Basic Statistics

1 Introduction

What is statistics? It has a broad definition. In this lab course, we will define it narrowly and simply as: the analysis of data.

Define data: the set (S) of the results (s_1, s_2, ...) of an experiment (E) after repeating it for a certain times (n).

e.g. E = throwing a dice
$n = 7$
$s_1 = 1$, $s_2 = 2$, $s_3 = 3$, $s_4 = 5$, $s_5 = 5$, $s_6 = 6$, $s_7 = 2$
$S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$

The key in this definition is what we call an experiment:

- it gives random result
  $$R = \{r_1, r_2, etc.\}$$
  e.g. for the dice, $R = \{1, 2, 3, 4, 5, 6\}$
  caution: this is the set of all possible result, so each element is different,
  to be distinguished with the data defined before, where the same element can appear more than once

- there is "a rule" governing the probability of appearance of each result $P(r)$:
  e.g. a perfect dice, $P(r) = 1/6$ for each result;
  a cheating dice with lead inside so that its center of mass being
  biased opposite to the face of No.3: $P(r_3) > 1/6$
  this "rule" is also called probability distribution

Define analysis of data: from the set of experimental results we get, to draw some conclusion about the experiment itself.

e.g. throwing the cheating dice for 1000 times
for 200 times we get No.3 → $P(r_3) = 1/5$
the centre of mass biased opposite to face No.3

This short section gives a qualitative description of the concepts in statistics. Later we will talk about each concepts in detail and illustrate them with more examples.

2 Probability distribution

The "rule" or the probability distribution that governs the random result of experiment is usually not arbitrary. It follows some pattern. More specifically, it's a function of the result with some parameters $a, b, ...$

$$P(r) = f(r|a, b, ...)$$

e.g. for a perfect dice, there is no parameter and $f(r) = 1/6$ for all r
for a cheating dice, $P(r) = f(r|x, y, z)$
where (x, y, z) is the coordinate of the centre of mass.

There are several typical probability distributions in everyday life and in experiment. They will be the topics of this section.

2.1 Properties of probability

Before introducing typical probability distributions, some basic properties of probability need to be mentioned.

First, the probabilities of all possible results should be summed up to 1.

$$\sum_{r}^{r\in R} P(r) = 1$$

Second, assume there are two independent experiments E_1 and E_2, each has possible result $r_1 \in R_1$ and $r_2 \in R_2$ and probability distribution $P_1(r_1)$ and $P_2(r_2)$. Then the combined experiment of the two has possible result of:

$$r = r_1 \times r_2 \in R_1 \otimes R_2$$

and the probability distribution is:

$$P(r) = P(r_1 \times r_2) = P_1(r_1) \times P_2(r_2)$$

This is called the multiplication rule.

e.g. throwing two dices a and b
$r = (a, b) \in R = \{$(1,1), (1,2), (2,1), ...$\}$
$P((1, 2)) = P_a(1) \times P_b(2) = 1/36$

Besides, there are two common quantities that can be used to characterize a probability distribution $P(r)$: the mean value $\bar{r}$ and the standard deviation σ. They are defined as following:

$$\bar{r} = \sum_{r} P(r) \times r$$
$$\sigma^2 = \sum_{r} P(r) \times (r - \bar{r})^2$$

The mean value represents the average value of result r if we repeat the experiment for infinite times. It is the location of the center of the probability distribution. The standard deviation characterizes the broadness of the probability distribution.

2.2 Binomial distribution

Let's start from a simple example: tossing a coin and counting the number of head.

In this case, the random result contains 2 elements: $R = \{$0 head, 1 head$\}$. And the probability distribution is:

$$P(0) = P(tail) = 0.5$$
$$P(1) = P(head) = 0.5$$

Now we toss the coin for n times.There are more possible results than tossing the coin for once:

$$R = \{0\ head,\ 1\ head,\ 2\ heads,\ ...,\ n\ heads\}$$

What is the probability distribution $P(r)$ $(r \in R)$ now?

$P(r)$ can be written as $f(m|n, p)$, which is the probability to get m heads up out of the n tosses. m $(0 \leq m \leq n)$ is the index for the random result $r \in R$ and p is the chance of head up in 1 toss. m is the variable, while n and p are the parameters.

To calculate $f(m|n,p)$, consider the case of the 1st to the mth tosses giving head up and the following n - m tosses giving tail up. The probability of this result r' to happen is, according to the multiplication rule:

$$\begin{aligned} P(r') =& \underbrace{P(head) \times P(head) \; ... \; P(head)}_{m \text{ factors}} \times \underbrace{P(tail) \times P(tail) \; ... \; P(tail)}_{n-m \text{ factors}} \\ =& P(head)^m P(tail)^{(n-m)} \\ =& p^m (1-p)^{n-m} \end{aligned}$$

However, to get m heads, it is not necessarily that the first m tosses give head. It can be any m tosses out of the n tosses giving head and each of these results have the same probability as above. If there are in total N such results:

$$P(r) = N \times P(r')$$

N can be calculated from the knowledge of permutation:

$$N = C_n^m = \frac{n!}{(n-m)!m!}$$

where $n!$ is defined as $n! = 1 \cdot 2 \cdot ... \cdot n$.

Summarizing the above discussion, we get the final formula for the probability distribution of tossing a dice for n times:

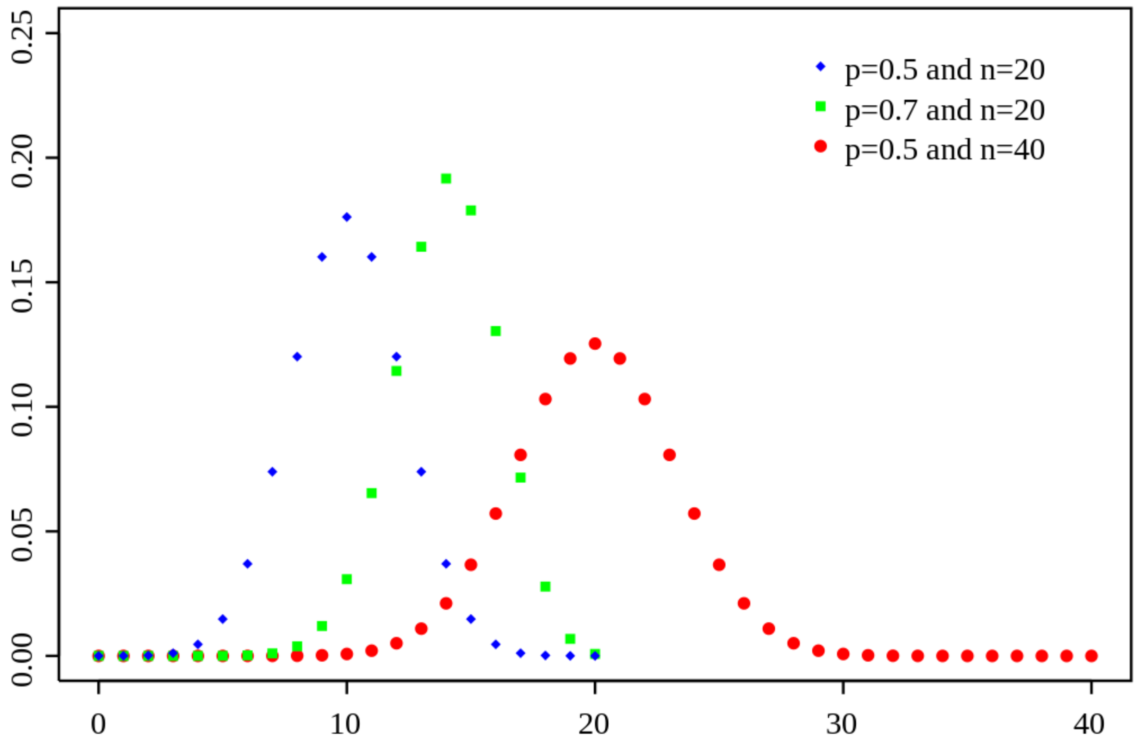
$$P(r) \equiv f(m|n,p) = C_n^m p^m (1-p)^{n-m}, \; m = 0, 1, ..., n$$

The distribution $f(m|n)$ is called binomial distribution. It describes a more general type of question than tossing a coin: the number of successes m in a sequence of n independent experiments with the probability p of one experiment being success.

The mean value and standard deviation of binomial distribution can be calculated using their definitions shown before:

$$\begin{aligned} \bar{m} &= np \\ \sigma &= \sqrt{np(1-p)} \end{aligned}$$

Below is a plot showing the binomial distribution in different parameter values of n and p. We can find the mean values in each cases indeed represent the centers of the distributions. It can also be seen that binomial distribution is symmetric in terms of its mean value. The red distribution's standard deviation, according to the definition, is calculated to be $\sqrt{2}$ times that of the blue distribution. This relationship can also be read from the plots.



2.3 Poisson distribution

Poisson distribution is actually the extreme case of binomial distribution when the number of trials n is infinite large while the success probability p is infinite small so that the mean value np of the binomial is finite.

In this case, the Poisson probability distribution will not depends on the two parameters n and p as in the case of binomial distribution, but just one parameter which is the mean value defined as below:

$$\lambda = np$$

With the above assumption and definition, the Poisson distribution can be derived from the binomial distribution as:

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Here the variable k is used to represent the random result, to be distinguished from m used for binomial distribution. And λ is the parameter of the distribution.

For example, for $\lambda = 4$, the probability distributions can be calculated as:

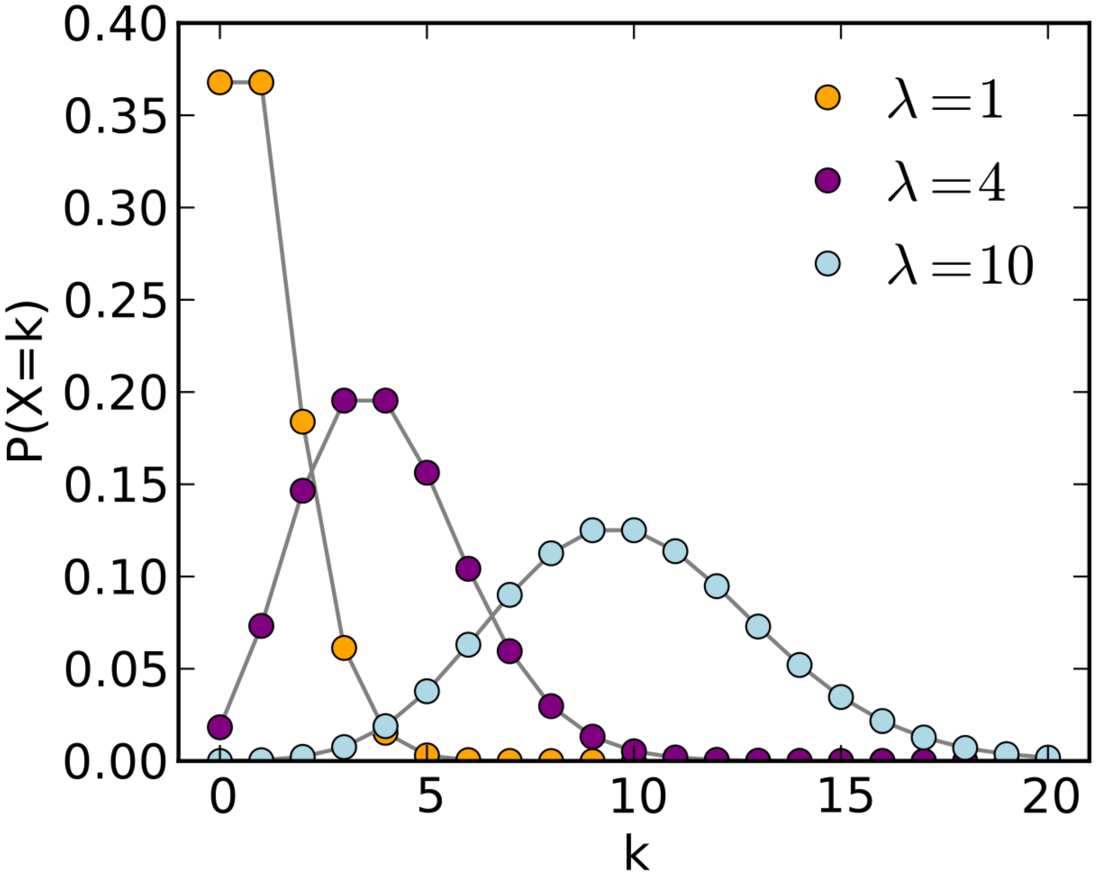
	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$P(k\|4)$	0.018	0.073	0.147	0.195	0.195	0.156
	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$	$k = 11$
$P(k\|4)$	0.104	0.060	0.030	0.013	0.005	0.002

Use the definition of the mean value and standard deviation, they can be calculated for the Poisson distribution:

$$\bar{m} = \lambda$$

$$\sigma = \sqrt{\lambda}$$

Below is a plot showing the Poisson distribution in different parameter values of λ. If you compare the Poisson distribution with $\lambda = 10$ and the binomial distribution with $n = 20$ and $p = 0.5$, you can find that they are quite similar but with obvious difference.



Mathematically, the trials n should be infinite large and the probability p should be infinite small. But in practise, we can use the Poisson distribution to describe experiments that have very large number of trials and very tiny probability of success.

For example, scientists in high energy physics usually collide two beams of particles to create some super rare interaction. In this case, the experiment is the collision of two particles from the two beams. The success is defined as the happening of the rare interaction. If each collision is independent from each other, the number of success in a certain amount of collisions will distribute as a binomial according to the definition of binomial distribution. However, as the name "super rare" suggests, the chance of having this interaction to happen is very rare, which means p is very tiny.

And again, since the chance is so small, we need to make the two beams so intense and collide them so frequently that the number of trials n gets extremely large. So the two conditions for using the Poisson distribution instead of binomial distribution are satisfied and the probability distribution of the number of rare interactions we observe can be described by Poisson.

2.4 Gaussian distribution

Before introducing Gaussian distribution, we need to extend the concept of probability distribution. In the above, the random results of the experiments we discussed are all discrete. What if they are continuous?

In this case, it's hard to define the probability for a particular value $P(r)$, like what we did for discrete random result, since there are infinite number of possibilities if the results are continuous and in turns, the probability for one particular result should be infinite small.

What we can do now is to define the density of the probability $p(x)$ at the vicinity of the particular result x. (We used x to represent the continuous random result, to distinguish with the case of discrete random result r). So when we divide equally the range of the continuous random result $[a, b]$ into n finite pieces, we can talk about the probability of a particular piece $P(i)$ as in the case of discrete random result:

$$P(i) = \int_{x_i}^{x_i+dx} p(x)dx$$

where x_i is the starting point of the piece i an dx the length of the piece.

$p(x)$ is called the probability density distribution or probability density function (p.d.f). It has the usual properties of the probability introduced before:

$$\int p(x)dx = 1$$

and if z is the combination of x and y, which are independent from each other:

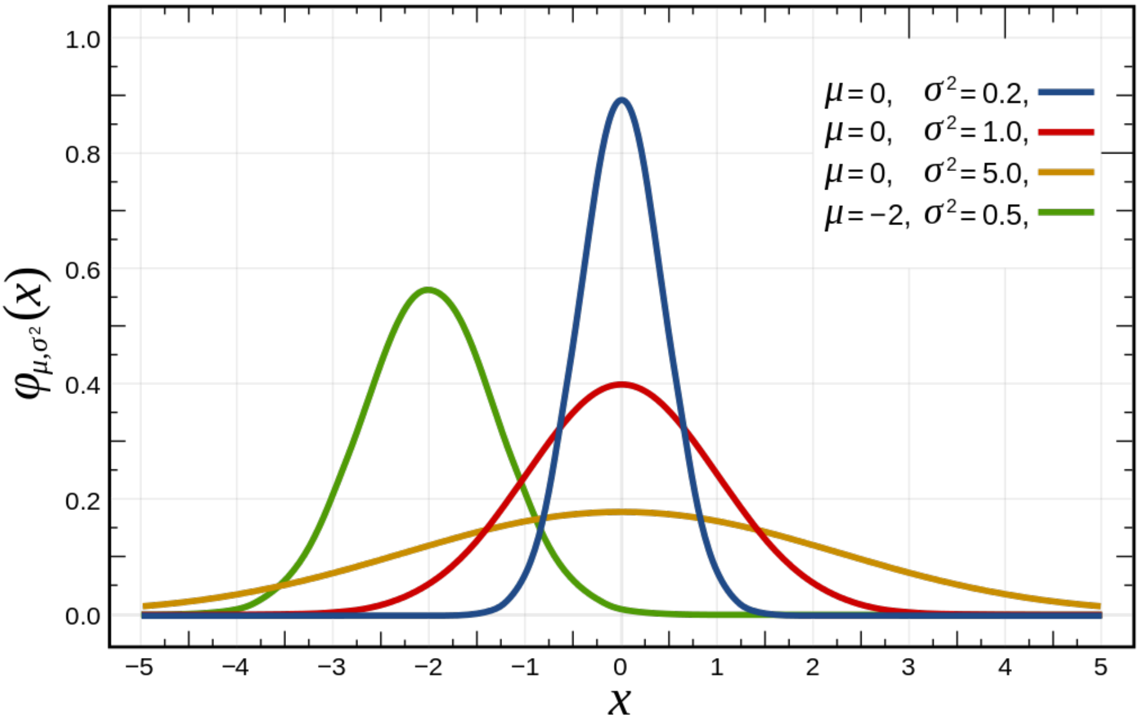
$$p(z) = p(xy) = p(x)p(y)$$

Now we can introduce the Gaussian distribution, which is used to describe the probability density distribution of continuous random result:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean value and σ the standard deviation (which is easy to validate using the corresponding definitions). The factor before the exponential is to normalize the distribution so that the total probability is one.

Below is a plot showing the Gaussian distribution in different parameter values of μ and σ. As the definitions of μ and σ suggest, they represent the location of the center and width of the distribution.



If there is a very sophisticated experiment, whose result is the sum of many independent random processes, the probability density distribution of this result will be the Gaussian.

For example, in high energy physics, the energy of a particle is usually measured by the so-called calorimeter. The calorimeter is an array of small sub-calorimeters and the energy of the particle is deposited in a bunch of these sub-calorimeters. So the measured energy of the particle is the sum of the measured energy in each sub-calorimeters. However, the measurement in each sub-calorimeter itself is a random process (e.g. due to electronic noise). So the particle energy is the sum of many independent random processes and it will distribute like a Gaussian.

More generally speaking, the Gaussian is usually used to describe the variation of the measured value x from the true value μ when the measurement resolution is σ.

3 Measurement and its error

3.1 Measurement

In physical measurement, stochastic processes are always involved. Each time when we measure something, we are actually performing an experiment which gives random result and the result follows some sort of probability distribution $P(r)$ or probability density distribution $p(x)$.

If the measurement result is random in itself, then what are we measuring? It must be something that is certain. And the answer lies at the purpose of the measurement: usually we perform a measurement (for one time or multiple times) to infer from the measured value(s) about the value of the parameter in $P(r)$ or $p(x)$. In short, measurement is a kind of parameter inference.

For example, when we use a ruler to measure the length of a table, we intuitively think that the read from the tick mark tells us the true length of the table. This is already some kind of inference (although quite simple). To clearly see the inference we made implicitly, we need to know that the measured length is a random result following a Gaussian distribution whose parameter μ is the true length and σ the resolution of the ruler. And we are inferring that the true length takes the value of the measured length. This is usually valid since σ is always so small that we don't care about the possible difference between the true length and the measured length.

3.2 Measurement error

However, in some physical experiment, the inference is not so easy to make as measuring the length of a table. People need to carefully analyze the observed data. And more importantly, we should assign error to our inferred parameter value.

A nice example is the measuring of the decay rate of some isotope. If in time period T, we observed N decay events, then the measured decay rate is $\Gamma = N/T$. Is it the true decay rate?

To answer the question, we need to know that the number of observed decay events in time interval T follows a Poisson distribution. If assuming the true decay rate to be Γ_t, then the number of observed events N follows:

$$P(N|\lambda) = \frac{(\lambda)^N e^{-\lambda}}{N!}$$

where $\lambda = \Gamma_t \cdot T$ is the expected number of decay events that will be observed.

Then from the formula $\Gamma = N/T$, we know the measured decay rate Γ is also random, which means that it is not necessarily the true decay rate $\Gamma_t = \lambda/T$ due to the fluctuation of N! Thus we are uncertain about whether we should claim that the measured decay rate is indeed the true decay rate.

To express this uncertainty about the real position of the true value, we introduce the error of the measurement or error of the inferred parameter σ. And in this case, we say $\Gamma_t = \Gamma \pm \sigma_\Gamma$. σ is defined to be the standard deviation of the probability distribution or probability density distribution (that's why they share the same denotation) of the measured result. In this example, the measured result is Γ and its distribution is a Poisson scaled by a constant time T. So its uncertainty is: $\sqrt{\lambda}/T$, which is approximately $\sqrt{N}/T$. So the true decay rate is measured to be $N/T \pm \sqrt{N}/T$. (This is just a simplified definition of measurement error. For a more rigorous one, we need to introduce the concept of likelihood, which is too complicated for this course.)

Besides, we usually introduce the concept of relative error which is useful. In this example, it is $\sigma_\Gamma/\Gamma = \sqrt{N}/N = 1/\sqrt{N}$, which means the more decay event we observed, the more precise measurement we can make: e.g when $N = 100$, the relative uncertainty is 10%, while when $N = 10000$, it becomes 1% (so to improve the precision by 1 order, you need to increase the size of the data set by 2 orders).

3.3 Error propagation

In physical measurement, it is often required to derive a physical quantity from the combination of several measured quantities. Consider the simple case below:

$$z = f(x, y)$$

where z is the final quantity we are interested in, x and y are two independently measured quantities, and f is the function relationship between z and x and y. If x and y are measured to be $x_0 \pm \sigma_x$ and $y_0 \pm \sigma_y$, what is the uncertainty of z: σ_z?

This is a problem of error propagation. It can be proved that:

$$\sigma_z = \sqrt{(\frac{df}{dx}\sigma_x)^2 + (\frac{df}{dy}\sigma_y)^2}$$

Take the previous decay rate measurement as example. The expected number of event decay is measured to be $N \pm \sqrt{N}$ and if the total time is also measured with some uncertainty $T \pm \Delta T$, then the decay rate, which is:

$$\Gamma = \frac{N}{T}$$

has the uncertainty of:

$$\sigma_\Gamma = \sqrt{(\frac{1}{T}\sqrt{N})^2 + (-\frac{N}{T^2}\Delta T)^2}$$

We can put some real numbers into the calculation to have a feeling of the error propagation. E.g. if $N = 100$ and $T = 10 \pm 0.5$, according to the formula, the relative uncertainty $\sigma_\Gamma/\Gamma = 11\%$, which is quite close to the relative uncertainty of N itself. While if $N = 10000$ and T is the same, the relative uncertainty σ_Γ/Γ will be 5.1%, which is close to the relative uncertainty of T. So we get the impression that the final propagated relative uncertainty is determined by the input random quantity which has dominating relative uncertainty.

Besides, there is a more efficient formula to calculate the propagated relative uncertainty if the function f is purely the product or ratio of several variables:

$$f(x_1, x_2, x_3, ...) = \frac{x_1 \cdot x_2 \cdot ...}{x_3 \cdot ...}$$

then we have:

$$\frac{\sigma_f}{f} = \sqrt{(\frac{\sigma_{x_1}}{x_1})^2 + (\frac{\sigma_{x_2}}{x_2})^2 + (\frac{\sigma_{x_3}}{x_3})^2 + ...}$$

Another interesting example is that when you want to measure something, like the length of a table, you can measure it for multiple times and take the average to reduce the measurement error. Intuitively this is correct. But what's the mathematic foundation behind it? This can be well explained by considering properly the error propagation.

Assume we measured the table length for n times, each time the result is $x_i \pm dx$, where x_i is the measured value varying each time and dx is the resolution of the ruler staying the same each time. So for the average:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

we can calculate its uncertainty by error propagation:

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{1}{n}\sqrt{dx^2 + dx^2 + ... + dx^2} \\ &= \frac{dx}{\sqrt{n}}\end{aligned}$$

So the error is reduced by a factor of $1/\sqrt{n}$!

3.4 Weighted mean

Let's review the last example, what if we group the first m measurements in one group and the rest $n-m$ the other ? In this case, the average of the two groups according to the formula are:

$$\begin{aligned}\bar{x}_m &= \frac{x_1 + ... + x_m}{m} \\ \bar{x}_{n-m} &= \frac{x_{m+1} + ... + x_n}{n-m}\end{aligned}$$

Then combining the results of the two groups again, the final average will be:

$$\begin{aligned}\bar{x} &= \frac{\bar{x}_m + \bar{x}_{n-m}}{2} \\ &= \frac{(n-m)(x_1 + .. + x_m) + m(x_{m+1} + ... + x_n)}{2m(n-m)}\end{aligned}$$

which is different from the case of averaging the measurements in one step if $n - m \neq m$!

There must be something wrong when $n - m \neq m$. Indeed, the group with less measurement should has less importance in the combination compared to the group with more measurements. But in the above combination, we treated them in equal foot.

To solve the problem, we introduce the concept of weighted arithmetic mean, which means we assign a smaller weight to the group with less measurements and larger weight to the one with more measurements:

$$\bar{x}_w = w_1 \cdot \bar{x}_m + w_2 \cdot \bar{x}_{n-m}$$

where $w_1 = m/n$ and $w_2 = (n-m)/n$. In this case, the recombined average is:

$$\begin{aligned}\bar{x} &= \frac{m}{n}\bar{x}_m + \frac{n-m}{n}\bar{x}_{n-m} \\ &= \frac{x_1 + .. + x_m + x_{m+1} + ... + x_n}{n}\end{aligned}$$

So we recovered the previous average.

Now we can define generally the weighted arithmetic mean as:

$$\bar{x}_w = \sum_i w_i x_i$$

From the error propagation formula, its error is:

$$\sigma_{\bar{x}_w} = \sqrt{\sum_i (w_i dx_i)^2}$$

Usually the weight is chosen as the inverse of squared error of individual measurement entering the averaging:

$$w_i = 1/\sigma_{x_i}^2$$

and then normalized to the total weights:

$$w_i = w_i / (\sum_j w_j)$$