Lecture 8.
The distribution of $\overline{X}$.
The Central Limit Theorem.

The distribution of any $\overline{X}$ is Normal…

… at least approximately.

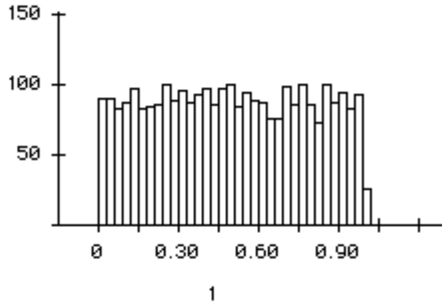
The Normal distribution is defined as the limiting distribution of a sum of similar things (same mean and variance) $\overline{X}$ is a sum of similar things so as $n \to \infty$ its distribution must get more Normal like.
Its just a question how big must n be for the approximation to be adequate?

If X itself is Normal $\overline{X}$ is Normal for any n – sums of Normals are Normal.
The more like the Normal the distribution of X is the smaller n has to be for practical use.
Eventually however all distributions succumb, it is surprising how fast this happens.

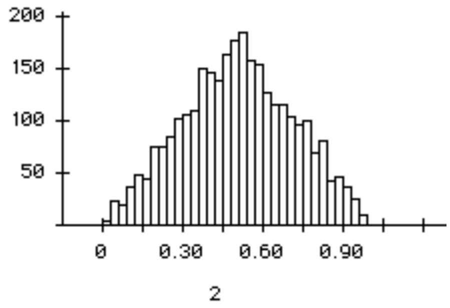
The proof of the CLT requires extra technology but we can show the effect numerically and give some pointers.

Consider averages of Uniforms.
3000 samples of 10 U(0,1) =Rand() were generated and averages of 1, 2, 5 and 10 were taken. The observed frequency distributions appear below:

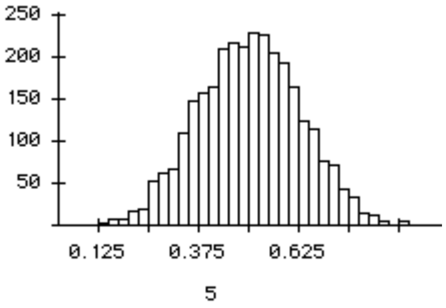
1 – Uniform



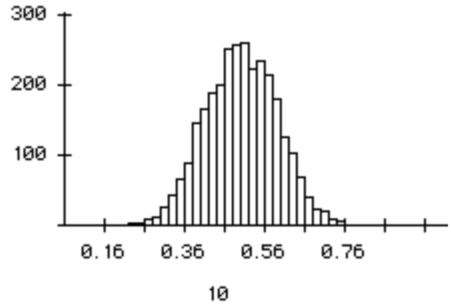
2 - triangular



Average of 5



Average of 10



As the number averaged increases the shape becomes more and more like the normal curve.

Why?

Because that is what the normal distribution is: it is *defined* as the limit distribution of sums of random variables.

The remarkable fact is that no matter what the underlying rv. is you always get a Normal.

But again it isn’t that strange.

There are ways other than the pdf to describe a random variable.
One way is by an infinite series of numbers called cumulants.

the first cumulant is the mean

the second the variance

the third is $E((x-\mu)^3)$

the fourth is $E((x-\mu)^4)-3\sigma^2$

They get more complicated after that.

A feature of cumulants (after the first one) is that when you take an average any random variables, its cumulant is smaller (in magnitude) than the cumulant of the original variable.

(so that averaging makes them tend to zero)

The normal is the unique distribution that has two non-zero cumulants the mean and the variance. The rest are zero.
That is why averages (rescaled to stop variance going to 0) tend to a normal distribution.

But this is just maths, what is remarkable is that it applies approximately in many practical situations.

We want to compare distributions of $\overline{X}$ for different X (and n) and compare them to the Normal. However the variance of $\overline{X}$ changes with n and the means might be different.

So to effectively compare these we need to standardise:

$$Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

Z is obtained from $\overline{X}$ by subtracting the mean of $\overline{X}$ and dividing by its standard deviation, this ensures that:

$$E(Z_n) = 0 \text{ and } V(Z_n) = 1.$$

So the Z_n arising from different distributions can be compared to N(0,1).

How do we compare two distributions – rather than use some uninterpretable summary measure we shall anticipate how we might use this result in practice.

P(-2<Z<2) this should be about 0.9545
X such that P(Z>X)=0.05 this should be about 1.64, X such that P(Z<X)=0.025 X should be about –1.64.
To examine more extreme properties we would need millions of simulations.

The method is to simulate from the distribution for n=10, and n=50, calculate Z for 1000 samples of each size and empirically establish

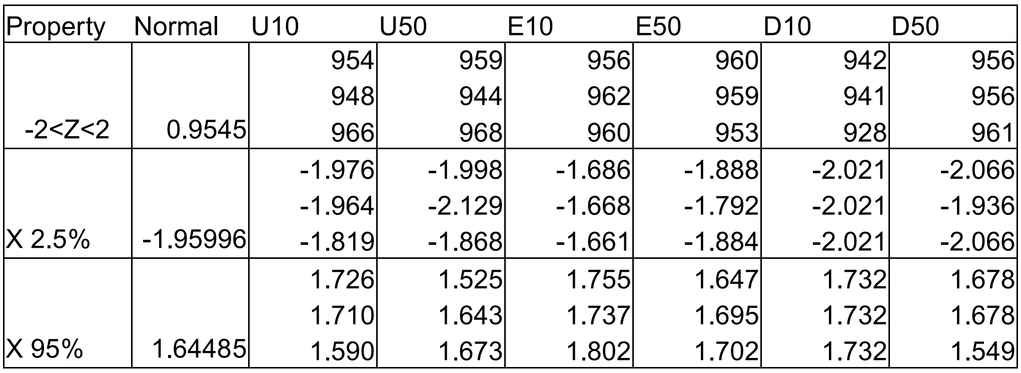
1. number/1000 that fall in the range (-2,2)
2. The 950th largest Z – 5% of Z are above that.
3. The 25th smallest Z 2.5% of Z are below that.

We shall use 3 distributions:

Uniform – nice and simple.
Car dealer – small discrete
Exponential – an un-Normal distribution.

VB programs (macros) were written behind an excel spreadsheet to implement the simulations.

By repeating the simulations 3 times we some idea of the accuracy of the results – we are doing only 1000 simulations not an infinity of them.



Property	Normal	U10	U50	E10	E50	D10	D50
-2<Z<2	0.9545	954	959	956	960	942	956
		948	944	962	959	941	956
		966	968	960	953	928	961
X 2.5%	-1.95996	-1.976	-1.998	-1.686	-1.888	-2.021	-2.066
		-1.964	-2.129	-1.668	-1.792	-2.021	-1.936
		-1.819	-1.868	-1.661	-1.884	-2.021	-2.066
X 95%	1.64485	1.726	1.525	1.755	1.647	1.732	1.678
		1.710	1.643	1.737	1.695	1.732	1.678
		1.590	1.673	1.802	1.702	1.732	1.549

Except for the Exponential 10, X 2.5% where the error is of the order of 15% (0.3 out of 2) the results are within a few percent of the Normal value.