Sampling from finite population

Hypergeometric distribution is appropriate when we are sampling without replacement from a finite population.

Example

A box contains 20 mobile phones of which 12 are good and 8 are defective. You took 5 phones at random from the box without replacement. Let X denote the number of phones that are good out of the 5 taken. $(X = 0, 1, 2, 3, 4, 5)$.

Example

There are 30 fish in a pond of which 20 are tilapia and 10 are catfish. A total of 7 fish are caught with a fishing rod. Let X denote the number of tilapias caught.

Hypergeometric pmf

Let the population have N items of which M are 'good' and $N - M$ are 'bad'. Let n items be drawn at random without replacement. Let X be the number of 'good' items drawn. Then, X has the following pmf:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \; ; \;\; x = 0, 1, 2, \cdots, n$$

[when $n < M$ and $n < N - M$.]

Here, $X \sim \text{hypergeometric}(N, M, n)$.

Exercise

There are 20 tilapia and 10 catfish in a small pond. You caught 5 fish from the pond with a net. What is the probability that exactly 2 of the fish are tilapia?

Solution

$$P(X = 2) = \frac{\binom{20}{2}\binom{10}{3}}{\binom{30}{5}} = 0.1600$$

Comparison with binomial distribution

Let a box contain 10 phones: 6 good and 4 bad. Let us draw 3 phones at random without replacement.

Each trial (draw) has two possible outcomes: good and bad (Bernoulli trial). However, since draws are done without replacement, probability of good (success) changes from draw to draw.

For the 1st draw: $P(G) = 6/10$.

For the 2nd draw: $P(G) = 6/9$ or $5/9$
(depending on the outcome of 1st draw)

and so on. Thus, conditions of binomial distribution are NOT fulfilled. If the draws are done with replacement, we will have a binomial setup.

Hypergeometric to binomial

If the population is large, there is almost no difference between sampling with replacement and sampling without replacement. (Think of taking a glass of water from an ocean. Does it matter whether you replace it or not?)

Let a box contain 10000 phones: 6000 good and 4000 bad. Let us draw 3 phones at random without replacement.

For the 1st draw: $P(G) = 6000/10000$.

For the 2nd draw: $P(G) = 6000/9999$ or $5999/9999$

All three numbers are almost equal. Thus,

When $N \to \infty$, hypergeometric(N, M, n) reduces to binomial$\left(n, p = \frac{M}{N}\right)$.

Exercise

There are 2000 tilapia and 1000 catfish in a small pond. You caught 5 fish from the pond with a net. What is the probability that exactly 2 of the fish are tilapia?

Solution

Let X be the number of tilapias caught.

Method 1

$X \sim$ hypergeometric$(N = 3000, M = 2000, n = 5)$.

$$P(X=2)=\frac{\binom{2000}{2}\binom{1000}{3}}{\binom{3000}{5}}=0.16458$$

Method 2

$$X\sim\text{binomial}\left(n=5,\ p=\frac{2000}{3000}=\frac{2}{3}\right)$$

$$P(X=2)=\binom{5}{2}(2/3)^2(1-2/3)^{5-2}=0.16461$$

Distribution of Sampling Statistic

A statistic (for example, $\bar{X}$) takes different values for different samples. That is, a statistic is a variable. The distribution of a statistic is called sampling distribution.

We will discuss the sampling distribution of $\bar{X}$.

Let X have a normal distribution with

mean = μ
variance = σ^2

Let a sample of size n be taken repeatedly from the above distribution and $\bar{X}$ be calculated. Then, $\bar{X}$ will have normal distribution with

mean = μ
variance = $\frac{\sigma^2}{n}$

In this case, for probability calculation,

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

Example

Let X follow normal distribution with mean 50 and standard deviation 5. Let a sample of size 25 be taken from the distribution. What is the probability that $\bar{X}$ will be less than 49?

Solution

Here, $\mu = 50$, $\sigma = 5$, $n = 25$, $\bar{X} = 49$.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{49 - 50}{5/\sqrt{25}} = -1$$

Probability = 0.1587 (from *Z*-table)
That is, in about 16% of cases, value of $\bar{X}$ will be less than 49.

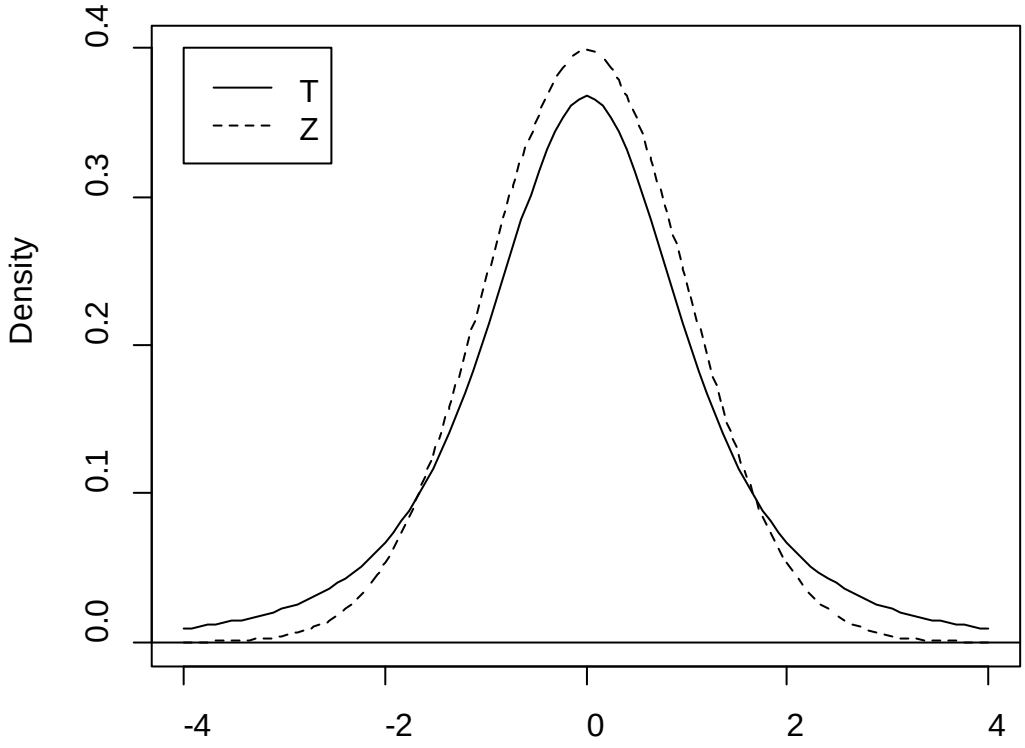
Some other sampling distributions

***t* Distribution**

When the distribution of X is normal and we are interested in the distribution of $\bar{X}$, usually σ is NOT known. Then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows t distribution with $n - 1$ degrees of freedom. Comparison of standard normal distribution and t distribution with 3 degrees of freedom is shown below.

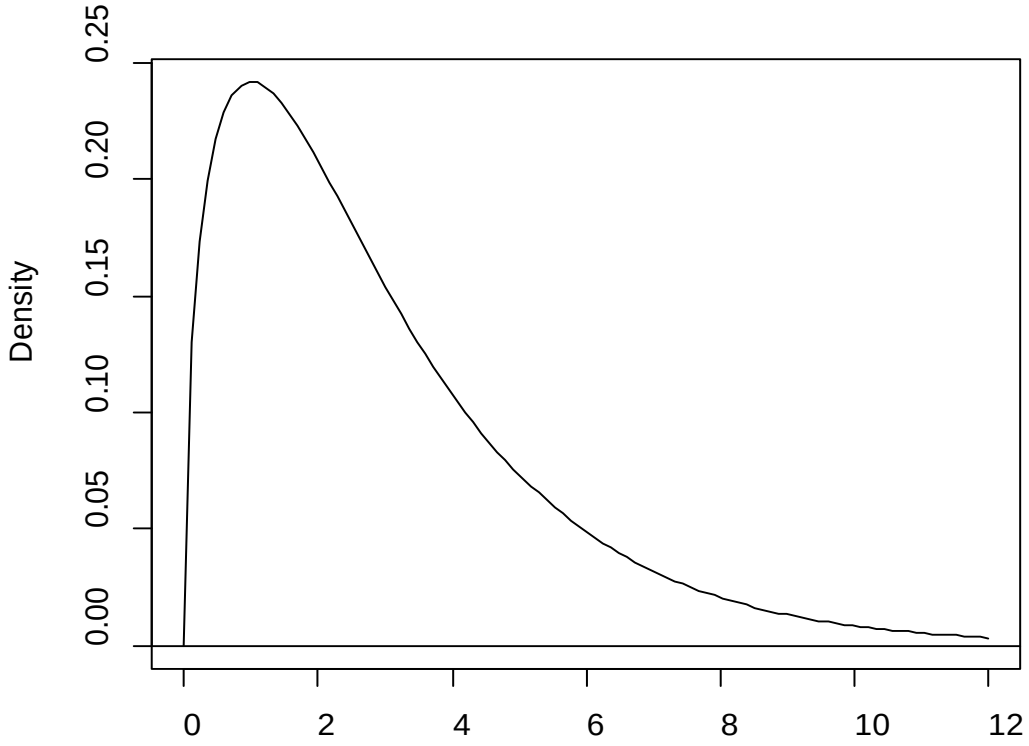


Chi-square Distribution

This distribution arises when our statistic of interest is the sample variance S^2. In particular, the statistic

$(n-1)\,S^2\,/\,\sigma^2$

follows χ^2 (chi-square) distribution with $n-1$ degrees of freedom. The following plot show a chi-square distribution with 3 degrees of freedom.

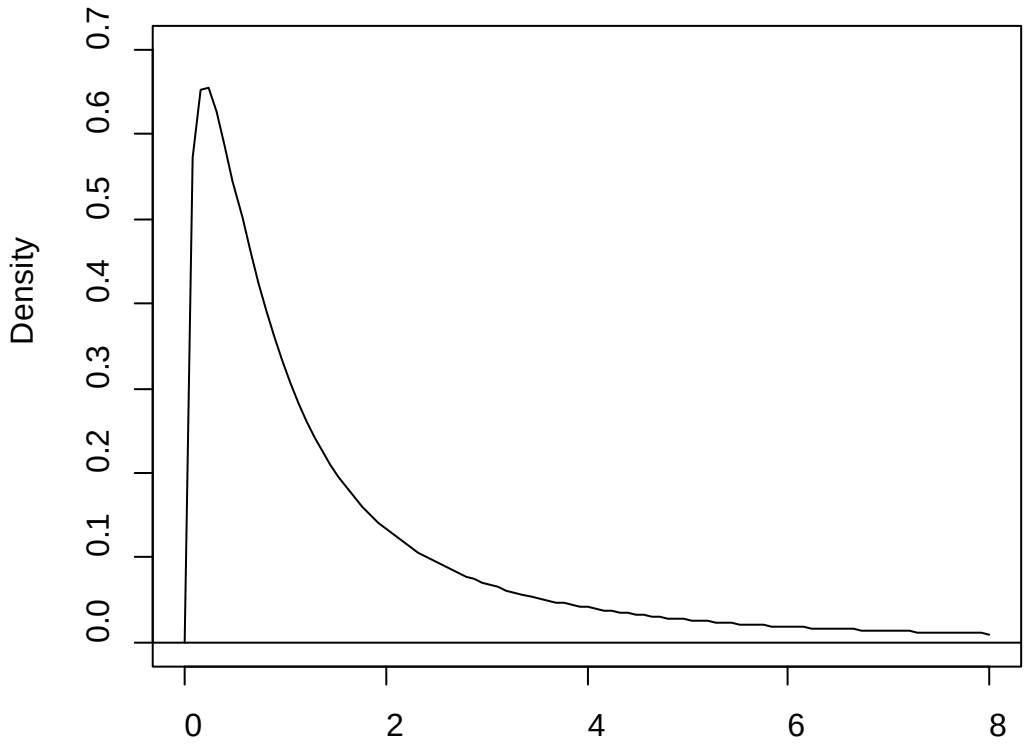


***F* Distribution**

Suppose we are comparing two independent normal populations. Let samples of sizes n_1 and n_2 be taken from the populations and the corresponding sample variances are S_1^2 and S_2^2, respectively. Then, the statistic

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

follows F distribution with degrees of freedom n_1-1 and n_2-1. The following plot shows an F distribution with degrees of freedom 3 and 3.



Central Limit Theorem

Let a large number of independent random observations be taken from any population with mean μ and variance σ^2. Then, mean (or sum) of the observations will follow a normal distribution approximately. That is,

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

approximately.

Example

The lifetime (in hours) of a particular type of electric bulb has mean 600 and standard deviation 40. You just bought 25 of these bulbs. What is the probability that the average lifetime of these bulbs will be more than 608 hours?

Solution

Here, μ = 600, σ = 40, $n = 25$ (large).

$P(\bar{X} > 616)$

$= P\left(\frac{\bar{X} - 600}{40/\sqrt{25}} > \frac{618 - 600}{40/\sqrt{25}}\right)$

$= P(Z > 2)$ (approximately)

$= 1 - \Phi(2)$

$= \ 0.0228$