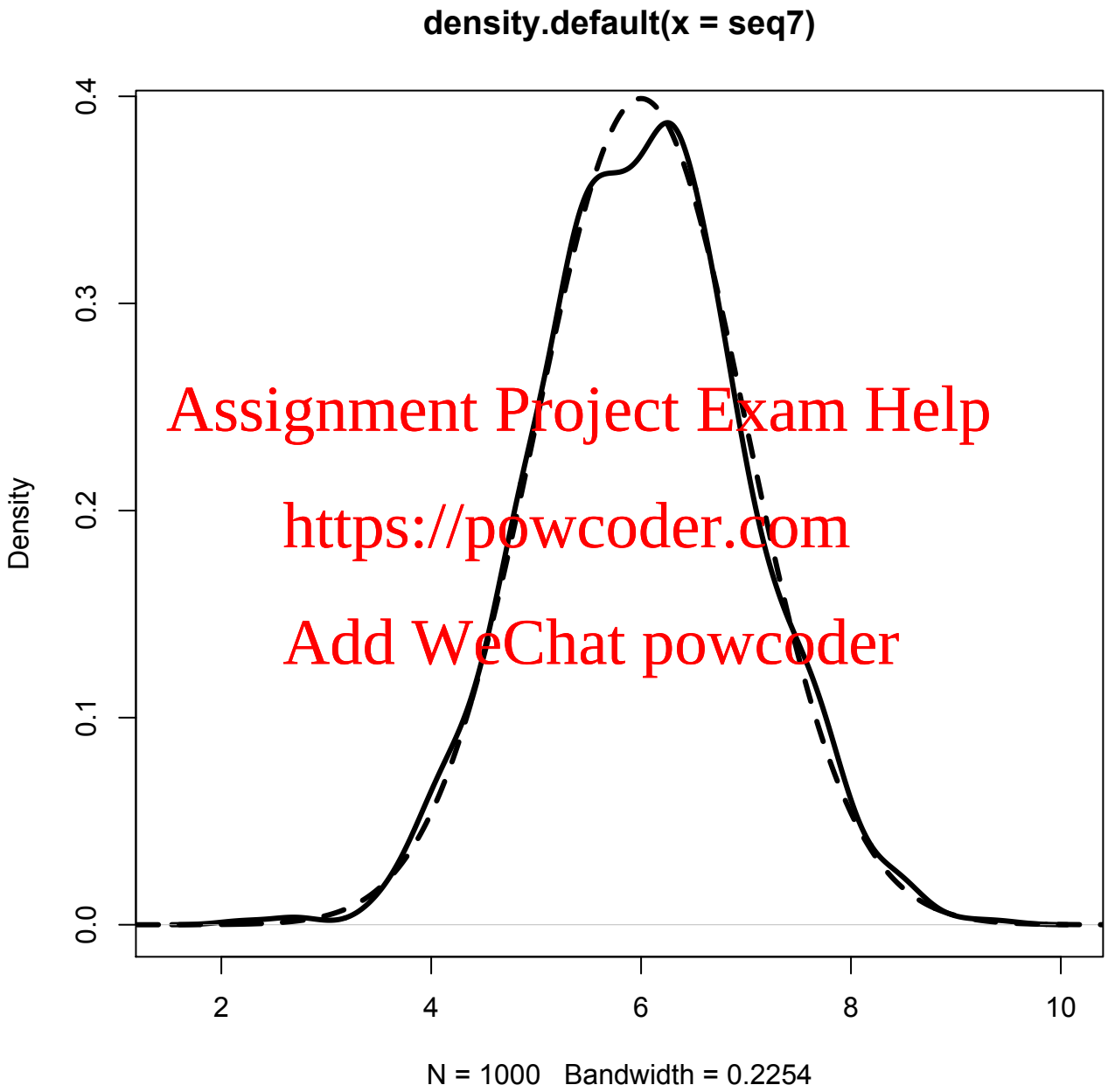
Is it not normal distribution, by the way?

```
> xx=seq(0,11,len=500)
> lines(xx,dnorm(xx,6,1),lwd=3,lty=2)
```


Well, who knows... (How did I know that $\mu = 6$ and $\sigma = 1$? By trial and error, let us say.)

It is useful when we are clueless; not that much for confirmations

For more, see Rizzo 12.2; and also 12.1



What is useful for confirmations? `qqplot()`

Plotting quantiles against quantiles

These quantiles can be either empirical or theoretical

Empirical quantiles: take original values $x_1, \ldots, x_n$, and *order* them:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots x_{(n)}$$

Theoretical quantiles: evaluate the quantile function on an equispaced grid

Regarding "equispaced": we have to avoid 0 and 1, and at the same time start close enough to those. A convenient R function used in this context to generate given number of equispaced points inside $(0, 1)$ is `ppoints`

```
> ppoints(5)
[1] 0.1190476 0.3095238 0.5000000 0.6904762 0.8809524
```

Compare to (after 0 and 1 are eliminated)

```
> (0:6)/6
[1] 0.0000000 0.1666667 0.3333333 0.5000000 0.6666667 0.8333333 1.0000000
```

Visual evaluation of `qqplot()`

We plot empirical quantiles against theoretical ones

```
> qqplot(seq7,qnorm(ppoints(length(seq7)),6,1))
```

and then we evaluate whether they lie approximately on a line; if the first and second argument of `qqplot()` correspond to *exactly* the same distribution, the line must be $y = x$ - it may help to add it to the picture


```
> abline(0,1)
```


Note: there is no universally set order in which theoretical and empirical quantiles should appear in the `qqplot()`. While some statisticians prefer the above one, the default in R (in `qqnorm()` function discussed below) is actually other way round


```
> qqplot(qnorm(ppoints(length(seq7)),6,1),seq7)
> abline(0,1)
```

Location and scale do not matter

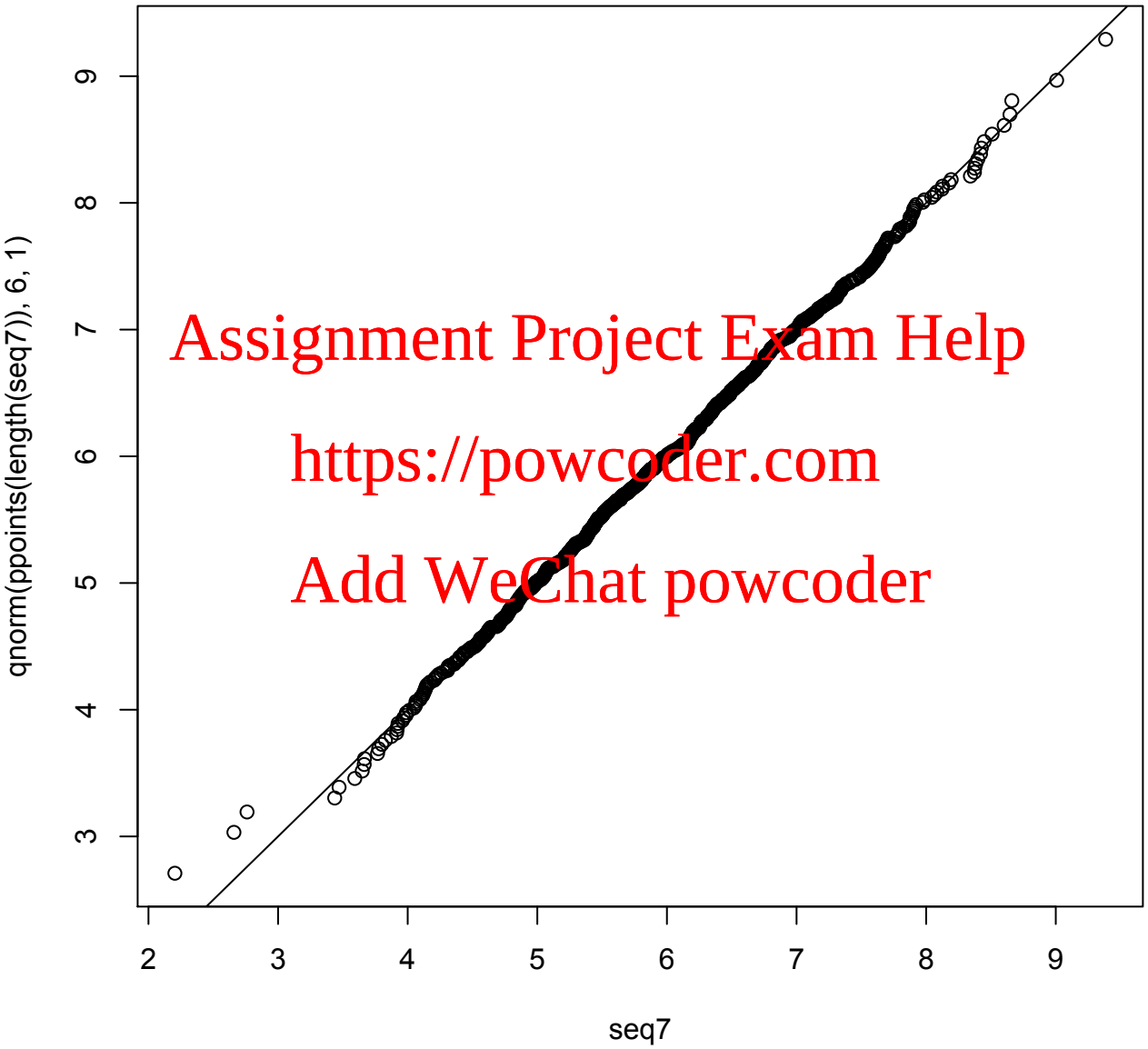
The beauty of `qqplot()` is that we do not have to know the location and scale of the theoretical distribution

That is: if the distribution is not exactly what we thought, but that of a linear tramsformation (not the distribution of x, but that of $a + bx$ for some a and b - which we do not need to know), we still get points lying approximately on a line

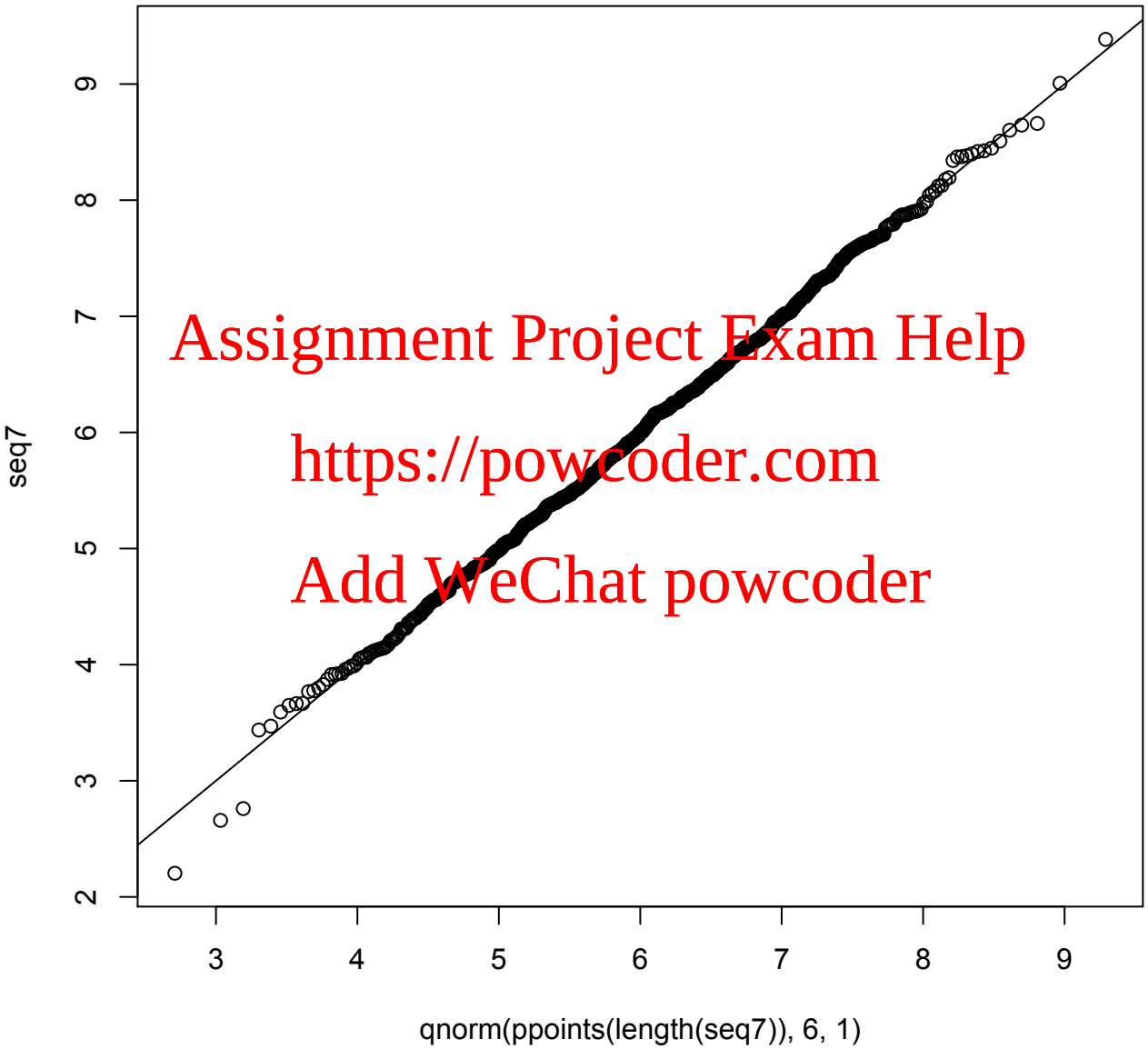
However, the line to fit would be not $y = x$ then, but a different one... If we do not know a and b, we have to do some trial and error then


```
> qqplot(seq7,qnorm(ppoints(length(seq7))))
> abline(0,1,lty=2)
> abline(-6,1)
```

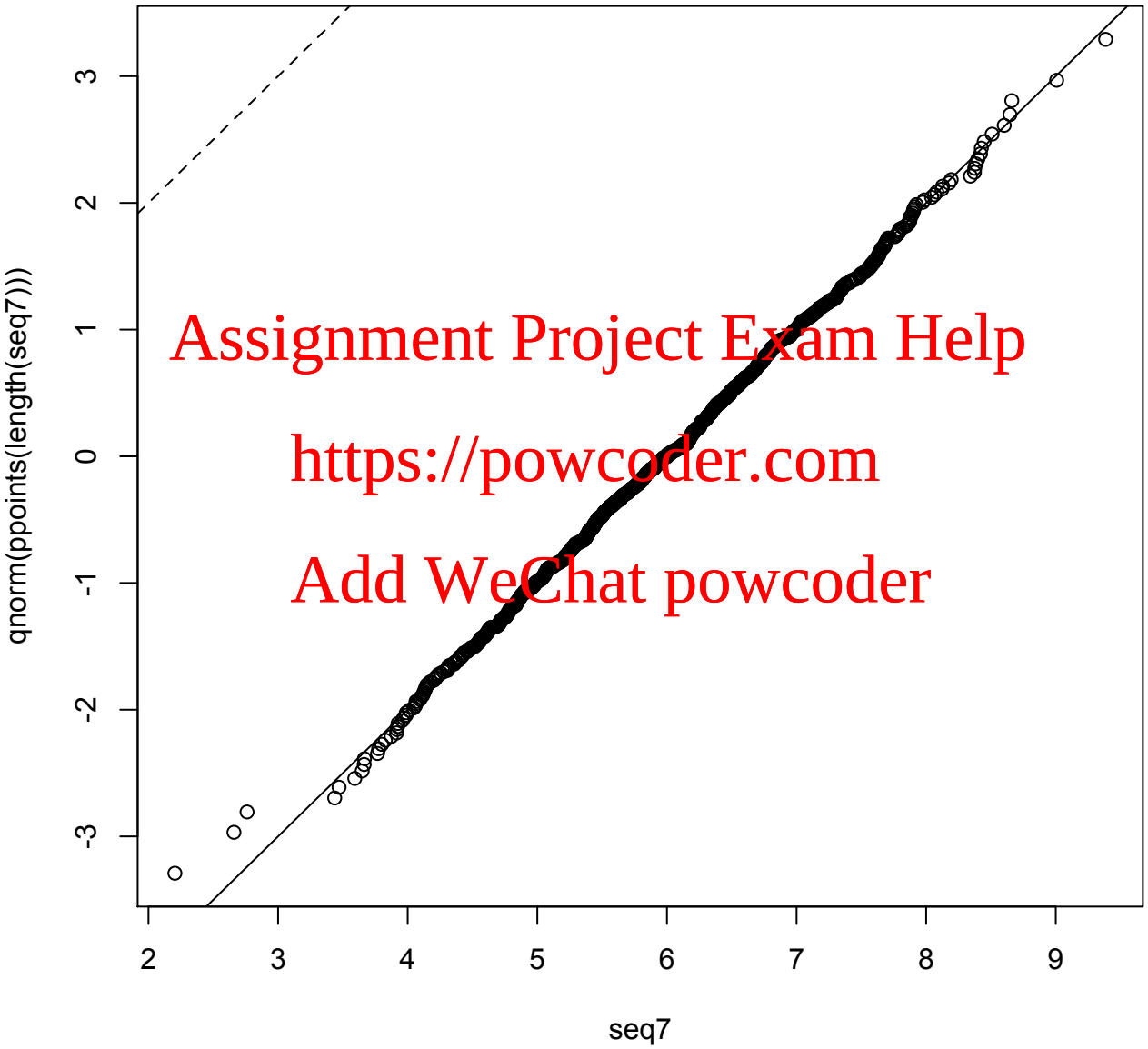
The first picture



And its default R variant



The second picture



Normal distribution is even easier

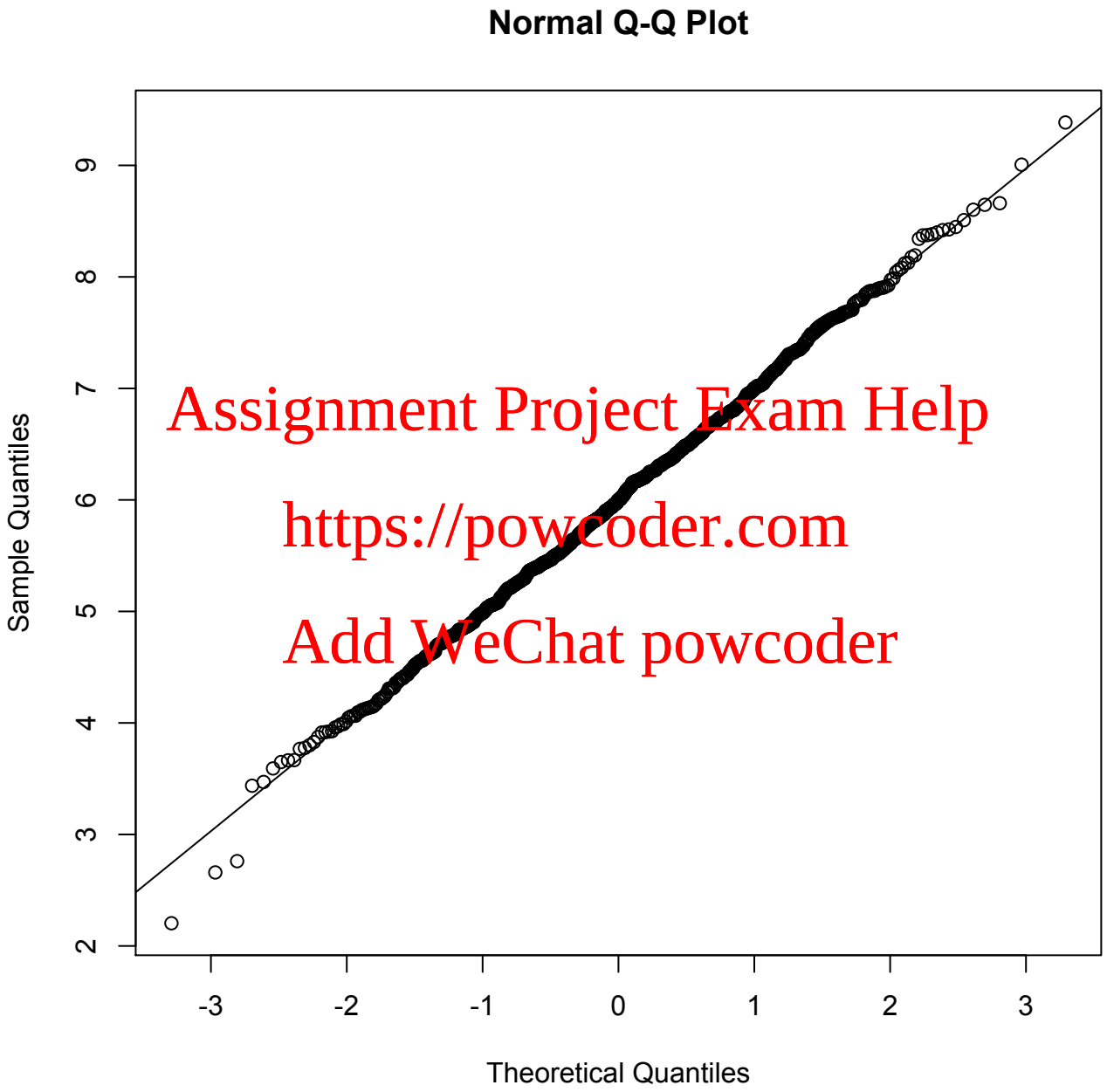
The above scheme works for general distributions; for the normal, it made even more streamlined via special functions `qqnorm()` and `qqline()`

```
> qqnorm(seq7)
> qqline(seq7)
```

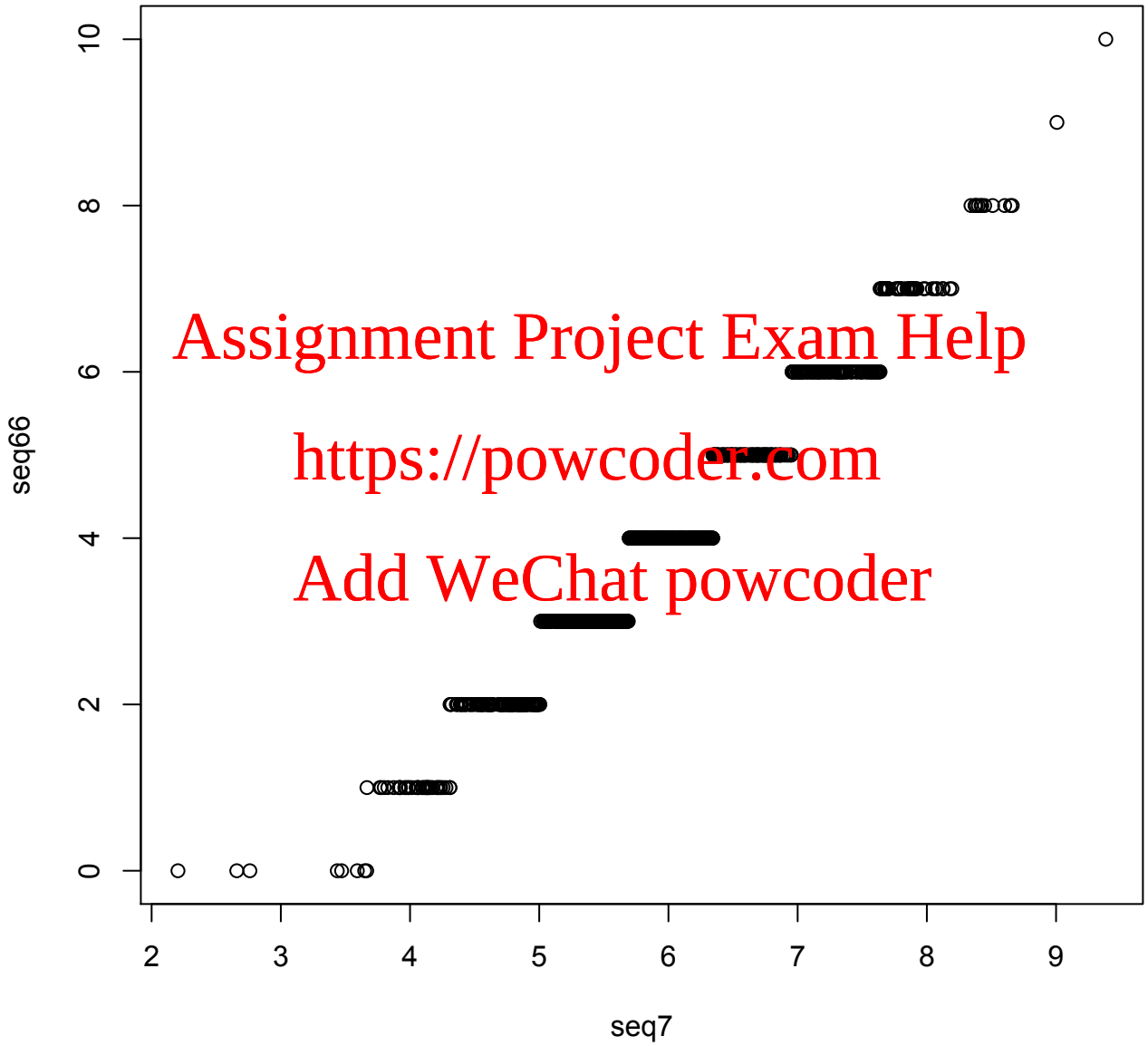

Finally, we can also compare in this way two sets of empirical quantiles - that is, whether two sets of numbers are likely to have the same distribution


```
> qqplot(seq7,seq66)
```

The third picture



Finally: do they have the same distribution?



Appendix: after all, what is "probability"?

Philosophical interpretations can be varied, and lengthy...

...but in our practice here we stick to the following principles:

• we are to evaluate probabilities only in a well-defined repeatable situation ("chance setup") - which we call, for the lack of better words, a *random experiment*: the latter has outcomes in some well-defined *sample space* (perhaps the lack of better words, or just some customary terminology); we consider an event as a particular subset of this sample space, and only two possibilities may happen: it occurs or it does not;

• *probability* is then something that is approached by the relative frequency of occurrence of an event in a number of repetitions of the experiment; *we believe that increasing the number of repetitions gives us better hold* of the said probability

Remember: the more repetitions, the closer we are supposed to be. We will come back to this.

Appendix: some technological recollections

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

If X and Y are independent, then $E(XY) = E(X)\,E(Y)$

$$\mathrm{Var}(X) = E\left((X - E(X))^2\right) = E(X^2) - (E(X))^2 \geqslant 0$$

$$\mathrm{Cov}(X, Y) = E[(X - E(X))(Y - E(Y)]$$

$$\mathrm{Cov}(X, X) = E[(X - E(X))(X - E(X))] = \mathrm{Var}(X)$$

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)$$

but only when X_i are independent; otherwise

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{Cov}(X, Y) + \mathrm{Var}(Y)$$

Transformations

$$\begin{aligned}\mathrm{Cov}(X+a, Y+b) &= \mathrm{E}[(X+a-\mathrm{E}(X+a))(Y+b-\mathrm{E}(Y+b)] \\ &= \mathrm{E}[(X+a-(\mathrm{E}(x)+a))(Y+b-(\mathrm{E}(Y)+b)] \\ &= \mathrm{E}[(X-\mathrm{E}(x))(Y-\mathrm{E}(Y)] = \mathrm{Cov}(X,Y)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(aX, bY) &= \mathrm{E}[(aX-\mathrm{E}(aX))(bY-\mathrm{E}(bY)] \\ &= \mathrm{E}[(aX-a\mathrm{E}(X))(bY-b\mathrm{E}(Y)] = ab\,\mathrm{Cov}(X,Y)\end{aligned}$$

$$\mathrm{Var}(X+a) = \mathrm{Var}(X) \qquad \mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$$

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \qquad -1 \leqslant \mathrm{Cor}(X,Y) \leqslant 1$$

Densities and cumulative distribution functions

A density, f, of a random variable X, gives the probabilities:

$$P[X \in A] = \int_A f(x)\,dx$$

For instance, the cumulative distribution function F of X is

$$F(z) = P[X \leqslant z] = P[X \in (-\infty, z]] = \int_{-\infty}^{z} f(x)\,dx$$

In fact, F is the antiderivative of f: $F'(x) = f(x)$ (roughly)

but beware: the antiderivative is also any function $F(x) +$ constant

So, how do we know a constant? From the fact that a cumulative distribution function $F(x)$ "starts" at 0 and "ends up" at 1

Precisely, $F(x) \to 0$ if $x \to -\infty$ and $F(x) \to 1$ if $x \to +\infty$

Conditional, marginal...

Conditional probabilities: the definition $P(A|B) = \frac{P(A\&B)}{P(B)}$

Conditional densities: $P[X|Y = y] = \frac{f(x,y)}{g(y)}$

where $f(x,y)$ is the *joint* distribution of X and Y

and $g(y)$ is the *marginal* (simply, the) distribution of Y

We can get the marginal from the joint via integration/summation

$$g(y) = \int f(x,y)\,dx \qquad \text{and, of course} \qquad h(x) = \int f(x,y)\,dy$$

In the discrete situation, we have probability mass functions instead of densities; the definitions remain the same, only instead of integrals we have to use sums

$$g(y) = \sum_x f(x,y) \qquad \text{and, of course} \qquad h(x) = \sum_y f(x,y)$$