# Probability Theory Lecture 13

Padmini Mukkamala

Budapest University of Technology and Economics

December 6, 2020

# Overview

# Point Estimators

Consider the following examples:

- Given a sample of 100 people and what they voted for during elections, we want to predict what percentage of the population voted for Candidate 1.
- Given a sample of 50 cola bottles, we want to estimate the average volume of cola in the advertized 1.75 liter bottle.
- Computer chip manufactures have very stringent size requirements. With a sample of 100 chips we want to estimate the average length of a certain type of chip along with an estimate for the standard deviation from this mean.

# Point Estimators cont.

In all the above cases, there was a parameter $\theta$ of a random variable $X$ to be determined, and we took a random sample and tried to estimate it. In the previous examples, $\theta$ was average percentage of people voting for candidate 1, average volume of cola in bottles, the average length and standard deviation of the length of a chip.
In each of these cases, the random variable $X$ is if vote of a person is for candidate 1 or not, volume of cola in a bottle, length of chip.
We can see that what we are trying to estimate is $\theta = E(X)$ in the first two cases and in the last case we also want to additionally estimate $\theta = \sigma_X$.

# Point Estimators cont.

The first step to estimation is collecting a sample. Let $X_1, .., X_n$ be $n$ randomly and independently chosen samples. In the above examples it is votes of randomly chosen $n$ people, volume of cola in randomly chosen $n$ bottles and length of $n$ randomly chosen chips. We use the sample to find an estimate for $\theta$, notice that we could estimate it as the mean $\hat{\theta} = \frac{X_1 + ... + X_n}{n}$ and we can find the standard deviation of this sample to estimate the standard deviation in the third example.
Here $\hat{\theta}$ is called the **Point estimate** for the population parameter $\theta$.

# Interval Estimates

The point estimate is a single value estimate. Sometime, we prefer to give an interval in which we think $\theta$ would lie in. This is called the **Interval estimate**. Also, we would mention the probability that $\theta$ will be in this interval. This probability is termed as the **Confidence level** of our interval estimate.

# Interval Estimates - Sigma known

If we knew the standard deviation $\sigma$ of $X_i$, then for our sample $X_1, ..., X_n$, if our point estimate was $\hat{\theta} = \frac{\sum X_i}{n}$, then we can use CLT to conclude that $\hat{\theta}$ must have the distribution $N(\theta, \frac{\sigma^2}{n})$.
Since $\Phi(1.65) = 0.95$, so we can say that probability that $\hat{\theta}$ is in the interval $[\theta - 1.65\frac{\sigma}{\sqrt{n}}, \theta + 1.65\frac{\sigma}{\sqrt{n}}]$ is 90%.
Turning this argument on its head, we can see that a 90% confidence interval for $\theta$ is $[\hat{\theta} - 1.65\frac{\sigma}{\sqrt{n}}, \hat{\theta} + 1.65\frac{\sigma}{\sqrt{n}}]$ is 90%.

# Interval Estimates - Sigma known

Summary: Let $\theta$ be the expected value of a population parameter that we want to estimate. Let a sample of $n$ items have mean $\hat{\theta}$. Suppose we know from previous data that the standard deviation of the population is $\sigma$. Let the confidence level we require of our interval estimate be $1-\alpha$, where $0 < \alpha < 1$. Further let $c$ be a constant such that $\Phi(c) = 1 - \frac{\alpha}{2}$, where $\Phi$ is the CDF of $N(0,1)$. Then, the confidence interval estimate for $\theta$ with $1-\alpha$ confidence is given by

$$[\hat{\theta} - c\frac{\sigma}{\sqrt{n}}, \hat{\theta} + c\frac{\sigma}{\sqrt{n}}]$$

# Interval Estimates - Sigma known

Example: Suppose that for a sample of 100 bottles of cola, the sample average volume was 1.77 litres. If we know from past data that the standard deviation is 0.2L, give a 90% confidence interval for the volume of cola bottles.
Using what we did before, it will be $[1.737, 1.803]$.

# Interval Estimates - Sigma unknown

When $\sigma$ is not known, then we use the sample standard deviation. But in this case, we cannot apply CLT. But still, with some considerable analysis, it can be shown that the point estimate has a **Student's t-distribution**. Then these tables are used in a similar manner as before to find the confidence interval.
The "Student" in the Student's t-distribution is William Sealy Gosset.

# Hypothesis Testing

In many situations, we have to evaluate a statement about some quality or production parameter, and our answer must be of the form true or false. For the previous examples, our question could be, 1- does the first candidate have more than half the votes, 2- are cola bottles filled to a minimum of 1.75L volume, and 3- are the chip lengths exactly 1mm. These statements which evaluate to true or false are called **Hypothesis**.

# Hypothesis Testing cont.

In statistical analysis, it is frequent to call the hypothesis we believe to be true as **Null Hypothesis** denoted by $H_0$, while its oppposite or contrary statement as **Alternative Hypothesis**, denoted by $H_1$. The Null hypothesis is one of the following three,
$H_0 : \mu \leq \mu_0$ or $\mu \geq \mu_0$ or $\mu = \mu_0$,
where $\mu_0$ is the threshold we desire, e.g. in the case of volume of cola bottles, we want them to have volume at least 1.75, so our Null hypothesis $H_0$ will be $\mu > 1.75$.

# Hypothesis Testing cont.

Our idea is to find first an interval $[\mu_0 - a, \mu_0 + a]$. We will call this the acceptance interval. If the sample mean $\frac{\sum X_i}{n}$ is in this interval, we will accept the null hypothesis, otherwise we will reject it.
We identify the error as **Type I** error if we rejected $H_0$, but it was true. Often in hypothesis testing, we want to base our acceptance interval such that we can put a limit on the probability of a Type I error.

# Hypothesis Testing cont.

The **significance Level** $\alpha$ of a Hypothesis test is defined as the maximum probability of a Type I error.
Consider $H_0 : \mu \geq \mu_0$.
Here, Type I error happens if the sample mean is not in the acceptance interval, but the population mean is greater than $\mu_0$.
For example, if the sample mean of the cola bottles was 1.6 but the population mean is 1.78.

# Hypothesis Testing cont. - sigma known

Let $\alpha = 0.05$, then, in the Cola example, we want the probability of error to be atmost 0.05. For the worst case scenario, let us assume that the population mean was exactly 1.75. Then using CLT, we can find the cut off $a$ so that the probability that a sample of 100 bottles will fail the hypothesis is exactly 0.05. Since $\Phi(0.95) = 1.65$, so we know that the cutoff $a = 1.65\frac{\sigma}{\sqrt{n}}$. Let us assume once again that $\sigma = 0.2$. Then, we get that we should reject the null hypothesis if the sample mean is less than $1.75 - 1.65\frac{0.2}{10} = 1.717$.
So with a cut-off of 1.717, I know that if the sample mean is less than 1.717 and I reject the hypothesis, my probability of error is at most 5%.

# Hypothesis Testing cont. - sigma known

Summary: Let $H_0 : \mu \geq \mu_0$ be a null hypothesis and let $\alpha$ be the significance level. Let the population standard deviation be given as $\sigma$ and further let our sample consist of $n$ items whose mean is $\mu$. Let $c$ be a constant such that $\Phi(c) = 1 - \alpha$, where $\Phi$ is the CDF of $N(0, 1)$. Then, we reject the Null Hypothesis if

$$\mu < \mu_0 - c\frac{\sigma}{\sqrt{n}}$$

Then our probability of Type I error is at most $\alpha$.

# Hypothesis Testing cont. - sigma unknown

When $\sigma$ is not known, then we use the sample standard deviation. Again here we use the **Student's t-distribution** in the place of standard normal distribution.
The "Student" in the Student's t-distribution is William Sealy Gosset.

Thank you for your attention! Wish you Good Luck in all future endeavors!