Point Estimation and Sampling Distributions

Anastasiia Kim

April 15, 2020

Statistical Inference

- A random sample is collected on a population to draw conclusions, or make statistical inferences, about the population.
- We choose a random sample of n members of the population:
  - a random sample consists of n independent r.v.s $X_1, X_2, ..., X_n$
  - every X_i has the same probability distribution
  - the r.v.s are independent and identically distributed (i.i.d.) if each random variable has the same probability distribution as the others and all are independent.
- We use the observed sample values to estimate characteristics/parameters, like the mean or variance, of the distribution.

Statistics

A *statistic* is a random variable whose value can be be computed from the values of the random sample $X_1, X_2, ..., X_n$.
Examples of statistics of interest:

- sample sum $\sum_{i=1}^{n} X_i$
- sample mean $\sum_{i=1}^{n} X_i / n$
- $\sum_{i=1}^{n} X_i^2$
- $min\ X_i$
- $max\ X_i$

Point estimators

Let θ be a parameter of the distribution of X:

- a statistic used to estimate θ is called an estimator, and is denoted by $\hat{\theta}$
- an estimate is a numerical value of an estimator for a particular collection of observed values of a random sample

Important: an estimator is a random variable, and an estimate is a number.
An estimator is unbiased (it is fluctuates around the right value) if

$$E(\hat{\theta}) = \theta$$

The bias of the estimator is $E(\hat{\theta}) - \theta$

Sample mean $\bar{X}$

- The sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ is a point estimate for the population mean μ
- Let $X_1, ..., X_n$ be a random sample of size n from a distribution with mean μ, then the statistic $\bar{X}$ is an unbiased estimator for μ:

$$E(\bar{X}) = E(\sum_{i=1}^{n} X_i/n) = \frac{1}{n} E(\sum_{i=1}^{n} X_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

$$\hat{\mu} = \bar{X}$$

Sample mean $\bar{X}$

- A desirable property of an estimator is that it has small variance for large sample sizes to ensure that estimates will be precise with large probability.
- Let $\bar{X}$ be the sample mean based on a random sample of size n from a distribution with mean μ and variance σ^2. Then, the variance is

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

  the larger the sample size, the larger the probability that our estimates are close to the true mean μ.
- The standard error of the sample mean is $\sqrt{Var(\bar{X})} = \sigma/\sqrt{n}$.
- The estimated standard error of $\bar{X}$ when σ is unknown would be

$$\hat{\sigma}_{\bar{X}} = S/\sqrt{n}$$

Sample variance

The statistic

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$

is the sample variance. S^2 is an unbiased estimator for σ^2:

$$Var(S^2) = \sigma^2$$

The Mean Squared Error

- In cases when a biased estimator used, the mean squared error (MSE) of the estimator can be important

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (bias)^2$$

- MSE is important criterion for comparing two estimators
- the estimator with the smaller MSE is more efficient
- sometimes biased estimators are preferable to unbiased estimators because they have smaller mean squared error
- we may be able to reduce the variance of the estimator considerably by introducing a relatively small amount of bias.

The Mean Squared Error

- Sometimes biased estimators are preferable to unbiased estimators because they have smaller mean squared error
- We may be able to reduce the variance of the estimator considerably by introducing a relatively small amount of bias
- In figure: $\hat{\theta}_1$ is a biased estimator of true value θ, whereas $\hat{\theta}_2$ is an unbiased estimator
- An estimate based on $\hat{\theta}_1$ would more likely be close to the true value of θ than would an estimate based on $\hat{\theta}_2$

The law of large numbers (LLN)

The LLN states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value.

- Let $X_1, X_2, ..., X_n$ be i.i.d. r.v.s with a finite expected value $E(X_i) = \mu$, then for any $\epsilon > 0$, the weak law of large numbers (WLLN) states that the sample average converges in probability towards the expected value

$$\bar{X}_n \xrightarrow{P} \mu, \quad n \to \infty$$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

The Sampling Distribution

The distribution of the random variable $\bar{X}$ is called the sampling distribution of $\bar{X}$

- there is randomness in the $\bar{X}$ value we get from a random sample because of sampling variability
- For a random sample $X_1, X_2, ..., X_n$ drawn from any distribution with mean $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$, we have

$$E(\bar{X}) = \mu \quad \text{and} \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

- What is the probability distribution of $\bar{X}$?

The Sampling Distribution of $\bar{X}$

If the n r.v.s $X_1, X_2, ..., X_n$ are drawn from a Normal distribution, each $X_i \sim Normal(\mu, \sigma^2)$, then

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

The central limit theorem (CLT)

What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?

For i.i.d. r.v.s $X_1, X_2, ..., X_n$ with finite expected value $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$ the sample mean is approximately normal

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right) \quad , n \to \infty$$

The random variable

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

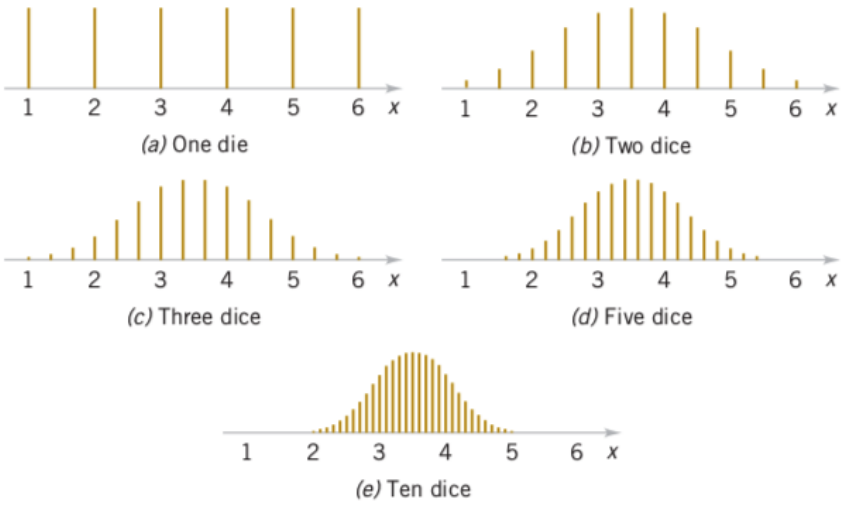
converges in distribution to the standard normal random variable as n goes to infinity:

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

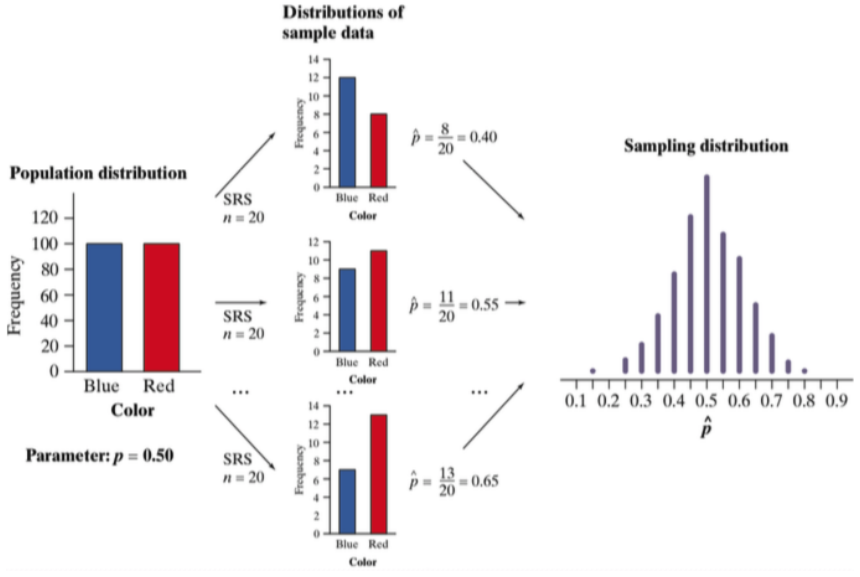
$\Phi(x)$ is the standard normal cdf

- for the CLT is that it does not matter what the distribution of the X_i is
- the distribution can be discrete, continuous, or mixed random variables.

The Sampling Distribution of average scores from throwing dice



The Sampling Distribution of sample proportion



Example

- The shape of the pdf gets closer to the normal pdf as n increases:

Assumptions:

- $X_1, X_2 \ldots$ are iid Uniform(0,1).
- $Z_n = \dfrac{X_1 + X_2 + \ldots + X_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$.

$Z_1 = \dfrac{X_1 - \frac{1}{2}}{\sqrt{\frac{1}{12}}}$

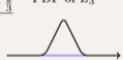
PDF of Z_1

$Z_2 = \dfrac{X_1 + X_2 - 1}{\sqrt{\frac{2}{12}}}$

PDF of Z_2

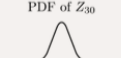
$Z_3 = \dfrac{X_1 + X_2 + X_3 - \frac{3}{2}}{\sqrt{\frac{3}{12}}}$

PDF of Z_3



$Z_{30} = \dfrac{\sum_{i=1}^{30} X_i - \frac{30}{2}}{\sqrt{\frac{30}{12}}}$

PDF of Z_{30}



CLT

How to Apply The Central Limit Theorem (CLT)

Here are the steps that we need in order to apply the CLT:

1. Write the random variable of interest, Y, as the sum of n i.i.d. random variable X_i's:

$$Y = X_1 + X_2 + \ldots + X_n.$$

2. Find EY and $\mathrm{Var}(Y)$ by noting that

$$EY = n\mu, \qquad \mathrm{Var}(Y) = n\sigma^2,$$

where $\mu = EX_i$ and $\sigma^2 = \mathrm{Var}(X_i)$.

3. According to the CLT, conclude that $\frac{Y-EY}{\sqrt{\mathrm{Var}(Y)}} = \frac{Y-n\mu}{\sqrt{n}\sigma}$ is approximately standard normal; thus, to find $P(y_1 \leq Y \leq y_2)$, we can write

$$\begin{aligned} P(y_1 \leq Y \leq y_2) &= P\left(\frac{y_1 - n\mu}{\sqrt{n}\sigma} \leq \frac{Y - n\mu}{\sqrt{n}\sigma} \leq \frac{y_2 - n\mu}{\sqrt{n}\sigma}\right) \\ &\approx \Phi\left(\frac{y_2 - n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{y_1 - n\mu}{\sqrt{n}\sigma}\right). \end{aligned}$$

The service time

A bank teller serves customers standing in the queue one by one. Suppose that the service time X_i for customer i has mean $E(X_i) = 2$ minutes and $Var(X_i) = 1$. We assume that service times for different bank customers are independent. Let Y be the total time the bank teller spends serving 50 customers. Find $P(90 < Y < 110)$

$$Y = X_1 + X_2 + ... + X_n$$

$$P(90 < Y < 110) = P\left(\frac{90 - n\mu}{\sqrt{n}\sigma} < \frac{Y - n\mu}{\sqrt{n}\sigma} < \frac{110 - n\mu}{\sqrt{n}\sigma}\right)$$

By the CLT, $\frac{Y-n\mu}{\sqrt{n}\sigma}$ is approximately standard normal, so

$$P(90 < Y < 110) = P\left(\frac{90 - 50(2)}{\sqrt{50}} < \frac{Y - n\mu}{\sqrt{n}\sigma} < \frac{110 - 50(2)}{\sqrt{50}}\right) = \Phi(\sqrt{2}) - \Phi(-\sqrt{2}) = 0.84$$

How many sandwiches?

You have invited 64 guests to a party. You need to make sandwiches for the guests. You believe that a guest might need 0, 1, or 2 sandwiches with probabilities 0.25, 0.5, and 0.25, respectively. You assume that the number of sandwiches each guest needs is independent from other guests. How many sandwiches should you make so that you are 95% sure that there is no shortage?

Let X_i be the number of sandwiches that the ith person needs, and let $Y = X_1 + X_2 + ... + X_{64}$. Need to find $P(Y \leq y) = 0.95$.

- $E(X_i) = 1$, $E(X_i^2) = 1.5$
- $Var(X_i) = E(X_i^2) - (E[X_i])^2 = 0.5$, $\sigma_{X_i} = \sqrt{Var(X_i)} = 1/\sqrt{2}$
- $EY = 64(1) = 64$, $Var(Y) = 64(0.5) = 32$, $\sigma_Y = 4\sqrt{2}$
- Applying the CLT to find y:

$$0.95 = P(Y \leq y) = P\left(\frac{Y - n\mu}{\sqrt{n}\sigma} < \frac{y - 64}{4\sqrt{2}}\right) = \Phi(\frac{y - 64}{4\sqrt{2}})$$

$$y = 73.3$$

If you make 74 sandwiches, you are 95% sure that there is no shortage.