8. Limit Theorems. Sampling Distribution.

The law of large numbers

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent random variables, each with the same distribution (or *identically distributed*) and mean μ. Then for every $\epsilon > 0$

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) \to 0$$

as n increases to infinity.

The law of large numbers: light versions

Consider a sample $X_1, \dots, X_n$ from a general population that has population mean μ. The Law of Large Numbers tells us that when sample size n is large (typically $n \geq 30$) then the sample mean (a random variable) is close to the population mean (a population parameter), that is,

$$\bar{X} \approx \mu.$$

Moreover, as sample size n increases, the size of fluctuations of $\bar{X}$ around μ is getting smaller.

The Law of Large Numbers also can be used to show that for large n sample variance s^2 is close to population variance σ^2, that is,

$$s^2 \approx \sigma^2$$

The central limit theorem

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent identically distributed random variables with mean μ and standard deviation σ. Then for any a and b

$$P\left(a < \frac{X_1 + \cdots + X_n - \mu n}{\sigma\sqrt{n}} < b\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-y^2/2}\, dy$$

as n increases to infinity.

The central limit theorem: light versions

Consider again a sample $\boldsymbol{X_1, \dots, X_n}$ from a general population that has population mean $\boldsymbol{\mu}$ and population standard deviation $\boldsymbol{\sigma}$. The Central Limit Theorem tells us that when sample size $\boldsymbol{n}$ is large (typically $\boldsymbol{n \geq 30}$) then

- the sample mean is *approximately* normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2/n}$, or

$$\boldsymbol{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

- the sum $\boldsymbol{X_1 + \cdots + X_n}$ is also *approximately* normally distributed with mean $\boldsymbol{n\mu}$ and variance $\boldsymbol{n\sigma^2}$, or

$$\boldsymbol{X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2)}$$

Sampling distribution of sample mean

- In real life calculating parameters of populations is often impossible because populations are very large.
- Rather than investigating the entire population, we take a *small* sample, calculate a statistic related to the parameter of interest, and then we try to make an inference.
- The sampling distribution of the statistic is the tool that tells us how close is the statistic to the parameter.

Sampling distribution of sample mean

Consider a sample $\boldsymbol{X_1, \dots, X_n}$ from a population with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$. Mathematically, we deal here with $\boldsymbol{n}$ independent identically distributed random variables. The sample mean is a random variable that is given by the following formula:

$$\overline{\boldsymbol{X}} = \frac{\boldsymbol{X_1} + \cdots + \boldsymbol{X_n}}{\boldsymbol{n}}$$

That is, the sample mean is the arithmetic average of observations in the sample.

Properties of sample mean

- The mean of the sample mean is given by

$$\mu_{\overline{X}} = \mu$$

- The variance of the sample mean is given by

$$\sigma_{\overline{X}}^2 = \sigma^2/n$$

- If the original data are normal, $\overline{X}$ is normal as well.

- If the data are non-normal but the sample size is sufficiently large ($n \geq 30$), then, by the Central Limit Theorem, $\overline{X}$ is approximately normally distributed.

Sampling distribution of sample proportion

- The parameter of interest for qualitative data is the proportion of times a particular outcome (a success) occurs.
- To estimate population proportion p we use sample proportion

  $$\hat{p} = \frac{X}{n}$$

  where X is the number of successes in the sample, and n is the sample size.
- Random variable X has a binomial distribution. Moreover, $E(\hat{p}) = p$ and $Var(\hat{p}) = p(1-p)/n$.
- However, in the case of a large sample size we prefer to use the normal approximation to the binomial distribution to make inferences about p. Again, the Central Limit Theorem is the reason why it works.

Normal approximation to binomial distribution

Normal approximation to the binomial works best when

- the sample size $\boldsymbol{n}$ is large
- and the probability of success, $\boldsymbol{p}$, is not too close either to $\mathbf{0}$ or $\mathbf{1}$

For a good approximation we typically need $\boldsymbol{np(1-p) > 5}$.

In such cases, approximately

$$\hat{\boldsymbol{p}} \sim \boldsymbol{N}\left(\boldsymbol{p}, \frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}\right)$$

Sampling Distribution of the difference between two sample means

What is the sampling distribution of difference between two sample means, when two random samples are drawn independently from two normal populations?

Sampling Distribution of the difference between two means

Consider one sample $\boldsymbol{X_1, \dots, X_n}$ from a population with mean $\boldsymbol{\mu_X}$ and variance $\boldsymbol{\sigma_X^2}$, and another one, $\boldsymbol{Y_1, \dots, Y_m}$, from another population with mean $\boldsymbol{\mu_Y}$ and variance $\boldsymbol{\sigma_Y^2}$. We assume that these two samples are independent.

Then if both sample sizes are large (typically, both $\boldsymbol{n \geq 30}$ and $\boldsymbol{m \geq 30}$) the difference between two sample means is normally distributed:

$$\boldsymbol{\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)}$$

Sampling distribution of $\overline{X}$: exercise

Exercise 1. A manufacturing process is designed to produce bolts with a 0.5-inch diameter. Once every day, a random sample of 36 bolts is selected and the bolt diameters are recorded. If the resulting sample mean is less than 0.49 inches or greater than 0.51 inches, the process is shut down for adjustment. The standard deviation is 0.02 inches.

1. What is the probability that the manufacturing line will be shut down unnecessarily (that is, when the true process mean really is 0.5 inches)?
2. Recalculate the probability of the same event for a random sample of 100 bolts.

Sampling distribution of $\widehat{\boldsymbol{p}}$: exercise

Exercise 2. An article reported that in a *large* study carried out in the state of New York, approximately 30% of the study subjects lived within 1 mile of a hazardous waste site. Let $\boldsymbol{p}$ denote the true proportion of all New York residents who live with 1 mile of such a site and suppose that $\boldsymbol{p} = .3$.

1. What are the mean value and standard deviation of $\widehat{\boldsymbol{p}}$ based on a random sample of size 400?
2. When sample size is 400, what is $\boldsymbol{P}(.25 \leq \widehat{\boldsymbol{p}} \leq .35)$?
3. When sample size is 900, what is $\boldsymbol{P}(.25 \leq \widehat{\boldsymbol{p}} \leq .35)$?