Large-Sample Confidence Interval for a Population Proportion

It is often necessary to construct confidence intervals on a population proportion.

- **Suppose a random sample of size n has been taken from a large (possible infinite) population and that X (≤n) observations in this sample belong to a class of interest.**
- **Then** $\hat{p} = \frac{X}{n}$ **is a *point estimator* of the proportion of the population p that belongs to this class.**
- **Note that n and p are the parameters of a binomial distribution.**
- **We know that the sampling distribution of** $\hat{p}$ **is approximately normal with mean p and variance p(1-p)/n, if *p* is <u>not too close</u> to either 0 or 1 and if n is relatively large.**

Normal Approximation for a Binomial Proportion

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal.

To construct the confidence interval on p, note that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$$

so

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1-\alpha$$

This may be rearranged as

$$P\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p}+z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

The quantity $\sqrt{p(1-p)/n}$ is called the *standard error of the point estimator* $\hat{p}$. Unfortunately, the upper and lower limits of the confidence interval contain the unknown parameter p. A satisfactory solution is to replace p by $\hat{p}$ in the standard error, which result in.

$$P\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

Approximate Confidence Interval on a Binomial Proportion

If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate 100(1-α)% confidence interval on the proportion p of the population that belong to this class is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Example: The fraction of defective integrated circuits produced in a photolithography process is being studied. A random sample of 85 circuits is tested, revealing 10 defectives.

Calculate a 95% CI on the fraction of defective circuits produced by this particular tool.

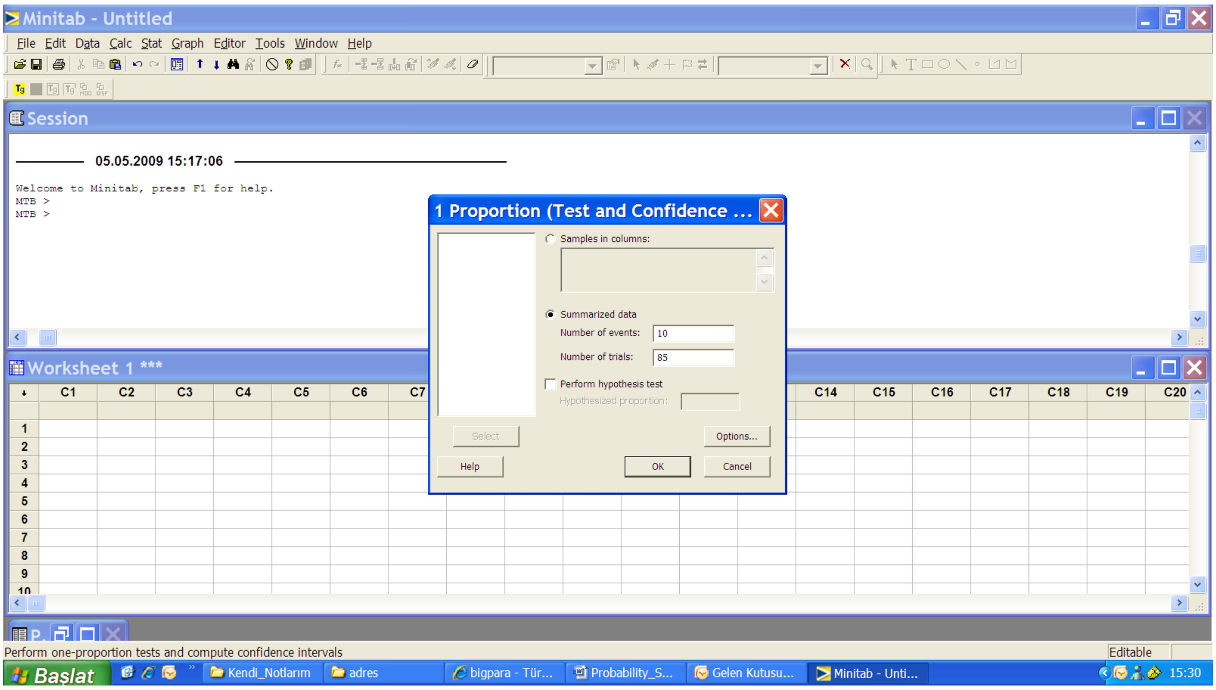
Point estimate $\hat{p} = \frac{X}{n} = \frac{10}{85} = 0.117647$

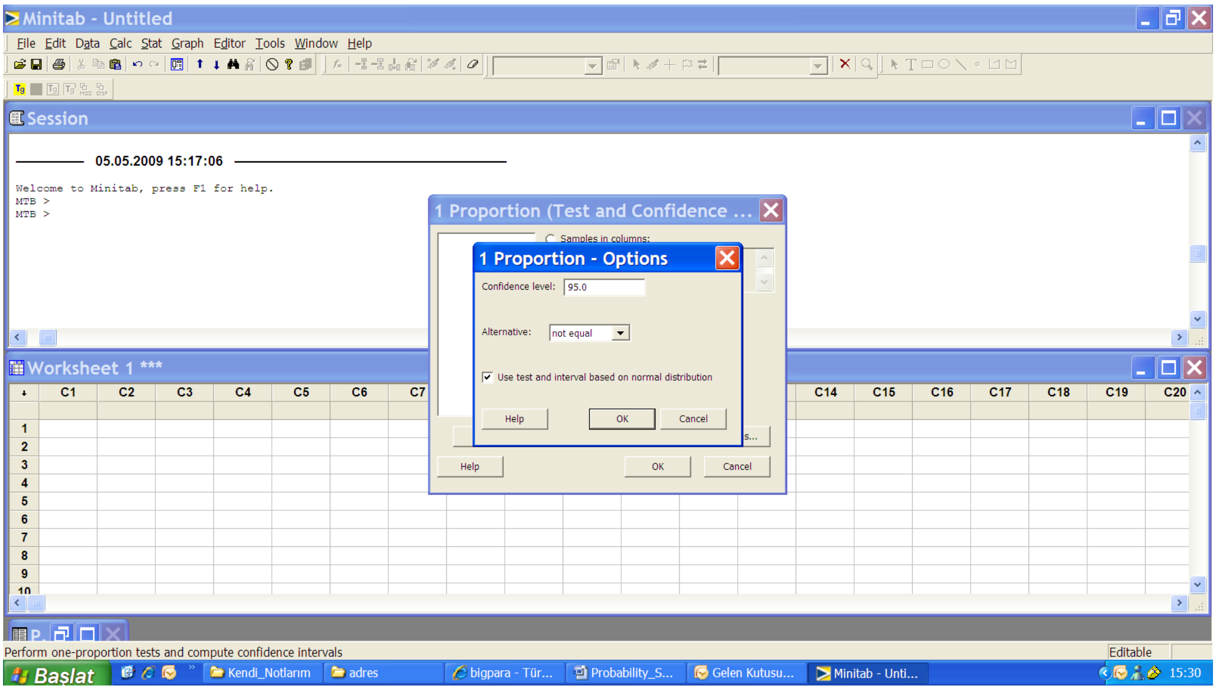
$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.117647 - 1.96\sqrt{\frac{0.117647(1-0.117647)}{85}} \le p \le 0.117647 + 1.96\sqrt{\frac{0.117647(1-0.117647)}{85}}$$

which simplifies to

$$0.049153 \le p \le 0.186141$$






```
MTB > POne 85 10;
SUBC>   UseZ.
```

Test and CI for One Proportion

```
Sample    X   N  Sample p         95% CI
1        10  85  0.117647  (0.049153; 0.186141)

Using the normal approximation.
```

```
MTB > Random 100 c1-c100;
SUBC>   Bernoulli 0.3.
MTB > POne C1-C100;
SUBC>   UseZ.

Test and CI for One Proportion: C1; C2; C3; C4; C5; C6; C7; C8; ...

Event = 1

Variable   X    N  Sample p        95% CI
C1        34  100  0.340000  (0.247155; 0.432845)
C2        37  100  0.370000  (0.275372; 0.464628)
C3        39  100  0.390000  (0.294403; 0.485597)
C4        23  100  0.230000  (0.147518; 0.312482)
C5        34  100  0.340000  (0.247155; 0.432845)
C6        39  100  0.390000  (0.294403; 0.485597)
C7        23  100  0.230000  (0.147518; 0.312482)
C8        30  100  0.300000  (0.210183; 0.389817)
C9        30  100  0.300000  (0.210183; 0.389817)
C10       29  100  0.290000  (0.201064; 0.378936)
C11       34  100  0.340000  (0.247155; 0.432845)
C12       24  100  0.240000  (0.156293; 0.323707)
C13       25  100  0.250000  (0.165131; 0.334869)
C14       35  100  0.350000  (0.256516; 0.443484)
C15       33  100  0.330000  (0.237840; 0.422160)
C16       34  100  0.340000  (0.247155; 0.432845)
C17       21  100  0.210000  (0.130169; 0.289831)
C18       32  100  0.320000  (0.228572; 0.411428)
C19       25  100  0.250000  (0.165131; 0.334869)
C20       24  100  0.240000  (0.156293; 0.323707)
C21       32  100  0.320000  (0.228572; 0.411428)
C22       40  100  0.400000  (0.303982; 0.496018)
C23       42  100  0.420000  (0.323264; 0.516736)
C24       33  100  0.330000  (0.237840; 0.422160)
C25       25  100  0.250000  (0.165131; 0.334869)
C26       28  100  0.280000  (0.191998; 0.368002)
C27       31  100  0.310000  (0.219353; 0.400647)
C28       27  100  0.270000  (0.182986; 0.357014)
C29       27  100  0.270000  (0.182986; 0.357014)
C30       32  100  0.320000  (0.228572; 0.411428)
C31       34  100  0.340000  (0.247155; 0.432845)
C32       31  100  0.310000  (0.219353; 0.400647)
C33       31  100  0.310000  (0.219353; 0.400647)
C34       37  100  0.370000  (0.275372; 0.464628)
C35       30  100  0.300000  (0.210183; 0.389817)
C36       29  100  0.290000  (0.201064; 0.378936)
C37       23  100  0.230000  (0.147518; 0.312482)
C38       25  100  0.250000  (0.165131; 0.334869)
C39       37  100  0.370000  (0.275372; 0.464628)
C40       27  100  0.270000  (0.182986; 0.357014)
C41       37  100  0.370000  (0.275372; 0.464628)
C42       33  100  0.330000  (0.237840; 0.422160)
C43       32  100  0.320000  (0.228572; 0.411428)
C44       31  100  0.310000  (0.219353; 0.400647)
C45       31  100  0.310000  (0.219353; 0.400647)
C46       38  100  0.380000  (0.284866; 0.475134)
```

C47	26	100	0.260000	(0.174029; 0.345971)
C48	23	100	0.230000	(0.147518; 0.312482)
C49	34	100	0.340000	(0.247155; 0.432845)
C50	32	100	0.320000	(0.228572; 0.411428)
C51	35	100	0.350000	(0.256516; 0.443484)
C52	34	100	0.340000	(0.247155; 0.432845)
C53	27	100	0.270000	(0.182986; 0.357014)
C54	31	100	0.310000	(0.219353; 0.400647)
C55	30	100	0.300000	(0.210183; 0.389817)
C56	32	100	0.320000	(0.228572; 0.411428)
C57	26	100	0.260000	(0.174029; 0.345971)
C58	23	100	0.230000	(0.147518; 0.312482)
C59	33	100	0.330000	(0.237840; 0.422160)
C60	29	100	0.290000	(0.201064; 0.378936)
C61	23	100	0.230000	(0.147518; 0.312482)
C62	29	100	0.290000	(0.201064; 0.378936)
C63	41	100	0.410000	(0.313602; 0.506398)
C64	30	100	0.300000	(0.210183; 0.389817)
C65	32	100	0.320000	(0.228572; 0.411428)
C66	34	100	0.340000	(0.247155; 0.432845)
C67	29	100	0.290000	(0.201064; 0.378936)
C68	22	100	0.220000	(0.138809; 0.301191)
C69	28	100	0.280000	(0.191998; 0.368002)
C70	30	100	0.300000	(0.210183; 0.389817)
C71	29	100	0.290000	(0.201064; 0.378936)
C72	41	100	0.410000	(0.313602; 0.506398)
C73	30	100	0.300000	(0.210183; 0.389817)
C74	28	100	0.280000	(0.191998; 0.368002)
C75	29	100	0.290000	(0.201064; 0.378936)
C76	27	100	0.270000	(0.182986; 0.357014)
C77	32	100	0.320000	(0.228572; 0.411428)
C78	28	100	0.280000	(0.191998; 0.368002)
C79	27	100	0.270000	(0.182986; 0.357014)
C80	35	100	0.350000	(0.256516; 0.443484)
C81	32	100	0.320000	(0.228572; 0.411428)
C82	40	100	0.400000	(0.303982; 0.496018)
C83	25	100	0.250000	(0.165131; 0.334869)
C84	27	100	0.270000	(0.182986; 0.357014)
C85	30	100	0.300000	(0.210183; 0.389817)
C86	31	100	0.310000	(0.219353; 0.400647)
C87	32	100	0.320000	(0.228572; 0.411428)
C88	32	100	0.320000	(0.228572; 0.411428)
C89	32	100	0.320000	(0.228572; 0.411428)
C90	31	100	0.310000	(0.219353; 0.400647)
C91	25	100	0.250000	(0.165131; 0.334869)
C92	32	100	0.320000	(0.228572; 0.411428)
C93	31	100	0.310000	(0.219353; 0.400647)
C94	30	100	0.300000	(0.210183; 0.389817)
C95	40	100	0.400000	(0.303982; 0.496018)
C96	35	100	0.350000	(0.256516; 0.443484)
C97	33	100	0.330000	(0.237840; 0.422160)
C98	39	100	0.390000	(0.294403; 0.485597)
C99	29	100	0.290000	(0.201064; 0.378936)
C100	30	100	0.300000	(0.210183; 0.389817)

Using the normal approximation.

Sample Size for a Specified Error on a Binomial Proportion

Since $\hat{p}$ is the point estimator of p, we can define the error in estimating p by $\hat{p}$ as $E = |p - \hat{p}|$. Note that we are approximately 100(1-α)% confident that <u>this error is less than</u>

$$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

In situations where the sample size can be selected, we may choose n to be 100(1-α)% confident that the error is less than some specified value E. If we set

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

and solve for n, the appropriate sample size is

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

An estimate of p is required to use this equation. If an estimate $\hat{p}$ from a previous sample is available, it can be substituted for p in the equation.

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

<u>This means that a preliminary sample can be taken, $\hat{p}$ computed, and then the given equation used to determine how many additional observations are required to estimate p with the desired accuracy.</u>

Example: Consider the situation in the previous example.

How large a sample is required if we want to be 95% confident that the error in using $\hat{p}$ to estimate p is less that 0.05?

$$0.049153 \le p \le 0.186141$$

CI=0.186141-0.049153=0.136988

Using $\hat{p} = \frac{10}{85} = 0.117647$ **as an initial estimate of p, we find from**

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.96}{0.05}\right)^2 0.117647(0.882353)$$
$$= \left(\frac{1.96}{0.05}\right)^2 0.10381 = 159.512 \cong 160$$

160*0.117647=18.82>>>18

```
MTB > POne 160 18.
```

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	18	160	0.112500	(0.068054; 0.171962)

CI=0.171962-0.068054=0.1039

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	19	160	0.118750	(0.073031; 0.179219)

CI=0.179219-0.073031=0.1062

Another Approach (Upper Bound)

Another approach to choosing n uses the fact that the sample size

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

will always be a maximum for p=0.5

[that is p(1-p)≤0.25 with equality for p=0.5] , and this can be used to obtain an upper bound on n.

In other words, we are at least 100(1-α)% confident that the error in estimating p by $\hat{p}$ is less than E if the sample size is

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (0.25)$$

If we wanted to be at least 95% confident that our estimate $\hat{p}$ of the true proportion p was within 0.05 regardless of the value of p, we would use this equation to find the sample size

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (0.25) = \left(\frac{1.96}{0.05}\right)^2 (0.25) \cong 385$$

385*0.117647=45.29>>>45

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	45	385	0.116883	(0.086545; 0.153260)

CI=0.153260-0.086545=0.0667