***Learning Statistics with R - University of Adelaide***

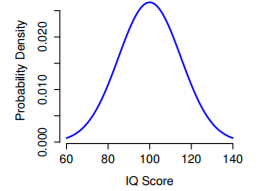
***Part IV – Statistical Theory***

**Chapter 10 – Estimating Unknown Quantities From A Sample**

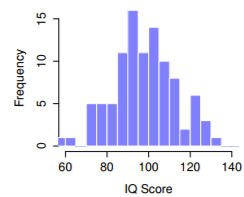
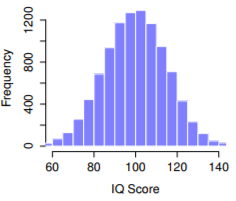
* Role of descriptive statistics = to concisely summarize what we DO know
* Role of inferential statistics = to learn what we do NOT know *from what we do*
* 2 big ideas: **estimation** and **hypothesis testing**.
* Estimation theory doesn’t make sense until you understand sampling

**10.1 - Samples, populations and sampling**

* ALL learning requires you to make assumptions
* 1st task: To come up w/ some fairly general assumptions about data that make sense 🡪 where **sampling theory** comes in
* **Probability Theory** = foundation upon which all statistical theory builds + **Sampling Theory** = the frame around which you can build the rest of the house.
* **Sampling theory** plays a huge role in *specifying the assumptions upon which statistical inferences rely.*
* Drawing inferences *from the sample* + *about the population.*
* Typical data set available to us is finite + incomplete.
* A **sample** is a *concrete*, while a **population** is a more abstract + refers to the set of ALL POSSIBLE observations you want to draw conclusions about (is generally much bigger than the sample)
* Ideal world = researcher begins a study w/ clear idea of what the **population of interest** b/c the process of designing a study + testing hypotheses about data the study produces depends on the population about which you want to make statements.
* usually researchers have vague ideas of what the population is + designs the study as best they can on that basis.
* Irrespective of how population is defined, the critical point is the **sample** = a subset of a population
* Goal = to use knowledge of the sample to draw inferences about the properties of the population.
* Relationship between the 2 depends on the *procedure by which the sample was selected* = **sampling method**
* The fact that the same sampling procedure can lead to different results each time we take a sample (i.e. sample of 5 cards w/out replacement repeated) means we refer to it as a **random process**
* Simplified: *a process has an element of randomness to it whenever it is possible to repeat the process and get different answers each time.*
* A procedure in which every member of the population has the same chance of being selected is a **simple random sample**.
* If we know that a sampling scheme is biased, that a sample doesn’t tell you very much about a population
* A simple random sample makes the data analysis much easier.
* Can be done **with or without replacement**
* w/ replacement = can observe the same population member multiple times in a sample
* Most statistical theory is based on the assumption the data arise from a SRS *with* replacement
* Rarely matters in real life if population of interest is large (> 10 entities), the difference between w/ + w/out replacement is negligible
* Almost impossible to obtain SRS from most populations
* **Stratified sampling** =population is divided into several different subpopulations, or **strata**.
* collect a SRS *from each of the strata*.
* sometimes easier to do than SRS, especially when population is already divided into distinct strata + can also be more efficient, especially when some subpopulations are rare.
* Ex: studying schizophrenia 🡺 better to divide population into 2 strata (schizophrenic + not-schizophrenic) + sample an equal number of people from each group.
* If you selected people randomly, you’d get so few schizophrenic people in the sample the study would be useless.
* This is **oversampling** = makes a deliberate attempt to over-represent rare groups.
* **Snowball sampling** = especially useful when sampling from a hidden/hard to access population (especially common in social sciences)
* Ex: opinion poll among transgender people + team might only have contact details for a few trans folks
* Stage 1: Survey starts by asking them to participate + at the end of the survey, participants are asked to provide contact details for other people who might want to participate.
* Stage 2 = new contacts are surveyed + process continues until we have sufficient data
* Big advantage to snowball sampling = gets you data in situations that might otherwise be impossible to get any.
* Main disadvantage = sample is *highly* NON-random + is so in ways that’re difficult to address
* procedure can also be unethical if not handled well b/c hidden populations are often hidden for a reason.
* might end up outing people who don’t want to be outed + can be intrusive to use people’s social networks to study them.
* Very hard to get people’s informed consent before contacting them, yet in many cases the simple act of contacting them + saying “we want to study you” can be hurtful.
* Social networks are complex things + just b/c you can use them to get data doesn’t always mean you should.
* **Convenience sampling** = samples are chosen in a way convenient to the researcher + not selected at random from population of interest.
* Snowball sampling = 1 type of convenience
* Studies on undergraduate psychology students automatically means data are restricted to a single subpopulation.
* Students usually get to pick studies they participate in, so a sample is a **self-selected subset** of psychology students, NOT a *randomly*-selected subset.
* In real life, most studies are convenience samples of 1 form or another
* Sometimes a severe limitation, but not always
* It *can* matter if data are not a SRS but it’s not quite as bad as it sounds.
* Some types of biased samples are entirely unproblematic
* **Stratified Sampling** = actually know what the bias is b/c YOU created it *deliberately*, often to *increase the effectiveness* of the study
* There’re statistical techniques you can use to adjust for biases you’ve introduced
* More generally though, remember that random sampling = a *means to an end, not the end in itself.*
* A bias in sampling method is only a problem if it causes you to draw the wrong conclusions.
* Don’t need sample to be randomly generated in *every* respect: only to the phenomenon of interest
* Ex: For a memory study, can sample randomly from all human beings currently alive w/ 1 exception
* Can only sample people born on a Monday OR Can only sample randomly from Australia.
* To generalize results to the population of ALL living humans, study 1 is better b/c we have no reason to think being born on a Monday has any interesting relationship to memory capacity
* Being Australian might matter b/c Australia is a wealthy, industrialized country w/ a very well-developed education system.
* People will have had life experiences much more similar to experiences of those who designed the tests + this shared experience might easily translate into similar beliefs the test
* When designing studies, it’s important to think about *what population you care about +* to try hard to sample in a way *appropriate to THAT population*.
* Usually forced to put up w/ a “sample of convenience”, but if so at least spend some time thinking about what the dangers of this practice might be.
* Secondly, if criticizing a study b/c of a sample of convenience rather than laboriously sampling randomly from the entire human population, have the courtesy to offer a specific theory as to how this might have distorted the results.
* Everyone in science is aware of this issue + does what they can to alleviate it.
* In most cases, populations scientists care about are *concrete* things that *actually exist* in the real world
* *Statisticians*, however, are interested in real world data + real science in the same way scientists are, but they also operate in the realm of *pure abstraction* in the way mathematicians do.
* As a consequence, **statistical theory** tends to be a bit abstract in how a population is defined.
* In much the same way psychological researchers operationalize abstract theoretical ideas in terms of concrete measurements, statisticians operationalize the concept of a “population” in terms of *mathematical objects* they know how to work w/ = **probability distributions**.
* Ex: IQ scores
* To a psychologist, population of interest = a group of actual humans who have IQ scores.
* A statistician “simplifies” this by **operationally defining** the population as: *the probability distribution of IQ scores*



* IQ tests are designed so the average IQ = 100 + the SD = 15 + the distribution of scores is normal
* These values = **population parameters** (characteristics of the entire population)
* Select 100 + 1K people at random from this normal population + administer an IQ test for an SRS

* Histogram is roughly right shape, but a very crude approximation to the true population distribution, + sample mean = fairly close to population mean but not identical.
* **Sample statistics** are properties of a sampled data set, + although they can be fairly similar to the *true* population values, they are NOT the same.
* **Sample statistics** = things you can calculate from a data set + **population parameters** = things you want to learn about.