Statistics for Computing

Lecture 6A

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn Semester 2013

Statistical Inference

- Statistics and Population parameters
- Random Sampling
- Properties of Estimators
- Estimation (Point and Interval)
- Confidence Intervals
- The Central Limit Theorem
- Standard Errors

Statistical Inference : Definitions

- A ***population*** consists of an entire set of objects, observations, or scores that have something in common. For example, a population might be defined as students in a university.
- Some populations are only hypothetical. Consider an experiment where a die is thrown 100 times and the sum of the scores was recorded.
- The researcher might define a population as the sums that would result if this experiment was repeated an infinite number of times.
- The population is hypothetical in the sense that it is not reasonable to repeat this experiment indefinitely.
- The distribution of a population can be described by several parameters such as the mean and standard deviation.

Statistical Inference : Sample

- A sample is a subset of a population.
- Suppose we are interested in some characteristic of a population (e.g. amount of time spent on the internet)
- Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available.
- To be properly representative of a population, a sample should be both ***random*** and sufficiently large.

Statistical Inference : Random Sampling

- In random sampling, each item or element of the population has an equal chance of being chosen at each draw.
- A sample is random if the method for obtaining the sample meets the criterion of randomness (each element having an equal chance at each draw).

Statistical Inference : Biased Sampling

- A biased sample is one in which the method used to create the sample results in samples that are systematically different from the population.
- For instance, consider a market research project on attitudes of attendees towards an event they attended.
- Collecting the data by publishing a questionnaire and asking people to fill it out and send it in would produce a biased sample.
- People interested enough to spend their time and energy filling out and sending in the questionnaire are likely to have different attitudes about the event than those not taking the time to fill out the questionnaire.

Statistical Inference : Parameters

- A parameter is a numerical quantity measuring some aspect of a population of scores.
- The population mean μ and population variance σ^2 are commonly used parameters.
- Another commonly used parameter is the population proportion π.
- (Remark : greek letters are used to designate parameters.)
- Parameters are rarely known and are usually estimated by ***statistics*** computed from samples.

Statistical Inference : Statistics

- The most common use of the word ‘statistics’is for describing a wide range of techniques and procedures for analyzing, interpreting and displaying data.
- In a second usage, a “statistic” is defined as a numerical quantity (such as the sample mean) calculated from a sample.
- Sample mean $\bar{x}$ and sample standard deviation s are types of statistics.
- These statistics are used to estimate population parameters.

Statistical Inference : Estimators

- Three important attributes of statistics as estimators are:
    - unbiasedness,
    - consistency,
    - relative efficiency.
- A statistic is unbiased if, in the long run, it's value is reasonably close to the parameter it is estimating.
- An estimator is consistent if the estimator tends to get closer to the parameter it is estimating as the sample size increases.

Statistical Inference : Estimators

- The efficiency of a statistic is the degree to which the statistic is stable from sample to sample.
- That is, the less subject to sampling fluctuation a statistic is, the more efficient it is.
- ***Sampling fluctuation*** refers to the extent to which a statistic takes on different values with different samples.
- That is, it refers to how much the statistic's value fluctuates from sample to sample.

Statistical Inference : Inferential Statistics

- Inferential statistics are used to draw inferences about a population from a sample.
- Consider an experiment in which 10 subjects who performed a task after 24 hours of sleep deprivation scored 12 points lower than 10 subjects who performed after a normal night's sleep.
- Is the difference real or could it be due to chance?
- How much larger could the real difference be than the 12 points found in the sample?
- These are the types of questions answered by inferential statistics.

Statistical Inference : Estimation

- When a parameter is being estimated, the estimate can be either a single number or it can be a range of numbers.
- When the estimate is a single number, such as a sample mean, the estimate is called a ***point estimate***.
- When the estimate is a range of values, the estimate is called an ***interval estimate***.
- ***Confidence intervals*** are used for interval estimation.
- As we will soon see, point estimates are not usually as informative as confidence intervals.

Statistical Inference : Confidence Intervals

- Confidence intervals allow us to use sample data to estimate a parameter value, such as a population mean.
- A confidence interval is a range of values for which we can be confident (at a specific level) that parameter value (such as the population mean) lies within.
- A confidence level will have a specified level of confidence, commonly 95%.
- The 95% confidence interval is a range of values which contains the parameter value of interest with a probability of 0.95.
- We can expected that a 95% confidence interval will not contain the parameter value of interest with a probability of 0.05.

Statistical Inference : Confidence Intervals

- It is natural to interpret a 95% confidence interval on the mean as an interval with a 0.95 probability of containing the population mean.
- However, the proper interpretation is not that simple.
- Consider the case in which 1,000 studies estimating the value of μ in a certain population all resulted in estimates between 30 and 40.
- Suppose one more study was conducted and the 95% confidence interval on μ was computed to be $40 \leq \mu \leq 50$ (based on that one study).
- The probability that μ is between 40 and 50 is very low, the confidence interval not withstanding.

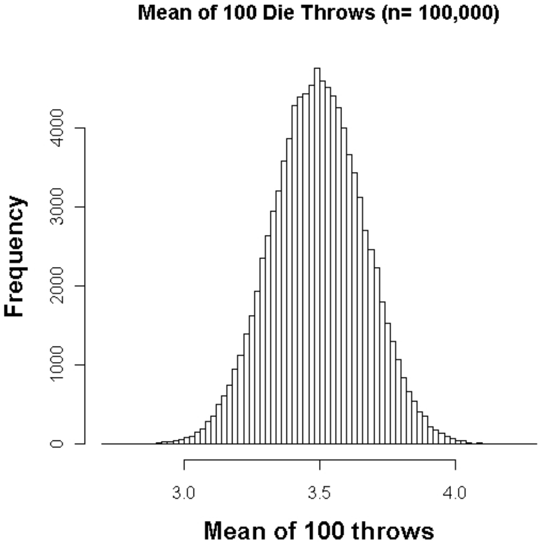
Central Limit Theorem

- Before we can begin computing confidence intervals, we must introduce the ***Central Limit Theorem***.
- Suppose random sample of size n are drawn from any distribution, with the distribution having a mean of μ (equivalently $E(X)$) and variance of σ^2 (i.e. standard deviation of σ).
- Also suppose that the sample size is large (i.e. $n > 30$).
- The sample means tend to form a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- We call the standard deviation of the sample means the ***standard error***
- Standard error is commonly denoted as $S.E.$

Central Limit Theorem

- Recall from earlier lectures, an experiment was carried out where the sum of 100 throws of a die were recorded.
- The underlying distribution of the die values is not normally distributed. (Actually discrete uniform between 1 and 6.)
- Nonetheless the distribution of the sum of 100 throws was normally distributed. Necessarily the distribution of the average score for 100 throws is normally distributed.

Distribution of means



Exercise

From previous lecture, we know the following properties of the dice distribution.
(Remark: In this case we know the variance, but that is not always the case.)

- Mean (Expected Value) $E(X) = \mu = 3.5$
- Variance $V(X) = \sigma^2 = 2.9166$
- Standard deviation $= \sigma = 1.707$

Compute the standard error $S.E.(\bar{x})$ for the mean value $\bar{x}$ of die values:

- when the die is thrown 25 times
- when the die is thrown 225 times.

Exercise

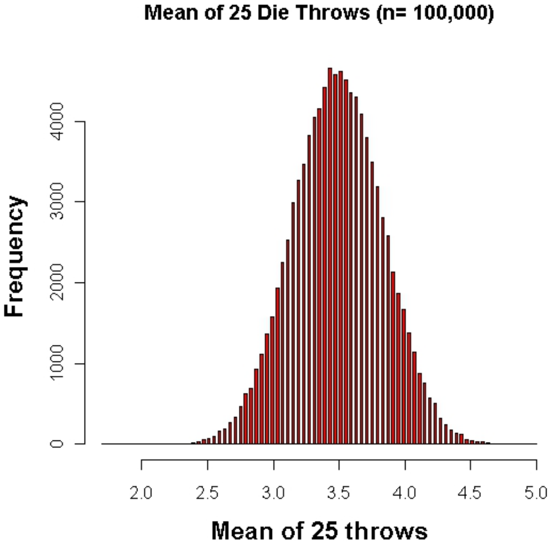
- When the die is thrown 25 times n = 25
- Therefore the standard error is

$$\frac{\sigma}{\sqrt{n}} = \frac{1.707}{\sqrt{25}} = \frac{1.707}{5} = 0.3415.$$

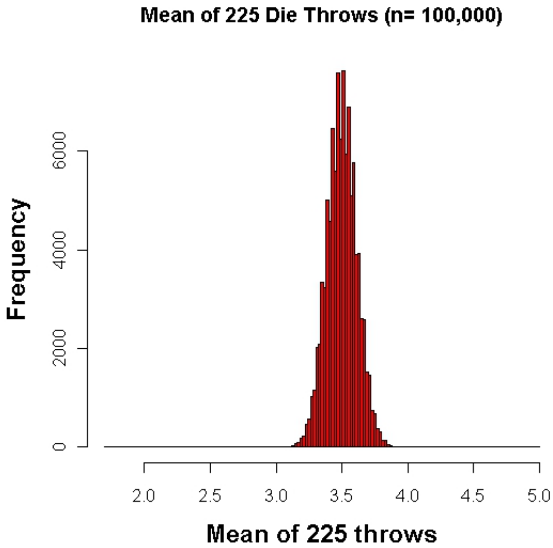
- When the die is thrown 225 times: n = 225
- Therefore the standard error is

$$\frac{\sigma}{\sqrt{n}} = \frac{1.707}{\sqrt{225}} = \frac{1.707}{15} = 0.1138.$$

Distribution of means



Distribution of means



Exercise

- Compare the two histograms on the previous slides. These horizontal range of value is the same for both histograms.
- We can see that with a larger sample size ($n = 225$), the distribution of sample means are clustered closely around the 3.5 mark, and have much less dispersion than distribution of sample means with a sample size $n = 25$.