MATH 460: Probability & Statistics — January_2019

Ch 6 Estimation Theory

Class 6 Notes — **Dr. Basilio**

Wed Jan_30 ∪ Thurs Jan_31

Southern California Institute of Technology

* * *

Guiding Question(s)

(1) How can we use sample data statistics to estimate the values of the entire population?

(2) How can we use sample data to test hypotheses (or claims) about the entire population?

(3) How can we estimate with confidence?

Chapter 6: Estimation Theory

Confidence Intervals

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a **point estimate** of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion (as measured over a long length of time, say a season).

Definition 1: Inferential-Statistics

- **POPULATION vs SAMPLE:** In statistics, we frequently want to understand data that comes from a particular group. For example, suppose I want to have a better understanding of all college student's in CA. The **population** is the entire group we are looking to study. In this case, our population is all students who are attending college in CA. Since it is generally not possible to collect data on an entire population, we collect data from a smaller group or subset taken from the population. This smaller group is called the **sample**.
- **PARAMETER vs STATISTIC:** A number that represents a characteristic of the population is called a **parameter**. A number that represents a characteristic of the sample is called a **statistic**.
- **INFERENTIAL STATISTICS:** We use **sample data** to make generalizations about an unknown population parameter. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter.
- **CONFIDENCE INTERVALS:** We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called **confidence intervals**.
- **ASSUMPTION:** For $n > 30$ the sample can be assumed to be nearly a normal distribution.

Definition 2: Confidence Level for Population Proportion

- **POINT ESTIMATE:** a single value used to estimate a population parameter.
- **CONFIDENCE INTERVAL:** is a range (or an interval) or values used to estimate the true value of a population parameter.
- **CONFIDENCE LEVEL:** is the probability $1-\alpha$ (so if $\alpha = 0.05$, then $1-\alpha = 0.95$) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. We use the notation $C = 1-\alpha$ to denote confidence levels when representing it as a decimal, and $C \cdot 100\%$ as a percent.
- **CRITICAL VALUE:** is the number on the borderline separating the likely region from the unlikely region. The number $z_{\alpha/2}$ is a critical value that is a z-score with the property that it is at the border that separates an area of $\alpha/2$ in the right tail of the standard normal distribution.
- **How to find the Critical Values:** Let z_C be the z-score such that the area between the interval $[-z_\alpha, z_\alpha]$ is C. To do this, you can use the inverse Normal distribution to find z_α but it is not simply C because we want the middle area. The formula is:

$$\boxed{z_{\alpha/2} = \text{invNorm}\left(\frac{1+C}{2}\right)}$$

Table 6-1

Confidence Level	99.73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

- **CONFIDENCE INTERVAL FOR PROPORTION p:** The population proportion p is within the following interval with a confidence level of C:

$$\boxed{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}} \quad \text{or} \quad \boxed{\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)}$$

Example 1: Confidence Intervals

(a) A Gallup poll in which 1487 adults were surveyed and 43% of them said they had a Facebook page. Find the best point estimate of the proportion of all adults who have a Facebook page.

Solution: The best point estimate is the sample proportion given by $\hat{p} = 0.43$. □

(b) The most common confidence levels are 90% (so $C = 0.9$ and $\alpha = 0.1$), 95% (so $C = 0.95$ and $\alpha = 0.05$), and 99% (so $C = 0.99$ and $\alpha = 0.01$). Use your calculator to verify the critical values.

Activity 1: Confidence Interval for proportion

(a) Find the 95% confidence interval for the population proportion p for all adults with a Facebook page.

(b) The drug OxyContin is used to treat pain but it is dangerous due to its addictive properties. In clinical trials, 227 subjects were treated with OxyContin and 52 of them developed nausa. Construct at 95% confidence interval of the percentage of OxyContin users who develop nausea.

Definition 3: Confidence Intervals for Population Mean

- **ASSUMPTION:** For $n > 30$ the sample can be assumed to be nearly a normal distribution.
- **POPULATION vs SAMPLE:** When you conduct a survey and calculate the sample mean, $\bar{x}$, and the **sample standard deviation**, s. You would use $\bar{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.
- **STANDARD ERROR OF MEAN:** The **standard error of the mean** is given by $\frac{s}{\sqrt{n}}$.
- **CONFIDENCE INTERVAL FOR POPULATION MEAN:** the interval that contains the mean of the population mean μ with a confidence of $C \cdot 100\%$ is given by

$$\boxed{\bar{x} \pm z_c \cdot \frac{s}{\sqrt{n}}} \text{ or } \left[\bar{x} - z_c \cdot \frac{s}{\sqrt{n}}, \quad \bar{x} + z_c \cdot \frac{s}{\sqrt{n}}\right] \tag{1}$$

In general:

- The more confident (larger C) we are that we know where the population mean, μ is, then the bigger the interval will be!
- The shorter the interval predicting where the population mean μ is, then we will be less confident (smaller C)!

Example 2: Confidence Intervals

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes.
Suppose we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100.
The standard deviation for the sample mean is $\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$.
The empirical rule, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean μ. So $z_C = 2$. Two standard deviations is $2 \cdot 0.1 = 0.2$. The sample mean $\bar{x}$ is likely to be within 0.2 units of μ.
Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.
Suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between $\bar{x} - 0.2 = 2?0.2 = 1.8$ and $\bar{x} + 0.2 = 2 + 0.2 = 2.2$.
We say that we are 95% confident that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. The 95% confidence interval is $[1.8, 2.2]$.
The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (95-100%).

Activity 2: Confidence-Interval

Find a $C \cdot 100\%$ confidence interval for μ for the given values:

(a) $C = 0.95$, $\bar{x} = 75$, $s = 13.2$, and $n = 57$

(b) $C = 0.99$, $\bar{x} = 315$, $s = 63$, and $n = 100$

Activity 3: Confidence-Interval

Below are the number of times per year 38 randomly selected employees for a large company feel overworked.

$$15, 23, 50, 31, 5, 27, 47, 43, 135, 164, 80, 123, 20, 34, 45, 56, 7, 12, 15,$$
$$16, 18, 64, 79, 84, 19, 32, 34, 56, 200, 0, 16, 61, 31, 52, 61, 70, 365, 105$$

(a) Find a 85% confidence interval for μ for the population mean of this data.

(b) Find a 90% confidence interval for the population mean of this data.