# What happens when we don’t know population standard deviation

## t-Distribution

* William Gossett of Guinness Breweries
* Published “[The Probable Error of the Mean](http://www.york.ac.uk/depts/maths/histstat/student.pdf)”
* Why “Student”?
* Ronald Fisher Extended the Concept

## Degrees of Freedom

### First way of looking

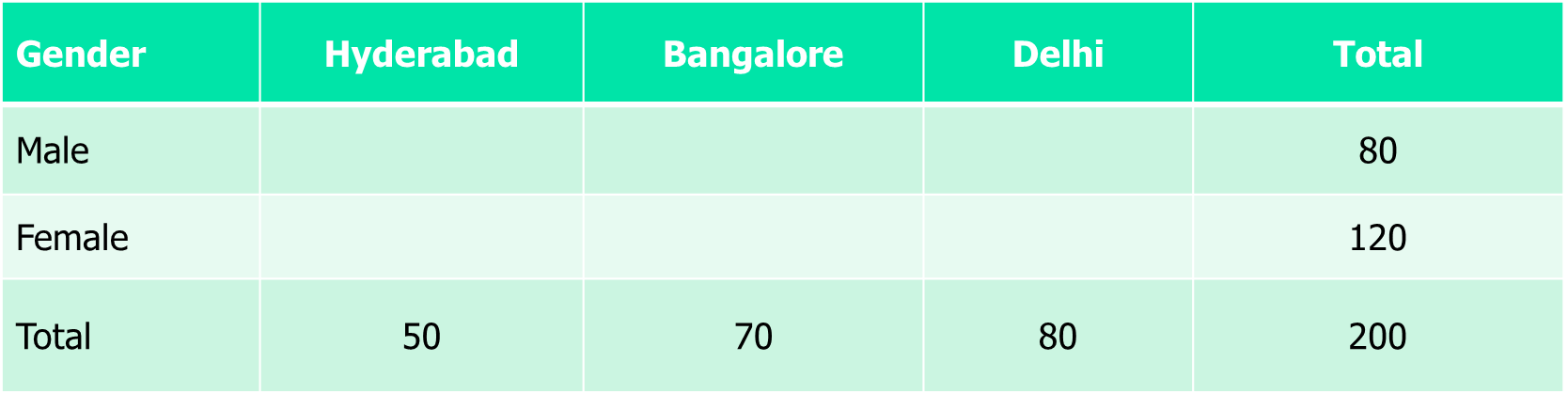
Suppose we have 83 students.

1. First person comes he has 83 choices.
2. Next one who comes has 82 choices.
3. …Second last person has 2 choices
4. Last point has no choice. 0 freedom available.

So if there are n people in a class n-1 people had freedom. Last one doesn’t have any freedom.

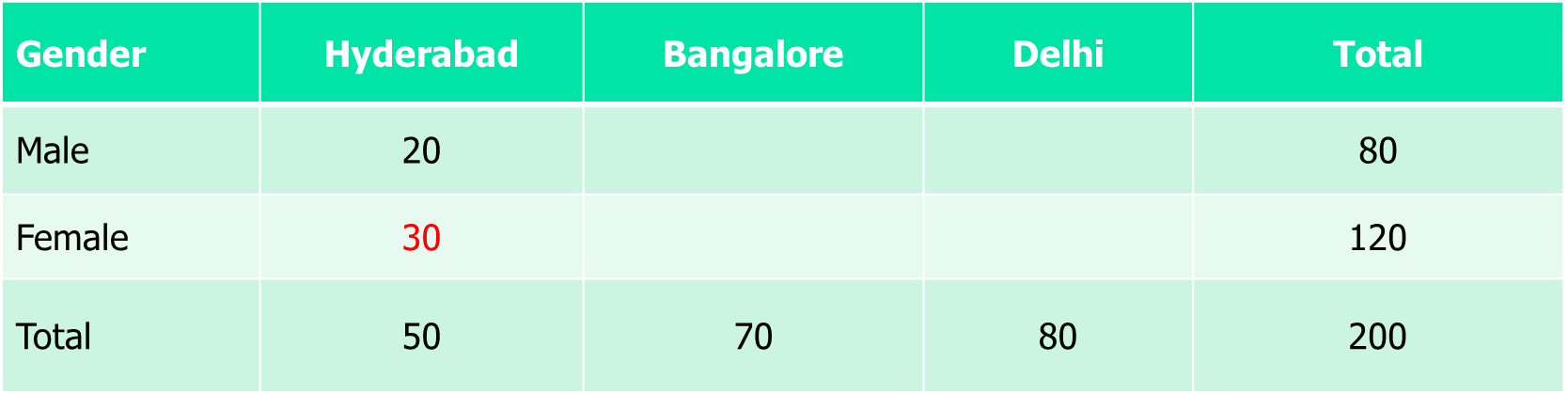
### Second way of thinking

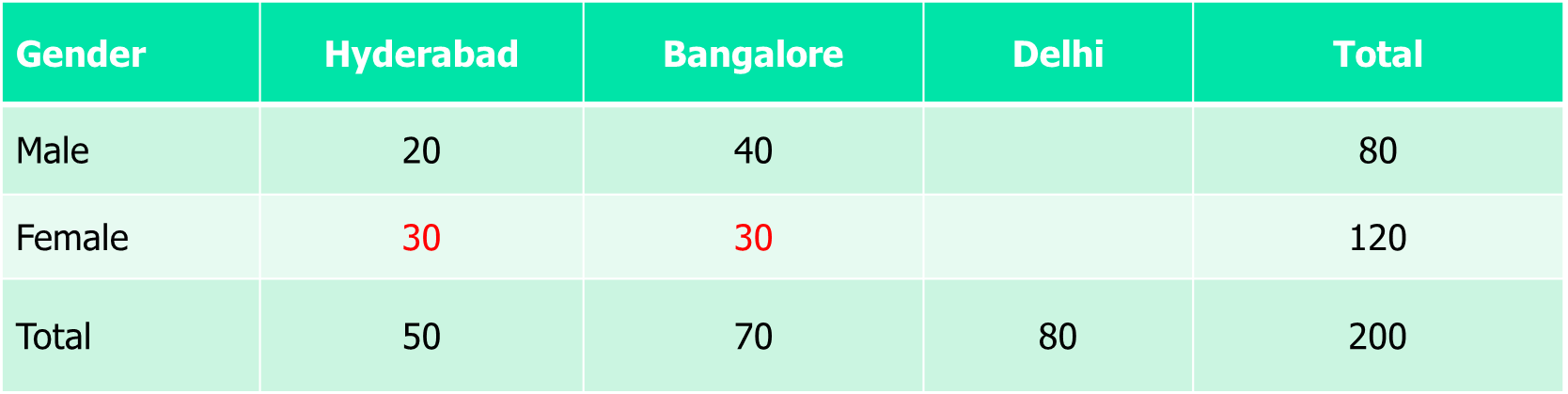
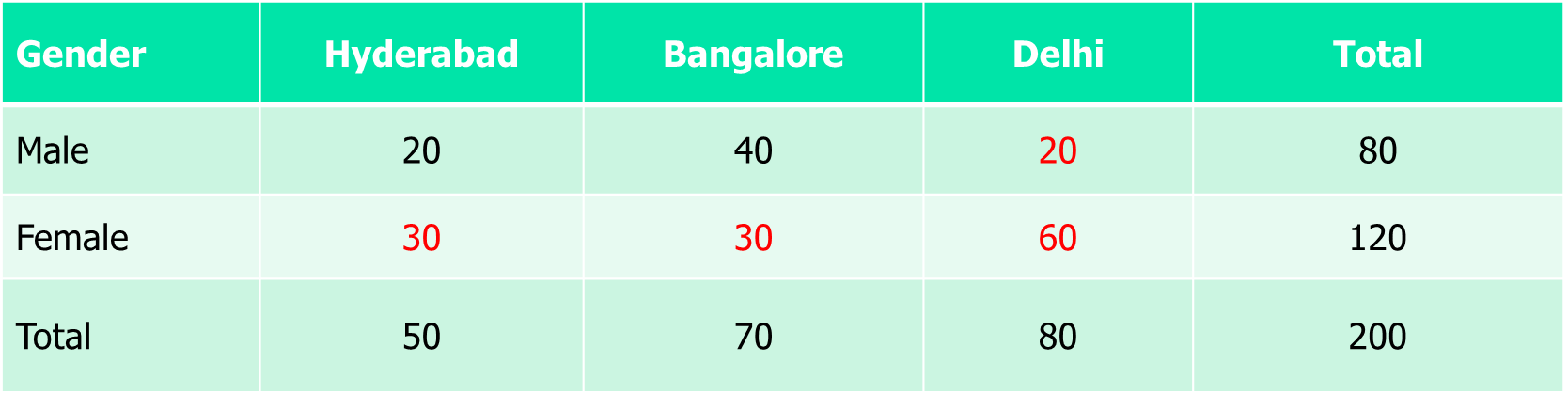
I want to collect samples from Hyderabad, Bangalore, Delhi. Each sample should have male and female populations as below:



Now there are 6 cells. Now how many cells can you fill with a choice upon keeping the constraints?

1. Suppose Male in Hyderabad = 25 That gives Female in Hyderabad with no choice



1. Bangalore Male 40 is choice Female Bangalore has no choice: 
2. Now Chennai has no freedom:

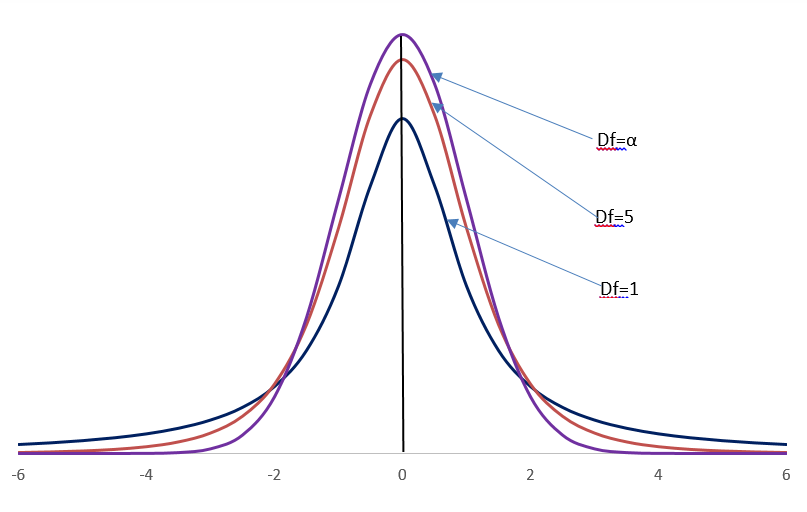
Thus formula for degrees of freedom = **(row -1)\*(column-1) =** 1\*2 = 2

## t - Distribution computation steps

1. We don't know µ and σ for the population
2. We take a sample of size n out of that population and calculate the sample mean = X\_bar
3. From the same sample we calculate the sample standard deviation = S. We make and estimation that the population standard deviation σ will be a number very close to S.
4. We calculate the standard deviation of sample means as SX\_bar = S/sqrt(n).
5. The degree of freedom = n-1
6. So we calculate the probabilities between our set confidence interval through t-distribution function for the given degree of freedom, sample mean and standard deviation of sample means.

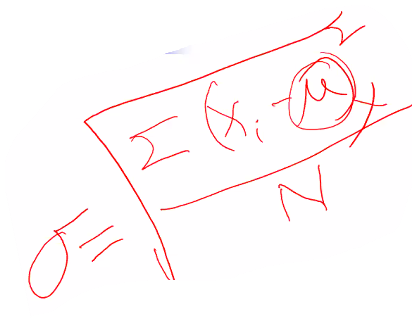
# Effect of changing degrees of freedom

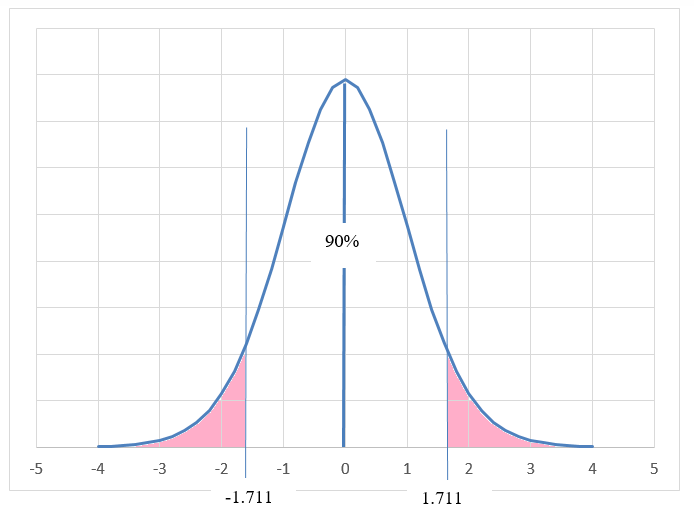
Increasing degrees of freedom will make the t-distribution will tend to become normal distribution.



# So what does it tell

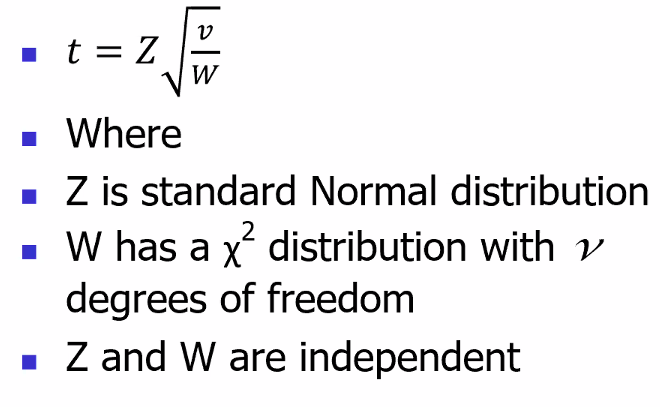
For a population: The problem is we don’t know µ



1. If you don’t know σ, calculate the sample standard deviation as: *S* =
2. Since we don’t know σ we are calculating S, we are bringing in more uncertainty in our estimation.
3. To account for the estimation t-distribution takes up the adjustments in 2 ways:
   1. Use degrees of freedom n-1 to calculate S which means S is slightly higher than the formula
   2. Use the t table.
4. 
5. There are observations:
   1. Remember from saree sagar example that Z0.90 for normal distribution was ±1.645
   2. Here the width from the mean is ±1.711.
   3. This means that with t – distribution we are increasing the width of the uncertainty within which µ is expected to lie. That means we are playing it safe in estimating µ.

# What did we understand?

If we know σ, there is no uncertainty. If σ is not known there is another uncertainty in estimating σ along with µ. t-distribution accounts for that uncertainty.



This formula takes care of the uncertainties.

See the session 4 video first part to understand which variable takes care of what uncertainties.

# Proportion of Orders (p)

In Saree Sagar example we did the following:

Till now we have estimated µ for two cases:

1. When the population standard deviation is known.
2. When the population standard deviation is unknown.

Now consider the situation: Out of 2000 customers reached only 144 reverted. So the hit rate was 7.2%. That gives rise to following 2 questions:

1. What will happen if we take another sample of 2000 customer. Will we get the same hit rate?
2. What will be the estimated hit rate throughout the tenure of business?

This gives rise to the need of estimation of hit rate a.k.a **Population Proportion (denoted as π).**

## Distribution of Sample proportion

* The formula for mean is
* Consider Q as a binomial random variable (no. of successes)
* Q~B{nπ,nπ(1-π)} where π is the probability of success (population proportion)
* n is the number of trials and each trial results in a success or failure
* Consider each success as 1 and each failure as 0 (i.e., Xi = {0,1}) and *=> Q* measures number of successes. [1,0,0,1,1,0,0,0,1,0…]
* Sample proportion, which is mathematically identical to In other words p is number of successes/total.

## Derivation of distribution of p

Below we will prove that p follows a normal distribution with mean = π and standard deviation .

* n = number of trials = sample size in our case = 2000 in our case.
* Similarly for variance in variance everything gets squared so 1/n2.
* So

So

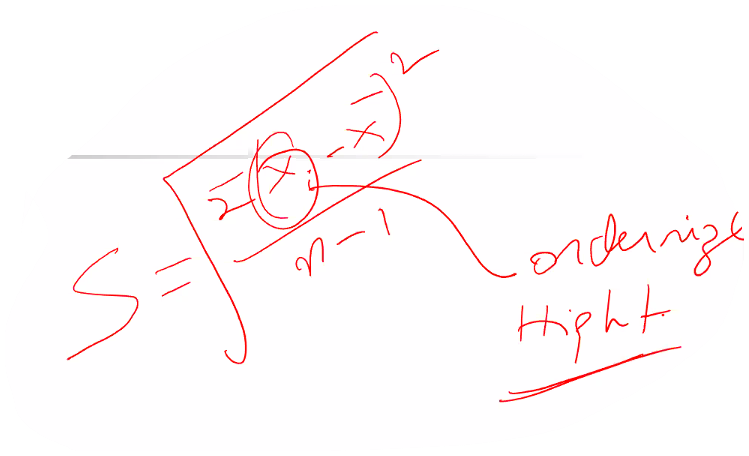
## Estimating population proportion π

Now here we have a vicious cycle:

* We are trying to estimate π for that we need to know and to know we need π
* We do not know π but we use our best guess (point estimate) *p* i.e. the hit rate of the samples selected.
* When we didn’t know σ during t-test we used a proxy for σ as the sample standard deviation S. Then we calculated the sample means distribution statistics using the mean and standard deviation of sample means as mean of sample and standard deviation of the sample/sqrt(sample size).
* Similarly define an estimate of standard deviation of population proportion with . This implies the standard deviation of sample proportions is .
* The (1 - a)% confidence interval can then be specified as: 
* What is the 90% confidence interval for the proportion of SariSagar offers accepted?

## Why aren’t we using t – distribution

See the formula for calculating sample std. dev. in t – distribution.



In t distribution this X\_bari is a quantifiable variable. Like height, order price etc.

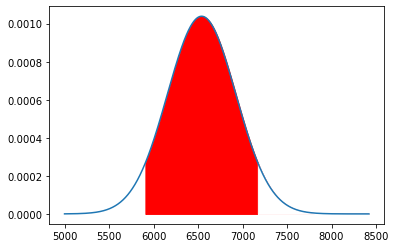
Now here while thinking about p we are dealing with p we are dealing with a nominal variable whether sari was bought or not. So for nominal variables we can’t use t distribution.

# How big a sample size should be?

## When σ is known

For the sari sagar suppose we are told that to estimate the sample size such that 90% Confidence interval falls within ±500 INR.

In other words it is telling giving us the following constraints:



X1

X2

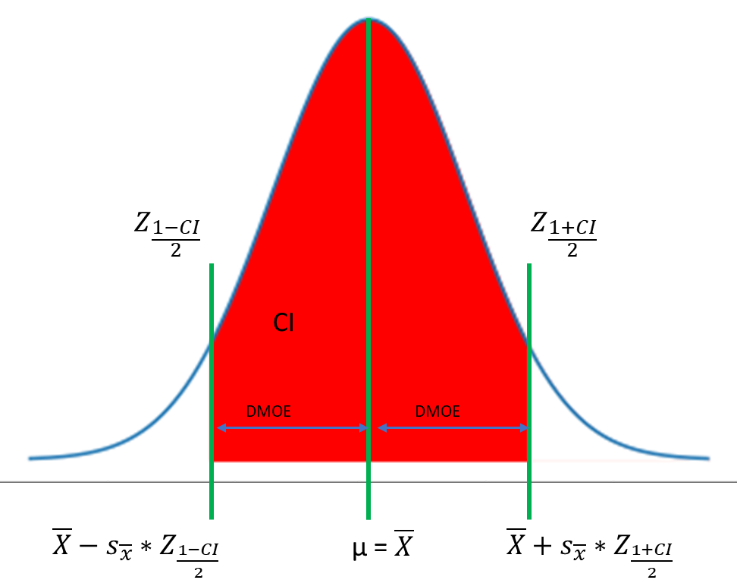
90%

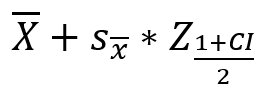
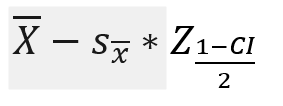
X2 – X1 is the Desired Margin of Error\*2 or DMOE\*2

### Proof

Let’s prove the generalized formulae for calculation:

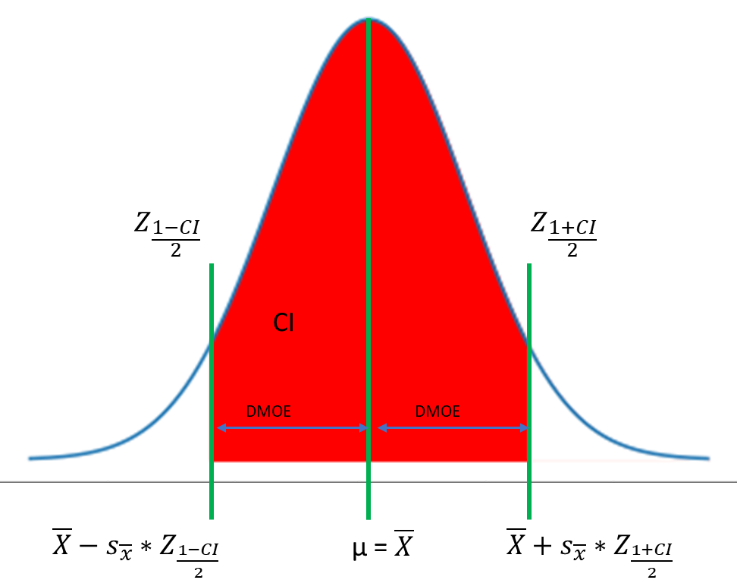
We want the confidence interval to lie within ±DMOE.



1. What are the given parameters in this case?
   1. CI: The confidence interval is predetermined.
   2. 2 x DMOE = Upper - Lower limit i.e. - ≤ 2 x DMOE is known
   3. is not known but as per point b we will see it will be cancelled in the inequation.
   4. is known n we will have to find out.
2. From the above diagram we can write that

1. cancelling from the LHS.
2. replace with and taking common.
3. rearranging equation 4.
4. Thus .
5. Due to symmetry of normal distribution
6. Thus

The other way to remember the above equation is that in the figure for normal distribution



The length = standard deviation of sample means x

## When σ is not known

1. We have to estimate σ from the sample standard deviation = s
2. Then standard deviation of sample means = .
3. Now there is a problem I got into. A vicious cycle. We can estimate σ only after we know s. To know s we have to collect a sample of n observations. But what will be n? n is something that we are ourselves trying to solve.
4. For t distribution in case σ is not known: we have to use t value instead of Z value. And to get t value for a certain confidence interval we have to know the number of samples to be taken to know the degrees of freedom.
5. So to get out of this dependency loop we take a **pilot sample** of whatever size conduct a trial to estimate S and error at appropriate value of and n pairs.
6. If I come up with an n the t value for that corresponding n will be different.
7. So from the pilot sample we need to iterate over to get an n for the corresponding t value by either increasing or decreasing the value of n by 1.

## Minimum Sample size for estimating proportion

What is the minimum sample size required to estimate the sample proportion within DMOE% of the population proportion with confidence interval CI?

1. Replace with as σ is estimated as where p = sample success rate.
2. Now p is not known if we are not doing sampling and going through the pain of reaching out customers while at the same time we need to know what should be the appropriate sample size.
3. To overcome this difficulty we take help of mathematics concepts of maximum value. p(1-p) is maximum at p = 0.5 thus giving a maximum variance of 0.25. Rest all values of p is good. But p = 0.5 is the worst case. i.e. we launch the product and have a 50% hit rate through out our life.

# Summary of Population parameter and sample statistics

