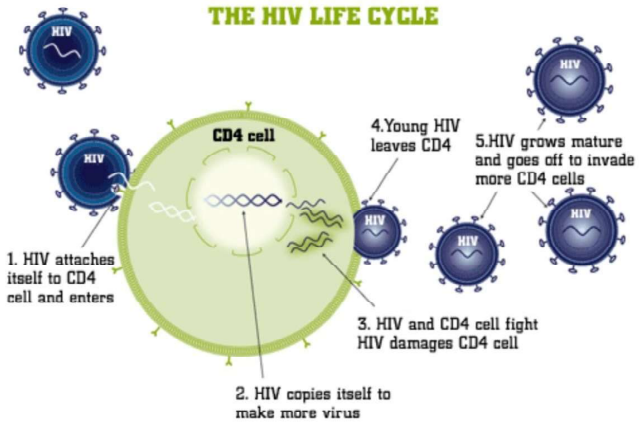
t-Distribution - Example

A researcher wants to examine CD4 counts for HIV+ patients at her clinic. She randomly selects a sample of 25 HIV+ patients and measures their CD4 levels. Calculate a 95% CI for population mean given the following sample results:



Variable	*n*	$\bar{x}$	SE of Mean	*s*	Min	Q1	Median	Q3	Max
CD4 (cells/μl)	25	321.4	14.8	73.8	208.0	261.5	325.0	394.0	449.0

t-Distribution - Example

Variable	*n*	$\bar{x}$	SE of Mean	*s*	Min	Q1	Median	Q3	Max
CD4 (cells/μl)	25	321.4	14.8	73.8	208.0	261.5	325.0	394.0	449.0

$$CI(0.05)\colon\ \bar{x} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

$$321.4 - t_{25-1,\frac{0.05}{2}}\frac{73.8}{\sqrt{25}} \leq \mu \leq 321.4 + t_{25-1,\frac{0.05}{2}}\frac{73.8}{\sqrt{25}}$$

$$t_{24,0.025} = 2.064$$

$$\therefore CI(0.05)\colon [290.85, 351.95]$$

t-Distribution - Example

What does $CI(0.05)$: $[290.85,351.95]$ mean in the business context given that US Government classifies AIDS under three official categories of CD4 counts:

- Asymptomatic: $\geq$ 500 cells/μl
- AIDS related complex (ARC): 200-499 cells/μl
- AIDS (Stage 3 infection): < 200 cells/μl

Application of t-test

- **Your company wants to improve sales. Past sales data indicate that the average sale was $100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work? Test your hypothesis at a 5% alpha level.**

- Step 1: Write your null hypothesis statement (How to state a null hypothesis).
- The accepted hypothesis is that there is no difference in sales, so:
  H_0: μ = $100.

Application of t-test

- Step 2: Write your alternate hypothesis. This is the one you're testing. You think that there *is* a difference (that the mean sales increased), so:
  H_1: μ > $100.

- Step 3: Identify the following pieces of information you'll need to calculate the test statistic. The question should give you these items:

- **Given:**
- **The sample mean**($\bar{x}$). This is given in the question as $130.
- **The population mean**(μ). Given as $100 (from past data).
- **The sample standard deviation**(s) = $15.
- **Number of observations**(n) = 25.

Application of t-test

- Step 4: Insert the items from above into the t score formula.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- t = (130 – 100) / ((15 / √(25))
  t = (30 / 3) = 10
  This is your **calculated t-value.**

Application of t-test

- Step 5: Find the t-table value. You need two values to find this:
- The alpha level: given as 5% in the question.
- The degrees of freedom, which is the number of items in the sample (n) minus 1: 25 – 1 = 24.
- Look up 24 degrees of freedom in the left column and 0.05 in the top row. The intersection is 1.711.This is your one-tailed critical t-value.
- What this critical value means is that we would expect most values to fall under 1.711. If our calculated t-value (from Step 4) falls within this range, the null hypothesis is likely true.
- Step 6: Compare Step 4 to Step 5. The value from Step 4 **does not** fall into the range calculated in Step 5, so we can reject the null hypothesis. The value of 10 falls into the rejection region.
- In other words, it's highly likely that the mean sale is greater. The sales training was probably a success.

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

EXAMPLE FOR T – TEST

- A car company claims that their Super Spiffy Sedan averages 31 mpg. You randomly select 8 Super Spiffies from local car dealerships and test their gas mileage under similar conditions.
- You get the following MPG scores:
- MPG: 30 28 32 26 33 25 28 30
- Does the actual gas mileage for these cars deviate significantly from 31 (alpha = .05)?

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

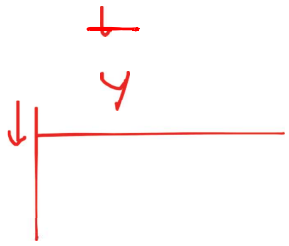
Testing Hypotheses about a Variance

What are null and alternate hypotheses?

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

How many degrees of freedom?

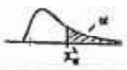
Since n=20, df=19.



Testing Hypotheses about a Variance

What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



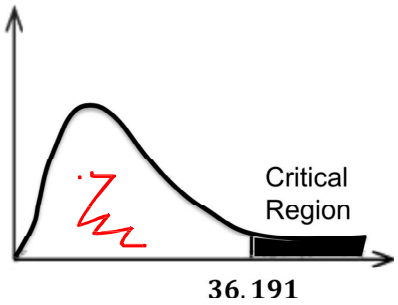
ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	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	0.554	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	3.074	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

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

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.

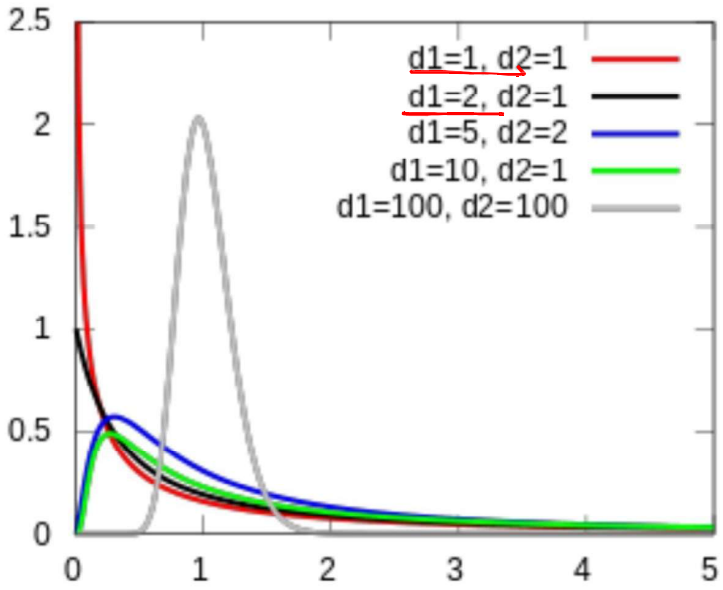
F distribution

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

F distribution

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

F distribution



Hypothesis test for 2 sample variances

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

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

How do you test this?

Hypothesis test for 2 sample variances

Data

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

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

Hypothesis test for 2 sample variances

What are null and alternate hypotheses?

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

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

Two-tailed.

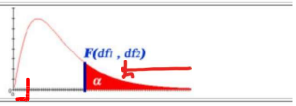
What are numerator and denominator degrees of freedom?

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

Hypothesis test for 2 sample variances

Reading an F-table.

F Table for $\alpha = 0.025$



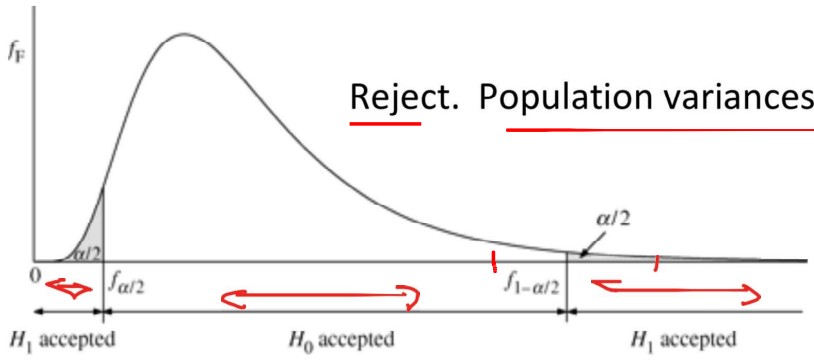
/	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	
df_2=1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274	976.7079	984.8668	993.1028	997.2492	1001.414	1005.598	1009.800	1014.020	101
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980	39.4146	39.4313	39.4479	39.4562	39.465	39.473	39.481	39.490	3
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	1
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	
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	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	
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	
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	
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	
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	
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	
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	

$$F_{0.025,9,11} = 3.5879; F_{0.975,9,11} = \frac{1}{F_{0.025,9,11}} = 0.2787$$

Hypothesis test for 2 sample variances

$$F_{0.025,9,11} = 3.5879;\ F_{0.975,9,11} = \frac{1}{F_{0.025,9,11}} = 0.2787;\ F_{observed} = 5.62$$

Will you reject the null hypothesis or not?



Reject. Population variances are not equal.

Hypothesis test for 2 sample variances

What are the business implications?

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

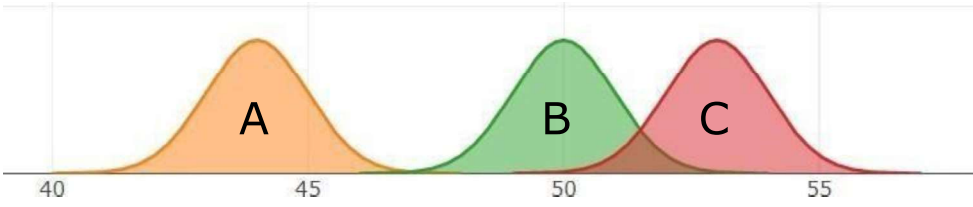
Applications of F Distribution

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

ANOVA
ANALYSIS OF VARIANCE

ANOVA

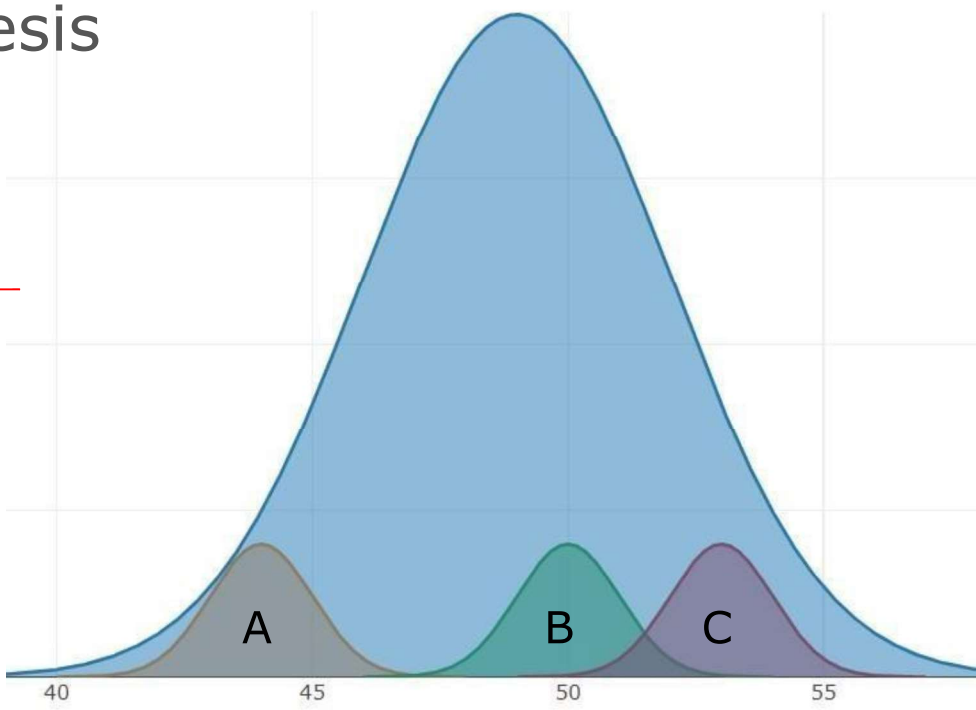
- In the previous section we tested samples to see if they are the representative of the population.
- What if we had three (or more) samples?
- Could we find if they are from same population?



ANOVA

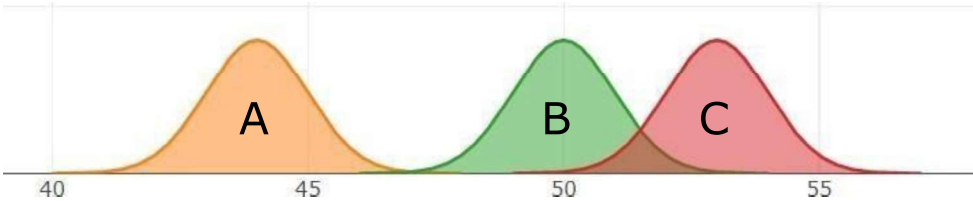
- Our null hypothesis would look like:

$$H0: \mu_A = \mu_B = \mu_C$$



ANOVA

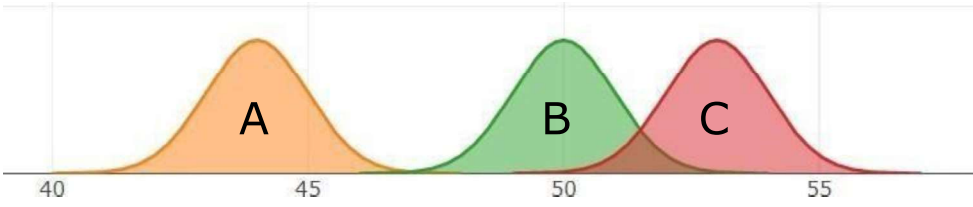
- This is where ANOVA comes in!
- We compute an F value, and compare it to a critical value determined by our degrees of freedom (the number of groups, and the number of items in each group)



ANOVA

Let’s work with somedata:

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43



ANOVA

First calculate the sample mean

Next calculate the overall mean

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$F =$

ANOVA considers two types of variance:

- Between Groups
  - how far group meansstray from the total mean
- Within Groups
  - how far individual values stray from their respective group mean

ANOVA

The F value we're trying to calculate is simply the ratio between these two variances!

$$F = \frac{VarianceBetweenGroups}{VarianceWithinGroups}$$

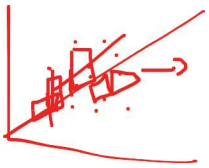
ANOVA

Recall the equation for variance:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{SS}{df}$$

Here $\Sigma(x - \bar{x})^2$ is the "sum of squares" SS and $n - 1$ is the "degrees of freedom" df

ANOVA



So the formula for the F value becomes:

$$F = \frac{VarianceBetweenGroups}{VarianceWithinGroups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

SSG=SumofSquaresGroups

SSE=SumofSquaresError

df_{groups}=degreesofreedom(groups)

df_{error}=degreesoffreedom(eror)

ANOVA

Sum of Squares Group

$(\mu_A - \mu_{TOT})^2 = (44-49)^2$

$(\mu_B - \mu_{TOT})^2 = (50-49)^2$

$(\mu_C - \mu_{TOT})^2 = (53-49)^2$

$SSG = 420$

s

= 25

= 1

=16

42

Multiply by the number of items in each group:

$42 \times 10 = 420$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	44	50	53
μ_{TOT}	49		

ANOVA

$SSG = 420$

$df_{groups} = 2$

Degrees of Freedom Groups

$$df_{groups} = n_{groups} - 1$$
$$= 3 - 1$$
$$= 2$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

Sum of Squares Error

$SSG = 420$

$df_{groups} = 2$

$SSE = 3300$

$(x_A-\mu_A)^2$	$(x_A-\mu_A)^2$	$(x_B-\mu_B)^2$	$(x_B-\mu_B)^2$	$(x_C-\mu_C)^2$	$(x_C-\mu_C)^2$
49	64	144	16	9	1
256	121	529	36	100	0
64	25	361	361	25	100
1	25	196	1	1	144
16	441	49	49	16	100
	1062		**1742**		**496**
				TOTAL	**3300**

$(37-44)^2 = (-7)^2 = 49$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$SSG = 420$

$df_{groups} = 2$

$SSE = 3300$

$df_{error} = 27$

Degrees of Freedom Error

$$df_{error} = (n_{rows} - 1) * n_{groups}$$

$$= (10 - 1) * 3$$

$$= 27$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$SSG = 420$
$df_{groups} = 2$
$SSE = 3300$
$df_{error} = 27$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43
$\mu_{A,B,C}$	44	50	53
μ_{TOT}	49		

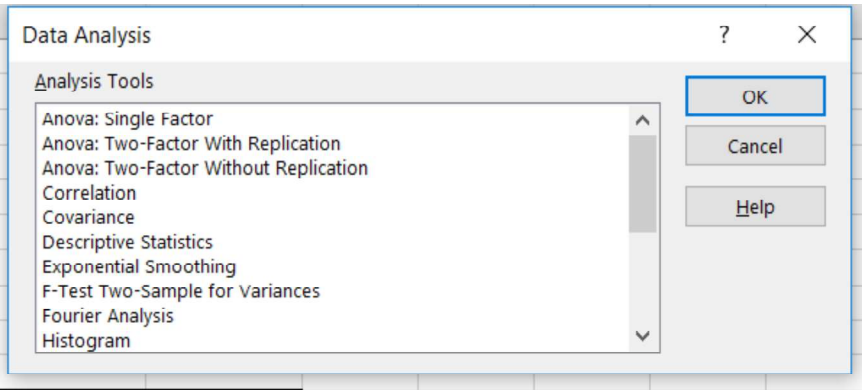
Plug these into ourfor mula:

$$F = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}} = \frac{\frac{420}{2}}{\frac{3300}{27}} = \frac{210}{122.2} = 1.718$$

ANOVA WITH EXCEL DATA ANALYSIS

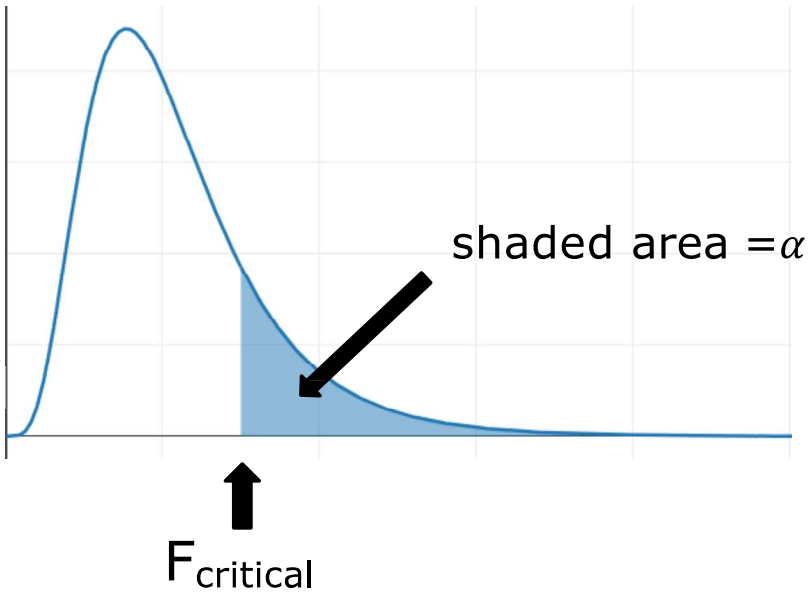
	A	B	C	D	E
1	Anova: Single Factor				
2					
3	SUMMARY				
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*
5	GroupA	10	440	44	118
6	GroupB	10	500	50	193.5555556
7	GroupC	10	530	53	55.11111111
8					
9					
10	ANOVA				
11	*Source of Variation*	*SS*	*df*	*MS*	*F*
12	Between Groups	420	2	210	1.71[illegible]1818
13	Within Groups	3300	27	122.2222	
14					
15	Total	3720	29		
16					

P-value	*F crit*
0.198430533	3.354130829



F DISTRIBUTION

F-DISTRIBUTION



F-DISTRIBUTION

Look up our critical value from an F-table

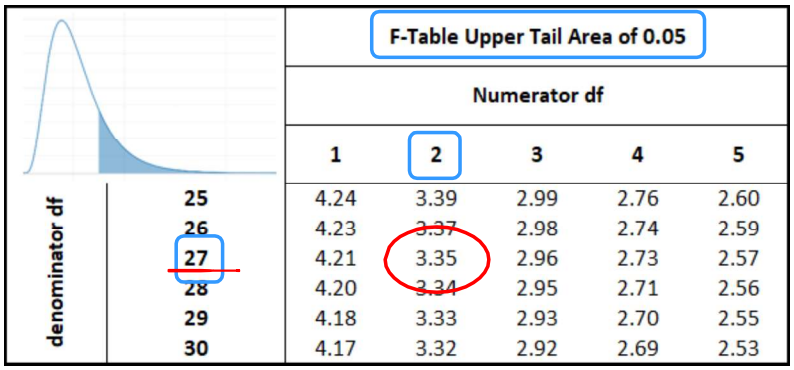
use a table set for

95% confidence

find numerator df

find denominator df

critical value =3.35



F-Table Upper Tail Area of 0.05						
		Numerator df				
		1	2	3	4	5
denominator df	25	4.24	3.39	2.99	2.76	2.60
	26	4.23	3.37	2.98	2.74	2.59
	27	4.21	3.35	2.96	2.73	2.57
	28	4.20	3.34	2.95	2.71	2.56
	29	4.18	3.33	2.93	2.70	2.55
	30	4.17	3.32	2.92	2.69	2.53

F-SCORES IN MS EXCEL

- In Microsoft Excel, the following function returns an F-score:

α	df1	df2	Formula	Output Value
0.05	2	27	=FINV(A2,B2,C2)	3.3541308285292

F-SCORES IN PYTHON

```
>>> from scipy import stats
>>> stats.f.ppf(1-.05,dfn=2,dfd=27)
3.354130828529198 6
```

ANOVA

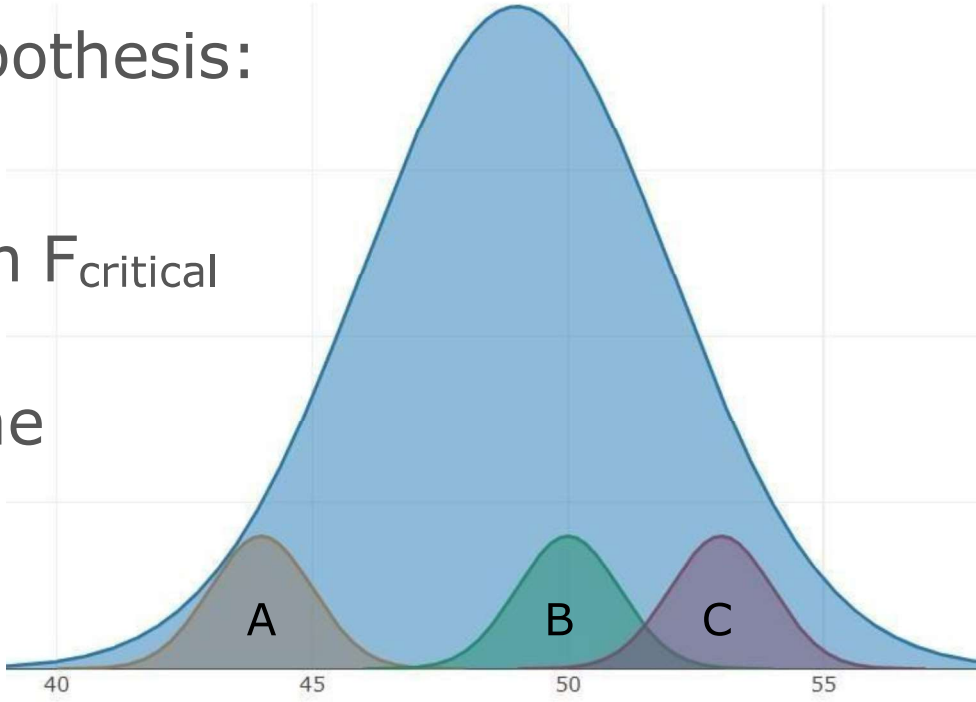
Recall our null hypothesis:

H_0: $\mu_A = \mu_B = \mu_C$

Since F is less than $F_{critical}$

1.718<3.354

wefail to reject the null hypothesis!



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

Groups	*Count*	*Sum*	*Average*	*Variance*
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

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

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