Introductory Notes on Probability Theory

Grigor Aslanyan

October 28, 2009

1 Definition of probabilities

Let us imagine a class of 100 physics students that have to take a physics test. Each student wants to know her overall standing among everybody else, so she wants to know "the overall picture" of the whole class, in other words how the class did *on average*. So instead of looking at 100 numbers she just checks the *average* score of the class and gets a very good picture of the whole class without any hard work. Now suppose she is more curious and she wants to know if her classmate sitting next to her got a higher score than herself or not. Here she has a problem because the scores are not posted with names, so she has no way to know the exact score of her classmate. But she can still make some wise guesses. Suppose only 7 people out of 99 remaining students (she has to exclude herself since she knows her score exactly) got higher scores. Then she may guess that the chances that her classmate got a higher score than herself are slim. How does she arrive at that conclusion? What really matters in this case is the ratio of the number of people with higher scores to the total number of people, and we will define that ratio to be the *probability* of the fact that her classmate got a higher score than herself:

$$P(\text{her classmate has a higher score}) = \frac{\text{Number of students with higher score}}{\text{Total number of students}} \qquad (1)$$

By definition we see that this number is always between 0 and 1. If she was good enough to be the best in class, i.e. nobody got a higher score, then the probability becomes 0, but in this case she knows with certainty that her classmate didn't get a higher score. So we can conclude that **the facts with probability** 0 **are definitely false**. On the other hand, if she was the worst, then 99 out of 99 others got a higher score and the probability becomes 1, however now she will certainly know that her classmate got a higher score. So **the facts with probability** 1 **are definitely true**. Note that the whole analysis that we did for the classmate sitting next to her will be true for any other of her classmates, so we can say that the probability defined above is really the probability of a *randomly chosen student* to have a higher score:

$$P(\text{a random student has a higher score}) = \frac{\text{Number of students with higher score}}{\text{Total number of students}} \qquad (2)$$

Now imagine that she is very social and after two weeks she makes friends with 40 students in her class and feels free to ask them their test scores. Some of them will have higher scores, some of them lower, but we can understand that the higher the probability is for one single student to have a higher score, the more people among her 40 friends will turn out to have higher scores. A good guess of the number of people with higher scores among these 40 will be just $40 \times P$(a random student has a higher score). Of course, the more people she asks the more accurate that guess becomes, and if she asks everybody then this guess will become exact just by our definition of probability.

Let us try to generalize the idea of probability. In the example above we can think about experiments and outcomes. The experiment was to ask a random student her score, and the outcome was the answer of the question "Is that score higher than a given value?" - true or false. We were able to calculate the probability of the positive outcome simply because we knew all the outcomes (we knew all the scores), but in general we may not know that. However, we may still be able to figure out the probability by some other ways. Consider tossing a coin. We are interested in the probability of getting a head. In this case it makes no sense to consider all of the outcomes of all coin tosses that have happened or will ever happen in the universe, but still, if we imagine a very large number of coin tosses, then we expect to get a tail in approximately half of them. The ratio of the number of heads to the total number of tosses will be the closer to 0.5 the more experiments we make. So although there is no finite number of experiments, we can still find the probability just by considering a "very large number" of experiments. There are even cases where we are able to make but a single experiment with a given system, but we can still talk about probabilities. For example, we may ask the question "What is the probability that it will rain tomorrow in La Jolla?". Depending on the weather conditions that we have today it is possible to find an answer to that question, but of course we will not be able to do the experiment (i.e. check if it rains the next day) more than once since we cannot reproduce the conditions that we have today once again or a very large number of times. But we can still imagine a large number or experiments and ask ourselves what the ratio of the systems where it actually rains the next day would be to the total number, i.e. the probability of raining tomorrow (after all, we did not have to really toss coins to figure out the probability of getting a head, imagination and intuition were enough). So we can in general define the probability that a given event A happens by the results of a large number of (real or imaginary) experiments:

$$P(A) = \frac{\text{Number of experiments where } A \text{ happens}}{\text{Total number of experiments}} \tag{3}$$

where formally we have to take the limit of the total number going to infinity.

2 Adding probabilities and normalization

Let us consider our example of physics students once again. Now imagine that a given student wants to know the probability that her classmate got the same score as herself. Let us call that *event* A. By our definition of probabilities above, we have

$$P(A) = \frac{\text{Number of students with equal score}}{\text{Total number of students}} \tag{4}$$

If we define *event* B to be the case when her classmate has a higher score, then we have

$$P(B) = \frac{\text{Number of students with higher score}}{\text{Total number of students}} \tag{5}$$

Now let event *event* C be the one when her classmate has an equal or higher score, which simply means event A *or* event B happens. We get

$$P(C) = \frac{\text{Number of students with equal or higher score}}{\text{Total number of students}} \tag{6}$$

But clearly

$$\text{Number of students with equal or higher score} = \text{Number of students with equal score} + \text{Number of students with higher score} \tag{7}$$

Plugging in (6) we get

$$P(C) = \frac{\text{Number of students with higher score}}{\text{Total number of students}} + \frac{\text{Number of students with equal score}}{\text{Total number of students}} = P(A) + P(B)$$

The important point above was that events A and B were mutually exclusive, i.e. there is no case when they happen simultaneously, that was why (7) was true. Generalizing, we can say that whenever events A and B are mutually exclusive we have the rule

$$P(A \text{ or } B) = P(A) + P(B) \tag{8}$$

Let us now consider the *event* B' which is the case when the classmate has a lower score. We then have

$$P(B') = \frac{\text{Number of students with lower score}}{\text{Total number of students}} \tag{9}$$

Clearly B' is mutually exclusive with both A and B, so we can apply our addition rule (8) to all three events A, B, and B' to get

$$P(A \text{ or } B \text{ or } B') = P(A) + P(B) + P(B') \tag{10}$$

However, these three events, being mutually exclusive, exhaust all of the possibilities, in other words the event A or B or B' is true with certainty, i.e. it has probability 1. Therefore

$$P(A) + P(B) + P(B') = 1 \tag{11}$$

The equation (11) is called the *normalization condition*. Of course, in this example the normalization condition can be checked directly by using equations (4), (5), and (9) and the fact that the total number of students is equal to the sum of the number of students with higher score, equal score, and lower score.

We can formulate the normalization condition for the general case as follows. Suppose that all of the possible outcomes of a given experiment are events A_1, A_2, ... , A_n, such that they are mutually exclusive. Then the sum of their probabilities must be equal to 1:

$$\sum_{i=1}^{n} P(A_i) = 1 \tag{12}$$

In general, we may not be able to calculate all of the different probabilities and check the normalization condition explicitly, but we have to use the normalization condition in addition to what we know to figure out the probabilities. Let us consider the example of tossing a coin, and calculate the probabilities of possible outcomes in a different way. We know that there are two possible outcomes - head and tail. We will use only the fact, that neither of these events is more preferable, i.e. they must have the same probability:

$$P(\text{head}) = P(\text{tail}) \tag{13}$$

But we also know that the normalization condition must be true, which in this case becomes

$$P(\text{head}) + P(\text{tail}) = 1 \tag{14}$$

Plugging (13) in (14) we get

$$2P(\text{tail}) = 1$$

$$P(\text{tail}) = \frac{1}{2}$$

And then (13) gives

$$P(\text{head}) = \frac{1}{2}$$

in agreement with our previous result!

3 Average and standard deviation

Before we formulate the rules for calculating averages in general, let us consider the example of the physics class once again. Suppose now that the score is an integer between 0 and 30 for each of the students, and we want to know the average score. Let us enumerate the students from 1 to 100 and let the score of student i be S_i. Then, by definition, the average score is (in general, we denote the average of a quantity Q by $< Q >$)

$$< S >= \frac{\sum_{i=1}^{100} S_i}{100} \tag{15}$$

The order in which we add all the different scores does not matter, so we may choose to first add all the scores of 0, then 1, and so on, up to scores of 30. Let us assume that the number of students with score x is n_x (of course, it could be 0 for some of the scores). Then,

if we just sum add the scores of students with a given score x, we just get $n_x x$. To get the total sum, we have to add $n_x x$ for all possible scores. We can rewrite the average score as

$$< S >= \frac{\sum_{x=0}^{30} n_x x}{100} = \sum_{x=0}^{30} \frac{n_x x}{100} = \sum_{x=0}^{30} \frac{n_x}{100} x$$

But let us recall that, by definition, the ratio of the number of students with a given score x to the total number of students is the probability of a random student to have score x:

$$P(x) = \frac{n_x}{100}$$

So we can rewrite the average score in the following way

$$< S >= \sum_{x=0}^{30} P(x)x \tag{16}$$

This way of calculating the average can be easily generalized, since it depends neither on the numbers of different events nor on the total number of events, it only depends on the probabilities of all different possibilities. So we can consider an experiment where we are measuring some quantity x, and all the possible outcomes are x_1, x_2, ... , x_n. If we denote the probability of the outcome x_i to be $P(x_i)$ then we can write the average of x as

$$< x >= \sum_{i=1}^{n} P(x_i)x_i \tag{17}$$

We may also be interested in calculating the average of some given function of x, call it $f(x)$. The different possible values of $f(x)$ are $f(x_1)$, $f(x_2)$, ... , $f(x_n)$, and the probability $P(f(x_i))$ of the value $f(x_i)$ is, of course, the same as for x to have the value x_i, i.e. $P(x_i)$

$$P(f(x_i)) = P(x_i)$$

We can now use the rule (17) to find the average of $f(x)$

$$< f >= \sum_{i=1}^{n} P(f(x_i))f(x_i)$$

$$< f >= \sum_{i=1}^{n} P(x_i)f(x_i) \tag{18}$$

Using (18), it is easy to see that the average of a sum of two functions is the sum of the averages. Indeed:

$$< f + g >= \sum_{i=1}^{n} P(x_i)(f(x_i) + g(x_i)) = \sum_{i=1}^{n} P(x_i)f(x_i) + \sum_{i=1}^{n} P(x_i)g(x_i) =< f > + < g > \tag{19}$$

Similarly, the average of a constant times a function is equal to the constant times the average of the function:

$$< cf >= \sum_{i=1}^{n} P(x_i) cf(x_i) = c \sum_{i=1}^{n} P(x_i) f(x_i) = c < f > \tag{20}$$

where c is any number. Finally, if the function f is a constant itself, i.e. it is the same for all the values of x, then the average is clearly equal to that constant.

Once we know the average, the next question of interest is how close a random measurement is to the average value. For example, if all of the students in our example of physics class get the same score of 20, then the average is clearly 20, but if half of the students gets 30 and the other half gets 10, then the average is still going to be 20. Here the difference is that in the first case we know with certainty that any randomly chosen student will have a score of 20, while in the second case a randomly chosen student will have a score that differs from the average by 10. So by just stating that the average is 20 we do not get any idea about how close a random measurement will be to that average value. That is why we want to know "on average, how much a random measurement will differ from the average value". We want another quantity that will describe this. The simplest thing to do would be to take the average of the difference from the average value, i.e. for a general function f to consider

$$\sigma_f =< f - < f >>$$

However, using the rules that we derived to calculate averages, we get

$$\sigma_f =< f > - << f >>=< f > - < f >= 0$$

where we used the fact that $<< f >>=< f >$ since $< f >$ is just a constant number. It is not hard to understand the reason of this failure. When we take $f - < f >$ it can be both positive and negative, and all the positive values cancel the negative ones when averaged out. What we are really interested in is the absolute value of that difference, so we can simply take the square of that quantity to make sure it is always positive and then take the average, and afterwards take the square root. We will call that quantity the *standard deviation* of f, or the *uncertainty* of f:

$$\sigma_f = \sqrt{< (f - < f >)^2 >} \tag{21}$$

Using our rules, let us simplify (21) a little bit:

$$\sigma_f^2 =< (f - < f >)^2 >=< f^2 - 2f < f > + < f >^2 >=< f^2 > - < 2f < f >> + << f >^2 >$$

$$\sigma_f^2 =< f^2 > -2 < f >< f > + < f >^2 =< f^2 > -2 < f >^2 + < f >^2$$

$$\sigma_f^2 =< f^2 > - < f >^2 \tag{22}$$

Let us calculate the standard deviation of the scores of the students for the two examples mentioned above. When all of the students get a score of 20, the probability of that score is 1, while for any other score it is 0, so the average score is

$$< S >= P(20) \times 20 + 0 = 1 \times 20 = 20$$

Let us now calculate $< S^2 >$:

$$< S^2 >= P(20) \times 20^2 + 0 = 1 \times 400 = 400$$

and finally, using (22) we can calculate the uncertainty of the score:

$$\sigma_S^2 =< S^2 > - < S >^2 = 400 - 20^2 = 0$$

and that is what we expect since the score of any student is 20 without any uncertainty. Now let us consider the second case. The probability of the score 30 is 0.5, so it is for the score 10, and it is 0 for any other score. For the average we have

$$< S >= P(30) \times 30 + P(10) \times 10 + 0 = 0.5 \times 30 + 0.5 \times 10 = 20$$

while for $< S^2 >$ we get

$$< S^2 >= P(30) \times 30^2 + P(10) \times 10^2 = 0.5 \times 900 + 0.5 \times 100 = 500$$

so the uncertainty squared is given by

$$\sigma_S^2 =< S^2 > - < S >^2 = 500 - 400 = 100$$

and the uncertainty is

$$\sigma_S = \sqrt{100} = 10$$

just as we would expect.

4 Probability distributions

Returning to the example of the physics class, let us now assume that the professor is using a more sophisticated grading system and each grade can be any real number between 0 and 30, not just integers (for example, the first problem may be worth π points). Now we do not have only finitely many possible outcomes, that is why it makes not a lot of sense to talk about the probability of a given exact score, that would be 0 if we take the reasonable assumption that the probabilities of scores very close to each other should be more or less the same. Indeed, let us consider the scores "very close" to say 20. There are infinitely many of them, we can take for example 20.1, 20.01, 20.001, and so on. So if they have the same non-zero probability then their sum would be infinity, which is not allowed since the normalization condition requires the sum of all possible different probabilities to be 1. However, we can talk about the probabilities of scores in some small range of the given score. It is reasonable to ask "What is the probability that a randomly chosen student will have a score in the range 20 to 20.1?" Let us denote the probability of the score between a and b by $P(a, b)$. By our assumption that very close scores should have similar probabilities, we do expect that, for example

$$P(20, 20.1) \approx P(20.1, 20.2)$$

since these are two ranges of the same size and close to each other. Let us now use our addition rule of probabilities to calculate $P(20, 20.2)$:

$$P(20, 20.2) = P(20, 20.1) + P(20.1, 20.2) \approx P(20, 20.1) + P(20, 20.1) = 2P(20, 20.1)$$

So by doubling the small range we doubled the probabilities. It is not hard to understand that the probability of a score around a given value is proportional to the small range itself that we take around that value:

$$P(20, 20 + \epsilon) \propto \epsilon$$

where ϵ is a small number. We will call the proportionality coefficient the *probability distribution* at value 20 and will denote it by $p(20)$:

$$P(20, 20 + \epsilon) = p(20)\epsilon$$

We can see that the bigger the probability distribution at 20 the more probable are the scores around 20. In the same way we define the probability distribution at any real-valued score x between 0 and 30, so the probability distribution becomes a function of the score defined on the whole range of scores $[0, 30]$. Thus, in general

$$P(x, x + \epsilon) = p(x)\epsilon \tag{23}$$

Given the probability distribution $p(x)$ we should be able to find the probability of the score in any (not necessarily small) range $[a, b]$. We choose a small ϵ and divide the range $[a, b]$ into small ranges of size ϵ. Let

$$t_0 = a, t_1 = a + \epsilon, t_2 = a + 2\epsilon, \ldots, t_n = b$$

Then using our addition rule of probabilities we can write:

$$P(a, b) = P(t_0, t_n) = P(t_0, t_1) + P(t_1, t_2) + \cdots + P(t_{n-1}, t_n) = \sum_{i=1}^{n} P(t_{i-1}, t_i)$$

But each of the ranges $[t_{i-1}, t_i]$ is of small size ϵ so we can use (23) to get

$$P(t_{i-1}, t_i) = p(t_{i-1})\epsilon$$

$$P(a, b) = \sum_{i=1}^{n} p(t_{i-1})\epsilon$$

In the limit $\epsilon \to 0$ the sum above becomes the integral of the function $p(x)$ from a to b (that is the Riemann definition of the integral), so finally we get

$$P(a,b) = \int_a^b p(x)\,dx \tag{24}$$

Using (24), it is not hard to get the normalization condition in this case. Indeed, any score is in the range $[0, 30]$ with certainty, so $P(0, 30) = 1$ and the normalization condition takes the form

$$\int_0^{30} p(x)\,dx = 1 \tag{25}$$

Finally, let us calculate the average score. We again divide the whole range into small pieces:

$$t_0 = 0, t_1 = 0 + \epsilon, t_2 = 0 + 2\epsilon, \ldots, t_n = 30$$

and use the general rule (17) to calculate the average score:

$$< x >= \sum_{i=0}^{n} P(t_{i-1}, t_i)t_{i-1} = \sum_{i=0}^{n} p(t_{i-1})\epsilon t_{i-1} = \sum_{i=0}^{n} [p(t_{i-1})t_{i-1}]\epsilon$$

In the limit $\epsilon \to 0$ the sum above again becomes an integral, but now of the function $p(x)x$. We get

$$< x >= \int_0^{30} p(x)x\,dx \tag{26}$$

Now that we have a good understanding of probability distributions for our example, it is straightforward to generalize the idea. Consider an experiment where a quantity x is being measured, and it can have any real value inside the range $[A, B]$ (where A may be equal to $-\infty$ and B may be equal to $+\infty$). We define the probability distribution $p(x)$ on the range $[A, B]$ such that the probability of getting a value in a small neighborhood ϵ of a point x is given by

$$P(x, x + \epsilon) = p(x)\epsilon \tag{27}$$

Then the probability of getting a value inside any range $[a, b]$ (not necessarily small) is given by

$$P(a,b) = \int_a^b p(x)dx \tag{28}$$

The normalization condition takes the form

$$\int_A^B p(x)\,dx = 1 \tag{29}$$

The average value of x is given by

$$< x >= \int_A^B p(x)x\,dx \tag{30}$$

while the average of any function $f(x)$ is given by

$$< f >= \int_A^B p(x) f(x)\, dx \tag{31}$$

The formula (22) is still valid for the standard deviation:

$$\sigma_f^2 =< f^2 > - < f >^2 \tag{32}$$

but here we will need to evaluate two integrals, one for $< f^2 >$ and another one for $< f >$, before we can calculate the standard deviation.

Comparing the equations above with the corresponding equations for the case of finitely many possible outcomes, we can see that all we had to do was to consider probability distributions instead of probabilities, and to replace the finite sums by integrals!

5 Normal distribution

In this section we will consider one of the most important probability distributions in physics and statistics, called *normal distribution* or *Gaussian distribution*, and will apply our general formulas to do actual calculations for this specific case. The normal distribution is defined on the whole range $(-\infty, +\infty)$ of real numbers and is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \tag{33}$$

where μ and σ are two constants, $\sigma \neq 0$ (we will understand their significance later). Let us examine the graph of this distribution, fig. 1, which has the shape of bell. We can see that the distribution reaches its maximum at point μ, and decreases as we go away in both directions. So if an experiment has the normal distribution as the probability distribution of its outcomes, then we can see that most of the measurements will result in values close to μ, evenly distributed on both sides of μ, therefore we do expect that the average of x should be equal to μ. The other thing that we can notice is that the "width" of the curve is proportional to σ. The smaller σ is, the denser will the outcomes be clustered around μ, so we do expect that the standard deviation should be proportional to σ (we will explicitly calculate the average and the standard deviation of x later in this section, and will get that the standard deviation is exactly equal to σ, that is why we chose the letter σ to denote that constant). We can see that normal distribution is very useful to describe experiments the outcomes of which are clustered around some average value. For example, when we are measuring some quantity in the lab we usually get experimental errors for various reasons, so if we measure the same thing multiple times, then we get values that are clustered around the correct value, but not exactly equal to it. The bigger our experimental errors are, the wider these values are spread around the correct value. The probability distribution describing the values that we get is normal, with the average equal to the correct value, and the standard deviation equal to the uncertainty of our measurements. Formally, the normal distribution is what we get if we add a large number of random variables[1]. That clarifies the reason why

[1]This statement is the so-called *central limit theorem* in statistics, which we will not prove here.

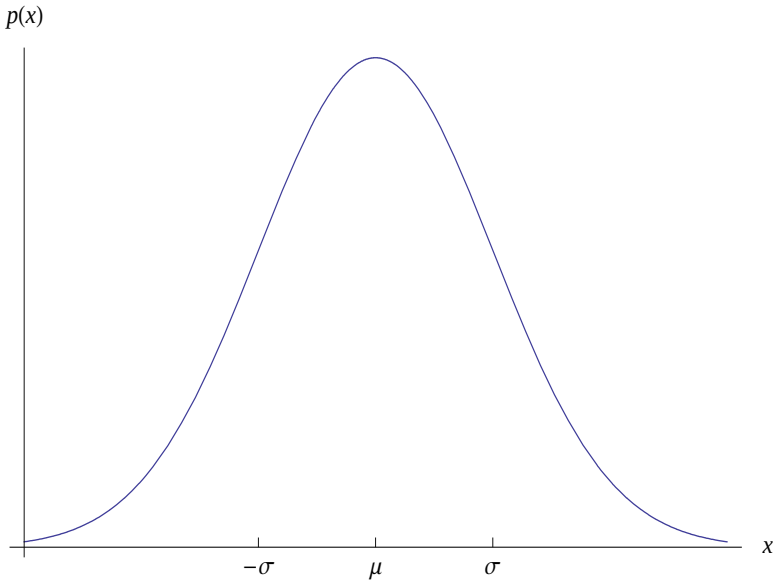

Figure 1: Normal distribution.

the experimental measurements of a physical quantity are distributed normally, since there are usually a large number of unknown sources of error, which are added together.

Let us now evaluate some integrals that will be useful in calculations with Gaussian distributions. Consider first

$$I_0(\alpha) = \int_{-\infty}^{+\infty} e^{-\alpha x^2}\, dx \tag{34}$$

where α is any positive number. We will need to use a non-obvious trick to calculate this integral. The variable x in the equation above is just a dummy variable of integration, so we are free to call it anything else. Let us rewrite the same equation but with x replaced by y:

$$I_0(\alpha) = \int_{-\infty}^{+\infty} e^{-\alpha y^2}\, dy \tag{35}$$

Now let us multiply both sides of equations (34) and (35) together. We get

$$(I_0(\alpha))^2 = \int_{-\infty}^{+\infty} e^{-\alpha x^2}\, dx \int_{-\infty}^{+\infty} e^{-\alpha y^2}\, dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dx\, dy\, e^{-\alpha(x^2+y^2)} \tag{36}$$

We treat x and y in (36) as cartesian coordinates and then switch to polar coordinates:

$$x = r\cos\phi$$

$$y = r\sin\phi$$

Then the integration measure becomes (just by calculating the Jacobian of the coordinate transformation above):

$$dx\, dy = r\, d\phi\, dr$$

We also have

$$x^2 + y^2 = r^2$$

Plugging in (36) we get

$$(I_0(\alpha))^2 = \int_0^{+\infty} r\,dr \int_0^{2\pi} d\phi\, e^{-\alpha r^2}$$

The integrand does not depend on ϕ, so the integral over ϕ just gives a factor of 2π:

$$(I_0(\alpha))^2 = 2\pi \int_0^{+\infty} r\,dr\, e^{-\alpha r^2}$$

On the other hand

$$r\,dr = \frac{1}{2} d(r^2)$$

so we get

$$(I_0(\alpha))^2 = \pi \int_0^{+\infty} d(r^2)\, e^{-\alpha r^2}$$

We make a change of the integration variable $r^2 = t$ to finally get an integral that is straightforward to evaluate:

$$(I_0(\alpha))^2 = \pi \int_0^{+\infty} dt\, e^{-\alpha t} = \pi \frac{e^{-\alpha t}}{-\alpha}\bigg|_{t=0}^{+\infty} = \frac{\pi}{\alpha}$$

So finally

$$I_0(\alpha) = \int_{-\infty}^{+\infty} e^{-\alpha x^2}\,dx = \sqrt{\frac{\pi}{\alpha}} \tag{37}$$

Now let us consider a more general integral, namely

$$I_n(\alpha) = \int_{-\infty}^{+\infty} x^n e^{-\alpha x^2}\,dx \tag{38}$$

for any integer $n \geq 0$ (clearly this definition agrees with our definition of I_0 above for $n = 0$). For odd n the function $x^n e^{-\alpha x^2}$ is an odd function, i.e. changes the sign if the sign of x is changed, so the integral becomes 0 (the integral of any odd function in any range $[-A, A]$, which is symmetric around 0, is 0). For even n we can get the result by simply taking the derivative of $I_0(\alpha)$ with respect to α a few times. Indeed, let us consider

$$\frac{dI_0(\alpha)}{d\alpha} = \frac{d}{d\alpha}\int_{-\infty}^{+\infty} e^{-\alpha x^2}\,dx = \int_{-\infty}^{+\infty} \frac{d}{d\alpha} e^{-\alpha x^2}\,dx = \int_{-\infty}^{+\infty} (-x^2) e^{-\alpha x^2}\,dx = -I_2(\alpha)$$

$$I_2(\alpha) = -\frac{dI_0(\alpha)}{d\alpha} \tag{39}$$

It is easy to see that if we take another derivative (with a minus sign), then we get $I_4(\alpha)$ and so on. In general

$$I_{2n}(\alpha) = (-1)^n \frac{d^n I_0(\alpha)}{d\alpha^n} \tag{40}$$

But these can be evaluated using the result that we have already obtained for $I_0(\alpha)$. For our future use, let us calculate $I_2(\alpha)$:

$$I_2(\alpha) = -\frac{dI_0(\alpha)}{d\alpha} = -\frac{d}{d\alpha}\sqrt{\frac{\pi}{\alpha}} = -\sqrt{\pi}\frac{d\alpha^{-1/2}}{d\alpha} = \frac{1}{2}\sqrt{\pi}\alpha^{-3/2} = \frac{1}{2}\sqrt{\frac{\pi}{\alpha^3}} \tag{41}$$

Let us now get back to the analysis of the normal distribution. The first thing that we want to check is if it satisfies the normalization condition. We have

$$\int_{-\infty}^{+\infty} p(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}\,dx$$

The first step in evaluating integrals with Gaussian distributions is the change of variable $x - \mu = t$. Clearly $dx = dt$ and t varies in the same range $(-\infty, +\infty)$. Let us also denote $\alpha = 1/(2\sigma^2)$. We get

$$\int_{-\infty}^{+\infty} p(x)\,dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-\alpha t^2}\,dt = \frac{1}{\sqrt{2\pi\sigma^2}} I_0(\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}}\sqrt{\frac{\pi}{\alpha}} = \sqrt{\frac{1}{2\sigma^2\alpha}} = 1$$

So indeed, the normalization condition is satisfied. In fact, when writing a probability distribution, the overall constant factor (e.g. $\frac{1}{\sqrt{2\pi\sigma^2}}$ in our case) is determined by the normalization condition[2]. Let us now calculate the average and the standard deviation of x with the normal distribution $p(x)$:

$$< x >= \int_{-\infty}^{+\infty} xp(x)\,dx = \int_{-\infty}^{+\infty} x\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}\,dx$$

We again do the change of variable $x - \mu = t$ and denote $\alpha = 1/(2\sigma^2)$. We get

$$< x >= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (t+\mu) e^{-\alpha t^2}\,dt = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} te^{-\alpha t^2}\,dt + \mu\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-\alpha t^2}\,dt$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} I_1(\alpha) + \mu \cdot 1 = 0 + \mu = \mu$$

[2]One important application of that is the determination of the overall constant factor of the so-called wave functions in quantum mechanics. The wave function is a complex-valued function of the position such that its squared absolute value gives the probability distribution of the position of the particle under consideration. In quantum mechanics the wave-function is determined by solving a linear differential equation, called the *Schrödinger equation*, which means that if we have a solution, then any constant times that solution is again a solution. So by the physical theory itself there is no way of determining the overall constant factor of the wave-function, and one has to impose the normalization condition on the probability distribution associated with the wave-function to determine that constant.

just as we expected. For evaluating the standard deviation of x we first need to calculate $< x^2 >$. We will use the same change of variable and α below:

$$< x^2 >= \int_{-\infty}^{+\infty} x^2 p(x)\, dx = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (t+\mu)^2 e^{-\alpha t^2}\, dt$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (t^2 + 2\mu t + \mu^2) e^{-\alpha t^2}\, dt$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} t^2 e^{-\alpha t^2}\, dt + 2\mu \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} t e^{-\alpha t^2}\, dt + \mu^2 \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-\alpha t^2}\, dt$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} I_2(\alpha) + 2\mu \frac{1}{\sqrt{2\pi\sigma^2}} I_1(\alpha) + \mu^2 \cdot 1 = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{2} \sqrt{\frac{\pi}{\alpha^3}} + 0 + \mu^2$$

$$= \frac{1}{2\alpha} \sqrt{\frac{1}{2\sigma^2\alpha}} + \mu^2 = \frac{1}{2\alpha} \cdot 1 + \mu^2 = \frac{2\sigma^2}{2} + \mu^2 = \sigma^2 + \mu^2$$

Finally, the standard deviation of x is given by

$$\sigma_x^2 =< x^2 > - < x >^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

$$\sigma_x = \sigma$$

as promised above.