Estimating The Mean Of Exponential and

Binomial Distributions

These two distributions are examples where there is only one parameter

to determine, the mean, since the variance is fixed, once the mean is.

The normal distribution does not have such a constraint.

Exponential Distribution

Suppose you are measuring the decay times of a radioactive material.

Physics tells us that the time interval between successive emission of

radiation is a random variable with exponential distribution. That is,

the time to the next emission T is such that

P[T>t]=e-tμ

where μ is the mean of this variable.

We measure these times to emission n times,

obtaining a sample T1,T2,…Tn

that is very reasonably made up of independent identically distributed

variables, thanks to a law of nature.

To estimate μ, we need to get a handle on

the distribution of the sample mean, or, equivalently, of the sum of

these variables. It turns out that the distribution of the sum of

exponential random variables is a χ2

distribution, whose values are available in any spreadsheet, as well as

in tables on line and in any statistics textbook, since the same

distribution is used to estimate the variance of a normal sample!

Remark 1. χn2

is the distribution of the sum of the squares of n

independent standard normal variables. Note that for each such

variable, say Xk2

has true mean 1 (since E⁡Xk=0,

E⁡Xk2

is the variance of Xk,

hence equal to 1). Also, the variance is E⁡[Xk2-1]2=E⁡Xk4-1,

and calculations show that that's equal to 3-1=2.

Incidentally, the applicability of this to the estimation of the

variance follows from the observation that, if the true mean is known,

(Xk-μ)2σ2

is the square of a standard normal variable.

Precisely, if we consider the variables 2μTk,

they happen to be distributed according to the χ22

distribution, and hence their sum is distributed according to χ2n2

(the proof requires a bit of calculus, and we'll just take the statement

for granted). It also turns out that the true mean of a χ2n2

variable is 2n, and

its variance is 4n

(since we are adding independent variables, both the mean and the

variance equal the sum of the individual means and variances), thanks to

the remark above. If n is large (for

example, n≥50,

so that 2n≥100⁡),

we can apply the Central Limit Theorem (and, indeed, you can check in a

table that χn2

looks more and more like a normal distribution), and conclude that

2μ∑kTk-2n4n=1n∑kTk-12μ

is approximately distributed like a standard normal variable, so that,

for example,

P[|T‾-12μ|<zα/2]=α

Solving for μ, produces a confidence

interval of level α, and, as you can see,

we end up using our familiar tables without having to go to a Student

distribution, whose applicability is questionable, since it would not

only require that the mean be approximately normal, but also that the

sample variance be approximately χ2,

a much less likely situation

Note 2. The exact estimate, without using a normal

approximation, is actually just as easy:

2nT‾χ2n,1-α/22<μ<2nT‾χ2n,α/22

and is definitely preferable

Note that we can use the same approach to the problem of testing for the

mean. However, if we want to be sloppy, we can use the

“usual” testing strategy, using a normal distribution, since

assuming that the true mean is μ, implies

that the true standard deviation is also equal to μ.

In other words, we could use as “test statistics”

nT‾-μμ

instead of the, more correct,

T‾-12μ

However, since

μ2nT‾

has a χ2n2

distribution, the better test relies on this, rather than those

approximately normal choices.

Binomial Distribution

Consider a coin tossing experiment, where we count the number of heads.

Assuming the coin may be biased, call p the

probability of coming up heads in a single toss. If the tosses are

independent and identical (same coin, no cheating), we saw that the

probability of ending up with k heads, in

n tosses is given by

P[X=k]=n!k!(n-k)!pk(1-p)n-k

(n!, read as

“n factorial”, is the product

of all integers form 1 to n:

1⋅2⋅…⋅(n-1)⋅n).

The true mean for X is n⁡p,

and its true variance is n⁡p(1-p).

Since X is actually the sum of n

independent random variables (each equal to 1

if the toss came up heads, and 0 if it came

up tails), the Central Limit Theorem tells us that 1nX=X‾

will be well described by a normal distribution for n

sufficiently large, and p not too close to

0 or 1. You

may recall seeing one or more “rule of thumb” suggested for

determining whether normality is a reasonable approximation. Note that

X‾ has true

mean p, and true variance p(1-p)n.

If we want to use this experiment to estimate p,

we can use the normal approximation, but since p

is unknown, we do not know the variance. One could think of using a

Student distribution for this purpose, but this is not recommended, as

conditions are not quite right, and one feels that with the variance

being so closely tied to the mean, we may avoid that.

A common approach is to use the sample mean in place of p

in the variance formula. Thus, if (as is often written), the sample mean

is p^, one

would use the normal distribution, with unknown mean p,

and “known” variance p^(1-p^)n.

As we already discussed in this course, the result is a confidence

interval that may be too small or too large, with no way of knowing

which way it may be “wrong”.

A more conservative approach starts from the observation that, since

0⩽p⩽1,

p(1-p)⩽14.

Hence, using 14 as

the value of the variance sets us up in the “worst case”,

and will result in a confidence interval which will be no smaller than

the “true” one. In other words, we would estimate

p^-zα/2⋅14n⩽p⩽p^+zα/2⋅14n

There is actually yet another way to proceed, but it is not popular,

since it is not nearly as straightforward and automatic, as it requires

more manipulations. Writing out the fact that the sample mean is

approximately normal, we have that an interval estimate at confidence

level α is given by

p^-zα/2p(1-p)n⩽p⩽p^+zα/2p(1-p)n

p^ is again the

sample mean here. Looking at the inequalities as two quadratic

inequalities in p, we can solve them and

find a more precise interval estimate than either of the methods

mentioned above:

zα/2np2+p(1-zα/2n)-p^⩽0

zα/2np2-p(1+zα/2n)+p^⩽0

You would get an estimate by finding all values of p

that satisfy both inequalities.

This issue is not present when testing for p,

because in this case we are assuming that p

is a given value, so that the corresponding variance for the sample mean

has to be p(1-p)n.