6. Good Old Average

Consider the following problem. We have a large number of measurements for the same quantity and we want to use them to find a better estimate of this quantity than using one of the measurements alone. Assume that the measurements come from a robot which fired its sonar 10 times and got 10 different distances x_i for $i = 1..10$. We want to find the most accurate distance to the object in from of the sonar. Just about everybody would suggest the average of all the measurements

$$\hat{x} = \frac{\sum_{i=1}^{i=10} x_i}{10}$$

where the hat in $\hat{x}$ indicates that this quantity is an estimate. This is the best we can do given the absence of any other information about the sensor. So we can state that

$$\hat{x} = \bar{x}$$

where the bar in $\bar{x}$ means average.

Although the average is usualy the safest thing to use and quite often the best, it is not always the case that there is a single definition of "average". Consider a slightly more complicated example. We have a robot with shaft encoders (little thingies on the wheels of a robot that send "clicks" to the robot brain to tell it how far it has gone) that send a click every distance $S = 10cm$. A certain function in the robot controller receives a series of numbers t_1, t_2 etc which are the time intervals between the clicks, and we have to find a way to estimate the velocity $\hat{v}$ of the robot. There are two ways to apply the "average" idea we stated above.

$$\hat{v} = \frac{\sum_{i=1}^{N} \frac{S}{t_i}}{N}$$

and

$$\hat{v} = \frac{S}{\frac{\sum_{i=1}^{N} t_i}{N}}$$

in other words, average the velocities or average the times between the clicks. Some people will instinctively prefer the one over the other, but the ones that know better (or just do not care) will say "I don't know". One should do a rather sophisticated modeling of the system to decide which (and if) one of the above is best. If we cannot construct a rich enough model, then we can just rely on empirical evaluation (lots of experiments) or just plain instinct.

6.1. Average and Expected Value

These two are more or less the same thing, if you are a common mortal. Their principal difference is that the "Expected Value" sounds more sophisticated. A secondary but certainly fundamental difference is that the *average* is what we actually compute from a set of measurements, whereas the *expected value* is what we expect to get from a measurement. If, for example, we place the robot 30cm away from the wall, the expected value of the sonar reading will be 30cm, although none of the measurements will be exactly 30, most probably not even the average. But if we take many measurements the average will converge to 30 for a properly calibrated sonar sensor (or to be honest, for a sensor that shows some respect for elementary statistics).

Since the average usually converges to the expected value and the expected value is usually what we try to estimate, we cannot go very wrong by using the average as an estimate of the expected value, if we have more data than we need.

OK, we know how to take the average: we sum all up and divide by their number. How do we compute the expected value? The books say papoulis

$$\mu_x = \mathrm{E}\{x\} = \int_{-\infty}^{\infty} xp(x)dx \tag{6.1}$$

where $p(x)$ is the Probability Density Function (pdf) of x. That's not that useful though if we do not have a pdf or simply do not know how to compute integrals. In practice we often need neither (phewww). It is usually enough to know four rules for the expected value

(1) The expected value of a constant is itself e.g $\mathrm{E}\{c\} = c$.

(2) The expected value of the expected value of a measurement is the expected value of the measurement e.g. $\mathrm{E}\{\mathrm{E}\{x\}\} = \mathrm{E}\{x\}$

(3) The expected value of the sum of two measurements is the sum of the expected values (very much like the average) e.g. $\mathrm{E}\{x + y\} = \mathrm{E}\{x\} + \mathrm{E}\{x\}$.

(4) The expected value of the product of two measurements is the product of the expected values *if* the measurements are statistically independent (very much *unlike* the average) e.g. $\mathrm{E}\{x\,y\} = \mathrm{E}\{x\}\,\mathrm{E}\{x\}$

The discrete version of Eq. (6.1) for the expected value (which is for random variables whose value is always an integer like a pair of dice) is not very different

$$\mu_x = \mathrm{E}\{x\} = \sum_i ip_i \tag{6.2}$$

where p_i is the probability of outcome i.

As it will become apparent in the next sections it is useful to define not only the expected value of our measurement but the expected value of x^2, $(x - \bar{x})^2$ or any other function of x as well. In general we want to compute $\mathrm{E}\{f(x)\}$ for any function $f(x)$ that might be useful to us.

6.2. Standard Deviation and Variance

So far we answered the question “how far is the object”. The next question we will be asked is how good were the data. There are many ways to do this but we are seeking one that is simple, mathematically tractable, generally applicable and makes sense. Few definitions would fill the bill better than the estimate of the variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{i=N} (x_i - \mathrm{E}\{x\})^2}{N}$$

which assumes that we know the expected value $\mathrm{E}\{x\}$. If we do not, and usually we do not, then the estimate of the variance is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{i=N} (x_i - \bar{x})^2}{N}.$$

The above expression has the tendency to slightly underestimate the variance because $\bar{x}$ is the center of gravity of all the data points x_i and as a result the sum of square distances from all the data points is always no more than the sum of square distances from the true expected value. We will not worry about this now for many reasons. One is that most often the effect is small. Another is that for data with many dimensions any action that corrects this bias will be expensive. Yet another is that in estimation we care mainly with the relative magnitude of variances, so the effecto of this bias is smaller. Finally, and by far the most important and scientific reason is that nobody cares. We do not want to be geeks, do we?

The variance has all the nice properties we want. It is simple, and makes sense. If the data vary a lot from the expected value, then the variance is large. If they cluster tightly around the expected value the variance is small. It is definitely mathematically tractable (whereas if we used absolute value instead of square it would not be) and it appears naturally in many probabilistic models.

Quite often it is more convenient to use the square root form of the variance which is called *standard deviation*

$$\sigma = \sqrt{\sigma^2}.$$

It is not very hard to prove that any measurement x_i is with very high probability within $\pm 3\sigma$ from the expected value. If the probability model is Gaussian (a nice bell curve) this probability is 99.5%.

If you think that the estimate of the variance defined above makes sense, you are ready to see the definition of the variance itself:

$$\sigma^2 = \mathrm{E}\left\{(x - E\{x\})^2\right\}. \tag{6.3}$$

6.3. Covariance

Consider two random variables x and y that are not independent, with means μ_x and μ_y and variances ${\sigma_x}^2$ and ${\sigma_y}^2$. It is obvious that

$$C_{xy} = \mathrm{E}\left\{(x-\mu_x)(y-\mu_y)\right\}$$

cannot be simplified using the rule about products. It is a number that describes the dependance of the random variables x and y and as such it has a name: *covariance*. It is closely related to the more familiar *correlation*, which is nothing more than a normalized covariance. Since for our purposes only covariance is needed, we treat correlation as a poor relative.

Let's do a simple example of two non-independent random variables, x and y. Assume that $y = x + w$ where w is a random variable which is independent of x with mean μ_w and variance ${\sigma_w}^2$. Then the mean of y is

$$\mu_y = \mu_x + \mu_w$$

the variance ${\sigma_y}^2$ is

$$\begin{aligned}
{\sigma_y}^2 &= \mathrm{E}\left\{(y-\mu_y)^2\right\} = \\
&\mathrm{E}\left\{((x-\mu_x)+(w-\mu_w))^2\right\} = \\
&\mathrm{E}\left\{(x-\mu_x)^2\right\} + \mathrm{E}\left\{(w-\mu_w)^2\right\} + 2\mathrm{E}\{(w-\mu_w)(x-\mu_x)\} \\
&= {\sigma_x}^2 + {\sigma_w}^2
\end{aligned}$$

and the covariance is

$$\begin{aligned}
C_{xy} &= \mathrm{E}\left\{(x-\mu_x)(y-\mu_y)\right\} = \\
&\mathrm{E}\{(x-\mu_x)((x-\mu_x)+(w-\mu_w))\} = \\
&\mathrm{E}\left\{(x-\mu_x)^2\right\} + \mathrm{E}\{(w-\mu_w)(x-\mu_x)\} = {\sigma_x}^2
\end{aligned}$$

It is obvious from the above definition that we can write

$$C_{xx} = {\sigma_x}^2$$

and this will be a usefull alternative to the traditional σ notation.

6.4. Expected Value and Variance of the Average

We expect our measurements to have some randomness in them. What happens to this randomness when we take the average? We are sure that it decreases, but by how much? This sounds like the next question to answer.

As a warm up exercise let's compute the expected value of the average.

$$\mathrm{E}\{\bar{x}\} = \mathrm{E}\left\{\frac{\sum_{i=1}^{i=N} x_i}{N}\right\}$$

and we know that the expected value of the sum is the sum of the expected values, so

$$\mathrm{E}\{\bar{x}\} = \frac{\sum_{i=1}^{i=N} \mathrm{E}\{x_i\}}{N}$$

and since the expected value of every measurement is the same (unless we assume that the first 5 measurements are for practice and after that the sonar gets tired)

$$\mathrm{E}\{\bar{x}\} = \frac{\sum_{i=1}^{i=N} \mathrm{E}\{x\}}{N} = \mathrm{E}\{x\} = \mu_x$$

or in other words the expected value of the average is the same as the expected value of x, which is exactly what we suspected all along (and if we did not suspect such a thing, we should have suspected it!)

Then, how about the variance of the average.

$$\sigma_{avg}{}^2 = \mathrm{E}\left\{(\bar{x} - \mu_x)^2\right\} = \mathrm{E}\left\{\left(\frac{\sum_{i=1}^{i=N} x_i}{N} - \frac{\sum_{i=1}^{i=N} \mu_x}{N}\right)^2\right\} =$$

$$\mathrm{E}\left\{\left(\frac{\sum_{i=1}^{i=N} x_i - \mu_x}{N}\right)^2\right\} =$$

$$\mathrm{E}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{(x_i - \mu_x)(x_j - \mu_x)}{N^2}\right\}$$

we can apply the rule that says that the expected value of the sum is the sum of the expected values

$$\sigma_{avg}{}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\mathrm{E}\left\{(x_i - \mu_x)(x_j - \mu_x)\right\}}{N^2}$$

and we have to apply the rule about products now. This double summation has two kinds of terms: the ones where $i \neq j$ and the ones where $i = j$. For every term that has $i \neq j$, we use the assumption that the measurements are statistically independent so

$$\mathrm{E}\left\{(x_i - \mu_x)(x_j - \mu_x)\right\} = \mathrm{E}\{(x_i - \mu_x)\}\mathrm{E}\{(x_j - \mu_x)\} = (\mu_x - \mu_x)(\mu_x - \mu_x) = 0.$$

That's great the majority of terms have vanished! And we are left with the terms for which i and j are equal

$$\sigma_{avg}{}^2 = \sum_{i=1}^{N} \frac{\mathrm{E}\left\{(x_i - \mu_x)^2\right\}}{N^2}$$

and with the help of Eq. (6.3) we see that

$$\sigma_{avg}{}^2 = \frac{\sigma^2}{N}$$

which quantifies what we expected. If for example we take 100 measurements then the variance of the average is 100 times smaller and the standard deviation is 10 times smaller. If the sonar had a nominal accuracy of ±10 then the averaging will improve the accuracy to ±1. And that is true only if the measurements are independent. If they are not (e.g. the noise is mostly dues to high or low temperature, and the measurements were taken one after the other in a short period of time, the temperature did not have any random fluctuation and the measurements are not independent.

7. Weighted Average

Now assume that we mount two identical sonar sensors on the robot and both are looking into the same direction. The only difference between them is that their internal software uses averaging to improve accuracy in both but the one averages 4 measurements and the other averages 9. It is obvious that we trust the second sonar more than the first and that the averaging should be weighted

$$\hat{x} = \frac{4x_1 + 9x_2}{4 + 9}. \tag{7.1}$$

While this happens to be the best estimate given the lack of other information, it is nonetheless the result of gut feeling. And like many other outcomes of the gut it might not be ideal.

The problem with Eq. (7.1) is that we do not always have raw data to average but only a few equations that involve random measurements and each one alone is

indeterminate. So we have to devise something that can be applied in general.

Since we are not the only ones that need this kind of general estimator, mathematicians and statisticians worked hard and invented a whole range of such things called estimators. These are general techniques by which we design functions that compute the quantity we want. These estimators have names like “Bayesian”, “Maximum Likelihood” or “χ^2”. Although all three have their respective ecological niches, they are quite similar in many ways. We could use any of the three for our case but we will opt for the simplest one, the χ^2.

We want to estimate a parameter, in this case $\hat{x}$ which is the same as the expected value of x. We form the expression

$$\chi^2 = \frac{(x_1 - \hat{x})^2}{{\sigma_1}^2} + \frac{(x_2 - \hat{x})^2}{{\sigma_2}^2} \tag{7.2}$$

and we try to minimize it. In more general settings we would form the expression

$$\sum_i \frac{(f(p) - x_i)^2}{\sigma_i}$$

where p is the parameter we want to estimate and $f(p)$ is a function that returns the expected value of x given the parameter p.

The more we look at Eq. (7.2) the more sense it makes. If we find an $\hat{x}$ that is as close as possible to the data points x_i with preference to points with small σ_i then Eq. (7.2) achieves minimum.

We now try the standard way to minimize it. We take the derivative with respect to the unknown and set it to zero.

$$\frac{\partial}{\partial \hat{x}} \left(\frac{(x_1 - \hat{x})^2}{{\sigma_1}^2} + \frac{(x_2 - \hat{x})^2}{{\sigma_2}^2} \right) = 2\,\frac{(x_1 - \hat{x})}{{\sigma_1}^2} + 2\,\frac{(x_2 - \hat{x})}{{\sigma_2}^2} = 0$$

and some simple algebra gives us

$$\hat{x} = \frac{\dfrac{x_1}{{\sigma_1}^2} + \dfrac{x_2}{{\sigma_2}^2}}{\dfrac{1}{{\sigma_1}^2} + \dfrac{1}{{\sigma_2}^2}} \tag{7.3}$$

We notice that we can get the Eq. (7.1) from (7.3) if we replace ${\sigma_1}^2$ with $\sigma^2/4$ and ${\sigma_2}^2$ with $\sigma^2/9$. It is always nice to see a formula derived from our gut feeling coincide with one derived by some kind of fancy math.

We now turn our attention to the quality of the estimate $\hat{x}$ and compute its variance

$$\sigma_{\hat{x}}{}^2 = \mathrm{E}\left\{(\hat{x} - \mu_x)^2\right\} =$$

$$\mathrm{E}\left\{\left(\frac{\dfrac{x_1}{\sigma_1{}^2} + \dfrac{x_2}{\sigma_2{}^2}}{\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}} - \mu_x \frac{\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}}{\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}}\right)^2\right\} =$$

$$\mathrm{E}\left\{\left(\frac{\dfrac{x_1 - \mu_x}{\sigma_1{}^2} + \dfrac{x_2 - \mu_x}{\sigma_2{}^2}}{\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}}\right)^2\right\}$$

and if we expand the square we can apply the rule of the expected value of a sum

$$\frac{\mathrm{E}\left\{\left(\dfrac{x_1 - \mu_x}{\sigma_1{}^2}\right)^2\right\}}{\left(\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}\right)^2} + 2\frac{\mathrm{E}\left\{\left(\dfrac{x_1 - \mu_x}{\sigma_1{}^2}\right)\left(\dfrac{x_2 - \mu_x}{\sigma_2{}^2}\right)\right\}}{\left(\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}\right)^2} + \frac{\mathrm{E}\left\{\left(\dfrac{x_2 - \mu_x}{\sigma_2{}^2}\right)^2\right\}}{\left(\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}\right)^2} =$$

$$\frac{\dfrac{\sigma_1{}^2}{\sigma_1{}^4}}{\left(\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}\right)^2} + 0 + \frac{\dfrac{\sigma_2{}^2}{\sigma_2{}^4}}{\left(\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}\right)^2}$$

which finally leads to

$$\sigma_{\hat{x}}{}^2 = \frac{1}{\dfrac{1}{\sigma_1{}^2} + \dfrac{1}{\sigma_2{}^2}}. \tag{7.4}$$

The above expression is nice and simple and makes a lot of intuitive sense. It tells us that whenever we average two measurements we always reduce the variance if we use the proper weighting (we could do the calculations using improper weighting and see that we might increase the variance) because $\sigma_{\hat{x}}{}^2$ is always less than either $\sigma_1{}^2$ or $\sigma_2{}^2$. And if we combine two measurements, one with standard deviation of say 1 and the other of standard deviation of 10, the relative weights in the averaging will be 1 and 1/100 respectively and as a result the variance will change by only 1 percent (and the standard deviation by half percent), which is negligible as one would expect.

The above hold for one dimensional data only. If our samples x_i are vectors, then things are slightly different, but only slightly: the variances σ^2 are not scalars anymore, they are matrices. And what comes to mind when the number of dimensions goes up is the “dimensionality curse”, or the rapidly increasing difficulty of the problem as the dimensionality increases. This is not the case here. The increased dimensionality decreases performance but allows some spectacular solutions to some estimation

problems.

8. Weighted Average in Multiple Dimensions

With the notable exception of drainpipe inspection robots all robots live in multidimensional worlds. Their position or state is a vector and not a number. An autonomous robotic platform would need at least three numbers to describe its position: x, y and orientation θ. A flying robot would need 6 such numbers: three for the position and three for the orientation. But having a whole set of numbers to describe the state of the robot means that we have a whole set of numbers to estimate. It also means that Goddess Chance has a whole set of numbers to infect with randomness. And this means multidimensional statistics, which in turn means vectors and matrices.

The first thing in order is to calm the population. Matrices are not that much harder and in most cases we can get away with 2×2 matrices which are simple to visualize. Our measurements will be again x_i, the average will be $\bar{x}$, the estimate will be $\hat{x}$. The only difference is that now they are vectors. The variance will be a little different though because it is a matrix and we will use the symbol $\mathbf{C}$.

A measurement x_i has two components now (say the north facing sonar and the west facing sonar)

$$x_i = \begin{bmatrix} p_i \\ q_i \end{bmatrix}$$

and the average of such a vector is the vector averages

$$\bar{x} = \begin{bmatrix} \bar{p} \\ \bar{q} \end{bmatrix}$$

and the expected value is similarly a vector

$$\mu_x = \mathrm{E}\{x\} = \begin{bmatrix} \mathrm{E}\{p\} \\ \mathrm{E}\{q\} \end{bmatrix} = \begin{bmatrix} \mu_p \\ \mu_q \end{bmatrix}$$

and so is the estimate

$$\hat{x} = \begin{bmatrix} \hat{p} \\ \hat{q} \end{bmatrix}$$

so the only partycrusher is the variance which is usually called “The Variance-Covariance Matrix”. Before we start thinking nasty things about this matrix, lets see the definition. It really makes sense.

$$\mathbf{C} = \mathrm{E}\left\{(x-\mu_x)(x-\mu_x)^T\right\} = \begin{bmatrix} E\left\{(p-\mu_p)^2\right\} & E\left\{(p-\mu_p)(q-\mu_q)\right\} \\ E\left\{(p-\mu_p)(q-\mu_q)\right\} & E\left\{(q-\mu_q)^2\right\} \end{bmatrix} = \begin{bmatrix} C_{pp} & C_{pq} \\ C_{pq} & C_{qq} \end{bmatrix}.$$

The two diagonal elements C_{pp} and C_{qq} are old friends. They are the one dimensional variances of the elements of the vector x which we expected. The thing that we might not expect are the two (actually one since they are equal) off-diagonal elements C_{pq}. These are the *Covariances*! Which at least explains the name of the matrix.

When we have two or more random variables, then it is important to have not anly their variances but their covariances as well. Consider the case when x is the vector of measurements from two sonars and that the error is mostly due to air drafts that affect the speed of sound. In general drafts make the distances appear larger as measured by the sonar. If the draft during a particular measurement is weaker than average then both p and q will be less than their expected value and both $(p-\mu_p)$ and $(q-\mu_q)$ will be negative and so their product will be positive. If the draft is greater than usual then both $(p-\mu_p)$ and $(q-\mu_q)$ will be positive and again their product will be positive. Since this thing is positive, no matter what the draft is, the expected value will be positive. And that's why we say they are positively correlated. If, on the other hand, in a different scenario, whenever $(p-\mu_p)$ was positive $(q-\mu_q)$ was negative and the opposite, the expected value of $(p-\mu_p)(q-\mu_q)$ would be negative and we would say that they are negatively correlated.

So what is the use of this apart from being a license for scientific lies. There is quite a bit of use. Consider our sonars taking measurements in that drafty lab. If we get an independent measurement of one of the distances and we find it 10% larger then we can not only correct this measurement but also the measurement of the other sonar. This is particularly useful in complicated multidimensional problems.

While we can do many things in two dimensions that look like magic, the derivations are too much trouble. So we skip a great deal of them and just mention the results.

In multiple dimensions the weighted average is

$$\bar{x} = \left(\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1}\right)^{-1}\left(\mathbf{C}_1^{-1}x_1 + \mathbf{C}_2^{-1}x_2\right)$$

a formula not very different from Eq. (7.3). And the variance is

$$\mathbf{C}_{\bar{x}} = \left(\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1}\right)^{-1}.$$

again a formula not very different than Eq. (7.4). And last but not least we need to guess the formula for χ^2. Following again our gut feeling, which is of proven quality

$$\chi^2 = (x_1 - \hat{x})\mathbf{C}_1^{-1}(x_1 - \hat{x})^T + (x_2 - \hat{x})\mathbf{C}_2^{-1}(x_2 - \hat{x})^T$$

Let's look at a simple example. A robot knows that its position in the x and y direction is 5 and 7 with a variance of 1 and 10 respectively. It can also fire its sonars and get its distance from the walls and find that its x and y position is 3 and 5 with variance of 10 and 1 respectively. We can compute the estimate of its position using the above formulas. The variance-covariance matrix for the position of the robot is

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$$

and the variance-covariance matrix for the sonar measurement is

$$\mathbf{C}_2 = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}.$$

These are nice diagonal matrices and we can invert them easily. We compute the variance covariance matrix first because the exact same expression appears in the average.

$$\mathbf{C}_{avg} = \left(\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1}\right)^{-1} = \left(\begin{bmatrix} 1 & 0 \\ 0 & .1 \end{bmatrix} + \begin{bmatrix} .1 & 0 \\ 0 & 1 \end{bmatrix}\right)^{-1} =$$

$$\begin{bmatrix} 1.1 & 0 \\ 0 & 1.1 \end{bmatrix}^{-1} = \begin{bmatrix} 0.91 & 0 \\ 0 & 0.91 \end{bmatrix}$$

The best estimate is then

$$\hat{x} = \begin{bmatrix} 0.91 & 0 \\ 0 & 0.91 \end{bmatrix}\left(\begin{bmatrix} 1 & 0 \\ 0 & .1 \end{bmatrix}\begin{bmatrix} 5 \\ 7 \end{bmatrix} + \begin{bmatrix} .1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 5 \end{bmatrix}\right) =$$

$$\begin{bmatrix} 0.91 & 0 \\ 0 & 0.91 \end{bmatrix}\begin{bmatrix} 5+.3 \\ 5+.7 \end{bmatrix} = \begin{bmatrix} 4.81 \\ 5.18 \end{bmatrix}$$