An intuitive derivation of a sample size calculation.

Jesse Hoey and Robby Goetschalckx

June 8, 2010

This note will give an intuitive derivation of a standard formula for computing the necessary sample size for an experiment. Typical sample size calculators (found online at e.g. `www.surveysystem.com/sscalc.htm`) ask for two things:

- the confidence level (usually given as 95% or 99%)
- the confidence interval (usually given as a std. deviation or percentage)

The calculator then computes the necessary sample size: the number of subjects you need in your experiment to achieve the confidence interval with the given confidence level. In this note, we will discover what these two fellows with similar names really are.

We assume that we are trying to measure some quantity, X, of a population, and that the values of this quantity X are *normally distributed* across the population (see question 1 below). For example, we may be trying to compute the average weight, X, of all gnus in Africa. We assume that the distribution of gnu weights follows a normal curve, that is $x \sim \mathcal{N}(\mu, \sigma)$, where $\mathcal{N}$ is a Gaussian curve, μ is the mean value of x across the entire population of gnus, and σ^2 is the variance of x about the mean across the population.

Now, as gnus are difficult to find, we want to measure as few as possible and still be able to say with some degree of confidence that our calculated average is somewhat meaningful, i.e. it is reasonably close to the true average we'd get if we really sought out all gnus and measured all their weights (so we'd get *exactly* μ if there were infinitely many gnus). So let's imagine an experiment (call it experiment number 1) where we find N randomly selected gnus, measure the weight of the i^{th} gnu, x_i, and take the mean, $\tilde{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i$. Now, if we do this experiment again (experiment number 2), with another set of N randomly selected gnus, we'll get another mean $\tilde{\mu}_2$. Notice that $\tilde{\mu}_2 \neq \tilde{\mu}_1$ in general (since the actual set of gnus used - the samples - for each experiment are different). So we have

$$\begin{array}{ll}
\text{Experiment 1:} & \text{take mean of N randomly picked gnus} \rightarrow \tilde{\mu}_1 \\
\text{Experiment 2:} & \text{take mean of another N randomly picked gnus} \rightarrow \tilde{\mu}_2 \\
\text{Experiment 3:} & \text{take mean of another N randomly picked gnus} \rightarrow \tilde{\mu}_3 \\
 & \vdots
\end{array}$$

Now it just so happens that these means, $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3 \ldots$ are also distributed as a normal, and this normal has mean μ and variance $\frac{\sigma^2}{N}$ (where μ and σ^2 here are the mean and variance of the entire population!). This can be derived quite easily by using the fact that the variance of a sum of uncorrelated random variables is equal to the sum of their variances:

$$var(\sum_{i=1}^{M} x_i) = \sum_{i=1}^{M} var(x_i) \tag{1}$$

This is known as the *Bienaymé formula* (see Appendix A for a proof). Using this, we can derive that (see question 2 below):

$$var(\tilde{\mu}_i) = \frac{\sigma^2}{N} \tag{2}$$

So this means that the mean we calculate in each experiment with a sample size of N will fall within $\pm\frac{\sigma}{\sqrt{N}}$ of the true population mean roughly 66% of the time. Why 66%? Because 66% of the area under a normal curve lies within one standard deviation σ of the mean. Similarly, about 95% of the area under a normal curve lies within two standard deviations 2σ of the mean, and 99% of the area under a normal curve lies within three standard deviations 3σ of the mean. In fact, there is a formula for computing how much area is under the normal curve within n standard deviations, and we'll call this formula $f(n)$, so that $f(1) = 66, f(2) = 95, f(3) = 99$, etc (see Appendix B). Aha! Now we see our old friend the *confidence level* showing up again! The sample size calculation is trying to choose N such that we have a *confidence level* chance that our mean value is "close" to the true population mean, where "close" is defined as some number of standard deviations, n.

Therefore, if we want our experiment to be within Δ of the true mean, we want

$$\frac{n\sigma}{\sqrt{N}} < \Delta$$

so, we want a sample size of

$$N > \left(\frac{n\sigma}{\Delta}\right)^2$$

As long as we specify Δ as a fraction of σ, then we are all good. The ratio $\frac{\Delta}{\sigma}$ is known as the *confidence interval* (sometimes quoted in percentage as $100 \times \frac{\Delta}{\sigma}$).

Suppose we are trying to find evidence for the hypothesis that gnus are not the same weight as emus. We would do the same experiment as above for both gnus and emus, using a sample size of N such that, if gnus and emus are, actually, not the same weight, then the weight difference will show up in the two mean values we compute. That is, if the actual mean weight of gnus is μ_{gnu} and for the emus its μ_{emu}, then we want our *estimates* of these two quantities with a sample size of N, $\tilde{\mu}_{gnu}$ and $\tilde{\mu}_{emu}$, to be close enough to the true values so that the difference will show up.

Let us assume that the variances of the weights of emus and gnus are the same $\sigma_{emu} = \sigma_{mu} = \sigma$, then, with N samples,

- $\tilde{\mu}_{gnu}$ will be within $n\sigma$ of μ_{gnu} $f(n)$% of the time, and
- $\tilde{\mu}_{emu}$ will be within $n\sigma$ of μ_{emu} $f(n)$% of the time, and

then, if μ_{gnu} and μ_{emu} are further apart than $2n\sigma$, we will expect to see $\tilde{\mu}_{gnu} \neq \tilde{\mu}_{emu}$ at least $f(n)\%$ of the time!. Call $\Delta m = |\mu_{gnu} - \mu_{emu}|$, then we want

$$\Delta m > \frac{2n\sigma}{2\sqrt{N}}$$

and thus that

$$N > \left(\frac{n\sigma}{\Delta m}\right)^2$$

But wait! We don't know what Δm is, since it involves knowing μ_{gnu} and μ_{emu}, which is what we're trying to measure! And, we don't know σ either! This is where the black magic starts that this note will not cover. However, all we really need to know is the ratio of $\frac{\sigma}{\Delta m}$, which we may be able to figure out.

Questions:

1. Why do we assume the population is normally distributed? In what cases would this not be true?
2. Using the Bienaymé formula (Equation 1), derive Equation 2

Answers:

1. Central limit theorem: the sum of infinitely many white noise sources is a Gaussian noise source. The distribution of the sum of infinitely many dice throws is a Gaussian. In cases where the noise is not white.
2. let $z = \sum_{i=1}^{N} x_i$, and $\bar{z} = \frac{1}{M}\sum_{i=1}^{M} z$, then

$$\begin{aligned}
var(\tilde{\mu}_i) &= var(\frac{1}{N}\sum_{i=1}^{N} x_i) && \text{(definition of mean)}\\
&= var(\frac{z}{N}) && \text{(definition of } z\text{)}\\
&= \frac{1}{M^2}\sum_{i=1}^{M}(\frac{z}{N} - \frac{\bar{z}}{N})^2 && \text{(definition of variance)}\\
&= \frac{1}{M^2}\frac{1}{N^2}\sum_{i=1}^{M}(z - \bar{z})^2 && \text{(distributivity)}\\
&= \frac{1}{N^2}var(z) && \text{(definition of variance)}\\
&= \frac{1}{N^2}\sum_{i=1}^{N} var(x_i) && \text{(Bienaymé formula)}\\
&= \frac{\sigma^2}{N} && \text{(assumption: all } x_i \text{ have same variance)}
\end{aligned}$$

A Proof of Bienaymé's formula

Let $X_1, X_2, \ldots, X_m$ be a set of m variables with relative means $\mu_1, \ldots, \mu_m$ and assume that they are uncorrelated:

$$i \neq j \Rightarrow \sum_{x_i} \sum_{x_j} P(x_i)P(x_j)(x_i - \mu_i)(x_j - \mu_j) = 0 \tag{3}$$

Theorem 1. ***Bienaymé's formula***
$Var\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} Var(X_i)$

Proof. For notational easy we will write $\vec{x}$ for the combined variable $(x_1, \ldots, x_m)$.
We start with working out the average of $(\sum_{i=1}^{m} X_i)$.

$$\sum_{\vec{x}} P(\vec{x}) \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} \sum_{\vec{x}} P(\vec{x}) x_i = \sum_{i=1}^{m} \mu_i \tag{4}$$

This means the average of the sum of the variables is just the sum of the averages. Now we can use this in the definition of the variance:

$$\begin{aligned}
Var\left(\sum_{i=1}^{m} X_i\right) &= \sum_{\vec{x}} P(\vec{x}) \left(\left(\sum_{i=1}^{m} x_i\right) - \left(\sum_{i=1}^{m} \mu_i\right)\right)^2 \\
&= \sum_{\vec{x}} P(\vec{x}) \left(\sum_{i=1}^{m} (x_i - \mu_i)\right)^2 \\
& \quad \text{Expanding and changing one of the } i\text{'s to a } j \text{ to avoid confusion:} \\
&= \sum_{\vec{x}} P(\vec{x}) \left(\sum_{i=1}^{m} (x_i - \mu_i)\right) \left(\sum_{j=1}^{m} (x_j - \mu_j)\right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{\vec{x}} P(\vec{x})(x_i - \mu_i)(x_j - \mu_j) \qquad (5)
\end{aligned}$$

Using the fact that the variables are uncorrelated (equation (3), we see that the term equals 0 if $i \neq j$, which simplifies equation (5) to:

$$\begin{aligned}
Var\left(\sum_{i=1}^{m} X_i\right) &= \sum_{i=1}^{m} \sum_{\vec{x}} P(\vec{x})(x_i - \mu_i)^2 \\
&= \sum_{i=1}^{m} \sum_{x_i} P(x_i)(x_i - \mu_i)^2 \\
&= \sum_{i=1}^{m} Var(X_i) \qquad (6)
\end{aligned}$$

□

B Derivation of confidence interval function

To derive the function $f(n)$, we first derive an expression for the area under a Gaussian curve $\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx$ between any two points a and b. The expression will contain the error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$, and the complementary error function $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$

$$\begin{aligned} p(-\sigma < x < \sigma) &= \int_a^b \mathcal{N}(x;\mu,\sigma)dx \\ &= \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \end{aligned}$$

`change variables` $t = \frac{x-\mu}{\sqrt{2}\sigma}$

$$= \frac{1}{\sqrt{\pi}}\int_{\frac{a-\mu}{\sqrt{2}\sigma}}^{\frac{b-\mu}{\sqrt{2}\sigma}} e^{-t^2}dt$$

use $\int_a^b f(x)dx = \int_a^\infty f(x)dx - \int_b^\infty f(x)dx$

$$\begin{aligned} &= \frac{1}{\sqrt{\pi}}\left[\int_{\frac{a-\mu}{\sqrt{2}\sigma}}^{\infty} e^{-t^2}dt - \int_{\frac{b-\mu}{\sqrt{2}\sigma}}^{\infty} e^{-t^2}dt\right] \\ &= \frac{1}{2}\left[\mathrm{erfc}\left(\frac{a-\mu}{\sqrt{2}\sigma}\right) - \mathrm{erfc}\left(\frac{b-\mu}{\sqrt{2}\sigma}\right)\right] \qquad (7) \end{aligned}$$

Next, we use the expression derived above to write an expression for the probability that a sample drawn from a Gaussian PDF falls within n standard deviations of the mean. To do this, we use the following facts: (1) that this answer will be the same regardless of the value of the mean; (2) that $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$; and (3), that $\mathrm{erf}(x)$ is an odd function. Writing Equation 7 with with $a = -n\sigma$ and $b = n\sigma$, we get

$$\begin{aligned} f(n) = \int_{-n\sigma}^{n\sigma} \mathcal{N}(x;0,\sigma)dx &= \int_{-n\sigma}^{n\sigma} \mathcal{N}(x;\mu,\sigma)dx \\ &= \frac{1}{2}\left[\mathrm{erfc}\left(\frac{-n\sigma-\mu}{\sqrt{2}\sigma}\right) - \mathrm{erfc}\left(\frac{n\sigma-\mu}{\sqrt{2}\sigma}\right)\right] \end{aligned}$$

setting $\mu = 0$ and using the fact that $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$

and so $\mathrm{erfc}(-x) <= 1 - \mathrm{erf}(-x) = 1 + \mathrm{erf}(x)$

$$\begin{aligned} &= \frac{1}{2}\left[2\mathrm{erf}\left(\frac{n}{\sqrt{2}}\right)\right] \\ &= \mathrm{erf}\left(\frac{n}{\sqrt{2}}\right) \end{aligned}$$

This expression can be used directly as $f(n)$ if we use a numerical approximation to erf, as in Matlab. Alternatively, we can use the fact that $f(n)$ will be the same regardless of the value of the mean, and the following approximations for erf.

$$\text{erf}(\frac{1}{\sqrt{2}}) = 0.6827$$

$$\text{erf}(x) = 1 - \frac{e^{-x^2}}{\sqrt{\pi}x}\left[1 - \frac{1}{2x^2}\right]$$

As an example, using the above expression with different values of n, we get

n	answer
1	0.6827
2	0.9595
4	0.9999
6	0.99999999
8	0.999...to at least 15 digits