

Exploring CLT, Part 1: Proving Standard Error

Zachary Fierstadt

November 28, 2011

Abstract

When dealing with large data sets, one strives to make accurate inferences based on samples, often with the aim of making critical decisions based on evidence, not just "educated guesses" or optimistic hindsight leading to "we think we improved the situation." How much did that new taxonomy actually improve the search experience? How confident are we that the new relevancy model lead to increased click-through for those top 100 skus? How do we approximate our bandwidth utilization on a 10 gigabit link without collecting every single packet?

Sampling is a practical way of dealing with large amounts of data. In many cases, we lack the resources (whether it be data itself, time, CPU cycles etc.) to measure every member of a given population - so if we sample, we want to quantify how close we are to "reality". Part 1 of this blog series will be an introduction to CLT and end with a proof for standard error.

1 Central Limit Theorem

Before we dive into standard error, let's discuss Central Limit Theorem (CLT). At first glance, the CLT seems unintuitive to the casual observer - it states: given any population, random samples of its mean will converge to a normal distribution. At the surface, even stranger is what it implies: regardless of distribution of the underlying population (i.e. whether its bimodal, positively skewed, negatively skewed, non-normal, etc.), random sampling of the mean will approximate a normal distribution, given a large enough sample size.

To give a visual example, here is a probability

distribution that could represent a given population:

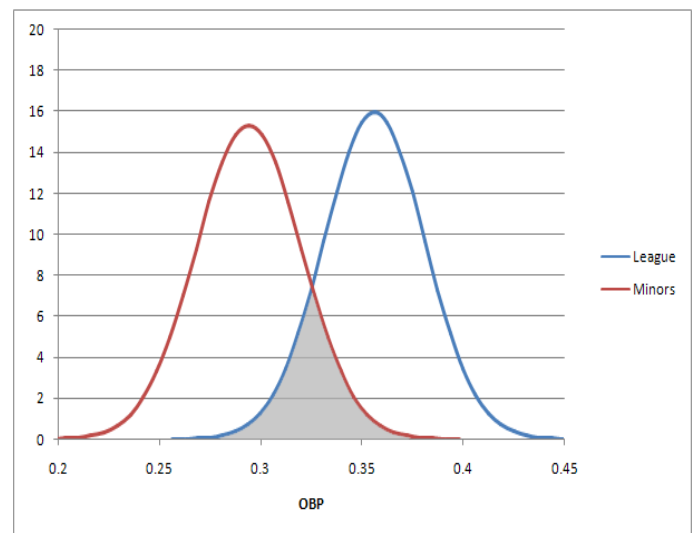


Figure 1: Bimodal Distribution

In the figure above, we have a bimodal distribution, with two peaks or local maximas (often a result of merging two unimodal distributions). If we did not have access to the underlying population, how could we accurately gather statistics about this distribution? At first glance, the two modes appear to be 15 and 16. If we were to take a small random sample of this population, would we be able to deduce the bimodality of the underlying distribution?

Probably not. But according to the CLT, with a limited sample we could very closely approximate μ , or its mean - which is powerful, because it implies that no matter how volatile the underlying distribution is, we can always calculate a very important statistic about the underlying population.

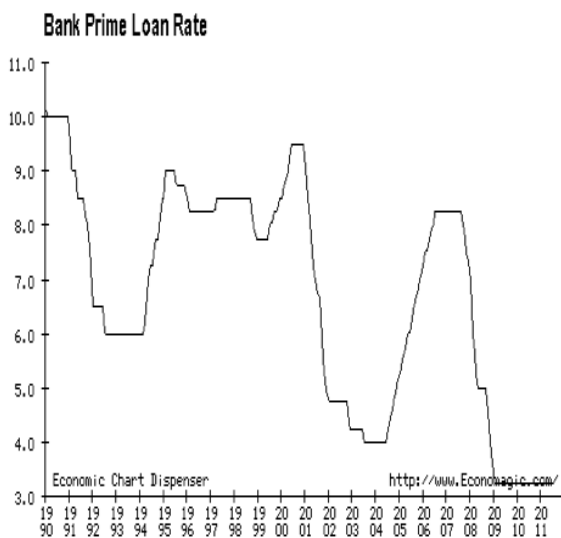


Figure 2: Non-Normal Prime Loan Rate

Figure 2 exemplifies another non-normal distribution. The distribution does not resemble a bell curve, data is not symmetrically clustered about the center. But regardless of the non-normality of the distributions depicted in Figures 1 and 2, the CLT tells us that given a large enough sample size, the sample means of both populations will always approach a normal distribution, as depicted below:

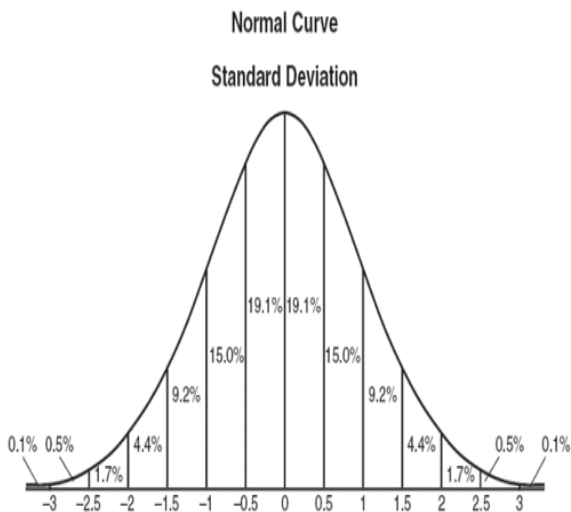


Figure 3: Normal Distribution

Regardless of how non-normal the probability distribution is, the CLT postulates that samples of the mean will converge to a bell-curve as in Figure 3. The CLT provides us with a valuable tool because we usually do not have the ability to survey the entire population - and thus it may be difficult to determine some other measure of central tendency (i.e. median or mode). Because we know the sample mean will always converge to a normal distribution, given a large enough sample size we can not only approximate the mean of the underlying population (despite its "funky" distribution), we can also determine how close we are to the true mean - so we can actually measure how close to reality our approximation is.

But how can we be certain the CLT is true?

2 Variance and Standard Deviation

The phrase "sample size" has been mentioned several times now - you may be wondering how much sample size affects the approximation of μ . To help us understand this, let's start with the concept of variance. Variance is a measure of the dispersion of a given data set around its mean, given by:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

In other words, the variance gives us the average squared difference between an observed event and the mean. There are a few side effects of squaring the deltas. Squaring the deviations allows one to take negative values into account as part of the aggregate deviation. Additionally, the squares magnify large deviations, and shrink small deviations - so that large outliers greatly increase the resulting variance, and fractional differences between the event and the mean are marginalized. Why not just average the deltas between each event and the true mean, you might ask? This is actually another statistic, called the mean deviation. Mean absolute deviation, or MD, is given by:

$$D_i = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)$$

Some statisticians argue that MD is superior to variance for samples that contain small errors or large outliers, since σ^2 can virtually eliminate the purported existence of minor errors in samples, and lend itself to “exploding” when there are long-tailed distributions - the outliers exponentially add to the variance. This debate is outside of the scope of this blog, but worth studying.

Once we’ve calculated variance, we end up with squared units - and often we want to represent the spread of the data in its original units, (i.e. feet instead of square footage). The positive square root of the variance, known as the standard deviation, gives us the dispersion of the data in original units.

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

Standard deviation, or σ , attempts to quantify the dispersion of the data set about its mean, and has pragmatic applications outside of science - i.e. one method of comparing the risk between two different equities is juxtaposition of their respective standard deviations. If the the σ of one stock is higher than another, it implies there is more volatility or price movement over a given period of time, and thus a riskier investment. If one were looking for long-term investment, a stable stock (less price movement) with a smaller σ might be a consideration. Investors seeking risk might choose stocks with higher values for σ , since they could potentially make gains off price fluctuations in the short-term.

Another example might be optimization of an e-commerce platform. If a firm wanted to understand the effectiveness of their product search engine, they might sample click-through on each search result for a high-value query over a one-hour period. While μ could potentially be a decent value of central tendency in this case, it may not reveal the entire picture. What if average clicks per result was 2.89, given the page had 10 displayed results? This could mean every result is receiving roughly 3 clicks; or one of the results is receiving 20 clicks, while the rest are receiving 1. The distribution of the latter would ultimately be a positively skewed, long-tailed distribution, whereby μ would not be a

good measure of central tendency. Key stakeholders in the project might request an efficient way of determining the “spread” of the clicks over the top N results - and σ would be one measure of this spread.

In the vein of the previous application for e-commerce, lets do an example calculation of σ . Suppose over a one-hour period we received the following clicks for 10 search results displayed on a page:

12, 4, 9, 15, 12, 16, 17, 3, 5, 12

Their sum is 105 and their mean is 10.5. Upper-management might deem 8 clicks per keyword is satisfactory to hit their revenue numbers, but they want an indicator of how spread out the data is over this measurement. The deviations from the mean are:

1.5, -6.5, -1.5, 4.5, 1.5, 5.5, 6.5, -7.5, -5.5, 1.5

To eliminate the negative values, the deviations are squared, giving:

2.25, 42.25, 2.25, 20.25, 2.25, 30.25, 42.25, 56.25, 30.25, 2.25

The sum of these squared deviations is 230.5. The mean of these deviations is 23.05, which is the variance of the original values. The standard deviation is the positive square root, or 4.8. In otherwords, the average deviation from the mean (10.5) is +/- 4.8, or between 5.7 and 15.3 clicks. Yielding on the side of caution, upper-management might decide a smaller σ is required so that the spread is more tightly wrapped around the target 8 clicks per result.

Thus far we’ve calculated standard deviation for an entire population - i.e. we’ve had access to all the elements in the data set, and were able to calculate the population’s true μ and σ . But what if we don’t have access to the entire population, and we must use samples to approximate these statistics? Can we even be sure we can arrive anywhere close to the true mean and variance if we use samples?

3 Sample Statistics

In many cases, it will be impractical, if not impossible, to sample every member of a given population.

When we don't have the ability to directly calculate population statistics such as μ and σ , we have to sample. This is where we would like to validate CLT - which postulates that given a large enough sample size, random samples of its population will converge to a normal distribution, and the sample mean will converge to the true population mean.

Let us start with a basic example with a small population. Suppose we have built a search engine for IT products and we have tallied click-through for 6 of the top results for a given query. The total clicks for each result over a 24-hour interval are tallied and sorted:

260, 326, 412, 512, 515, 787

The mean click-through for the population above is 468.7. If we were to take samples of size two and average them, what would our sample distribution look like? The fifteen possible random samples and their respective means are tabulated below:

X_1, X_2	\bar{x}
260, 326	293
260, 412	336
260, 512	386
260, 515	387.5
260, 787	523.5
326, 412	369
326, 512	419
326, 515	420.5
326, 787	556.5
412, 512	462
412, 515	463.5
412, 787	599.5
512, 515	513.5
512, 787	649.5
515, 787	651

Each of the enumerations above has the same probability, 1/15, of being selected as a random sample. While the original values range from 260 to 787, the sample means only range from 293 to 651 - and the majority of the values lay in the middle region of that range. If the CLT is true, we should see this range tighten and the midpoint approach the true mean as the sample size gets larger.

X_1, X_2, X_3, X_4, X_5	\bar{x}
260, 326, 412, 512, 515	405
260, 326, 412, 512, 787	459.4
260, 326, 412, 515, 787	460
260, 326, 512, 515, 787	480
260, 412, 512, 515, 787	497.2
326, 412, 512, 515, 787	510.4

From the table above, one can see that the values for \bar{x} become more tightly concentrated about the mean as the sample size increases. Additionally, after sample size increased, values became much closer to μ , or 468.7. This is intuitive at this point - as our sample size increases, we become closer and closer to the actual calculation of μ .

4 Verifying Sample Mean

In the previous section, we gathered sample statistics about a population over two different sample sizes, and we observed that as sample size increased, we approached the true mean of the population. How can we better establish this?

1. **Expected value** of a random variable is the weighted average of all possible values the variable can take; in other words, the long-term average outcome of a given scenario:

$$\mathbf{E}(X) = \sum_x xP(x) = \mu \quad (1)$$

2. **Sample mean** is the sum of random samples divided by the sample size, given by:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

Before we continue, let's go over linear combinations. A linear combination is any expression of the form $C = aX + bY$, where a and b are constants. Consider the following:

$$C = \frac{1}{4}X + \frac{3}{4}Y \quad (3)$$

In this case, $\frac{1}{4}$ and $\frac{3}{4}$ are linear weights, and are held constant regardless of variables X and Y . If we had to determine the mean of this linear combination, how would we do it?

$$\mathbf{E}(C) = \frac{1}{n}(\frac{1}{4}X_1 + \frac{3}{4}Y_1 + \frac{1}{4}X_2 + \frac{3}{4}Y_2 \dots + \frac{1}{4}X_n + \frac{3}{4}Y_n) \quad (4)$$

$$= \frac{1}{n}(\frac{1}{4}X_1 + \frac{1}{4}X_2 \dots + \frac{1}{4}X_n) + \frac{1}{n}(\frac{3}{4}Y_1 + \frac{3}{4}Y_2 \dots + \frac{3}{4}Y_n) \quad (5)$$

$$= \frac{1}{4}(\frac{1}{n} \sum_{i=1}^n X_n) + \frac{3}{4}(\frac{1}{n} \sum_{i=1}^n Y_n) \quad (6)$$

$$= \frac{1}{4} \mathbf{E}(X) + \frac{3}{4} \mathbf{E}(Y) \quad (7)$$

As it turns out, the expectation of a linear combination of random variables is the linear combination of the expectations of those random variables. Notice that when we went from equation 5 to equation 6, the constants in the respective linear combinations were pulled through the expectations. This is actually one of the laws of linear combination - lets define them:

$$\mathbf{E}(a) = a \quad (8)$$

$$\mathbf{E}(aX) = a \mathbf{E}(X) \quad (9)$$

$$\mathbf{E}(aX + bY) = a \mathbf{E}(X) + b \mathbf{E}(Y) \quad (10)$$

The laws above, while not generally proven by our example, should seem intuitive after having worked through equations 4 through 7. A more general proof would be trivial - simply replace the linear weights in equation 3 with terms a and b , and you should be able to verify the laws above.

Now that we have an understanding of linear combinations and their expectations, we can define the expectation of the sample mean - which is itself a linear combination.

$$\mathbf{E}(\bar{X}) = \mathbf{E}(\frac{1}{n}(X_1 + X_2 + \dots + X_n))$$

Look familiar? It's a linear combination. Since \mathbf{E} is a linear operator, the equation above can be rewritten further:

$$= \frac{1}{n}(\mathbf{E}(X_1) + \mathbf{E}(X_2) + \dots + \mathbf{E}(X_n)) \quad (11)$$

$$= \frac{1}{n}(\mu_{x_1} + \mu_{x_2} + \dots + \mu_{x_n}) \quad (12)$$

$$= \frac{n\mu}{n} \quad (13)$$

$$= \mu \quad (14)$$

Thus the mean of the sample means is the population mean. While a given sample mean may be smaller or larger than the population mean, the expected value, or long term average, of the sample means will be the underlying population mean. You can quickly validate this - find the means of the sample means in the previous two tables and you'll find that they equal 468.7. If one sums all the combinations of the possible sample means, after some rearranging it's obvious that you're just summing the population means.

For example, if the set 1,2,3 represents a population, and you have a sample size of 2, there are $3C2$ possible sample means:

X_1, X_2	$P(\bar{X})$
1,2	1/3
1,3	1/3
2,3	1/3

In this case,

$$\mathbf{E}(\bar{X}) = \frac{1}{3}(\frac{1}{2}(1+2) + \frac{1}{2}(1+3) + \frac{1}{2}(2+3)) \quad (15)$$

$$= \frac{1}{6}(1+2+3+1+2+3) \quad (16)$$

$$= \frac{\frac{1}{3}(1+2+3) + \frac{1}{3}(1+2+3)}{2} \quad (17)$$

$$= \frac{\mu + \mu}{2} \quad (18)$$

$$= \mu \quad (19)$$

5 Verifying Sample Variance

So at this point you should be convinced that the mean of the sample means approaches the population mean. In the previous section, we observed that given enough samples, the sampling distribution will approach μ regardless of sample size; however, with larger samples, the sampling distribution is tighter and the individual sample means are much closer to the true mean.

Given a random sample, how can we gauge how close we are to the true mean of the population? As our sample size becomes larger, our sampling distribution more closely approximates a normal distribution - in other words, the spread tightens. Since our sample means are random variables themselves, we can use variance to calculate how tight the sampling distribution is.

Before we calculate the variance of the sample mean, let's establish a few laws that will help us prove the variance of \bar{X} .

1. **The variance of a linear function** is the variance of a random variable that takes the form $Y = aX + b$.

$$\text{Var}(aX + b) = E[(aX + b - E(aX + b))^2] \quad (20)$$

$$= E[(aX + b - (aE(X) + b))^2] \quad (21)$$

$$= E[(aX + b - aE(X) - b)^2] \quad (22)$$

$$= E[(aX - aE(X))^2] \quad (23)$$

$$= E[a^2(X - E(X))^2] \quad (24)$$

$$= a^2 E[(X - E(X))^2] \quad (25)$$

$$= a^2 \text{Var}(X) \quad (26)$$

2. **The variance of a linear combination** is the variance of an expression that takes the form $aX + bY$. The proof for this is not as trivial as the one for the expectation for a linear combination, and is out of the scope of this blog (for now).

Note: this applies to a linear combination of independent random variables; it does not apply for combinations of dependent variables.

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) \quad (27)$$

Now we can prove out the variance of the sample mean, given a sample size of n .

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n (X_i)\right) \quad (28)$$

$$= \frac{1^2}{n} \text{Var}(X_1 + X_2 + \dots + X_n) \quad (29)$$

$$= \frac{1^2}{n} (\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)) \quad (30)$$

$$= \frac{1^2}{n} (n\sigma^2) \quad (31)$$

$$= \frac{\sigma^2}{n} \quad (32)$$

To find the standard deviation of the sample means, also known as the **standard error** of the sample means, simply take the square root of the variance:

$$\sigma = \sqrt{\left(\frac{\sigma^2}{n}\right)} \quad (33)$$

$$= \frac{\sigma}{\sqrt{n}} \quad (34)$$

So this equation should no longer be a grand mystery to the statistics novice - though variance of linear combinations needs to be proven out.

Stay tuned for Part 2.