Probability and Statistics
for Final Year Engineering Students

By Yoni Nazarathy, Last Updated: May 2, 2011.

Lecture 2:
Sampling Distributions: Independence, Binomial, HG, Normal and the CLT

Random samples, independence and statistics

Two random variables (RVs), X and Y are **independent** if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ or alternatively,

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F(x)F(y).$$

In case of discrete RVs:

$$P(X = x, Y = y) = P(X = x)P(Y = y) = p(x)p(y),$$

And in case of continuous RVs:

$$P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y) = P(x \leq X \leq x + \Delta x)P(y \leq Y \leq y + \Delta y) \cong f(x)f(y).$$

Note that in the above, the notation $P(C, D)$, can be read as $P(C \cap D)$.

Independence can also be defined for an arbitrary number of random variables, e.g. in the discrete case, independence of $X_1, X_2, \ldots, X_n$ is:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i).$$

Mathematically, independence is a useful property because it allows us to easily evaluate the distribution of to two (or more) random variables by multiplying distributions of individual random variables.

Example: Let I_1 be the result of the "first" coin flip and let I_2 be the result of the "second" coin flip. In this case it is very sensible to assume that $I_1$and I_2 are independent.

Example: Let I_1 be the result of the "first" kick towards the goal of a football player and let I_2 be the result of the "second" kick. In this case, one may still "assume" independence from a modeling perspective, but the validity of the model is sometimes questionable. Why?

Example: Assume a box with 3 marbles, one of which is red and the other blue. You blindly pick marbles out of the box sequentially (not returning them), yielding the random variables I_1, I_2, I_3, where I_i=1 if the marble in the i'th pick is red and 0 otherwise. In this case it is obvious that the random variables are not independent: The distribution of the second pick depends on the result of the first pick etc… We can

see this more precisely by calculating the probability of each sequence of I_1, I_2, I_3 (there are 8 sequences in total).

P(0,0,0)= P(0,1,1)=P(1,0,1)=P(1,1,0)=P(1,1,1)=0. Why? Whereas,

P(1,0,0)=1/3 *1 *1 = 1/3

P(0,1,0)=2/3 * 1/2*1 =1/3

P(0,0,1)=2/3 * 1/2 *1 =1/3.

Now looking at the distribution of each result individually it is clear that, $P(I_i = 1) = 1/3$ for i=1,2,3. So we have that

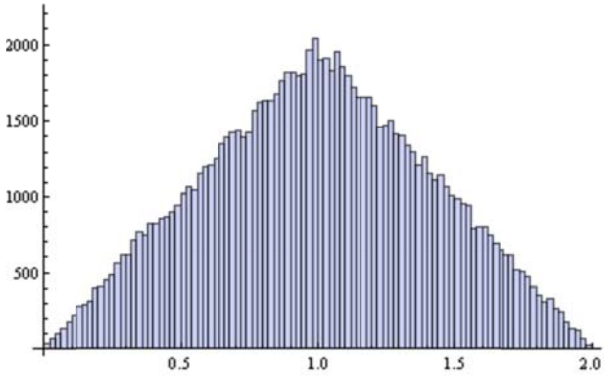
$$P(1,0,0) \neq \frac{1}{3}\frac{2}{3}\frac{2}{3},$$

and thus these three RVs are not independent.

Example: Let U_i, i=1,2,3 be independent uniform [0,1] random variables. Let

$X_1 = U_1 + U_2$ and $X_2 = U_2 + U_3$. Are X_1 and X_2 independent?

First let's look at the distribution of X_1 (the distribution of X_2 is the same). Here is a histogram of the distribution taken from 100,000 simulated samples of X_1:



It thus appears that the density is,

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 1-x & 1 < x \leq 2 \\ 0 & otherwise \end{cases}.$$

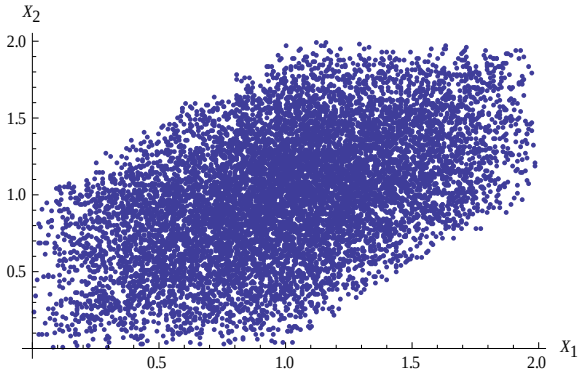
This density is obtained by assuming a general triangular shape (based on the simulation) with a maximal height equal to a:

$$f(x) = \begin{cases} ax & 0 \leq x \leq 1 \\ 2a - ax & 1 < x \leq 2 \\ 0 & otherwise \end{cases},$$

and then finding the value of a, that yields $\int_0^2 f(x)dx = 1$. It turns out that a=1. (Note that this distribution may be found without simulation by calculating the "convolution of two uniform densities" – but we do not discuss this in the current subject).

It is not surprising that the range of values which the random variable takes is [0,2]. Thinking about it, it should also make sense that values "near the edge" are less likely than values near 1.0. Why?

Now, here is a scatter plot of 10,000 generated samples of both random variables (remember that each has the triangular distribution):



We see for example that if $X_1 > 1.5$ then $X_2 > 0.5$. This has to be the case because $U_2 > 0.5$.

Thus, $P(X_1 > 1.5, X_2 < 0.5) = 0 \neq P(X_1 > 1.5)P(X_2 < 0.5) > 0$. And thus the random variables are not independent. This is not surprising because both random variables jointly depend on U_2. (Note though that there are cases where two random variables depend on a third random variable yet are still independent).

Independence of a random sample:

Our brief introduction to the concept of independence is primarily for discussing a data sample. Remember that we use two examples:

- Weights: $X_1, X_2, \dots, X_n$.
- Success indications: $I_1, I_2, \dots, I_n$.

Mathematically viewing each of the samples as independent random variables allows us to quite easily apply basic tools of probability for knowing the distribution of statistics generated from the samples (as shown below).

On the other hand, in cases where the samples are not independent, one has to be much more careful in the analysis (there are many forms of dependence and finding the "correct" form is not an easy matter). It is thus often a goal in "sampling data" to obtain a **random sample**. In practice this often requires randomizing the data, and choosing sampling data points at random. Note that in cases where

it is clear to us that the sample is not independent, we can still assume it is independent for purposes of statistical analysis, but this is at a cost of model inaccuracies which are often hard to quantify.

For purposes of illustration (of the basic concepts), we shall mostly be interested in the sample proportion statistic, $\hat{p}$. Observe that,

$$n\,\hat{p} = \sum_{i=1}^{n} I_i.$$

We will now see how this sum is distributed.

The Binomial distribution:

Let $Y = \sum_{i=1}^{n} I_i$. When $I_1, I_2, \ldots, I_n$ is an independent sequence with,

$$P(I_i = x) = \begin{cases} (1-p) & x = 0 \\ p & x = 1' \end{cases}$$

For some $p \in [0,1]$, we call Y a **Binomial(n,p)** random variable. It counts the **number of successes in a sequence of n independent trials**.

Example: A door to door salesperson visits exactly 10 houses every day, at each house he has a probability of 0.1 of succeeding in making a sale. The number of sales he makes a day is binomially distributed with parameters n=10, p=0.1

What is the PDF of a Binomial(n,p) random variable? Observe first that the range of values which such an RV can take is {0,1,…,n}. Let us start with the case of 0 and n.

$$p(0) = P(Y = 0) = P(all\ trials\ failed) = (1-p)^n.$$

Where did we use here the independence of the trials? Similarly,

$$p(n) = P(Y = n) = P(all\ trials\ were\ a\ success) = p^n.$$

Consider now, $p(1) = P(Y = 1) = P(one\ success\ and\ n-1\ failures)$. One now needs to partition this event based on which of the trials was the success (also indicating that the others were failures). There are n possibilities for this, and for each of them we have a probability of $p(1-p)^{n-1}$ of realizing the single success and n-1 failures thus,

$$p(1) = P(Y = 1) = P(one\ success\ and\ n-1\ failures) = np(1-p)^{n-1}.$$

Similarly

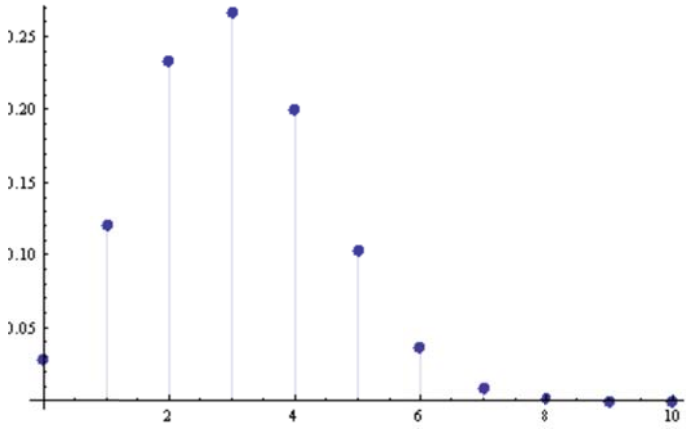
$$p(n-1) = P(Y = n-1) = P(one\ failure\ and\ n-1\ successes) = n(1-p)p^{n-1}.$$

Consider now an arbitrary $k \in \{0,1,..,n-1,n\}$. What is p(k)? Here we need to partition the event of {k successes and n-k successes} into $\binom{n}{k}$ events, each indicating which subset of the trials was a success.

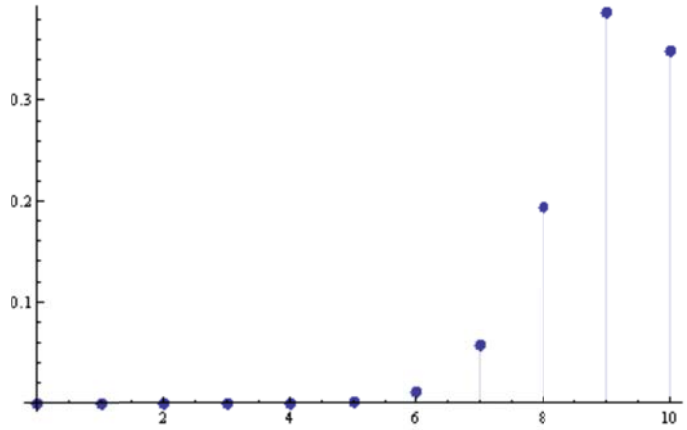
Then the probability of realizing this "sub-event" is $p^k(1-p)^{n-k}$. We have finally found the formula for the PDF of a Binomial(n,p) random variable:

$$p(k) = P(Y = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k \in \{0, \dots, n\} \\ 0 & otherwise \end{cases}.$$

Here is a plot of p(k) with n=10, p=0.3:



And here is the same with p=0.9:



What is the expectation (mean) and variance of Y? Are there simple formulas? It turns out that,

$$E[Y] = np \text{ and } Var(Y) = np(1-p).$$

Is the formula for the mean intuitive? Explain it.

The straightforward way to do these calculations is to attempt to simplify the expressions:

$$E[Y] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}, and \;\; Var(Y) = \sum_{k=0}^{n} (k - E[Y])^2 \binom{n}{k} p^k (1-p)^{n-k}.$$

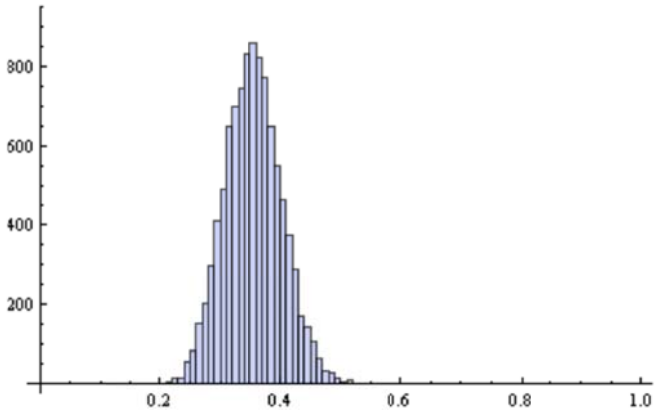
This is doable (but requires some basic algebraic and sum manipulation skills), e.g. Mathematica can do it for you:

```
In[14]:= Sum[k n!/((n - k)! k!) p^k (1 - p)^(n-k), {k, 0, n}]

Out[14]=
n p

In[19]:= Sum[(k - n p)^2 n!/((n - k)! k!) p^k (1 - p)^(n-k), {k, 0, n}] // Simplify

Out[19]=
-n (-1 + p) p
```

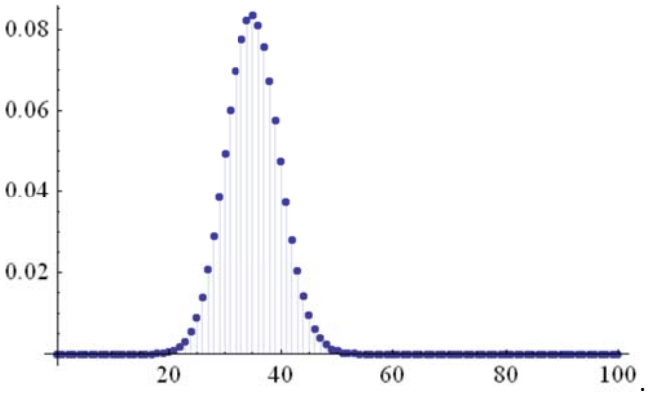
Alternatively we can immediately obtain the mean and variance of Y by using the fact that Y is a sum of independent random variables each with mean p and variance p(1-p).

Back to the distribution a statistic:

In the last section (last week) we illustrated the distribution of $\hat{p}$ by using simulation. For example for the case of n=100 samples and an actual success probability of p=0.35 we illustrated the following histogram for the proportion estimate:



Can we now obtain this theoretically? Here is the PDF of a Binomial(100,0.35) RV:



So now we know how our proportion estimate is distributed. The next Section (next week) will make use of that for statistical purposes (confidence intervals, hypothesis testing and planning of experiments). But first we go on and study the hypergeometric distribution as well as the Normal distribution and the CLT.

One way of breaking down the independent sampling assumption: The Hypergeometric distribution:

Note: this section may be skipped without loss of continuity of the main idea of the course.

We begin by an example of the binomial distribution:

Example: Consider a box with 10 balls, 3 of which are red. We are taking 4 balls out of the box, one by one at random, yet **returning** the ball taken every time. We count the number of red balls that were picked out. The number of red balls is distributed Binomial(4,3/10).

Assume now that in the example above, balls are **not returned** to the box, i.e. after taking a ball out of the box the probability of success "changes" since it depends on the number of successes so far. Denote by W the number of red balls. Let us find the PDF of W:

$$p(x) = P(W = x) = \frac{\binom{3}{x}\binom{7}{4-x}}{\binom{10}{4}}.$$

Why? Here we put in the denominator the total number of subsets of balls which we may end up taking out. In the numerator we count the number of subset of balls that satisfy the property of having x red balls.

For what range of values is the above PDF valid (outside of this range P(W=x)=0) ? We may have x=0. This will occur if all balls picked are not red. We may also have x=1, or x=2 or x=3. But it is impossible to have x=4 or higher because there are only 3 red balls.

Let us now go to the general case (with arbitrary numbers): Assume that we are randomly sampling n items, **without replacement** from a population of N+M items, were N items are of "type 1" and M items are of "type 2". We let W denote the random variable which counts **how many type 1 items we sampled**. Then we call the distribution of W a Hypergeomteric distribution, denoted **Hypergeometric(n,N,M)** and we have:

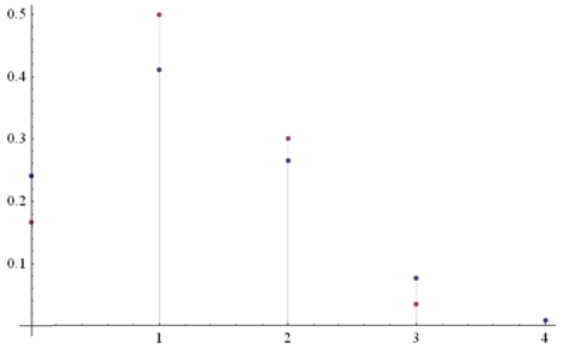
$$p(x) = P(W = x) = \begin{cases} \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{N+M}{n}} & Max(0, n-M) \le x \le Min(n, N) \\ 0 & otherwise \end{cases}.$$

Observe that if $n \le M$ and $n \le N$ then like the Binomial, x may take values 0,…,n, but otherwise, there may be more restrictive lower and/or upper limits on these values.

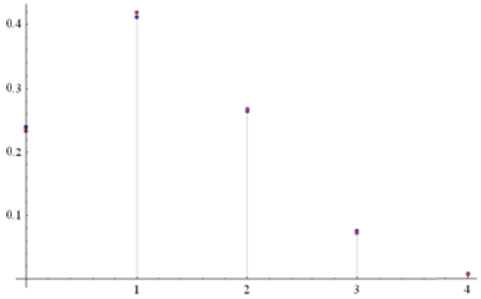
The mean of the Hypergeomteric distribution is: $E[W] = n\,\frac{N}{N+M}$. Compare this to the Binomial.

For small populations (small N+M) or when the proportion $\frac{N}{N+M}$ is either very close to 0 or 1, it is necessary to distinguish between the Hypergeometric distribution and a Binomial(n, $\frac{N}{N+M}$), yet when N+M is large, it practically does not matter if we are sampling with replacement or not.

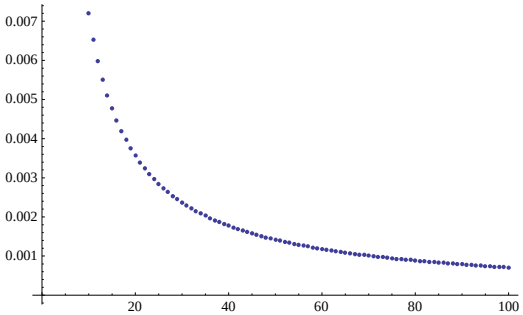
The graph below illustrates the PDFs of Binomial(4,3/10) vs. Hypergeometric(4,3,7) (guess which is which – note that the red PDF does not take a value for 4) :



This graph, is a Hypergeometric(4,30,70) plotted jointly with the same Binomial as above.

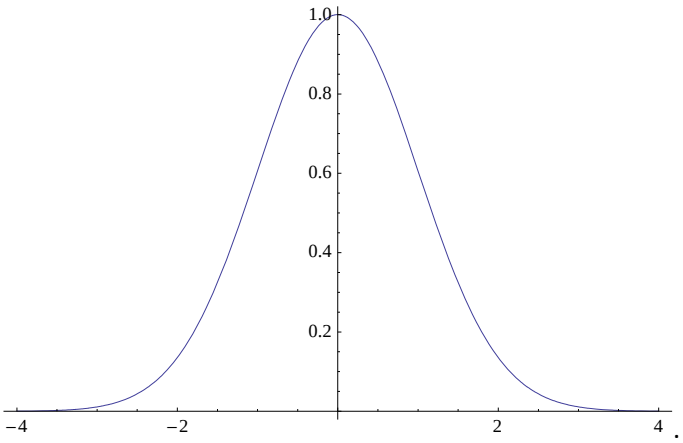


In fact, one can show (starred exercises) that the sequences of distributions of the form: Hypergeomteric(4,3a,7a) , for a=1,2,3…. converges to binomial(4,3/7) distribution. The graph below illustrates this numerically by plotting the maximal absolute distance between the distributions as a function of the variable a.



The Normal (Gauassian) Distribution:

The bell shaped curve $g(x) = e^{-\frac{x^2}{2}}$ is central to probability and statistics for reasons described below. The shape of the curve is as follows:



Here are some observations regarding g(x):

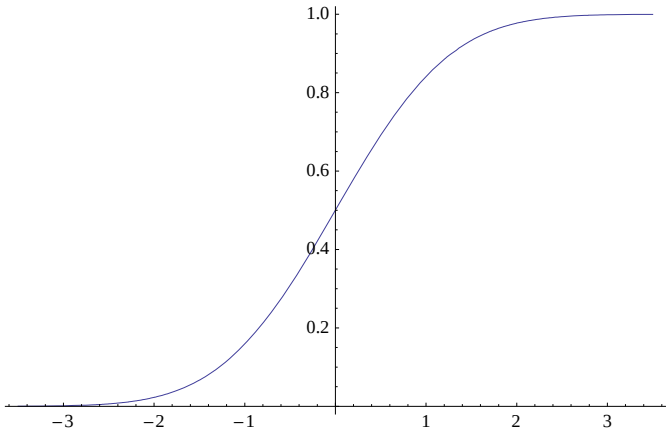
- It is positive for all x, yet for |x|>4, it is very close to 0.
- It is symmetric about 0.
- There is a maximum at x=0 attaining the height 1.
- Taking 2 derivatives of g(x), and solving g''(x)=0 we find that there are two inflection points at x=-1 and x=1.
- There is no analytic expression for $\int g(x)dx$, yet it can be shown that $\int_{-\infty}^{\infty} g(x)dx = \sqrt{2\pi}$.

We wish to have a continuous distribution with a density that appears as the bell curve, we thus need to modify g(x) and obtain a density function:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

We call this the density of the **standard normal** distribution. It is a density function because $\int_{-\infty}^{\infty} f(x)dx = 1$.

The CDF , F(x), associated with f(x) is typically denoted as $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$. Here is its graph:



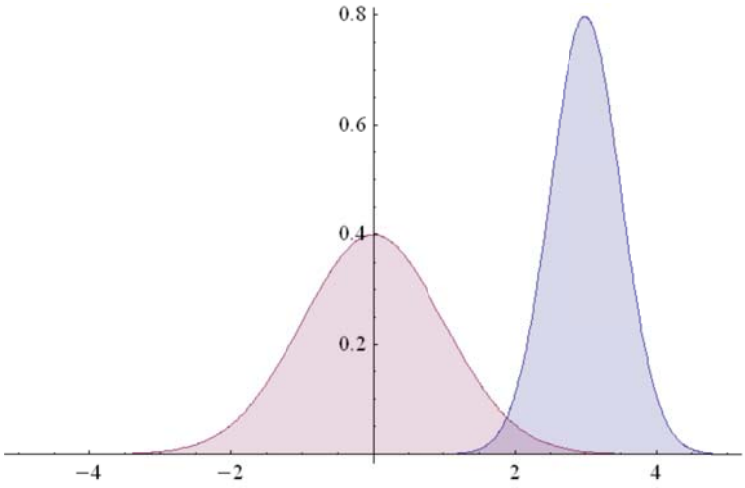
One needs to numerically evaluate $\Phi(x)$. The values are typically recorded in "normal distribution tables" (available in any statistics text-book or in the course web-site). Note that $\Phi(-x) = 1 - \Phi(x)$) so it is enough to record values for x>0. (What is $\Phi(0)$)?

We typically use Z denote a standard normal random variable. We have that

$$\text{E[Z]}=\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dt = 0 \text{ and Var(Z)} = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dt = 1.$$

So "standard normal" means that the mean is 0 and the variance is 1.

One can use a standard normal random variable to generate **arbitrary normal random variables** by doing the transformation: $X = \mu + \sigma Z$ (do you remember what such a transformation did for a uniform [0,1] random variable?). Here is a plot of the PDF of Z (in red) and the PDF of X=3+0.5Z (in blue):



Here μ is the mean of the RV X and σ^2 is the variance. Further, the CDF of X is $\Phi(\frac{x-\mu}{\sigma})$ and the PDF is

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Why? The following lemma shows how to get the distribution (CDF or PDF) of a linear transformation.

Lemma: Let V be a random variable with CDF F(x) and density f(x) and let a>0.
Then Y=b+aV has a CDF $F\left(\frac{x-b}{a}\right)$ and density $\frac{1}{a}f(\frac{x-b}{a})$.

Proof: $F_V(x) = P(V \leq x) = P(b + aV \leq b + a\,x) = P(Y \leq b + ax)$.

Thus, $F_Y(x) = P(Y \leq x) = F_V\left(\frac{x-b}{a}\right)$.

Then the density of Y is $f_Y(x) = \frac{d}{dx}F_Y(x) = \frac{d}{dx}F_V\left(\frac{x-b}{a}\right) = \frac{1}{a}f_V(\frac{x-b}{a})$.

Q.E.D.

Further, in tutorial 1, you have shown that E[Y]=b+aE[V] and Var(Y)=a^2Var(V). So in the case of an arbitrary **Normal random variable, the mean is $\boldsymbol{\mu}$ and the variance is $\boldsymbol{\sigma^2}$**. Note that we call σ, the **standard deviation**.

The central limit theorem: CLT:

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having the same distribution with mean μ and variance σ^2. Then:

I. $Y_n = \sum_{i=1}^{n} X_i$ is asymptotically normally distributed with mean $n\mu$ and variance nσ^2.
II. Alternatively, the sample mean $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ is asymptotically normally distributed with mean μ and variance $\left(\frac{\sigma}{\sqrt{n}}\right)^2$.
III. Alternatively, there is the case where $X_1, X_2, \ldots$ is a Bernoulli (binary) sequence with success probability p, denoted $I_1, I_2, \ldots$. Then $E[I_i] = p$ and $Var(I_i) = p(1-p)$ and $B_n = \sum_{i=1}^{n} X_i$is a Binomial(n,p) random variable. Then following (I), B_n is asymptotically normally distributed with mean np and variance $np(1-p)$.
IV. Alternatively, the sample proportion: $\hat{p}_n = \frac{\sum_{i=1}^{n} I_i}{n}$ is asymptotically normally distributed with mean p and variance $\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)^2$.

The CLT shows that normal random variables appear “naturally in life”. This is because many real life phenomena can be modeled as additions of independent random variables. In statistics, the CLT is crucial because it allows us to use the normal distribution to handle statistical statements regarding the sample mean and/or the sample proportion. Next week.

We end with a numeric illustration of the CLT.