Statistical Methods for Decision Making
Residency-II

For BABI Program[1]

[1] Great Lakes Institute of Management, Gurgaon

February 16, 2018

Recap of Residency-I

- Descriptive Statistics
  - Measures of Central Tendency: **Mean, Median, Mode**
  - Measures of Dispersion/Scale: **Variance, Standard Deviation**
  - Measures of Shape: **Skewness and Kurtosis**

- Coefficient of Variation (Comparison of movie actors - who is a better bet at box office?)

- Data Visualization
  - Histogram
  - Five Number summary and Boxplots
  - Identification of outliers using Boxplots

- Probability (Amazon Sale offer, Speeding Cars, Preowned car certification)
  - Joint Probability
  - Conditional Probability
  - Bayes' Theorem
- Probability Distribution of a random variable:
  - Continuous Distribution (Properties of Normal Distribution $N(\mu, \sigma^2)$, Z-Score)
  - Discrete Distribution (Bernoulli p, and Binomial Distribution $B(n, p)$)
  - Expected Value of a random variable
- Sampling Distribution and Central Limit Theorem

Topics for Residency-II

Statistical inference is the process of drawing conclusions about the entire population based on the information in a sample.

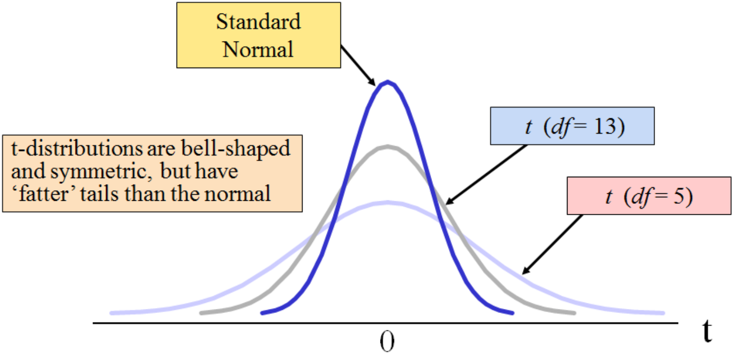
- Inferential Statistics
  - Test of a mean - test whether CocaCola 330ml can **on average** contains 330ml.
  - Test of proportions - the **proportion of** bottles in the population with less than two liters of Coke is 1%.
  - Test of variance -
  - Test of equality of means - there is no difference in the **average** scores of males and females in a test
  - Test of equality of proportions - the proportion of boys and girls in the age group 12-16 who play online games is the same
  - Test of equality of variances - the variances of stock returns of ITC and HLL are the same
  - Test of Association -
  - Paired difference test - Strength of concrete after two and seven days of setting

Case Studies

- Travel Management
- Diamond Case Study
- Housing Prices Case Study
- Cash Transfer Case Study
- Golf Ball Case Study
- Internet Usage
- Luggage Delivery in a hotel
- Strength of Concrete at two different points in time

Sampling Distribution

1. Sample: Random Draws of size N from a population (simple random sampling)
2. Step 1: Draw a random sample of Size N from the population.
3. Step 2: Calculate the population parameter of interest. Here, it is either mean or proportion
4. Step 3: Repeat Steps 1 and 2 for large number of times (say 10,000 times)
5. Sampling distribution is a plot of those sample means (or proportions). Draw a histogram
6. If $N > 30$ and population distribution is also normal, Sampling distribution will be normal
7. If $N > 30$ and population distribution is not normal, sampling distribution tends to be normal
8. if $N < 30$ and population distribution is not normal, sampling distribution will have a t-distribution (symmetric with thicker tails)



t and Z distributions

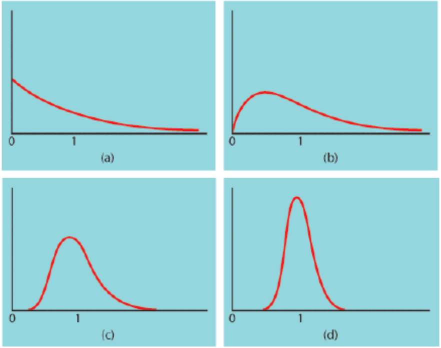
Central Limit Theorem: For random samples with a sufficiently large sample size from **any population distribution**, the distribution of sample means or proportions are normally distributed and are centered at the value of the population parameter μ (or p) and with variance equal to $\frac{\sigma^2}{N}$ (or $\frac{p(1-p)}{n}$).

CLT

Central Limit Theorem: When randomly sampling from **any** population with mean μ and standard deviation σ, **when *n* is large enough**, the sampling distribution of $\overline{x}$ is approximately normal: ~ $N(\mu, \sigma/\sqrt{n})$.

Population with strongly skewed distribution

Sampling distribution of $\overline{x}$ for n = 2 observations



Sampling distribution of $\overline{x}$ for n = 10 observations

Sampling distribution of $\overline{x}$ for n = 25 observations

Central Limit Theorem

Population Parameters and Sample Statistics

	Population Parameter	Sample Statistic
mean	μ	$\overline{X}$
standard deviation	σ	s
difference in two means	$\mu_1 - \mu_2$	$\overline{X_1} - \overline{X_2}$
test of proportions	p	$\hat{p}$
difference of proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$

Note: The population parameters are either known or unknown. Examples?

Point Estimate and Interval Estimate

Point Estimate:

We use the statistic from a sample as a point estimate for a population parameter.

Interval Estimate:

An interval estimate gives a range of plausible values for a population parameter.

Confidence Interval:
A confidence interval is an interval estimate, computed from a sample, that has a predetermined chance of capturing the value of the population parameter.

95% Confidence Interval Using the Standard Error:

If the sampling distribution is relatively symmetric and bell-shaped, a 95% confidence interval can be estimated using

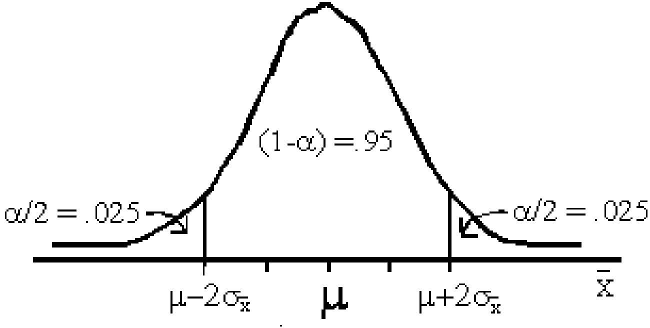
$$\text{Statistic} \pm 2 \times \frac{\sigma^2}{N}$$

Note: Standard error is the standard deviation of sampling distribution $\frac{\sigma^2}{N}$.

Interpreting Confidence Level:

The confidence level indicates how sure we are that our interval contains the population parameter. For example, we interpret a 95% confidence interval by saying we are 95% sure that the interval contains the population parameter.

The 95% confidence interval for μ



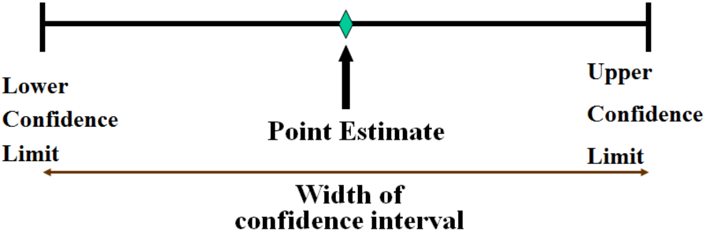
Confidence Interval

To sum it up...

A **point estimate** is a single number,

How much uncertainty is associated with a point estimate of a population parameter?

An interval estimate provides more information about a population characteristic than does a point estimate. It provides a confidence level for the estimate. Such interval estimates are called confidence intervals.



Confidence Interval

Computing the confidence intervals

We know that in 95% of the samples, the true population proportion(or mean) will fall within in 2 standard errors of the sample mean. Where does the 2 come from:

For a bell curve 95% of the data will be between ±1.96 standard deviations. So approximated to 2.

If population variance, σ is known and $N > 30$

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}$$

If population variance is unknown or $N < 30$ or both then we use t-stat and sample standard deviation s.

$$\overline{X} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{N}}$$

Hypothesis Testing

Example - I

Two Scenarios:

- What if a drug is passed when the intended effects of the drug are suspect?
- What are the implications of not letting an effective drug hit the market?

Read the following News:

`https://prescriptiondrugs.procon.org/view.resource.php?resourceID=005528`

`http://www.rediff.com/money/report/drugs-ban-these-are-the-worst-hit-brands-companies/20160321.htm`

`https://www.eurekalert.org/pub_releases/2009-06/l-igd062609.php`

Example - II

According to Toyota website, the average mileage per gallon for the 2013 Toyota Prius is between 48 miles per gallons to 51 miles per gallon. We would like to test the claim that the mean miles per gallon for all 2013 Toyota Prius is, at the worst case possible, greater than 48 miles per gallon?

https://www.wyzant.com/resources/blogs/369788/example_of_hypothesis_testing_in_real_life

http://www.fuelly.com/

Example - III

- **NFL**: https://www.mathbootcamps.com/using-nfl-understand-hypothesis-testing

- **NBA**: https://squared2020.com/2015/11/01/hypothesis-testing-is-nba-scoring-up-this-year/

Example - IV

You are the commercial loan officer at a bank, in the process of reviewing a loan application recently filed by a local firm. Examining the firm's list of assets, you notice that the largest single item is 3 million dollars in accounts receivable. You have heard enough scare stories, about loan applicants manufacturing receivables out of thin air, that it seems appropriate to check whether these receivables actually exist.

http://www.kellogg.northwestern.edu/faculty/weber/decs-430/decs-430%20session%204/hypothesis_testing.htm

Test of a single Mean

Experian Marketing Services reported that the typical American spends a mean of 144 minutes (2.4 hours) per day accessing the Internet via a mobile device. (Source: The 2014 Digital Marketer, available at ex.pn/1kXJjfX.) In order to test the validity of this statement, you select a sample of 30 friends and family. The results for the time spent per day accessing the Internet via mobile device (in minutes) are stored in InternetMobileTime.xls

a. Is there evidence that the population mean time spent per day accessing the Internet via mobile device is different from 144 minutes? Use the p-value approach and a level of significance of 0.05.

b. What assumption about the population distribution is needed in order to conduct the t test in (a)?

Problem II, Test of Proportions

Salesforce ExactTarget Marketing Cloud conducted a study of U.S. consumers that included 205 tablet owners. The study found that 134 tablet owners use their tablet while watching TV at least once per day. (Source: “New Mobile Tracking & Survey Data: 2014 Mobile Behavior Report,” bit.ly/1odMZ3D.) The authors of the report imply that the survey proves that more than half of all tablet owners use their tablet while watching TV at least once per day.

Test the hypothesis that more than half of all tablet owners use their tablet while watching TV at least once per day.

Problem III, Test of two Sample Means

A hotel manager looks to enhance the initial impressions that hotel guests have when they check in. Contributing to initial impressions is the time it takes to deliver a guest's luggage to the room after check-in. A random sample of 20 deliveries on a particular day were selected in Wing A of the hotel, and a random sample of 20 deliveries were selected in Wing B. The results are stored in Luggage.

Analyze the data and determine whether there is a difference between the mean delivery times in the two wings of the hotel.

Problem IV, Test of Paired Sample Means

The file Concrete1.xls contains the compressive strength, in thousands of pounds per square inch (psi), of 40 samples of concrete taken two and seven days after pouring.

(Data extracted from O. Carrillo Gamboa and R. F. Gunst, “Measurement Error Model Collinearities,” Technometrics, 34 (1992): 454 — 464.)

At the 0.01 level of significance, is there evidence that the mean strength is lower at two days than at seven days?

Steps to hypothesis Testing

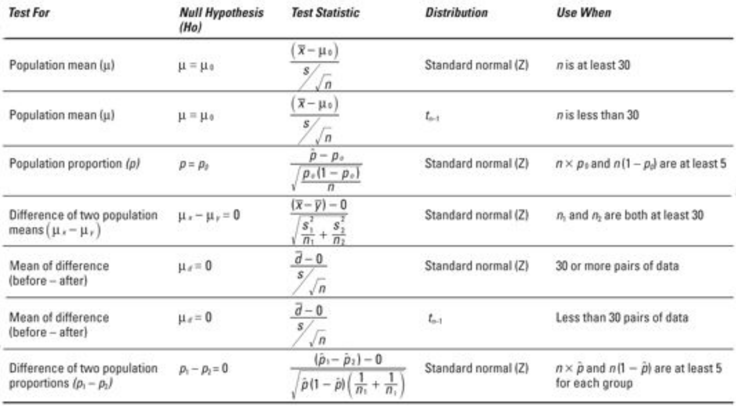
Statistical Tests: A statistical test uses data from a sample to assess a claim about a population.

- Formulate the null and alternative hypotheses. Alternative Hypothesis is what you want to prove.
- Specify the level of significance and select the sample size.
- Select the test statistic to be used and determine its distribution under the assumption that the null hypothesis is true.
- Collect the sample and compute the value of the test statistic.
- Determine the critical value(s) and specify the rejection region.
- Reject the null hypothesis if the test statistic falls in the rejection region.
- Interpret this decision in the context of the problem.

Cheat Sheet for Hypothesis Testing

1	Type of test	Test of mean μ or test of proportion p
2	Type of sample	one sample or two sample
3	Specify Null as	Status Quo
4	Specify Alternate as	Whatever you want to prove
5	Tails	If $H_a < \mu$ then left tail, $H_a > \mu$ is right tail , not equal is both
6	Confidence level	Fix the alpha (probability of type—I error)
7	Is sigma known? and/or Is $n > 30$?	If answer “Yes” to both , then Z otherwise t test
8	“critical values” or areas	Based on 6 and 7, find “critical values” or areas that define Left, Right or Both tails
9	Acceptance Region	Shade the area where H_a is “true”
10	Test Statistic	Calculate test statistics using fomula
11	Inference	If test statistic in the above step is in shaded portion then H_a holds, not otherwise

Reference:
`http://www.math.wayne.edu/~bert/courses/1020/hypothesis.testing.pdf`



Test For	*Null Hypothesis (Ho)*	*Test Statistic*	*Distribution*	*Use When*
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$	Standard normal (Z)	n is at least 30
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x}-\mu_0)}{s/\sqrt{n}}$	t_{n-1}	n is less than 30
Population proportion *(p)*	$p = p_0$	$\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Standard normal (Z)	$n \times p_0$ and $n(1-p_0)$ are at least 5
Difference of two population means $(\mu_x - \mu_y)$	$\mu_x - \mu_y = 0$	$\frac{(\bar{x}-\bar{y})-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	Standard normal (Z)	n_1 and n_2 are both at least 30
Mean of difference (before – after)	$\mu_d = 0$	$\frac{\bar{d}-0}{s/\sqrt{n}}$	Standard normal (Z)	30 or more pairs of data
Mean of difference (before – after)	$\mu_d = 0$	$\frac{\bar{d}-0}{s/\sqrt{n}}$	t_{n-1}	Less than 30 pairs of data
Difference of two population proportions $(p_1 - p_2)$	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_1}\right)}}$	Standard normal (Z)	$n \times \hat{p}$ and $n(1-\hat{p})$ are at least 5 for each group

Hypothesis testing

t-Test using R

The R-Command for t-Test:

$t.test(x, y = NULL, alternative = c(\text{``}two.sided\text{''}, \text{``}less\text{''}, \text{``}greater\text{''}), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)$

Source: `https://www.rdocumentation.org/packages/stats/versions/3.4.3/topics/t.test`

Proportion test using R

$prop.test(x, n, p = NULL, alternative = c(''two.sided'', ''less'', ''greater''), conf.level = 0.95, correct = TRUE)$

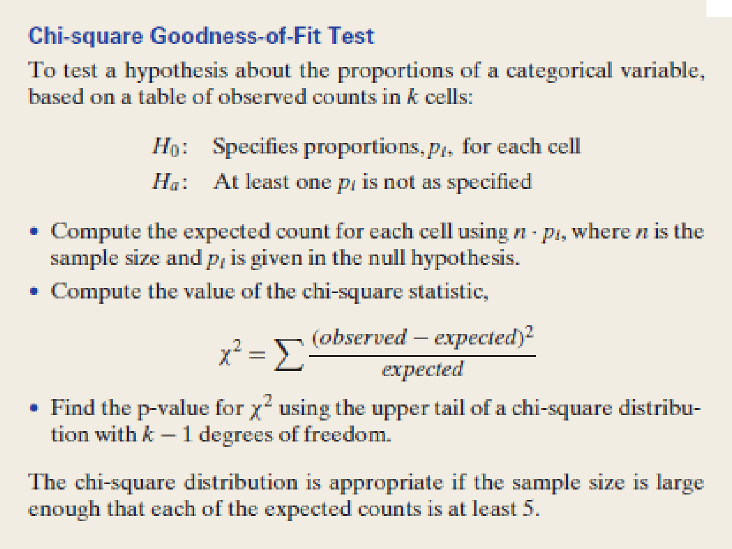
Source: `https://www.rdocumentation.org/packages/stats/versions/3.4.3/topics/prop.test`

Chi-square Statistic

The chi-square statistic, denoted with the Greek χ^2, is found by comparing the observed counts from a sample with expected counts derived from a null hypothesis. The formula for computing the statistic is

$$\chi^2 = \sum \frac{(Obeserved - Expected)^2}{Expected}$$

where the sum is over all cells of the table.



Chi-square Goodness-of-Fit Test

To test a hypothesis about the proportions of a categorical variable, based on a table of observed counts in k cells:

$$\begin{aligned} H_0 &: \quad \text{Specifies proportions, } p_i, \text{ for each cell} \\ H_a &: \quad \text{At least one } p_i \text{ is not as specified} \end{aligned}$$

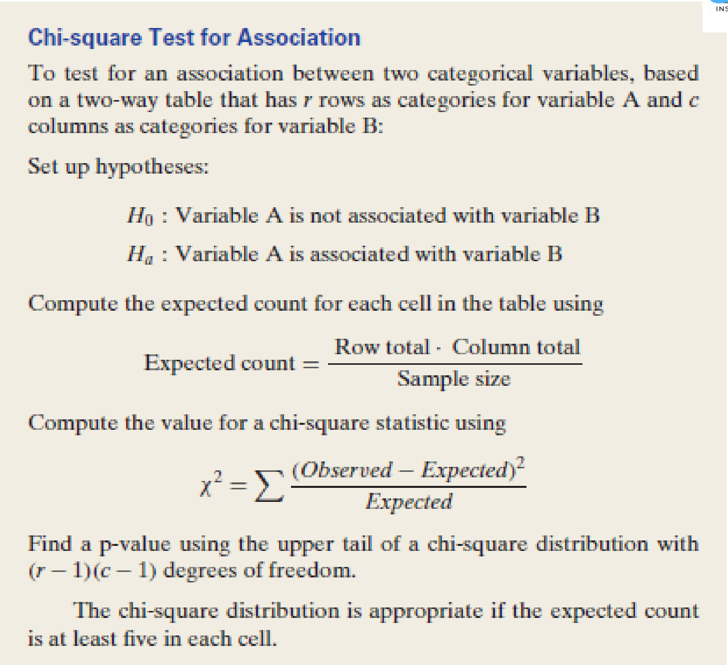
- Compute the expected count for each cell using $n \cdot p_i$, where n is the sample size and p_i is given in the null hypothesis.
- Compute the value of the chi-square statistic,

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

- Find the p-value for χ^2 using the upper tail of a chi-square distribution with $k-1$ degrees of freedom.

The chi-square distribution is appropriate if the sample size is large enough that each of the expected counts is at least 5.

Chi-square goodness—of—fit test



Chi-square Test for Association

To test for an association between two categorical variables, based on a two-way table that has r rows as categories for variable A and c columns as categories for variable B:

Set up hypotheses:

$$H_0 : \text{Variable A is not associated with variable B}$$

$$H_a : \text{Variable A is associated with variable B}$$

Compute the expected count for each cell in the table using

$$\text{Expected count} = \frac{\text{Row total} \cdot \text{ Column total}}{\text{Sample size}}$$

Compute the value for a chi-square statistic using

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Find a p-value using the upper tail of a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

The chi-square distribution is appropriate if the expected count is at least five in each cell.

Happy Learning!!!