



Inspire...Educate...Transform.

Foundations of Probability and Statistics for Data Science

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

Dr. Sridhar Pappu

Executive VP – Academics, INSOFE

December 24, 2017

TWO-SAMPLE t -TEST FOR MEANS

CSE 7315C



- 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, $\bar{x}_1 - \bar{x}_2$, is normally distributed for large sample sizes (both n_1 and $n_2 \geq 30$) whatever the population distribution.

Also, $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ [Recall $E(X-Y)=E(X)-E(Y)$]

and $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ [Recall $Var(X-Y)=Var(X)+Var(Y)$]

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of the difference}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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, $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

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

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ with } (n_1 + n_2 - 2) \text{ degrees of freedom.}$$

Note: Variance is average variation (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

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2; \text{ Test statistic, } t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

The degrees of freedom in this case are calculated as:

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

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.

DANGEROUS ANTIBIOTICS

■ Antibiotics may make the body more prone to infection, a new study has warned.

■ Exactly how the resident "good" microbes in the gut protect against pathogens and how antibiotic treatments foster growth of disease-causing microbes have been poorly understood

■ Now, researchers studied a mouse model and identified the chain of events that occur within the gut lumen after antibiotic treatment that allow "bad" bugs to flourish.

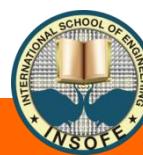
■ The finding has profound implications, expanding the current view of how microbes interact with each other at the gut surface

■ The process begins with antibiotics depleting "good" bacteria in the gut, including those that break-down fibre from vegetables to create butyrate

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.

CSE 7315C



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).

CSE 7315C



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

$$n_2 = 15$$

$$\bar{x}_2 = 0.633$$

$$s_2 = 0.216$$

$$s_2^2 = 0.0467$$

$$n_1 = 15$$

$$\bar{x}_1 = 0.931$$

$$s_1 = 0.202$$

$$s_1^2 = 0.0408$$

CSE 7315C



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 $\alpha = 0.05$, determine if there is a significant difference between the two groups.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}; t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ with } (n_1 + n_2 - 2) \text{ df.}$$

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

$$t = \frac{0.931 - 0.633}{0.209 * \sqrt{\frac{1}{15} + \frac{1}{15}}} = 3.91$$

You can find the p-value for this t-score or knowing that the t-score is way more than the critical value for 28 df (~ 2) at this significance level, you see that it is in the critical region in the right tail.

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ Rewriting:}$$

$$(\bar{x}_1 - \bar{x}_2) - ts_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + ts_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

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

95% CI: (0.142, 0.454)

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

HYPOTHESIS TEST

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1421489 0.4551844

CONFIDENCE INTERVAL

sample estimates:

mean of x mean of y

0.9313333 0.6326667



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

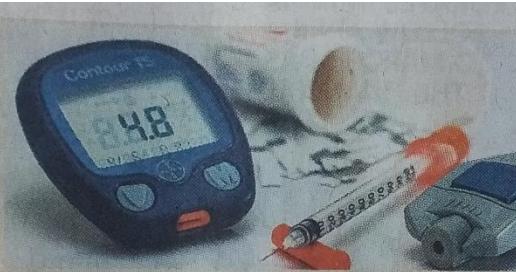
60 children were studied

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.

CSE 7315C



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 12 diverticulosis patients with comparable severity in 20-44 age group is given two treatments. In Treatment A, bran is given in the same dosage in three meals during the day and in Treatment B, it is given in one meal.

The alimentary transit times of marker pellets through the alimentary canal are recorded.

The clinician wants to test at 5% Significance Level (or 95% Confidence Level) if the transit times are shorter when bran is given in 3 meals.

CSE
7315C



Two-Sample t-Test for Paired Data

Treatment times (h)			
Patient	Treatment A	Treatment B	A-B
1	63	55	8
2	54	62	-8
3	79	108	-29
4	68	77	-9
5	87	83	4
6	84	78	6
7	92	79	13
8	57	94	-37
9	66	69	-3
10	53	66	-13
11	76	72	4
12	63	77	-14
TOTAL	842	920	-78
MEAN	70.17	76.67	-6.5

CSE 7315C



Two-Sample t-Test for Paired Data

Mean of the differences, $\bar{d} = -6.5$

Standard Deviation of the differences, $s_d = 15.1$

Standard Error of the mean, $SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = 4.37$

$$t = \frac{\bar{d}}{SE(\bar{d})} = \frac{-6.5}{4.37} = -1.487$$

$$t_{11,0.025} = 2.20099$$

Comparing the absolute t-value, we cannot reject the null hypothesis that the mean transit time between the two forms of treatment is the same.

Numbers in each row of the table are values on a t-distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264833	0.717558	1.439756	1.943181	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782286	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770333	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753056	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10992	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7068
27	0.255858	0.683695	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	—	—	80%	90%	95%	98%	99%	99.9%

Two-Sample t-Test for Paired Data

The 95% CI for the mean difference is given by $\bar{d} \pm t_{n-1, \frac{\alpha}{2}} * SE(\bar{d})$

$$-6.5 - 2.201 * 4.37 \leq D \leq -6.5 + 2.201 * 4.37$$

95% CI: (-16.1, 3.1).

As zero is included in the CI, we cannot reject the null hypothesis.

Business Decision

Although zero is included in CI, the range is very wide, which should lead the business to conduct a larger study to be sure.

CSE 7315C



CI and HT – R Output

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

Paired t-test

```
data: treatmentA and treatmentB  
t = -1.4868, df = 11, p-value = 0.1652  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-16.122519 3.122519  
sample estimates:  
mean of the differences  
-6.5
```

χ^2 DISTRIBUTION

CSE 7315C



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 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

$$X^2 = 38.272$$

Is this high?

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

χ^2 distribution

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

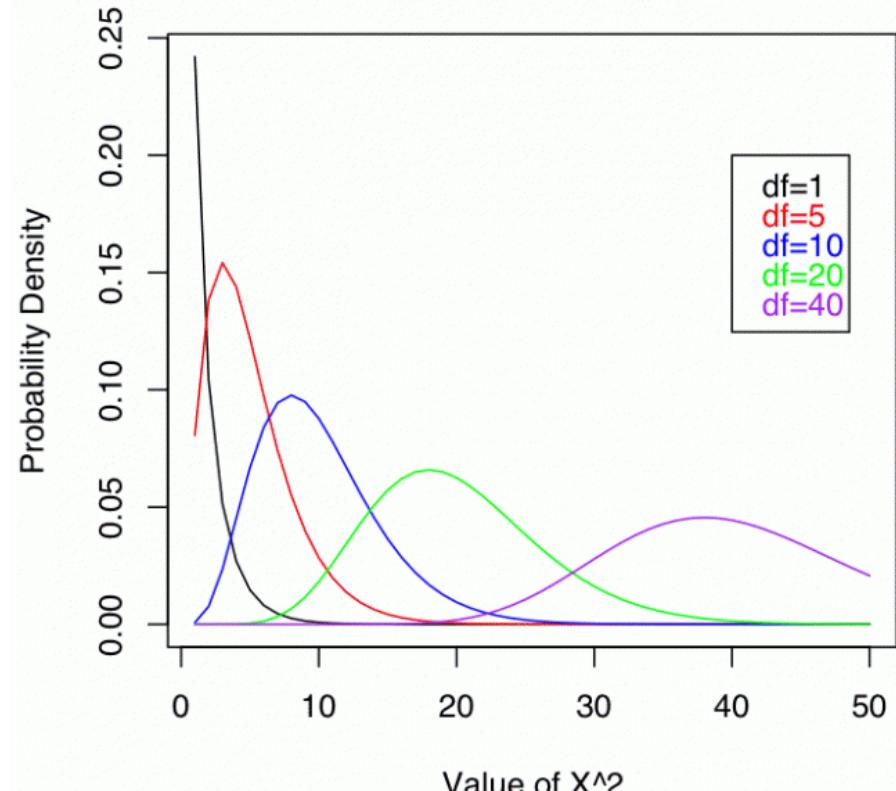
$$Z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

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

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

The shape depends on number of squared deviates added together.

Examples of chi-squared distributions



CE 73156

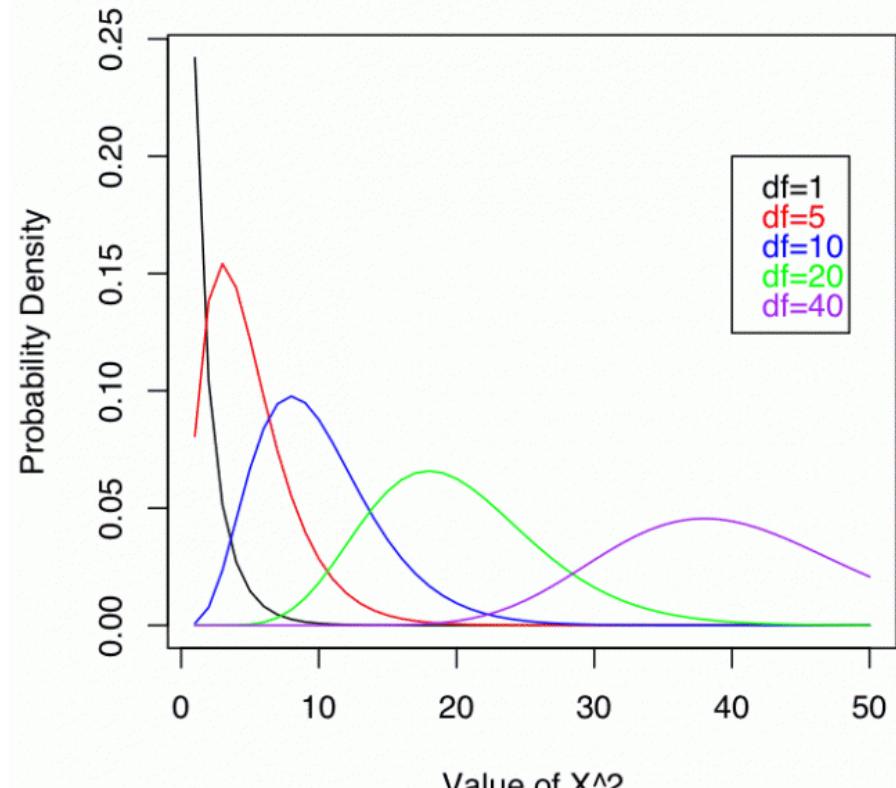


χ^2 distribution

$X^2 \sim \chi_{(\nu)}^2$, 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 .

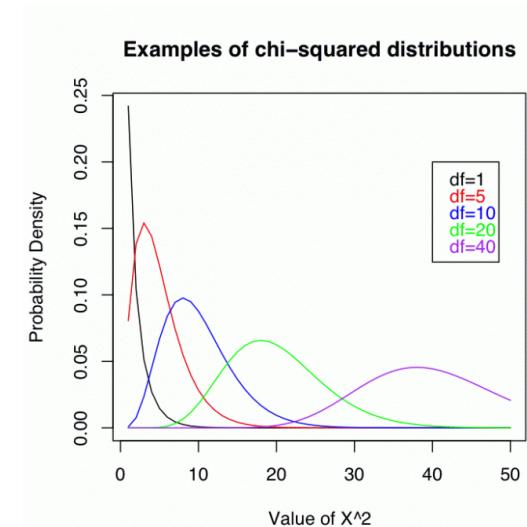
Examples of chi-squared distributions



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 5th. Therefore, there are $5-1=4$ degrees of freedom.

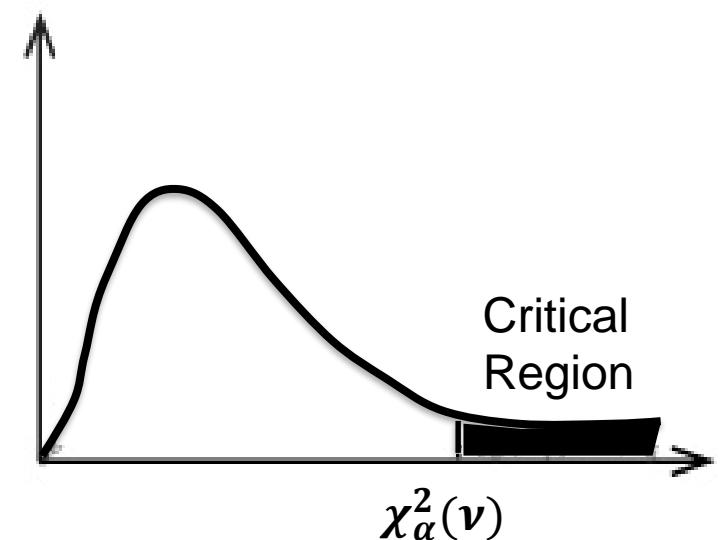
$\nu = (\text{number of classes}) - (\text{number of restrictions})$, or

$\nu = (\text{number of classes}) - 1 - (\text{number of parameters being estimated from sample data})$

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 $\chi^2_{\alpha}(v)$.
The critical region is to its right.



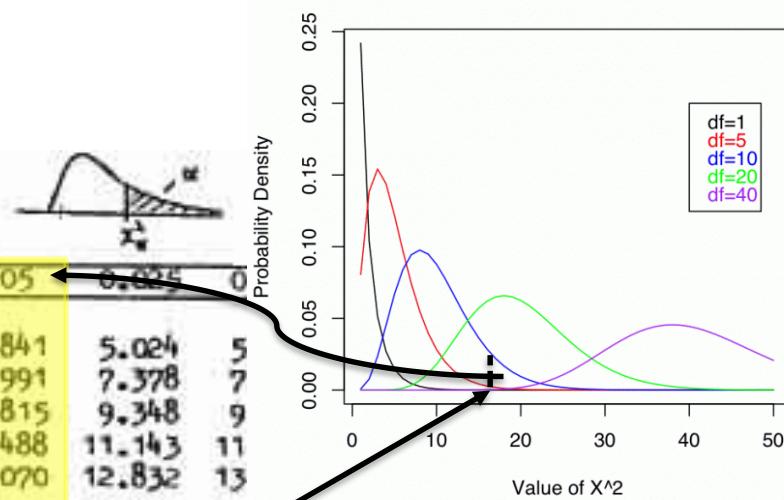
Why?

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

Using χ^2 probability tables

TABLE OF CHI-SQUARE DISTRIBUTION

α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0	Probability Density		
1	0.0393	0.03157	0.03628	0.03982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5			
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7			
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9			
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11			
5	0.412	0.594	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13			
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.022	16.014	16.790	22.427
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.555	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.099	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	9.260	10.196	11.293	11.688	13.091	14.848	17.187	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620



Uses of χ^2 distribution

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

CSE 7315C



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.



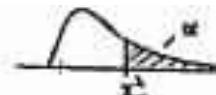
What are the expected frequencies and degrees of freedom?

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

$$v = 4$$

What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0393	0.03157	0.03628	0.03982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0217	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.279	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.024	3.571	4.178	4.466	5.226	6.244	7.807	15.812	18.510	21.776	23.770	24.461	26.140	28.140	33.264

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

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

Expected frequency = Probability * Total

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

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

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

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 $\lambda = 2.39$

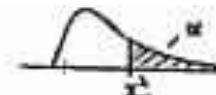
# of calls per 5-min interval	Expected Probability	Expected Frequency
0	0.0916	13.74
1	0.2190	32.85
2	0.2617	39.25
3	0.2085	31.27
4	0.1246	18.69
5	0.0595	8.93
6 or more	0.0351	5.26
TOTAL		150.00

$$v = 7 - 2 = 5$$

How do you find this value?

Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0393	0.03157	0.03628	0.03982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0217	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.279	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.024	3.571	4.178	4.466	5.226	6.244	7.807	15.812	18.510	21.226	23.220	24.001	26.140	28.120	33.264

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

Step 4: Calculate the test statistic χ^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

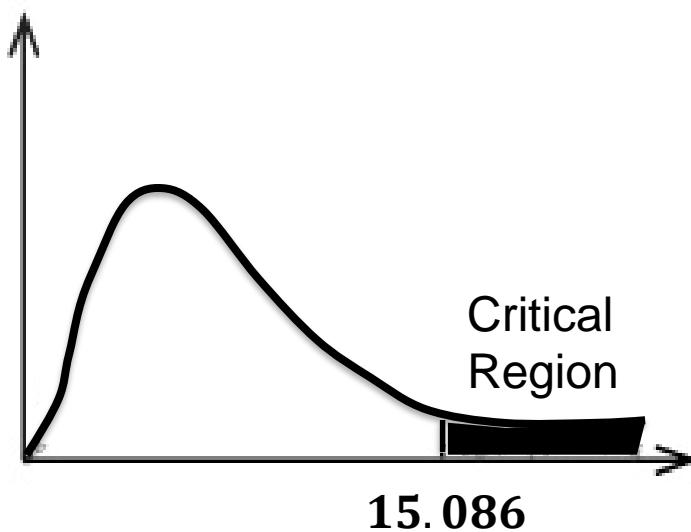
$$\chi^2 = 10.47$$

CSE 7315C



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

$\chi^2 = 10.47$, which is less than the critical value of 15.086. It 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.

CSE 7315C



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.



Possible Outcomes	Croupier A	Croupier B	Croupier C	Observed Results
Win	43	49	22	
Draw	8	2	5	
Lose	47	44	30	

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	2	5	15
Lose	47	44	30	121
Total	98	95	57	250

$$P(\text{Win}) = \frac{\text{Total Wins}}{\text{Grand Total}} = \frac{114}{250}$$
$$P(A) = \frac{\text{Total } A}{\text{Grand Total}} = \frac{98}{250}$$

If croupier and the outcome are independent,

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

χ^2 independence test

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	2	5	15
Lose	47	44	30	121
Total	98	95	57	250

Expected Frequency of Win and A

$$\begin{aligned} &= \text{Grand Total} \times \frac{\text{Total Wins}}{\text{Grand Total}} \times \frac{\text{Total A}}{\text{Grand Total}} = \frac{\text{Total Wins} \times \text{Total A}}{\text{Grand Total}} \\ &= \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}} \end{aligned}$$

χ^2 independence test – Finding expected frequencies

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	2	5	15
Lose	47	44	30	121
Total	98	95	57	250

	Croupier A	Croupier B	Croupier C
Win	$(114*98)/250$	$(114*95)/250$	$(114*57)/250$
Draw	$(15*98)/250$	$(15*95)/250$	$(15*57)/250$
Lose	$(121*98)/250$	$(121*95)/250$	$(121*57)/250$

χ^2 independence test – Calculating χ^2

	Observed	Expected	$\frac{(O - E)^2}{E}$
A	43	44.688	0.0638
	8	5.88	0.7644
	47	47.432	0.0039
	49	43.32	0.7447
B	2	5.7	2.4018
	44	45.98	0.0853
C	22	25.992	0.6131
	5	3.42	0.7299
	30	27.588	0.2109
$\sum O = 250$		$\sum E = 250$	$\sum \frac{(O - E)^2}{E} = 5.618$

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.

TABLE OF CHI-SQUARE DISTRIBUTION



α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
v	0.0393	0.03157	0.03628	0.03982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
1	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
2	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
3	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
4	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
5	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
6	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
7	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
8	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
9	2.156	2.558	3.099	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
10	2.603	3.053	3.609	3.816	4.575	5.528	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
11	3.074	3.571	4.128	4.404	5.226	6.204	7.802	16.812	18.510	21.022	23.220	24.041	26.242	28.244	33.264
12	3.564	4.091	4.748	5.064	5.926	6.904	8.502	18.812	20.510	23.022	25.220	26.041	28.242	30.244	35.264

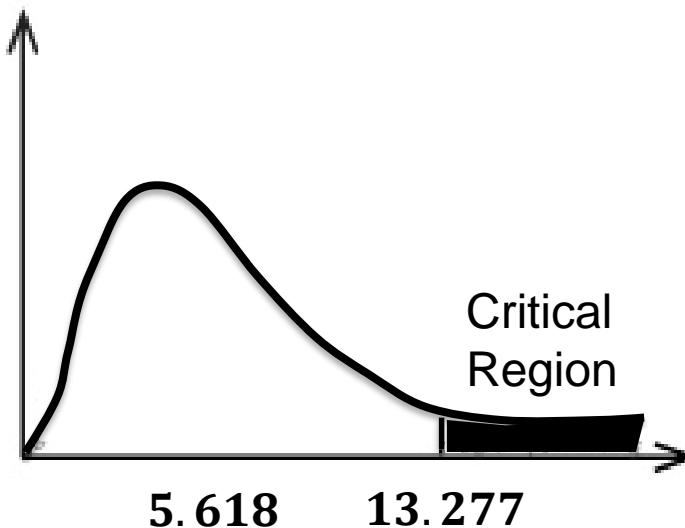
CSE 7315C

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



χ^2 independence test – Decision

Since calculated $\chi^2 = 5.618$, 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

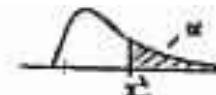
EXPECTED

	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

$$v = 4$$

Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0393	0.07157	0.07628	0.07982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0217	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.024	3.571	4.178	4.466	5.226	6.244	7.807	15.812	18.510	21.226	23.220	24.021	26.120	28.120	33.264

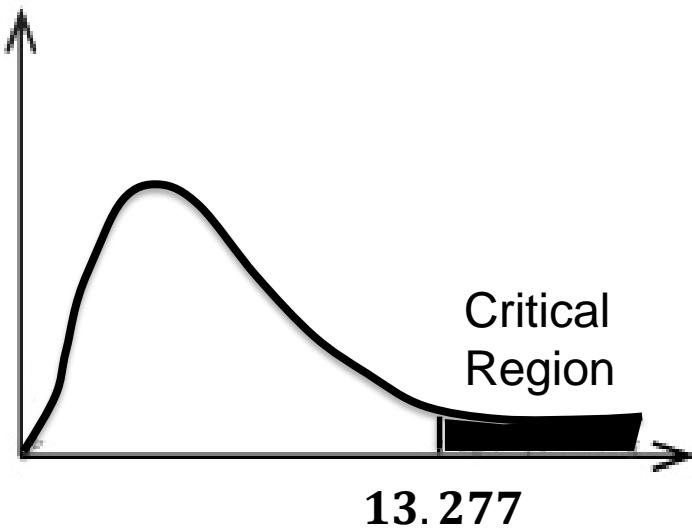
$\chi^2_{1\%}(4) = 13.277$. This means the critical region is $X^2 > 13.277$.

Step 4: Calculate the test statistic χ^2

	Observed	Expected	$\frac{(O - E)^2}{E}$
Pain relievers	26	40.94	
	41	29.74	
	24	20.32	
	95	66.58	
Tranquilizers/ Sedatives	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

$\chi^2 = 59.41$, 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.

CSE 7315C



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?

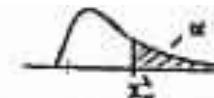
Since n=20, df=19.



Testing Hypotheses about a Variance

What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0393	0.03157	0.03628	0.03982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.594	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.099	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315

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

CSE 7315C



Testing Hypotheses about a Variance

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

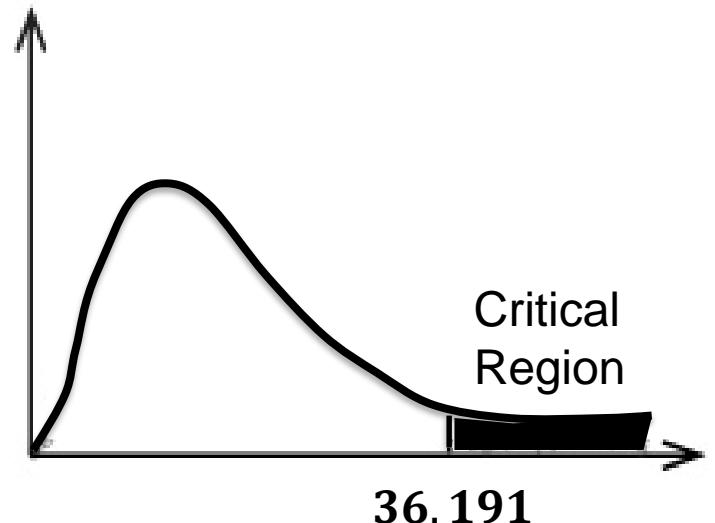
No.

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?

CSE7315G

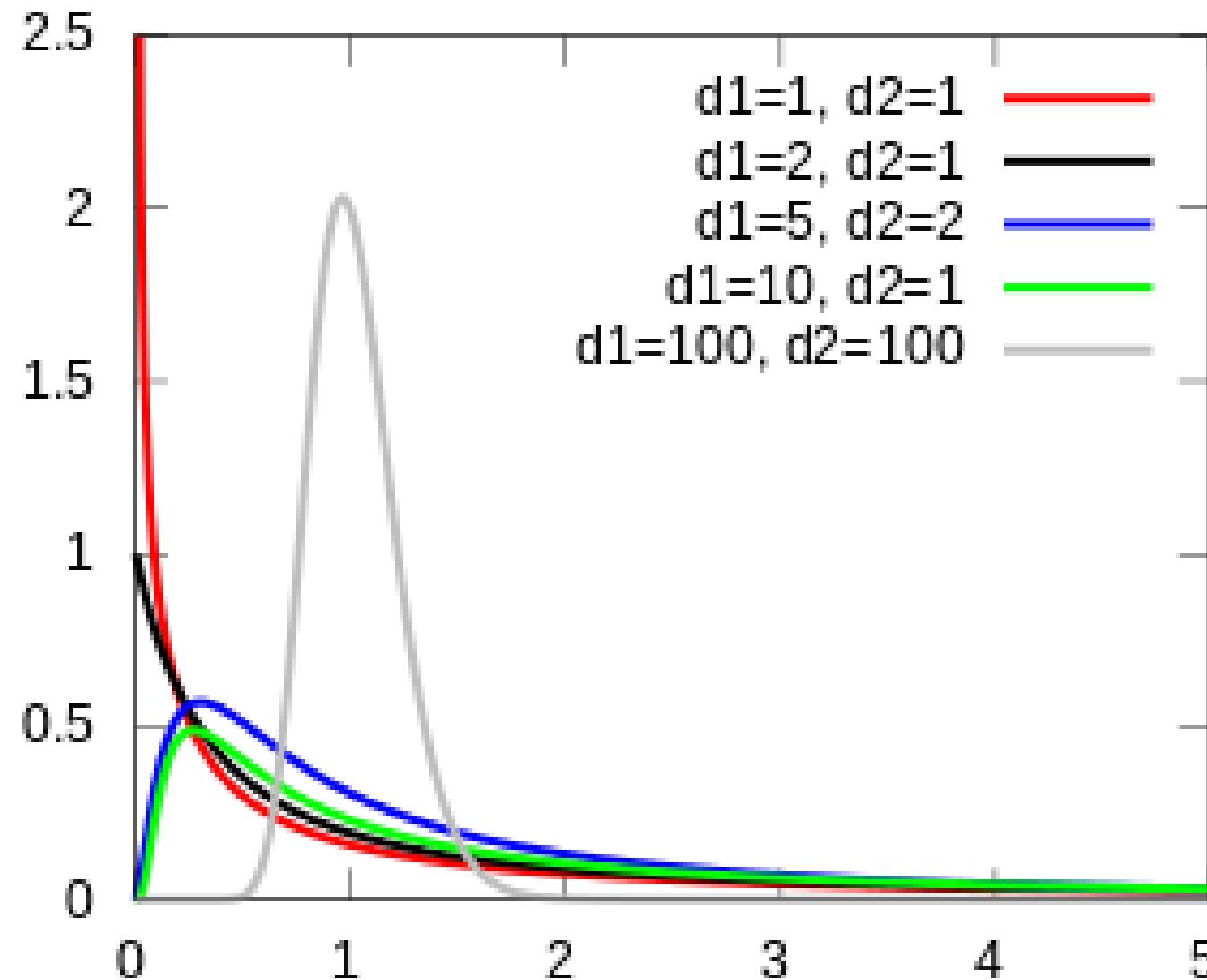


F distribution

- Ratio of 2 variance estimates: $F = \frac{s_1^2}{s_2^2} = \frac{\text{est.}\sigma_1^2}{\text{est.}\sigma_2^2}$
- 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.,

$$F = \frac{\frac{\chi_{\nu_1}^2}{\nu_1}}{\frac{\chi_{\nu_2}^2}{\nu_2}}$$

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?

CSE 7315C



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
$s_1^2 = 0.11378$	$n = 10$	$s_2^2 = 0.02023$	$n = 12$

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

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

Reading an F-table.



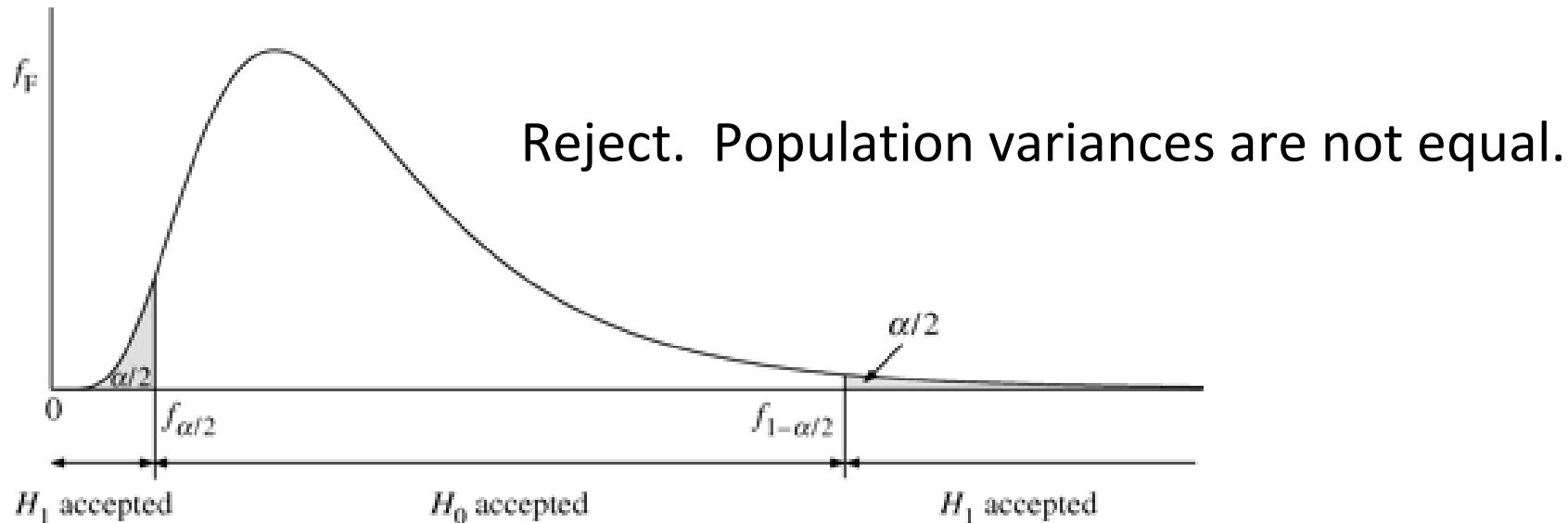
/	df ₁ =1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df ₂ =1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274	976.7079	984.8668	993.1028	997.2492	1001.414	1005.598	1009.800	1014.020	1018.258
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980	39.4146	39.4313	39.4479	39.4562	39.465	39.473	39.481	39.490	39.498
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189	14.3366	14.2527	14.1674	14.1241	14.081	14.037	13.992	13.947	13.902
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439	8.7512	8.6565	8.5599	8.5109	8.461	8.411	8.360	8.309	8.257
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192	6.5245	6.4277	6.3286	6.2780	6.227	6.175	6.123	6.069	6.015
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613	5.3662	5.2687	5.1684	5.1172	5.065	5.012	4.959	4.904	4.849
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.6658	4.5678	4.4667	4.4150	4.362	4.309	4.254	4.199	4.142
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951	4.1997	4.1012	3.9995	3.9472	3.894	3.840	3.784	3.728	3.670
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260	3.9639	3.8682	3.7694	3.6669	3.6142	3.560	3.505	3.449	3.392	3.333
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.6209	3.5217	3.4185	3.3654	3.311	3.255	3.198	3.140	3.080
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879	3.5257	3.4296	3.3299	3.2261	3.1725	3.118	3.061	3.004	2.944	2.883
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736	3.2773	3.1772	3.0728	3.0187	2.963	2.906	2.848	2.787	2.725

$$F_{0.025, 9, 11} = 0.2556; F_{0.975, 9, 11} = 3.5879 \quad R\text{-code: } qf(\alpha, df_1, df_2)$$

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?



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.

CSE 7315C



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)

CSE 7315C



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., $t_{(\nu)}^2 = F_{(1,\nu)}$

CSE 7315C

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

$$\begin{aligned}
 &= (2 - 4.96)^2 + 5 * (3 - 4.96)^2 + 4 * (4 - 4.96)^2 + 7 * (5 - 4.96)^2 + 5 * (6 - 4.96)^2 \\
 &+ 4 * (7 - 4.96)^2 + (8 - 4.96)^2 = \mathbf{62.96}
 \end{aligned}$$

When there are m groups and n members in each group, the degrees of freedom are $mn - 1$, 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

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

When there are m groups and n members in each group, the degrees of freedom are $m(n - 1)$, 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 $m - 1$.

SST = SSW + SSB

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

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?

The drug formulations have different impacts 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.

F Table for $\alpha = 0.10$

N	$df_1=1$	2	3	4	5	6	7	8	9	10	12
$df_2=1$	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50196
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632	2.37888
10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731	2.32260	2.28405
11	3.22520	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.30400	2.27350	2.24823	2.20873
12	3.17655	2.80680	2.60552	2.48010	2.39402	2.33102	2.28278	2.24457	2.21352	2.18776	2.14744
13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.23410	2.19535	2.16382	2.13763	2.09659
14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.15390	2.12195	2.09540	2.05371
15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621	2.05932	2.01707
16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.12800	2.08798	2.05533	2.02815	1.98539
17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839	2.00094	1.95772
18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467	1.97698	1.93334
19	2.98990	2.60561	2.39702	2.26630	2.17596	2.10936	2.05802	2.01710	1.98364	1.95573	1.91170
20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.03970	1.99853	1.96485	1.93674	1.89236
21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797	1.91967	1.87497
22	2.94858	2.56131	2.35117	2.21927	2.12794	2.06050	2.00840	1.96680	1.93273	1.90425	1.85925
23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888	1.89025	1.84497
24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625	1.87748	1.83194
25	2.91774	2.52831	2.31702	2.18424	2.09216	2.02406	1.97138	1.92925	1.89469	1.86578	1.82000

The df are 2 for numerator and 24 for denominator.

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 the drug formulations do have different impacts on migraine pain relief.

Anova:Single Factor

SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Group1	9	32	3.555556	1.027778
Group2	9	50	5.555556	2.777778
Group3	9	52	5.777778	0.694444

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	26.963	2	13.4815	8.987654	0.0012	2.5383
Within Groups	36	24	1.5			
Total	62.963	26				

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.

Anova: Single Factor

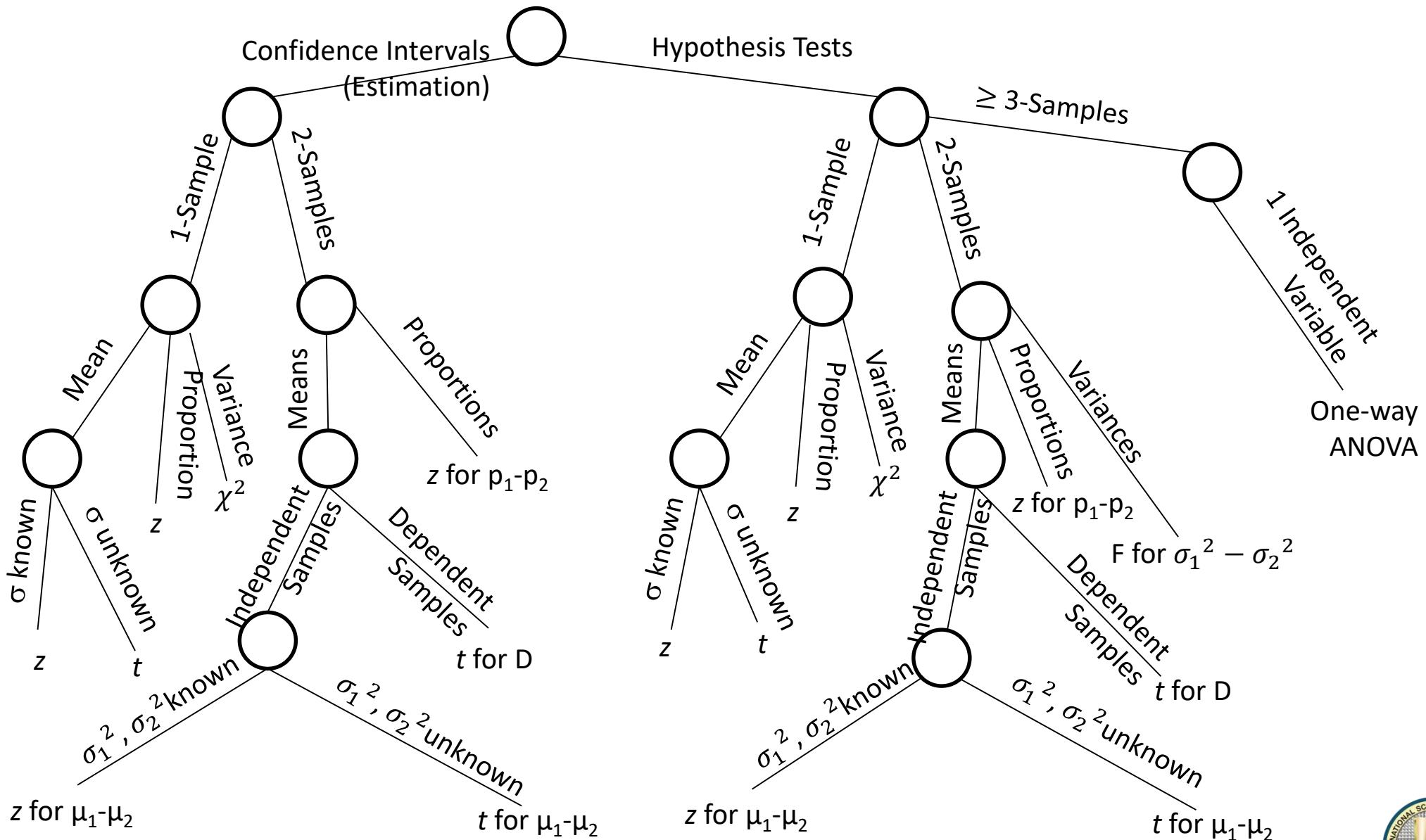
SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Financial	8	92.68	11.585	26.2528
Energy	8	110.77	13.8463	23.6879
Utilities	8	71.3	8.9125	20.5247

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	97.593	2	48.7965	2.07747	0.1502	3.4668
Within Groups	493.26	21	23.4885			
Total	590.85	23				

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, 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

CSE 7315C



THE STORY

CSE 7315C



We started with a high level overview of various Data Science problems, applications and skills needed for a Data Scientist. We also saw how Statistics were all around us and how people misuse them.

We understood 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. 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.

We started with 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.

We then looked at variety of ways this data 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 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.

We then studied various statistical tests to test hypotheses.

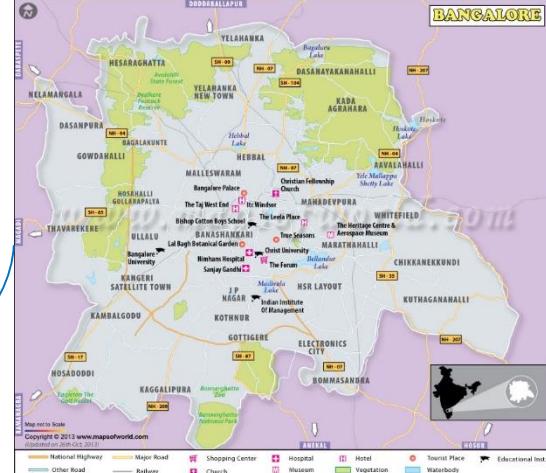


We looked at how to analyze results and find differences between what we expect and what we get, through χ^2 Distributions (goodness-of-fit). χ^2 Distributions were also useful in studying variable independence and in testing hypotheses about a population variance.

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

Some good resources

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



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.