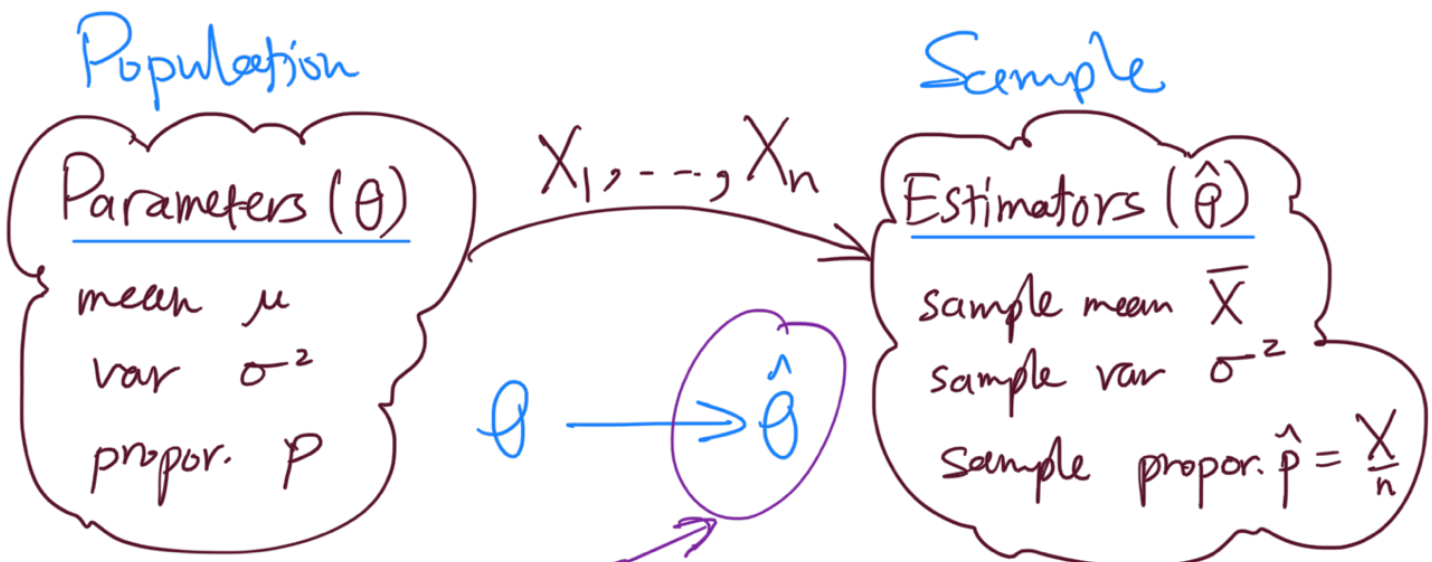
Concepts of statistical inference

- → Confidence Interval (Today!).
- → Tests of Hypothesis (Next time!).

Estimation of confidence interval.



$\hat{\theta}$ is called a point estimator for θ.

$\theta = \hat{\theta} \pm E$, where E is the error.

If the sample size $(n) \longrightarrow \infty$, then sample $\rightarrow$ population, and hence $\hat{\theta} \longrightarrow \theta$, no error!

In this case, $\bar{X} \rightarrow \mu$, $S^2 \rightarrow \sigma^2$ and $\hat{p} \rightarrow p$.

When $P\left(\theta \in (\hat{\theta}-E, \hat{\theta}+E)\right) = 1-\alpha$, $(0<\alpha<1)$, we call $(\hat{\theta}-E, \hat{\theta}+E)$ the $(1-\alpha)100\%$ confidence interval (C.I.) for the parameter θ.

<u>Def.</u> Let $X_1, X_2, \dots, X_n$ be a random sample and θ be an unknown population parameter. Let L and U be funcs of $X_1, X_2, \dots, X_n$. The interval (L, U) is called $(1-\alpha)100\%$ C.I. for θ if $P(L<\theta<U) = 1-\alpha$.

Here, $\alpha \in (0,1)$ is called the significant level.

$1-\alpha$ is called the confidence coefficient.

L is called the lower confidence limit (LCL).

U 〃 〃 〃 upper 〃 〃 (UCL).

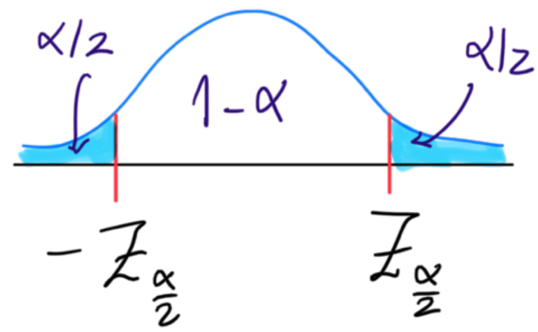
<u>Interval estimation of μ.</u>

<u>Remark</u>: If $X_1, X_2, \dots, X_n \overset{r.s.}{\sim} N(\mu, \sigma^2)$, then the $(1-\alpha)100\%$ C.I. for μ is

$$\boxed{\left(\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)}.$$

<u>Proof</u>: If $X_1, X_2, \ldots, X_n \overset{r.s.}{\sim} N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

or $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.



Now the $(1-\alpha)100\%$ C.I. for μ satisfies

$P(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}) = 1-\alpha$. Note that

$$-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}} \Leftrightarrow -Z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < Z_{\frac{\alpha}{2}}$$

$$\Leftrightarrow -Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow -\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

Thus, the $(1-\alpha)100\%$ C.I. for μ satisfies

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

This completes the proof.

For simplicity, (L, U) can be written as $\bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$.

$Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ is called the error of estimation.

$\frac{\sigma}{\sqrt{n}}$ is called the standard (or margin) error.

In general, we have the following fact.

Fact: The $(1-\alpha)100\%$ C.I. for μ is

$$\boxed{\begin{array}{ll} \text{i)} & \bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ if } \sigma \text{ is known.} \\ \text{ii)} & \bar{X} \pm T_{\frac{\alpha}{2}}(n-1) \frac{S}{\sqrt{n}} \text{ if } \sigma \text{ is unknown } (\& \ n < 30). \\ \text{iii)} & \bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \text{ if } \sigma \text{ is unknown but } n \geq 30. \end{array}}$$

The t-dist. with d.f. $(n-1)$.

Ex. The salaries of teachers in Jordan for 1990-2000 are normally distributed with standard deviation 50 JD. The average salary based on a sample of 400 teachers for 1990-2000 was 215 JD per month.

a) What is the point estimate of the mean 1990-2000 salaries and its S.E.?

b) Give a 90% C.I. for the mean 1990–2000 salaries.

c) 〃 〃 95% 〃 〃 〃 〃 〃 – 〃 〃.

<u>Soln.</u> $X_1, X_2, \ldots, X_{400} \overset{r.s.}{\sim} N(\mu, 50^2)$. And $\bar{X} = 215$.

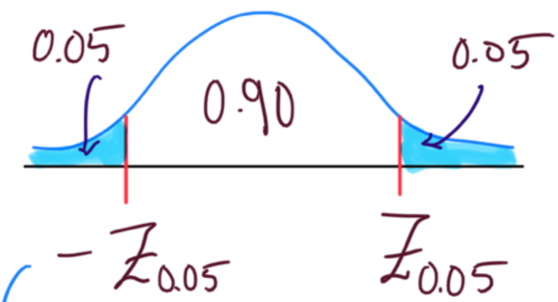
a) Point estimate for μ is $\bar{X} = 215$.

$$S.E. = \sigma/\sqrt{n} = 50/\sqrt{400} = 50/20 = 2.5.$$

b) $1 - \alpha = 0.90$.

So, $\alpha = 0.10$,

and $\alpha/2 = 0.05$.



$= -1.64$ (from tables).

So, $Z_{0.05} = 1.64$.

The 90% C.I. for μ is

$$\bar{X} \pm Z_{0.05} \cdot \frac{\sigma}{\sqrt{n}} = 215 \pm (1.64)\left(\frac{50}{\sqrt{400}}\right).$$

$$L = 215 - (1.64)(2.5) = 210.90.$$

$$U = 215 + (1.64)(2.5) = 219.10.$$

$\therefore$ The 90% C.I. for μ is $(210.90, 219.10)$.

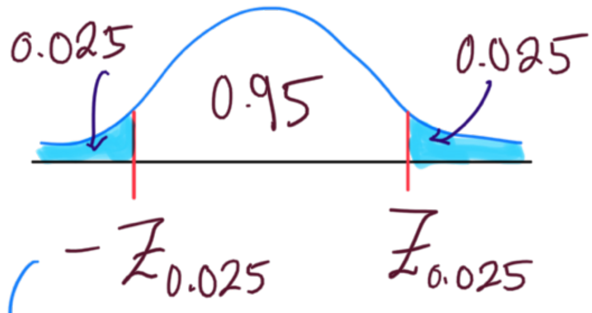
هذا يعني إننا ع ثقة مقدارها ٩٠٪

أنه الـ mean بين هذين الرقمين.

c) $1-\alpha = 0.95$.

So, $\alpha = 0.05$,

and $\alpha/2 = 0.025$.



$= -1.96$ (from tables).

So, $Z_{0.025} = 1.96$.

The 95% C.I. for μ is

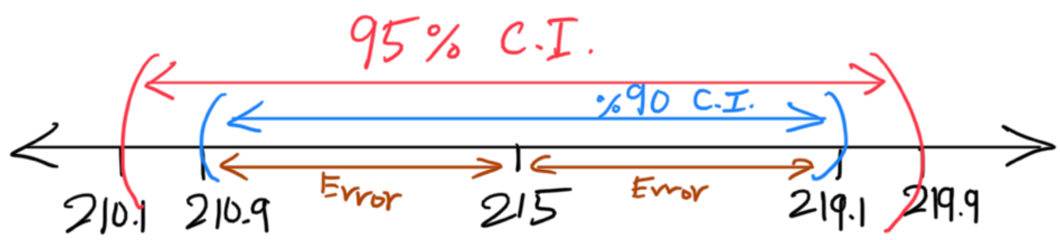
$$\bar{X} \pm Z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} = 215 \pm (1.96)\left(\frac{50}{\sqrt{400}}\right).$$

$$L = 215 - (1.96)(2.5) = 210.10.$$

$$U = 215 + (1.96)(2.5) = 219.90.$$

$\therefore$ The 95% C.I. for μ is $(210.10 \,,\, 219.90)$.

هذا يعني أننا على ثقة مقدارها ٩٥٪
أن ال mean بين هذين الرقمين.



Note that the 90% C.I. is shorter than 95% C.I.

The shortest interval provides more information about the population mean.

هناك أيضاً علاقة عكسية بين قيمة (α) وطول الفترة.

Ex. The mean cholesterol levels in a general population are normally distributed. A sample of 16 persons is taken under a test with sample mean $\overline{X} = 220$ mg/dl and standard deviation $S = 25$ mg/dl. Give 90% C.I. for the population mean μ.

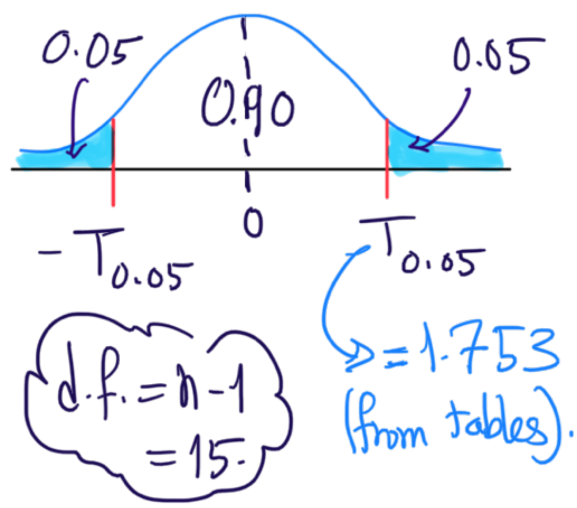
Soln. $n = 16$, $\overline{X} = 220$, $S = 25$.

$1 - \alpha = 0.90$, so $\alpha = 0.10$ and $\alpha/2 = 0.05$.

σ is unknown and $n < 30$.

The 90% C.I. for μ is

$$\overline{X} \pm T_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = 220 \pm 1.753 \left(\frac{25}{\sqrt{16}}\right).$$



$$\text{L.C.L.} = 220 - 1.753 \times \frac{25}{4} = 209.04.$$

$$\text{U.C.L.} = 220 + 1.753 \times \frac{25}{4} = 230.96.$$

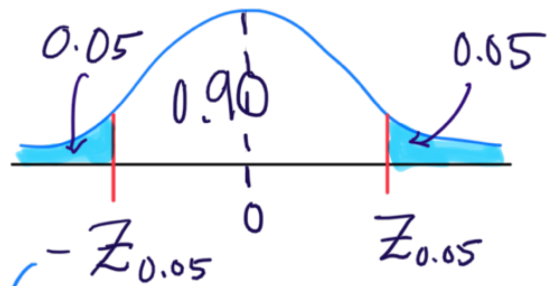
Thus, the 90% C.I. for μ is $(209.04, 230.96)$.

Ex. A random sample of 400 people with a professional degree taken showed that their mean monthly salary is 450 JD with a standard deviation of 100 JD. Give 90% C.I. for the mean monthly salary.

Soln. $n = 400$, $\bar{X} = 450$, $S = 10$.

$1 - \alpha = 0.90$, so $\alpha = 0.10$ & $\frac{\alpha}{2} = 0.05$.

σ is unknown and $n > 30$.



$= -1.64$ (from tables).

So $Z_{0.05} = 1.64$.

The 90% C.I. for μ is

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} = 450 \pm (1.64) \frac{100}{\sqrt{400}}.$$

$$L.C.L. = 450 + 1.64 \frac{100}{\sqrt{400}} = 441.8.$$

$$U.C.L. = 450 - 1.64 \frac{100}{\sqrt{400}} = 458.2.$$

Thus, the 90% C.I. for μ is $(441.8, 458.2)$.

Sample size to estimate μ.

Recall that the error of estimation is

$$E = Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

Then $\dfrac{E}{Z_{\frac{\alpha}{2}}} = \dfrac{\sigma}{\sqrt{n}}$ or $\dfrac{\sqrt{n}}{\sigma} = \dfrac{Z_{\alpha/2}}{E}$.

So, $\sqrt{n} = \dfrac{Z_{\alpha/2}}{E} \cdot \sigma$, hence $n = \left(\dfrac{Z_{\alpha/2}}{E}\right)^2 \sigma^2$.

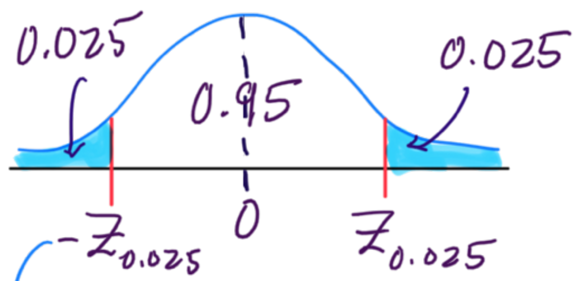
This proves the following fact.

Fact: To estimate μ by $(1-\alpha)100\%$ C.I., we need a sample of size $\boxed{n = \left(\dfrac{Z_{\alpha/2}}{E}\right)^2 \sigma^2}$.

Ex. A researcher wants to estimate the average loss of people who are on a new diet plan. In a preliminary study, the population standard deviation σ of weight losses is about 5 kgs. How large a sample should be to estimate the mean weight loss by 95% C.I. to within 1.5 kgs?

Soln. $\sigma = 5$, $1-\alpha = 0.95$, so $\alpha = 0.05$, and hence $\alpha/2 = 0.025$. Want: n.

$E = 1.5$, $\mu - 1.5 \longleftarrow \mu \longrightarrow \mu + 1.5$



$-Z_{0.025} = -1.96$ (from tables).

So, $Z_{0.025} = 1.96$.

$$n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 \cdot \sigma^2$$

$$= \left(\frac{1.96}{1.5}\right)^2 (5)^2$$

$$= 42.68.$$

Thus, $n \simeq 43$.

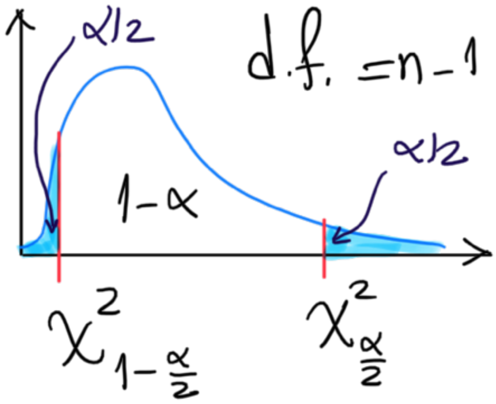
Exc. Suppose that Jordan Bureau of Census in 2004 wants to estimate the mean size μ of all Jordan families by 99% C.I.. It is known that the standard deviation σ for the sizes of all families is 1.5. How large a sample size should be bureau select to estimate μ within 0.02 of the population mean?

Final Ans. $n \simeq 37,153$.

Interval estimation for σ^2.

The $(1-\alpha)100\%$ C.I. for σ^2 is

$$\boxed{\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}},\ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\right)}.$$



$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Ex. Quality-control engineer wishes to study the weight variation of a new product. A sample of 10 items is taken and provided

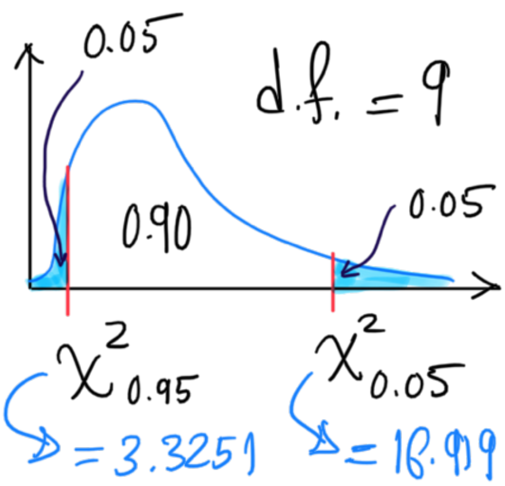
$$\bar{X} = 0.6 \text{ kgs} \quad \text{and} \quad S = 0.4 \text{ kgs}.$$

Assume that the distribution of the weights can be modeled as a normal distribution. Find a 90% C.I. for the variance of all items.

Soln. $n=10$, $\bar{X}=0.60$, $S=0.40$.

$1-\alpha = 0.90$, so $\alpha = 0.10$

and $\alpha/2 = 0.05$.



The 90% C.I. for σ^2 is

$$\left(\frac{(n-1)\,S^2}{\chi^2_{0.025}},\ \frac{(n-1)\,S^2}{\chi^2_{0.95}}\right)=\left(\frac{(10-1)\,(0.40)^2}{16.919},\ \frac{(10-1)\,(0.40)^2}{0.3251}\right)$$

$$=(0.09,\ 0.43).$$

Interval estimation for p.

The $(1-\alpha)\,100\%$ C.I. for p is $\boxed{\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$.

Ex. It was believed in the Arab World that 50% of persons are smoking. During the year 2000, a sample of 1000 persons showed that the number of smokers is 620. Establish 95% C.I. for proportion p of smokers.

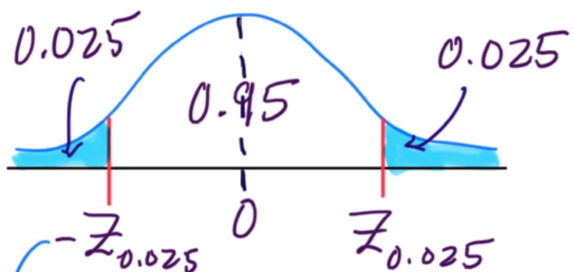
Soln. $n=1000$, $X=620$.

$\hat{p}=\frac{X}{n}=\frac{620}{1000}=0.62.$

$1-\hat{p}=1-0.62=0.38.$

$1-\alpha=0.95$, so $\alpha=0.05$,

hence $\alpha/2=0.025$.



$= -1.96$ (from tables).

So, $Z_{0.025}=1.96$.

The 95% C.I. for p is $\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

$$L.C.L. = 0.62 - (1.96)\sqrt{\frac{(0.62)(0.38)}{1000}} = 0.59.$$

$$U.C.L. = 0.62 + (1.96)\sqrt{\frac{(0.62)(0.38)}{1000}} = 0.65.$$

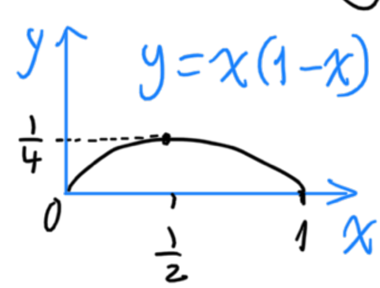
Thus, the 95% C.I. for p is $(0.59, 0.65)$.

Sample size to estimate p.

Fact: To estimate p by $(1-\alpha)100\%$ C.I., we need a sample of size $\boxed{n = \left(\dfrac{Z_{\alpha/2}}{E}\right)^2 p^*(1-p^*)}$,

where $p^* = \begin{cases} \hat{p} & \text{if } \hat{p} \text{ is known from a previous study.} \\ \frac{1}{2} & \text{if } \hat{p} \text{ is not know from a previous study.} \end{cases}$

Why $\frac{1}{2}$? As the func. $p^*(1-p^*)$ attains its maximum at $p^* = 1/2$.



Ex. Assume that it is required to estimate the proportion of patients suffering a bad reaction from taking a certain medication p by 95% C.I.. Determine the sample size needed if the error of estimation is about 0.10 in the following cases:-

a) no prior information about p.

b) previous study showed that p is approximately by 0.20.

Soln. $E = 0.10$, $1-\alpha = 0.95$, $\alpha = 0.05$, $\alpha/2 = 0.025$.

$Z_{0.025} = 1.96$. Want: n.

a) $p^* = 0.5$, so $n = \left(\frac{1.96}{0.10}\right)^2 (0.5)(1-0.5) \cong 97$.

b) $p^* = 0.2$, so $n = \left(\frac{1.96}{0.10}\right)^2 (0.2)(1-0.2) \cong 62$.

Summary.

The $(1-\alpha)100\%$ C.I.

for μ

for σ^2

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}\right)$$

for p

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 p^*(1-p^*)$$

σ is known

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2 \cdot \sigma^2.$$

σ is unknown but $n<30$

$$\bar{X} \pm T_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$$

σ is unknown but $n \geq 30$

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

Searching keywords:

- Estimation of confidence interval
- Interval estimation for
- Sample size to estimate
- The University of Jordan الجامعة الأردنية
- Principles of Statistics مبادئ الإحصاء
- Baha Alzalg بهاء الزالق

References: See the textbook in the course website

http://sites.ju.edu.jo/sites/Alzalg/Pages/131.aspx

For any comments or concerns, please use my email to contact me.

د. بهاء محمود الزالق
The University of Jordan
Dr. Baha Alzalg
baha2math@gmail.com

This material might not be used for commercial purposes.