Select Reading Question Responses (9/14/2021)

> This is more of a logistical question but I noticed that the book always says "statistical software was used to calculate __" but never tells us how to do it in R. I've had to google the actual functions and was wondering if there was a way we could figure these out easier? Or if you could provide the basic functions for us?

We've seen three continuous distributions in this class: standard normal, χ^2, and t. For each of these three, R provides functions to go from observations of these distributions to percentiles (names begin with `p`), and functions that go from percentiles to observations (names begin with `q`). These functions are the following, where you fill in the question marks with the appropriate numbers.

- Standard normal distribution: `pnorm(?)` and `qnorm(?)`
- χ^2 distribution: `pchisq(?, df = ?)` and `qchisq(?, df = ?)`
- t distribution: `pt(?, df = ?)` and `qt(?, df = ?)`

I think those are the main things that the book has invoked "statistical software" for so far.

> Why does a sample size of 30 serve as a good standard for helping to determine if a distribution is normal? Why was the rule of thumb of a sample size of 30 not described in earlier representations of the success-failure conditions?

> I am curious how this value of 30 is reached, and similarly how values like 10 are found for the success-failure condition. Are they arbitrary, or is there a statistical justification/proof for these values?

These heuristics have to do with the central limit theorem, but this is not so easy to describe without getting into somewhat complicated math. Let me try to explain this, but be forewarned that even my short-ish explanation here isn't so easy (especially if you haven't seen limits in a calculus class before).

Suppose $X_1, X_2, \ldots$ are all independent and identically distributed random variables with mean μ and standard deviation σ. Let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

The general version of the central limit theorem says that

$$\lim_{n\to\infty} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges to a standard normal distribution $N(0,1)$.

This is a mathematically precise result with a formal proof and so forth. Loosely interpretted, it means that when n is "very big,"

$$(\bar{X}_n - \mu)/(\sigma/\sqrt{n}) \approx N(0,1).$$

There's a question of what "very big" means. This depends very much on what application you have in mind *and* what the distribution of the X_i's is like.

For example, one important special case of the central limit theorem comes up when we're studying a proportion. In this case, each X_i is a Bernoulli random variable. For instance, maybe we're interested in the proportion of Americans who speak Spanish at home. In this case, "success" means that a given American speaks Spanish at home. In other words, X_i represents picking out a random American, and X_i outputs 1 if that American speaks Spanish at home and 0 if they don't. All of these are Bernoulli random variables; if p is the true proportion of Americans who speak Spanish at home, then each X_i has expected value $\mu = p$ and standard deviation $\sigma = \sqrt{p(1-p)}$. Now if I pick out a sample of n Americans, the random variable

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

models the proportion of my sample who speak Spanish at home (the numerator counts how many sampled Americans speak Spanish at home, and the denominator is the number of Americans sampled). In other words, each $\bar{X}_n$ is a binomial random variable. The central limit theorem then says that

$$\lim_{n\to\infty} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

converges to a standard normal distribution. In other words,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

or, by rearranging terms,

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

where we recall that $\mu = p$ and $\sigma = \sqrt{p(1-p)}$ for Bernoulli random variables. When is n "big enough" that the error in the above approximation is too small to matter? It really depends on how much precision you need for whatever application you have in mind, but for *most* applications, people have found that n is "big enough" when the success-failure condition is satisfied (ie, when $np \geqslant 10$ and $n(1-p) \geqslant 10$).

The second important special case of the central limit theorem comes up when we're studying numbers that come from a normal random variable. For example, suppose we're interested in the average height of adult American women. In this case, each X_i represents the height of a given American woman. The expected value μ of X_i is the population

average height of all American women, and the standard deviation σ of X_i is the population standard deviation. Now if I pick out a sample of n American women, the random variable

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

represents the average height of my sample. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

follows a t-distribution with $n - 1$ degrees of freedom, and the central limit theorem says that, when n is "big enough,"

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1),$$

or, by rearranging again,

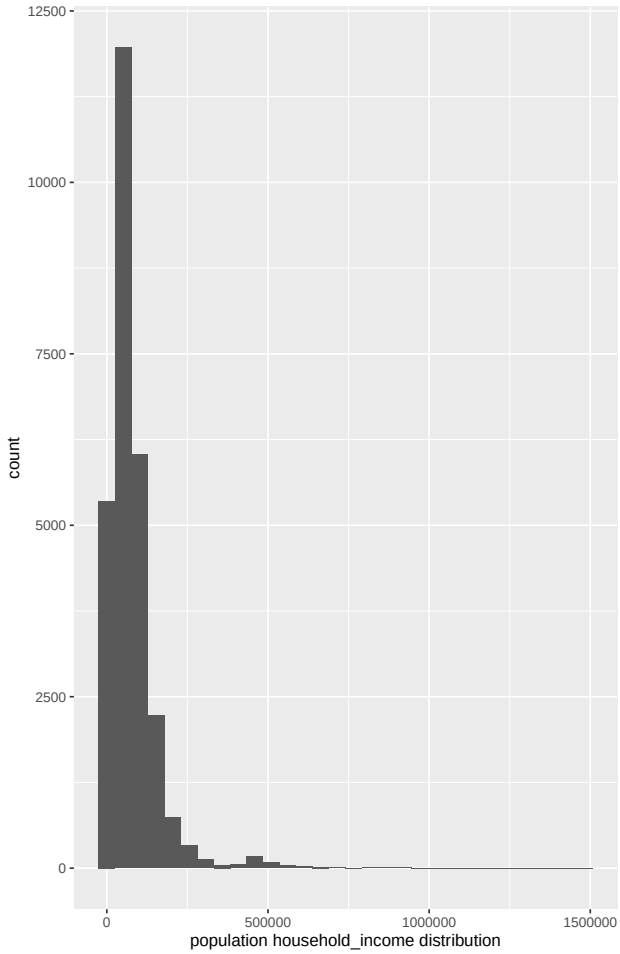
$$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

When is n "big enough" that the error in this approximation is too small to matter? Again, it really depends on how much precision you need for the application you have in mind. But, for *most* applications, people have found that $n \geqslant 30$ is "big enough."

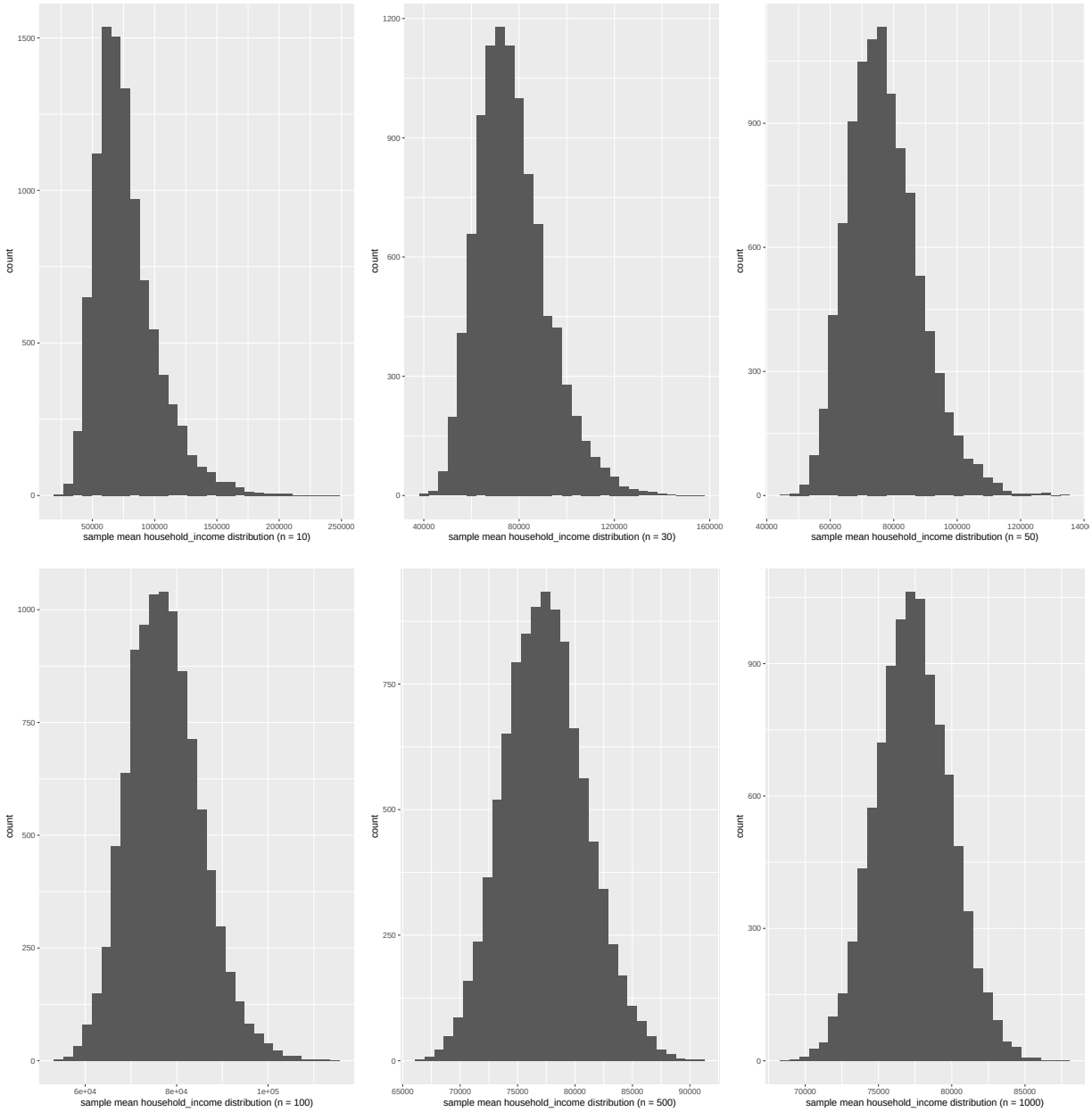
Technically the previous paragraph above only works verbatim when the X_i's are normal. If they're not "too far" from being normal, you can safely continue to use the same $n \geqslant 30$ heuristic. But if your X_i's are neither Bernoulli nor normal, then neither of the heuristics from the previous two paragraphs is guaranteed to hold! If the X_i's are *really weird*, it could be that you need *much bigger* values of n before the error in approximation becomes small.

For example, the distribution of household incomes in the US is *extremely* skewed. If each X_i represents the household income of a given American household, then the central limit theorem still guarantees that $\bar{X}_n \approx N(\mu, \sigma/\sqrt{n})$ when your sample size n is "big enough." Here $\bar{X}_n$ represents your sample mean household income, μ is the popoulation mean household income in the US and σ is the population standard deviation of household income in the US. But, because each X_i is so far from being either Bernoulli or normal, you need *much* bigger values of n here before the sampling distribution starts to look normal.

The US Census Bureau provides household incomes from 2019 for a sample of 27256 households. This distribution looks like this:



Let's pretend like this is a *population* distribution. We'll draw random samples from this population of 27256 households and look at the sampling distribution of the mean household income. Below are the sampling distributions for sample means for various sample sizes n. The $n = 10$ distribution is very skewed, but even the $n = 30$ and $n = 50$ distributions are somewhat skewed. It's not till $n = 500$ that the distribution starts looking decidedly normal to me. The point is that the $n \geqslant 30$ is a heuristic that only applies if you're drawing from a population that looks normal to begin with. If your population is extremely non-normal, you might need much bigger sample sizes to get roughly normal sampling distributions of means.


count
sample mean household_income distribution (n = 10)
sample mean household_income distribution (n = 30)
sample mean household_income distribution (n = 50)
sample mean household_income distribution (n = 100)
sample mean household_income distribution (n = 500)
sample mean household_income distribution (n = 1000)