STATISTICS

INTRODUCTION TO INFERENTIAL STATISTICS

Instructor:

Prof. Dr. Doğan Nadi LEBLEBİCİ

Source: Kaplan, Robert M. Basic Statistics for the Behavioral Sciences, Allyn and Bacon, Inc., Boston, 1987. SENTENCES IN THIS POWER POINT PRESENTATION ARE USUALLY BORROWED FROM KAPLAN'S BOOK.

INFERENTIAL STATISTICS

A Hypothetical Study of the Effects of a New Arthritis Drug upon Daily Functioning: Does the Drug Work?

DRUG GROUP				**PLACEBO GROUP**			
Patient	**Function**			**Patient**	**Function**		
	Before Treatment	After Treatment	Change		Before Treatment	After Treatment	Change
Ayşe	5	6	+1	Saliha	3	3	0
Fatma	4	6	+2	Umut	4	4	0
Nurcan	3	4	+1	Gülcan	5	8	+3
Ali	5	5	0	Neşet	5	6	+1
Mehmet	5	4	- 1	Münevver	6	4	- 2
Rıfat	6	8	+2	Hasan	5	5	0
			μ=.83 ***S=1.17***				***μ=.33*** ***S=1.63***

DECISION RULES

A decision is a choice among several alternatives. Scientists must make decisions about the meaning of scientific observation in systematic way. There are rules for doing this.

Samples and Populations

A population is a set of all possible observations of a specific type.

A sample is a subset of observations drawn from the population.

A Sample must be representative of the population.

Rules and Decisions

Two principles dominate in criminal proceedings:

1. The accused is innocent until proven guilty.

2. A conviction must be based on evidence that is beyond a reasonable doubt.

INTRODUCTION TO INFERENTIAL STATISTICS

Rules and Decisions

Scientific evidence is based on similar principles:

1. In comparison of two sample groups, it is assumed that groups do not differ until there is substantial evidence that they are not the same. The assumption that the groups do not differ is called the *null hypothesis*.

2. A null hypothesis is assumed to be correct until we have evidence beyond a reasonable doubt that it is incorrect. Under these circumtances, we reject the null hypothesis in favor of an *alternative hypothesis* that states that the observed differences are large enough that is unlikely they occured by chance.

INTRODUCTION TO INFERENTIAL STATISTICS

Rules and Decisions / Sampling Error

Select a sample from a population and calculate the mean, and then, repeat it for several times. It is likely that means of each experiment will be different.

INTRODUCTION TO INFERENTIAL STATISTICS

Parameters and Statistics

Population include every member of a defined class. The mean and the standard deviation of a population are referred to as *parameters*.

A subset of the population is a *sample*. Means, standard deviations, and other values that describe characteristics of the sample are known as *statistics*.

INTRODUCTION TO INFERENTIAL STATISTICS

Parameters and Statistics

	POPULATION	SAMPLE
DEFINITION	**All elements with same definition.**	**A subset of the population usually drawn to represent it in an unbiased fashion.**
Descriptive characteristics	**Parameters**	**Statistics**
Symbols for mean	**μ**	$\overline{X}$
Symbol for standard deviation	**σ**	***S***

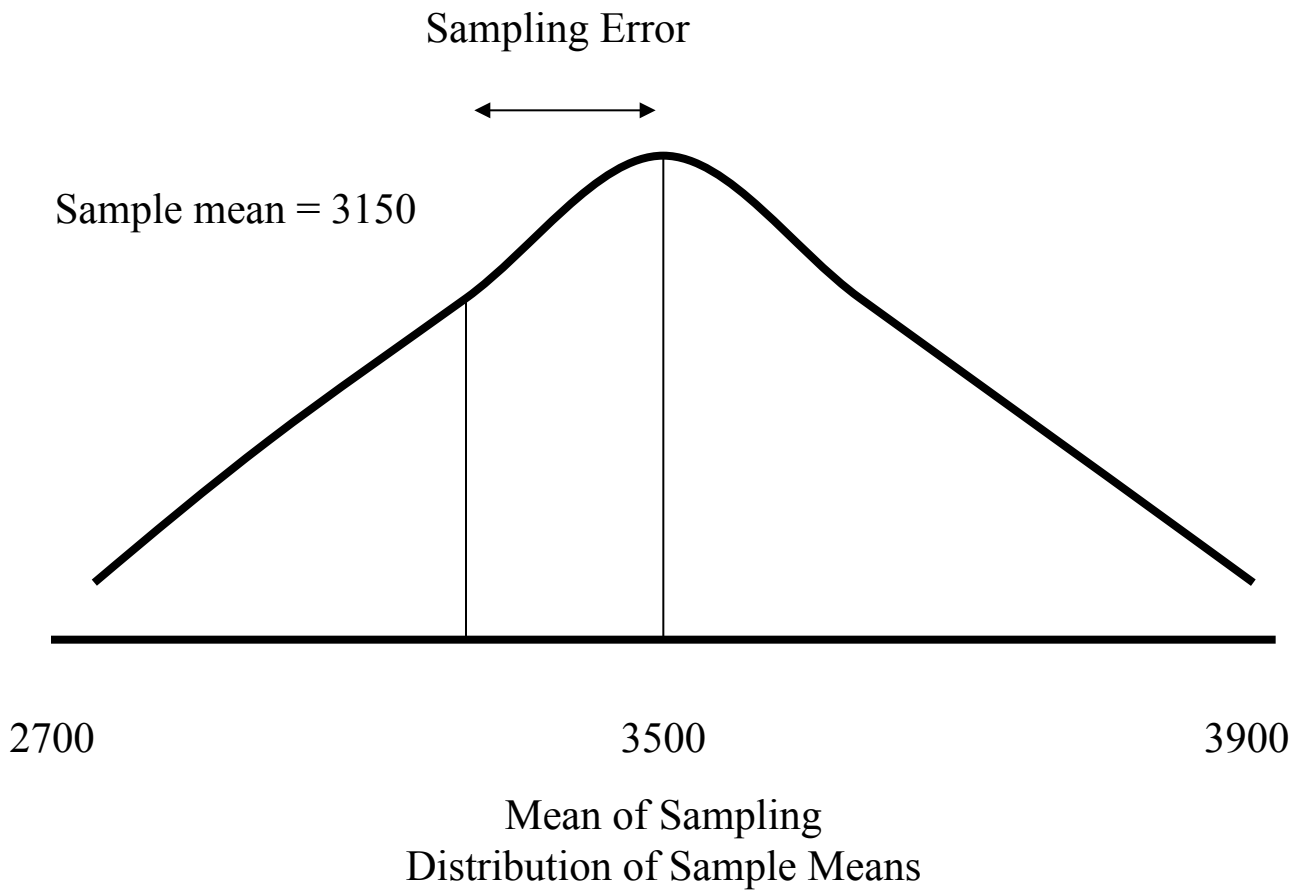
INTRODUCTION TO INFERENTIAL STATISTICS

Sampling Distributions

Sampling distribution is defined as the theoretical probability distribution of values that could be obtained from some statistic in random samples where a particular sample size is taken from a population.

The difference between any sample mean and the mean of the sampling distribution is known as sampling error.

Sampling Distributions



Sampling Distributions

In the real world, it is difficult and very expensive to measure every member of the population. Instead we use a sample mean to estimate the population mean. However, sample mean may not be exactly the same as the population mean. We could close it by repeating random sampling of a particular size and use them to create a sampling distribution. The mean of this sampling distribution would be an unbiased estimate of population mean.

Sampling Distributions

Nevertheless, it still is difficult and very expensive as measuring every member of the population. Thus in the real world we most often only one sample mean.

The first step in estimating how well a sample mean (which is statistic) represents a population mean (which is a parameter) is to find standard error of the mean. This is an estimate of the standard deviation of the sampling distribution.

Sampling Distributions

The formula for the standard error of the mean is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \Rightarrow S_{\bar{X}} = \frac{S}{\sqrt{N}}$$

INTRODUCTION TO INFERENTIAL STATISTICS
Sampling Distributions

Babies	Weight	Sample	$\overline{X}$		$(X-\overline{X})^2$	X²				
Ceren	8,000	8,000	**4,600**	3,400	11,56	64,000				
Aysu	8,000	8,000	**4,600**	3,400	11,56	64,000				
Cenk	3,000	7,000	**4,600**	2,400	5,76	49,000				
Tuna	7,000	6,000	**4,600**	1,400	1,96	36,000				
Ongun	6,000	5,000	**4,600**	0,400	0,16	25,000				
Mert	5,000	4,000	**4,600**	-0,600	0,36	16,000				
Suzi	5,000	4,000	**4,600**	-0,600	0,36	16,000				
Muzi	5,000	4,000	**4,600**	-0,600	0,36	16,000				
Fuzi	4,000	0,000	**4,600**	-4,600	21,16	0,000				
Dizy	9,000	0,000	**4,600**	-4,600	21,16	0,000				
Mary	4,000			$\sum(X-\overline{X})^2$	74,4	286				
Obua	0,000									
Logan	0,000									
Ayşe	6,000									
Fatma	5,000									
Nihal	4,000									
Ahmet	7,000									
Mehmet	3,000	46,000								
Mean	**4,944**	**4,600**								

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = 2,875 \quad \text{St.Deviation}$$

$$S_{\overline{x}} = \frac{S}{\sqrt{N}} = 0,91 \quad \text{St. ERROR}$$

Confidence Intervals and Confidence Limits

By calculating the standard error, we can estimate how far the sample mean is from population mean. We may guess that sample mean is most probably near to the population value. It is possible that sample mean is an overestimate or an underestimate of the population mean. The ranges that are likely to capture the population mean are called confidence intervals. Confidence intervals are bounded by confidence limits.

A confidence interval is defined as a range of values with a specified probability of including the population mean. A confidence interval is typically associated with certain probability level.

Confidence Intervals and Confidence Limits

For example 95 % confidence interval has 95 % chance of including the population mean. A 99 % confidence interval is expected to capture the true mean in 99 of each 100 cases.

Confidence limits are defined as the values or points that bound the confidence interval. The task of defining this interval requires that we first obtain a sample mean and then define an interval around it that most probably captures the population mean.

INTRODUCTION TO INFERENTIAL STATISTICS

Confidence Intervals and Confidence Limits

We use the fallowing formula to calculate the confidence interval:

$$CI = \overline{X} \pm Z_{\alpha} S_{\overline{X}}$$

Z_{α} is the Z-score for a certain probability and S_x is the standard error of the mean.

INTRODUCTION TO INFERENTIAL STATISTICS

Confidence Intervals and Confidence Limits

For our example, the confidence interval for 95 % level is:

$$CI = 4{,}60 \pm 1{,}96 x 0{,}91$$

CI = 4,60±1,78 => Upper CL is 6,38
Lower CL is 2,82

INTRODUCTION TO INFERENTIAL STATISTICS

Confidence Intervals and Confidence Limits

For our example, the confidence interval for 60 % level is:

$$CI = 4{,}60 \pm 0{,}84x0{,}91$$

CI = 4,60±0,76 => Upper CL is 5,36
Lower CL is 3,84

Confidence Intervals and Confidence Limits

Many years ago, mathematicians proved that the sampling distribution of sample means is a normal distribution. This is true whether or not the population from which the samples are drawn is normally distributed. Staticians have proved that as the sample size increases, the more likely it is that the sampling distribution will be normal. In other words, as greater number of sampling units are included, the greater is the likelihood that the distribution will approach normality. Further, the distribution of sample means will be normal even though the population distribution may not be normal. This is known as the Central Limit Theorem. For sample smaller than 30, some assumptions about this normality of sampling distribution are not appropriate. Therefore, for samples less than 30 it is better to use *t* values (will be discussed later).

INTRODUCTION TO INFERENTIAL STATISTICS

Confidence Intervals and Confidence Limits

We use the fallowing formula to calculate the confidence interval with *t* values:

$$CI = \overline{X} \pm t_{\alpha} S_{\overline{X}}$$

t_{α} is a score for a certain probability like *Z*-score and S_x is the standard error of the mean.

INTRODUCTION TO INFERENTIAL STATISTICS

Confidence Intervals and Confidence Limits

For our example, the confidence interval for 95 % level (for two tailed test) is:

$$CI = 4{,}60 \pm 2{,}262x0{,}91$$

CI = 4,60±2,06 => Upper CL is 6,66

Lower CL is 2,54