Inspire…Educate…Transform.

Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Hypothesis Tests, Statistical Tests: t, Chi-Squared, F

Prof Anuradha Sharma

Mentor, INSOFE

January 5, 2020

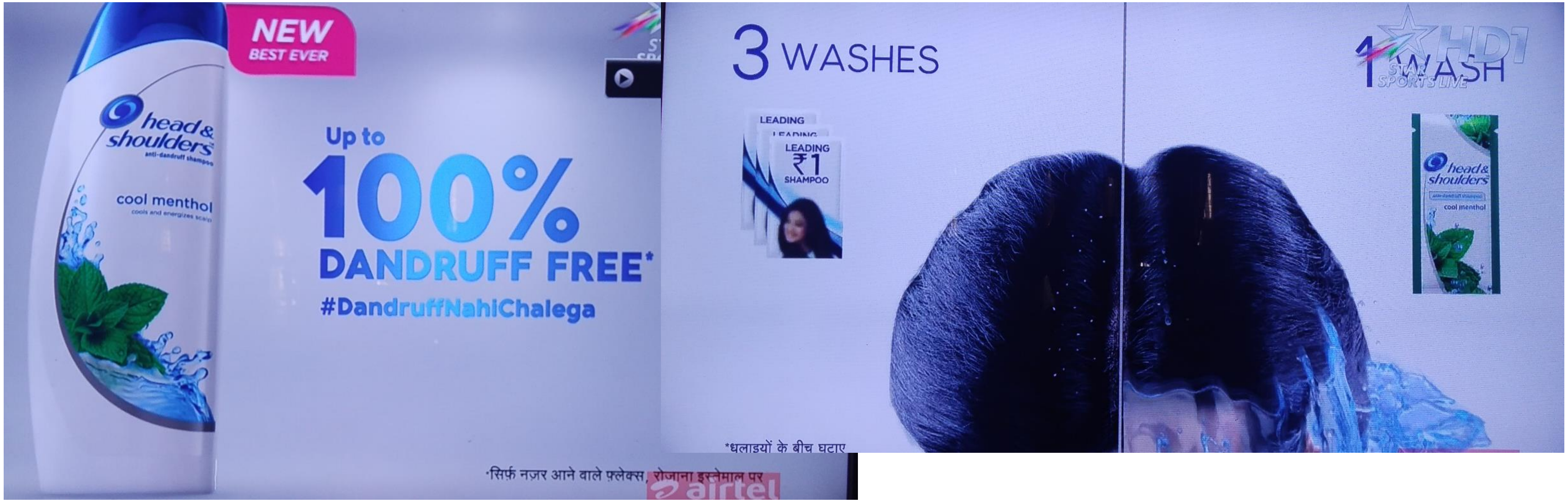
Content : Dr. Sridhar Pappu

INFERENTIAL STATISTICS

HYPOTHESIS TESTS

CSE 7215c

Hypothesis tests give a way of using samples to test whether statistical claims are likely to be true or not about a population.

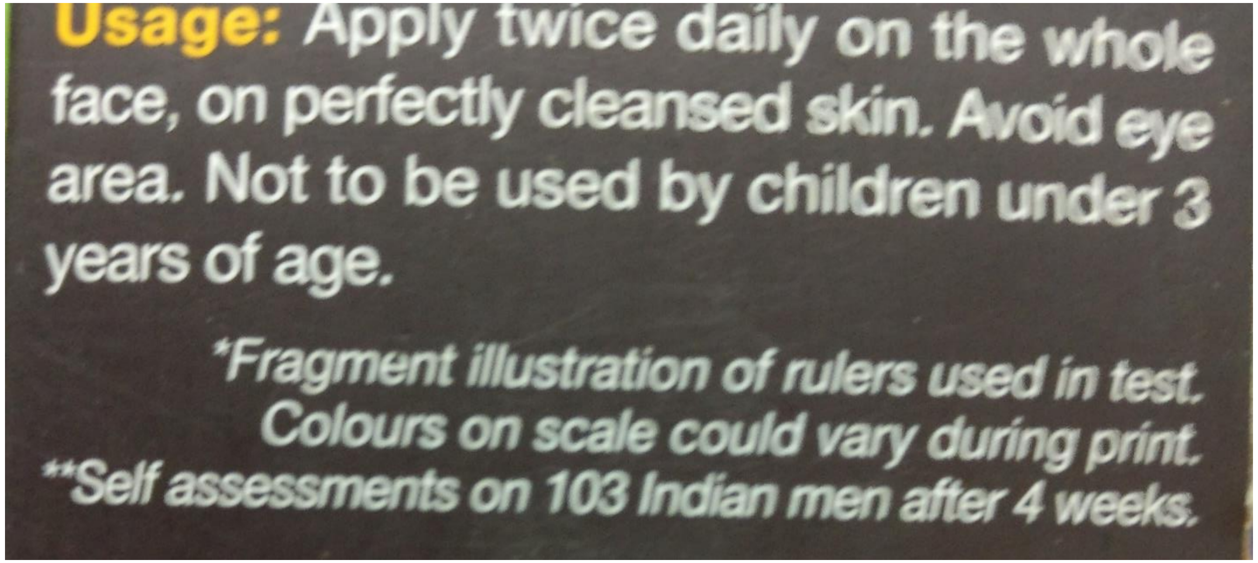


CSE 7215c

Hypothesis tests give a way of using samples to test whether statistical claims are likely to be true or not about a population.

CSE 7215c

Hypothesis tests give a way of using samples to test whether statistical claims are likely to be true or not about the population.



Not sure if claims are true?
Are they telling lies or is their sample biased?

- Examine claim
- Examine evidence
- Make decision! Hypothesis testing process

CSE 7215c

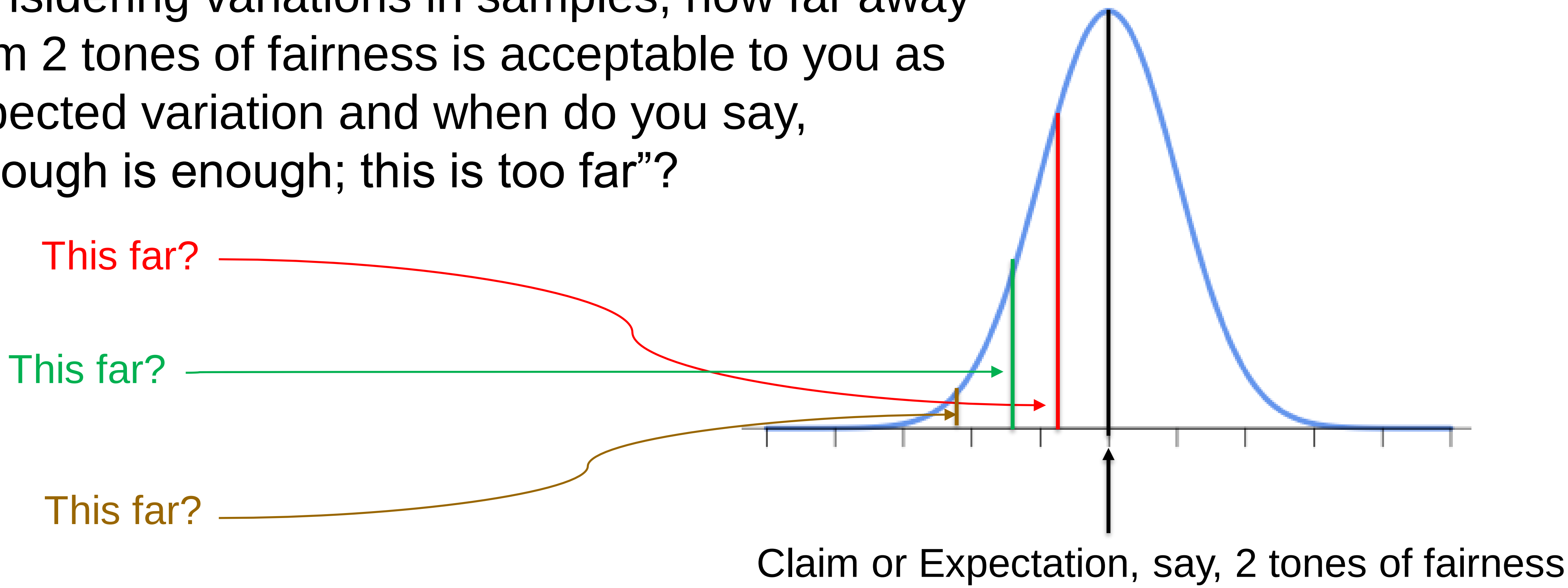
Examples of Hypotheses

- Two hypotheses in competition:
  - H_0: The NULL hypothesis, usually the most conservative.
  - H_1 or H_A: The ALTERNATIVE hypothesis, the one we are actually interested in.
- Examples of NULL Hypothesis:
  - The coin is fair
  - The new drug is no better (or worse) than the placebo
- Examples of ALTERNATIVE hypothesis:
  - The coin is biased (either towards heads or tails)
  - The coin is biased towards heads
  - The coin has a probability 0.6 of landing on tails
  - The drug is better than the placebo

CSE 7215c

Hypothesis Testing Process

Considering variations in samples, how far away from 2 tones of fairness is acceptable to you as expected variation and when do you say, “enough is enough; this is too far”?



We need something that allows us to measure how far an observation is from what I expect to see if the null hypothesis is true

CSE 7215c

Step 1: Decide on the hypothesis and choose the test statistic(z, t, F or others)

Garnier Men PowerWhite improves fairness by 2 tones within 4 weeks. This is called Null Hypothesis and is represented by H_0.

In this case, H_0: Tone = 2

If Null Hypothesis is rejected based on evidence, an Alternate Hypothesis, H_1, needs to be accepted. We always start with the assumption that Null Hypothesis is true.

In this case, H_1: Tone < 2

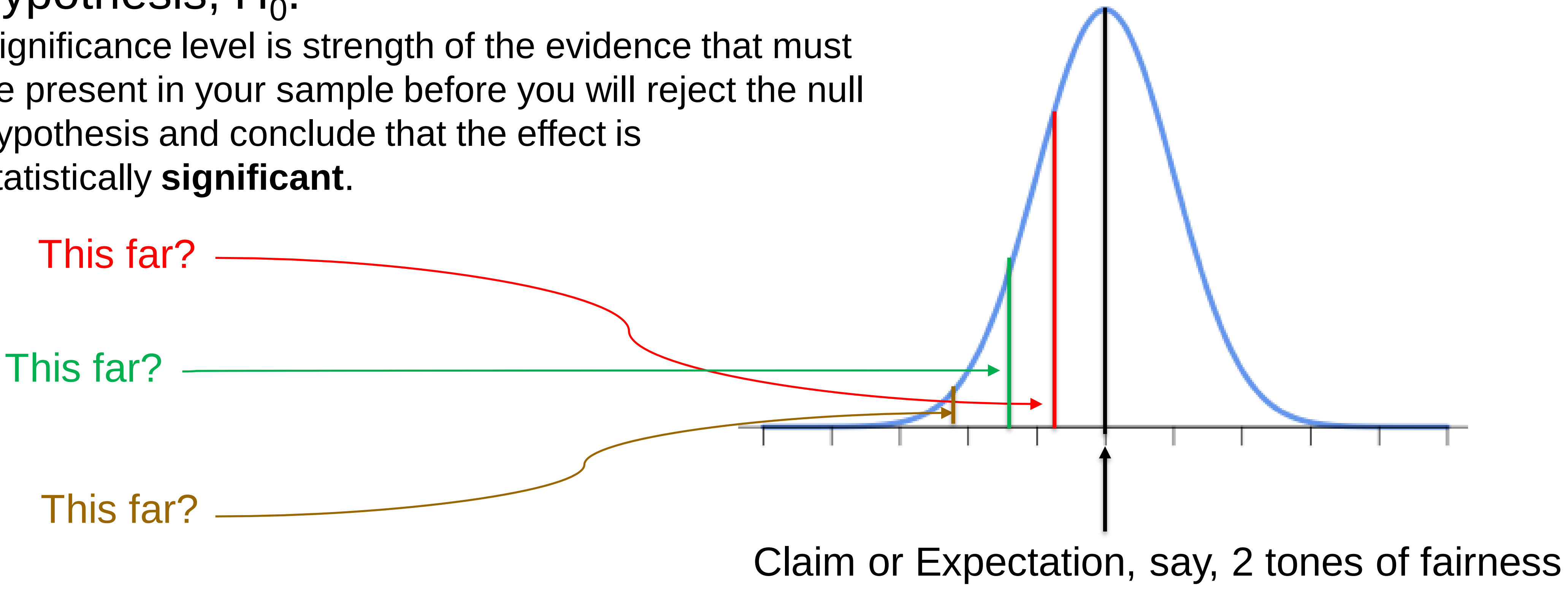
The test statistic allows us to measure how far an observation is from what I expect to see if the null hypothesis is true, its value is calculated if its distribution is known

CSE 7215c

Step 2: Specify the critical region(certain level of certainty or standard to be set)

Then we must decide on the **Significance Level, α**. It is a measure of how unlikely you want the results of the sample to be before you reject the null hypothesis, H_0.

Significance level is strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically **significant**.



CSE 7215c

Step 2: Specify the critical region

If X represents tones of fairness achieved, the critical region is defined as $P(X < c) < \alpha$ where $\alpha = 5\%$. Critical region is sample values improbable enough to consider rejecting null hypothesis

Recall that in a 95% CI, a 5% of the CI's of the samples will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 2 tones of fairness increased, is rejected.

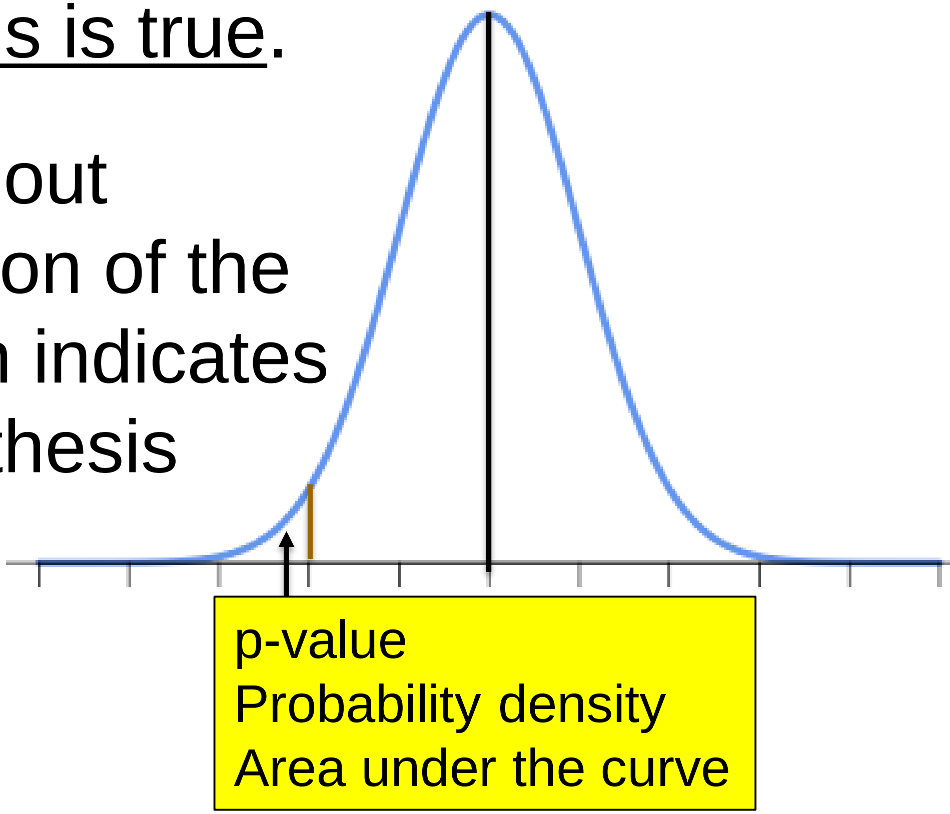
That is the reason 5% or 0.05 is called the Significance Level. In a 99% CI, 0.01 is the Significance Level.

CSE 7215c

Step 3: Find the *p*-value(how rare our results are assuming the null hypothesis is true)

p-value is the probability of getting ***only by chance*** a value at least as extreme as the one in the sample under the assumption that the null hypothesis is true.

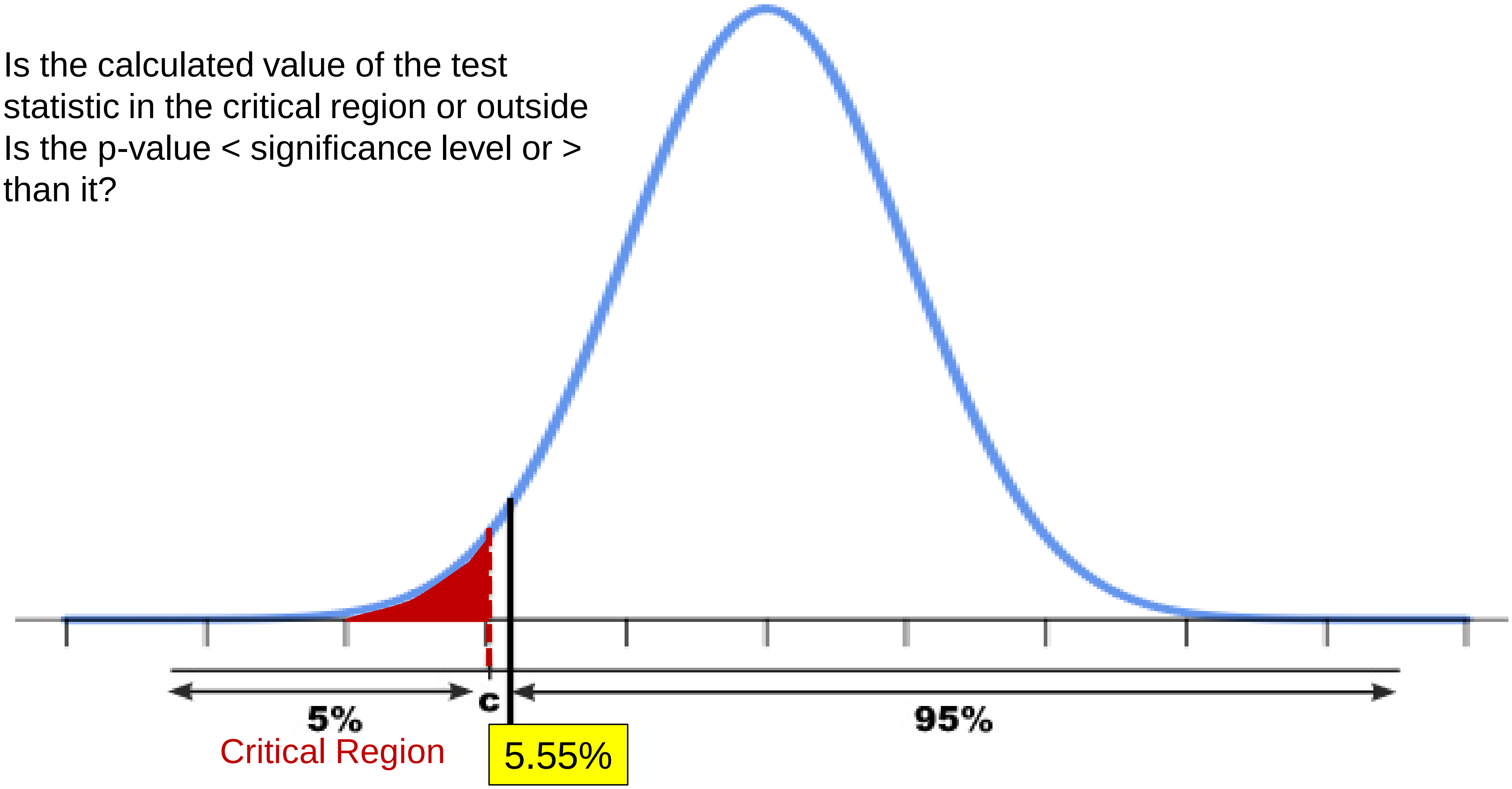
It is a way of taking the sample and working out whether the result falls within the critical region of the hypothesis test. A value in the critical region indicates presence of a real effect when the null hypothesis represents presence of no effect.



CSE 7215c

Step 4: Is the sample result in the critical region?

- Is the calculated value of the test statistic in the critical region or outside
- Is the p-value < significance level or > than it?



CSE 7215c

Step 5: Make your decision

There isn't sufficient evidence to reject the null hypothesis and so, the claims of the company are "accepted".

CSE 7215c

Attention Check

In hypothesis testing, do you assume the null hypothesis to be true or false?

True.

If there is sufficient evidence against the null hypothesis, do you “fail to reject” it or reject it?

Reject it

CSE 7215c

Attention Check

If the p-value is less than 0.05 for the above significance level, will you “fail to reject” or reject the null hypothesis?

Reject it.

Do you need weaker evidence or stronger to reject the null hypothesis if you were testing at the 1% significance level instead of the 5% significance level?

Stronger.

CSE 7215c

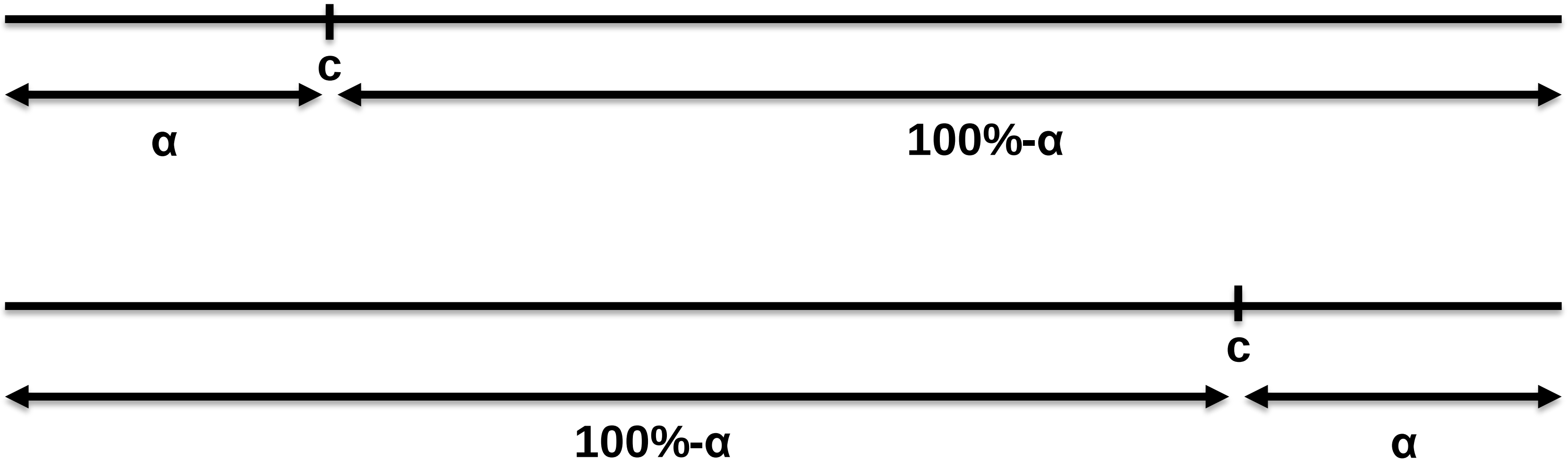
Critical Region Up Close

One-tailed tests

The position of the tail is dependent on H_1.

If H_1 includes a < sign, then the lower tail is used.

If H_1 includes a > sign, then the upper tail is used.



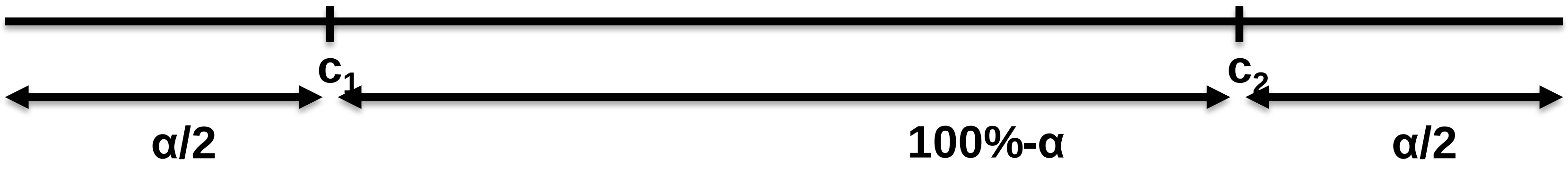
CSE 7215c

Critical Region Up Close

Two-tailed tests

Critical region is split over both ends. Both ends contain α/2, making a total of α.

If H_1 includes a ≠ sign, then the two-tailed test is used as we then look for a change in parameter, rather than an increase or a decrease.



CSE 7215c

Critical Region Up Close

For each of the scenarios below, identify what type of test you would require.

- Garnier Men PowerWhite hypothesis test as discussed till now.

  One-tailed/Lower-tailed

- If we were checking whether significantly more or significantly less than 2 tones fairer result was achieved, i.e., H_1: Fairness tone improvement $\neq 2$.

  Two-tailed test

- The coin is biased.

  Two-tailed test

- The coin is biased towards heads with probability 0.8.

  One-tailed/Upper-tailed

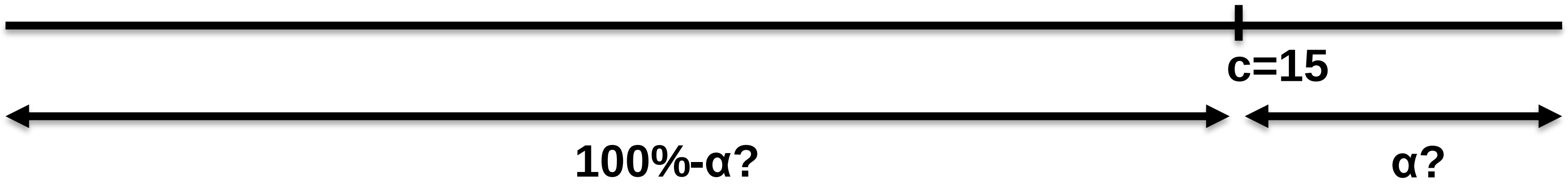
CSE 7215c

The Missing Link in the Google Interview

Q. What is the probability of getting 15 or more heads?

A. $$P(X \geq 15) = P(X = 15) + P(X = 16) + P(X = 17) + P(X = 18) + P(X = 19) + P(X = 20) = 0.021$$

What can you now say about the coin being biased or not?



At 5 % level of significance, we reject the NULL hypothesis (that the coin is fair), At 1%, we fail to reject the NULL hypothesis(that the coin is fair)

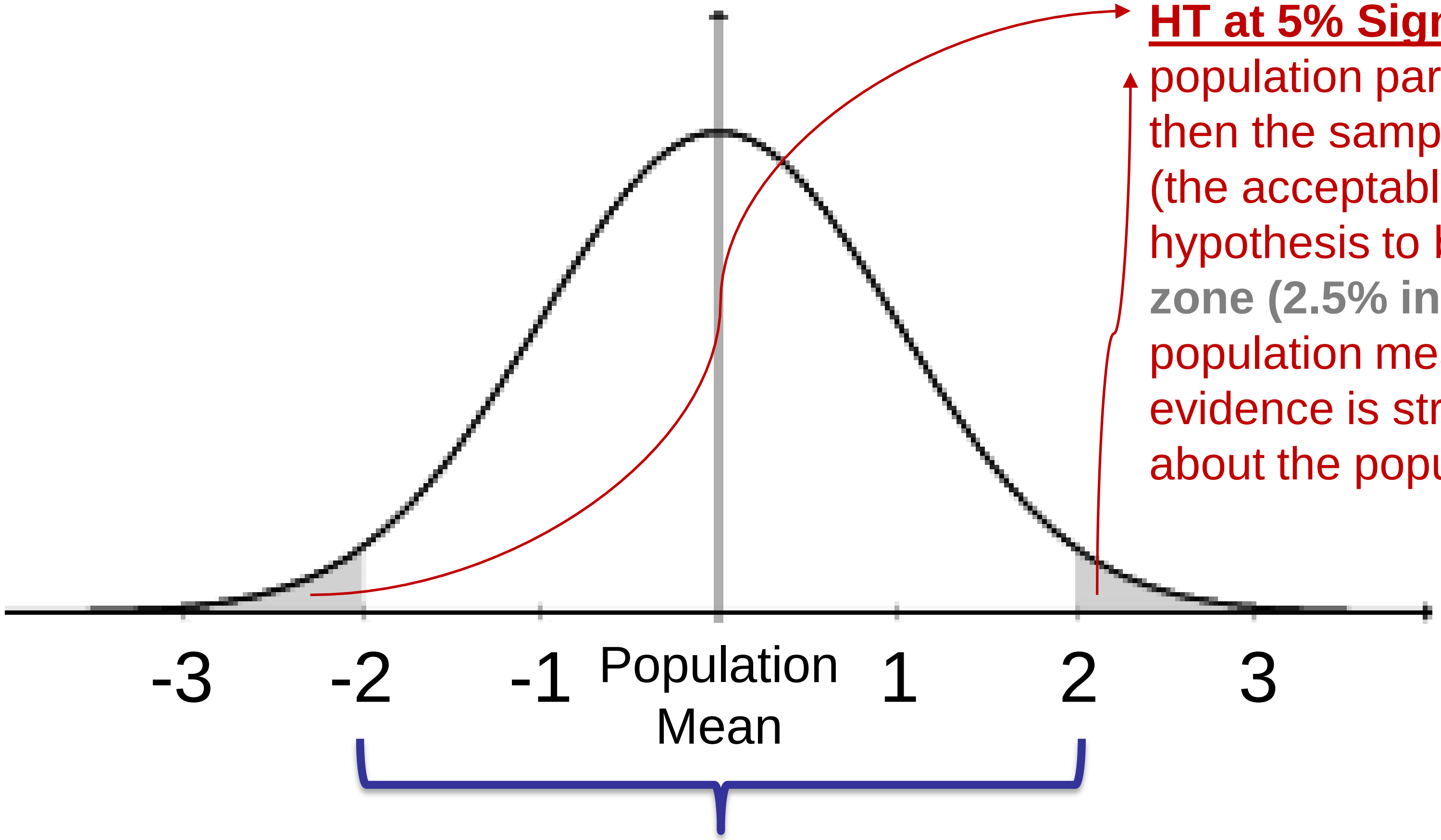
CSE 7215c

The hypothesis test doesn’t answer the question whether the company is telling the truth or not or if Garnier Men PowerWhite really works or not or if the coin is biased or not.

It only states whether the evidence is enough to reject the null hypothesis or not at the chosen significance level.

CSE 7215c

Confidence Intervals and Hypothesis Testing – Two Ways of Inferring the Same



HT at 5% Significance Level: If the true population parameter (e.g., mean) is as shown, then the sample must be within $\pm 2SE$ limits from it (the acceptable normal variation for the null hypothesis to be true). If the sample is in the **5% zone (2.5% in each tail shown in gray)**, then the population mean cannot be as shown (i.e., the evidence is strong to reject the null hypothesis about the population mean).

95% CI: If the true population parameter (e.g., mean) is as shown, then 95% of the samples will contain it within the range $\bar{x} \pm 2SE$. If the sample is in the **5% zone (2.5% in each tail shown in gray)**, then the true population parameter cannot be as shown (i.e., it will not lie in the range $\bar{x} \pm 2SE$.)

CSE 7215c

INFERENTIAL STATISTICS

COMMON TEST STATISTICS

CSE 7215c

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

z, t: Closely related to Sampling Distribution of **Means**

χ^2, F: Closely related to Sampling Distribution of **Variances**

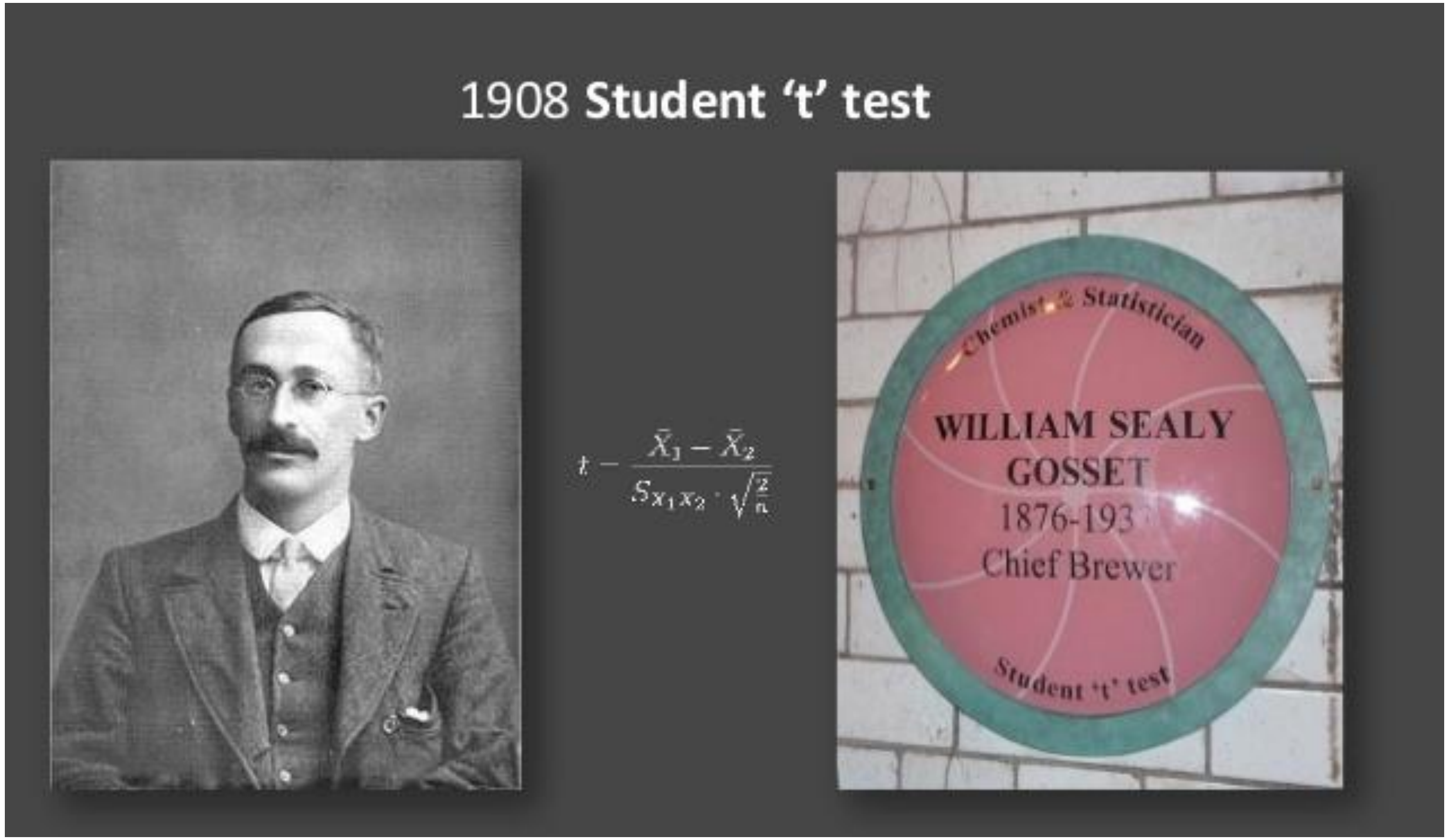
All are derived from Normal Distribution

CSE 7215c

ONE-SAMPLE *t*-TEST FOR MEANS

CSE 7215c

t-Distribution

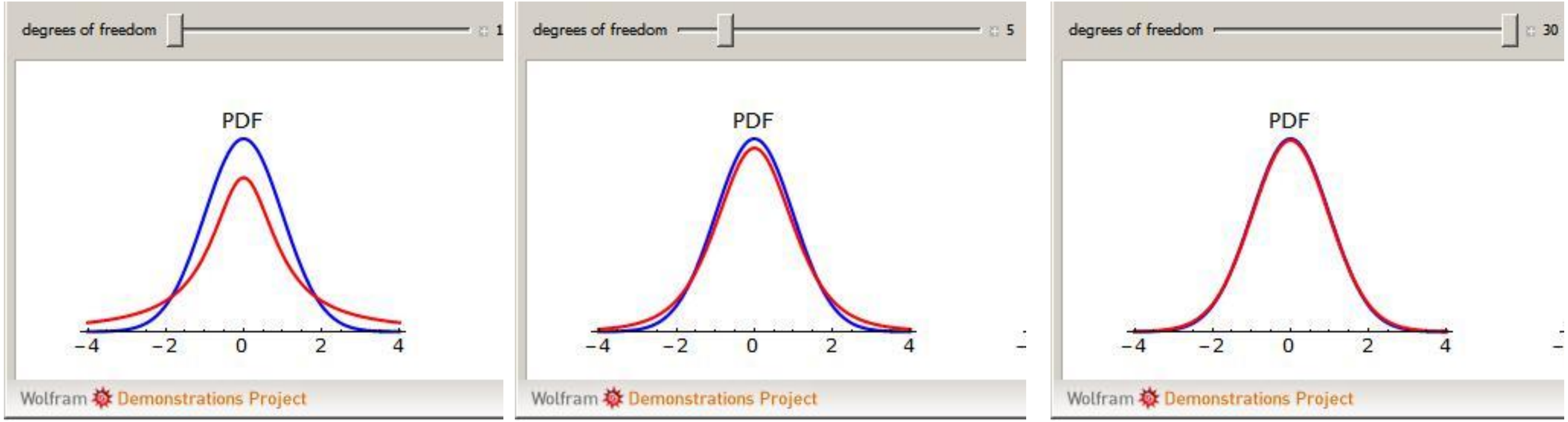


Ref: http://image.slidesharecdn.com/2013-ingenious-ireland-theingeniousirishiet-slideshow-130524065705-phpapp01/95/2013-ingeniousirelandthe-ingenious-irishietslideshow-43-638.jpg?cb=1369825611
Last accessed: October 31, 2015

CSE 7215c

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

CSE 7215c

t-Distribution

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

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

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

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

CSE 7215c

Properties of *t*-Distribution

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

CSE 7215c

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 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).

CSE 7215c

t-Test in Pharmaceutical Industry

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):

99.2	100.1	100.0	100.0	99.5
99.4	99.3	100.3	99.9	99.2

CSE 7215c

t-Test in Pharmaceutical Industry

What are null and alternate hypotheses?

$H_0 : \mu_1 = 100;\ H_1 : \mu_1 \neq 100$

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

Two-tailed.

What are the degrees of freedom?

$\nu = 10 - 1 = 9$

CSE 7215c

t-Test in Pharmaceutical Industry

Mean, $\bar{x}$ = 99.69 mg

Standard deviation, s = 0.41

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

CSE 7215c

t-Test in Pharmaceutical Industry

Mean, $\bar{x}$ = 99.69 mg, Standard deviation, s = 0.41

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

$$\bar{x} - t_{\left(\frac{\alpha}{2},\nu\right)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\left(\frac{\alpha}{2},\nu\right)} \frac{s}{\sqrt{n}}$$

$$99.69 - 2.262 * \frac{0.41}{\sqrt{10}} \leq \mu \leq 99.69 + 2.262 * \frac{0.41}{\sqrt{10}}$$

$99.40 \leq \mu \leq 99.98$

The batch mean is 99.69 mg with an error of +/-0.29 mg. The researcher is 95% confident that the average potency of the batch of tablets is between 99.40 mg and 99.98 mg.

CSE 7215c

t-Test in Pharmaceutical Industry

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

Note: conf.level = 1 - α

How do you get the sample t-value?

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{99.69 - 100}{\frac{0.41}{\sqrt{10}}} = -2.3784$$

```
        One Sample t-test

data:  dosage
t = -2.3784, df = 9, p-value = 0.04134
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 99.39515 99.98485
sample estimates:
mean of x
    99.69
```

HYPOTHESIS TESTING APPROACH

CONFIDENCE INTERVAL APPROACH

Can population mean be equal to 100mg?

R code: pt(-2.3784,9) = 0.02067

All software output double the calculated value for 2-tailed tests to enable easy check against the significance level (α) instead of with $\frac{\alpha}{2}$.

Reject H0 at 5% level of significance since p(.04134)<alpha(.05). You can compare the calculated value of t to the critical value of t and make that decision i.e. -2.38 lies in the left critical region beyond -2.26. Population mean is not contained in the confidence interval thus reject H0.

CSE 7215c

TWO-SAMPLE *t*-TEST FOR MEANS

CSE 7215c

- 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?
  - Is one version of web page leading to better product sales in an A/B testing for an eCommerce company?
  - Does the average cost of a two-bedroom flat differ between Delhi and Mumbai? 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?

CSE 7215c

Two-Sample t-Test

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

In 2-sample t-test, **difference in means** of the two samples is studied.

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

The degrees of freedom are calculated as:

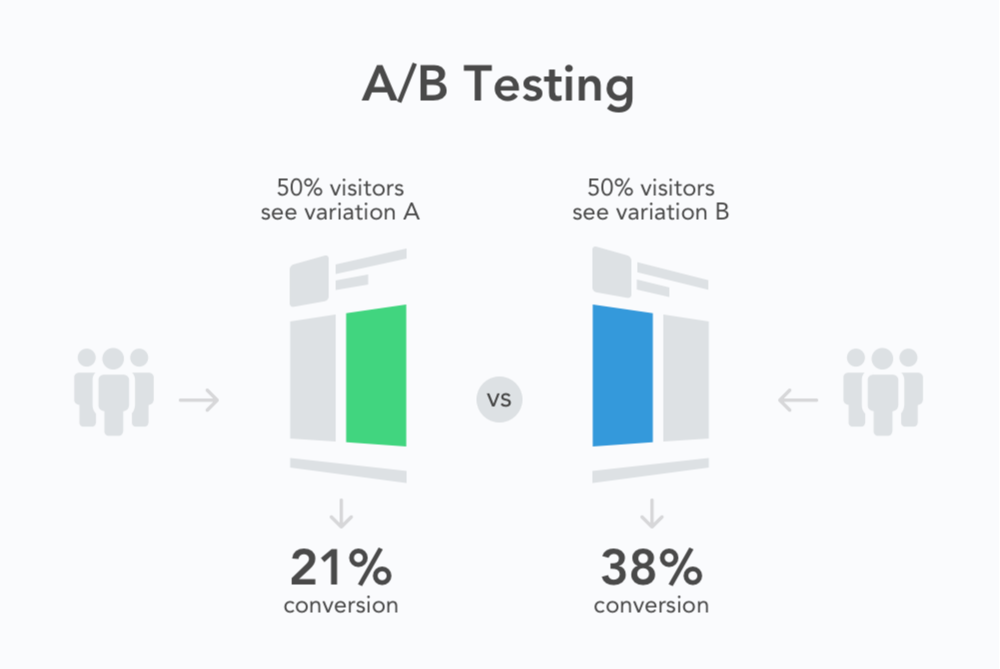
$\nu = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\left[\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}\right]}$, rounded off to the nearest integer.

R code: t.test(data1, data2, conf.level=0.95)

CSE 7215c

t-Test: A/B Testing in eCommerce Industry

Multiple factors affect our decision to click on an action button like buying a product in an eCommerce platform. These factors could range from colours, placement of product images, labels, etc. A/B testing is a standard approach to testing if one version of the webpage leads to better conversions than another.



200 people (100 male and 100 female) were randomly shown a different version (A or B) of the webpage and their purchases tracked. Is there a significant difference in the purchase amounts between the 2 variants?

CSE 7215c

t-Test: A/B Testing in eCommerce Industry

Amount Purchased (INR)	
5003	3400
2345	3476
2341	453
3489	5643
321	5643
236	349
712	761
2190	4321
324	987
451	549
2315	579
713	5612
453	651
678	670
902	654
910	2398
2110	549
312	982
1035	1129
	837
	1057

	Male	**Female**
Total	100	100
Purchased (# of people)	15	25
Purchased Variant A	10	9
Purchased Variant B	5	16
Amount Purchased Variant A	12243	14597
Amount Purchased Variant B	9456	31244

CSE 7215c

t-Test: A/B Testing in eCommerce Industry

What is the null hypothesis?

$H_0: \mu_1 - \mu_2 = 0$ (Variant A is not different from Variant B)

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?

Variant A leads to higher sales than Variant B.

CSE 7215c

t-Test: A/B Testing in eCommerce Industry

```
ttest2U <- t.test(VariantA, VariantB, conf.level = 0.95)
```

```
        Welch Two Sample t-test

data:  VariantA and VariantB
t = -1.0306, df = 35.299, p-value = 0.3097
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1560.1965   509.2692
sample estimates:
mean of x mean of y
 1412.632  1938.095
```

HYPOTHESIS TEST

CONFIDENCE INTERVAL

Can true difference in means be equal to 0?

R code: pt(-1.0306,35.299) = 0.154869
All software output double the calculated value for 2-tailed tests to enable easy check against the significance level (α) instead of with $\frac{\alpha}{2}$.

Fail to Reject H0 at 5% level of significance since p(.3097)>alpha(.05). You can compare the calculated value of t to the critical value of t and make that decision.
True diff of mean is zero. Zero is contained in the confidence interval thus fail to reject H0.

CSE 7215c

t-Test: A/B Testing in eCommerce Industry

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

Fail to reject. That means Variant A has no significant difference from Variant B in terms of purchasing behaviour of people.

CSE 7215c

χ^2 DISTRIBUTION

CSE 7215c

χ^2 Test for Categorical Variables' Independence

χ^2 distribution uses a test statistic to check the independence of two categorical variables. It computes the difference between the **expected** and the **observed frequencies**, 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.

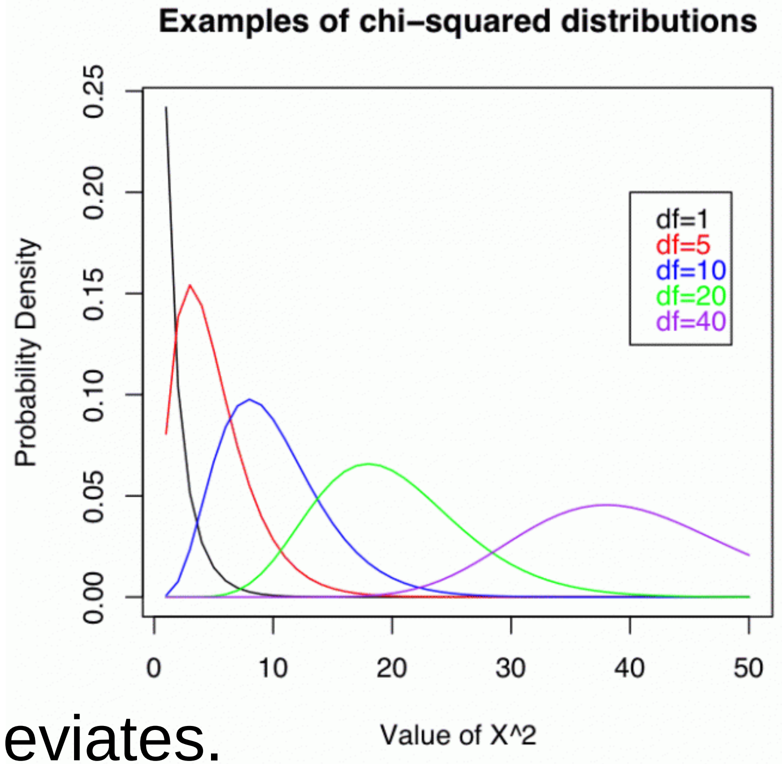
CSE 7215c

χ^2 Distribution

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

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

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



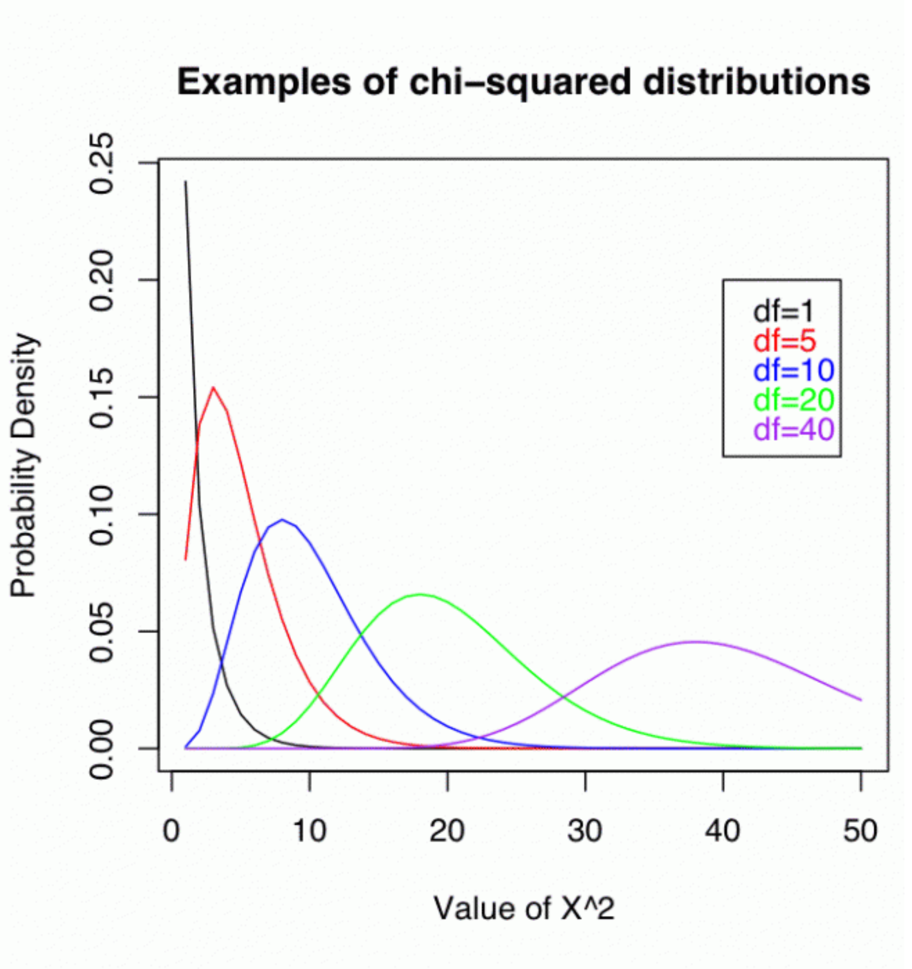
Thus χ^2 distribution is a distribution of the squared deviates.

CSE 7215c

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.



72 15c

Properties of X^2 Random Variable

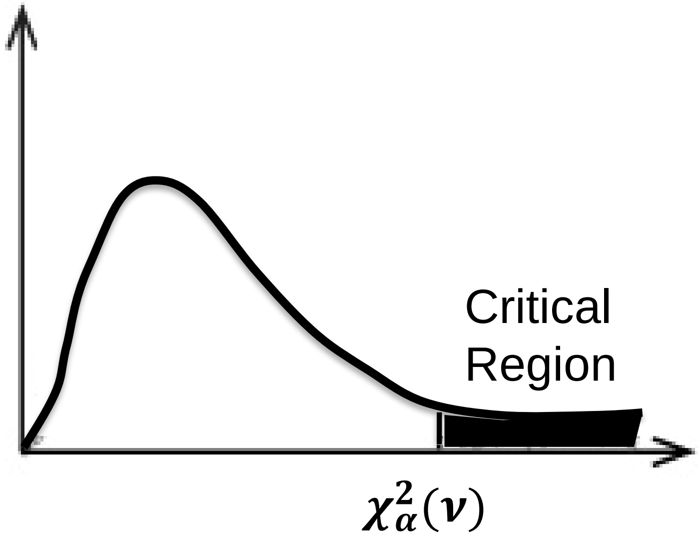
- A X^2 random variable takes values between 0 and ∞.
- Mean of a χ^2 distribution is ν.
- Variance of a χ^2 distribution is 2ν.
- The shape of the distribution is skewed to the right.
- As ν increases, Mean gets larger and the distribution spreads wider.
- As ν increases, distribution tends to normal.

CSE 7215c

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.

CSE 7215c

χ^2 Independence Test in Gaming Industry

You are a casino manager and suspect fraud. You think you are losing more money from one of the croupiers (dealers) on the blackjack tables. You want to test if the outcome of the game is dependent on which croupier is leading the game.

CSE 7215c

χ^2 Independence Test in Gaming Industry

Possible Outcomes

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

Observed Results

CSE 7215c

χ^2 Independence Test in Gaming Industry

The null hypothesis assumes that choice of croupier is independent of the outcome, which is rejected if there is sufficient evidence against it.

A **contingency table** is drawn to find the expected frequencies using probability.

CSE 7215c

χ^2 Independence Test in Gaming Industry

	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,

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

CSE 7215c

χ^2 Independence Test in Gaming Industry

	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

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

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

CSE 7215c

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

CSE 7215c

χ^2 Independence Test – Calculating X²

	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
	$\sum O = 258$	$\sum E = 258$	$\sum \frac{(O-E)^2}{E} = 1.6155$

CSE 7215c

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.

CSE 7215c

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)

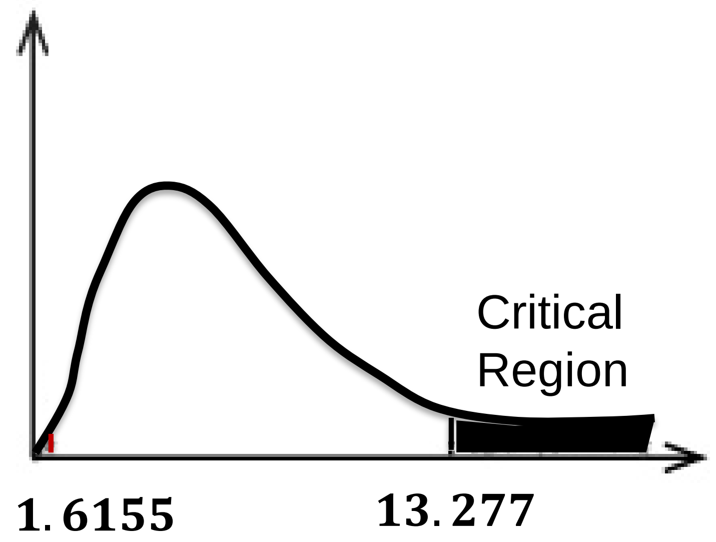
Fail to Reject H0 at 1% level of significance since calculated value of Chisq(1.6155) is less than the critical to the critical value of Chisq(13.277).
This can also be done via p-values comparison to alpha(.01)

CSE 7215c

χ^2 Independence Test – Decision

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

R: pchisq(1.6155,4,lower.tail=FALSE) = 0.806. This is >0.01.



Fail to Reject H0 at 1% level of significance since p-value(.806) is greater than .01

CSE 7215c

χ^2 Independence Test in Policymaking

There is widespread abuse of prescription drugs for a variety 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."

CSE 7215c

χ^2 Independence Test in Policymaking

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 α = 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

CSE 7215c

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.

CSE 7215c

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

$\nu = 4$

CSE 7215c

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)

CSE 7215c

Step 4: Calculate the Test Statistic X^2

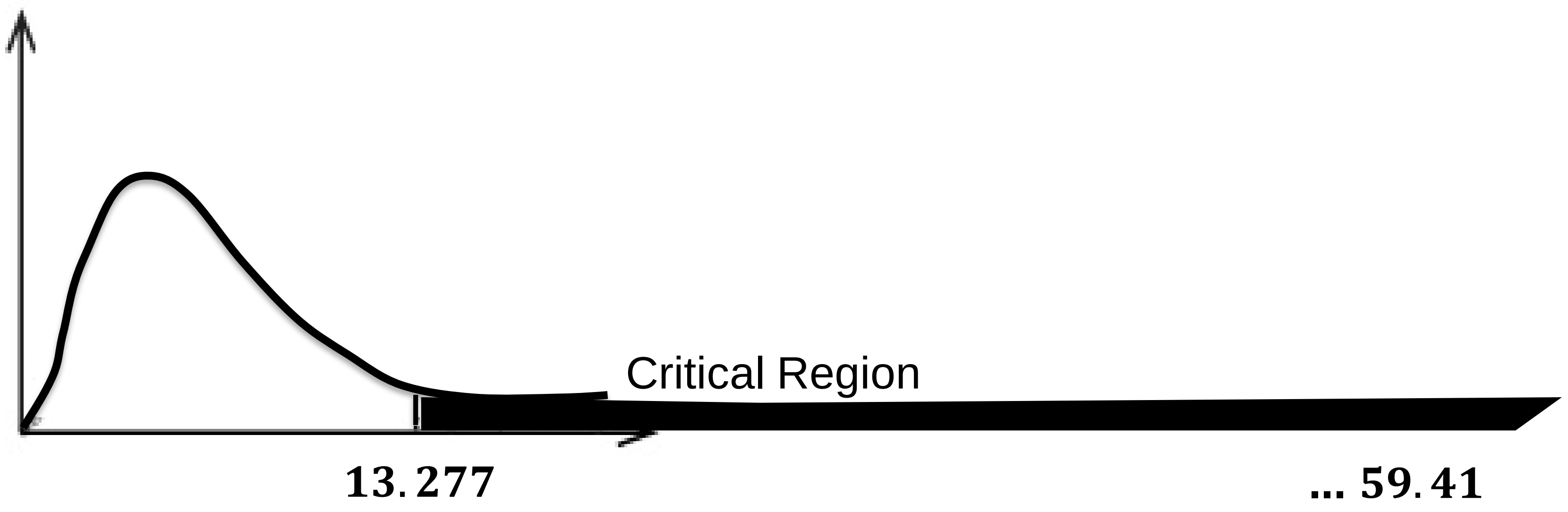
	Observed	Expected	$\frac{(O-E)^2}{E}$
	26	40.94	
Pain relievers	41	29.74	
	24	20.32	
	95	66.58	
Tranquilizers/ Sedatives	40	48.38	
	13	33.05	
	18	31.49	
Stimulants	20	22.88	
	32	15.63	
	$\sum O = 309$	$\sum E = 309$	$\sum \frac{(O-E)^2}{E} = 59.41$

CSE 7215c

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.

R: pchisq(59.41,4,lower.tail=FALSE) = 3.86×10^{-12}. This is <0.01.



Reject H0 at 1% level of significance since calculated value of Chisq(59.41) is in the critical region or greater than the critical value of Chisq(13.277).

CSE 7215c

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 7215c

F DISTRIBUTION

CSE 7215c

F Distribution

- 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?

CSE 7215c

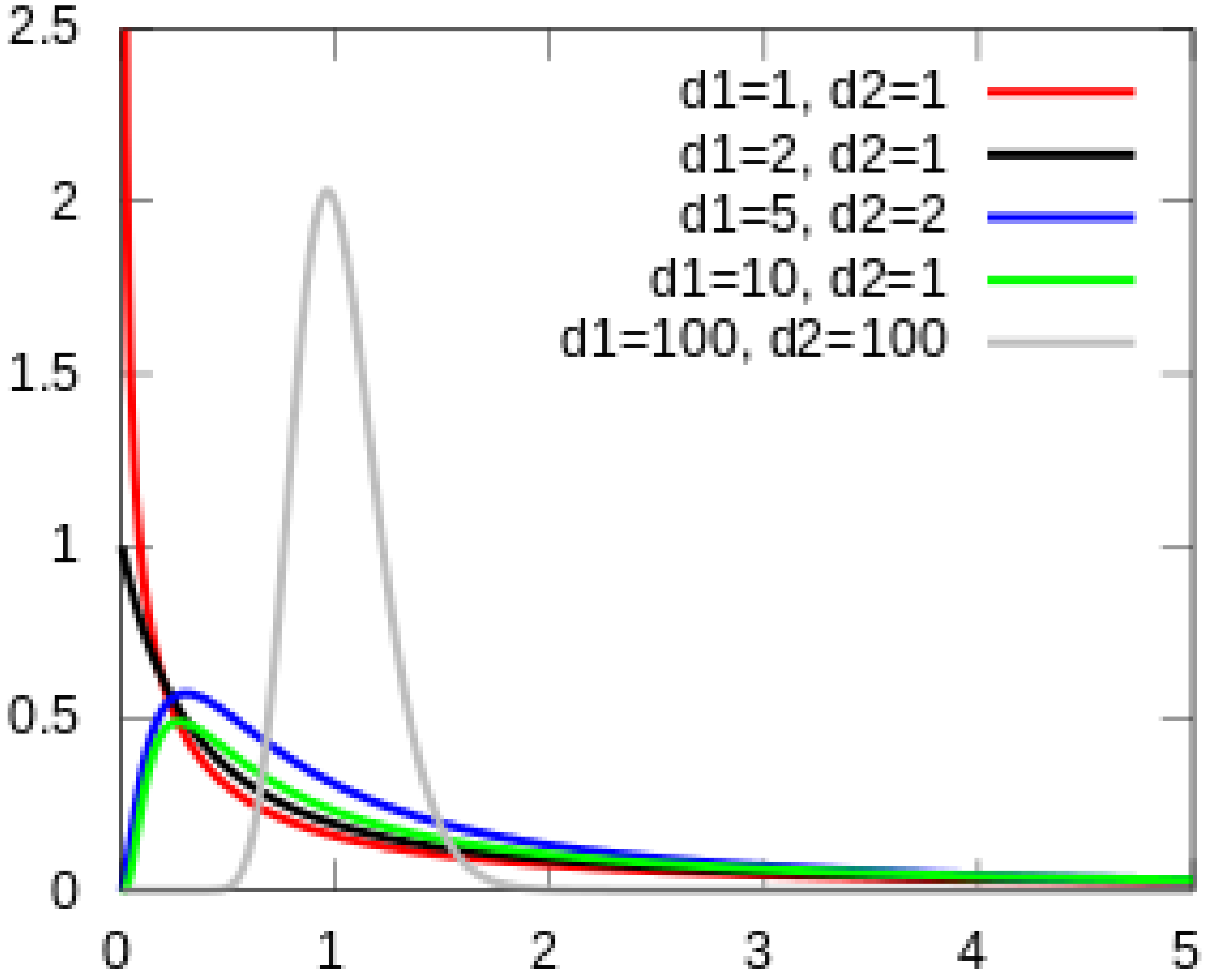
F Distribution

- 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.
- Another equation for Chi-Squared is $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$.
- So, F is also a ratio of 2 chi-squares, each divided by its degrees of freedom, i.e.,

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

CSE 7215c

F Distribution



CSE 7215c

F Test in Manufacturing Industry

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 7215c

F Test in Manufacturing Industry

Ratio of sample variances,

$$F = \frac{s_1^2}{s_2^2} = \frac{0.11378}{0.02023} = 5.62$$

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

CSE 7215c

F Test in Manufacturing Industry

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?

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

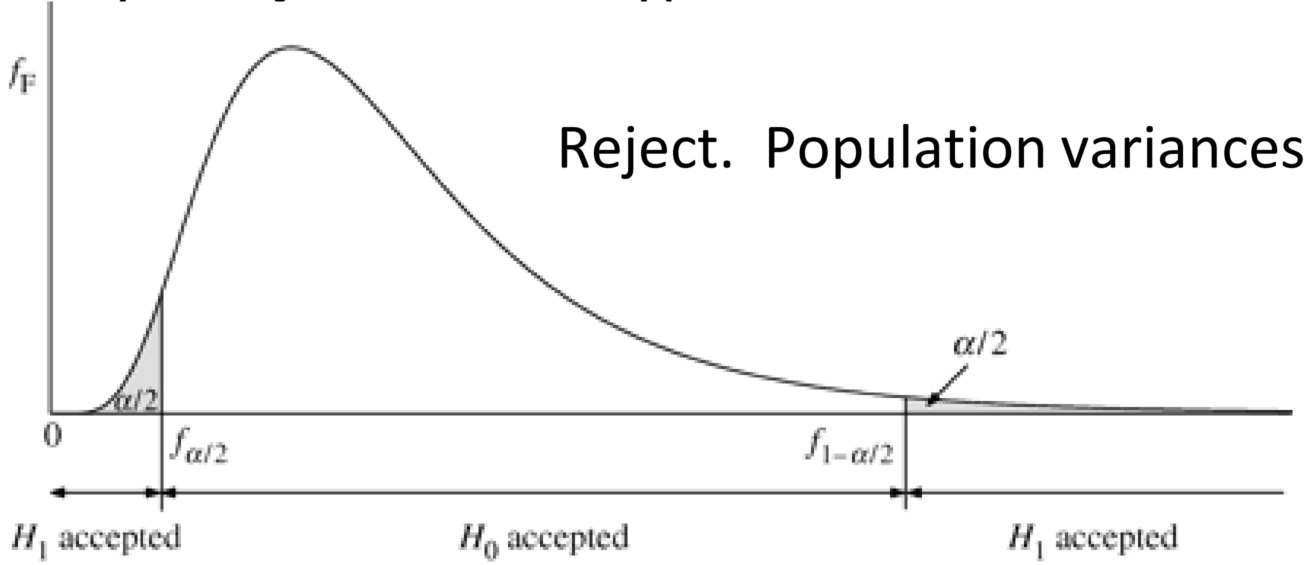
CSE 7215c

F Test in Manufacturing Industry

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

R-code: qf(α, df_1, df_2)   *pf(5.62,9,11,lower.tail=FALSE) = 0.0047*

Will you reject the null hypothesis or not?



Reject. Population variances are not equal.

Reject H0 at 5% level of significance since calculated value of F(5.62) is in the right critical region or greater than the critical value of F(3.5879).
P value(.0047) <.05 thus reject null hypothesis

CSE 7215c

F Test in Manufacturing Industry

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 7215c

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 7215c

Thought Process on When to Use a Particular Test

What do you want to do?

- Description – Summary statistics, Various plots, Correlations
- Intervention (differences between groups) – *t*-test, Chi-square, F-test
- Prediction – Linear regression, Logistic regression *(next module)*

CSE 7215c

THE STORY

CSE 7215c

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.

CSE 7215c

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 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.

CSE 7215c

We studied Binomial (discrete) and Normal (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.

CSE 7215c

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 about populations from sample data.

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

CSE 7215c

χ^2 Distribution was useful in studying (in)dependence between categorical variables.

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

CSE 7215c

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.

CSE 7215c

Some good resources

- http://onlinestatbook.com
- http://stattrek.com
- http://www.khanacademy.org
- http://www.statsoft.com/Textbook
- http://vassarstats.net/textbook
- Applied Business Statistics by Ken Black
- Statistics For Business: Decision Making and Analysis by Robert Stine and Dean Foster
- The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition by Trevor Hastie, Robert Tibshirani and Jerome Friedman

CSE 7215c

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, 3A Main Road
Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

PUNE

Kirloskar - Pune
S. L. Kirloskar Center for Executive Education,
Kirloskar Corporate Office, 8th Floor,
Cello Platina, Model Colony, Shivaji Nagar – 411005

MUMBAI

Kanakia Wall Street, 4th Floor, Andheri-Kurla Road
Chakala, Andheri East, Mumbai - 400093

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.