Basic Physical Statistics & Calculations
Draft 2009-04-03

My purpose is to give some *insight* into batch data calculations involving averages (means) and the variability about those averages (variances). Along the way I introduce the vocabulary of basic statistics and some of the math symbols used in calculations. At the end I will re-interpret these calculations in terms of the statistical idea of the *Expected Value of Random Variables*. Since Random Variables use expectations, averages, and variability estimates extensively, having insight and skills in calculating these will be a natural lead in.

For additional insight, I will relate perspectives from engineering to show that the statistics calculations for averages and variances are exactly the same as parallel engineering calculations. In fact, the physical approach is probably a better introduction to statistics than the textbook approach. So, take a look at the text book format first to get an idea of what needs to be done, and then go to the physical interpretation. (See "Basic Physical Insight" on page 7).

Various functions of means and averages match up between engineering and statistics, although the interpretations differ since the engineering perspective introduces masses, forces, and torques (that is, a 'turning force'). If you like a 'hands on' perspective though, this is the way to learn the material. (In that section, I note Newton's laws for linear and rotational motion as background for the interested reader).

> An aside: I firmly believe that the perceptions and experiences of physical 'pushes' and 'pulls' form the basis for all the behavior I can think of, with thought itself mediated by this perception. So, I do better when tasks are physically interpreted.

Before I start in about means and variances, let me say a word about the underlying context of statistics, *sample populations* and *total populations*.

Setting up (part of) the Context for an Analysis: a Population

A *population* is a set of measurements, counts, or observations on a collection of *entities* of interest. This says that the *measurements of some feature/property of the entities* is the population, not the entities themselves.This distinction lets me consider different kinds of measurement on the same set of entities, which is very common. If I talked about the set of weights of a certain collection of people (the entities), the population is the set of weights, not the people. For those same entities, I could also *measure* their ages, another population, or *count* the number of people with brown eyes yet another population. Another very common example of these ideas is a survey presented to a set of people (the entities). The researcher may extract multiple populations from these surveys such as demographic data as well as preference data or even open-ended question data. (See the tutorial on this site *Exploratory Discourse Analysis* for a perspective on non-numerical, textual analyses).

For another example, suppose I have a collection of invoices I am analyzing - now consider all of the errors found on these invoices as the population. The entities in this case are the invoices, and measurements taken on those invoices, the errors, comprise the population. Or, consider the meteorological entities 'clouds' with their color or density as the measures of interest.

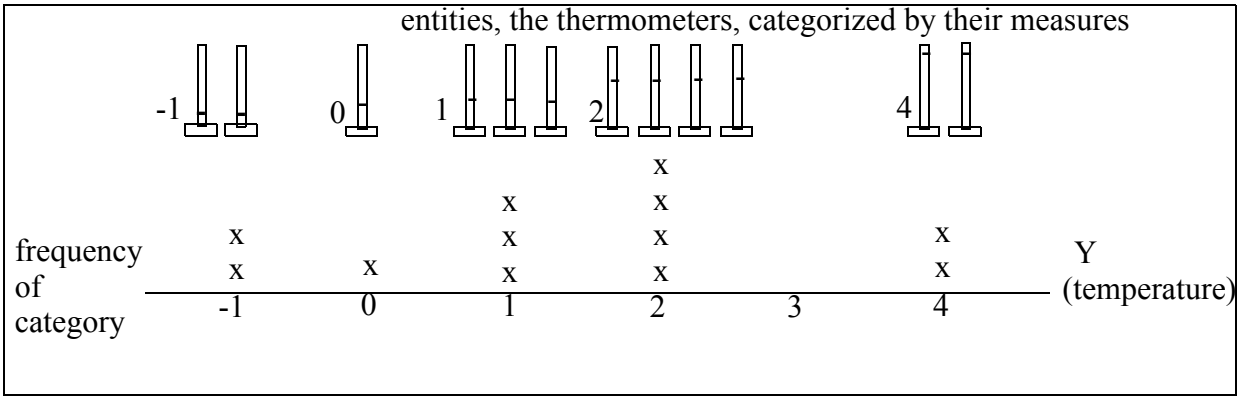
The discussion below makes use of the concept of temperature along with entities (thermometers) with measurement in terms of degrees. The temperature degree measurements, a batch of numbers, is the population. This isn't the whole story of course, since there is a lot more having to do with

the *context* in which the entities are identified, such as time and place, as well as the procedure (protocol) of taking observations or measurements/counts on them.

A Multi - View of a Batch of Numbers

An underlying theme of this tutorial suggests that the 'batch' and the numbers that comprise it, can be viewed in three useful ways. (A detailed example of these ideas follows.). Often you will get a batch of numbers, out of context, such as y = {-1, -1, 0, 1, 1, 1, 2, 2, 2, 2, 4, 4}. What do these numbers mean, and why should I care? So, finding out their *source/context* is the first task of the analyst.

- First, the batch of numbers was or will be, gathered from a *context (a set of entities and their setting),* while the batch of numbers themselves are the population. Notice that the numbers have usually lost their association with the entities they came from. Still, we notice that some numbers may be repeated, suggesting that these may be meaningfully grouped as representing a category of underlying entities, as shown in the next bullet. (Actually, the numbers are from temperature measurements (Farenheit degrees) at a northern city that is discussed below).
- Second, numbers in the batch identify *entities*, but are often not uniquely traceable as mentioned above. For example, the unique temperatures in the batch above identify *categories of entities*, such as the 4 entities having a specific temperature of '2' degrees. In math, this would be called an 'onto' mapping. The diagram below shows this connection between an entity and its' measure. Categories of entities having the same measure 'map' into the temperatures that show up in the batch of numbers.



- Third, each of the categories of entities has a count or measure of their frequency of occurrence.These frequencies represent the *frequency of occurrence* of each entity category.(For continuous distributions, the relative frequency of occurrence shades over into a *density of occurrence* measure).

Taken together, these three views give me a deeper appreciation of the batch itself and so tends to increase my emotional involvement and interest that, for me, is a precondition to really taking the trouble to find out something! This is the attitude of the *Taoist* scientist as described by Abraham Maslow [Motivation, 1970]. I must be involved with the data in order to first appreciate it, and then to understand it.

You will see in these notes how the calculations shown and the perspectives suggested, help to understand the somewhat abstract ideas of *data distributions*, *statistical means and variances, random variables, expectations, probability mass functions*, and their continuous extensions, *probability density functions*.

Textbook-Style Calculations for the Mean and Variance

This section is in two parts: I calculate the *average* (*mean)* both for a sample population and a total population, and then calculate the *variance* and *standard deviation* both for a sample population and total population.

Use the Same Calculation for both Sample Mean and Population Mean

Here is a (sorted) batch of observations of temperatures, introduced earlier, recorded on 12 different January days in Albany N.Y. (where my grandchildren live). Brrrr!

y = {-1, -1, 0, 1, 1, 1, 2, 2, 2, 2, 4, 4}.

I can interpret this either as a *sample* population of daily temperatures or as the *total population* of daily temperatures. The interpretation depends on whether I want to generalize to other days from this *sample* or if this is the total set of temperatures (the total *population*) that I will consider.

Either way, sample or total population, the average of this batch calculates the same.

> What I want to most strongly point out (again!) is that this batch of numbers is not just a batch of numbers! It has a context, daily temperatures for a particular city. That context is further broken down into *entity categories* represented by the various temperatures, -1, 0, 1, 2, and 4, together with the *frequency* of *occurrence* of those entities. This is a subtle point that is overlooked if I just start manipulating the pure numbers. The take home idea is that there are *two* kinds of numbers implicit in any batch: one kind *identifies* a category of entities within a context, such as a -1 temperature, while a second kind of number refers to the *measure* or *count* of occurrences of that category, namely '2' occurrences of the temperature '-1'. This count of occurrences is often called *frequency of occurrence,* or just *frequency.* In the continuous case, relative frequency shades into a measurement called the *density* of occurrences

I will calculate the mean or average of this batch of temperature numbers under the assumption that it is a sample and then re-calculate its mean under the assumption that it is the total population. The calculations are identical, but the interpretation of the results differ.

Vocabulary Note: For a sample, the resulting calculated average is called a sample *statistic* while for a population, the calculated average is called a population *parameter*.

Let me lay out this sorted batch in what looks like a histogram (also called a *dot plot*), showing the count of each temperature. The *frequency of occurrence at that temperature* is shown by the x's. That is, there were two days of -1 degrees, 1 day of zero degrees, 3 days of 1 degree, 4 days of 2 degrees, zero days of 3 degrees and 2 days of 4 degrees. Let me index the distinct individual values that have observations so that I can compactly refer to them in equations (don't worry, I'll also write them out in full!). Note: This illustrates the idea that these values identify *categories of entities/locations*. Secondarily, there is the property of their frequency of occurrence.

(I left out indexing the 3 degree day on the grounds that it had zero observations).

y[1] = -1, y[2] = 0, y[3] = 1, y[4] = 2, y[5] = 4

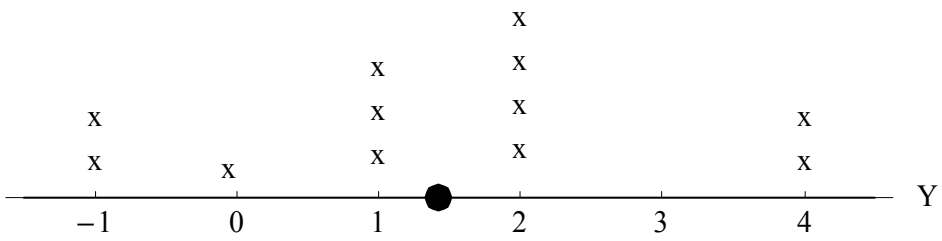


FIGURE 1. Distribution of Temperatures, The Large Dot' is at the Average (Mean) Temperature

Sample Mean Calculation (the sample mean is also called the sample average)

I will use *ybar* as the location of the sample average (it turns out to be 1.42 degrees in this example). There are a total of 12 occurrences of temperatures, with frequencies of occurrence as shown in the figure above. That is, at y[1], which refers to temperature -1, the frequency of occurrence is '2'. For y[2], which refers to zero degrees, there is '1' occurrence, while for y[3]= 1 degree, there are '3' occurrences and so on. The total frequency refers to the total number of occurrences of observations, namely '12'.

Here is the compact formula that shows how to calculate the average. The capital Greek symbol Σ is called sigma, and is suggestive of the operation of summation. So, in words, this says, take each y[i], multiply it by its' frequency of occurrence, add these all up and divide by the total number of occurrences. That's it!

$$\text{ybar} = \left(\sum_{i=1}^{5} y[i] * \text{frequency}[\, y[i]\,] \right) \Big/ \text{totalFrequency}$$

ybar = (-1 * (2) + 0 * (1) + 1 *(3) + 2 * (4) + 4 * (2)) / 12

ybar = 17/12 = 1.42 degrees

Note that *ybar* is what is called a *sample statistic* and is an *estimate (called a point estimate)* of the true underlying population mean, which is the unknown μ.

Another way to look at it: If I divide each individual frequency of occurrence by the *totalFrequency* of '12', I get the relative frequency at each temperature location, which is a common way of setting up the average calculation.

$$\text{ybar} = \sum_{i=1}^{5} y[i] * \text{relativeFrequency}[\, y[i]\,]$$

ybar = -1 * (2/12) + 0 * (1/12) + 1 *(3/12) + 2 * (4/12) + 4 * (/12)

ybar = 1.42 degrees

Population Mean (the population mean is also called the population average)

Here is the population mean calculation, which is identical with the sample calculation. The interpretation differs though, since this result, μ, is called a *parameter* of the population. A parameter

is something about this batch of numbers that doesn't change since I know everything about this population. Note the symbol change from *ybar* to μ to indicate the *parameter* status of the result.

$$\mu = \left(\sum_{i=1}^{5} \mathbf{y[i] * frequency[\, y[i]\,]} \right) \Big/ \mathbf{totalFrequency}$$

Another way to look at it: If I divide each frequency by the *totalFrequency* I get the relative frequency at each location, which is a common way of setting up the average calculation.

$$\mu = \sum_{i=1}^{5} \mathbf{y[i] * relativeFrequency[\, y[i]\,]}$$

μ = -1 * (2/12) + 0 * (1/12) + 1 *(3/12) + 2 * (4/12) + 3 * (0/12) + 4 * (2/12)

μ = 1.42 degrees

Finally, if the relative frequencies are stable, and unchanging, I can interpret them as the probabilities of occurrences of each temperature. This is often called the *probability mass function* (which shows how physical ideas turn up in even these abstract calculations. I will expand on the idea of number frequencies being represented by masses in the section of physical interpretations).

$$\mu = \sum_{i=1}^{5} \mathbf{y[i] * probability[\, y[i]\,]}$$

Calculate the Sample and Population Variance, Just a Little Different

Sample Variance

There is some difference between the sample variance calculations and population variance calculations. That difference is in using the *totalFrequency -1* in the denominator for the sample variance calculations versus using just the *totalFrequency* for the population variance calculation. It is also necessary to replace the sample mean, *ybar*, by the population mean, μ, in the formulas.

When the sample size gets to be more than say, 30, the frequency adjustment is negligible but, for smaller sample sizes it's best to use *totalFrequency -1*.

Ok, here is the *sample variance* calculation for the temperature batch data set of Figure 1:

$$\mathbf{s^2} = \left(\sum_{i=1}^{5} \mathbf{(y[i] - ybar)^2 * frequency[\, y[i]\,]} \right) \Big/ \mathbf{(totalFrequency - 1)}$$

$= (-1 - 1.42)^2 * 2 + (0- 1.42)^2 * 1 + (1- 1.42)^2 * 3 + (2 - 1.42)^2 * 4 + (4-1.42)^2 * 2)/11$

$= 2.63$

Sample Variance = 2.63 degrees^2

This is the average squared distance from *ybar*.

The sample standard deviation, s, is the square root of s^2

s = Sqrt[2.63] = 1.62 degrees

This is the average distance from *ybar*.

Population Variance

If I assume I am dealing with the *total population* of temperatures, my calculations above showed that the population mean was 1.42, the same as the sample mean. That is the number I will use for the μ below.

$$\sigma^2 = \left(\sum_{i=1}^{5} (y[i] - \mu)^2 \times \text{frequency}[y[i]] \right) / \text{totalFrequency}$$

$= (-1 - 1.42)^2 * 2 + (0- 1.42)^2 * 1 + (1- 1.42)^2 * 3 + (2 - 1.42)^2 * 4 + (4-1.42)^2 * 2)/12$

$= 2.409$ degrees^2

This differs from the sample variance, s^2, due to the denominator difference plus the replacement of ybar by μ. Next, I have scaled each frequency by the total, which gives me a relative frequency.(Note that the relative frequencies add up to unity).

$$\sigma^2 = \sum_{i=1}^{5} (y[i] - \mu)^2 \times \text{relativeFrequency}[y[i]]$$

$= (-1 - 1.42)^2 *2 /12 + (0- 1.42)^2 * 1/12 + (1- 1.42)^2 * 3/12 + (2 - 1.42)^2 * 4/12 + (4 -1.42)^2 * 2/12$

$= 2.409$ degrees^2

If these relative frequencies settle down, they are called the *probability mass functions* for reasons you will understand when you read "Basic Physical Insight" on page 7.

The population standard deviation, σ, is the square root of σ^2

σ = Sqrt[2.409] = 1.55 degrees

This is the average distance from the mean temperature.

(An Optional Aside) The Mean Interpreted as a 'Center of Numbers'

This is a slightly different 'take' on the interpretation of the temperature batch of numbers. (You will see the similarity of the *center of numbers* to the *center of gravity* discussion in the next section) We already know that the number of x's is the frequency of numbers *at a temperature location*. That is, the number of x's shows the *number of that number*, as in Figure 1. So, the calculation for *ybar* (or μ), could also be thought of as the *location* such that if the total numbers were concentrated there, they would have the same 'turning effect' (*torque*) as the spread out distribution.

In other words, if I took the total count of numbers, which is '12', and placed that count at one lo-

cation, where would that location be so that it could represent the same effect as the spread out distribution? That location is *ybar*. What is really going on here is that I am thinking of the count of numbers at a location as the 'number-mass' (number of units) at that location and using physical intuition as to where that numerical mass should be placed to give an equivalent 'turning' effect. This is the really center of gravity idea, but a little more abstract.

ybar * TotalCountOfNumbers =

-1 *(count of number of -1's)

+ 0 * (count of number of 0's)

+ 1 * (count of number of 1's)

+ 2 * (count of number of 2's)

+ 3 * (count of number of 3's)

+ 4 * (count of number of 4's)

ybar * TotalCountOfNumbers= -1 * (2) + 0 * (1) + 1 *(3) + 2 * (4) = 3 * (0) + 4 * (2)

**So, this left hand side balances the right hand side, if I only knew where ybar was!,

So I solve for ybar and get: **

ybar * 12 = -1 * (2) + 0 * (1) + 1 *(3) + 2 * (4) = 3 * (0) + 4 * (2)

ybar = -1 * (2/12) + 0 * (1/12) + 1 *(3/12) + 2 * (4/12) = 3 * (0/12) + 4 * (2/12)

ybar = 17/12 = 1.42 degrees

Naturally, this is the same as the previous results. You will see another way to calculate this mean in the section below.

Basic Physical Insight

The Mean as Center of Gravity

For a physical interpretation of the above temperature set of Figure 1, replace the horizontal axis, where the temperatures are laid out, with an extraordinarily strong but light plastic rod, marked off using the temperature numbers as locations along the rod.(See"The Batch Distribution Balance (fulcrum) point: Center of Gravity = Mean = Average" on page 8).

Further, place a number of *unit masses* at those locations corresponding to the frequencies at that location. That is, at the location of '-1' on the rod, place 2 mass units. At the location marked '0', place one unit, at location '1', place 3 mass units, at '2' place 4 mass units, while at '4', place 2 units. Where would this collection of 12 masses balance, if you could put a *fulcrum* (point of support) somewhere between -1 and 4, under the rod? In the diagram below I have shown arrows pointing down to indicate the effects of gravity acting on the masses at each location. The arrows are in proportion to the weight at that point. Remember that on a see-saw (teeter totter?) a lighter child, of weight, say, 60 pounds who is at 4 feet from the fulcrum (position of support) can balance a mom weighting 120 pounds at a distance of 2 feet. So the distances times weights *balance and therefore so do the participants/masses.* That is, the sum of the *distance x weight*(s) on one side of the fulcrum must be matched with a sum of *distance x weight(s)* on the other side, if the see-saw doesn't rotate.

The Mean of a Discrete Distribution

Now, let me bit more general using the figure below. Where must the *fulcrum* be placed in the diagram below in order for it to balance? That balance point, the fulcrum point I have been talking about, is the center of gravity as well as the statistical mean, μ! Let me denote that unknown location by the symbol 'μ' and write the balance equation it must satisfy, then I'll solve for where μ must lie.

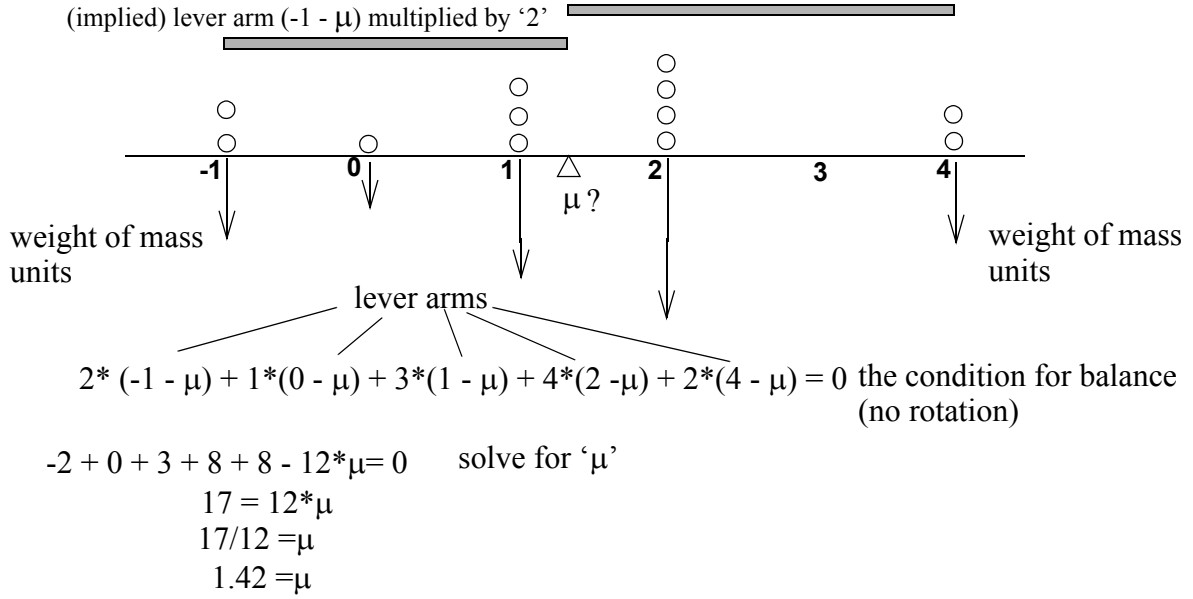


FIGURE 2. The Batch Distribution Balance (fulcrum) point: Center of Gravity = Mean = Average

You can see that this balance point, μ = 1.42, is the same as our old friend *ybar* from previous calculations. In general though, the population mean μ, will not be the same as the sample mean *ybar*.

Next, the same problem is expressed a little more generally. Notice especially that mass[y[i]] is just the number of mass units at a particular location y[i]. So, for example, at y[4], which is "2", there are 4 mass units. The mass count is therefore 4 units, at location y=2. The total mass is 12 units.

The two equations below are just a formal way of writing the equation of Figure 2 on page 8 above. I introduce it since it will generalize to the continuous case considered next.

$$\sum_{i=1}^{5} (y[i] - \mu) * mass[y[i]] = 0$$

the balance condition
(no rotation around y = μ)

$$totalMass = \sum_{i=1}^{5} mass[y[i]]$$

FIGURE 3. Formal Equation for Population Mean (μ)

The Mean of a Continuous Distribution - You Need a Density Function

If the distribution of the values is not discrete but instead there are too many to count individually, you will need to use a density function to represent that spread. Recall the (discrete) mean calculation above when there were only 12 masses and 'μ' was the location of the balance point such that there was no rotation around an axis through μ (think of that rotation axis as perpendicular to the page, coming out towards you).

Same thing for the continuous case. I will again find a location μ, such that there is no rotation of the distribution of masses that lie on either side of it. In the continuous case, the masses are 'smeared out, rather than being concentrated in 'lumps'.(See "A Special Case for Finding the Mean of a Continuous Distribution" on page 10).

Now consider that those masses were spread along the 'y' axis as in the figure. Each of those sets of masses is now broken up into a (rectangular!) pile of sand. For example, suppose that first set of 2 unit masses were broken up into lots and lots of sand particles and piled up on the y axis in the rectangular shape shown, same for the other masses.

Mass Density Function f[y] (exactly equivalent to the Probability Density Function)

Since there are now too many grains of sand to count individually at each 'y' location, I estimate the number of particles (mass) by using a *rectangle* of *height = density function at the point y*, times a small length along the y-axis, denoted as 'dy'. The density function is defined to be the mass in an interval *scaled by the total mass*. For example, the density function around the y location of '1', denoted as f[1], is 3/12 pounds per unit length. Similarly, the density function, f[4] around the location y=4, is 2/12 pounds per unit length. That is, the density function, written as f[y], depends on where I am along the axis while the 'dy' denotes a small interval centered on each such 'y' location.

> This product of the density function (a height) at a location 'y', times an interval length around this y value gives me the (relative) mass enclosed by that rectangle. Note that the dimensions of the density function is *relative mass per unit length*.
>
> The relative mass idea is exactly equivalent to an element of Probability given by a Probability Density Function times an interval length. That product encloses a relative amount of probability. That is, if f[y] is now interpreted as a Probability Density Function,(*reflecting a density of numbers at a y location*), and multiplied by 'dy', you get f[y]*dy. This is the equivalent of a probability relative element, rather than a physical relative mass element, both at the end of a lever arm 'y'.

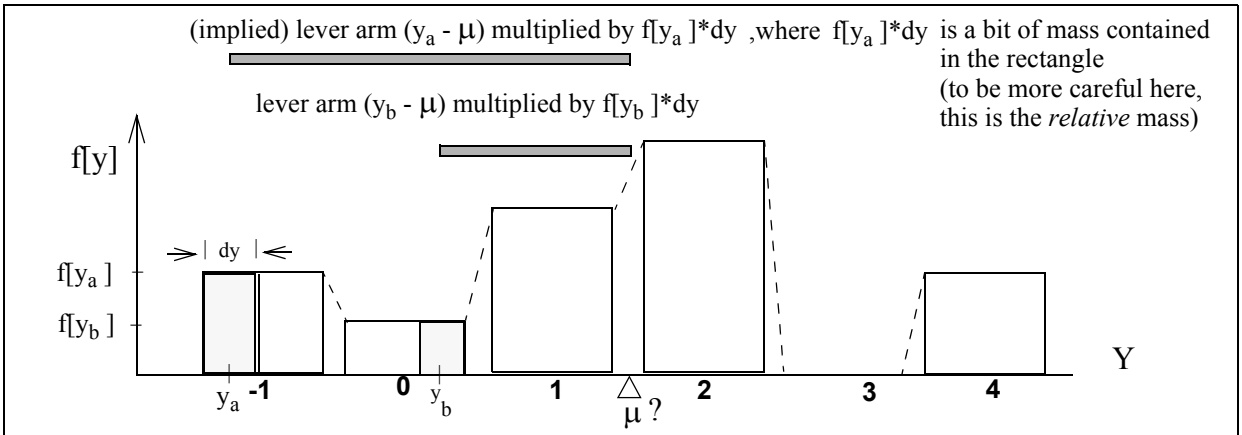
So, in the diagram below, when y is a general point very near y= -1, say, y_a, the density would be 2/12 pounds/unit length and so the mass of the rectangle centered at y_a and of (small) base length 'dy', would be 2/12 * dy. In other words, f[y] * dy represents the relative amount of mass contained in that rectangle.

A Plan to Find the Center of Gravity in the Continuous Case

Now, to find μ, the plan is an extension of the discrete case. I am still looking to calculate the point μ such that the distribution will balance. This means that the sum of the lever arms times a mass element on one side of the balance point μ, matches the sum of the lever arms times mass elements on the other side. (Note that I am interpreting each mass element at the end of a lever arm as acting under the influence of gravity so as to produce a turning force, called a *torque*).

1. Divide up the y-axis into finer and finer subdivisions with a 'y; value at the center of each subdivision (these subdivisions are denoted by 'dy'). This translates to having more and more individual y values in the center of each subdivision which mathematically equates to having more and more indexes, 'i' along with shorter and shorter intervals.

2. The (relative) mass around a very small interval centered on a y value is given by the contents of the rectangle of height f[y] and base length, 'dy'.
3. Multiply the 'lever arm', which is of length '(y - μ)', by the relative mass at 'y', given by f[y]* dy. This gives the 'torque' due to that lever arm times that amount of relative mass acted upon by gravity, at the end of the lever arm.
4. Add up all of these torques, picking μ such that they sum to zero. Note that μ must be solved for.
5. Solve for that location μ, that will satisfy the sum of torques being zero
6. In the diagram below, I have labeled a couple of arbitrary points along the Y axis, y_a and y_b just to have something to reference.



$$\int (y - \mu)\, f[y]\, dy = \int y\, f[y]\, dy + \int -\mu\, f[y]\, dy = 0$$ (balance condition) (sum of torques = 0)

$$\int y\, f[y]\, dy = \int \mu\, f[y]\, dy \quad \text{but} \quad \int f[y]\, dy = 1$$ since f[y] is a relative density function. This integral simply adds up all the relative mass, which always totals to '1'.

$$\mu = \int y\, f[y]\, dy$$ this is the location solution for μ

FIGURE 4. A Special Case for Finding the Mean of a Continuous Distribution

The Mean as Center of Mass

The above discussion shows how to calculate the mean as the *center of gravity*. The *mean* is the point where a distribution of weights would balance, under gravity. This idea of balance under gravity is a general one and can be used confidently in statistical work, but, there is yet another way to look at these masses as well, one that doesn't involve gravity.

The *center of mass* concept has meaning even in deep space, where there is no gravity. For that description I need to reference Newton's laws explicitly. For this derivation, I don't need a rigid ruler, or any other prop at all, just the mass distribution is enough to work with. This application of Newton's laws finds *the point associated with the mass distribution that acts as if all the mass*

were concentrated there, as far as Newtons laws go. Specifically, if I can think of the batch of numbers as a "particle" then this description can be helpful.

I will denote that center of mass point by *ybar.* (This will again turn out to be at location 1.42 relative to an axis referenced to zero just as calculations in previous sections showed). That's the location where the mass distribution may be assumed to be concentrated. Physically, *ybar* is the point such that if the total mass were concentrated there (all 12 units in our case), it would have the same effect as the spread out distribution, as far as Newton's physical equations of motion are concerned. So, where is *ybar*? The equation below sets up to find that concentration point.

ybar * TotalMass =

-1 * (number of unit masses at location -1)

\+ 0 * (number of unit masses at location 0)

\+ 1 * (number of unit masses at location 1)

\+ 2 * (number of unit masses at location 2)

\+ 3 * (number of unit masses at location 3) * note I didn't need to include this since no mass here*

\+ 4 * (number of unit masses at location 4)

ybar * TotalMass= -1 * (2) + 0 * (1) + 1 *(3) + 2 * (4) + 3 * (0) + 4 * (2)

ybar * 12 = -1 * (2) + 0 * (1) + 1 *(3) + 2 * (4) = 3 * (0) + 4 * (2)

ybar = (-1 * (2) + 0 * (1) + 1 *(3) + 2 * (4) + 3 * (0) + 4 * (2)) / 12

= 17/12 = 1.42

Further insight into *ybar*, which illustrates the idea of *mass density:*

Divide each mass count at each location by the total mass and so get the *mass density* at each location. This is the fraction of the total mass at that location.

ybar = -1 * (2/12) + 0 * (1/12) + 1 *(3/12) + 2 * (4/12) + 3 * (0/12) + 4 * (2/12)

ybar = 17/12 = 1.42

When I divided the total mass of 12 into each of the mass counts, I can then talk about the *mass density* at a location, a concept that will come up repeatedly. In this discrete case that will be the number of mass units at that location divided by the total mass. In statistics, this is called the *probability mass function.* We saw this more abstractly earlier as the *relative frequency of numbers at a given number location*!

When the number of mass particles smears out, I have to talk about a mass density using a *density function*.

So, let me indicate the *mass density* at a given location by "s[y_i]" where y_i takes on each of the location points.

s[-1] = 2/12

s[0] = 1/12

s[1] = 3/12

s[2] = 4/12

s[3] = 0/12

s[4] = 2/12

So, another way to look at *ybar* is that it is

ybar = Sum [each y location times its mass density at that location]

ybar = -1 * s[-1] + 0 * s[0] + 1* s[1] + 2 * s[2] + 3* s[3] + 4 * s[4]

1.42 (as before)

$$\mu = \sum_{i=1}^{5} y[i] * massdensity[y[i]]$$

Variance as the Moment of Inertia

Let me now calculate the Variability (variance) of the batch of temperature numbers. Now I am interested in how far away each temperature is from the overall average, or mean. (See Figure 1).

As in the discussion of the center of mass set-up, replace the horizontal axis, where the temperatures are laid out, with an extraordinarily rigid, light plastic rod, marked off using the temperature numbers as locations along the rod. Further, as before, place a number of *unit masses* at those locations corresponding to the frequencies at that location. That is, at the location of '-1' on the rod, place 2 mass units. At the location marked '0', place one unit, at location '1', place 3 mass units, at '2' place 4 mass units, while at '4', place 2 units.

Now, in addition to asking where this distribution of masses would balance (its center of mass), I will calculate its' moment of inertia about that *center of mass*. In physics and engineering, the moment of inertia is the rotational analog of mass in the linear case. In the linear case, the more massive an object, the harder it is to get it moving. For rotational motion, the inertia is a measure of how hard it is to get an object rotating.

Correspondence of Linear Motion with Rotational Motion

To appreciate the role of inertia in the engineering and physics world, it will help to refresh your memory on some fundamental mechanical laws and draw some correspondences between straight line motion and rotational motion.

Note: the physical moment of inertia is the same as the statistical variance *if* the moment of inertia is taken around the center of mass (the *mean* location of the masses). So, when I say *moment of inertia*, I am talking about the moment of inertia around the batch of numbers center of mass.

Recall Newton's second law for straight line motion:

Force = Mass * Acceleration, F = m* a

If I apply the same force to different masses, it is found experimentally that the more massive object *accelerates more slowly*, in an inverse relationship to the lighter one.

Recall Newton's second law for rotational motion

There is an analogous Newton equation (law) for rotational motion. The rotational analog to force is 'torque' which is defined as a force acting through a lever arm. (Think of a mechanic's *torque wrench*, or the process when you use a lug wrench to take off your cars wheel nuts. What you are doing when you remove the nuts is to apply a force at the end of your lug wrench, and that is 'torque'. Earlier I derived the center of gravity by illustrating it with a see-saw. Placing children on the see-saw so that it doesn't rotate is an example of equilibrium under torques.

Torque = Moment of Inertia * Angular Acceleration, Torque = I * alpha

where 'I' denotes the object's Moment of Inertia (analogous to mass) about an axis through its Center of Mass, and alpha is the rotational acceleration about that axis through the center of mass, in analogy with the linear acceleration 'a'

So, if I apply the same *torque* to an object with moment of inertia I_1 and to another object of moment if inertia I_2, we have:

torque = I_1 * $alpha_1$ = I_2 $alpha_2$

The above equation, verified by experiment, says that the inertias are inversely related to the rotational accelerations just as in the linear case. What this means is that it's harder to get the larger inertia object rotating! Similarly, a batch of numbers with a larger variance is harder to 'spin' than a batch with a smaller variance.

Sample Moment of Inertia

Now, I would like to re-introduce the center of gravity diagram you saw when calculating the mean. This diagram is similar except that

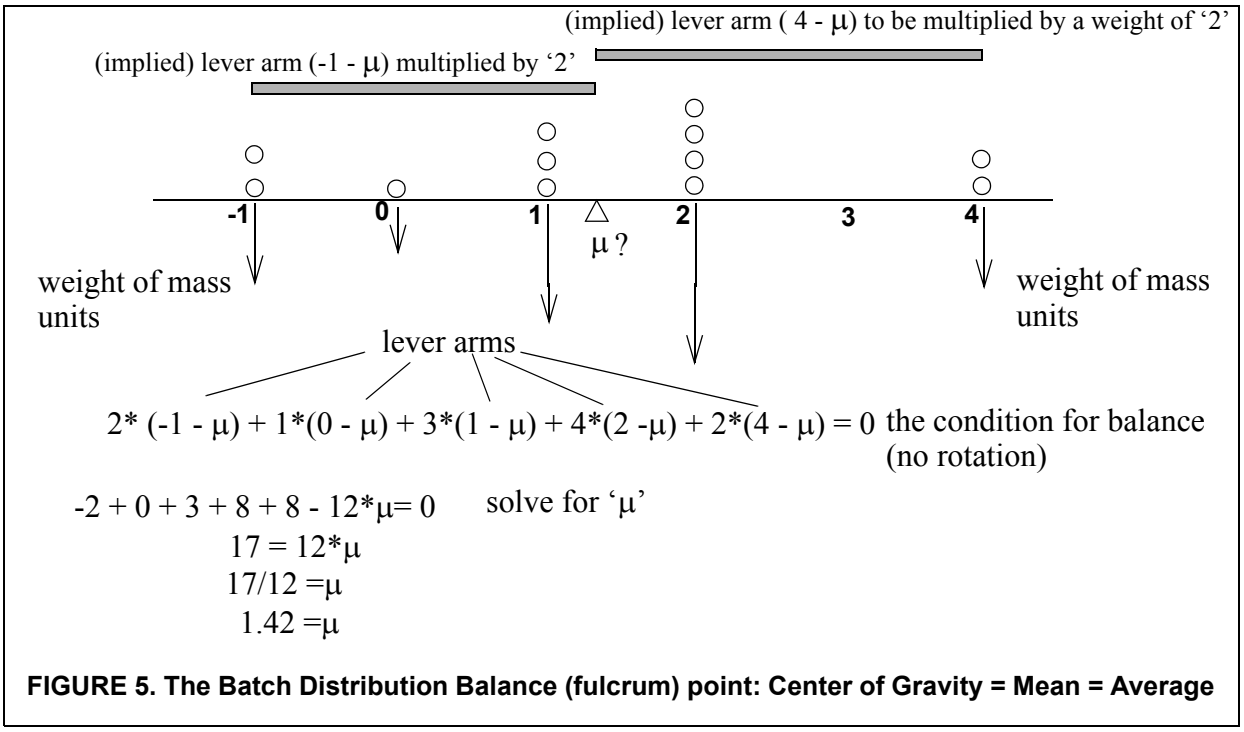


FIGURE 5. The Batch Distribution Balance (fulcrum) point: Center of Gravity = Mean = Average

$$s^2 = \left(\sum_{i=1}^{5} (y[i] - ybar)^2 * frequency[y[i]] \right) \Big/ (totalFrequency - 1)$$

Population Moment of Inertia

$$\sigma^2 = \left(\sum_{i=1}^{5} (y[i] - \mu)^2 \times \text{frequency}[y[i]] \right) / \text{totalFrequency}$$

$$\sigma^2 = \left(\sum_{i=1}^{5} (y[i] - \mu)^2 \times \text{relativeFrequency}[y[i]] \right)$$

$$\sigma^2 = \left(\sum_{i=1}^{5} (y[i] - \mu)^2 \times \text{probability}[y[i]] \right)$$

Finding the Moment of Inertia (Variance), the Continuous Case (*TBD***)**

This section will continue on with the approach shown in "A Special Case for Finding the Mean of a Continuous Distribution" on page 10. Now, the lever arm is squared before multiplying by the mass element. The variance equation now looks like:

$$\sigma^2 = \int (y - \mu)^2 f[y] \, dy$$

FIGURE 6. Moment of Inertia (Variance) of a Continuous Distribution

Random Variables and Expected Values

This section revisits the mean and variance calculations in the context of Random Variables and their Expected Values.

***Notation Conventions*:**

- I will use capital letters for random variables
- I will use lower case letters for realized values of the random variables

- RV will stand for Random Variable
- μ pronounced 'mu', is the population mean
- σ^2 pronounced 'sigma squared', is the population variance
- σ pronounced 'sigma', is the population standard deviation (this is the square root of the population variance).
- P[Y=y] means the probability of getting a realized value of 'y', for the random variable "Y".
- ybar is the sample average (this means that there are more entities that we did not measure, that is, the sample population is just part of the total population)
- s^2 is the sample variance
- s is the sample standard deviation, the square root of the sample variance

What is a Random Variable?

A *Random Variable* is a *rule*, defined on a Population (or Populations), that specifies (describes how to calculate) a number as a result of a random sampling scheme. The "random sampling" part just means that the selection is done without bias (every draw is equally likely, like in the Arizona lottery, each number has the same chance of being drawn). So, a *random sample* of a specified size, from a given population, means that every sample of this size has an equally likely chance of being chosen.

By definition then, the Random Variable is only the rule, it is *not* a number itself. So, the rule tells me *how* to determine a number, based on the sample outcome that I actually get. The number that I actually get is called the *realization* of the experiment. A Random Variable is pretty much like the functions you might have studied that look like, say, f[y1, y2] = y1 + y2 +5. Here, 'f' is the function or *rule* that says you can plug in various numbers for y1 and y2 (observational measurements) to get a resultant number. If you plug in f[9,3] you get 17, a *realization* of the function/rule. As the experimenter, you get to decide what the rule is going to be, although you will usually use only the few standard ones found in text books.

Next I define three Random Variables individually, that I will then combine into an overall Random Variable, representing the rule for calculating the *sample average*. The s*ample average* rule is found in all text books and is the main rule used by experimenters and analysts. (Assume the population is big enough so that you don't have to replace each observation).

Example, here is the rule for a Random Variable," Y1"

(Assume I have an appropriate population to draw from):

I *define* Y1 to be the rule that says:

1. Draw 3 times from the same Population. This will give me an observation vector of three components, that is, three numbers (when I actually carry this out).
2. Keep only the first number

(At this point, I only have a rule, no numbers yet). Now suppose I actually *do* the experiment and get an observation outcome vector, say, *yobserved*= {1, 3, 8} (Note that I am using lower case 'yobserved' here to show that this is the actual outcome (realization) of the RV). So, the *realization* of this experiment for the Random Variable Y1, is the number "1". A little more formally, I could write:

Y1[*yobserved*] = 1

Example 2 The Random Variable, "Y2"

Let me *define* Y2 as the rule that says

1. Draw three times from the same Population as above and so get an observation vector of three components
2. Record the second number
3. Now suppose I actually do the experiment and get the observation vector, say, *yobserved*= {2, 7, 5}

According to the rule Y2, the realization here is the number "7"

Example 3, Random Variable Y3

This is the same as Y1 and Y2 except, I keep the third component.

Example 3, Defining a Sample Average, "YBar" from Y1, Y2, and Y3

An extremely common Random Variable rule is the *Sample Average Rule* that I will write as "YBar" and define as follows (naturally, you could draw a bigger sample, according to your experimental plan):

1. Draw a sample of size 3 from the given population
2. Define YBar as the rule that combines the rules Y1, Y2, and Y3 as follows:
   YBar =(Y1 + Y2 + Y3)/3

This combined rule says: Take a sample of size three, record each component number, add them up and divide by three. Remember, this is just a formula that says what to do before you actually get numbers to work with. Now suppose I actually do the experiment and get an outcome vector, say, yobserved= {4,3,7}.

Applying the Random Variable "YBar" rule I get: *ybar* = (14)/3 = 4.67, which is the realized value. That is the *sample average*, *ybar*, that you are used to seeing.

The Population Mean and Variance of a Random Variable

Let Y be a random variable that says: select *all* of the values from a population of size 'N', and record the values. Let the probability of getting a particular value 'y', be denoted by P[Y=y]. In the common case where we sample in a random manner from a population of size N, then the probability of getting a particular value in that population is P[Y=y] = 1/N. In general, the probabilities need not be equal. Let me illustrate these ideas with two small examples.

**NOTE: Probability is just the long run *relative frequency* of observing a particular outcome.

Expected Value (Mean): Given Equal Probabilities for Every Measurement

Suppose I record all of the weights in a class of 5 people. The entities are the people, and I choose to measure only their weights. The weights are the population of concern (this is now separate from the people that I actually got the weights from). Let *yobserved*= {145, 160, 120, 210} be that population of weights. If I write these numbers on slips of paper, place them in a jar, and then draw one out, I would expect to see each number 1/4 of the time (over the long run, replacing each slip of paper after I draw it). Over the long run, if this ratio really settles down to 1/4, then I can say that the probability (frequency) of drawing 160 is 1/4. More formally, P[Y=160] = 1/4, for example. Y is the random variable that specifies the rule: select an individual, and record their weight.

Similarly, for the other three numbers: P[Y=145] = 1/4, P[Y=120]=1/4, P[Y=210] = 1/4.

The expected value of Y is then:

E[Y]

= 145 *P[Y=145] + 160 *P[Y=160] + 120 * P[Y=120] + 210 * P[Y=210]

= (145 + 160 + 120 + 210) / 4

= 158.75

(This last equality is what you would calculate, where all the probabilities are the same)

Expected Value: Using Relative Frequencies

Suppose I go into another classroom, having 4 people, and weigh them. As luck would have it, the weights turn out to be: yobserved= {150, 150, 180, 200}. Again I copy the values on slips of paper, put them in a jar and draw. Now chances are a little different since 150 appears twice. No problem: P[Y=150] = 2/4, P[Y=180]= 1/4 and P[Y=200] = 1/4

Expected value of Y is then:

E[Y]

= 150 * P[Y=150] + 180 * P[Y=180] + 200 * P[Y=200]

= 150 * 2/4 + 180 * 1/4 + 200 * 1/4

= 170

Notice though, that I could have written it out in full (so long as every measurement has the same probability):

150 * 1/4+ 150 * 1/4 + 180 * 1/4+ 200 * 1/4

Mean Calculations for a Population

E[Y] is a common way to indicate the population mean of the random variable "Y" and is read: The expected value of Y is. . ." . Often, the mean of the RV is simply written as μ.

To actually find out what E[Y] is, use one of the definitions below:
E[Y]=

1. $\sum$ measurements / number of measurements (for discrete populations)
2. $\sum$ measurement * relative frequency (for discrete populations)
3. $\int$ y* f[y] dy (for continuous populations, where f[y] is the probability density function for the random variable Y)

Variance Calculations for a Population

The Variance of a Random Variable Y is often written as Var[Y]. A shorter way to write it is: σ^2 (sigma squared)

To calculate the variance of the Random Variable Y for a population, use one of the definitions below:

Var[Y]=

1. $\sum$ (measurements - population mean)^2/ number of measurements (discrete distributions)

2. $\sum$ (measurement - population mean)^2 * relative frequency of measurement (discrete distributions)
3. $\int (y- \mu)^2 * f[y] dy$ (continuous distributions, where f[y] is the probability density function for the random variable Y)

Variance Calculations for a Sample

1. The sample variance, denoted by S^2 is calculated similarly as in the population case, but substituting the sample mean in place of the population mean and divide by n-1 rather than n.

Rules for Combining Random Variables (RVs)

Let me use a particular combination to illustrate these rules. I will only deal with linear combinations of random variables. That means that each random variable appears to the first power only and is multiplied by a real number. The rules actually work for the sum of any number of RVs. The a1, a2, and a3 are plain numbers like 1/3, -2, or 7.4. The Y1, Y2, Y3 are general Random Variable rules and "Y" itself is a Random Variable rule that combines the individual rules.

Let Y= a1* Y1 + a2 * Y2 + a3 * Y3

Rule 1: The population mean of Y, which can be written briefly as μ, is defined to be the Expected Value of Y and is given by:

E[Y] = E [a1 * Y1 + a2 * Y2 + a3 * Y3]

= a1 * E[Y1] + a2 * E[Y2] + a3 * E]Y3]

That is, the mean of a sum of random variables is the sum of the means.(This holds true whether or not the Y's are normal, it holds true for any sum of random variables, Binomial, Poisson or whatever. The kicker is that the sum of random variables is not usually the same type of distribution as its components. For normal RVs however, the sum is again *normal*)

Rule 1*: Special Case that is Very Important! If Y1, Y2 and Y3 are *normal* RVs then Y is also a normally distributed RV-- that is, the linear sum of normal random variables results in a new random variable that is also normally distributed. The properties of this new variable, Y, are listed next.

Rule 3: If the Y1, Y2, Y3 are independent RVs, then the variance of Y, which can also be denoted by σ^2, is given by:

Var[Y] = Var [a1 * Y1 + a2 * Y2 + a3 *Y3]

= $a1^2$* Var[Y1] + $a2^2$* Var[Y2] + $a3^2$ * Var[Y3]

A Few Examples:

Example 4 Adding A Few Independent Random Variables

Suppose E[Y1] = 5, E[Y2] = 9, Var[Y1] = 1, Var[Y2] = 6

Let Y = 2 Y1 + 3 Y2

E[Y] =E[2* Y1] + E[3* Y2] (check out the use of the rules here)

= 2 * E[Y1] + 3 *E[Y2]

= 2 * 5+ 3 * 9= 37

Var[Y] = Var[2*Y1] + Var[3*Y2]

= 4 * Var[Y1] + 9 * Var[Y2]

= 4 * 1 + 9 * 6 = 58

Let Y = (2 *Y1 + 3 Y2) / 5 = 2/5 Y1 + 3/5 Y2

Var[Y] = (2/5)^2 Var[Y1] + (3/5)^2 Var [Y2]

= 4/25 * 1 + 9/25 * 6

58/25

Summary

I have presented a few ideas on how to calculate means, variances, as well as some notes on Random Variable. I hope these thought will help you in your individual work as well.