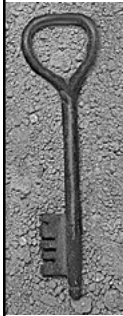


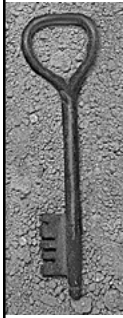


Probability Distributions & Central Limit Theorem



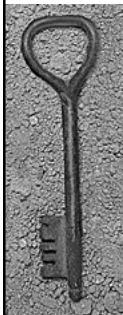
Review

- ◆ Probability:
 - discrete vs. continuous variables
- ◆ Probability Distribution
- ◆ Normal Distribution
 - Z-score \leftrightarrow probability
 - Standard normal distribution



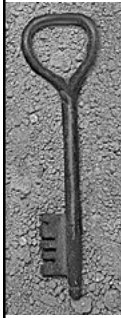
Today's topic

- ◆ Probability and inference
- ◆ The Central Limit Theorem



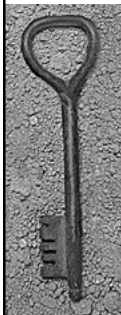
Probability and Inference

- ◆ The link between normal distributions and probabilities allows us to draw conclusions
- ◆ Assumption: known μ , σ
- ◆ Problem: we often do not have all information about the whole population, thus we usually do not know μ , σ



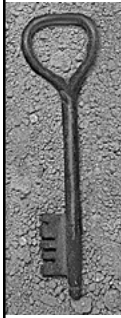
Probability and Inference

- ◆ But we wish to describe and understand large sets of people (or organizations or countries)
 - School achievement of American teenagers
 - Fertility of individuals in Indonesia
 - Behavior of organizations in the auto industry
- ◆ Problem: It is seldom possible to collect data on all relevant people (or organizations or countries) that we hope to study



Probability and Inference

- ◆ How can we calculate the mean or standard deviation for a population, without data on most individuals?
 - Without even knowing the total N of the population?
- ◆ IDEA: Maybe we can gain some understanding of large groups, even if we have information about only some of the cases within the group
 - We can examine part of the group and try to make intelligent guesses about what the entire group is like

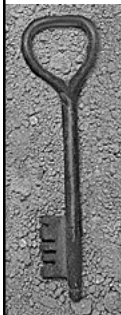


Sample and Population

- ◆ Population: The entire set of persons, objects, or events that have at least one common characteristic of interest to a researcher
 - Voting age Americans (their political views)
 - 6th grade students attending a particular school (their performance on a math test)
 - People (their response to a new AIDS drug)
 - Small companies (their business strategies)
- ◆ Sample: a subset of the population
 - Dataset is usually a sample
 - Many samples can be drawn from a population

GOG 502/PLN 504 Youqin Huang

7

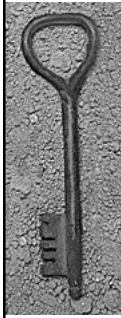


Statistical Inference

- ◆ Making statistical generalizations about a population from evidence contained in a sample
- ◆ When is statistical inference likely to work?
 - 1. When a sample is large
 - If a sample approaches the size of the population, it is likely to be a good reflection of that population
 - 2. When a sample is representative of the entire population
 - As opposed to a sample that is atypical in some way, and thus not reflective of the larger group
 - Use appropriate sampling design to avoid bias.

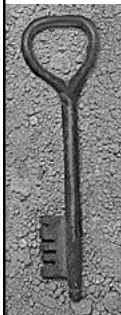
GOG 502/PLN 504 Youqin Huang

8



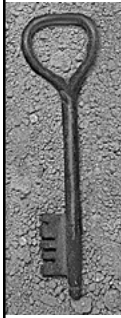
Biased Samples: Examples

- ◆ Biased samples can lead to false conclusions about characteristics of populations
- ◆ What are the problems with these samples?
 - Internet survey asking people the number of CDs they own (population = all Americans)
 - Telephone survey conducted on election day (pop = voting age Americans)
 - Survey of this class (population = UAlbany students, SUNY students, or all college students in the U.S.)



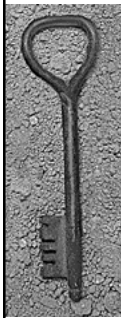
Statistical Inference

- ◆ Statistical inference involves two tasks:
 - 1. Using information from a sample to estimate properties of the population
 - 2. Using laws of statistics and information from the sample to determine how close our estimate is likely to be
 - We can determine whether or not we are confident in our assessment of a population



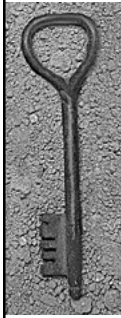
Statistical Inference: Example

- ◆ Population: Students in the United States
- ◆ Sample: Individuals in this classroom
- ◆ Question: What is the mean number of CDs owned by students in the U.S.?
 - Goal #1: Use information on students in this class to guess the mean number of CD's owned by students in the U.S.
 - Goal #2: Try to determine how close (or far off) our estimate of the population mean might be. Estimate the quality of the guess.
- ◆ Goal #2 prevents us from drawing inappropriate conclusions from #1



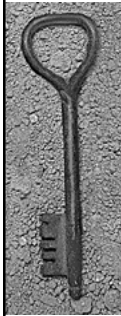
Jargons: Population vs. Sample

	Population	Sample
Characteristics	“parameters”	“statistics”
Characteristics are:	constant (one for population)	variables (different for each sample)
Notation	Greek (μ , σ)	Roman (, s)
Estimate	“hat”:	“point estimate” based on sample



Population and Sample

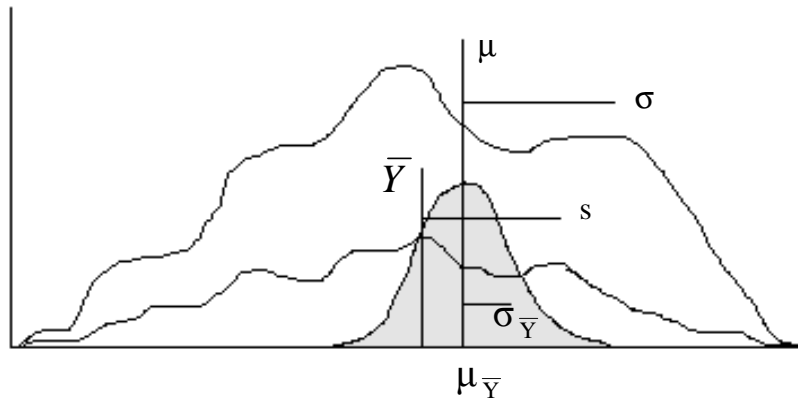
- ◆ Population parameters (μ , σ) are constants
 - There is one true value, but it is unknown
- ◆ Sample statistics (\bar{Y} , s) are variables
 - Up until now we've treated them as constants
 - There are many possible samples, and thus many possible values for each
 - In fact, the range of possible values makes up a distribution – the “sampling distribution”



Three Distributions

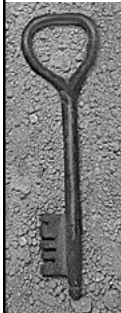
- ◆ Population Distributions:
 - “Unknown” distribution with unknown μ , σ or even N
- ◆ Sample Distribution:
 - Distribution of the data we observe (existing dataset)
 - Known mean (\bar{Y}) and s .
 - The larger the n is, the more the sample distribution resembles the population distribution, the closer the sample statistics (e.g \bar{Y}) fall to the population parameters (e.g. μ)
- ◆ Sampling Distribution
 - the distribution of estimates (e.g. \bar{Y}) created by taking all possible unique samples (of a fixed size) from a population
 - it allows one to get a sense of the range of estimates of the parameter

Three Distributions



GOG 502/PLN 504 Youqin Huang

15

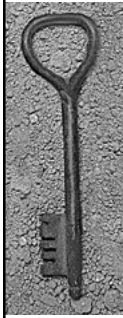


Sampling Distribution

- ◆ BUT, we rarely take many samples. We usually take only ONE sample.
- ◆ Sampling distribution is important because it allows us to compare sample statistics from SINGLE sample with their relevant sampling distribution, which allows us to make conclusions about the unknown population parameter (point and interval estimation)
- ◆ It turns out that that under some circumstances, the shape of the sampling distribution can be determined, which provides insight into the population mean and standard deviation
- ◆ So, even if there is only one sample, we may still be able to make conclusions about unknown population parameters.

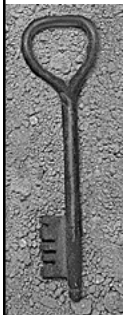
GOG 502/PLN 504 Youqin Huang

16



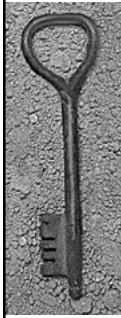
Four Key Sampling Distributions

- ◆ Standard normal (z)
 - ◆ Student's (t)
 - ◆ chi-square ()
 - ◆ Fisher's (F)
-
- ◆ Corresponding tables in appendix



Sampling Distribution Notation

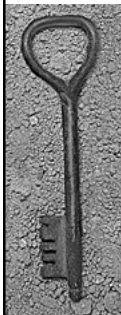
- ◆ Population mean and S.D. are: μ, σ
- ◆ Each sample has a mean and S.D.: \bar{Y}, s
- ◆ The sampling distribution of the mean (i.e., the distribution of mean-estimates) also has a mean and a S.D., aka the “standard error” $\mu_{\bar{Y}} \quad \sigma_{\bar{Y}}$
 - They are Greek because all possible samples represent a population
 - Using sub- \bar{Y} to indicate they are the mean and s.d of all possible \bar{Y} -bars



Example: Estimating the Mean

- ◆ Goal: the mean of a population (μ).
 - Plan A: Spend \$100 million to survey the entire population, if possible.
 - Plan B: Spend \$1,000 sampling a few hundred people.
- ◆ Estimate the mean

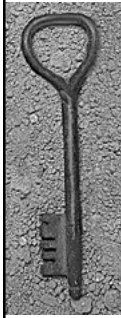
- ◆ How good is the estimate?
 - Using known sampling distribution to determine how likely the estimate is good.



The Central Limit Theorem

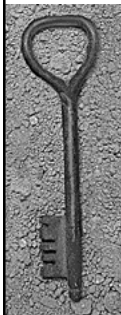
- ◆ The basis for drawing statistical inferences about the mean

- ◆ “If a population has a mean μ , and s.d. σ , the sampling distribution of the sample mean approaches a normal distribution with mean μ and s.e. σ divided by square root of n (), as the sample size n increases.”



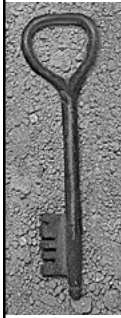
The Central Limit Theorem

- ◆ 1. As n becomes large, the sampling distribution of the mean approaches normality
 - Even if the population is not normally distributed!
- ◆ 2. The mean of the sampling distribution is the same as the population mean
- ◆ 3. The standard deviation of the sampling distribution (aka standard error) is equal to the standard deviation of population divided by the square root of n



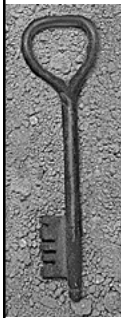
The Central Limit Theorem

- ◆ More simply, the sampling distribution of the mean (and thus all possible estimates of the mean) cluster around the true population mean
 - ◆ They cluster as a normal curve
 - ◆ The estimates are dispersed around the population mean by a knowable standard deviation (σ over root n)



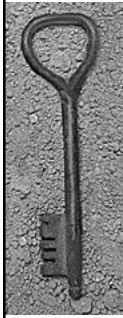
The Central Limit Theorem

1. As n grows large, the sampling distribution of the sample mean (\bar{Y}) approaches normality
- 2.
- 3.



The Central Limit Theorem

- ◆ The sampling distribution of mean becomes more “normal” as n increases ($n > 30$ is often sufficient)
- ◆ The sampling distribution of mean get narrower as n increases
- ◆ Enable us to make inferences using probability theory and properties of normal distribution even when the population distribution is highly irregular



Central Limit Theorem: Visually

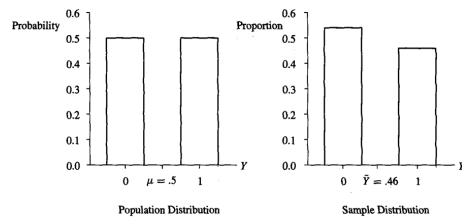
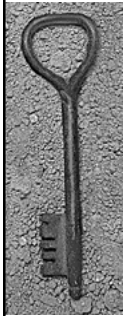
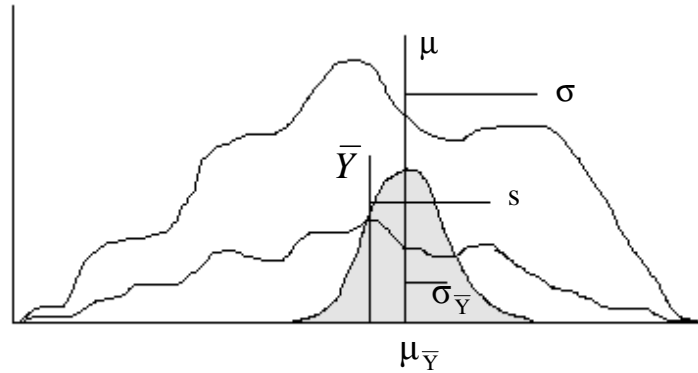


Figure 4.17 The Population ($N = 4$ million) and Sample ($n = 100$) Distributions of Candidate Preference

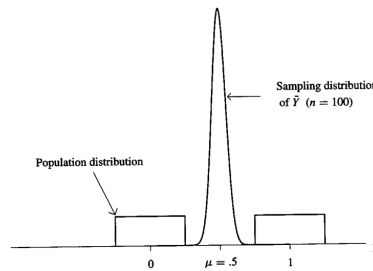
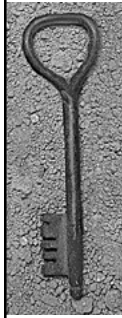
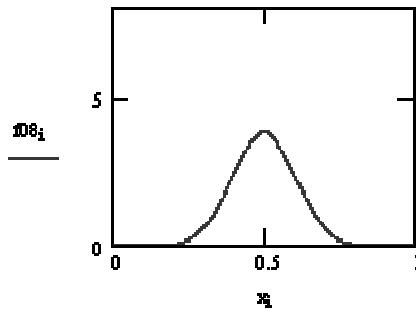


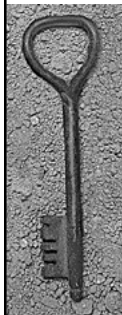
Figure 4.18 The Population Distribution and the Sampling Distribution of \bar{Y} for $n = 100$



Change in sampling distribution as the sample size increases for a uniform population distribution:

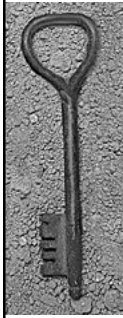


Distribution of Xbar when sample size is 8



Why is C.L.T. Important?

- ◆ We know that the mean from the sample falls somewhere in the sampling distribution
 - Which has mean μ , standard deviation
 - We don't know exactly where
- ◆ If we can estimate σ , we can estimate $\sigma_{\bar{x}}$, which indicates how dispersed mean estimates are, and the range by which a mean estimate is likely to miss.
- ◆ The larger is the sample, the smaller is the range and the better is the estimate.



Summary

- ◆ Probability and statistical inference
- ◆ Some statistical notations
- ◆ Three distributions
 - Population distribution
 - Sample distribution
 - Sampling distribution
- ◆ The Central Limit Theorem