Math 321 – Chapter 4 – The Central Limit Theorem
(draft version 2019/06/20-13:23:22)

Contents

1 Independence for random variables

Independence. Let X and Y be real-valued random variables. We say X and Y are independent is the events $\{a < X < b\}$ and $\{c < Y < d\}$ are independent for all a, b, c, d. Intuitively, this means that knowing the value of one of the variables doesn't affect the probabilities for the other variable.

Example: Let X be the outcome for a roll of a 6-sided die, and Y be the outcome for a roll of a 20-sided die. It should be clear that they are independent random variables. The probability that the first die is 5 and the second die is 18 can be calculated as

$$P(X = 5 \text{ and } Y = 18) = P(X = 5) \cdot P(Y = 18) = \frac{1}{6} \cdot \frac{1}{20}.$$

Identically distributed. Random variables that have the same probability distribution function are called *identically distributed*.

Example: Consider random variables X ={the numerical outcome of a fair, green 6-sided die} and Y =the numerical outcome of a fair, red 6-sided die}. These two random variables

have the same possible values $\{1, 2, 3, 4, 5, 6\}$ and the same probabilities for each value; they are identically distributed.

Example: Consider a coin that is flipped a number of times. The outcome for each coin flip is identically distributed.

The two above examples are also independent. Random variable that are independent and identically distributed are called... *independent and identically distributed* or for short, iid.

2 Law of large numbers

Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (iid) random variables with mean μ. That is, $E(X_i) = \mu$ for $i = 1, 2, \ldots, n$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \text{ converges in probability to } \mu \text{ as } n \to \infty.$$

Another way to phrase this is that for all positive numbers $\epsilon > 0$,

$$P\left(\left|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right| > \epsilon\right) \longrightarrow 0 \text{ as } n \to \infty.$$

How to interpret this: Remember that the actual value of $\frac{1}{n}\sum_{i=1}^{n} X_i$ is still random, so it could vary quite a bit if we were to sample different X_i's. But as the sample size n gets very large, the chance of the sample average deviating too far from the mean μ is very low.

We have already seen this in class in several cases.

Example: Let's consider n coin flips with $P(H) = p$. Let X_i be the outcome of the i^{th} coin flip, i.e. $X_i = 1$ if the i^{th} coin flip is heads and $X_i = 0$ if it is tails. Then the X_i's are independent, and they are all Bernoulli random variables with probability of success p. So they are iid (independent and identically distributed). Thus if we take the average value of n coin flips: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, as we increase n, this average will get very close to p (which is the μ in this case). We say that $\overline{X}_n$ converges to p in probability as $n \to \infty$. Recall that you observed this in one of the R activity assignments.

3 Central limit theorem

Let $X_1, X_2, \ldots, X_n$ be iid with mean μ and variance σ^2. Note that we are not assuming anything about their distribution other than the mean (expected value) and variance. They could be drawn from a normal distribution, exponential distribution, binomial, Poisson, or any other distribution that has a finite mean and variance.

Then when the sample size n is large, the sample mean will be approximately normally distributed with mean μ and variance σ^2/n.

Remember that $X_1, X_2, \ldots, X_n$ is a random sample. So for a fixed sample size n, we can draw a variety of different random samples. Each sample will have a different sample mean $\overline{X}_n$. So by drawing many many different samples, we will have many many different sample means. If we create a histogram for our dataset of sample means, it will look very much like the normal distribution with mean μ and variance σ^2/n.

This is called the central limit theorem (CLT).

This is a remarkable fact!!!! This is one of the reasons why the normal distribution is so important to all of statistics.

If the $X_1, X_2, \ldots, X_n$ are iid and $X_i \sim N(\mu, \sigma^2)$, in other words if they come from a normal distribution exactly, then $\overline{X}_n$ is exactly normally distributed.

Summary of CLT:

$$X_i \sim N(\mu, \sigma^2) \text{ for } i = 1, 2, \ldots, n \text{ then } \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

We also get that the sum of the X_i's are approximately normally distributed:

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right)$$

If the X_i are not normally distributed, this still holds approximately when the sample size is sufficiently large.

How large should the sample size be so that the central limit theorem is a good approximation? In most cases $n \geq 30$ is a general rule of thumb. Even with smaller sample sizes, the approximation may not be too bad.

If the underlying distribution of the X_i is extremely skewed with a large probability for very far away outliers, then n in the 100s, 1000s or larger may be necessary. For example, if the X_i are exponentially distributed (a distribution that has a higher probability of large outliers), then a sample size of 100 or greater will be required for a decent approximation by the CLT.

3.1 Using the CLT to calculated probabilities for $\overline{X}_n$

Since we know that $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, we can now calculate the probability of the sample mean being in a particular range of values. It is now just a normal random variable!

Example: Human height is approximately normally distributed with mean 175 cm and standard deviation 7 cm. If 30 people are selected at random, approximate the probability that the mean height of the sample is greater than 182 cm.
Solution: $\overline{X}_{30} \sim N(\text{mean} = 175, \text{std. dev.} = 7/\sqrt{30})$.
Thus $P(\overline{X}_{30} \geq 182)$ =1-`pnorm(182,mean=175,sd=7/sqrt(30))`$\approx 2(10)^{-8}$. This is a minuscule probability! Note that the cut-off of 182 is only 1 standard deviation above the mean, so according to the 68-95-99.7 rule, there is a 17% chance of any individual data

point being above 182, but the mean of a sample of 30 data points will rarely be that far away from 175.

3.2 Justification for using CLT and the normal distribution

In scientific contexts, data is most often gathered from quantitative measurements. These measurements will generally have a minimum and maximum possible value set by that physical limitations inherent of the system of interest. For example, if we are measuring a length of an object on planet earth, then there is an absolute limit on the measurement that it is between zero and the size of the planet earth.

Generally the underlying variability of a measurement will be due to many individual, and potentially interacting, sources. These can include error in reading the measuring device (e.g. incorrect rounding or just copying the result incorrectly), variability or errors in the manufacture of the measuring devices itself (e.g. slight variability in the precise locations of markings on rulers), or variability in the chain pf physical processes involved in the measurement process itself (e.g. electrical interference in a circuit or signal loss/noise). Assume these sources of error are additive and each is sufficiently small. Let e_j be the error from error source j for $j = 1, ..., k$ each with mean zero and variance σ^2. Note that we make no assumptions except zero mean and finite variance. Then by the central limit theorem

$$\sum_{j=1}^{k} e_j \sim N\left(0, k\sigma^2\right)$$

Let μ be the "true" value for the physical property being measured (e.g. the exact length). Then our recorded measurement will be a random variable

$$X = \mu + e_1 + e_2 + \cdots + e_k = \mu + \sum_{j=1}^{k} e_j.$$

The expected value of X is

$$\mathrm{E}(X) = \mu + \mathrm{E}\left(\sum_{j=1}^{k} e_j\right) = \mu + \sum_{j=1}^{k} \mathrm{E}\left(e_j\right) = \mu$$

and the variance of X is

$$\mathrm{Var}(X) = \mathrm{Var}(\mu) + \mathrm{Var}\left(\sum_{j=1}^{k} e_j\right) = 0 + \sum_{j=1}^{k} \mathrm{Var}\left(e_j\right) = \sum_{j=1}^{k} \sigma^2 = k\sigma^2$$

and X is approximately normally distributed

$$X = \mu + \sum_{j=1}^{k} e_j \sim N\left(\mu, k\sigma^2\right).$$

Even if the sources of error do not have the same variance and are not all mean zero, as long as the sum of all sources of error is zero on average then the CLT still applies.

If the sources of error are not additive though, e.g. are multiplicative, then justifying the use of the CLT become more difficult. E.g. if $X = \mu \cdot (1 + e_1) \cdot (1 + e_2) \cdots (1 + e_k)$ then the CLT does not directly apply. In this case a logarithmic transform of the data may be useful

$$\ln X = \ln \mu + \sum_{j=1}^{k} \ln(1 + e_j)$$

If $\sum_{j=1}^{k} \ln(1 + e_j)$ can be justified to have mean zero and finite variance, then $\ln X$ may be approximately normally distributed.

4 Propagation of error

When an experimental or measurement apparatus is constructed, errors and variability may be introduced at many places. If these errors are not strictly additive, then it can become difficult to know how their variability will impact the variability of the total system or measurement. Propagation of error methodology gives us a way to estimate these effects.

4.1 Partial derivatives and multivariable functions

In order to make propagation of error calculations, we must introduce the concept of partial derivatives and multivariable. Let f be a function of x and y, $f(x, y)$, then the partial derivative of f with respect to x is denoted $\frac{\partial f}{\partial x}$ and it found using all standard differentiation rules and treating y as if it were a constant.

Example: $f(x, y) = x^2 y^3$, then $\frac{\partial f}{\partial x} = 2xy^3$ and $\frac{\partial f}{\partial y} = 3x^2 y^2$.

Example: $f(x, y) = x^2 + y^3$, then $\frac{\partial f}{\partial x} = 2x + 0$ and $\frac{\partial f}{\partial y} = 0 + 3y^2$.

4.2 Variance of a multivariable function

If $Z = f(X, Y)$ and we wish to estimate the variance of Z in terms of the variance of X and Y and we assume that X and Y are independent, then as long as we assume the variance of X and Y is small, then

$$\sigma_Z^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_X^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_Y^2.$$

If X and Y is not independent then the calculation is much more difficult.

Example: Resistance R can be calculated by measuring voltage V and current I and then using the equation $R = \frac{V}{I}$. If we know that are measurement of voltage has variance σ_V^2 and our measurement of current has variance σ_I^2 and that both of these variances are small (i.e. that our measurements are reasonably accurate), the the variance of our resistance calculation is given approximately by

$$\sigma_R^2 = \left(\frac{\partial R}{\partial V}\right)^2 \sigma_R^2 + \left(\frac{\partial R}{\partial I}\right)^2 \sigma_I^2 = \frac{1}{I^2}\sigma_R^2 + \frac{V^2}{I^4}\sigma_I^2.$$

So if we know that voltage is 9 ± 0.25 volts and current is 2 ± 0.13 milliamps, then we can estimate our error in resistance by

$$\sigma_R^2 = \frac{1}{2^2}(0.25)^2 + \frac{9^2}{2^4}(0.13)^2 = 0.35785$$

so that our resistance measurement with error bounds is

$$V/I \pm \sigma_R = 9/2 \pm \sqrt{0.35785} = 4.5 \pm 0.5982.$$

See the Wikipedia article Propagation of uncertainty for a great list of propagation of error formulae for a variety of multivariable functions of two variables.

5 Sampling Distributions

In discussing the central limit theorem, we are really talking about *sampling distributions*. A sampling distribution is the probability distribution of a sample statistic. A sample statistic is something that you calculate from a sample. For example the sample mean and sample variance are both two sample statistics. There are other sample statistics such as the sample median, sample skewness, and even many others.

If the X_i are normally distributed with mean μ and variance σ^2, then the sample mean is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$ as stated by the CLT. As long as n is large enough, we have that, approximately, $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ by the CLT. This is true regardless of the underlying distribution of the X_i.

Here we will look at another example of a sampling distribution.

Consider a discrete finite population: $\{1, 1, 1, 1, 1, 2, 2, 2, 3, 3\}$. We will sample from this population with replacement. Let X_i be the i^{th} individual selected. We'll just select a sample of size $n = 2$.

All possible samples of size 2, the probability of each, and the sample mean and sample variance for each sample are given in the table below.

Then we look for each possible value of $\overline{x}$ and add up the probabilities to create a probability mass function $f_{\overline{X}}(\overline{x})$ for $\overline{X}$. This is the sampling distribution for $\overline{X}$.

Similarly we can create a probability mass function $f_{S^2}(s^2)$ for the sample variance $\overline{S}^2$. This is the sampling distribution for S^2. Note that we are treating the sample variance

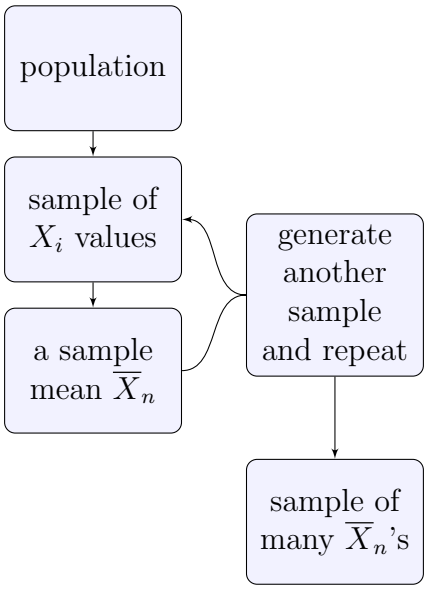

Figure 1: Gathering many sample means flow chart diagram.

	X_1	X_2	$P(X_1, X_2)$	$\overline{x}$	s^2
1	1.00	1.00	0.25	1.00	0.00
2	2.00	1.00	0.15	1.50	0.50
3	3.00	1.00	0.10	2.00	2.00
4	1.00	2.00	0.15	1.50	0.50
5	2.00	2.00	0.09	2.00	0.00
6	3.00	2.00	0.06	2.50	0.50
7	1.00	3.00	0.10	2.00	2.00
8	2.00	3.00	0.06	2.50	0.50
9	3.00	3.00	0.04	3.00	0.00

Table 1: All possible sample means for sample of size 2 from the given discrete probability distribution. The sampling distribution for $\overline{X}_2$ can be constructed from this table.

S^2 as a random variable now, because it's value depends on the particular sample that is gathered.

Here is an R code that you can use to generate such sampling distributions:

```
################################
## EDIT the parameters below
################################

# list possible x values (population)
# it's best to keep this between 2 and 5 total values
x=c(1,2,3)

# weights for each x
# (number of tickets in the box)
w=c(5,3,2)

# sample size,
```

$\overline{x}$	1.00	1.50	2.00	2.50	3.00
$f_{\overline{X}}(\overline{x})$	0.25	0.30	0.29	0.12	0.04

Table 2: Sampling distribution for the sample mean.

s^2	0.00	0.50	2.00
$f_{S^2}(s^2)$	0.38	0.42	0.20

Table 3: Sampling distribution for the sample variance.

```
# use n=2 up to 6
# beyond that it may
# take up too much computer memory
n=2

################################
## DO NOT edit below here (without risk!)
################################

p=w/sum(w) # turn weights into probabilities
mu=sum(x*p) # population mean
sigsq=sum(x^2*p)-mu^2 # population variance

# now we create the list of all samples of size n,
# calculate sample statistics (mean and varaince of each sample)
S=expand.grid(replicate(n,x,simplify=FALSE))
pr=expand.grid(replicate(n,p,simplify=FALSE))

prob=cbind(0*1:length(S[,1])+1)

# calculate all sample means and their probabilities
xbar=0*prob
for (j in 1:n){
    prob=prob*pr[,j]
    xbar=xbar+S[,j]
}

# append table with probabilities and sample means
S$prob=prob
S$xbar=xbar/n

# calculate all sample varainces
var=0*prob
for (j in 1:n){
    var=var+(S[,j]-S$xbar)^2
}

# append table with probabilities and sample variances
S$var=var/(n-1)

# construct sampling distributions
# for sample mean and sample variance
xbar_vals=as.numeric(names(table(S$xbar)))
var_vals=as.numeric(names(table(S$var)))

xbar_probs=0*1:length(xbar_vals)
```

```
for (k in 1:length(xbar_vals)){
    xbar_probs[k]=sum(S$prob[S$xbar==xbar_vals[k]])
}

var_probs=0*1:length(var_vals)
for (k in 1:length(var_vals)){
    var_probs[k]=sum(S$prob[S$var==var_vals[k]])
}

# construct sampling distributions
xbar_samp_distr=rbind(xbar_vals, xbar_probs)
var_samp_distr=rbind(var_vals, var_probs)

xbar_mean=sum(xbar_samp_distr[1,]*xbar_samp_distr[2,])
xbar_var=sum(xbar_samp_distr[1,]^2*xbar_samp_distr[2,])-xbar_mean^2

var_mean=sum(var_samp_distr[1,]*var_samp_distr[2,])
var_var=sum(var_samp_distr[1,]^2*var_samp_distr[2,])-var_mean^2

# plot resulting sampling distributions
par(mfrow=c(2,1))
barplot(xbar_samp_distr[2,],
    names.arg=as.character(xbar_samp_distr[1,]),
    main="sample mean sampling distribution")
barplot(var_samp_distr[2,],
    names.arg=as.character(var_samp_distr[1,]),
    main="sample variance sampling distribution")

print(xbar_samp_distr)
print(var_samp_distr)
```

6 Summary

R commands:

```
pnorm(x,mean=mu,sd=s/sqrt(n))
```

Notation and formulas:

LLN: Sample mean converges to μ (in probability) as sample size gets large:
$\overline{x}_n \to \mu$ as $n \to \infty$. (when n is large $\overline{x}$ is unlikely to vary to far from μ)

LLN: Sample proportion converges to p (in probability) as sample size gets large:
$\hat{p}_n \to p$ as $n \to \infty$. (when n is large $\hat{p}$ is unlikely to vary to far from p)

CLT: Sample mean is always approximately normally distributed as long as sample size is sufficiently large:
$\overline{X} \sim N(\mu, \sigma^2/n)$

CLT: Sample proportion is always approximately normally distributed as long as sample size is sufficiently large:
$\hat{p} \sim N(p, p(1-p)/n)$

Normal approximation to binomial: $X \sim Bin(n, p)$ then $X \sim N(np, np(1-p))$ as long as n is large and $np > 5$, $n(1-p) > 5$.

Normal approximation to Poisson: $X \sim Pois(\lambda)$ then $X \sim N(\lambda, \lambda)$ as long as λ is large.

These normal approximations are consequences of the CLT and are more inaccurate in the tails of the distributions.

Sampling distribution is the probability distribution of any quantity calculated from sample data, e.g. $\overline{x}$, s, $\tilde{x}$, etc. each have a sampling distribution.