Data Analysis Course

Testing of Hypothesis

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

- What is the need of testing
- Recap of sampling distribution
- Hypothesis testing
  - Five main steps in testing
    - Assumptions
    - Hypotheses
    - Test Statistic
    - P-value (P)
    - Conclusion:
- Testing example
- Types of errors
- Testing for Means
  - Z test
  - T test
- Testing for Proportions
- Test of independence

Inference

- **A cake**, (weighs 20 Kg) how do you decide about its taste? A piece of cake or after eating completely
- **A truck** half filled with oranges rest with apples. How would you verify that? By counting them?
- **Apple & Samsung** sell mobiles all over the world. If you want to find which one people prefer, Do you take the opinion poll from all the users around the world?
- **Product** manager claims that girls are buying their product more than boys? Is there any association between gender and buying or not buying?

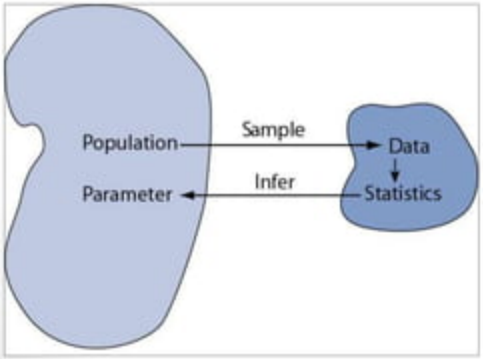
Inference

- **A cake**, (weighs 20 Kg) how do you decide about its taste? A piece of cake or after eating completely
  - You took a piece of cake, you are not completely satisfied with the taste –Would you recommend the cake?
  - You took a 10 gram piece of cake, you didn't like the taste—What would you say?
  - You took a 1 milligram piece of cake, you liked the taste—What would you say?
- **A truck** half filled with oranges rest with apples (owner claims 10,000 oranges & 10,000 apples) . How would you verify that? By counting them?
  - You randomly picked 200 fruits, 120 are apples and 80 are oranges. What would you say?
  - You randomly picked 200 fruits, 190 are apples and 10s are oranges. What would you say?

How bad in the sample is bad enough to say the population is also bad

Statistical Inference

- Inferences about a population are made on the basis of results obtained from a sample drawn from that population
- Want to talk about the larger population from which the subjects are drawn, not the particular subjects!



- A hypothesis test is a process that uses sample statistics to test a claim about the value of a population parameter.
- A verbal statement, or claim, about a population parameter is called a statistical hypothesis.
- Hypothesis: Proportion of apples = 0.5

Applications of testing

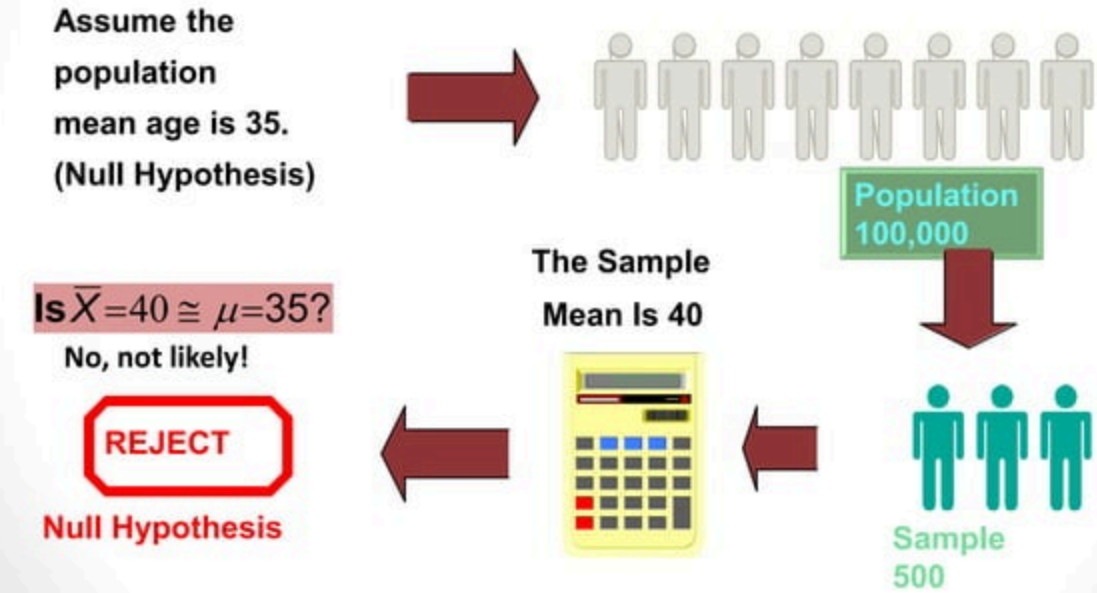
- Law and Forensics
  - Testing for discrimination (in admission, hiring, pay, promotion practices, etc.) – test of proportions
  - Paternity testing
  - Testing whether evidence found on a suspect came from the crime scene (blood, fiber, fingerprints, ...)
  - Indeed, testing whether the defendant is guilty or not
- Medicine and Health
  - Testing if a new drug is effective or ineffective-test of association
  - Testing if particulate matter in air pollution causes lung cancer
  - Testing if a particular gene is responsible for hemophilia
- Industry/Business/Economics
  - Testing if a production machine is 'in control' or not-test of means
  - Testing if a silicon wafer is good for use
- Science and Engineering
  - Testing which theory of gravitation is correct, based on dark matter
  - Testing whether men and woman differ according to a psychological trait
  - Testing if two species have a common ancestor

Hypothesis Testing – An example

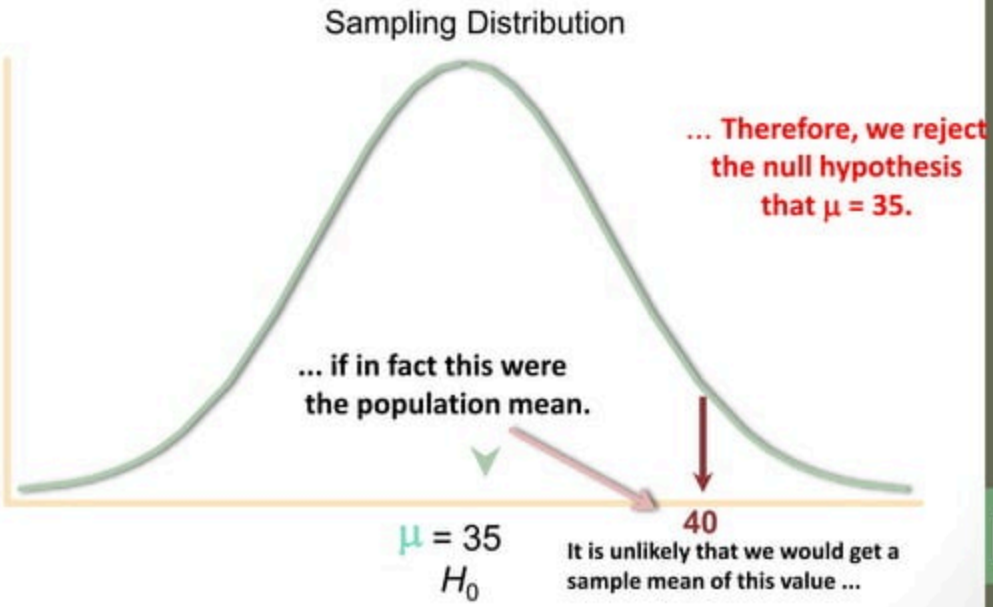
CEO of SBI claims employees mean age in SBI bank is 35 (292,215 employees). How can we prove or disprove it?

1. Take a random sample (500) and find their age
2. If it is near 35 then we say there is no evidence to reject that hypothesis
3. What if sample average age is lower or higher than 35 ?
4. How far is really far? We want to quantify the severity of deviation……
5. Lets find the probability of this occurrence
6. It if it is really low, then we say we reject null hypothesis

Hypothesis testing process

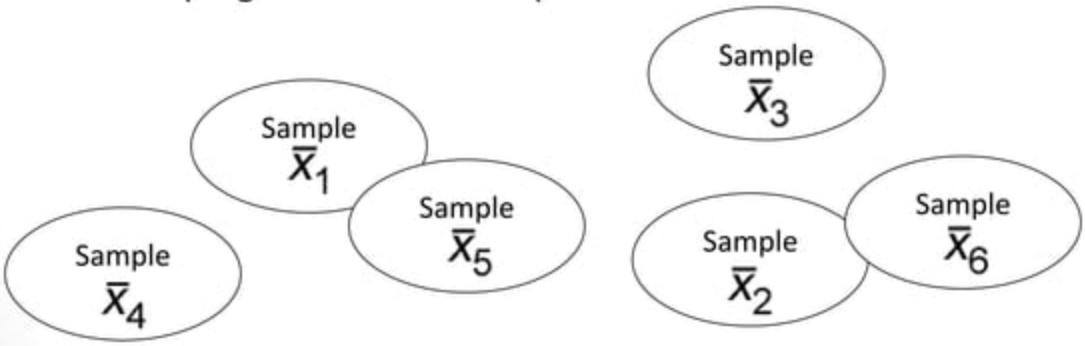


Reason for Rejecting H0



Sampling distribution -Recap

- A sampling distribution is the probability distribution of a sample statistic that is formed when samples of size n are repeatedly taken from a population.
- If the sample statistic is the sample mean, then the distribution is the sampling distribution of sample means.

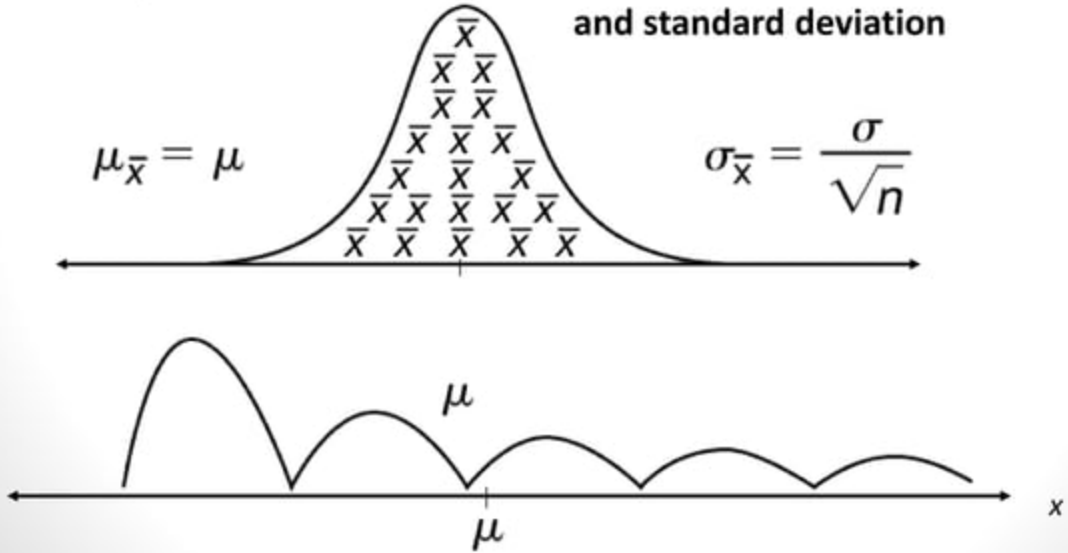


The sampling distribution consists of the values of the sample means, $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5, \bar{x}_6, \ldots$

Central Limit theorem -Recap

If a sample *n* (30) is taken from a population with ***any type distribution*** that has a mean = μ and standard deviation = σ

the ***sample means*** will have a **normal distribution and standard deviation**



Five Step in Testing of Hypothesis

1. Make Assumptions and meet test requirements.
2. State the null hypothesis.
3. Select the sampling distribution and establish the critical region.
4. Compute the test statistic.
5. Make a decision and interpret results.

Step 1: Make Assumptions and Meet Test Requirements

- Random sampling
  - Hypothesis testing assumes samples were selected using random sampling.
  - In this case, the sample of 500 cases was randomly selected from all major branches.
- Level of Measurement is Interval-Ratio.
  - Yes age is not a categorical variable
- Sampling Distribution is normal in shape.
  - What is the sampling distribution of age?
- This is a "large" sample ($n \geq 100$).

Step 2 State the Null Hypothesis

- H0: $\mu = 35$
- In other words, Ho: No difference between the sample mean and the population parameter
- In other words, The sample mean of 40 is really the same as the population mean of 35 – the difference is not real but is due to chance.
- In other words, The sample of 500 comes from a population that has average age of 35
- In other words, The difference between 35 and 40 is trivial and caused by random chance.

Step 2 (cont.) State the Alternate Hypothesis

- H1: $\mu \neq 35$
- Or H1: There is a difference between the sample mean and the population parameter
- Or The sample of 500 comes a population that does not have average age 35 In reality, it comes from a different population.
- Or The difference between 40 and 35 reflects an actual difference
- Or the average age of the population is more than 35

Step 3 Select Sampling Distribution and Establish the Critical Region

- What is the sampling distribution of population mean?
- What is alpha?
  - Probability of rejecting H0 when it is true
- α is the indicator of "rare" events.
- Any difference with a probability less than α is rare and will cause us to reject the H0.
- We started with H0 as true, we still want to reject the null if the statistic is beyond a certain value,
- We already know about some unlikely values of test statistic when null hypothesis is true
- for example if the average age of the sample is 60, we definitely want to reject null

Details later

Step 4: Use Formula to Compute the Test Statistic - Z for large samples (≥ 100)

- We got the sample average as 40, the age according to null hypothesis is 35, there is a difference of 5, is it due to chance?
- How bad is this difference of 5?

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$$

When the Population σ is not known, use the following formula:

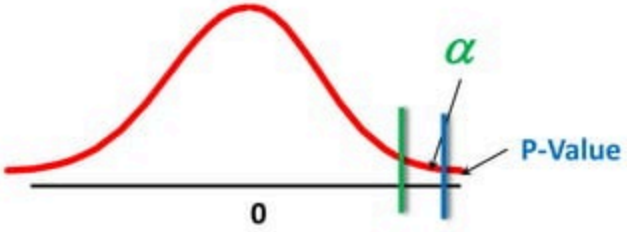
$$Z = \frac{\overline{X} - \mu}{s / \sqrt{n-1}} \qquad Z = \frac{40 - 35}{7.86 / \sqrt{500-1}}$$

Step 5 Make a Decision and Interpret Results

- The obtained Z score fell in the Critical Region, so we reject the H0.
  - If the H0 were true, a sample outcome of 14 would be unlikely.
  - Therefore, the H0 is false and must be rejected.

H_0: $\mu = 35$

H_1: $\mu > 35$



What does z of 14 mean? The probability of z being more than 14 is less than 0.000000001

- It is like getting more than 25 heads in a row when you toss a coin
- It is like drawing the same card more than 6 times from a shuffled deck
- If the average age of 40 is just by chance then compare that chance with above examples

What is P-Value

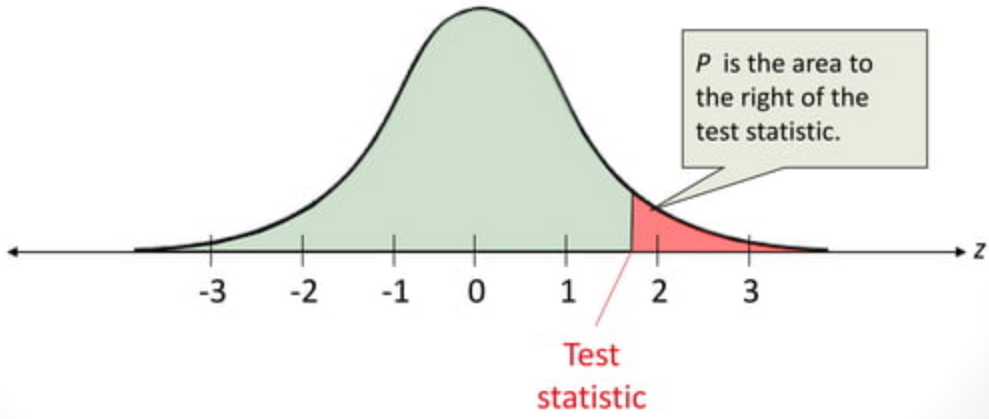
- If the observed statistic happens to be just a chance, p-values tells us what is the probability of that chance
- **The P-value answer the question:** What is the probability of the observed test statistic or one more extreme when H0 is true?
- Given H0, probability of the current value or extreme than this
- Given H_0 is true, probability of obtaining a result as extreme or more extreme than the actual sample
- The **observed significance level,** or **p-value** of a test of hypothesis is the probability of obtaining the observed value of the sample statistic, or one which is even more supportive of the alternative hypothesis, under the assumption that the null hypothesis is true.
- Smallest α the observed sample would reject H_0

P -value

- If the alternative hypothesis contains the greater-than symbol (>), the hypothesis test is a right-tailed test.

$H_0: \mu = k$

$H_a: \mu > k$



One tail & two tailed tests

- Two-tailed Test
  - If the alternative hypothesis contains the not-equal-to symbol (≠), the hypothesis test is a **two-tailed test**. In a two-tailed test, each tail has an area of 0.5*P*.
    - $H_0: \mu = k$
    - $H_a: \mu \neq k$
- Left-tailed Test
  - If the alternative hypothesis contains the less-than inequality symbol (<), the hypothesis test is a left-tailed test.
    - $H_0: \mu \geq k$
    - $H_a: \mu < k$
- Right-tailed Test
  - If the alternative hypothesis contains the less-than inequality symbol (<), the hypothesis test is a left-tailed test.
    - $H_0: \mu \geq k$
    - $H_a: \mu < k$

Types of errors

- No matter which hypothesis represents the claim, always begin the hypothesis test assuming that the null hypothesis is true.
- At the end of the test, one of two decisions will be made:
  - reject the null hypothesis, or
  - fail to reject the null hypothesis.
- A **type I error** occurs if the null hypothesis is rejected when it is true.
- A **type II error** occurs if the null hypothesis is not rejected when it is false.

Error Types

- Type I Error: Reject H0 when it is true
- Type II Error: Do not reject H0 when it is false

Test Result – *Reality*	Reject H_0	Don't Reject H_0
H_0 True	Type I Error	Correct
H_0 False	Correct	Type II Error

Level of Significance α

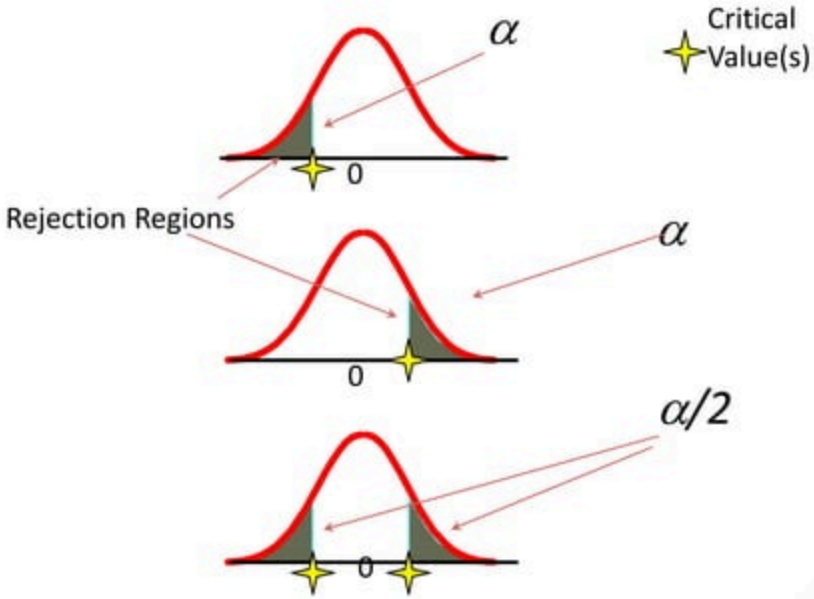
- Defines Unlikely Values of Sample Statistic if Null Hypothesis Is True
  - Called Rejection Region of Sampling Distribution
- Designated α (alpha)
- Typical values are 0.01, 0.05, 0.10
- Selected by the Researcher at the Start Provides the Critical Value(s) of the Test
- P(Type I error)
- Think of analogy with SBI average age

Level of Significance, α and the Rejection Region

H_0: $\mu = 35$
H_1: $\mu < 35$

H_0: $\mu = 35$
H_1: $\mu > 35$

H_0: $\mu = 35$
H_1: $\mu \neq 35$



Power of test 1-β

- P(Type II error) is β
- P(Type II error) = β depends on the true value of the parameter (from the range of values in Ha).
- The farther the true parameter value falls from the null value, the easier it is to reject null, and P(Type II error) goes down.
- Power of test = 1 - β = P(reject null, given it is false)
- In practice, you want a large enough n for your study so that P(Type II error) is small for the size of effect you expect.

Which error is bad?

- False negative
  - Miss what could be important
    - Testing a metal whether it is gold or not
    - Are these samples going to be looked at again?
- False positive
  - Waste resources following dead ends
    - Test whether a drug is deadly or not

Confidence Intervals

- Hypothesis testing focuses on where the sample mean is located
- Confidence intervals focus on plausible values for the population mean
- General Formula (1-α)% CI for μ

$$\left(\bar{X} - \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{1-\alpha/2}\sigma}{\sqrt{n}} \right)$$

- Construct an interval around the point estimate
- Look to see if the population/null mean is inside

Significance Test for Mean

For large samples $$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$$

For small samples $$t = \frac{\overline{y} - \mu_0}{se} \quad \text{where} \quad se = s / \sqrt{n}$$

- Sampling distribution for small samples is t
- The curve of the t distribution varies with sample size (the smaller the size, the flatter the curve)
- In using the t-table, we use "degrees of freedom" based on the sample size.
- For a one-sample test, df = n – 1.
- When looking at the table, find the t-value for the appropriate df = n-1. This will be the cutoff point for your critical region.

Lab: Significance Test for Mean

- It is known that the mean cholesterol level for the nation is 190. We test 100 only children and find that the sample average cholesterol level is 198 and suppose we know the population standard deviation $\sigma = 15$. does that signify that only children have an average higher cholesterol level than the national average?
- Given this sample what are the reasonable values for population mean?

- 50 smokers were questioned about the number of hours they sleep each day. We want to test the hypothesis that the smokers need less sleep than the general public which needs an average of 7.7 hours of sleep. If the sample mean is 7.5 and the population standard deviation is 0.5, what can you conclude?
- Given this sample what are the 95% confidence limits for population mean?

Test of Proportion

- Assumptions:
  - Categorical variable
  - Randomization
  - Large sample (but two-sided ok for nearly all n)

- Hypotheses:
  - Null hypothesis: H0: $\pi = \pi_0$
  - Alternative hypothesis: Ha: $\pi \neq \pi_0$ (2-sided)
  - Ha: $\pi > \pi_0$ Ha: $\pi < \pi_0$ (1-sided)
  - Set up hypotheses before getting the data
- Test statistic:

$$z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Lab: Test of Proportion

- Suppose a coin toss turns up 12 heads out of 20 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?
- Suppose that you interview 1000 exiting voters about who they voted for PM. Of the 1000 voters, 550 reported that they voted for Rahul Gandhi. Is there sufficient evidence to suggest that Rahul Gandhi will win the election at the .01 level?

Chi square test for Independence

- Chi square test of independence
- Is happiness independent of family income? A sample of 2955 families are studied

		Happiness			
		Very	Pretty	Not too	Total
Income	Above	272	294	49	615
	Average	454	835	131	1420
	Below	185	527	208	920

Chi-square statistic

- Summarize closeness of $\{f_o\}$ and $\{f_e\}$ by

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where sum is taken over all cells in the table.

- When H_0 is true, sampling distribution of this statistic is approximately (for large *n*) the *chi-squared probability distribution.*

Chi-square calculation

- In happiness and family income

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(272 - 189.6)^2}{189.6} + \ldots = 172.3$$

- $df = (3 - 1)(3 - 1) = 4$. P-value = 0.000 (rounded, often reported as $P < 0.001$). Chi-squared percentile values for various right-tail probabilities are in table on text p. 594.
- There is very strong evidence against H_0: independence (If H_0 were true, prob. would be < 0.001 of getting this large a χ^2 test statistic or even larger).
- For significance level $\alpha = 0.05$ (or $\alpha = 0.01$ or $\alpha = 0.001$), we reject H_0 and conclude that an association exists between happiness and income.

Lab Chi-square distribution

- Suppose that 125 children are shown three television commercials for breakfast cereal and are asked to pick which they liked best. The results are shown in table below. You would like to know if the choice of favorite commercial was related to whether the child was a boy or a girl or if these two variables are independent

	A	B	C	Totals
Boys	30	29	16	75
Girls	12	33	5	50
Totals	42	62	21	125

$$\chi^2 = \frac{(30-25.2)^2}{25.2} + \frac{(29-37.2)^2}{37.2} + \frac{(16-12.6)^2}{12.6} + \frac{(12-16.8)^2}{16.8} + \frac{(33-24.8)^2}{24.8} + \frac{(5-8.4)^2}{8.4}$$
$$= 0.914 + 1.808 + 0.917 + 1.371 + 2.711 + 1.376$$
$$= 9.098$$

- Suppose you conducted a drug trial on a group of animals and you hypothesized that the animals receiving the drug would show increased heart rates compared to those that did not receive the drug. You conduct the study and collect the following data.

	Heart Rate Increased	No Heart Rate Increase	Total
Treated	36	14	50
Not treated	30	25	55
Total	66	39	105

Further Reading

- Test of samples means for two populations
- Test of sample proportions for two populations
- Odds ration for test of association

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879