Inspire…Educate…Transform.

Foundations of Probability and Statistics for Data Science

t-Distribution, Chi-Square Distribution, F Distribution, ANOVA

Prof Anuradha Sharma

December 16, 2018

MATERIAL CONTENT FROM Dr. SRIDHAR PAPPU

Common Test Statistics for Inferential Techniques

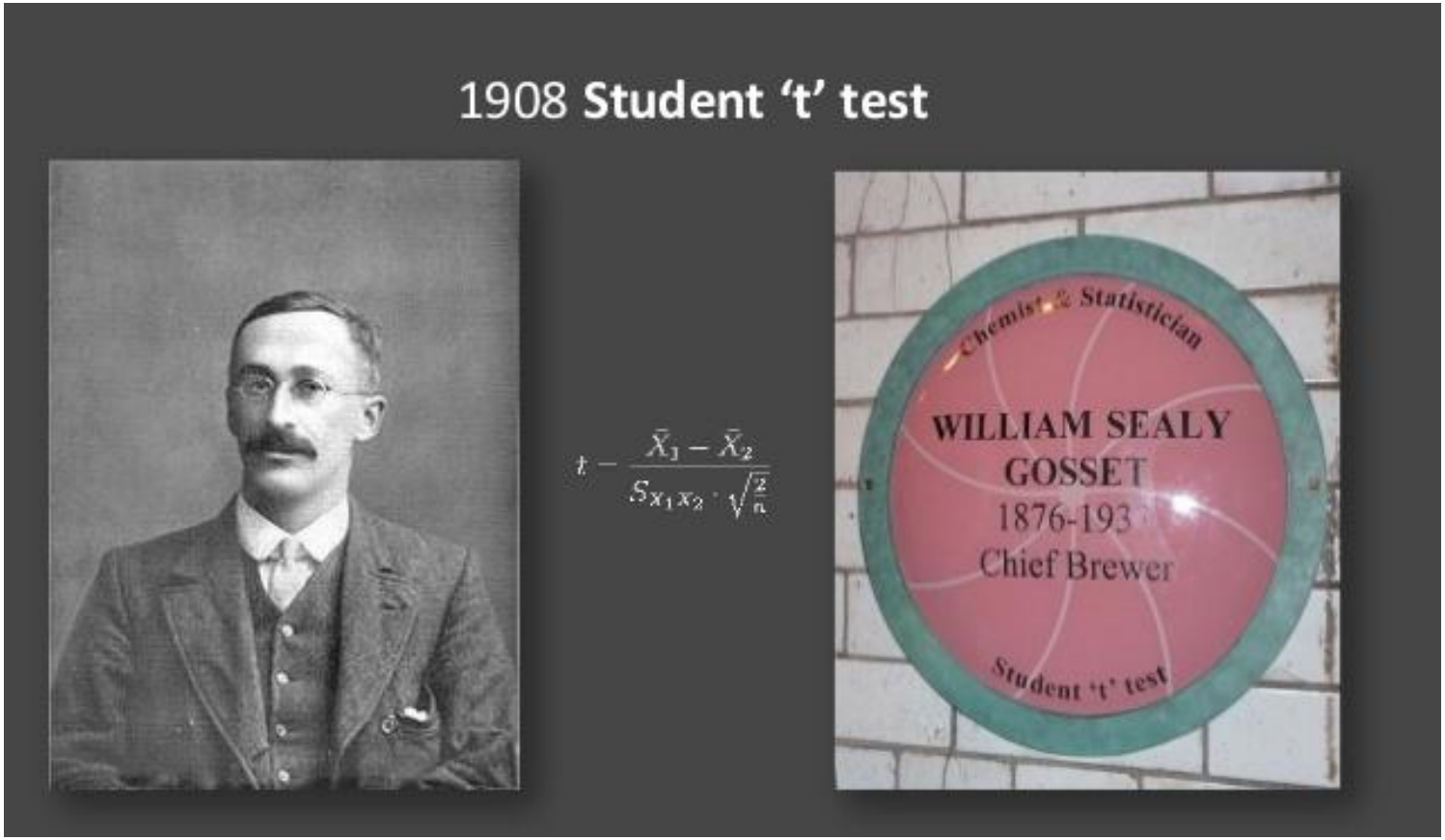
Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistics:

- z
- t
- χ^2 (Chi-squared)
- F

Closely related to Sampling Distribution of **Means**

- Closely related to Sampling Distribution of **Variances**
- Derived from Normal Distribution

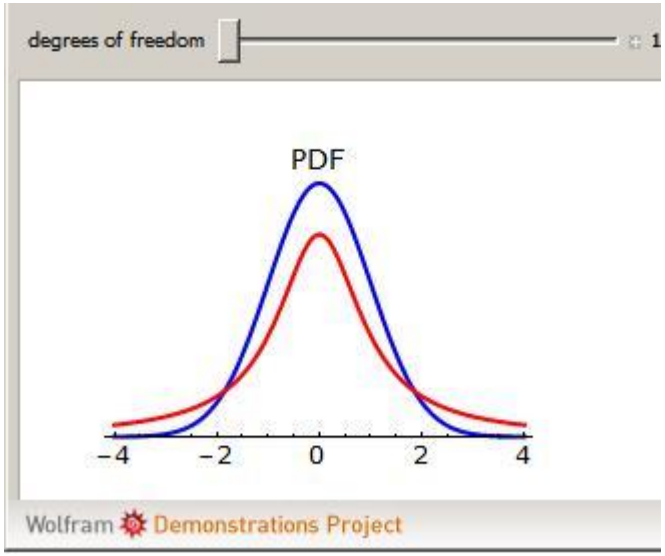
t-Distribution

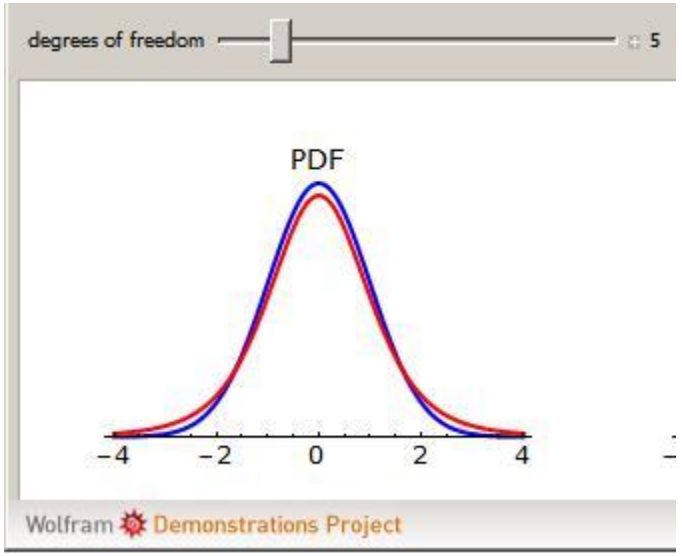


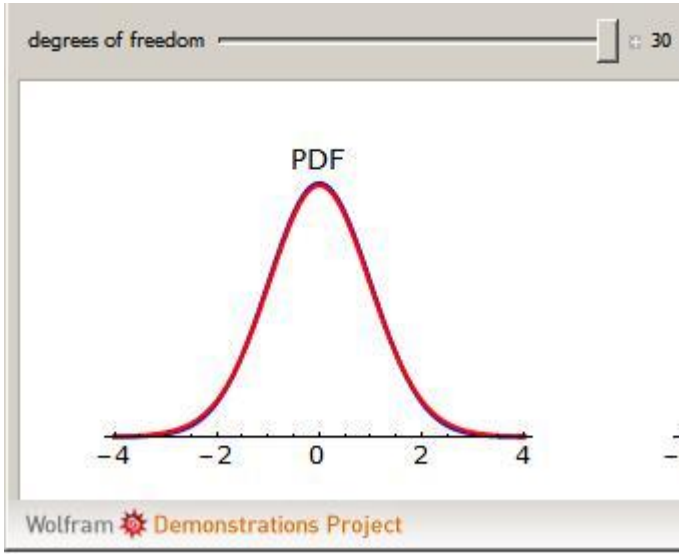
Ref: http://image.slidesharecdn.com/2013-ingenious-ireland-theingeniousirishiet-slideshow-130524065705-phpapp01/95/2013-ingeniousirelandthe-ingenious-irishietslideshow-43-638.jpg?cb=1369825611
Last accessed: October 31, 2015

t-Distribution

If the sample size is small (<30), the variance of the population is not adequately captured by the variance of the sample. Instead of z-distribution, t-distribution is used. It is also the appropriate distribution to be used when population variance is **not known**, irrespective of sample size.







Ref: "Comparing Normal and Student's t-Distributions" from the Wolfram Demonstrations Project
http://demonstrations.wolfram.com/ComparingNormalAndStudentsTDistributions/
Contributed by: Gary McClelland; Last accessed: August 11, 2017

t-Distribution [Excel Degree of Freedom]

$$t\ statistic\ (or\ t\ score), t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$$

Degrees of freedom, ν: # of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.*

Degrees of freedom is represented by "$\boldsymbol{\nu}$" which is the Greek alphabet "Nu"

When sample size is considered, degrees of freedom are *n-1*.

Recall the Infosys stock hypothetical data created to explain the concept of variance on Day 1 of this module.

* Roger E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, California: Brooks/Cole, 1968.

Properties of *t*-Distribution

- Mean of the distribution = 0
- Variance = $\frac{\nu}{\nu-2}$, where $\nu > 2$

- "ν " is the Greek alphabet "Nu" – Degrees of freedom

- Variance is always greater than 1, although it is close to 1 when there are many degrees of freedom (sample size is large)
- With infinite degrees of freedom, *t* distribution is the same as the standard normal distribution

Population Variance and Sample Variance

Population variance

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{n}$$

Sample variance

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

Confidence Interval to Estimate μ

- Population standard deviation UNKNOWN and the population normally distributed.
- $\bar{x} - t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}}$

  Recall: $t\ statistic\ (or\ t\ score), t = \frac{(\bar{x}-\mu)}{\frac{s}{\sqrt{n}}}$

  - Sample mean, standard deviation and size can be calculated from the data; *t* value can be read from the table or obtained from software.
  - α is the area in the tail of the distribution. For 90% Confidence Level, α=0.10. In a Confidence Interval, this area is symmetrically distributed between the 2 tails (α/2 in each tail).

t-Distribution - Example

The labeled potency of a tablet dosage form is 100 mg. As per the quality control specifications, 10 tablets are randomly assayed.

A researcher wants to estimate the interval for the true mean of the batch of tablets with 95% confidence. Assume the potency is normally distributed.

Data are as follows (in mg):

98.6	102.1	100.7	102.0	97.0
103.4	98.9	101.6	102.9	105.2

t-Distribution - Example

Mean, $\bar{x}$ = 101.24 mg

Standard deviation, *s* = 2.48. ($S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$)

n = 10

$\nu = 10 - 1 = 9$

At 95% level, $\alpha = 0.05$, and $\therefore, \frac{\alpha}{2} = 0.025$

R: qt(0.025,9) -> $t_{critical}$ = -2.262

t-Distribution - Example

Mean, $\bar{x}$ = 101.24 mg, Standard deviation, s = 2.48

n = 10, $\nu = 10 - 1 = 9$

$$\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$101.24 - 2.262 * \frac{2.48}{\sqrt{10}} \leq \mu \leq 101.24 + 2.262 * \frac{2.48}{\sqrt{10}}$$

$99.47 \leq \mu \leq 103.01$

The batch mean is 101.24 mg with an error of +/-1.77 mg (101.24 - 99.47 or 103.01-101.24) The researcher is 95% confident that the average potency of the batch of tablets is between 99.47 mg and 103.01 mg.

t-Distribution – Example – R

R code: t.test(dosage, conf.level = 0.95, mu = 100)

How do you get the sample t-value?

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{101.24 - 100}{\frac{2.48}{\sqrt{10}}} = 1.5824$$

```
        One Sample t-test

data:  potency
t = 1.5824, df = 9, p-value = 0.148
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  99.46735 103.01265
sample estimates:
mean of x
   101.24
```

HYPOTHESIS TESTING APPROACH

CONFIDENCE INTERVAL APPROACH

Can population mean be equal to 100mg?

TWO-SAMPLE *t*-TEST FOR MEANS

- Do two samples come from the same population?
- If they come from different populations, what is the difference in the means of the two populations?
  - Does the average cost of a two-bedroom flat differ between Bengaluru and Hyderabad? What is the difference?
  - What is the difference in the strength of steel produced under two different temperatures?
  - Does the effectiveness of Head & Shoulders anti-dandruff shampoo differ from Pantene anti-dandruff shampoo?
  - What is the difference in the productivity of men and women on an assembly line under certain conditions?
  - Does an antibiotic affect the efficacy of another drug being taken by a patient?

The Central Limit Theorem states that the difference in two sample means, $\overline{x_1} - \overline{x_2}$, is normally distributed for large sample sizes (both n_1 and $n_2 \geq 30$) whatever the population distribution.

Also, $\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$ *[Recall E(X-Y)=E(X)-E(Y)]*

and $\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$ *[Recall Var(X-Y)=Var(X)+Var(Y)]*

$$z = \frac{observed\ difference - expected\ difference}{SE\ of\ the\ difference} = \frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}}$$

This is the test statistic for a 2-sample z-test.

Two-Sample t-Test for Unpaired Data

$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$

Test statistic, $t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$

Assuming the two samples come from populations with the **same standard deviation** (Rule of thumb: The ratio between the higher *s* and the lower *s* is less than 2), pooled variance can be used to calculate SE.

Two-Sample t-Test for Unpaired Data

$$s_p{}^2 = \frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{(n_1 - 1) + (n_2 - 1)} \rightarrow \left[\frac{Total\ Variation}{Total\ df} = Variance\right]$$

$$t = \frac{(\overline{x_1} - \overline{x_2})}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ with $(n_1 + n_2 - 2)$ degrees of freedom.

Note: ***Variance*** is average variation. ***Variation*** is sum of squared deviations. The average is computed by dividing the **total variation** by the **degrees of freedom**. So, total variation is obtained by multiplying variance with degrees of freedom.

Two-Sample t-Test for Unpaired Data

Welch's t-test using Welch-Satterthwaite equation for df

$H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$; Test statistic, $t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$

for **unequal standard deviations** for the two populations.

The degrees of freedom in this case are calculated as:

$$\nu = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\left[\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}\right]}$$, rounded off to the nearest integer.

R code: t.test(data1, data2, alternative="two.sided", var.equal=FALSE)

Hypothesis Testing

Antibiotic rifampicin increases the amount of drug metabolizing enzyme present in the liver. This causes increase in the rate of elimination of a lot of other drugs.

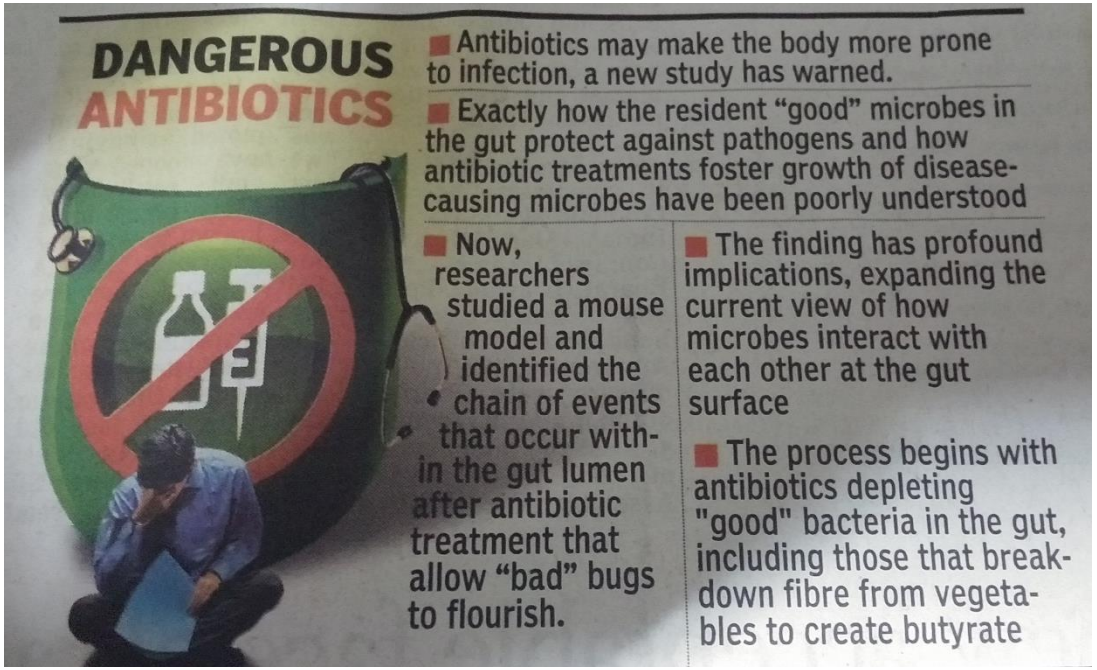


Image Source: Deccan Chronicle, Hyderabad edition, May 04, 2016

An experiment was conducted to study whether rifampicin affects the metabolic removal of the anti-asthma drug theophylline. A high elimination rate would mean inadequate treatment of the patient's asthma.

Hypothesis Testing

Two groups of 15 subjects were pre-treated with oral rifampicin (600 mg daily for 10 days) and a placebo, respectively. All of them were then given intravenous injection of theophylline (3 mg/kg of body weight).

Drug content was then measured from the blood samples and efficiency of removal of theophylline reported as clearance (in ml/min/kg).

Hypothesis Testing

Clearance of theophylline (ml/min/kg)					
Control Subjects			Treated Subjects		
0.81	0.56	0.46	1.15	1.15	0.92
1.06	0.45	0.43	1.28	0.72	0.67
0.43	0.88	0.37	1.00	0.79	0.76
0.54	0.73	0.73	0.95	0.67	0.82
0.68	0.43	0.93	1.06	1.21	0.82

$n_2 = 15$

$\overline{x_2} = 0.633$

$s_2 = 0.216$

$s_2{}^2 = 0.0467$

$n_1 = 15$

$\overline{x_1} = 0.931$

$s_1 = 0.202$

$s_1{}^2 = 0.0408$

Hypothesis Testing

What is the null hypothesis?

$H_0: \mu_1 - \mu_2 = 0$ (Rifampicin does not cause a change in theophylline clearance)

What is the alternative hypothesis?

$H_1: \mu_1 - \mu_2 \neq 0$

Is it a one-tailed test or a two-tailed test?

Two-tailed

What could be a possible hypothesis for a one-tailed test?

Rifampicin decreases theophylline clearance.

Hypothesis Testing

At α = 0.05, determine if there is a significant difference between the two groups.

$$s_p{}^2 = \frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{(n_1-1)+(n_2-1)}; t = \frac{(\overline{x_1}-\overline{x_2})}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ with } (n_1+n_2-2) \text{ df.}$$

$$s_p{}^2 = \frac{(15-1)*0.0408+(15-1)*0.0467}{(15-1)+(15-1)} = 0.04375; s_p = 0.209$$

$$t = \frac{0.931-0.633}{0.209*\sqrt{\frac{1}{15}+\frac{1}{15}}} = 3.9088$$

1-pt(3.9088,28) = 0.0002682. As this is less than 0.025 (two-tailed test), we reject the null hypothesis.

Note: All software will report twice this calculated value so that you can simply compare with 0.05 and not worry about having to compare with 0.025 in two-tailed tests.

Hypothesis Testing

Will you reject the null hypothesis or fail to do so?

Reject. That means rifampicin does affect theophylline clearance.

Does it increase or decrease theophylline clearance and by how much?

As the treated patients showed a higher clearance (0.931 ml/min/kg) compared to the control group (0.633 ml/min/kg), rifampicin increases clearance by about 0.298 ml/min/kg).

Confidence Intervals

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Rewriting:

$$(\overline{x_1} - \overline{x_2}) - t_{critical} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{x_1} - \overline{x_2}) + t_{critical} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$0.298 - 2.048 * 0.0763 \leq \mu_1 - \mu_2 \leq 0.298 + 2.048 * 0.0763$$

95% CI: (0.142, 0.454)

$t_{critical}$ = qt(0.05, 28)

Note zero difference is unlikely as at 95% Confidence Level, the difference ranges between 0.142 and 0.454 ml/min/kg, with a point estimate for the difference in mean clearance being 0.298 ml/min/kg.

Confidence Intervals and Hypothesis Test – R Output

ttest2 <- t.test(rifampicinT, rifampicinC, conf.level = 0.95, var.equal = TRUE)

```
        Two Sample t-test

data:  rifampicinT and rifampicinC
t = 3.9088, df = 28, p-value = 0.0005364
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1421489 0.4551844
sample estimates:
mean of x mean of y
0.9313333 0.6326667
```

HYPOTHESIS TEST

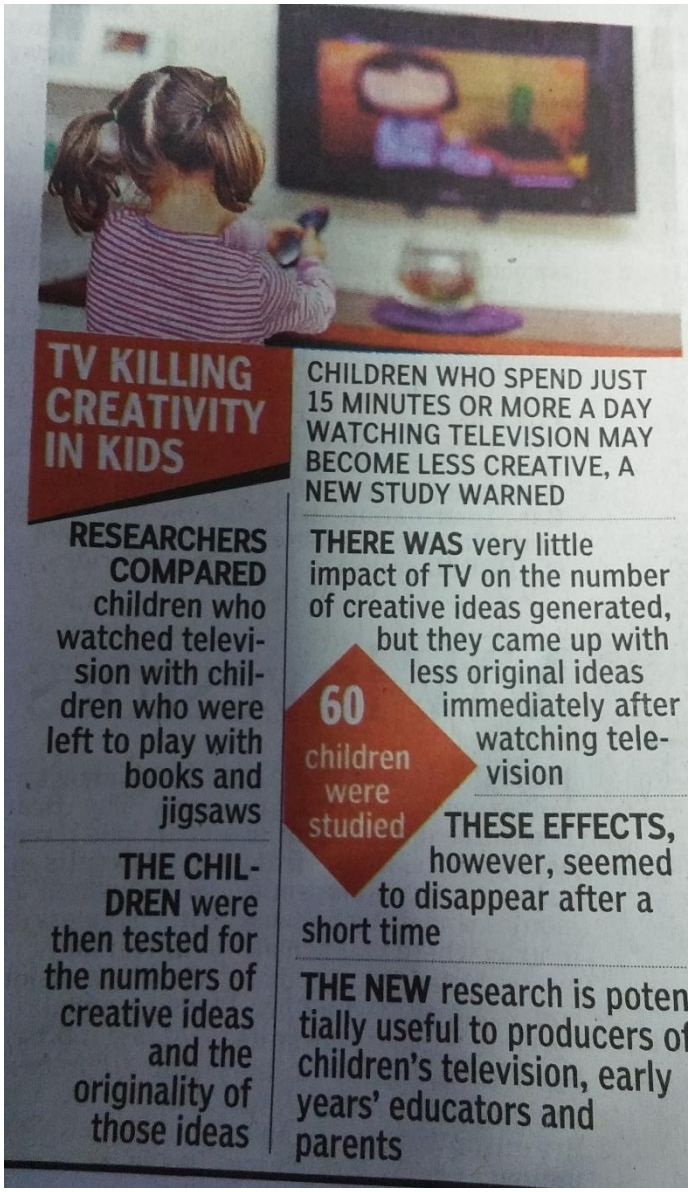
CONFIDENCE INTERVAL

Two-Sample t-Test for Paired Data

When the effects of two alternative treatments is to be compared, sometimes it is possible to make comparisons in pairs, where, e.g., the pair can be the same person at two different occasions or matched pairs where they are alike in all respects.

In unpaired t-test, **difference in means** is studied. In paired t-test, **mean of the differences** is studied.

Two-Sample t-Test for Paired Data



TV KILLING CREATIVITY IN KIDS

CHILDREN WHO SPEND JUST 15 MINUTES OR MORE A DAY WATCHING TELEVISION MAY BECOME LESS CREATIVE, A NEW STUDY WARNED

RESEARCHERS COMPARED children who watched television with children who were left to play with books and jigsaws

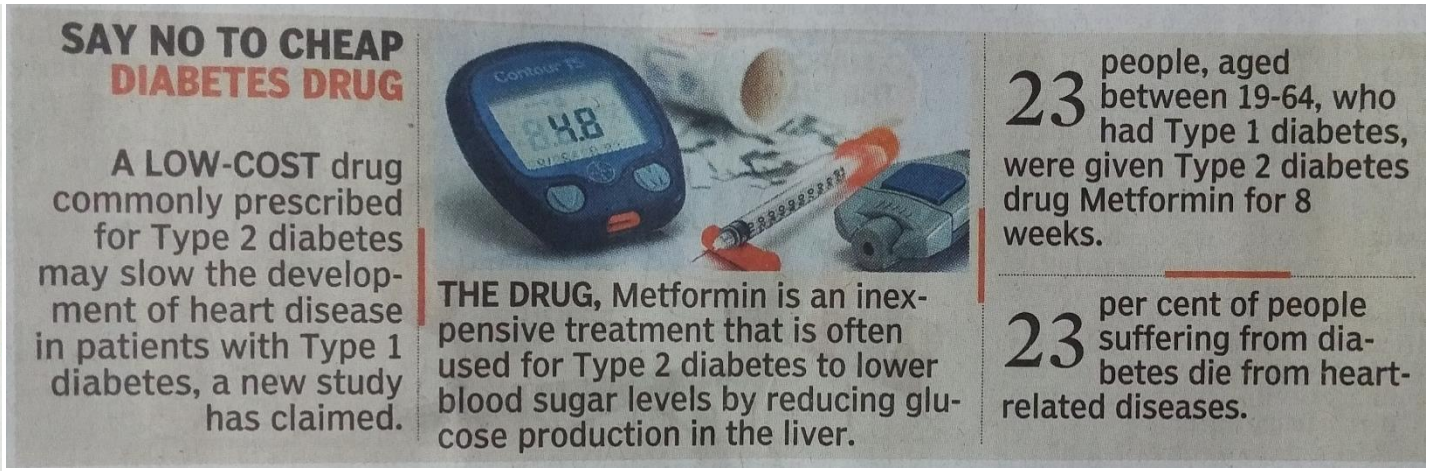
THE CHILDREN were then tested for the numbers of creative ideas and the originality of those ideas

THERE WAS very little impact of TV on the number of creative ideas generated, but they came up with less original ideas immediately after watching television

60 children were studied

THESE EFFECTS, however, seemed to disappear after a short time

THE NEW research is potentially useful to producers of children's television, early years' educators and parents



SAY NO TO CHEAP DIABETES DRUG

A LOW-COST drug commonly prescribed for Type 2 diabetes may slow the development of heart disease in patients with Type 1 diabetes, a new study has claimed.

THE DRUG, Metformin is an inexpensive treatment that is often used for Type 2 diabetes to lower blood sugar levels by reducing glucose production in the liver.

23 people, aged between 19-64, who had Type 1 diabetes, were given Type 2 diabetes drug Metformin for 8 weeks.

23 per cent of people suffering from diabetes die from heart-related diseases.

Two-Sample t-Test for Paired Data

How the study was carried out

Researchers studied a treatment group of 23 people aged 19-64 who had Type 1 diabetes for up to 23 years and had no evidence of heart disease.

Patients were given metformin at a dose they could tolerate, between one to three tablets a day, for eight weeks. Participants were advised to adjust their insulin to keep blood glucose levels safe.

Scientists measured patients' stem cells directly in the blood and also grew stem cells in a test tube, observing how they behaved. Another cell type was also counted to assess damaged blood vessels.

The participants were matched with nine patients within the same age bracket who took standard insulin treatment and 23 healthy non-diabetic people aged 20-64.

Experts found that the stem cells of patients who took metformin were able to promote the repair of the blood vessels and there was an improvement in how vascular stem cells worked.

Source: http://www.ncl.ac.uk/press/news/2016/08/metformintype1diabetes/
Last accessed: September 08, 2016

Two-Sample t-Test for Paired Data

A random sample of 54 people with high BP, high cholesterol and low vitamin-B12 with comparable severity in 30-50 age group is given two treatments. In Treatment A, 27 people were prescribed alternate day Atorva 5mg for cholesterol. In Treatment B, the other 27 people were additionally prescribed daily Vitamin-B12 dosage.

The BP of these 54 people was recorded 15 days after the start of the treatment. The recorded BP was taken as an average over 3 days.

The clinician wants to test at 5% Significance Level (or 95% Confidence Level) if B12 has any impact on BP.

Problem adapted from Sridhar Pappu's BP data based on his treatment cycle.

Two-Sample t-Test for Paired Data

Patient	Systolic (mmHg)		Difference (d)
1	145	120	25
2	130	135	-5
3	120	120	0
4	132	125	7
5	132	125	7
6	132	135	-3
7	142	123	19
8	120	128	-8
9	138	128	10
10	135	122	13
11	138	130	8
12	135	120	15
13	125	130	-5
14	135	125	10
15	135	122	13
16	135	125	10
17	135	125	10
18	132	120	12
19	130	120	10
20	120	125	-5
21	125	120	5
22	130	130	0
23	135	125	10
24	135	120	15
25	135	118	17
26	140	118	22
27	130	125	5

Mean Difference = 8.04

n = 27, NOT 54

Two-Sample t-Test for Paired Data

Mean of the differences, $\bar{d} = 8.04$

Standard Deviation of the differences, $s_d = 8.56$

Standard Error of the mean, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = 1.65$

$$t = \frac{\bar{d}}{SE(\bar{d})} = \frac{8.04}{1.65} = 4.87633462$$

$t_{0.025,26} = 2.05553$ *R: qt(0.025, 26, lower.tail = FALSE)*

p-value = 0.000023 *R: pt(4.87633462, 26, lower.tail = FALSE)*

Comparing the absolute t-value with the critical t-value (or corresponding p-value with the significance level), we cannot reject the null hypothesis that vitamin-B12 has NO effect on BP.

Two-Sample t-Test for Paired Data

The 95% CI for the mean difference is given by $\bar{d} \pm t_{(\frac{\alpha}{2},\nu)} * SE(\bar{d})$

$8.04 - 2.05553 * 1.65 \leq D \leq 8.04 + 2.05553 * 1.65$

95% CI: (4.65, 11.42).

As zero is NOT included in the CI, we can reject the null hypothesis that the difference between the two treatments is 0.

Business Decision

Doctors should evaluate if B12 deficiency is a factor before prescribing BP medication solely on the evidence of high BP.

Research on the subject: Courtesy Tanzeem Ahmed Nayaz, Batch 42

- https://academic.oup.com/ajh/article/24/11/1215/2281951
- https://www.ncbi.nlm.nih.gov/pubmed/26147383
- https://www.heartlungcirc.org/article/S1443-9506(08)00387-9/abstract

CI and HT – R Output

BREAK

ttest2P <- t.test(treatmentA, treatmentB, paired = TRUE, conf.level = 0.95)

```
        Paired t-test

data:  treatmentA and treatmentB
t = 4.8763, df = 26, p-value = 4.658e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.649171 11.424903
sample estimates:
mean of the differences
                8.037037
```

Chi Square Distribution

χ^2 DISTRIBUTION

So you modeled a situation using a probability distribution and got a good idea of how things will shape up in the long run. But what if what you expected and what you observed are not the same? How would you know if the difference is due to normal fluctuations or if your model was incorrect?

Let us say you are running a casino and the slot machines are causing you headaches. You had designed them with the following expected probability distribution, with X being the net gain from each game played.

x	-2	23	48	73	98
P(X=x)	0.977	0.008	0.008	0.006	0.001

You collected some statistics and found the following frequency (1000 players) of peoples' winnings.

x	-2	23	48	73	98
Frequency	965	10	9	9	7

You want to compare the actual frequency with the expected frequency.

x	-2	23	48	73	98
P(X=x)	0.977	0.008	0.008	0.006	0.001

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

Are these differences significant and if they are, is it just pure chance?

χ^2 test to the rescue

χ^2 distribution uses a test statistic to look at the difference between the expected and the actual, and then returns a probability of getting observed frequencies as extreme.

$X^2 = \sum \frac{(O-E)^2}{E}$, where O is the observed frequency and E the expected frequency.

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$$X^2 = \frac{(965-977)^2}{977} + \frac{(10-8)^2}{8} + \frac{(9-8)^2}{8} + \frac{(9-6)^2}{6} + \frac{(7-1)^2}{1}$$

$$\boldsymbol{X^2 = 38.272}$$

Value in my X- Axis. Is this high?

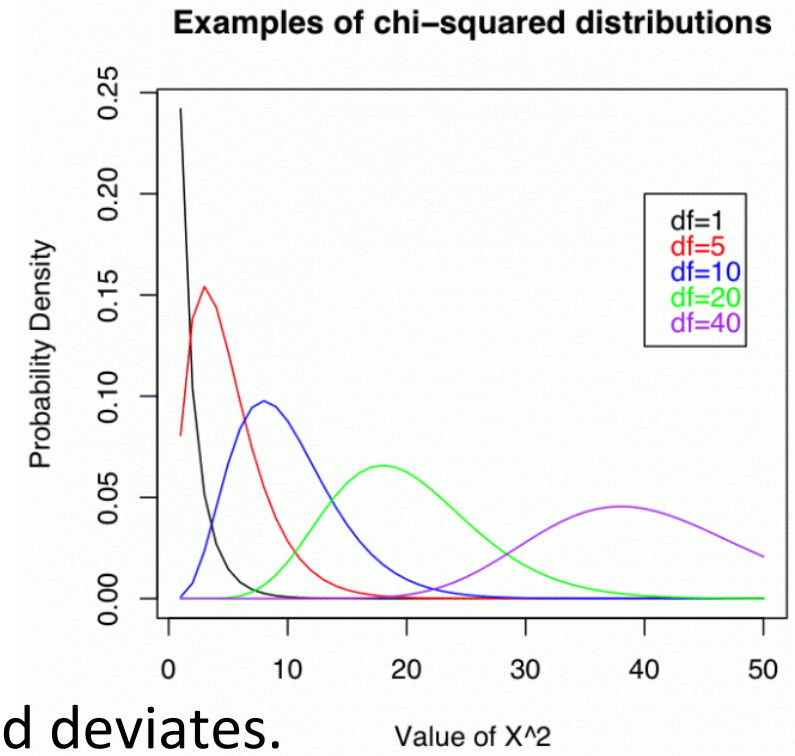
To find this, we need to look at the χ^2 distribution.

χ^2 distribution

$$Recall\ Z = \frac{X - \mu}{\sigma}$$

$$Z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

$$Z^2 = \chi^2_{(1)}$$



Thus χ^2 distribution is a distribution of the squared deviates.

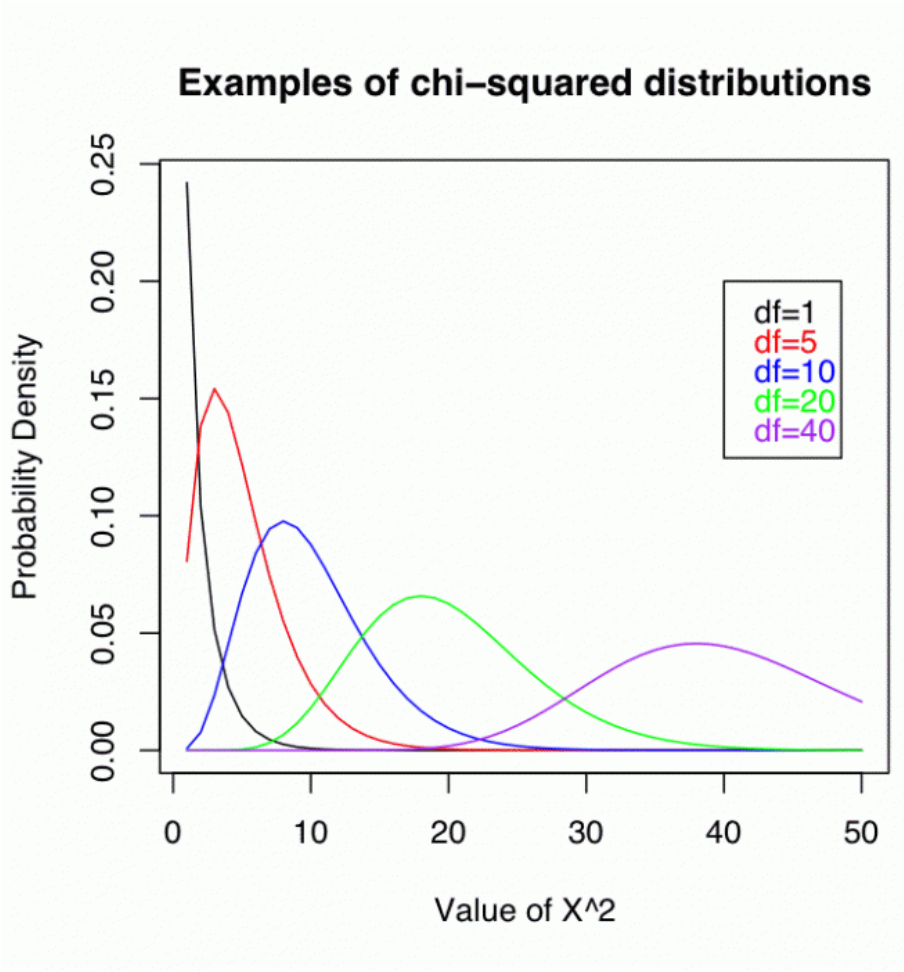
For example, $\chi^2_{(3)} = Z_1^2 + Z_2^2 + Z_3^2$, where Z_1, Z_2 and Z_3 are independent standard normal variables.

The shape depends on number of squared deviates added together.

χ^2 distribution

$X^2 \sim \chi^2_{(\nu)}$, where ν represents the degrees of freedom.

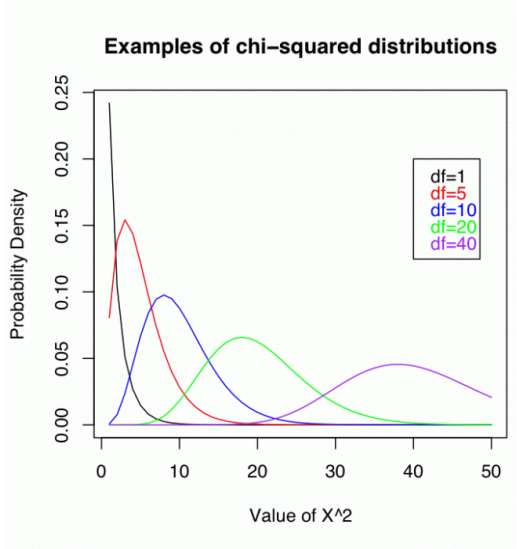
When ν is greater than 2, the shape of the distribution is skewed positively gradually becoming approximately normal for large ν.



Properties of X^2 random variable

- A X^2 random variable takes values between 0 and ∞.
- Mean of a χ^2 distribution is ν.
- Variance of a χ^2 distribution is 2ν.
- The shape of the distribution is skewed to the right.
- As ν increases, Mean gets larger and the distribution spreads wider.
- As ν increases, distribution tends to normal.

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1



In the above case, we had 5 frequencies to calculate. However, since the TOTAL expected frequency has to be equal to the TOTAL observed frequency (**RESTRICTION**), calculating 4 would give the 5^{th}. Therefore, there are 5-1=4 degrees of freedom.

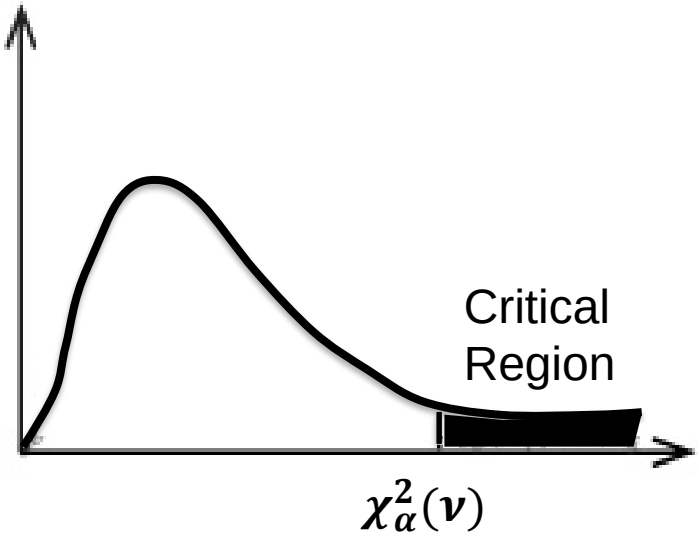
ν = **(number of classes) – (number of restrictions)**, or

ν = **(number of classes) – 1 – (number of parameters being estimated from sample data)**

How do we know the Significance of the difference?

One-tailed test using the upper tail of the distribution as the critical region.

A test at significance level α is written as $\chi^2_\alpha(\nu)$.
The critical region is to its right.



Why?

Higher the value of the test statistic, the bigger the difference between observed and expected frequencies.

Uses of χ^2 distribution

- To test **goodness of fit**.
- To test **independence** of two variables.
- To test hypothesis about **variance** of a population.

Steps to test Goodness-of-fit

You want to see if there is sufficient evidence at the 5% significance level to say the slot machines have been rigged.

What are the null and alternate hypotheses?

H_0: The slot machine winnings per game follow the described probability distribution, i.e., they are not rigged.

H_1: The slot machine winnings per game do not follow this distribution.

What are the expected frequencies and degrees of freedom?

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

$$\nu = 4$$

What is the critical region?

$\chi^2_{5\%}(4) = 9.488$. This means the critical region is $X^2 > 9.488$.

R code: qchisq(0.95,4) or qchisq(0.05,4,lower.tail=FALSE)

Note the syntax is exactly the same as for t-distribution: qt(p,df)

Is the test statistic inside or outside the critical region?

Since $X^2 = 38.27$ and the critical region is $X^2 > 9.488$, this means X^2 is **inside** the critical region.

Will you accept or reject the null hypothesis?

Reject. There is sufficient evidence to reject the hypothesis that the slot machine winnings follow the described probability distribution.

This sort of hypothesis test is called a **goodness of fit** test. This test is used whenever you have a set of values that should fit a distribution, and you want to test whether the data actually does.

χ^2 goodness of fit works for any probability distribution

Distribution	Condition	ν
Binomial	You know *p* (probability of success or the proportion of successes in a population)	$\nu = n - 1$
	You don't know *p* and have to estimate it from observed frequencies	$\nu = n - 2$
Poisson	You know λ	$\nu = n - 1$
	You don't know λ, and have to estimate it from observed frequencies	$\nu = n - 2$
Normal	You know μ and σ^2	$\nu = n - 1$
	You don't know μ and σ^2, and have to estimate them from observed frequencies	$\nu = n - 3$

The 108 Medical Emergency Service received calls during 150 5-minute intervals as follows. Is the distribution Poisson at $\alpha=0.01$?

# of calls per 5-min interval	Frequency
0	18
1	28
2	47
3	21
4	16
5	11
6 or more	9

Step 1: Decide H_0 and H_1

H_0: The frequency distribution is Poisson.

H_1: The frequency distribution is not Poisson.

Step 2: Find expected frequencies and degrees of freedom

# of calls per 5-min interval	Observed Frequency	Total Calls = # of calls/interval X Frequency
0	18	0
1	28	28
2	47	94
3	21	63
4	16	64
5	11	55
6 or more	9	54
TOTAL	**150**	**358**

Probability = $\frac{Expected\ Frequency}{Total\ \ Frequency}$

Expected frequency = Probability * Total frequency

$$P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

(r = 0, 1, 2, 3, 4, 5, 6)

$$\lambda = Avg\ no.\ of\ calls\ per\ 5\ \text{min}\ interval = \frac{Total\ \#\ of\ calls}{Total\ \#\ of\ intervals}$$

$$\lambda = \frac{358}{150} = 2.39$$

Step 2: Find expected frequencies and degrees of freedom

Expected frequencies are obtained by multiplying expected probabilities by the total frequency. $P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$, where $\lambda = 2.39$ and r =(0,1,2,3,4,5,6)

# of calls per 5-min interval	Expected Probability	Expected Frequency
0	0.0916	13.74
1	0.2190	32.85
2	0.2617	39.25
3	0.2085	31.27
4	0.1246	18.69
5	0.0595	8.93
6 or more	0.0351	5.26
TOTAL		**150.00**

How many df?

ν = 7-2 = 5

From Sample we lose one degree (n-1) but I also do not know λ so I lose another degree of free. Hence (n-2)

Step 3: Determine the critical region

$\chi^2_{1\%}(5) = 15.086$. This means the critical region is $X^2 > 15.086$.

R code: qchisq(0.99,5) or qchisq(0.01,5,lower.tail=FALSE)

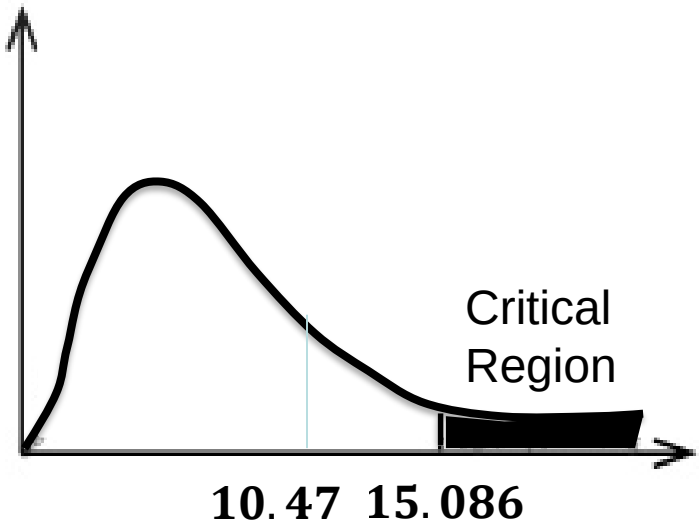
Step 4: Calculate the test statistic X^2

# of calls per 5-min interval	Observed Frequency	Expected Frequency	$\frac{(O-E)^2}{E}$
0	18	13.74	1.32
1	28	32.85	0.72
2	47	39.25	1.53
3	21	31.27	3.37
4	16	18.69	0.39
5	11	8.93	0.48
6 or more	9	5.26	2.66
TOTAL	**150**	**150**	**10.47**

$X^2 = 10.47$

Step 5: See whether the test statistic is in the critical region

X^2 = 10.47, which is less than the critical value of 15.086. It is NOT in the critical region. **Alternatively**, the p-value for the sample is more than the significance level, indicating the sample is NOT in the critical region.



Step 6: Make your decision

There is not enough evidence to reject the null hypothesis that the distribution is Poisson.

Business Implication: Now that 108 services management knows that the distribution is Poisson, it can plan the staffing of the call centre more efficiently.

χ^2 independence test

Your casino is facing another issue. You think you are losing more money from one of the croupiers on the blackjack tables. You want to test if the outcome of the game is dependent on which croupier is leading the game.

χ^2 independence test



	Croupier A	Croupier B	Croupier C
Win	43	49	22
Draw	8	10	5
Lose	47	44	30

Possible Outcomes

Observed Results

CSE 7315c

χ^2 independence test

The process is the same as before. The null hypothesis assumes that choice of croupier is independent of the outcome, and is rejected if there is sufficient evidence against it.

However, a **contingency table** has to be drawn to find the expected frequencies using probability.

χ^2 independence test

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	10	5	23
Lose	47	44	30	121
Total	98	103	57	258

$$P(Win) = \frac{Total\ Wins}{Grand\ Total} = \frac{114}{258}$$

$$P(A) = \frac{Total\ A}{Grand\ Total} = \frac{98}{258}$$

If croupier and the outcome are independent,

$$P(Win\ and\ A) = \frac{Total\ Wins}{Grand\ Total} X \frac{Total\ A}{Grand\ Total}$$

χ^2 independence test

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	10	5	23
Lose	47	44	30	121
Total	98	103	57	258

$$Expected\ Frequency\ of\ Win\ and\ A$$

$$= Grand\ Total\ X\ \frac{Total\ Wins}{Grand\ Total}\ X\ \frac{Total\ A}{Grand\ Total} = \frac{Total\ Wins\ X\ Total\ A}{Grand\ Total}$$

$$= \frac{Row\ Total\ X\ Column\ Total}{Grand\ Total}$$

χ^2 independence test – Finding expected frequencies

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	10	5	23
Lose	47	44	30	121
Total	98	103	57	258

	Croupier A	Croupier B	Croupier C
Win	(114*98)/258	(114*103)/258	(114*57)/258
Draw	(23*98)/258	(23*103)/258	(23*57)/258
Lose	(121*98)/258	(121*103)/258	(121*57)/258

χ^2 independence test – Calculating X^2

	Observed	Expected	$\frac{(O-E)^2}{E}$
A	43	43.302	0.0021
	8	8.736	0.0621
	47	45.961	0.0235
B	49	45.512	0.2674
	10	9.182	0.0728
	44	48.306	0.3839
C	22	25.186	0.4030
	5	5.081	0.0013
	30	26.733	0.3994
	$\sum O = 258$	$\sum E = 258$	$\sum \frac{(O-E)^2}{E} = 1.6155$

χ^2 independence test – Calculating ν

	Croupier A	Croupier B	Croupier C
Win			
Draw			
Lose			

We calculated 9 but really need to calculate 4 and figure out the rest using the total frequency of each row and column. In general, the degrees of freedom will be (*m*-1)(*n*-1) where *m* is the number of columns and *n* the number of rows.

χ^2 independence test – Determine critical region

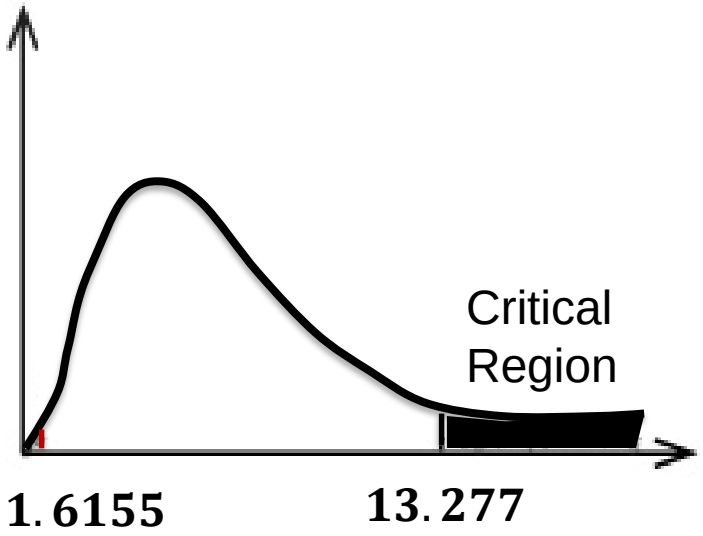
Let us say we need 1% significance level to see if the outcome is independent of the croupier.

$\chi^2_{1\%}(4) = 13.277$. This means the critical region is $X^2 > 13.277$.

R code: qchisq(0.99,4) or qchisq(0.01,4,lower.tail=FALSE)

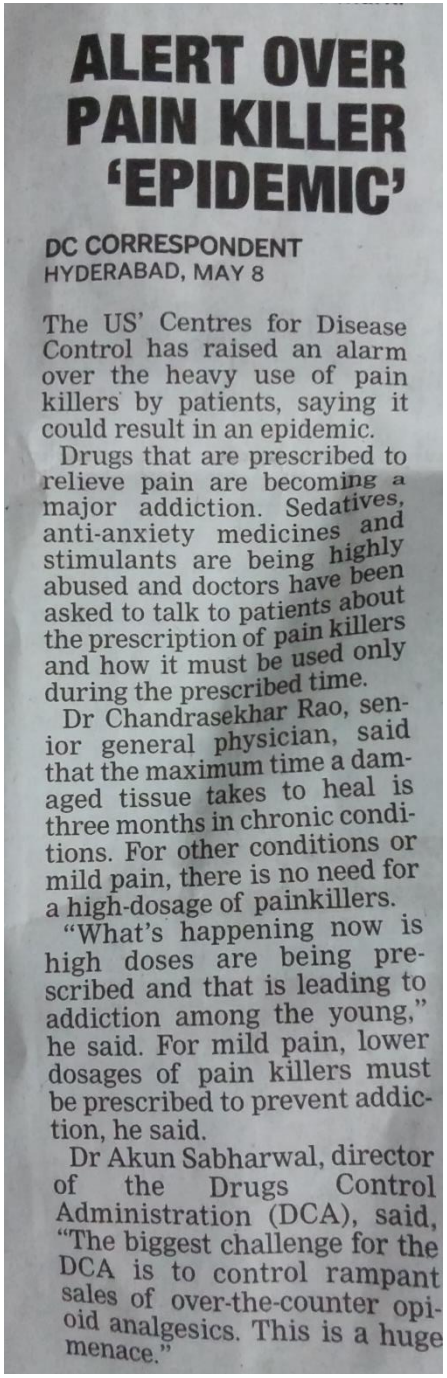
χ^2 independence test – Decision

Since calculated $X^2 = 1.6155$, it is outside the critical region, and hence we accept the null hypothesis.



χ^2 independence test

There is widespread abuse of prescription drugs for a number of reasons like getting high, reducing appetite, relieving tension, feeding an addiction, etc.



ALERT OVER PAIN KILLER 'EPIDEMIC'

DC CORRESPONDENT
HYDERABAD, MAY 8

The US' Centres for Disease Control has raised an alarm over the heavy use of pain killers by patients, saying it could result in an epidemic.

Drugs that are prescribed to relieve pain are becoming a major addiction. Sedatives, anti-anxiety medicines and stimulants are being highly abused and doctors have been asked to talk to patients about the prescription of pain killers and how it must be used only during the prescribed time.

Dr Chandrasekhar Rao, senior general physician, said that the maximum time a damaged tissue takes to heal is three months in chronic conditions. For other conditions or mild pain, there is no need for a high-dosage of painkillers.

"What's happening now is high doses are being prescribed and that is leading to addiction among the young," he said. For mild pain, lower dosages of pain killers must be prescribed to prevent addiction, he said.

Dr Akun Sabharwal, director of the Drugs Control Administration (DCA), said, "The biggest challenge for the DCA is to control rampant sales of over-the-counter opioid analgesics. This is a huge menace."

χ^2 independence test

The National Council on Alcoholism and Drug Dependence wants to understand if there is dependence of the type of prescription drug abuse on the age of the patient. A random poll of 309 patients is taken as shown below. At $\alpha = 0.01$, are the two variables independent?

	Pain relievers	Tranquilizers /Sedatives	Stimulants	TOTAL
21-34	26	95	18	139
35-55	41	40	20	101
>55	24	13	32	69
TOTAL	91	148	70	309

Step 1: Decide H_0 and H_1

H_0: Type of prescription drug abuse is independent of age.

H_1: Type of prescription drug abuse is not independent of age.

Step 2: Find expected frequencies and degrees of freedom

OBSERVED

	Pain relievers	Tranquilizers/ Sedatives	Stimulants	TOTAL
21-34	26	95	18	139
35-55	41	40	20	101
>55	24	13	32	69
TOTAL	91	148	70	309

EXPECTED

	Pain relievers	Tranquilizers/ Sedatives	Stimulants	TOTAL
21-34	40.94	66.58	31.49	139
35-55	29.74	48.38	22.88	101
>55	20.32	33.05	15.63	69
TOTAL	91	148	70	309

$\nu = (m-1)*(n-1)$

$m=3, n=3$

Where m = number of columns and n = number of rows

$\nu = 4$

Step 3: Determine the critical region

$\chi^2_{1\%}(4) = 13.277$. This means the critical region is $X^2 > 13.277$.

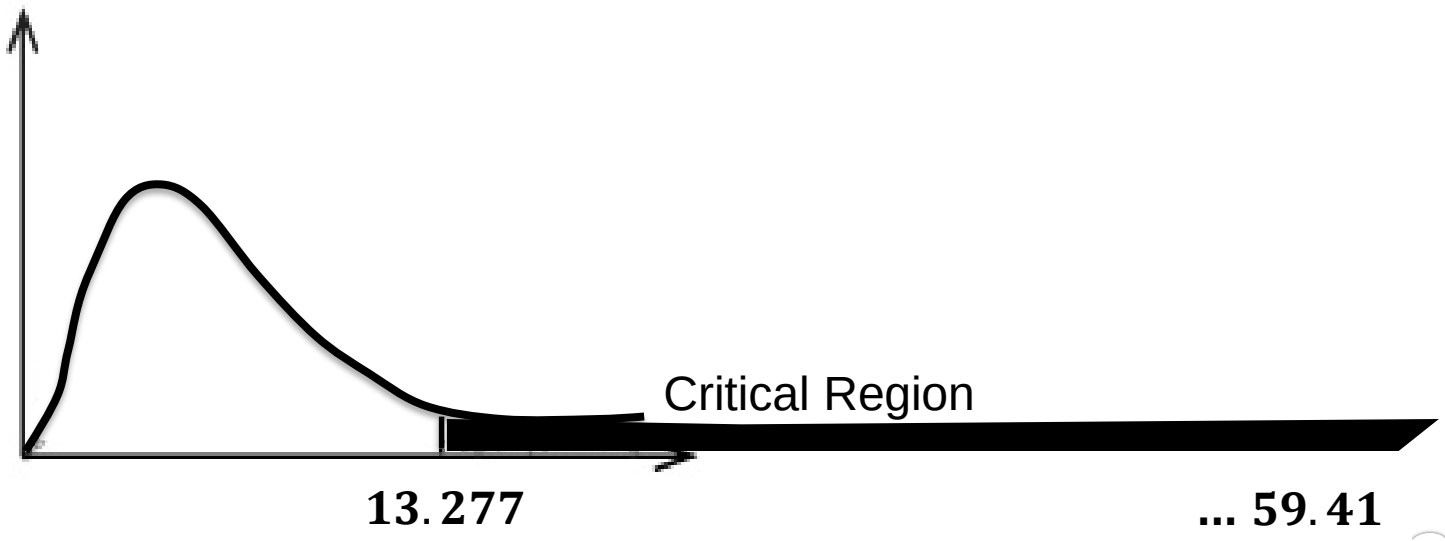
R code: qchisq(0.99,4) or qchisq(0.01,4,lower.tail=FALSE)

Step 4: Calculate the test statistic X^2

	Observed	Expected	$\frac{(O-E)^2}{E}$
Pain relievers	26	40.94	
	41	29.74	
	24	20.32	
Tranquilizers/ Sedatives	95	66.58	
	40	48.38	
	13	33.05	
Stimulants	18	31.49	
	20	22.88	
	32	15.63	
	$\sum O = 309$	$\sum E = 309$	$\sum \frac{(O-E)^2}{E} = 59.41$

Step 5: See whether the test statistic is in the critical region

X^2 = 59.41, which is greater than the critical value of 13.277. It is in the critical region.



Step 6: Make your decision

There is enough evidence to reject the null hypothesis that the type of prescription drug abuse and age are independent.

Testing Hypotheses about a Variance

Sample estimate of population variance is given by

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Multiplying the variance estimate by *n-1* gives the sum of squares. Dividing by population variance gives a random variable distributed as chi-squared with *n-1* degrees of freedom.

$$\therefore \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Recall $\chi^2 = \frac{\sum(x_i - \bar{x})^2}{\sigma^2}$

Testing Hypotheses about a Variance

A manufacturing company produces bearings of 2.65 cm in diameter. A major customer requires that the variance in diameter be no more than 0.001 cm^2. The manufacturer tests 20 bearings using a precise instrument and gets the below values. Assuming the diameters are normally distributed, can the population of these bearings be rejected due to high variance at 1% significance level?

Data: 2.69, 2.66, 2.64, 2.59, 2.62, 2.63, 2.69, 2.66, 2.63, 2.65, 2.57, 2.63, 2.70, 2.71, 2.64, 2.65, 2.59, 2.66, 2.62, 2.57

Testing Hypotheses about a Variance

What are null and alternate hypotheses?

H_0: $\sigma^2 \leq 0.001$; H_1: $\sigma^2 > 0.001$

How many degrees of freedom?

Since n=20, df=19.

Testing Hypotheses about a Variance

What is the critical region?

$$\chi^2_{0.01,19} = 36.191$$

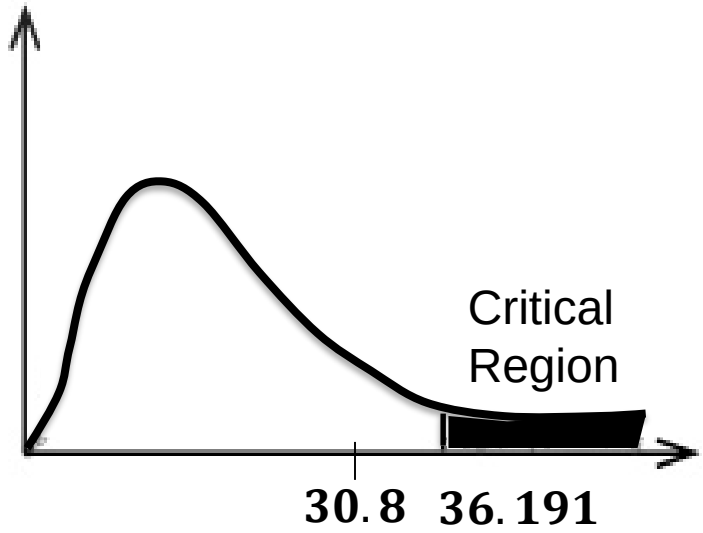
R code: qchisq(0.99,19) or qchisq(0.01,19,lower.tail=FALSE)

Testing Hypotheses about a Variance

BREAK

What is the observed χ^2 value?

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{19 * 0.001621}{0.001} = 30.8$$



Is it in critical region? $\chi^2_{0.01,19} = 36.191$

No.

Will you reject or fail to reject the null hypothesis?

Fail to reject.

What is the business decision?

The population variance is within specification limits required by the customer and hence the bearings can be shipped.

F DISTRIBUTION

F distribution

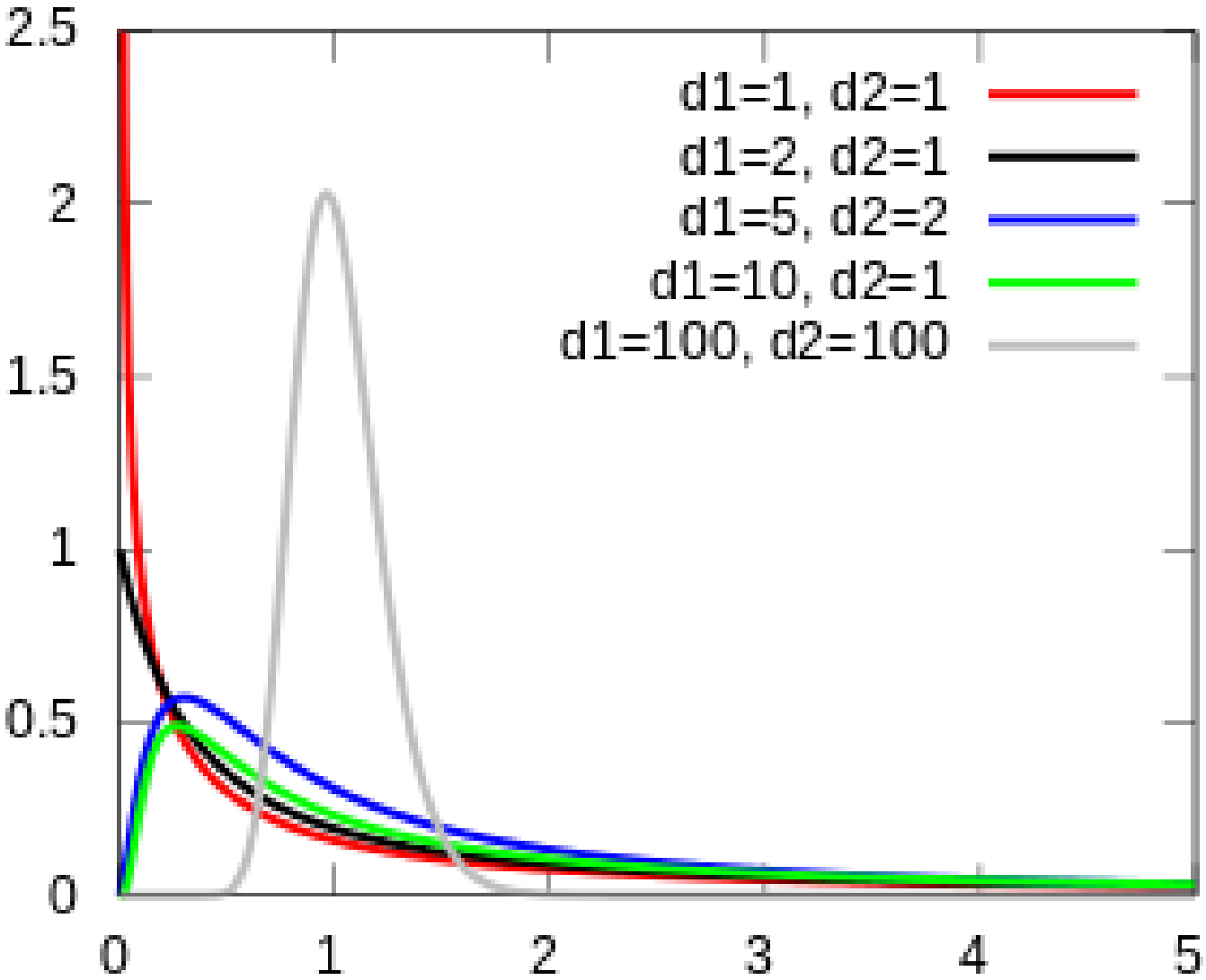
- χ^2 was useful in testing hypotheses about a single population variance.
- Sometimes we want to test hypotheses about difference in variances of two populations:
  - Is the variance of 2 stocks the same?
  - Do parts manufactured in 2 shifts or on 2 different machines or in 2 batches have the same variance or not?
  - Is the powder mix for tablet granulations homogeneous?
  - Is there variability in assayed drug blood levels in a bioavailability study?
  - Is there variability in the clinical response to drug therapy of two samples?

F distribution

- Ratio of 2 variance estimates: $F = \frac{s_1^2}{s_2^2} = \frac{est.\sigma_1^2}{est.\sigma_2^2}$
- Ideally, this ratio should be about 1 if 2 samples come from the same population or from 2 populations with same variance, but sampling errors cause variation.
- *Recall* $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$. So, F is also a ratio of 2 chi-squares, each divided by its degrees of freedom, i.e.,

$$F = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

F distribution



Hypothesis test for 2 sample variances

A machine produces metal sheets with 22mm thickness. There is variability in thickness due to machines, operators, manufacturing environment, raw material, etc. The company wants to know the consistency of two machines and randomly samples 10 sheets from machine 1 and 12 sheets from machine 2. Thickness measurements are taken. Assume sheet thickness is normally distributed in the population.

The company wants to know if the variance from each sample comes from the same population variance (population variances are equal) or from different population variances (population variances are unequal).

How do you test this?

Hypothesis test for 2 sample variances

Data

Machine 1		Machine 2	
22.3	21.9	22.0	21.7
21.8	22.4	22.1	21.9
22.3	22.5	21.8	22.0
21.6	22.2	21.9	22.1
21.8	21.6	22.2	21.9
		22.0	22.1
$s_1^2 = 0.11378$	$n = 10$	$s_2^2 = 0.02023$	$n = 12$

Ratio of sample variances, $F = \frac{s_1^2}{s_2^2} = \frac{0.11378}{0.02023} = 5.62$

Hypothesis test for 2 sample variances

What are null and alternate hypotheses?

$H_0: \sigma_1^2 = \sigma_2^2;\ H_1: \sigma_1^2 \neq \sigma_2^2$

Is it a one-tailed test or a two-tailed test?

Two-tailed.

What are numerator and denominator degrees of freedom?

$\nu_1 = 10 - 1 = 9; \nu_2 = 12 - 1 = 11$

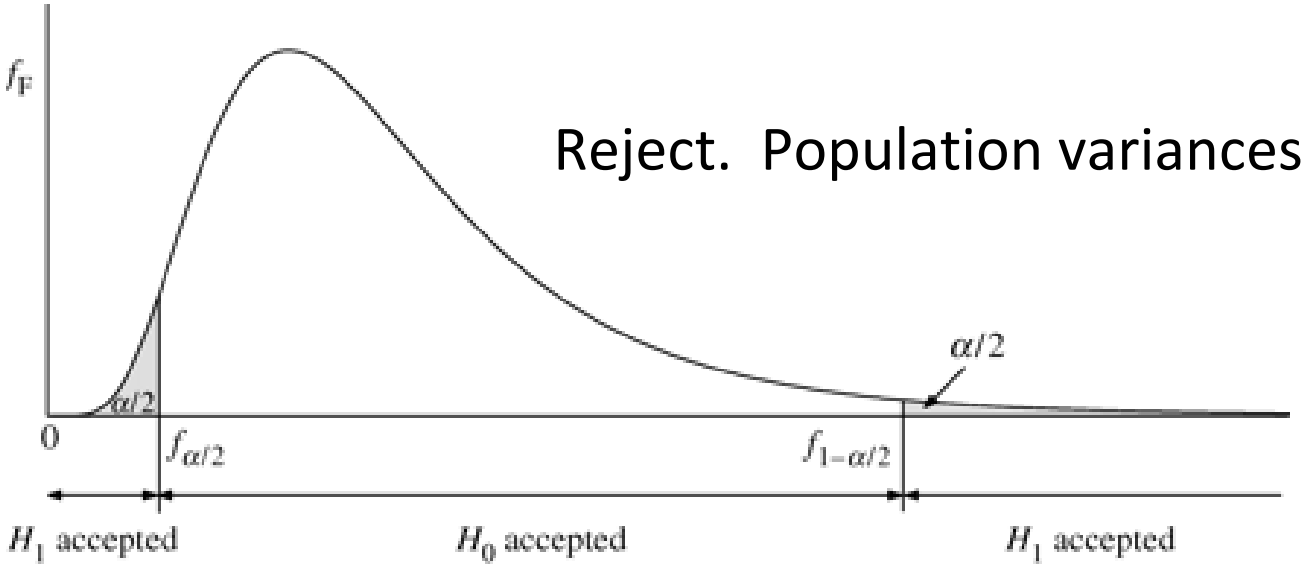
Hypothesis test for 2 sample variances

$F_{0.025,9,11} = 0.2556; F_{0.975,9,11} = 3.5879$ *R-code: qf(α, df_1, df_2)*

Hypothesis test for 2 sample variances

$F_{0.025,9,11} = 0.2556; F_{0.975,9,11} = 3.5879; F_{observed} = 5.62$

Will you reject the null hypothesis or not?



Reject. Population variances are not equal.

Hypothesis test for 2 sample variances

What are the business implications?

Variance in machine 1 is higher than in machine 2. Machine 1 needs to be inspected for any issues.

Applications of F Distribution

- Test for equality of variances.
- Test for differences of means in ANOVA.
- Test for regression models (slopes relating one continuous variable to another, e.g., Entrance exam scores and GPA)

Relations among Distributions – Children of the Normal

- χ^2 is drawn from the normal – N(0,1) deviates squared and summed.
- *F* is the ratio of 2 chi-squares, each divided by its *df*.
- A χ^2 divided by its *df* is a variance estimate, i.e., a sum of squares divided by the degrees of freedom.
- $F=t^2$. If you square *t*, you get an *F* with 1 *df* in the numerator, i.e., $t^2_{(v)} = F_{(1,v)}$

ANOVA

The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.

A pharmaceutical company tested 3 formulations of a migraine relief drug. 27 volunteers were randomly grouped in 3 groups. Each group was given a different drug formulation. The participants took the drug when they had the next migraine attack and recorded the pain on a scale of 1 to 10, 1 being no pain and 10 being extreme pain 30 minutes after taking the medicine.

We want to understand if the differences are due to within group differences or between group differences.

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56$ $\bar{X}_2 = 5.56$ $\bar{X}_3 = 5.78$

$$\bar{\bar{X}} = \frac{134}{27} = 4.96$$

$Total\ Sum\ of\ Squares, SST$
$= (2 - 4.96)^2 + 5 * (3 - 4.96)^2 + 4 * (4 - 4.96)^2 + 7 * (5 - 4.96)^2 + 5 * (6 - 4.96)^2$
$+ 4 * (7 - 4.96)^2 + (8 - 4.96)^2 = \mathbf{62.96}$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{mn - 1}$, since we can calculate one member knowing the overall mean.

How much of this variation is coming from within the groups and how much from between the groups?

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56 \quad \bar{X}_2 = 5.56 \quad \bar{X}_3 = 5.78 \quad \bar{\bar{X}} = \frac{134}{27} = 4.96$

$Total\ Sum\ of\ Squares\ Within, SSW$

$= (2-3.56)^2 + 4 * (3-3.56)^2 + 2 * (4-3.56)^2 + 2 * (5-3.56)^2 + (3-5.56)^2 + 2 * (4-5.56)^2 + (5-5.56)^2 + 2 * (6-5.56)^2 + 2 * (7-5.56)^2 + (8-5.56)^2 + 4 * (5-5.78)^2 + 3 * (6-5.78)^2 + 2 * (7-5.78)^2 = \mathbf{36.00}$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{m(n-1)}$, since we can calculate one member knowing the group mean.

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56 \quad \bar{X}_2 = 5.56 \quad \bar{X}_3 = 5.78 \quad \bar{\bar{X}} = \frac{134}{27} = 4.96$

$Total\ Sum\ of\ Squares\ Between, SSB$

$= 9 * (3.56 - 4.96)^2 + 9 * (5.56 - 4.96)^2 + 9 * (5.78 - 4.96)^2 = \mathbf{26.96}$

When there are *m* groups, the degrees of freedom are $\boldsymbol{m - 1}$.

SST = SSW + SSB

Also, for degrees of freedom, $\boldsymbol{mn - 1 = m(n - 1) + (m - 1)}$

What is the null hypothesis?

The population means of the 3 groups from which the samples were taken have the same mean, i.e., the drug formulations do not have different impacts on relieving migraine headache. $\mu_1 = \mu_2 = \mu_3$. Let us also have a significance level, $\alpha = 0.10$.

What is the alternate hypothesis?

At least one of the drug formulations has a different impact on migraine pain relief.

$$F - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.96}{2}}{\frac{36}{24}} = 8.9876$$

If numerator is much bigger than the denominator, it means variation **between** means has bigger impact than variation **within**, thus rejecting the null hypothesis.

The *df* are 2 for numerator and 24 for denominator.

R Code $F_{Critical}$ = qf(0.1,2,24) = 2.53833

F_c, the critical F-statistic, therefore, is 2.53833. 8.9876 is way higher than this and hence we reject the null hypothesis. That means at least one of the drug formulations has a different impact on migraine pain relief.

STOCK MARKET EXAMPLE

A stock analyst randomly selected 8 stocks from each of 3 industries, viz., Financial, Energy and Utilities. She compiled the 5-year rate of return for each stock.

The analyst wants to know if, at 0.05 Significance Level, there is a difference in the rate of return for any of the industries.

STOCK MARKET EXAMPLE

	5-Year Rates of Return		
	Financial	Energy	Utilities
	10.76	12.72	11.88
	15.05	13.91	5.86
	17.01	6.43	13.46
	5.07	11.19	9.9
	19.5	18.79	3.95
	8.16	20.73	3.44
	10.38	9.6	7.11
	6.75	17.4	15.7
xbar	**11.585**	**13.846**	**8.913**
s	**5.124**	**4.867**	**4.530**

What is the null hypothesis?

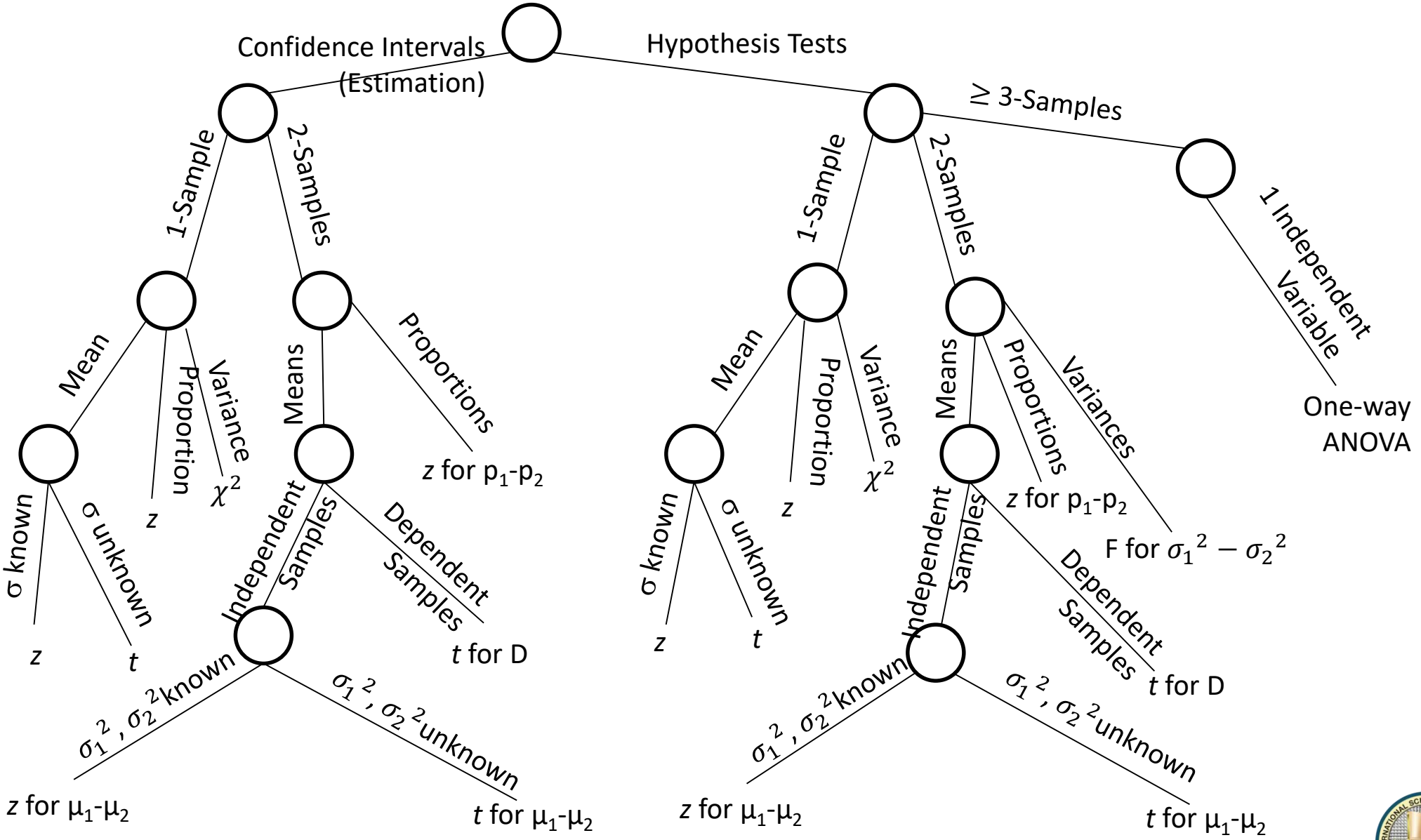
$\mu_1 = \mu_2 = \mu_3.\ \alpha = 0.05$

All 3 industries have the same average rate of return.

What is the alternate hypothesis?

At least one of the industries has a different rate of return than the others.

Tree Diagram Taxonomy of Inferential Techniques



Thought Process on When to Use a Particular Test

- What do you want to do?
  - Description – Summary statistics, Various plots, Correlations
  - Prediction – Linear regression, Logistic regression
  - Intervention (differences between groups) – *t*-test, Chi-square, F-test, ANOVA
- Is the dependent variable Categorical or Numerical?
  - Nominal – Chi-square, Logistic regression
  - Ordinal – Chi-Square
  - Dichotomous – Logistic regression
  - Numerical – *t*-test, ANOVA, Correlation, Multiple regression

THE STORY

We started by seeing how Statistics were all around us and how people misuse them. So, we wanted to understand data using statistics and to be able to make useful inferences using data.

After learning some important statistical terminology, we started understanding data by getting an average value to describe it. When Mean didn't work, we went to Median and then to Mode.

We then found that along with the average, we need to understand the spread because averages don't describe the data fully. We looked at Range, Interquartile Range, Variance and Standard Deviation.

We learned about the Box plot and how it can be used to identify outliers.

With basic understanding of statistics, we studied probability basics as it is the basis of all statistical inference. We learned Bayes Theorem. We also took an important detour to learn the Confusion Matrix as a tool to evaluate Classification models.

We then looked at variety of ways this data (or probabilities) is distributed and their properties, and looked at the expected values, their variance and the probabilities of various possible outcomes.

We studied both discrete and continuous probability distributions.

Then we saw how the Sampling Distributions of Means tend to normal distribution irrespective of how the population is distributed and learned how to describe populations based on available sample data. Central Limit Theorem helped us do these.

We then looked at Confidence Intervals to properly describe the conclusions about populations based on samples.

Then we studied Hypothesis Tests as another way of making inferences from sample data. Of course, there are errors in these tests too (Type I and Type II or False Positives and False Negatives, respectively).

We then studied various statistical tests to generate confidence intervals and test hypotheses.

We looked at how to analyze results and find differences between what we expect and what we get, through χ^2 Distributions (goodness-of-fit). χ^2 Distributions were also useful in studying variable independence and in testing hypotheses about a population variance.

We studied ANOVA, t-tests and F test as a means of understanding significant differences between means and variances.

CONGRATULATIONS! You are now prepared to make sense of the Statistical Methods for Decision Modeling, i.e., understand the outputs of statistical models you will learn next and take appropriate decisions.

Some good resources

- http://onlinestatbook.com
- http://stattrek.com
- http://www.khanacademy.org
- http://www.statsoft.com/Textbook
- http://vassarstats.net/textbook

Please Provide Module Feedbacks

Student ID	Name	Marks needed in Project for Certification	Marks needed in Project for Career Services	Overall Grade						
				Total	PHD	CUTe	MiTH	ROTe	# of Feedback	Attendance
[redacted]		Verify Participation	Not Eligible	43.00	46.80	39.85	67.20	41.09	1.00	0.00
		PGP	Eligible	80.00	80.80	70.73	88.30	70.59	9.00	5.00
		PGP	Eligible	66.00	68.50	55.73	64.20	67.07	5.00	5.00
		PGP	Eligible	66.00	64.00	56.47	51.50	71.71	9.00	5.00
		PGP	Not Eligible	57.00	48.40	57.60	66.30	66.23	2.00	5.00
		Cert. of Participation	Not Eligible	44.00	35.50	42.33	42.10	58.57	2.00	5.00
		PGP	Eligible	70.00	71.10	51.25	77.30	70.12	8.00	5.00

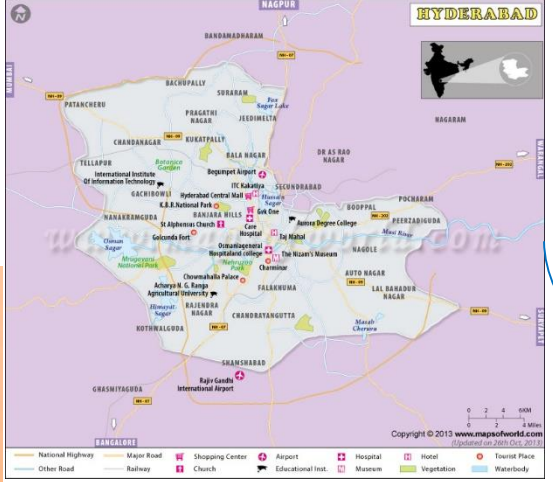
Batch 42 - CSE 7212c (Essential Engineering Skills in Big Data Analytics Using R and Python)

Modified on: Apr 09, 2018 | Created on: Apr 09, 2018

33

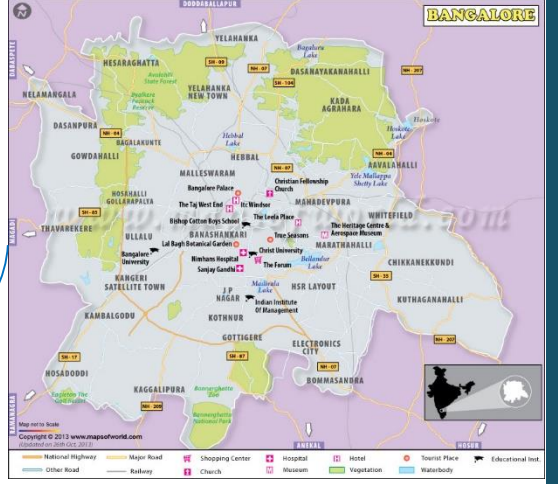
Responses

CSE 7315c



INSOFE

Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Floors 1-3, L77, 15th Cross Road, 3rd Main Road, Sector 6, HSR Layout, Bengaluru – 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.