$f_X(x)$ vs $f(x)$ ← more general notation, can refer to any PDF, does not specify which random variable the PDF is associated with.

↑ refers to the pdf of the random variable X. It describes the likelihood of X taking on the value x.

Week 4 Notes Part 2 CLT

Code

corrections to notes made in red. Emailed him + he admitted there are errors.

Author: Brendan Gongol

Last update: 03 January, 2023

1 Overview

In this part of the notes we discuss one of the fundamental facts of probability and statistics: the Central Limit Theorem. We will learn about the theorem and its implications and run a simulation in R as a practical illustration.

2 Central Limit Theorem

- in context of the CLT, a normal distribution is denoted as: $N(\mu, \sigma^2)$

The Central Limit Theorem (CLT) is a very important theorem in statistics that states that the distribution of a sum of n independent identically distributed (i.i.d.) random variables with finite variance, converges to a normal distribution as n increases.

↳ spread of values is limited.

In more formal way, if we have i.i.d. random variables $X_1, \ldots X_n$ (with arbitrary probability density ~~$fX(x)$!!~~ and define new random variable

(1) $f_x(x)$

i.i.d.'s

$$\tilde{X} \quad \cancel{X} = X_1 + \ldots + X_n$$

Here $\tilde{X}$ is the sum of the random variables X_n

then as long as $f_X(x)$ has finite variance, the distribution of $\tilde{X}$, $f_{\tilde{X}}(x)$, has the limit

(2) since (1) is the sum: I think equation 2 should actually be $f_{\tilde{X}}(x) \xrightarrow[n \to \infty]{} N(n\mu, n\sigma^2)$

$$f_{\tilde{X}}(x) \xrightarrow[n\to\infty]{} N(\mu, \cancel{n}\sigma^2)$$

incorrect. ~~This assumption~~ Since the $f_X(x)$ represents the original population distribution, use this function implies that the original distribution itself becomes normal. this is not what the CLT states.

where N(m,s2) denotes normal distribution with mean m and variance s^2 and the actual values of the parameters μ and σ^2 in the asymptotic distribution are the mean and variance of the distribution $f_{\tilde{X}}(x)$ of the random variables we are summing up (remember that those are i.i.d., so that they all have the same distribution $f(x)$). From the material presented in Part 1 of this week's notes, you should be able to understand/derive why the variance of the distribution (2) is equal to $\cancel{n}\sigma^2$, so this is not particularly new.

$f_x(x)$

What is new and important, and we will stress it once again, is that the limit of the distribution in (2) has a Gaussian shape, while the variable X does not have to follow a normal distribution. Mean and variance are defined for an arbitrary distribution, and as long as the variance is finite, the statement (2) holds: the sum of a large number of such arbitrary distributions is (almost) a normal. Note that some distributions have tails that are "too long" and their variance diverges; for instance the distribution $f(x) = 1/x^2$ has infinite variance. For those long-tailed distributions the CLT in its classical form does not hold.

We will omit the proof of the CLT, although it is not particularly hard or long. One very quick proof involves characteristic functions, and you can look it up if you are interested, but it is beyond the scope of this course.

Instead we will focus on the meaning and importance of the CLT, and, as it becomes our tradition, run a few numerical simulations in order to observe CLT in action.

While normal distribution is (fortunately) "convenient" in a sense that it allows for many analytical results to be obtained, both for normal itself and for related distributions (such as chi-square, t-distribution, etc. – we will discuss those in more detail soon), this convenience is not the main (or at least not the only) reason for the importance of this distribution. The reason is that many processes are (or can be thought of as) a sum of large number of "elementary" processes, so that according to the CLT their distributions should be close to normal.

One particular example we will consider here is the sample mean. As we have seen in Part 1, sample mean is a random variable defined as

(3)

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

where i.i.d. variables $X_1, \ldots, X_n$ represent the process of drawing our sample: n independent drawings from some underlying distribution. We already know from Part 1 that the variance of $\bar{X}$ is σ^2/n (the variance of the sum of n random variables in (1) is $n\sigma^2$, but we have additional factor $1/n$ in (3), which rescales the variance). This is an important result as it tells us something about the "typical spread" of the sample mean estimates around the true population mean. However, this is insufficient for precise quantification of the probability of error of any given size: if we only know the variance σ^2 for some (unknown) probability density distribution $f(x)$, we still cannot tell what percentage of the total mass of $f(x)$ is located, for instance, one σ away from the center, vs 1.375σ away, vs 10σ away. These numbers will be different for distributions with different functional forms (or "shapes").

Reference where a sequence of estimates or distributions approaches a normal distribution as the sample size grows infinitely large

However, with the CLT we now know that the random variable X in (1) has asymptotically normal distribution, and thus the sample mean X in (3) also follows (approximately) normal distribution. Using specific functional form of the normal distribution,

(4)

if x is close to μ → the exponent is near zero making $N(x|\mu,\sigma^2)$ larger which is a higher ~~probability~~ density probability.

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp[\frac{-(x-\mu)^2}{2\sigma^2}]$$

we can calculate precisely the probability density of any given value x, or the probability to observe a value $x > x_o$ above any specific threshold x_o. The latter probability actually brings us very close to understanding the calculation behind the t-test.

Let us postpone the discussion of the "full" t-test for just a little longer and instead consider the following (somewhat artificial) scenario. We have a random variable X with variance σ^2, which we know exactly (i.e. we happen to know the variance of the underlying distribution). However, we do not know the mean, so we follow the typical experimental procedure: we draw a sample of size n, calculate the sample mean, $\bar{X}$, and use it as an approximation to the underlying population mean μ. Let us ask the following question: if the mean of the underlying population is zero, how likely it is to observe sample mean of $\bar{X}$ (the value we just observed in our sample) or even larger? You should have recognized this question as precisely the one that t-test is supposed to answer: we postulate the null hypothesis ($\mu = 0$), and we want to calculate the probability to observe a sample mean at least as extreme as the one we actually see.

Now we know that for a random variable with $\mu = 0$ (this is our null) and known variance σ^2, the sample mean calculated from a sample of size n must be distributed as

(5)

$$N(x|0, \sigma^2/n) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} exp[\frac{-nx^2}{2\sigma^2}]$$

(remember that the distribution of the observed sample mean values is centered at true population mean, which we postulate to be 0 under our null hypothesis, the variance of the sample mean is σ^2/n, and according to CLT the distribution itself is (nearly) normal when n is sufficiently large). Hence the probability to observe, by random chance, a value of the sample mean equal to or greater than our observation $\bar{X}$ (which probability is exactly the p-value of the one-sided test!) is

where N is given by (5). This is not the t-test yet, but this example is very instructional as it shows how the things work.

3 CLT Illustration

Let us now run a quick example in R. For this exercise we will consider a Bernoulli process (coin toss) X with success probability 0.5. The random variable X is discreet and can take only two possible values, 0 and 1, with equal probabilities 0.5. Nothing can be probably as far from a normal distribution as the distribution of X. But what happens if we take a sample of size n (toss our $0/1$ coin n times) and sum the values observed in each of these n Bernoulli trials (which is equivalent to counting the number of 1's in the sample)? So what we want to study is the random variable

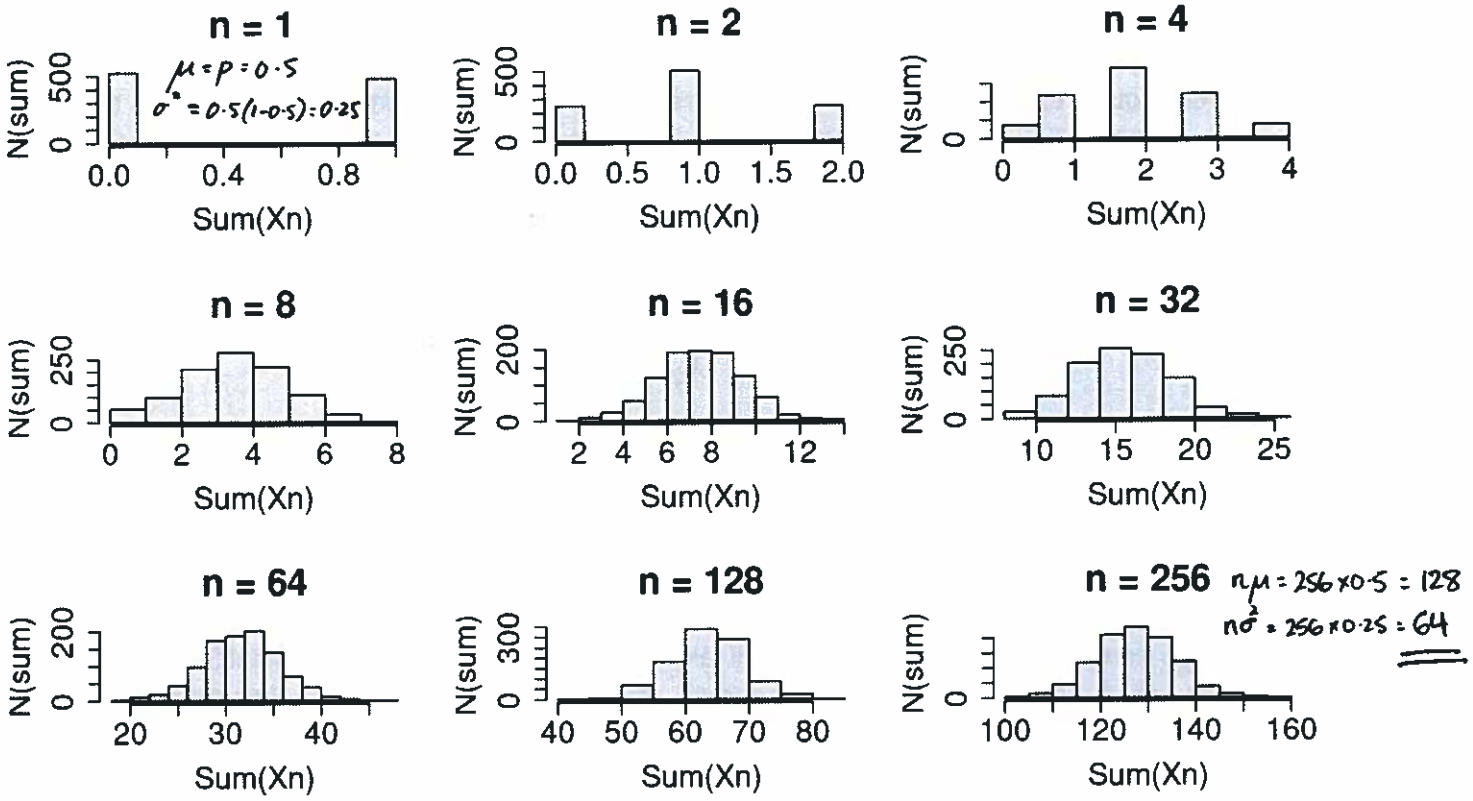
$$S = X_1 + X_2 + \ldots + X_n$$

where X_i are i.i.d. Bernoulli random variables.

We could run a complete simulation: build a sample by drawing n times from the Bernoulli distribution (which is simply choosing n values of 0 or 1 with probability 0.5 every time), add up the resulting n values in order to obtain the sum S for the given sample, then repeat this process multiple times in order to build the distribution of S numerically. However, this complete program is left for the exercise – you will have to do this in one of the homework problems. Here we will take a shortcut and use the exact result for the sum of n Bernoulli trials. As it was shown in previous weeks, the distribution of this sum (or equivalently the distribution of the observed number of 1's out of n 0/1 coin tosses) is the binomial distribution B(n,p) (p is equal to 0.5 in our case). Take a look at Week 2 Notes, Part 1 (Combinatorics, Binomial Distribution) if you need to refresh your memory. Since binomial distribution $B(n, p)$ already represents the distribution of the sum of n Bernoulli variables, we will just draw this distribution directly:

Hide

```
smpl.sizes <- 2^(0:8)              # different sample sizes n we want to try
n.sim <- 1000                      # we want to draw n.sim samples for each sample size n
old.par <- par(mfrow=c(3,3),ps=20) # draw 9 plots in 3x3 lattice
for ( i.smpl in smpl.sizes ) {     # for each sample size
 # if we drew n.sim examples of Bernoulli samples of size i.smpl
 # and every time summed up i.smpl values in each sample, the
 # resulting n.sim sums (that represent numbers of heads observed
 # after i.smpl coin tosses) would be distributed as binomial,
 # so instead of resampling we just use that directly:
 x.tmp <- rbinom(n.sim,i.smpl,0.5) # distribution of n.sim sums
 # plot the distribution (histogram) of n.sim sums S
 plot(hist(x.tmp,plot=F),
 main=paste("n =",i.smpl),
 xlab="Sum(Xn)",ylab="N(sum)")
 }
```



mean of a Bernoulli : $E[X] = p$
Variance : $Var(X) = p(1-p)$

$\xrightarrow{n \to \infty}$ $N(n\mu, n\sigma^2)$

Hide

```
par(old.par)
```

The resulting plot is shown below. At n=1 we are "summing up" a single Bernoulli variable (and thus the value of the sum can be either 0 or 1). The distribution such "sum" is just the distribution of the variable X itself (Bernoulli). At n=2 we perform 2 coin tosses, so the possible values of the sum (the number of heads) is 0, 1, or 2, each with its own probability, etc. When we start increasing the sample size, the distribution of the sum approaches normal distribution very quickly (of course we could run Shapiro test or at least draw a qq-plot in order to prove this more rigorously, but here we will just examine the distributions visually). Note that already at n=8 the distribution of the sum of 8 Bernoulli variables look pretty close to normal!

- The original random variable X has its own distribution $f_X(x)$ with a mean $= \mu$ and variance $= \sigma^2$ these values are fixed. This distribution may not be normal at all. $f_X(x)$ is the population distribution - the entire, true distribution of the random variable X. - It tells you the ~~entire~~ probability (or probability density) of every possible value that X can take across the entire population.
- The CLT doesnt say that X itself becomes normal as n samples ↑, Instead it says that if you take many independent samples of X and form either ther sum or their average, the distribution of that sum or average will tend to be normal as the ~~distn~~ number of samples increases.

→ for the **sum**: if you add up independent observations (X_n)

$$S = X_1 + X_2 + \dots + X_n$$

sum $\text{Var}(S)\uparrow$ (Var with more n) the sum will have a variance of $n\sigma^2$ (b/c variances of independent random variables add up)

the mean $= n\mu$ (each independent variable X_i has a mean of μ, so summing n independent samples simply scales the mean by n).

→ for the **Sample mean**:

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

$\text{Var}(\bar{X})\downarrow$ (with more n) when you calc the sample mean its variance becomes $\frac{\sigma^2}{n}$, thus each individual X might not be normal, the sample mean will be approx normally distributed around the mean μ.