STATS 8: Introduction to Biostatistics

Estimation

Babak Shahbaba
Department of Statistics, UCI

Parameter estimation

- We are interested in population mean and population variance, denoted as μ and σ^2 respectively, of a random variable.
- These quantities are unknown in general.
- We refer to these unknown quantities as *parameters*.
- We discuss statistical methods for parameter **estimation**.
- Estimation refers to the process of guessing the unknown value of a parameter (e.g., population mean) using the observed data.

Convention

- We use $X_1, X_2, \ldots, X_n$ to denote n possible values of X obtained from a sample randomly selected from the population.
- We treat $X_1, X_2, \ldots, X_n$ themselves as n random variables because their values can change depending on which n individuals we sample.
- We assume the samples are *independent and identically distributed* (IID).
- We use $x_1, x_2, \ldots, x_n$ as the specific set of values we have observed in our sample.
- That is, x_1 is the observed value for X_1, x_2 is the observed value X_2, and so forth.

Point estimation vs. interval estimation

- Sometimes we only provide a single value as our estimate.
- This is called **point estimation**.
- We use $\hat{\mu}$ and $\hat{\sigma}^2$ to denote the point estimates for μ and σ^2.
- Point estimates do not reflect our uncertainty.
- To address this issue, we can present our estimates in terms of a range of possible values (as opposed to a single value).
- This is called **interval estimation**.

Estimating population mean

- Given n observed values, $X_1, X_2, \ldots, X_n$, from the population, we can estimate the population mean μ with the sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- In this case, we say that $\bar{X}$ is an *estimator* for μ.

- The estimator itself is considered as a random variable since it value can change.

- We usually have only one sample of size n from the population $x_1, x_2, \ldots, x_n$.

- Therefore, we only have one value for $\bar{X}$, which we denote $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Law of Large Numbers (LLN)

- The **Law of Large Numbers (LLN)** indicates that (under some general conditions such as independence of observations) the sample mean converges to the population mean ($\bar{X}_n \to \mu$) as the sample size n increases ($n \to \infty$).

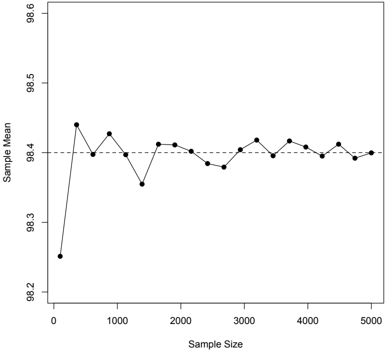
- Informally, this means that the difference between the sample mean and the population mean tends to become smaller and smaller as we increase the sample size.

- The Law of Large Numbers provides a theoretical justification for the use of sample mean as an estimator for the population mean.

- The Law of Large Numbers is true regardless of the underlying distribution of the random variable.

Law of Large Numbers (LLN)

- Suppose the true population mean for normal body temperature is 98.4F.



- Here, the estimate of the population mean is plotted for different sample sizes.

Estimating population variance

- Given n randomly sampled values $X_1, X_2, \ldots, X_n$ from the population and their corresponding sample mean $\bar{X}$, we estimate the population variance as follows:

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

- The sample standard deviation S (i.e., square root of S^2) is our estimator of the population standard deviation σ.

- We regard the estimator S^2 as a random variable.

- In practice, we usually have one set of observed values, $x_1, x_2, \ldots, x_n$, and therefore, only one value for S^2:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Sampling distribution

- As mentioned above, estimators are themselves random variables.
- Probability distributions for estimators are called **sampling distributions**.
- Here, we are mainly interested in the sampling distribution of $\bar{X}$.

Sampling distribution

- We start by assuming that the random variable of interest, X, has a normal $N(\mu, \sigma^2)$ distribution.

- Further, we assume that the population variance σ^2 is known, so the only parameter we want to estimate is μ.

- In this case,

$$\bar{X} \sim N\big(\mu, \sigma^2/n\big).$$

  where n is the sample size.

Sampling distribution

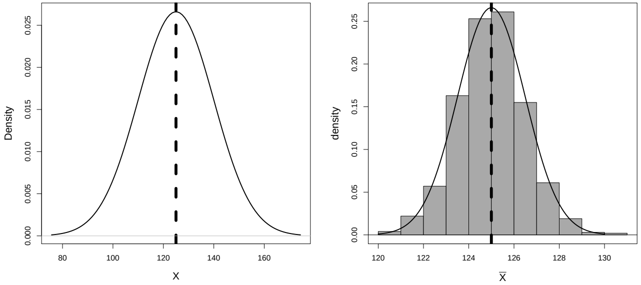


Figure: *Left panel*: The (unknown) theoretical distribution of blood pressure, $X \sim N(125, 15^2)$. *Right panel*: The density curve for the sampling distribution $\bar{X} \sim N(125, 15^2/100)$ along with the histogram of 1000 sample means.

Confidence intervals for the population mean

- It is common to express our point estimate along with its standard deviation to show how much the estimate could vary if different members of population were selected as our sample.

- Alternatively, we can use the point estimate and its standard deviation to express our estimate as a range (interval) of possible values for the unknown parameter..

Confidence intervals for the population mean

- We know that $\bar{X} \sim N(\mu, \sigma^2/n)$.
- Suppose that $\sigma^2 = 15^2$ and sample size is $n = 100$.
- Following the *68–95–99.7%* rule, with 0.95 probability, the value of $\bar{X}$ is within 2 standard deviations from its mean, μ,

$$\mu - 2 \times 1.5 \leq \bar{X} \leq \mu + 2 \times 1.5.$$

- In other words, with probability 0.95,

$$\mu - 3 \leq \bar{X} \leq \mu + 3.$$

Confidence intervals for the population mean

- We are, however, interested in estimating the population mean μ (instead of the sample mean $\bar{X}$).

- By rearranging the terms of the above inequality, we find that with probability 0.95,

$$\bar{X} - 3 \leq \mu \leq \bar{X} + 3.$$

- This means that with probability 0.95, the population mean μ is in the interval $[\bar{X} - 3, \bar{X} + 3]$.

Confidence intervals for the population mean

- In reality, however, we usually have only one sample of n observations, one sample mean $\bar{x}$, and one interval $[\bar{x} - 3,\ \bar{x} + 3]$ for the population mean μ.

- For the blood pressure example, suppose that we have a sample of $n = 100$ people and that the sample mean is $\bar{x} = 123$. Therefore, we have one interval as follows:

$$[123 - 3, 123 + 3] = [120, 126].$$

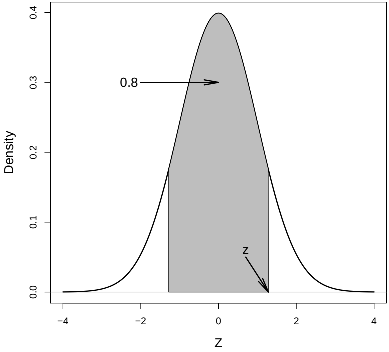
- We refer to this interval as our 95% **confidence interval** for the population mean μ.

- In general, when the population variance σ^2 is known, the 95% confidence interval for μ is obtained as follows:

$$\left[\bar{x} - 2 \times \sigma/\sqrt{n},\ \bar{x} + 2 \times \sigma/\sqrt{n}\right]$$

z critical value

- In general, for the given confidence level c, we use the standard normal distribution to find the value whose upper tail probability is $(1 - c)/2$.



z critical value

- We refer to this value as the z-critical value for the confidence level of c.

- Then with the point estimate $\bar{x}$, the confidence interval for the population mean at c confidence level is

$$[\bar{x} - z_{crit} \times \sigma/\sqrt{n},\ \bar{x} + z_{crit} \times \sigma/\sqrt{n}]$$

- We can use R or R-Commander to find z_{crit}.

Standard error

- So far, we have assumed the population variance, σ^2, of the random variable is known.

- However, we almost always need to estimate σ^2 along with the population mean μ.

- For this, we use the sample variance s^2.

- As a result, the standard deviation for $\bar{X}$ is estimated to be $s/\sqrt{n}$.

- We refer to $s/\sqrt{n}$ as the **standard error** of the sample mean $\bar{X}$.

Confidence Interval When the Population Variance Is Unknown

- To find confidence intervals for the population mean when the population variance is unknown, we follow similar steps as described above, but
  - instead of $\sigma/\sqrt{n}$ we use $SE = s/\sqrt{n}$,
  - instead of z_{crit} based on the standard normal distribution, we use t_{crit} obtained from a t-distribution with $n-1$ degrees of freedom.
- The confidence interval for the population mean at c confidence level is

$$\left[\bar{x} - t_{\text{crit}} \times s/\sqrt{n},\ \bar{x} + t_{\text{crit}} \times s/\sqrt{n}\right],$$

Central limit theorem

- So far, we have assumed that the random variable has normal distribution, so the sampling distribution of $\bar{X}$ is normal too.

- If the random variable is not normally distributed, the sampling distribution of $\bar{X}$ can be considered as *approximately* normal using (under certain conditions) the **central limit theorem** (CLT):

  > *If the random variable X has the population mean μ and the population variance σ^2, then the sampling distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n.*

- Note that CLT is true regarding the underlying distribution of X so we can use it for random variables with Bernoulli and Binomial distributions too.

Confidence Interval When for the Population Proportion

- For binary random variables, we use the sample proportion to estimate the population proportion as well as the population variance.

- Therefore, estimating the population variance does not introduce an additional source of uncertainty to our analysis, so we do not need to use a t-distribution instead of the standard normal distribution.

- For the population proportion, the confidence interval is obtained as follows:

$$[p - z_{\text{crit}} \times SE, p + z_{\text{crit}} \times SE],$$

where

$$SE = \sqrt{p(1-p)/n}.$$