How To Apply Statistical Techniques

We have had a quick look at the theory behind statistical sampling, and,

in general, how statistical tools might help in the measurement of

factors. However, the translation from theory to practice is very far

from trivial. In fact, the so-called Design of Experiments is a

full fledged research area, and the abundance of examples where the

practice did not quite live up to theoretical standards proves how

difficult this can be.

There are general classifications for various types of statistical

applications, and most of those you will find mentioned in textbooks

refer to variations of polling. We should not forget the important

applications from areas like quality control, survival analysis, or

scientific measurement.

In an introductory course, we can only sketch some general ideas

concerning this complex topic, trying to convey a feeling for the

considerable challenges that have to be met. Actual implementation of a

statistical study requires a much deeper study of all the elements

involved.

In this quick overview we will not address issues related to Census

taking. By definition, a census aims at collecting complete information

on a given population. This is trivial for really small populations

(e.g., taking a census in a classroom), and poses all sorts of very

peculiar statistical issues when performed on the scale of, say, a

national census (as well-known examples, you might think of issues like

undercounting of specific groups, or the care needed to cross-reference

collected data). Given that this is a highly specialized area, we will

leave it for dedicated studies, and concentrate on common statistical

activities.

Theory and Application of Sampling

Recall what the theory behind sampling theory is: we have a number of

observations that are supposed to observations of independent and

identically distributed random variables. That implies that

different numbers are the result of completely random fluctuations, and

not due to any bias or connection between the different outcomes.

Additionally, we are also assuming that the distribution we are talking

about is indeed the one we are interested in studying.

To see that it is not trivial to produce an experiment matching these

requirements, think of a "simple" problem like polling the

American population about, say, their favorite ice cream flavor. We are

interested in the proportion of the whole population, and that

means that the small sample we will actually interview should be

"representative" of the whole population: the likelihood of

interviewing a strawberry lover should be the same from the sample as

from the population. In the theoretical setup, we may think of the

population consisting of shy of 310 million colored balls (each flavor

corresponding to a color) in a big bowl, with a blindfolded child

picking one ball at a time, with the bowl thoroughly turned so the balls

are thoroughly mixed before each extraction. This ideal setup is the

model for the gold standard of sampling: simple random

sampling.

That's not how the respondents are actually going to be picked in

reality, and it is not quite obvious how to construct a decent

approximation to that abstract model.

First of all, it is very easy to do this completely wrong. Such bad

practices are so common that some have their own names. Typical examples

are convenience sampling (for example, you go somewhere

convenient, and interview the people you meet: typically the result of

your "poll" has no significance at all, since you picked a

specific type of people - those who go where you went - with very little

randomness, and to assume that the responses are independent is also

quite stretch), and voluntary response sampling, as in phone-in

or Internet polls (people who respond chose to respond, and were not

chosen "at random" at all, so their responses reflect their

responses only, and nothing can be inferred about a larger population)

Traditionally, a widespread method was to pick names at random from

phone books. It was not perfect (for example, ideally, we need a

nation-wide phone book, and you are excluding people who do not have a

phone), and it is less and less so, as a growing number of Americans opt

to rely on cell phone service only. The typical method used to go in

steps:

pick a telephone exchange at random

add random digits to the exchange to produce a phone number

pick one adult from each household you reach this way

There are complex issues behind each of these steps (each exchange has

to be weighed by the population that uses it, random digits produce non

existent or illegal phone numbers, you may not have the choice of

picking the adult at random: you may have to pick whoever happens to be

at home, and more), and the design of such a procedure to make it at

least reasonably close to a true random sampling is completely non

trivial. Even if a good design has been implemented, the issues of

under-coverage (people who have no land line, homeless people,

students in college dormitories, prison inmates, and so on) are

important. Independence between respondents is another possible problem,

but the, biggest problem is posed by people who are selected, but, for

some reason, do not respond. Depending on what we are actually asking,

this can be a big source of bias. Polling services keep devising ways to

"correct" for this problem, but, almost always, these ways are

kept as trade secrets, and not made public. This, as in all

"closed" situations, makes it impossible to subject the

methodology to peer review, and thus you have to trust the small team

that came up with the idea for it to be effective, without the

validation that proper scientific review guarantees.

As we can see, it is not that surprising if a poll fails to predict an

election correctly, but think of the many more surveys that have no

benchmark to be checked against, as a pre-election poll has.

How to pick anything at random?

Currently, the most convenient way to “pick something at

random” out of a list of candidates, is to use a random number

generator program (RNG) on a computer. As discussed in the Special

Topics section on Simulation, computers do not really come up

with truly random numbers, but, without addressing this quite difficult

issue, it turns out that, for all intent and purposes,

pretending that they are “truly” random has proved

to be a viable attitude.

What a RNG will provide you is a number between 0 and 1 with the

probability of it falling between a and

b, with 0⩽a<b⩽1,

being (theoretically) b-a.

Your spreadsheet produces such a number through the function

rand() (a similar function is part of the standard C

libraries, and for most other computing languages). An alternate

possibility is the function randbetween(x,y), where x and y are two real

numbers: it produces a “random integer”

between x and y

(here, “random” means that, theoretically, all the

qualifying numbers have equal likelihood to be chosen).

In a simple random sampling experiment you would list all candidates to

be chosen, and, since each should have the same probability of being

chosen, you could code each with a number, say, from 1

to n, and have randbetween(0.5,

n+1/2) choose one. Alternatively, you could divide the interval

[0,1]

in n equal subintervals of length 1n, connect each

candidate with one such interval, and choose the one corresponding to

the interval where the output of rand() falls.

Note that this ideal situation is not what you would face in most

practical cases. Even the phone sampling strategy mentioned above does

not fit this model, since you cannot assign the same probability to all

telephone exchanges - typical exchanges in Manhattan include a lot more

people than typical exchanges in Wyoming.

Variations to the Simple Random Sampling Method

To mitigate some of the difficulties in setting up a random sampling

that matches the theoretical ideal, variations have been devised, but

they lead to new difficulties, some of them very serious. Here are a few

examples.

Stratified Sampling

Ideally, the population is divided in strata, that is subgroups

that share some feature (e.g., by geography, by political orientation,

by gender, ...), and each subgroup is sampled randomly. The results are

combined, by weighing each stratum by its numerosity in the total

population. In theory, this should allow one to build a

representative sample of the population with fewer individuals.

As you can easily see, to produce a proper stratified sampling procedure

requires a lot of difficult work, as it necessitates to be able to

classify, for example, all Americans according to the strata we

selected, and to properly assess the number of individuals in each

stratum, as well as correctly assigning individuals. In a way, the

telephone strategy sketched above is a version of stratified sampling.

This is a (generically) recommended procedure, but you should not

underestimate the problems that have to be solved to make it effective.

Still, if conducted properly, it is a legitimate implementation of the

sampling paradigm. That is not the case, to varying degrees, of other

popular methods, as exemplified in the following.

Systematic Sampling

This is "simple" method that can easily produce very biased

samples, without you even being aware of it. The idea is to choose an

integer number, let's call it k, then pick one individual at

random, and proceed to poll every further k-th individual.

Supposedly, this method would not require you to know too many details

of the population you are sampling. As an example, some textbooks will

suggest this as a good way to test a production line, in quality

control. It is not difficult to figure out a situation where a glitch in

the production occurs periodically. A systematic sampling approach, in

this case, could either miss the problem completely, or blow it out of

proportion.

The problem is that the randomness of this method is shifted from the

random selection of each individual, to the selection of the sequence in

which you consider individuals, as this determines who the

k-th, 2k-th, 3k-th, and so on will be.

Setting up a true "random sequencing" is much harder than

successively extracting "random numbers". In fact, it is the

connected to the difficult problem of generating "good"

"random" sequences in a computer (an extremely challenging

problem in computer science). Incidentally, in theory, if the sequence

you are sampling was really random (this ensuring that the method does

not introduce bias), choosing k=1

would be just as effective as any other choice.

Cluster Sampling

This could be applied in a quality control setting, where, instead of

sampling n individual items, all items from a random selection

of batches (as in cans organized in 12-packs) are tested. While this

method seems to make it easy to sample a sufficiently large number of

items with much less effort, the gain may be completely illusory. Since

one could well suspect that items in each batch are not independent (for

example, production problems may affect all cans in a specific batch),

the size of the sample is not really determined by the number of

individuals tested, but rather by the number of batches.

Convenience Sampling

This method should not even be considered "sampling". It

consists in picking individuals that happen to be found. For example,

"sampling" the population of a city by going to a mall and

polling visitors, or "sampling" the student population of a

school by polling your class. In a way, you can put opt-in surveys, like

Internet polls, as convenience sampling too. In any case, results

obtained by this method are completely worthless and represent only the

specific group that was polled.

Related Experimental Designs

There are other observation methods that do not aim at producing a

random sample, in the sense we have discussed, and the techniques used

on these are not necessarily the same as in a sampling experiment. Here

are a few common examples.

Controlled Experiments

We discuss them in the next section.

Acceptance Sampling

Actually, this is not a procedure that fits into this lineup. It is a

rough and tumble approach to quality control. Here is a prototypical

example:

You receive supplies in batches of N items. You are not willing

to tolerate batches that have p = k/N

defective items, or more. Since checking for defective items is

destructive or, at the least, expensive and unwieldy, you

"sample" n items and reject the batch if one or more

sampled items are defective.

To make sense of this procedure, you have to choose n in terms

of your choice of k and N. You can quickly realize

that this is a standard problem in probability, but does not tie in well

with "standard" statistical procedures. Specifically, here is

how you would choose n, given N, and k. As an

example, consider N = 1000, k = 10, (p =

0.01). This is a binomial experiment: assume you are willing to live

with a q chance of accepting a batch that you should refuse

(let's take q = 0.05 - 5% - as an example). You need to choose

n such that the probability of that many items not being

defective, even if the overall percentage of bad items is p

(e.g., 1%) or higher, should be no greater than q (5%). Since

the probability of an item being good is 1-p

(or less), the probability of n items being

good, assuming that they are independent - a non trivial

assumption here, is no more than (1-p)n.

Now, we need (1-p)n⩽q,

or n⁡log(1-p)⩽log⁡q,

or (recall that probabilities are less than 1, hence logarithms are

negative, and we are dividing by a negative number),

n⩾log⁡qlog⁡(1-p)

(It does not matter what base you are using for the logarithms). For our

example, q=0.05,

p=.99,

hence,

n⩾298.07

which is quite large. To make this method cost effective, we would have

to either tolerate a lot more defective items, and/or be resigned to not

rejecting many more “bad” shipments. Regardless of this

uncomfortable number, you can appreciate how this procedure is not part

of the “sampling” list we just went over.

Experiments

Even though the basic theory underlying "experiments" (as

opposed to "observational studies") is the same, the practice

is somewhat different, and is fraught with different problems. An

experiment consists in actively subjecting individuals to some kind of

procedure, and observing the result. In an observational study, on the

other hand, individuals are passively observed. For example in a quality

control experiment, we might take a production sample and subject the

items in the sample to a stress test, to evaluate their sturdiness. In

this latter example, we might have a benchmark in mind - for example, we

might want to evaluate the likelihood that a certain appliance will fail

in the first year of operation, matching this likelihood with the costs

that we might incur because of our warranty. In other cases, typically

when testing for the medical effectiveness of a treatment, there is no

obvious benchmark, and to make the experiment meaningful, we need to

compare the result of the new treatment with those from an old,

reference, treatment, or even with no treatment at all. An additional

issue is that the observed results might be due to causes that we have

not identified in the first place. These "lurking" variables

can sway the result in a way that leads us to incorrect conclusions. The

methods in use in this area try to limit as much as possible these

pitfalls.

As we all know, this type of statistical study is very widespread, but

it does come with significant challenges. First of all, except in cases

like the stress test, in reference to the warranty costs example, there

is indeed the need to have a control group, in order to

evaluate if the procedure we are testing produces the desired effect or

not. Having two groups to compare, adds, to the usual problem of

producing a sample, the problem of how to assign individuals to the two

sub-samples. The latter question can be addressed by

randomizing the design: individuals are assigned to the

sub-groups at random. When comparing two treatments, it is also

important to avoid a bias due to people (or even only the experimenters)

knowing which treatment they are experiencing. Thus, the "golden

standard" is the double blind randomized design, where

only the designers of the experiment know who is getting what treatment.

There are several issues that are not easy to address even in the best

experiment. Firstly, there is the problem of how the individuals

participating in the experiment are chosen. This is rarely discussed (in

fact, in many cases, these are volunteers), but it does affect the scope

of any conclusion we might draw. The size of the groups is an obvious

issue, and practical considerations may force to work with relatively

small groups. In recent years, to address this problem, there has been a

growing number of “meta-studies”, studies that do not

perform experiments themselves, but rather aggregate the results from

many related experiments, in order to work with a sufficiently numerous

data set. The problem of properly merging disparate experiments,

performed under similar, but different circumstances, is obviously far

from trivial, and lies well beyond our scope.

Scientific Measurements

This is an area where things are more straightforward, since it usually

involves nature, and not living beings. When measuring a physical

quantity (e.g., the speed of light), the ideal model of observation can

be approximated very well if we are careful enough. Indeed, once we set

up the experimental apparatus, we should make sure that external

influences are reduced to a minimum (e.g., you will not set it up in a

building under which there is a subway line with trains running every

five minutes). Also we will have to make sure that our instruments are

not biased (think of a time measuring device that runs too fast). This

is not easy to set up and, indeed, good experimenters are not abundant.

A measurement will then be performed many times, in such a way that the

results may be viewed as observations that are independent and

identically distributed: each measurement is performed in exactly the

same manner, and we are careful to ensure that each measurement has no

influence on any other. A careful precise experiment will produce

different numbers when repeated (in fact, if it doesn't, we may conclude

that the observations are not independent), and having eliminated as

many perturbing factors as we can, we may feel confident that variations

in the measurements will be due to random uncontrollable effects. At

this point, we can apply standard statistical techniques, and come up

with estimates that are probably the most rigorous application of

statistical methods that we have. If you look up the value of physical

quantities as published in the scientific literature, you will find

statements like "the universal gravitational constant (in Newton's

gravitation law) is G=6.67259×10-11±8.5×10-15"

(using scientific notation). The number after ± is the standard

deviation of our estimate, that was obtained by interval estimation, and

the appropriate Student distribution.

One More Issue: Simultaneous Measurements

Here is yet another trap where inexperienced experimenters will easily

fall. In most situations, we will not be measuring one quantity only,

even if that is what we have been suggesting all along this course. The

reason for this reduced presentation is that simultaneous observation of

several quantities poses a host of additional issues. To be precise, if

we are observing more than one feature (and this is what will often

happen), the proper statistical goal would be the joint

distribution of the observed quantities. Since this is a

significantly more complex problem, it is tempting to bypass it, and

pretend that the data as coming from independent

factors. Unfortunately, this is often a ridiculously unrealistic

assumption. Depending on what you will actually work out of the raw

data, your conclusions might be fine, but they could also be

meaningless. You would be surprised to learn how often badly designed

experiments have led to totally meaningless conclusions, simply because

the researchers did not account for the fact that the various

measurements they took were correlated (in the intuitive, as

well as technical, sense). Of course, the most famous recent case of

mishandled dependencies is the wild underestimate of the likelihood of a

“snowball effect” in mortgage defaults in the years leading

to the crash of 2007. While the assumption was not of independence, the

lack of data prompted financial operators to “guess” the

dependencies using a method without solid empirical justification.

Conclusion

This has been, obviously, a very cursory exploration of the issues

involved in designing experiments. In real circumstances, the

considerations we made may have different impact. Unfortunately, it is

not always easy to sort out which situations are more affected by less

than careful design than others. The main conclusion that you may take

from this discussion is that extreme care should be applied when

evaluating the significance of a statistical study. A good design behind

the study will ensure that the results are very valuable, but a poor

design should be a red flag, as it is all too easy to draw unwarranted

conclusions from a poorly conducted experiment.