CDS501: PRINCIPLES & PRACTICES OF DATA SCIENCE & ANALYTICS

Statistical Hypothesis Testing

Copyright © 2021 Halim Noor, USM

Outline

- Hypothesis Testing
- t-Test
  - One-sample t-Test
  - Two-sample t-Test
- ANOVA
- Chi-square

Copyright © 2021 Halim Noor, USM

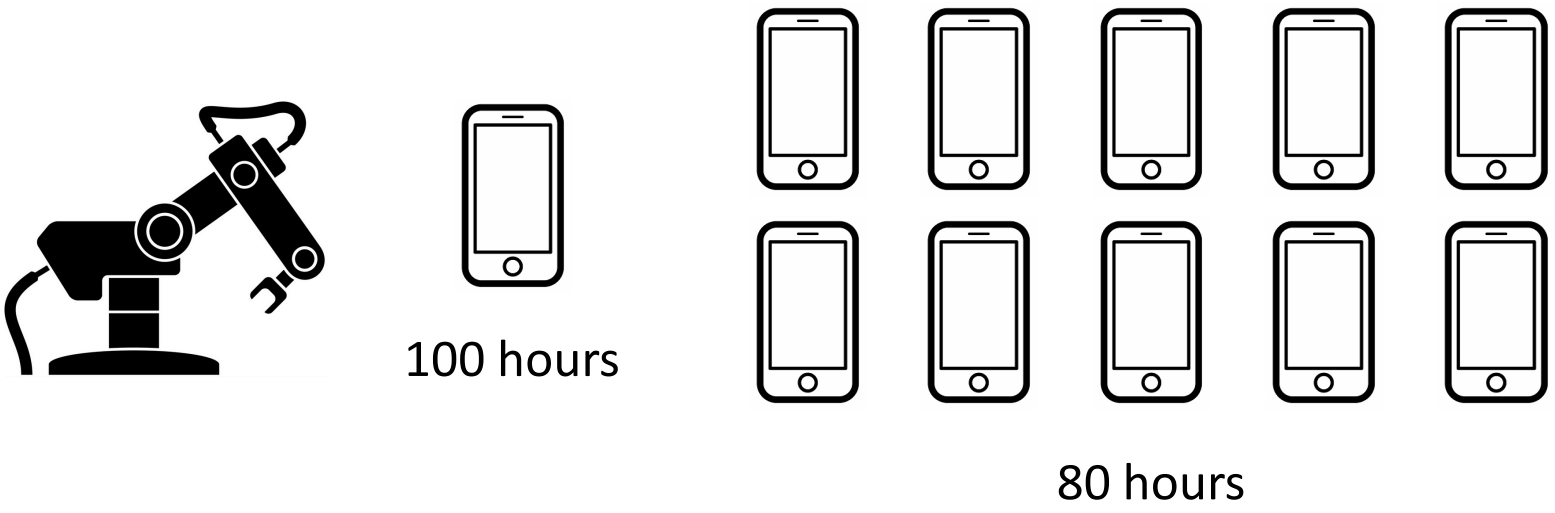
Hypothesis Testing

- A method to test whether a certain hypothesis is true or false
- Hypothesis test follows the following logic
  - An assumption or a claim is made
  - If the data contradicts the assumption or claim, then we can conclude the assumption made must be wrong

Copyright © 2021 Halim Noor, USM

Hypothesis Testing

- A mobile phone was designed and manufactured to have a battery lifetime of 100 hours.



Copyright © 2021 Halim Noor, USM

Hypothesis Testing

- If the battery lifetime across 10 units is 80 hours
  - Easy to conclude the phone was not properly designed
- If the battery lifetime across 10 units is 100 hours
  - Easy to conclude the phone was properly designed
- If the battery lifetime across 10 units is 99.9 hours or 100.1 hours
  - Perhaps easy to conclude the phone was properly designed

Copyright © 2021 Halim Noor, USM

Hypothesis Testing

- If the battery lifetime across 10 units is 95 hours? 97? 99? 102? … ?
- At which point would you start rejecting the assumption that the phone battery lifetime is 100 hours?
  - Use intuition
  - Hypothesis testing – size of the sample, variability in the sample and level of significance in conclusion

Copyright © 2021 Halim Noor, USM

t-Test

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

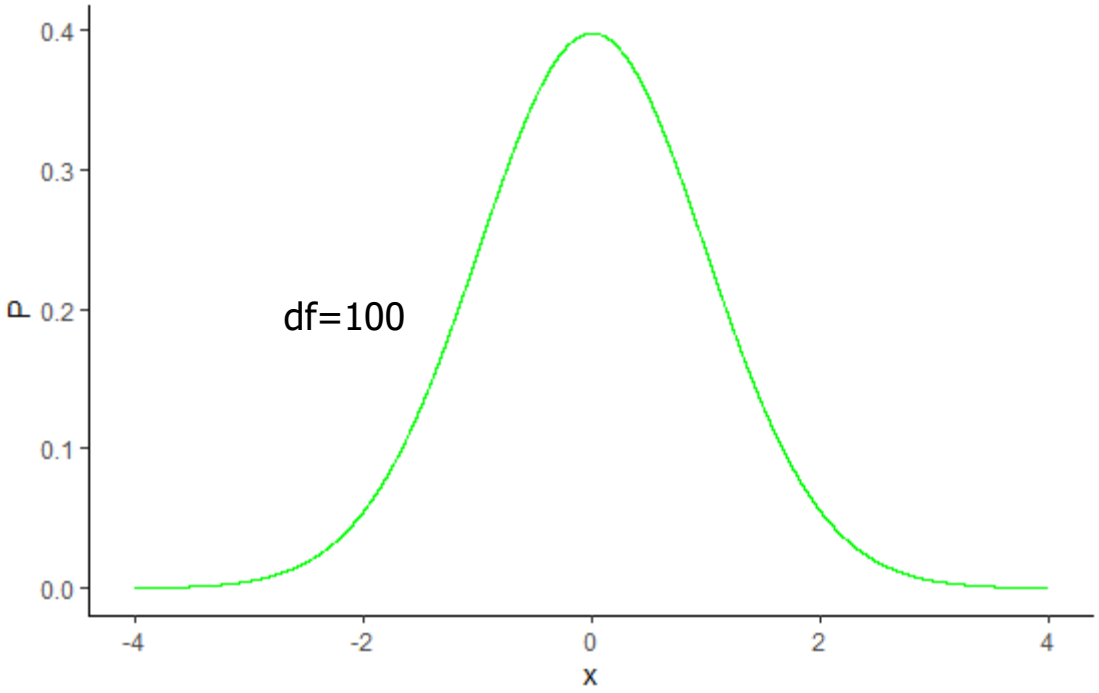
- Determine if the population mean is significantly different from the hypothesize mean
- t-value is calculated as follows

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

Copyright © 2021 Halim Noor, USM

t distribution

- A distribution that look almost identical to the normal distribution
- It is parameterized by degree of freedom (df)



Copyright © 2021 Halim Noor, USM

t distribution

- The df is the number of observations in the data that are free to vary when estimating statistical parameters

$$df = n - 1$$

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- A mobile phone was designed and manufactured to have a battery lifetime of 100 hours.
- As a production manager, you need to test that indeed the battery lifetime is 100 hours
- To do so, you randomly take 10 units and measure the battery lifetime

population mean, μ

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 100$ $n = 10$ $\bar{x} = 99$ $s = 2$

- Step 1: Formulate hypothesis

  Null Hypothesis H_0: $\mu = 100$

  Alternate Hypothesis H_1: $\mu \neq 100$

  Reject Null hypothesis if $\bar{x}$ is way above 100

  Reject Null hypothesis if $\bar{x}$ is way below 100

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 100 \qquad n = 10 \quad \bar{x} = 99 \quad s = 2$

- Step 2: Calculate t-value

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

$$t = \frac{99 - 100}{2/\sqrt{10}} = -1.581$$

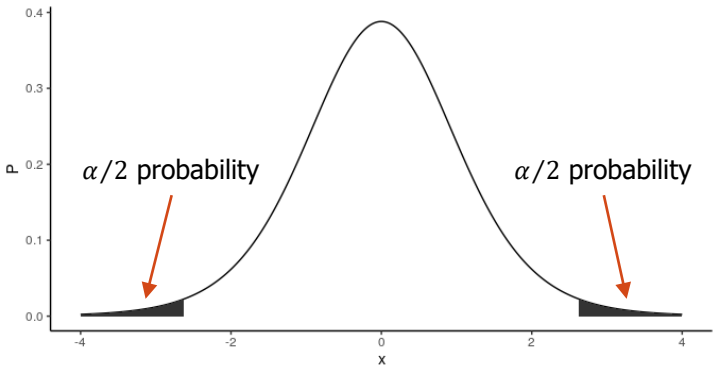
  - 99 way above 100 ≡ -1.581 way above 0
  - 99 way below 100 ≡ -1.581 way below 0

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 100$ $n = 10$ $\bar{x} = 99$ $s = 2$

- Step 3: Determine the Cut-off value for t-value
  - Specify the 'significance' level, $\alpha = 0.05$ (typical value)



Copyright © 2021 Halim Noor, USM

One-Sample t-Test

t Table

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										

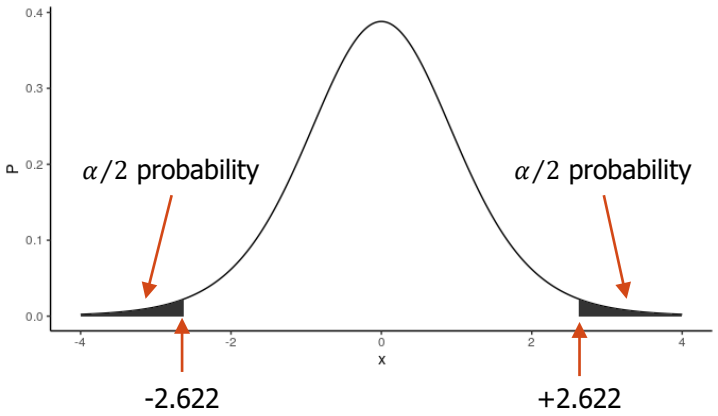
The value is between 2.262 and 2.821 (closer to 2.821)

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 100 \qquad n = 10 \quad \bar{x} = 99 \quad s = 2$

- Step 3: Determine the Cut-off value for t-value
  - Specify the 'significance' level, $\alpha = 0.05$ (typical value)



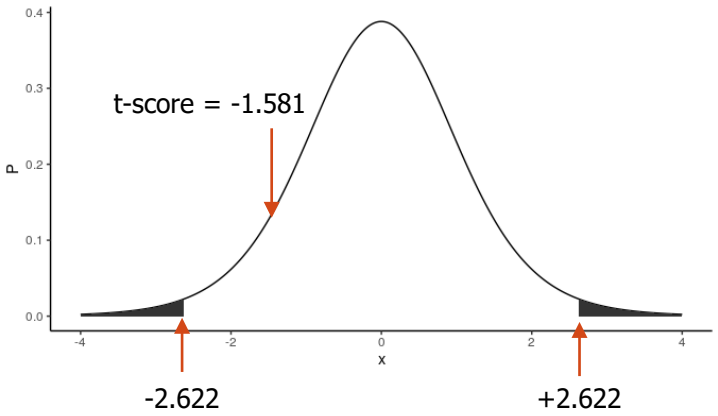

```
qt(0.05/2, df=9) = -2.622
qt(0.05/2, df=9, lower.tail=TRUE) = 2.622
```

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 100 \qquad n = 10 \quad \bar{x} = 99 \quad s = 2$

- Step 4: Check whether t-value falls in the rejection region

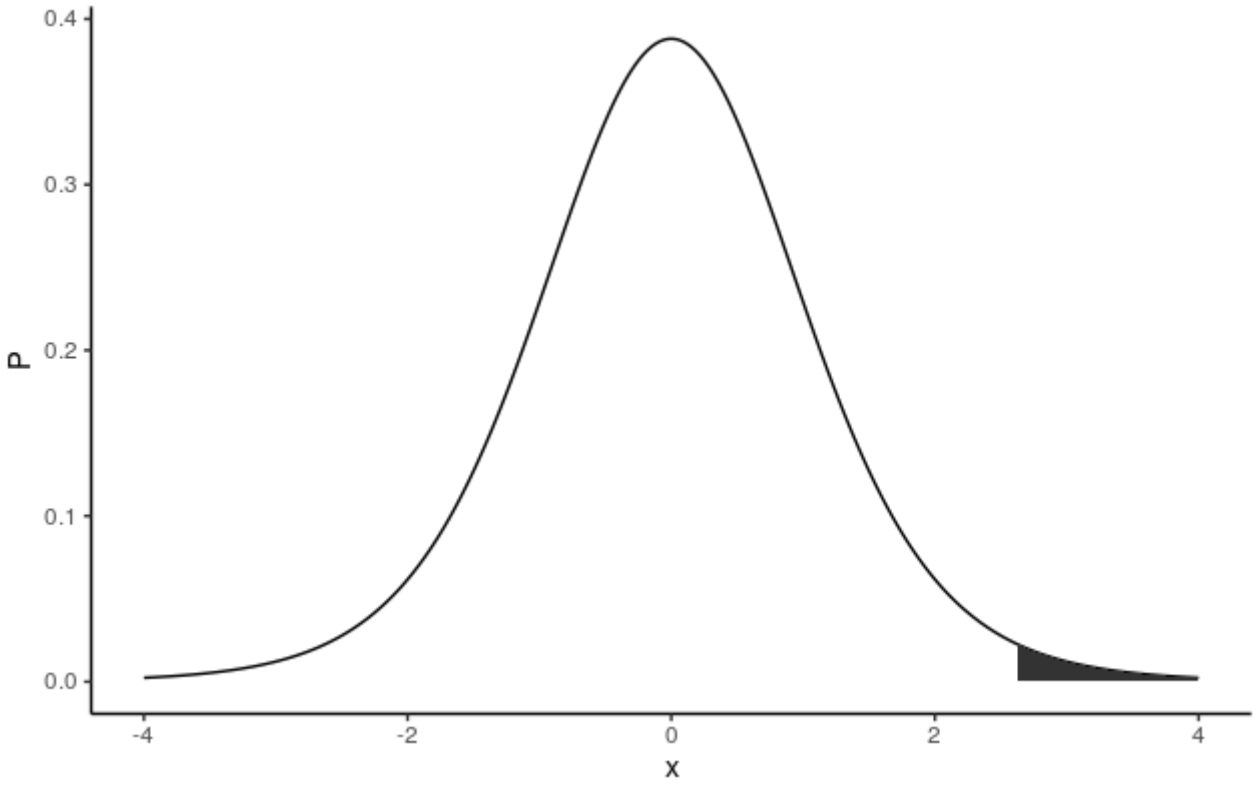


Conclusion:

- Since the t-score does not fall in the rejection region, we do not reject the Null hypothesis
- Our evidence is consistent with the claim made

Copyright © 2021 Halim Noor, USM

Single Tail Hypothesis Test



The entire rejection region on a single tail, either left tail or right tail

Copyright © 2021 Halim Noor, USM

Single Tail Hypothesis Test

- A new mobile phone design has on average an increase of 10 hours of battery lifetime
  - Increase in battery lifetime is 10 hours or more
- To test the claim
  - Random selection of 50 mobile phones
  - Measure the new phone's battery lifetime
  - Compare with the current phone to obtain the increase in battery lifetime
- Sample mean, $\bar{x} = 9.8$ hours
- Sample standard deviation, $s = 1.6$ hours

Copyright © 2021 Halim Noor, USM

Single Tail Hypothesis Test

- $\mu = 10$ $n = 50$ $\bar{x} = 9.8$ $s = 1.6$

- Step 1: Formulate hypothesis

Null Hypothesis	H_0:	$\mu \geq 10$
Alternate Hypothesis	H_1:	$\mu < 10$

Reject Null hypothesis

if $\bar{x}$ is way below 10 $\equiv$ if t-score is way below 0

Copyright © 2021 Halim Noor, USM

Single Tail Hypothesis Test

- Single tail hypothesis test with rejection on the left hand side

$$H_0: \quad \mu \geq \cdots$$

$$H_1: \quad \mu < \cdots$$

- Single tail hypothesis test with rejection on the right hand side

$$H_0: \quad \mu \leq \cdots$$

$$H_1: \quad \mu > \cdots$$

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 10$ $n = 50$ $\bar{x} = 9.8$ $s = 1.6$

- Step 2: Calculate t-value

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

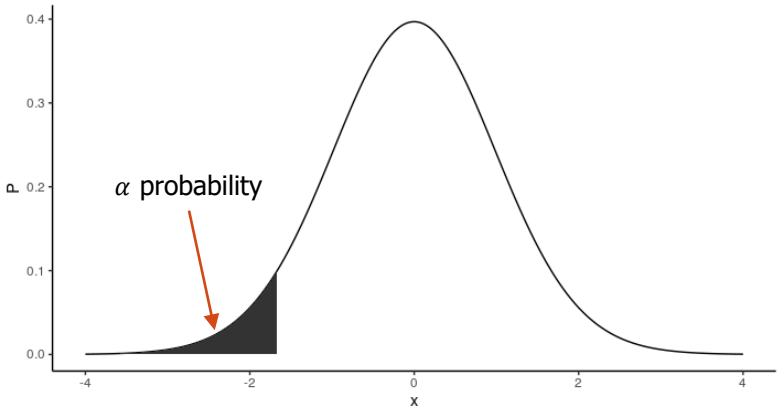
$$t = \frac{9.8 - 10}{1.6/\sqrt{50}} = -0.884$$

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 10$ $n = 50$ $\bar{x} = 9.8$ $s = 1.6$

- Step 3: Determine the Cut-off value for t-value
  - 'significance' level, $\alpha = 0.05$



Copyright © 2021 Halim Noor, USM

TABLE C *t* distribution critical values

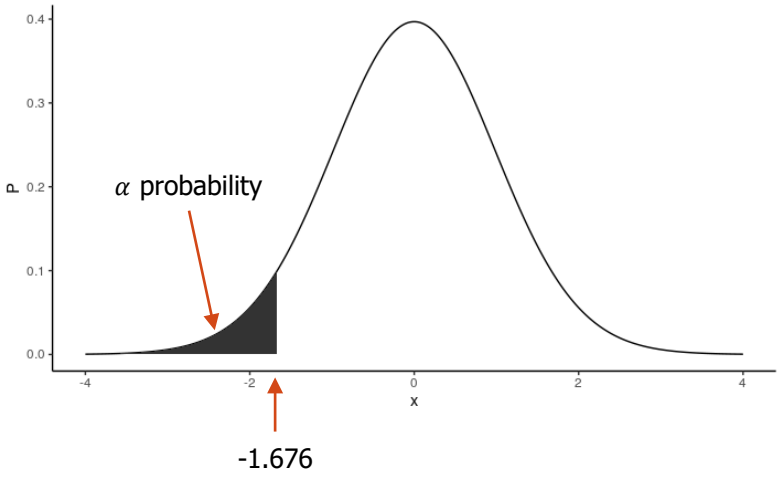
Degrees of freedom	Confidence level C											
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.710	15.890	31.820	63.660	127.300	318.300	636.600
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.090	22.330	31.600
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.210	12.920
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
One-sided *P*	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
Two-sided *P*	.50	.40	.30	.20	.10	.05	.04	.02	.01	.005	.002	.001

The value is between -1.684 and -1.676 (closer to -1.676)

One-Sample t-Test

- $\mu = 10$ $n = 50$ $\bar{x} = 9.8$ $s = 1.6$

- Step 3: Determine the Cut-off value for t-value
  - 'significance' level, $\alpha = 0.05$



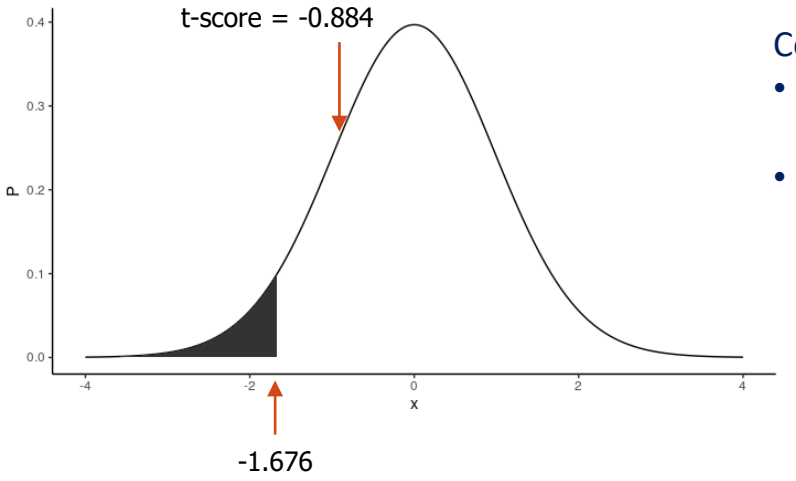

```
qt(0.05, df=49) = -1.676
```

Copyright © 2021 Halim Noor, USM

One-Sample t-Test

- $\mu = 10$ $n = 50$ $\bar{x} = 9.8$ $s = 1.6$

- Step 4: Check whether t-value falls in the rejection region



Conclusion:

- Since the t-score does not fall in the rejection region, we do not reject the Null hypothesis
- Our evidence is consistent with the claim made

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- Claim to be tested:

  Difference between the population mean mathematics score of 482 male students and 518 female students is 7.5 marks.

$\bar{x}_A = 68.728$ (male)

$\bar{x}_B = 63.633$ (female)

$s_A^2 = 206.102$ (male)

$s_B^2 = 239.985$ (female)

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- Step 1: Formulate the hypothesis

$$H_0: \quad \mu_{male} - \mu_{female} = 7.5$$

$$H_1: \quad \mu_{male} - \mu_{female} \neq 7.5$$

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- Step 2: Calculate the t-value

  Variance in the two populations

  Is the variance is more similar or more dissimilar

  Assuming Either equal variance or unequal variance

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- t-value between random sample A and B having similar variance is calculated as follows

$$t = \frac{\bar{x}_A - \bar{x}_B - \mu}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}}$$

where: $\bar{x}_A$ and $\bar{x}_B$ are the mean of sample A and B

μ is the hypothesized difference in population means

n_A and n_B are the sizes of sample A and B

s_p^2 is the pooled variance of the two samples

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

$n_A + n_B - 2$ is the degree of freedom

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- t-value between random sample A and B having different variance is calculated as follows

$$t = \frac{\bar{x}_A - \bar{x}_B - \mu}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

where: $\bar{x}_A$ and $\bar{x}_B$ are the mean of sample A and B

μ is the hypothesized difference in population means

n_A anood n_B are the sizes of sample A and B

s_A^2 and s_B^2 are the variance of sample A and B

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- Step 2: Calculate the t-value

$$t = \frac{\bar{x}_A - \bar{x}_B - \mu}{\sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}}} \text{ where } s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

$$t = -2.541$$

$$\bar{x}_A = 68.728$$

$$\bar{x}_B = 63.633$$

$$s_A^2 = 206.102$$

$$s_B^2 = 239.985$$

Copyright © 2021 Halim Noor, USM

Two-Sample Independent t-Test

- Step 3: Determine the Cut-off value for t-value
  - 'significance' level, $\alpha = 0.05$

$$t_{cut-off} = -1.962$$

```
qt(alpha/2, df=998)
```

- Step 4: Check whether t-value falls in the rejection region

  t-value falls in the rejection region, hence we reject the null hypothesis

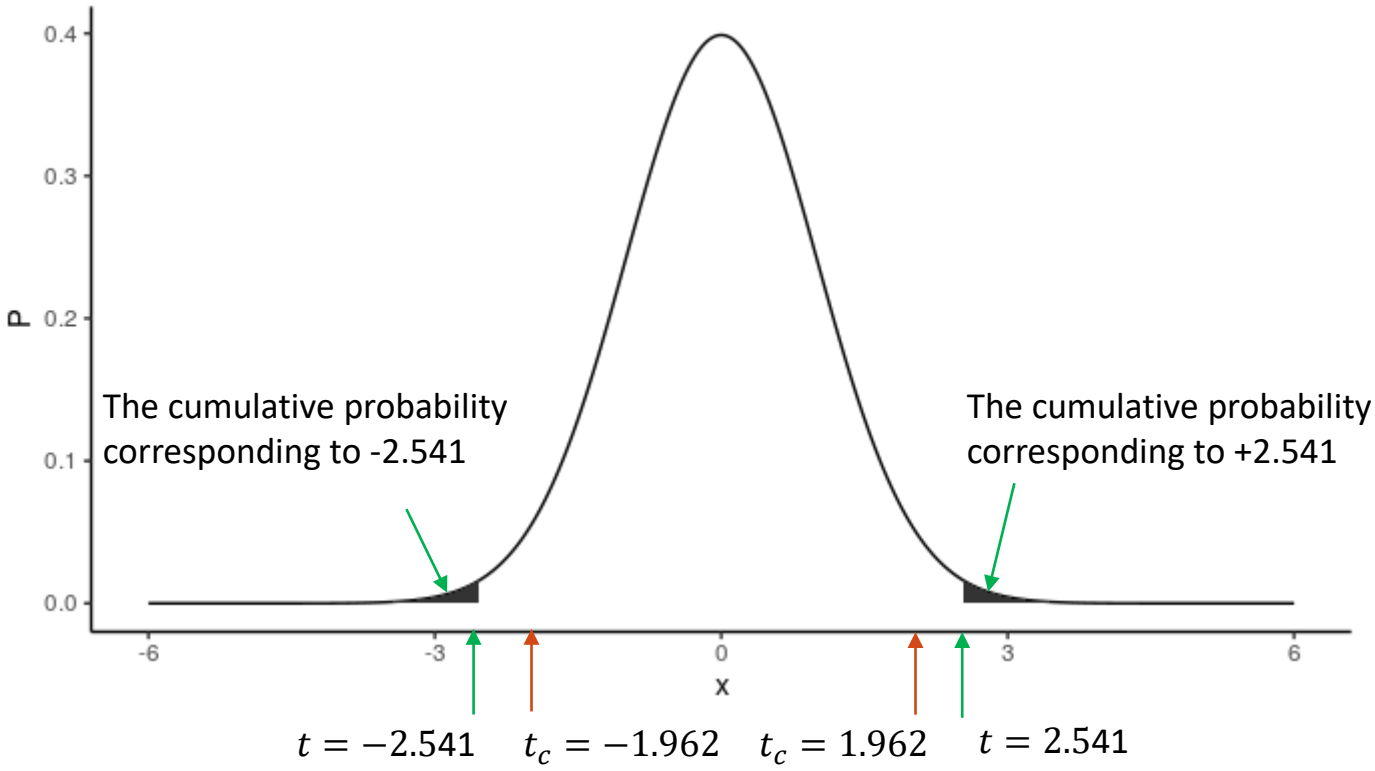
Copyright © 2021 Halim Noor, USM

The Notion of "p-value"

- -2.541 is below than the cut-off value (reject)
- -5.172 is way below than the cut-off value (strongly reject)
- To quantify this notion, the p-value of the hypothesis test is calculated

Copyright © 2021 Halim Noor, USM

The Notion of "p-value"



Copyright © 2021 Halim Noor, USM

The Notion of "p-value"

- Compare the p-value of the test to the significance level 0.05
- If p-value > 0.05: fail to (cannot) reject the null hypothesis
- If p-value ≤ 0.05: reject the null hypothesis

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

- Difference in means test for two random samples that are related
  - Measurement that are repeated under the same condition or by the same entity
  - Test scores that are performed by the same person before and after the training
  - Car test scores of different manufacturers but under the same test condition

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

- t-value between two dependent random sample A and B is calculated as follows

$$t = \frac{\mu_D - \mu}{s_D / \sqrt{n}}$$

where: μ_D is the mean difference between the paired values

μ is the hypothesized difference

s_D is the standard deviation difference between the paired values

n is the size of the paired values

degree of freedom is $n - 1$

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

- Claimed to be tested:

  Difference between the test scores before and after the computer skills training is greater or equal to 10 marks

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

Before	After
80.1	92.9
84.9	93.2
62.7	75.1
73.1	83.7
43.4	54.1
76.9	87.9
70.2	82.2
81.5	83.9
65.2	82.3
61.7	72.2

$\mu_D = 10.78$
$\mu = 10$
$s_D = 3.721$
$n = 10$

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

- Step 1: Formulate the hypothesis

$$H_0: \quad \mu_{aft} - \mu_{bef} \geq 10$$

$$H_1: \quad \mu_{aft} - \mu_{bef} < 10$$

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

- Step 2: Calculate the t-value

$$t = \frac{\mu_D - \mu}{s_D/\sqrt{n}}$$

$$t = 0.663$$

Copyright © 2021 Halim Noor, USM

Two-Sample Paired t-Test

- Step 3: Determine the Cut-off value for t-value
  - 'significance' level, $\alpha = 0.05$

$$t_{cut-off} = -1.833$$

- Step 4: Check whether t-value falls in the rejection region

  t-value does not fall in the rejection region, hence we fail to reject the null hypothesis

  p-value = 0.7379 `pt(t_val, df=9)`

Copyright © 2021 Halim Noor, USM

Analysis of Variance (ANOVA)

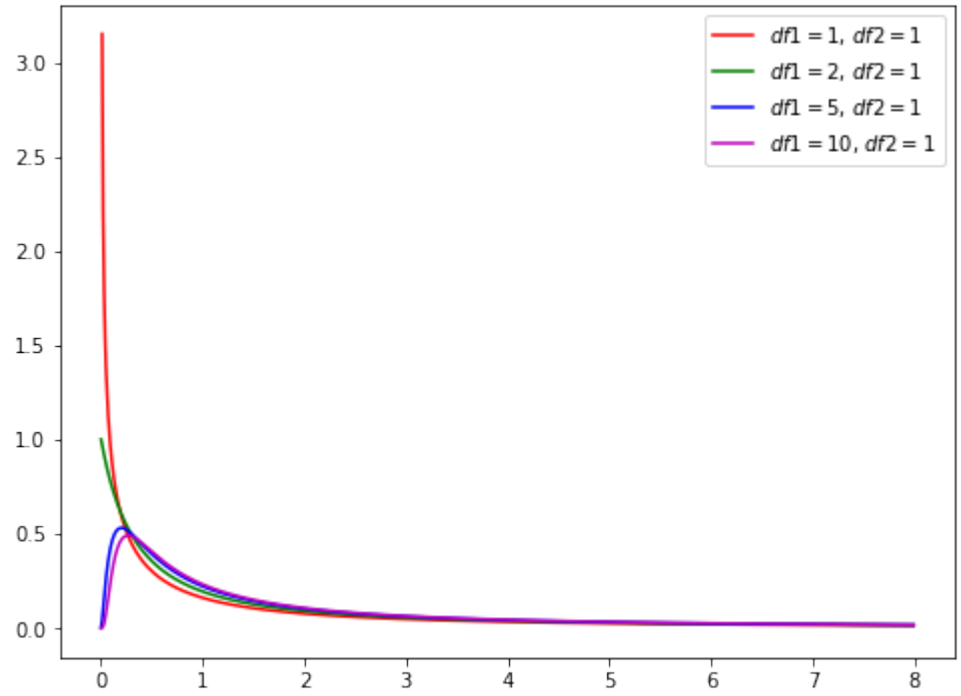
Copyright © 2021 Halim Noor, USM

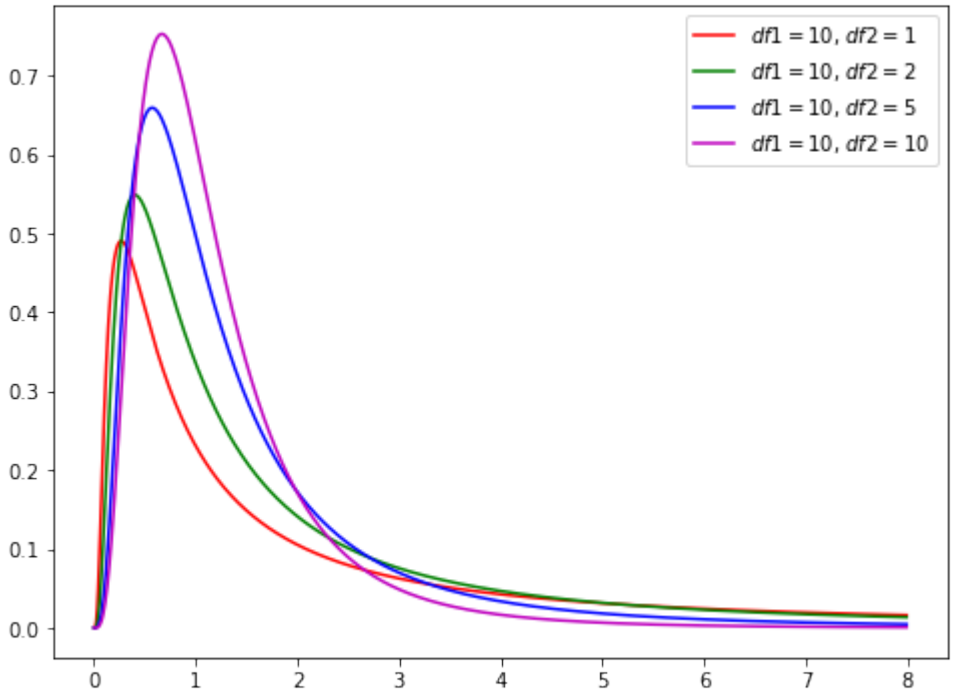
One-way ANOVA

- Extension of the t-test
- To determine whether the means of two or more groups are significantly different from each other
- Analysing the variation in the data
  - Compare the amount of variation between groups with the amount of variation within groups

Copyright © 2021 Halim Noor, USM

F-distribution





Parameterized by df1 and df2

Copyright © 2021 Halim Noor, USM

ANOVA

- Null hypothesis is the means of the different groups are the same
- Alternative hypothesis is at least one group mean is not equal to the others

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1: \quad \mu_1 \neq \mu_2 = \cdots = \mu_k \text{ or}$$

$$\mu_1 = \mu_2 \neq \cdots = \mu_k \text{ or}$$

$$\mu_1 = \mu_2 = \cdots \neq \mu_k$$

Copyright © 2021 Halim Noor, USM

ANOVA

- F-value between two groups is calculated as follows

$$F = \frac{MS_b}{MS_w}$$

MS_b: between group mean square

MS_w: within group mean square

Copyright © 2021 Halim Noor, USM

ANOVA

$$SS_b = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2 \quad MS_b = \frac{SS_b}{k-1}$$

$$SS_w = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_j^i - \bar{x}_j\right)^2 \quad MS_w = \frac{SS_w}{n_1 + n_2 + \cdots + n_k - k}$$

$\bar{x}_j$: mean of group j, $\bar{x}$: global mean

n_j: number of observations in group j

k: number of groups

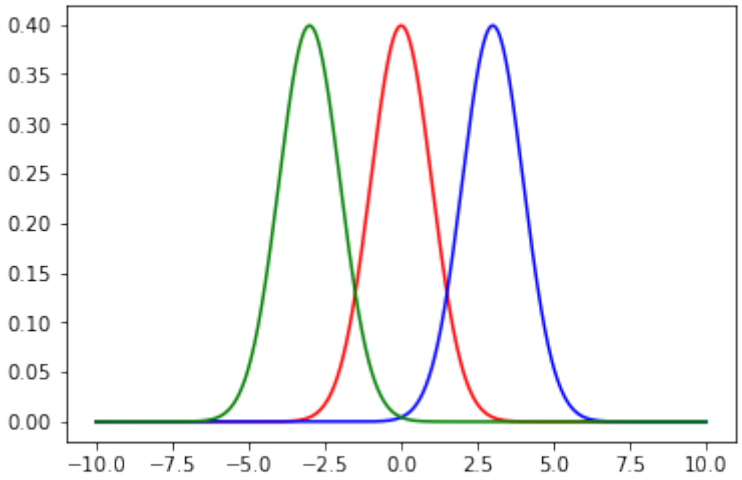
SS_b: between group sum of squared

SS_w: within group sum of squared

Copyright © 2021 Halim Noor, USM

Individual Mean and Grand Mean

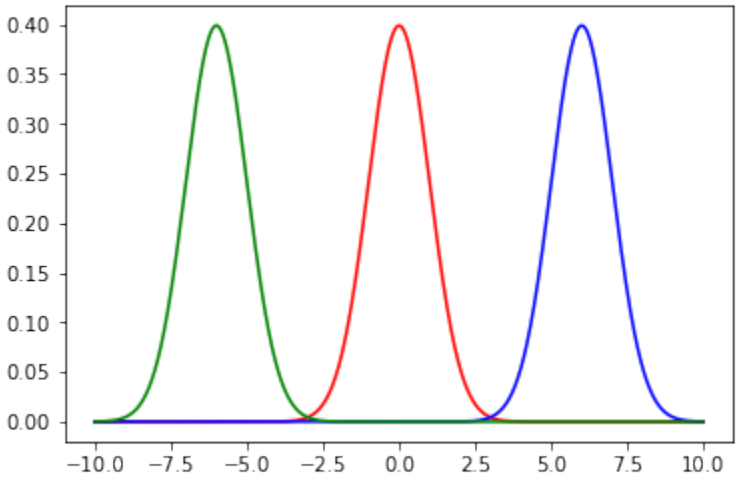
- Individual mean – mean of each group
- Grand mean – mean of all observations



Copyright © 2021 Halim Noor, USM

Between-Group Sum of Squared

- The deviation of group means from the grand mean

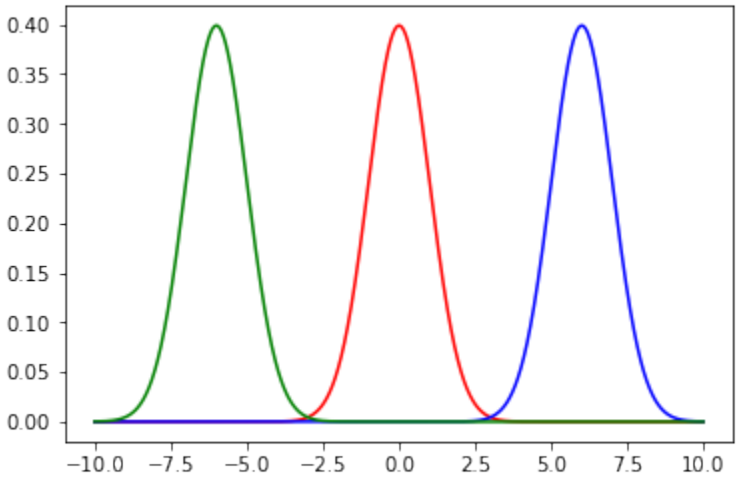


If SS_b is large, F-value will be large indicating significant difference between the groups' means

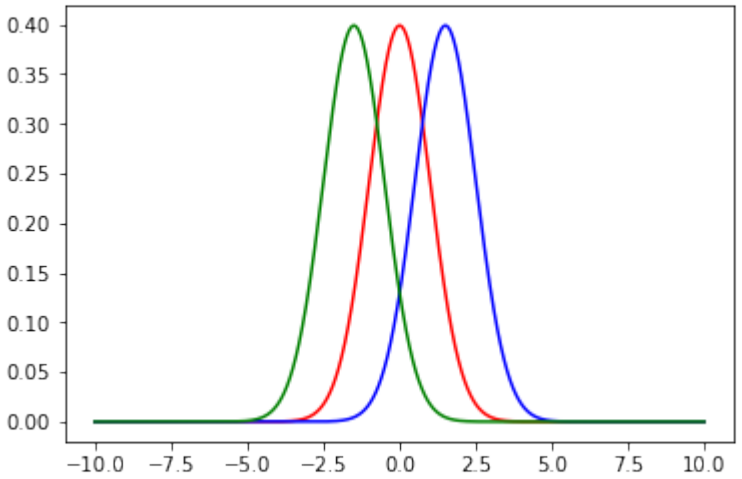
Copyright © 2021 Halim Noor, USM

Between-Group Sum of Squared

- The deviation of group means from the grand mean



If SS_b is large, F-value will be large indicating significant difference between the groups' means

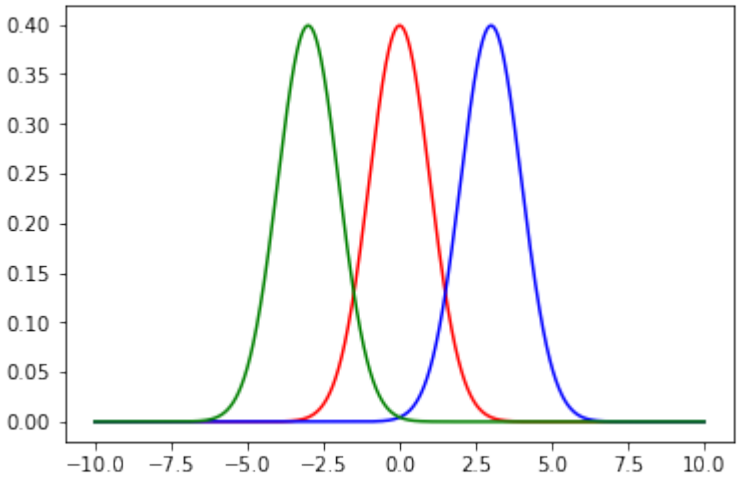


If SS_b is small, F-value will be small indicating no significant difference between the groups' means

Copyright © 2021 Halim Noor, USM

Within-Group Sum of Squared

- The variation of observations within a group

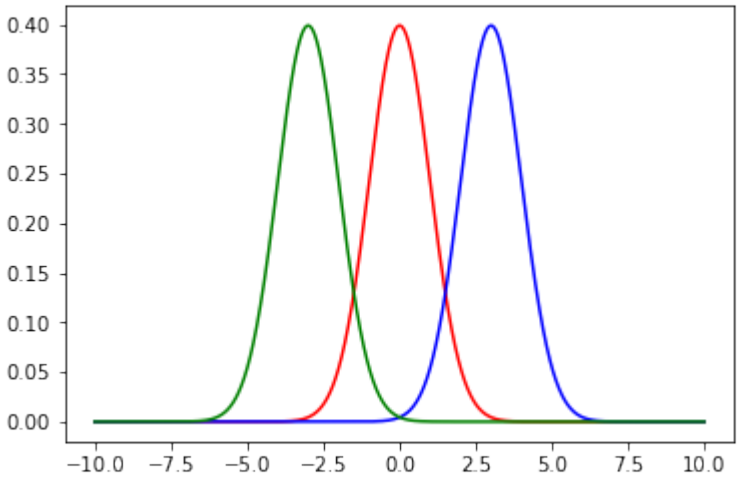


If SS_w is small, the groups seems to belong to different populations

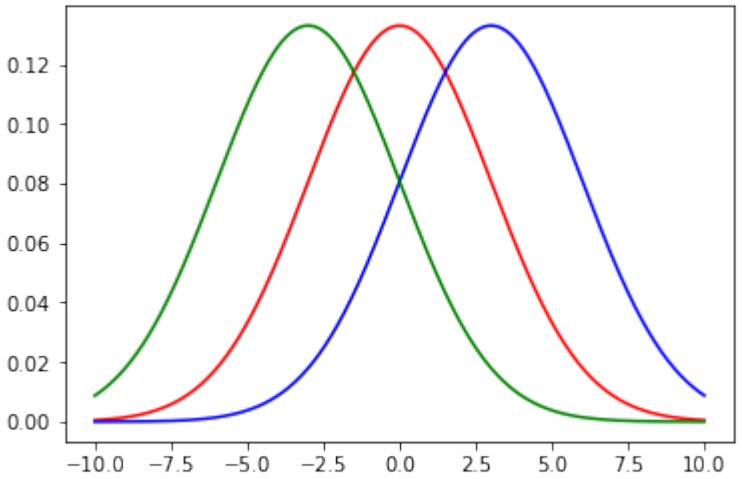
Copyright © 2021 Halim Noor, USM

Within-Group Sum of Squared

- The variation of samples within a group



If SS_w is small, the groups seems to belong to different populations



If SS_w is large (spread out), the groups overlap, and they become part of a big population

Copyright © 2021 Halim Noor, USM

ANOVA

- Given three batches of students, compare the scores using ANOVA

H_0: $\mu_{batch1} = \mu_{batch2} = \mu_{batch3}$

H_1: $\mu_{batch1} \neq \mu_{batch2} = \mu_{batch3}$ or

$\mu_{batch1} = \mu_{batch2} \neq \mu_{batch3}$ or

$\mu_{batch1} \neq \mu_{batch2} \neq \mu_{batch3}$

Copyright © 2021 Halim Noor, USM

ANOVA

Batch 1	Batch 2	Batch 3
47	55	54
76	69	63
40	59	82
64	74	52
58	82	70
40	81	51
78	80	82
88	35	57
46	60	47
66	87	59

Copyright © 2021 Halim Noor, USM

ANOVA

- Step 2: Calculate the F-value

$n_1 = n_2 = n_3 = 10$ (number of samples)

$k = 3$ (number of groups)

$\bar{x}_1 = 60.3$ $\bar{x}_2 = 68.2$ $\bar{x}_3 = 61.7$

$\bar{x} = 63.4$

Copyright © 2021 Halim Noor, USM

ANOVA

Between group mean square

$$df = k - 1 = 2$$

$$SS_b = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2 = 355.4$$

$$MS_b = \frac{SS_b}{k-1} = 177.7$$

Within group mean square

$$df = n_1 + n_2 + n_3 - k = 27$$

$$SS_w = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_j^i - \bar{x}_j\right)^2 = 6301.8$$

$$MS_w = \frac{SS_w}{n_1+n_2+n_3-k} = 233.4$$

F-value is $F = \frac{MS_b}{MS_w} = 0.7614$

Copyright © 2021 Halim Noor, USM

ANOVA

- Step 3: Determine the Cut-off value for F-value
  - 'significance' level, $\alpha = 0.05$

$$t_{cut-off} = 3.354$$

- Step 4: Check whether f-value falls in the rejection region

  t-value does not fall in the rejection region, hence we fail to reject the null hypothesis

  p-value = 0.4768 `qf(alpha, df1=2, df2=27, lower.tail = FALSE)`

Copyright © 2021 Halim Noor, USM

Chi-Square Test of Independence

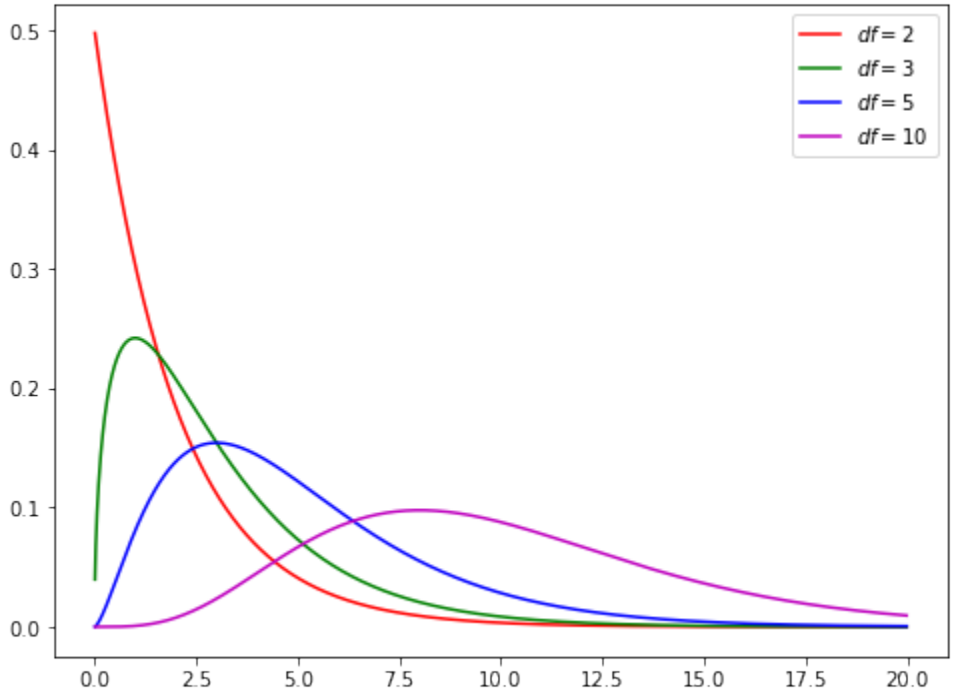
Copyright © 2021 Halim Noor, USM

Chi-Square Test

- Evaluates relationship between categorical variables
- Null hypothesis: The variables are independent
  - knowing one variable does not help to predict the other variable
- Alternative hypothesis: There variables are dependent
  - knowing one variable does help to predict the other variable

Copyright © 2021 Halim Noor, USM

Chi-Square Distribution



Copyright © 2021 Halim Noor, USM

Chi-Square Test

- χ^2-value of two categorical variables is calculated as follows

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

R is number of row, C is number of columns and E_{ij} is defined as

$$E_{ij} = \frac{row.sum_i col.sum_j}{N}$$

$$df = (R-1)(C-1)$$

Copyright © 2021 Halim Noor, USM

Chi-Square Test

- A survey is asking 563 individuals of their book type preference. Determine the relationship between the two variables.

Book type/genre	Comedy	Novel
Male	149	96
Female	70	248

Copyright © 2021 Halim Noor, USM

Chi-Square Test

- Step 2: Calculate the χ^2-value

Book type/genre	Comedy	Novel	
Male	149	96	245
Female	70	248	318
	219	344	563

Calculate the summation of rows and columns

Copyright © 2021 Halim Noor, USM

Chi-Square Test

- Step 2: Calculate the χ^2-value

Book type/genre	Comedy	Novel	
Male	$\frac{245 \times 219}{563} = 95.3$	$\frac{245 \times 344}{563} = 149.7$	245
Female	$\frac{318 \times 219}{563} = 123.7$	$\frac{318 \times 344}{563} = 194.3$	318
	219	344	563

Calculate the expected value for each entry:

Copyright © 2021 Halim Noor, USM

Chi-Square Test

- Step 2: Calculate the χ^2-value

Calculate the χ^2-value:

Book type/genre	Comedy	Novel	
Male	$\frac{(149-95.3)^2}{95.3}$	$\frac{(96-149.7)^2}{149.7}$	245
Female	$\frac{(70-123.7)^2}{123.7}$	$\frac{(248-194.3)^2}{194.3}$	318
	219	344	563

Book type/genre	Comedy	Novel	
Male	30.25	19.26	245
Female	23.31	14.84	318
	219	344	563

Chi-Square Test

- Step 2: Calculate the χ^2-value

$$\chi^2 = 30.25 + 19.26 + 23.31 + 14.84 = 87.66$$

- Step 3: Calculate the cut-off value ($\alpha = 0.05$)

$$\chi^2_{cut-off} = 3.84$$

- Step 4: Check whether t-value falls in the rejection region

  χ^2-value falls in the rejection region, hence we reject the null hypothesis

  Variables are dependent

Copyright © 2021 Halim Noor, USM

Chi-Square Test

- If you have one categorical variable from a single population, and you would like to determine whether the sample is consistent with a hypothesized distribution
- Link to video on eLearn.

Copyright © 2021 Halim Noor, USM

End

Copyright © 2021 Halim Noor, USM