Introduction to mathematical Statistics

1. Sample size estimation based on the large sample C.I. for p

From the interval $\left[\hat{p} - Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

$$L = \text{lengh of your } 100(1-\alpha)\% \text{ CI} = 2 \times Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$L, \alpha, \hat{p}$ are given and we are interested in sample size n. Therefore,

$$n = \frac{4(Z_{\alpha/2})^2 \hat{p}(1-\hat{p})}{L^2} \leq \frac{4(Z_{\alpha/2})^2 \left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right)}{L^2} = \frac{(Z_{\alpha/2})^2}{L^2}$$

(When $\hat{p} = \frac{1}{2}$, it has the maximum value.)

Example. $L = 0.02,\ \alpha = 0.05,\ \hat{p} = 0.54 \Rightarrow n = ?$

Example. $L = 0.02,\ \alpha = 0.05,\ \hat{p} = 0.5 \Rightarrow n = ?$

2. Sample size calculation for p based on the *maximum error E*.

Definition. $P(|\hat{p} - p| \leq E) = 1 - \alpha$

We want to estimate p **<u>within</u>** E with a probability of $(1-\alpha)$.

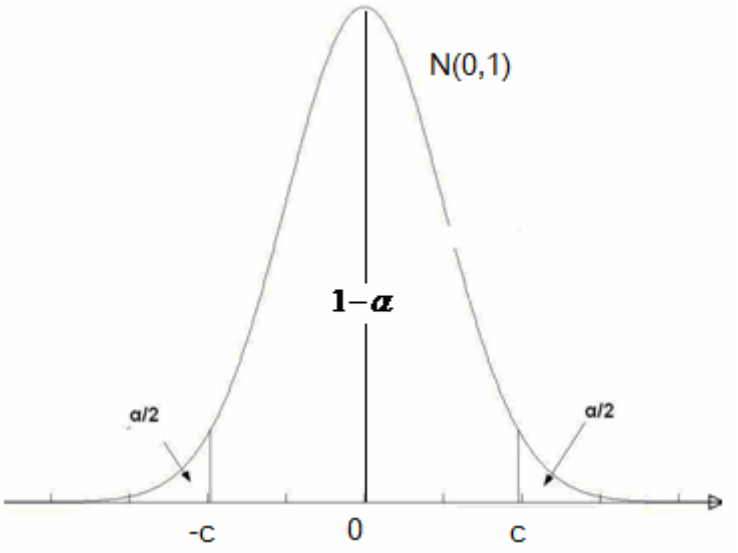
Derive the formula for n

$$P(|\hat{p}-p| \le E) = 1-\alpha$$

$$P(-E \le \hat{p}-p \le E) = 1-\alpha$$

$$P(-\frac{E}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{E}{\sqrt{\frac{p(1-p)}{n}}}) = 1-\alpha \text{ and } Z = \frac{p-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

$$Thus: \; P(-\frac{E}{\sqrt{\frac{p(1-p)}{n}}} \le Z \le \frac{E}{\sqrt{\frac{p(1-p)}{n}}}) = 1-\alpha$$



$$c = Z_{\alpha/2} = \frac{E}{\sqrt{\frac{p(1-p)}{n}}}$$

$$\therefore n = \frac{(Z_{\alpha/2})^2 p(1-p)}{E^2} \le \frac{(Z_{\alpha/2})^2}{4 \cdot E^2}$$

Recall we also derived n based on L - the length of the $100(1-\alpha)\%$ large sample confidence interval for p. Their relationship is

$$L = 2 \cdot E$$

$$P(|\hat{p}-p| \le E) = 1-\alpha$$
$$P(-E \le \hat{p}-p \le E) = 1-\alpha$$
$$P(-E-\hat{p} \le -p \le E-\hat{p}) = 1-\alpha$$
$$P(\hat{p}-E \le p \le \hat{p}+E) = 1-\alpha$$

$\therefore$ The $100(1-\alpha)\%$ confidence interval for p is $\left[\hat{p}-E, \hat{p}+E\right]$.

The length of the confidence interval is $L=\left(\hat{p}+E\right)-\left(\hat{p}-E\right)=2E$

Example. In order to estimate the percent of children with inadequate immunization to be within 0.05 of the true proportion with a probability of 98%

(a) How many children should be sampled?

Solution. $E=0.05,\ \alpha=1-0.98=0.02$

$$n=\frac{\left(Z_{0.01}\right)^2}{4\times(0.05)^2}$$
$$=\frac{\left(2.33\right)^2}{4\times(0.05)^2}\approx 543$$

(b) If the percentage of children with inadequate immunization is estimated to be $\leq 20\%$, then $n=?$

Solution. $\hat{p}=20\%$

$$n=\frac{\left(Z_{0.01}\right)^2(0.2)(1-0.2)}{(0.05)^2}=348$$

Sample size calculates for 1 population proportions based on the ***maximum error E***.

3. Confidence Interval for 1 population mean μ

There are 4 scenarios – we cover only the first 3 scenarios in our class.

1. **Normal population, σ^2 is known.**
   a. Point Estimator : $\overline{X}$
   b. Pivotal Quantity : $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

c. $100(1-\alpha)\%$ CI for μ : $P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha \Rightarrow \overline{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

d. Length of CI : $L = 2 \cdot Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

e. Sample size based on L : $n = \frac{4\left(Z_{\alpha/2}\right)^2 \sigma^2}{L^2}$

f. Sample size based on E

$P(\left|\overline{X} - \mu\right| \le E) = 1-\alpha$

$P(-E \le \overline{X} - \mu \le E) = 1-\alpha$

$P(\overline{X} - E \le \mu \le \overline{X} + E) = 1-\alpha$

$\therefore L = \left(\overline{X} + E\right) - \left(\overline{X} - E\right) = 2E$

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

2. Normal population, σ^2 is unknown.

a. Point Estimator : $\overline{X}$

b. Pivotal Quantity : $T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

c. $100(1-\alpha)\%$ CI for μ : $P(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) = 1-\alpha \Rightarrow \overline{X} \pm t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}$

d. Length of CI : $L = 2 \cdot t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$

e. Sample size based on L : $n = \frac{4\left(t_{n-1,\alpha/2}\right)^2 S^2}{L^2}$

f. Sample size based on E

$P(\left|\overline{X} - \mu\right| \le E) = 1-\alpha$

$P(-E \le \overline{X} - \mu \le E) = 1-\alpha$

$P(\overline{X} - E \le \mu \le \overline{X} + E) = 1-\alpha$

$\therefore L = \left(\overline{X} + E\right) - \left(\overline{X} - E\right) = 2E$

$\therefore n = \frac{\left(t_{n-1,\alpha/2}\right)^2 S^2}{E^2}$

3. **Any populations, large sample**
   a. Point Estimator : $\overline{X}$
   b. Pivotal Quantity : $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$ or $Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \dot{\sim} N(0,1)$
   c. $100(1-\alpha)\%$ CI for μ : $P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$

$$\Rightarrow \begin{cases} \overline{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \\ \overline{X} \pm Z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \end{cases}$$

   d. Length of CI : $L = 2 \cdot Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ or $L = 2 \cdot Z_{\alpha/2} \frac{S}{\sqrt{n}}$
   e. Sample size based on L : $n = \frac{4\left(Z_{\alpha/2}\right)^2 \sigma^2}{L^2}$ or $n = \frac{4\left(Z_{\alpha/2}\right)^2 S^2}{L^2}$
   f. Sample size based on E

$$P(\left|\overline{X} - \mu\right| \le E) = 1-\alpha$$

$$P(-E \le \overline{X} - \mu \le E) = 1-\alpha$$

$$P(\overline{X} - E \le \mu \le \overline{X} + E) = 1-\alpha$$

$$\therefore L = \left(\overline{X} + E\right) - \left(\overline{X} - E\right) = 2E$$

$$\therefore n = \frac{\left(Z_{\alpha/2}\right)^2 \sigma^2}{E^2} \text{ or } n = \frac{\left(Z_{\alpha/2}\right)^2 S^2}{E^2}$$

4. **There also exist other cases, but we don't cover those in our class.**

Now I will present more details for Scenario 2 mentioned above.

Scenarios 1 & 3 (easy)

Scenario 2 : normal population, σ^2 unknown

1. Point estimation : $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
2. $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

3. **Theorem.** Sampling from normal population
   a. $Z \sim N(0,1)$
   b. $W = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
   c. Z and W are independent.

Definition. $T = \dfrac{Z}{\sqrt{W/(n-1)}} = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

---------- **Derivation of CI, normal population, σ^2 is unknown** ----------

$\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$ is not a pivotal quantity.

$\overline{X} - \mu \sim N(0, \dfrac{\sigma^2}{n})$ is not a pivotal quantity.

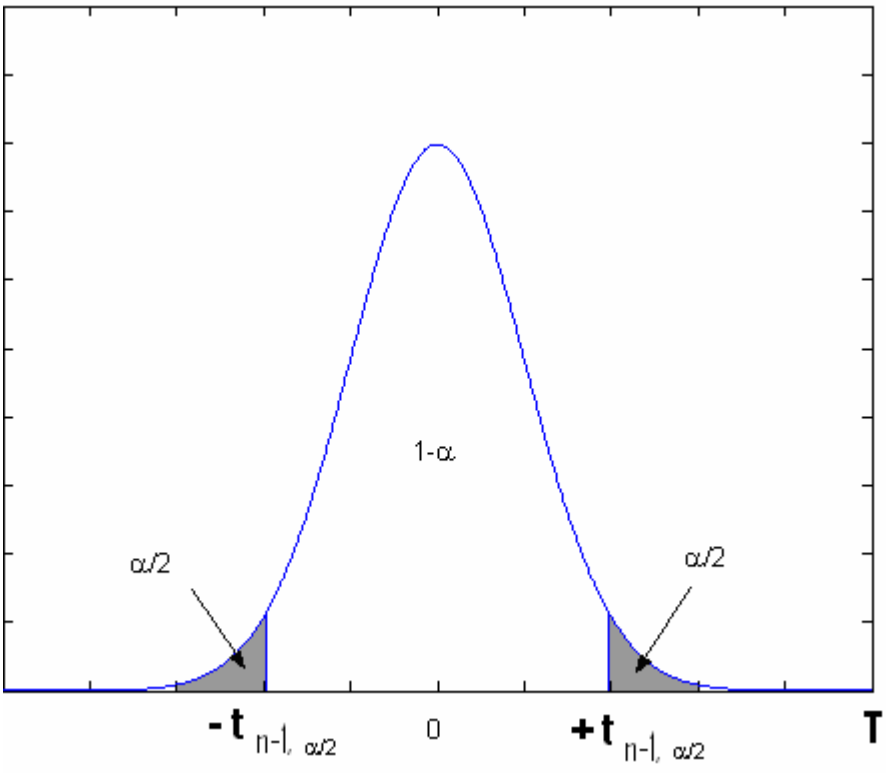
$Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ is not a pivotal quantity.

Remove σ !!!

Therefore $\boxed{T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}}$ is a pivotal quantity.

Now we will use this pivotal quantity to derive the 100(1-α)% confidence interval for μ.

We start by plotting the pdf of the t-distribution with n-1 degrees of freedom as follows:



The above pdf plot corresponds to the following probability statement:

$$P(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) = 1-\alpha$$

$$\Rightarrow P(-t_{n-1,\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}) = 1-\alpha$$

$$\Rightarrow P(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \bar{X}-\mu \le t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$\Rightarrow P(-\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le -\mu \le -\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$\Rightarrow P(\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \ge \mu \ge \bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$\Rightarrow P(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$\Rightarrow$ Thus the $100(1-\alpha)\%$ C.I. for μ when σ^2 is unknown is

$$[\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}].$$ (*Please note that $t_{n-1,\alpha/2} \geq Z_{\alpha/2}$)

Example. In a random sample of $n = 36$ parochial schools throughout the south, the average number of pupils per school is 379.2 with a standard deviation of 124. Use the sample to construct a 95% CI for μ, the mean number of pupils per school for all parochial schools in the south.

Solution. CI for μ, large sample

$n = 36,\ \bar{X} = 379.2,\ S = 124,\ \alpha$=0.05

95% CI for μ is $\bar{X} \pm Z_{0.05} \cdot \frac{S}{\sqrt{n}} = 379.2 \pm 1.96 \cdot \frac{124}{\sqrt{36}}$

$\therefore [338.7, 419.7]$

Example. In a psychological depth-perception test, a random sample of $n = 14$ airline pilots were asked to judge the distance between 2 markers at the other end of a laboratory.

The data (in test) are

2.7, 2.4, 1.9, 2.4, 1.9, 2.3, 2.2, 2.5, 2.3, 1.8, 2.5, 2.0, 2.2, 2.6

Please construct a 95% CI for μ, the average distance.

Solution.

(Note: we can perform the Shapiro-Wilk test to examine whether the sample comes from a normal population or not. This test is not required in our class. Here we simply assume the population is normal. I will always give you such information in the exams.)

CI for μ, small sample, normal population, population variance unknown.

$n = 14,\ \bar{X} = 2.26,\ S = 0.28,\ \alpha$=0.05

95% CI for μ is $\bar{X} \pm t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}} = 2.26 \pm 2.16 \cdot \frac{0.28}{\sqrt{14}}$

$\therefore [2.10, 2.42]$

Example. A federal agency has decided to investigate the advertised weight we printed on cartons of a certain brand of cereal. Historical data show that $\sigma = 0.75$ ounce. If we wish to estimate the weight within 0.25 ounce with 99% confidence, how many cartons should be sampled?

Solution. $E = 0.25,\ \sigma = 0.75,\ \alpha = 0.01$

$$n = \frac{\left(Z_{0.005}\right)^2 \left(0.75\right)^2}{\left(0.25\right)^2} \approx 60$$