Statistical and Mathematical Methods for Data Analysis

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics**, 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman

- **Probability Demystified**, Allan G. Bluman

- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce

- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson

- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

Readings for these lecture notes:

- Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer

- Elementary Statistics, Tenth Edition, Mario F. Triola

These notes contain material from the above resources.

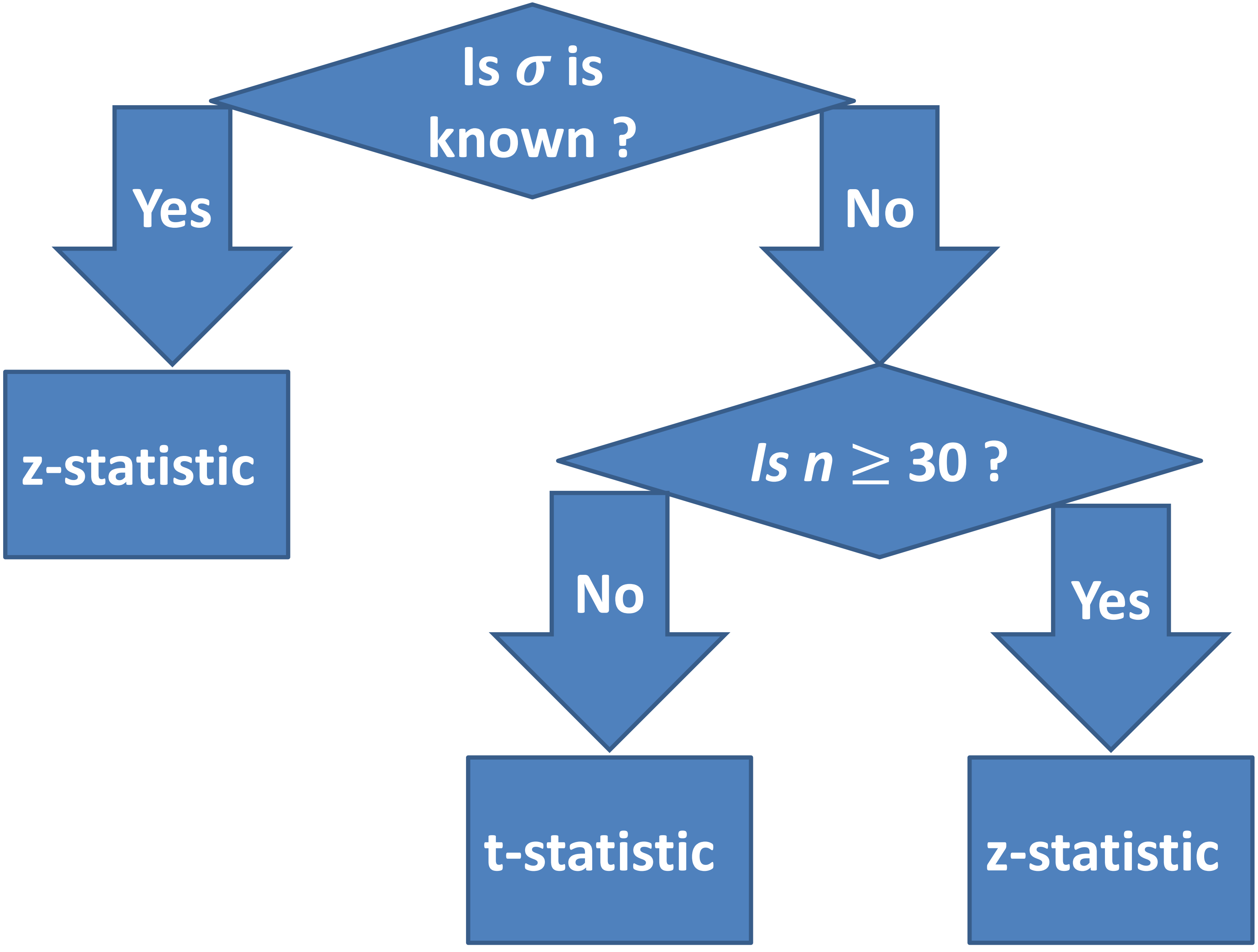
Is σ is known ?

Yes

If either the population is normally distributed or $n \geq 30$, then use the use the standard normal distribution or Z-test

No

If either the population is normally distributed or $n \geq 30$, then use the *t*-distribution or t-test


Is σ is known ?
Yes
No
z-statistic
Is $n \geq 30$?
No
Yes
t-statistic
z-statistic

The Case of σ Unknown [4]

$$\sum(x-\overline{x})^2 = \sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n} = \frac{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2}{n}$$

or $s^2 = \frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2\}$

where $t_{\alpha/2}$ is the t- value with $n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

The Case of σ Unknown [1]

Frequently, we must attempt to estimate the mean of a population when the **variance is unknown**. If we have a random sample from a normal distribution, then the random variable

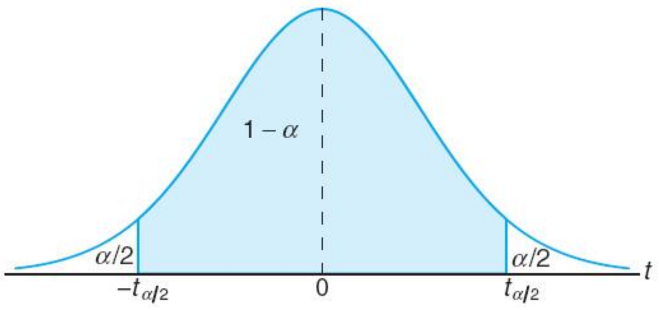
$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

has a **Student t-distribution** with **n – 1 degrees of freedom**. Here **s** is the sample standard deviation. In this situation, with **σ unknown**, T can be used to construct a confidence interval on μ.

The Case of σ Unknown [2]

$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$, where $T = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$

$$\Longrightarrow P\left(-t_{\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha$$



$$\Longrightarrow P\left(-\,t_{\alpha/2}\frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Longrightarrow P\left(-\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < -\mu < -\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Longrightarrow P\left(\bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} > \mu > \bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Longrightarrow P\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

The Case of σ Unknown [3]

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} - t_{(\alpha/2,n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\alpha/2,\, n-1)} \frac{s}{\sqrt{n}}$$

OR

$$\text{C.I} = \bar{x} \pm t_{(\alpha/2,n-1)} \frac{s}{\sqrt{n}}$$

where $s^2 = \dfrac{\sum (x-\bar{x})^2}{n-1}$

The Case of σ Unknown [4]

$$\sum(x-\bar{x})^2 = \sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)^2}{n} = \frac{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2}{n}$$

or $s^2 = \frac{1}{n(n-1)}\{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2\}$

where $t_{\alpha/2}$ is the t- value with $n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

The Case of σ Unknown [5]

We have made a distinction between the cases of **σ known** and **σ unknown** in computing confidence interval estimates. We should emphasize that for **σ known** we exploited the **Central Limit Theorem**, whereas for **σ unknown** we made use of the **sampling distribution** of the **random variable *T***.

However, the use of the *t* distribution is based on the premise that the **sampling** is from a **normal distribution**. As long as the distribution is approximately bell shaped, confidence intervals can be computed when σ^2 is unknown by using the *t*-distribution and we may expect very good results.

One-Sided Confidence Bounds on μ, σ^2 unknown [1]

If $\overline{X}$ is the mean of a random sample of size n from a population with unknown variance σ^2, the one-sided 100(1 − α)% confidence bounds for μ are given by

upper one-sided bound: $\overline{\mathbf{x}} + t_{(\alpha,\, n-1)} \dfrac{\mathbf{s}}{\sqrt{\mathbf{n}}}$

lower one-sided bound: $\overline{\mathbf{x}} - t_{(\alpha,\, n-1)} \dfrac{\mathbf{s}}{\sqrt{\mathbf{n}}}$

Critical Values of the t-Distribution

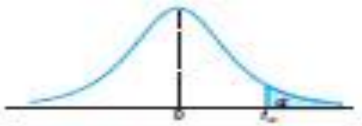


Table A.4 Critical Values of the t-Distribution

	α						
v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Critical Values of the t-Distribution

	α						
v	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

Example: The contents of seven similar containers of sulfuric acid are **9.8, 10.2, 10.4, 9.8, 10.0, 10.2**, and **9.6** liters. Find a **95%** confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

x	$x - \bar{x}$	$(x - \bar{x})^2$
9.8	-0.2	0.04
10.2	0.2	0.04
10.4	0.4	0.16
9.8	-0.2	0.04
10.0	0	0
10.2	0.2	0.04
9.6	-0.4	0.16
$\sum x =$ 70		$\sum (x - \bar{x})^2$ = 0.4800

$\bar{x} = \frac{\sum x}{n}$

$= \frac{70}{7}$

$= 10$

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

$= .48 / 6$

$= 0.0800$

$\Rightarrow$ **s = 0.28**

v = n – 1 = 6 degrees of freedom

$\alpha = 0.05$

$\Rightarrow \alpha/2 = 0.05/2 = \mathbf{0.025}$

$t_{(0.025,\, 6)} = \mathbf{2.447}$

95% confidence interval for μ is

$$\bar{x} - t_{(\alpha/2,\, n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{(\alpha/2,\, n-1)} \frac{s}{\sqrt{n}}$$

$$\Rightarrow 10.0 - \frac{(2.447)(0.283)}{\sqrt{7}} < \mu < 10.0 + \frac{(2.447)(0.283)}{\sqrt{7}}$$

$$\Rightarrow 9.74 < \mu < 10.26.$$

Alternative approach to compute s^2

x	x^2
9.8	96.04
10.2	104.04
10.4	108.16
9.8	96.04
10.0	100
10.2	104.04
9.6	92.16
$\sum x$ = 70	700.48

$n = 7$

$\sum x = 70$

$\sum x^2 = 700.4800$

$$\mathbf{s^2 = \frac{1}{n(n-1)}\{n \sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x)^2\}}$$

$$s^2 = \frac{1}{7(7-1)}\{7(700.4800) - (70)^2\}$$

$$= 0.0800$$

$\Rightarrow$ **s = 0.2828**

95% confidence interval for μ is

$$\bar{\mathbf{x}} - \mathbf{t}_{(\alpha/2,\, n-1)} \frac{\mathbf{s}}{\sqrt{\mathbf{n}}} < \boldsymbol{\mu} < \bar{\mathbf{x}} + \mathbf{t}_{(\alpha/2,\, n-1)} \frac{\mathbf{s}}{\sqrt{\mathbf{n}}}$$

$$\Rightarrow 10.0 - \frac{(2.447)(0.283)}{\sqrt{7}} < \mu < 10.0 + \frac{(2.447)(0.283)}{\sqrt{7}}$$

$$\Rightarrow 9.74 < \mu < 10.26.$$

Single Sample: Estimating a Proportion [1]

Point estimate of the parameter p: A point estimator of the proportion p in a binomial experiment is given by the statistic $\widehat{P} = X/n$, where **X represents the number of successes in n trials**.

Therefore, the sample proportion $\widehat{p} = x/n$ will be used as the point estimate of the parameter p.

Confidence Intervals for Proportions p:

100(1 – α)% confidence interval for p is

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}} < p < \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}}$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

OR

$$C.I = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Single Sample: Estimating a Proportion [2]

Example: In a random sample of **n = 500** families owning television sets in the city of Hamilton, Canada, it is found that **x = 340** subscribe to **HBO**. Find a **95% confidence interval** for the actual proportion of families with television sets in this city that subscribe to HBO.

Solution: The point estimate of p is $\hat{p} = 340 / 500 = 0.68$. **$z_{0.025} = 1.96$**.

95% confidence interval for p is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.68 - 1.96\sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96\sqrt{\frac{(0.68)(0.32)}{500}}$$

$$\Rightarrow 0.6391 < p < 0.7209$$

Calculating sample size using $\hat{p}$

If $\hat{p}$ is used as an estimate of p, we can be **100(1 – α)%** confident that the **error** will be less than a specified **amount e** when the sample size is approximately

$$n = \frac{\hat{p}\hat{q}\, z^2_{\alpha/2}}{e^2}$$

Example: How large a sample is required if we want to be **95% confident** that our **estimate of p** in the previous example is within **0.02** of the true value?

Solution:

$$n = \frac{\hat{p}\hat{q}\ z^2{}_{\alpha/2}}{e^2}$$

$$n = \frac{(0.68)(0.32)(1.96)^2}{(0.02)^2}$$

$$n = 2089.8 \approx 2090$$

Calculating sample size without prior knowledge about $\widehat{p}$ [1]

If $\widehat{p}$ is used as an estimate of **p**, we can be **at least 100(1 – α)%** confident that the error will not exceed a specified amount e when the sample size is

$$n = \frac{z^2_{\alpha/2}}{4e^2}$$

Calculating sample size without prior knowledge about $\hat{p}$ [2]

Example: How large a sample is required if we want to be at least 95% confident that **our estimate of p** in the previous example within 0.02 of the true value?

Solution: We shall now assume that no preliminary sample has been taken to provide an estimate of p. Consequently, we can be at least 95% confident that our sample proportion will not differ from the true proportion by more than **0.02** if we choose a sample of size

$$n = \frac{(1.96)^2}{4(0.02)^2} = 2401$$

Two Samples: Estimating the Difference between Two Proportions

$100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$$

OR

$$\text{C.I} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$$

Example: A certain change in a process for manufacturing component parts is being considered. Samples are taken under both the existing and the new process so as to determine if the new process results in an improvement. If **75 of 1500** items from the existing process are found to be defective and **80 of 2000** items from the new process are found to be defective, find a **90% confidence interval** for the true difference in the proportion of defectives between the existing and the new process.

Solution

Let p_1 and p_2 be the true proportions of defectives for the existing and new processes, respectively.

$\hat{p}_1$ **= 75/1500 = 0.05** and $\hat{p}_2$ **= 80/2000 = 0.04**

The point estimate of $p_1 - p_2$ is

$\hat{p}_1 - \hat{p}_2 = 0.05 – 0.04 = 0.01$

$z_{0.05}$ **= 1.645**.

90 % C.I for $p_1 - p_2$ is

$$z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}} = 1.645\sqrt{\frac{(0.05)(0.95)}{1500}+\frac{(0.04)(0.96)}{2000}}$$

$$= 0.0117$$

$$\text{C.I} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Substitute values in the formula, we get

$-0.0017 < P_1 - P_2 < 0.0217$

Two Samples: Estimating the Difference between Two Means [1]

Confidence Interval for $\mu_1 - \mu_2$, when **σ_1^2 and σ_2^2 known**

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or

$$\text{C.I} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Two Samples: Estimating the Difference between Two Means [2]

Example: A study was conducted in which two types of engines, A and B, were compared. Gas mileage, in miles per gallon, was measured. **Fifty** experiments were conducted using engine type A and **75** experiments were done with engine type B. The gasoline used and other conditions were held constant. The average gas mileage was **36 miles** per gallon for engine A and **42 miles** per gallon for engine B. Find a **96% confidence interval** on $\mu_B - \mu_A$, where μ_A and μ_B are population mean gas mileages for engines A and B, respectively. Assume that the **population standard deviations** are **6** and **8** for engines A and B, respectively.

Solution: The point estimate of $\mu_B - \mu_A$ is $\bar{x}_1 - \bar{x}_2 = 42 - 36 = 6$. Using $\alpha = 0.04$, we find $z_{0.02} = 2.05$ from Table A.3. Hence, with substitution in the formula above, the 96% confidence interval is

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or

$$\textbf{C.I} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_1 - \mu_2 < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$3.43 < \mu_B - \mu_A < 8.57.$$

Two Samples: Estimating the Difference between Two Means [4]

or

C.I = (3.43, 8.57)

Two Samples: Estimating the Difference between Two Means [3]

Assumption: Population Variances Unknown but Equal ($\sigma_1^2 = \sigma_2^2$)

Pooled Estimate of Variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Confidence Interval for $\mu_1 - \mu_2$, σ_1^2 **and** σ_2^2 **unknown but equal**

$$(\bar{x}_1 - \bar{x}_2) - t_{(\alpha/2,\, n1 + n2 - 2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{(\alpha/2,\, n1 + n2 - 2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Assumption: Population Variances Unknown but Equal

$$\mathbf{C.I} = (\bar{x}_1 - \bar{x}_2) \pm t_{(\alpha/2,\ n1 + n2 - 2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example: The article "Macro invertebrate Community Structure as an Indicator of Acid Mine Pollution," published in the Journal of Environmental Pollution, reports on an investigation undertaken in Cane Creek, Alabama, to determine the relationship between selected physiochemical parameters and different measures of macro invertebrate community structure. One facet of the investigation was an evaluation of the effectiveness of a numerical species diversity index to indicate aquatic degradation due to acid mine drainage. Conceptually, a high index of macro invertebrate species diversity should indicate an unstressed aquatic system, while a low diversity index should indicate a stressed aquatic system.

Example (cont.)

Two independent sampling stations were chosen for this study, one located downstream from the acid mine discharge point and the other located upstream. For **12 monthly samples** collected at the downstream station, the species diversity index had a mean value $\bar{x}_1 = 3.11$ and a standard deviation $s_1 = 0.771$, while **10 monthly samples** collected at the upstream station had a mean index value $\bar{x}_2 = 2.04$ and a standard deviation $s_2 = 0.448$. Find a **90%** confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with **equal variances**

Solution:

Our point estimate of $\mu_1 - \mu_2$ is

$\overline{x_1} - \overline{x_2} = 3.11 - 2.04 = 1.07$

$s^2_p = \dfrac{(12-1)(0.771^2) + (10-1)(0.448^2)}{12+10-2} = 0.417$, $s_p = 0.646$

95 % for Confidence Interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - t_{(\alpha/2,\ n1+n2-2)}\ s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 <$$

$$(\bar{x}_1 - \bar{x}_2) + t_{(\alpha/2,\ n1+n2-2)}\ s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$t_{(\alpha/2,\ n1+n2-2)} = t_{(0.05,20)} = 1.725$

$0.593 < \mu_1 - \mu_2 < 1.547.$

Two Samples: Estimating the Difference between Two Means [3]

Assumption: Population Variances Unknown but Unequal ($\sigma_1^2 \neq \sigma_2^2$)

100(1- α)% for Confidence Interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - t_{(\alpha/2,\, v)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{(\alpha/2,\, v)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)]}$$

Two Samples: Estimating the Difference between Two Means [4]

Assumption: Population Variances Unknown but Unequal ($\sigma_1^2 \neq \sigma_2^2$)

$$\text{C.I} = (\bar{x}_1 - \bar{x}_2) \pm t_{(\alpha/2,\, v)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)]}$$

Unknown and Unequal Variances [2]

Example: A study was conducted by the Department of Zoology at the Virginia Tech to estimate the difference in the amounts of the chemical orthophosphorus measured at two different stations on the James River. Orthophosphorus was measured in milligrams per liter. **Fifteen samples** were collected from station 1, and **12 samples** were obtained from station 2. The 15 samples from station 1 had an **average** orthophosphorus content of **3.84** milligrams per liter and a standard deviation of **3.07** milligrams per liter, while the **12** samples from station 2 had an average content of **1.49** milligrams per liter and a standard deviation of **0.80** milligram per liter.

Find a **95% confidence interval** for the difference in the true average orthophosphorus contents at these two stations, assuming that the observations came from normal populations with **different variances**.

Unknown and Unequal Variances [3]

Given

For station 1: $\overline{x_1} = 3.84$, $s_1 = 3.07$, and $n_1 = 15$.

For station 2, $\overline{x_2} = 1.49$, $s_2 = 0.80$, and $n_2 = 12$.

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)]} = 16.3 \approx 16.$$

Our point estimate of $\mu_1 - \mu_2$ is $\overline{x_1} - \overline{x_2} = 3.84 - 1.49 = 2.35$.

Using $\alpha = 0.05$, $t_{(0.025,\ 16)} = 2.120$ for $v = 16$ degrees of freedom.

95% confidence interval for $\mu_1 - \mu_2$ is

$$0.60 < \mu_1 - \mu_2 < 4.10.$$