STATISTICS

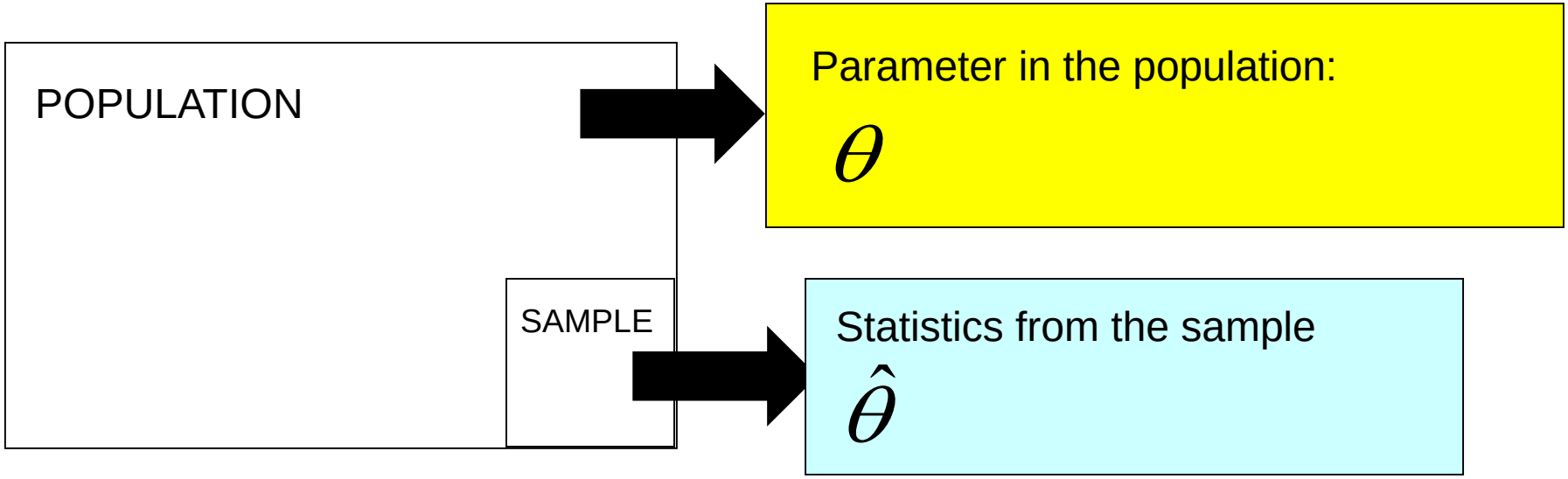
Estimation

Paweł Strzelecki
Lucas van der Velde

Warsaw School of Economics
Spring Semester 2018

Lecture is based on J.T. Mc Clave, P.G. Benson,T. Sincich: Statistics for Business and Economics, 11th Edition, 2010

The idea of estimation



How to estimate the values of population parameters on the basis of the sample?

- Point estimators
- Interval estimators (confidence intervals)

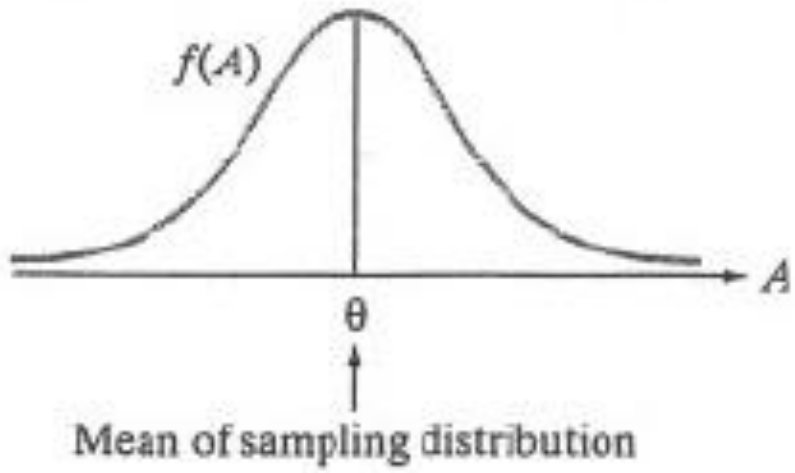
Point estimator and its properties

- A **point estimator** of a population parameter is a rule on how to use the sample data to calculate an estimate of the population parameter (one number)

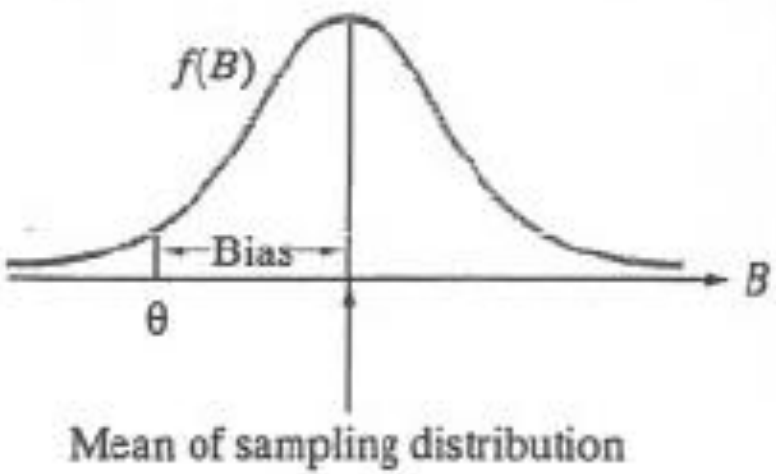
- **Properties of point estimators:**
  - Unbiased
  - Efficiency
  - Consistent

Biased and unbiased estimators

The estimator is unbiased if its expected value equals to the population parameter. Otherwise it is **biased**



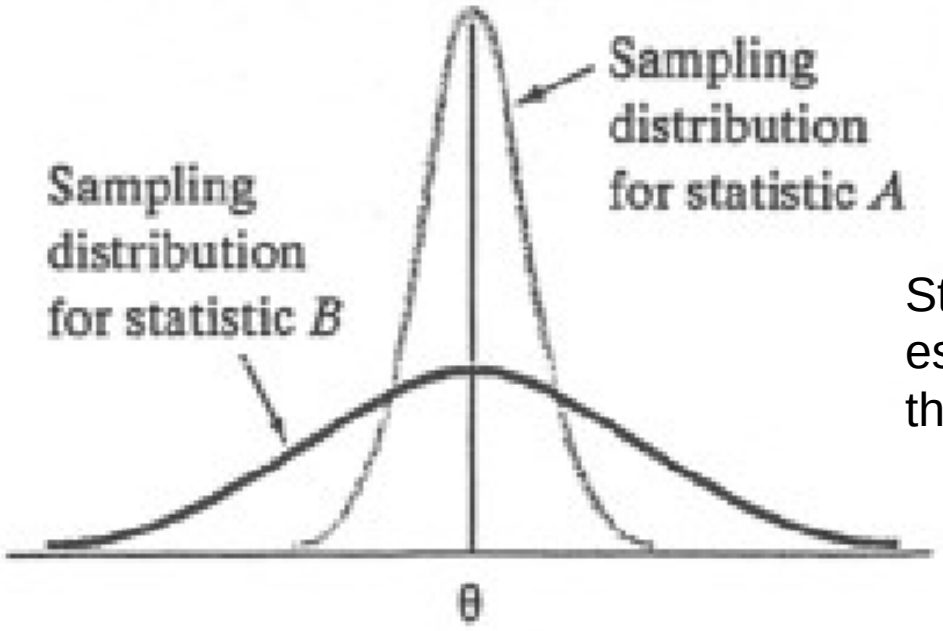
a. Unbiased sample statistic for the parameter θ



b. Biased sample statistic for the parameter θ

Efficiency of estimators

If two estimator are unbiased it is usually better to chose an estimator with lower variance (minimum variance), i.e. a more **efficient estimator**



Statistics A is a more efficient estimator of population parameter θ than statistics B

Consistency of estimators

- A **point estimator** is consistent if its values tend to become closer to the population parameter as the sample size becomes larger → larger samples give better estimates

Point Estimators

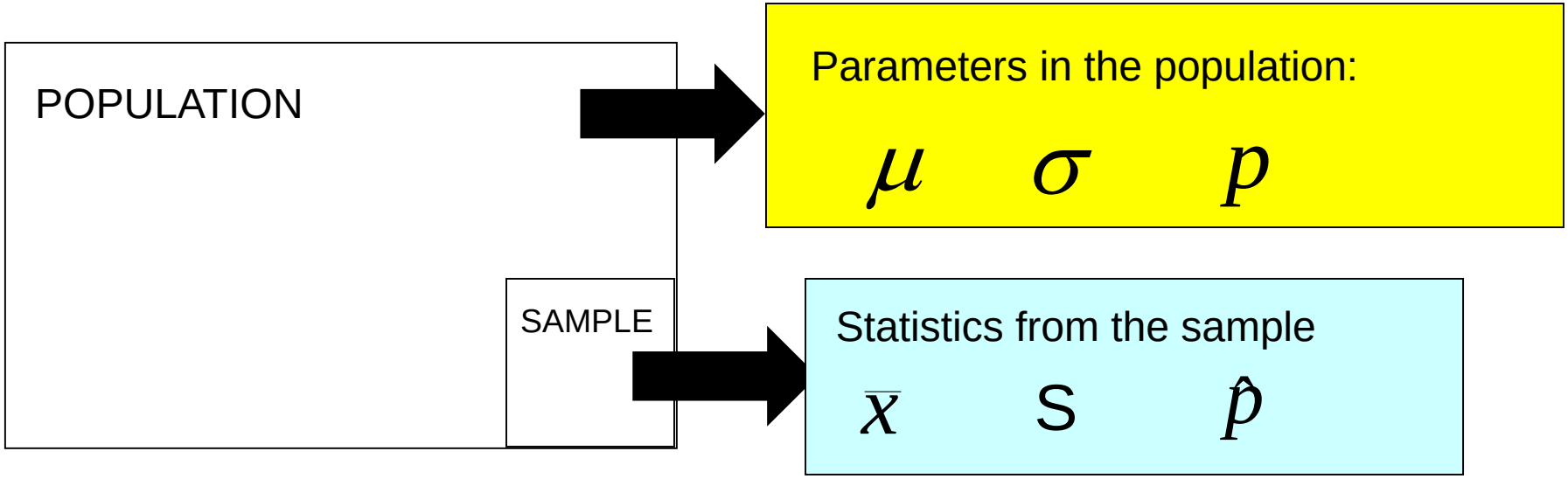
$$\bar{x} = \frac{\sum_i x_i}{n}$$ minimum variance unbiased estimator of μ

$$\hat{p} = \frac{x}{n}$$ minimum variance unbiased estimator of p

$$\hat{S} = \sqrt{\frac{1}{n-1}\sum (x_i - x)^2}$$ minimum variance unbiased estimator of σ

$$\tilde{S} = \sqrt{\frac{1}{n}\sum (x_i - x)^2}$$ biased estimator in smaller samples (asymptotically unbiased)

Point Estimators



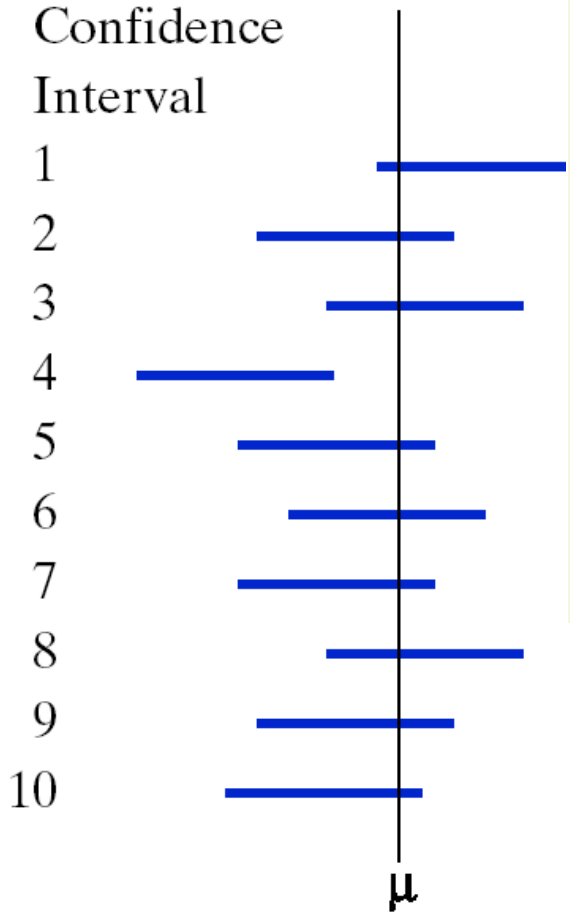
Is it enough to say that the population parameters equals the relevant sample statistics?

What about sampling error?

Interval estimators

- Point estimators are simple but they do not inform us about the reliability of the estimation.
- **Interval estimators** are rules that allow us to compute **confidence intervals**, i.e. intervals which include an unknown population parameter with a certain confidence.
- **Confidence coefficient** is the probability that a randomly selected confidence interval encloses the population parameter (1-α)
- Confidence coeficients usually range from 0.9 to 0.99.

Confidence interval



Interpretation of a Confidence Interval for a Population Mean

When we form a $100(1-\alpha)\%$ confidence interval for μ, we usually express our confidence in the interval with a statement such as, "We can be $100(1-\alpha)\%$ confident that μ lies between the lower and upper bounds of the confidence interval," where for a particular application, we substitute the appropriate numerical values for the confidence and for the lower and upper bounds. *The statement reflects our confidence in the estimation process rather than in the particular interval that is calculated from the sample data.* We know that repeated application of the same procedure will result in different lower and upper bounds on the interval. Furthermore, we know that $100(1-\alpha)\%$ of the resulting intervals will contain μ. There is (usually) no way to determine whether any particular interval is one of those that contain μ, or one that does not. However, unlike point estimators, confidence intervals have some measure of reliability, the confidence coefficient, associated with them. For that reason they are generally preferred to point estimators.

Interval estimators

- For a mean
  - If a sample was drawn from a normally distributed population with known standard deviation
  - If a sample was drawn from a normally distributed population with unknown standard deviation
    - Large sample
    - Small sample
  - If a sample was drawn from a population with unknown distribution
- For a proportion
  - If a sample consists of 100 or more elements

Interval estimators

- For a mean
  - If a sample was drawn from a normally distributed population with known standard deviation
  - If a sample was drawn from a normally distributed population with unknown standard deviation
    - Large sample
    - Small sample
  - If a sample was drawn from a population with unknown distribution

In these cases we proceed in a very similar way, using the fact that the sample mean is normally distributed

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \qquad \text{or} \qquad \bar{x} \sim N(\mu, \frac{S}{\sqrt{n}})$$

Interval estimators

- For a mean
  - If a sample was drawn from a normally distributed population with known standard deviation
  - If a sample was drawn from a normally distributed population with unknown standard deviation
    - Large sample
    - Small sample
  - If a sample was drawn from a population with unknown distribution

In this case we proceed differently, using the fact that the standardised sample mean has a t-distribution

$$\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} \sim t-Student$$

Confidence interval for population mean

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Standard error

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

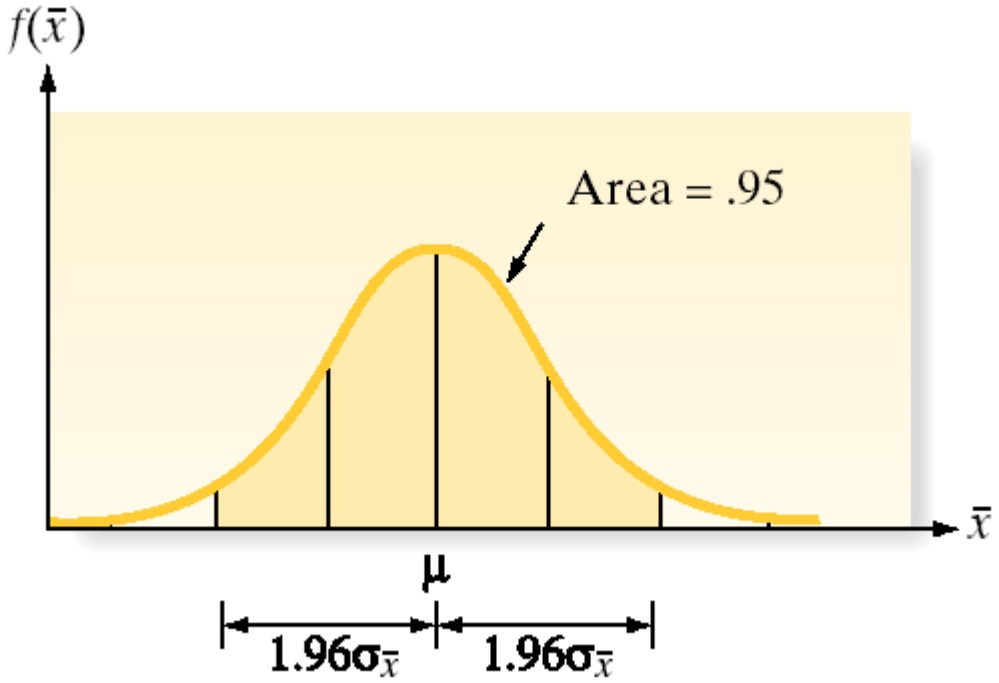
$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

$$P(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$

Sampling error
SE

$$P(\bar{x} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot \sigma / \sqrt{n}) = 1 - \alpha$$

Confidence interval - construction



If the sampling distribution is known we can calculate the interval in which the statistics from the sample will be with assumed probability

Confidence interval for population mean

- If a sample was drawn from a normally distributed population with known standard deviation

$$P(\bar{x} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot \sigma / \sqrt{n}) = 1 - \alpha$$

- If a sample was drawn from a normally distributed population with unknown standard deviation
  - Large sample

$$P(\bar{x} - z_{\alpha/2} \cdot S / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot S / \sqrt{n}) = 1 - \alpha$$

  - Small sample

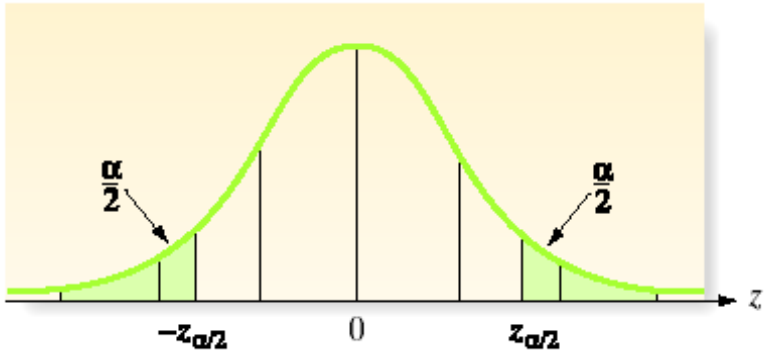
    ?????

- If a large sample was drawn from a population w/unknown distribution

$$P(\bar{x} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot \sigma / \sqrt{n}) = 1 - \alpha$$

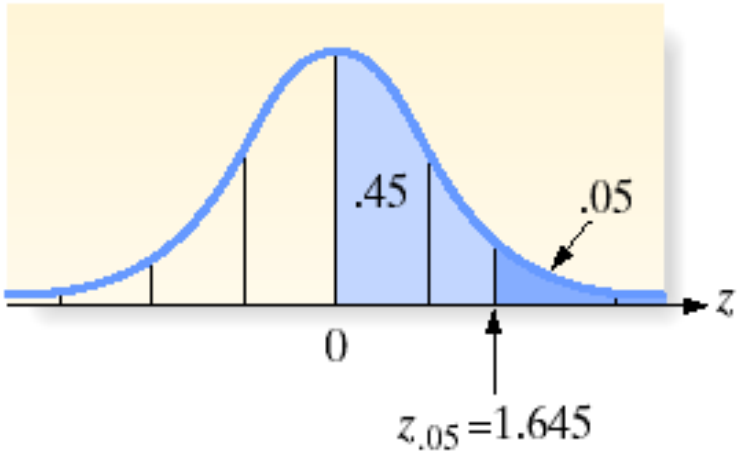
How to find $z_{\alpha/2}$?

$$\overline{x} \pm z_{\alpha/2}\sigma_x$$



Commonly used Values of α

Confidence Level 100(1 − α)	α	α/2	$z_{\alpha}/2$
90%	.10	.05	1.645
95%	.05	.025	1.96
99%	.01	.005	2.575



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

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

How to find $z_{\alpha/2}$?

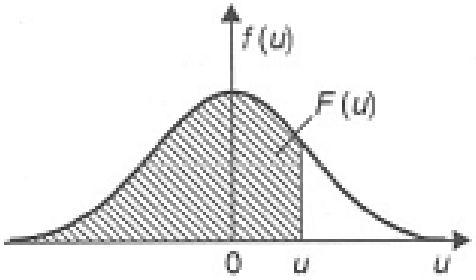


Table 1. Cumulative normal distribution

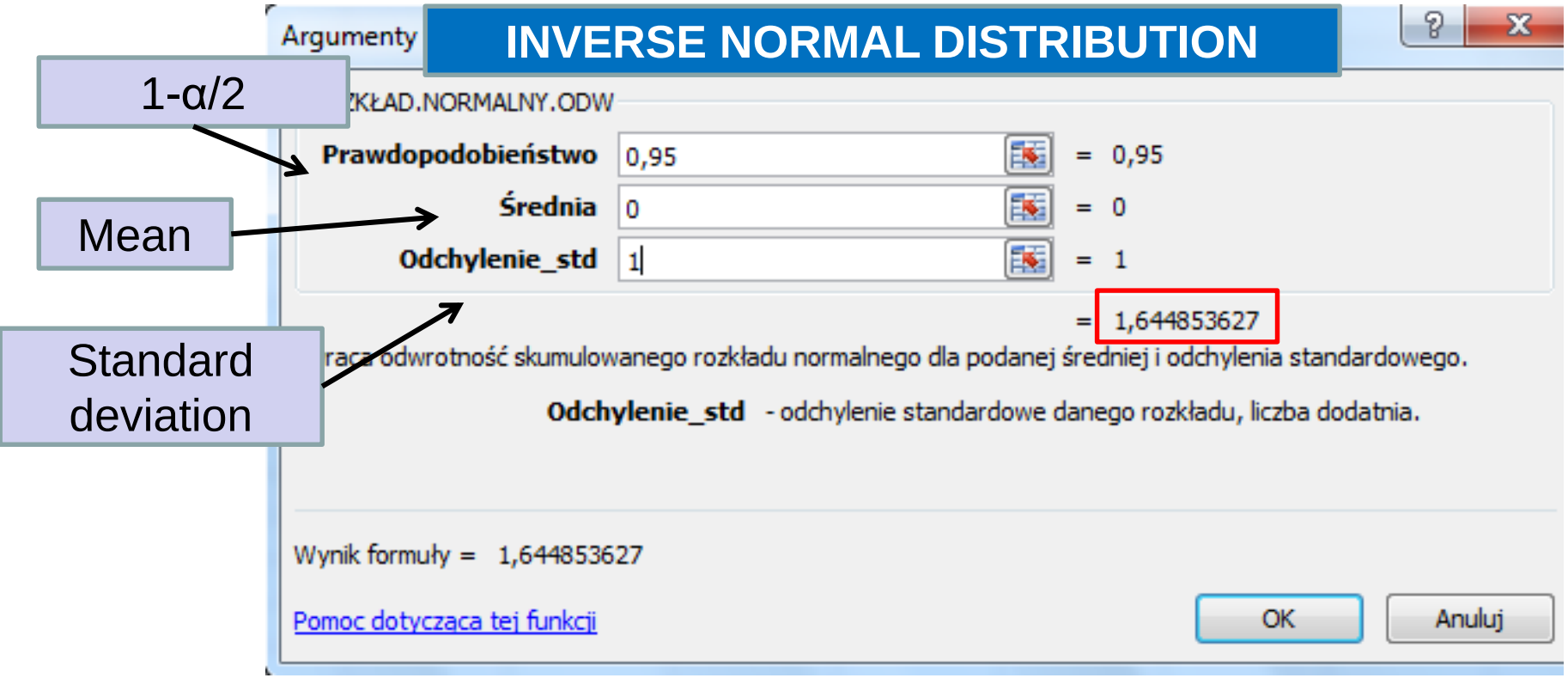
u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	u
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408	1,5
1,6	,94520	,94630	,94738	,94845	,94950	,95053	,95154	,95254	,95352	,95449	1,6
1,7	,95543	,95637	,95728	,95818	,95907	,95994	,96080	,96164	,96246	,96327	1,7
1,8	,96407	,96485	,96562	,96638	,96712	,96784	,96856	,96926	,96995	,97062	1,8
1,9	,97128	,97193	,97257	,97320	,97381	,97441	,97500	,97558	,97615	,97670	1,9
2,0	,97725	,97778	,97831	,97882	,97932	,97982	,98030	,98077	,98124	,98169	2,0
2,1	,98214	,98257	,98300	,98341	,98382	,98422	,98461	,98500	,98537	,98574	2,1
2,2	,98610	,98645	,98679	,98713	,98745	,98778	,98809	,98840	,98870	,98899	2,2
2,3	,98928	,98956	,98983	,9²0097	,9²0358	,9²0613	,9²0863	,9²1106	,9²1344	,9²1576	2,3
2,4	,9²1802	,9²2024	,9²2240	,9²2451	,9²2656	,9²2857	,9²3053	,9²3244	,9²3431	,9²3613	2,4
2,5	,9²3790	,9²3963	,9²4132	,9²4297	,9²4457	,9²4614	,9²4766	,9²4915	,9²5060	,9²5201	2,5
2,6	,9²533					975	,9²6093	,9²6207	,9²6319	,9²6427	2,6
2,7	,9²653					020	,9²7110	,9²7197	,9²7282	,9²7365	2,7
2,8	,9²744					814	,9²7882	,9²7948	,9²8012	,9²8074	2,8
2,9	,9²813					411	,9²8462	,9²8511	,9²8559	,9²8605	2,9
3,0	,9²865					856	,9²8893	,9²8930	,9²8965	,9²8999	3,0
3,1	,9³032					1836	,9³2112	,9³2378	,9³2636	,9³2886	3,1
3,2	,9³3129	,9³3363	,9³3590	,9³3810	,9³4002	,9³4230	,9³4429	,9³4623	,9³4810	,9³4991	3,2
3,3	,9³5166	,9³5335	,9³5499	,9³5658	,9³5811	,9³5959	,9³6103	,9³6242	,9³6376	,9³6505	3,3
3,4	,9³6631	,9³6752	,9³6869	,9³6982	,9³7091	,9³7197	,9³7299	,9³7398	,9³7493	,9³7585	3,4

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

$$P(Z < 1.65) = 0.95$$

How to find $z_{\alpha/2}$?

Using Excel:



Example

Suppose a large bank wants to estimate the average amount of money owed by its delinquent debtors (i.e., debtors who are more than 2 months behind in payment). To accomplish this objective, the bank plans to randomly sample 100 of its delinquent accounts and to use the sample mean, $\bar{x}$, of the amounts overdue to estimate μ, the mean for *all* delinquent accounts.

⇨ see file OVERDUE.xls

Example

Suppose a large bank wants to estimate the average amount of money owed by its delinquent debtors (i.e., debtors who are more than 2 months behind in payment). To accomplish this objective, the bank plans to randomly sample 100 of its delinquent accounts and to use the sample mean, $\bar{x}$, of the amounts overdue to estimate μ, the mean for *all* delinquent accounts.

Solution:

Based on the data we compute $\bar{x} = 223.28$

$S = 90.34$

$$P(\bar{x} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot \sigma / \sqrt{n}) = 1 - \alpha$$

We need to replace σ with S:

$$223.28 - 1.96 \cdot 90.34 / \sqrt{100} < \mu < 223.28 + 1.96 \cdot 90.34 / \sqrt{100}$$

$\mu = 223.28 \pm 17.7$ with confidence 0.95

Small sample confidence interval for population mean

$$T = \frac{\bar{x} - \mu}{S / \sqrt{n}} \sim t - Student$$

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$$

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

Sampling error
SE

$$P(\bar{x} - t_{\alpha/2;n-1} \cdot S / \sqrt{n} < \mu < \bar{x} + t_{\alpha/2;n-1} \cdot S / \sqrt{n}) = 1 - \alpha$$

Confidence interval for population mean

- If a sample was drawn from a normally distributed population with known standard deviation

$$P(\bar{x} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot \sigma / \sqrt{n}) = 1 - \alpha$$

- If a sample was drawn from a normally distributed population with unknown standard deviation
  - Large sample

$$P(\bar{x} - z_{\alpha/2} \cdot S / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot S / \sqrt{n}) = 1 - \alpha$$

  - Small sample

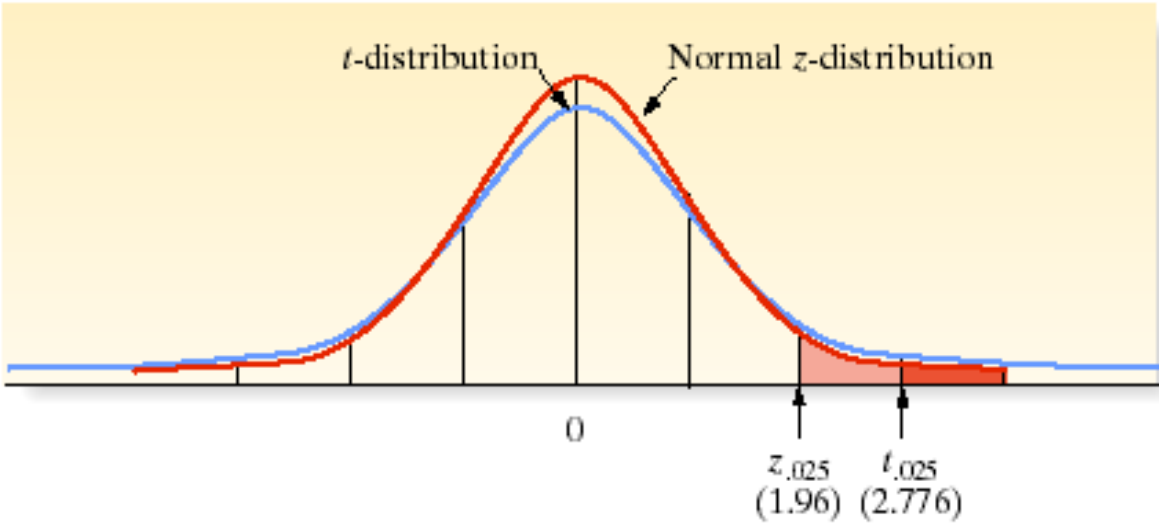
$$P(\bar{x} - t_{\alpha/2;n-1} \cdot S / \sqrt{n} < \mu < \bar{x} + t_{\alpha/2;n-1} \cdot S / \sqrt{n}) = 1 - \alpha$$

- If a sample was drawn from a population with unknown distribution

$$P(\bar{x} - z_{\alpha/2} \cdot S / \sqrt{n} < \mu < \bar{x} + z_{\alpha/2} \cdot S / \sqrt{n}) = 1 - \alpha$$

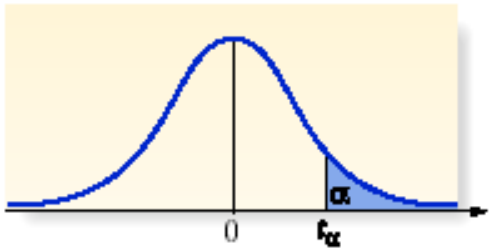
Small-sample confidence interval for population mean

- Student's t-distribution once more



Wider confidence intervals with the same $1 - \alpha$

How to find $t_{\alpha/2}$?



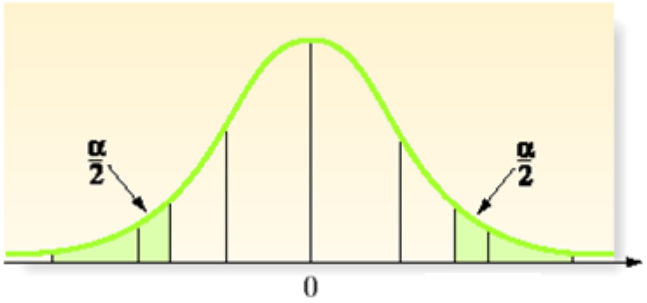
$$1-\alpha = 0.9$$

$$P(T > t_{\alpha/2}) = \alpha/2$$

$$P(T > 2.015) = 0.05$$

Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.13	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	21.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.132	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

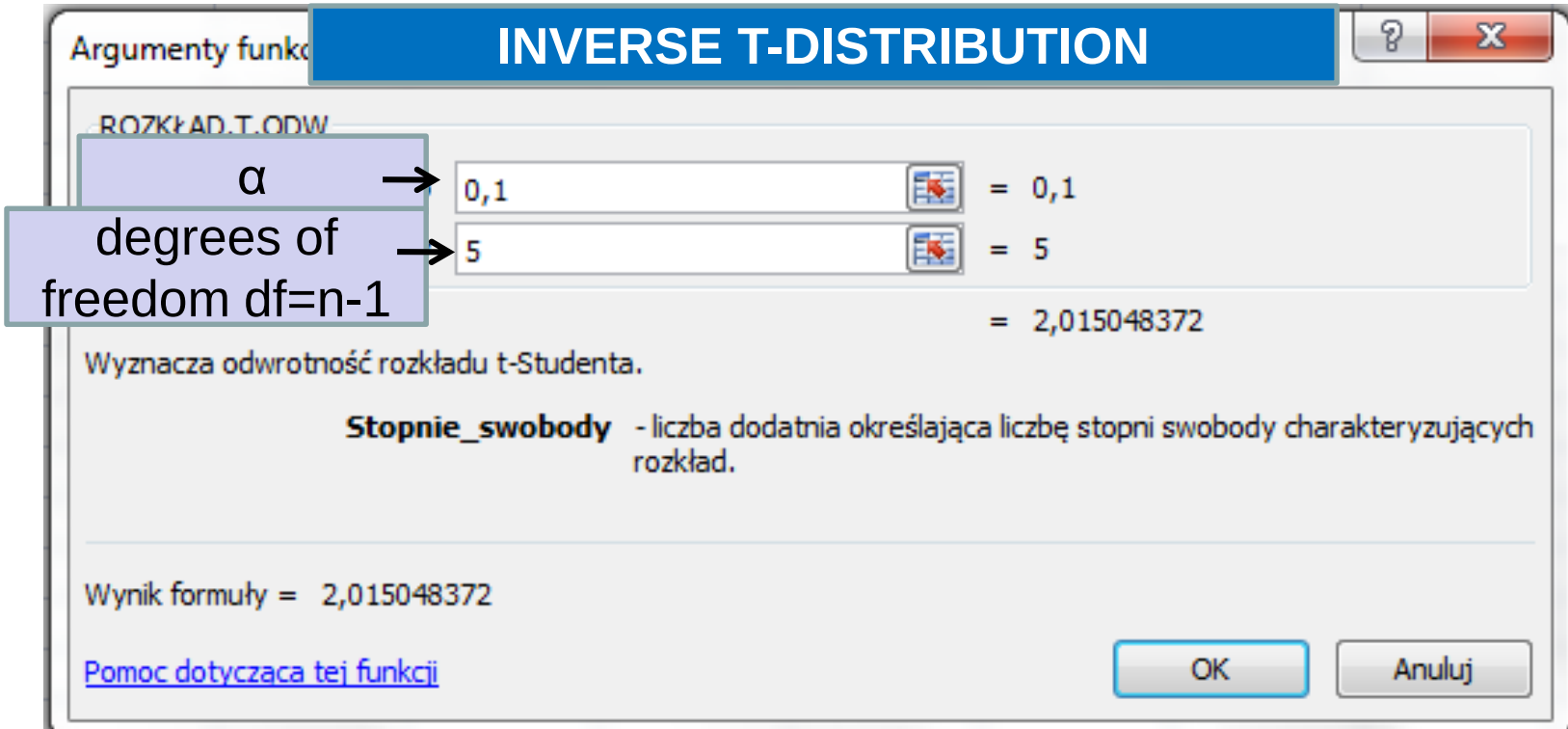
How to find $t_{\alpha/2}$?



$$1-\alpha = 0.9$$

$$P(|T| > t_{\alpha/2}) = \alpha$$

$$P(|T| > t_{\alpha/2}) = P(T > 2.015) + P(T < -2.015) = 0.1$$



Example

Problem Some quality-control experiments require *destructive sampling* in order to measure some particular characteristic of the product. The cost of destructive sampling often dictates small samples. For example, suppose a manufacturer of printers for personal computers wishes to estimate the mean number of characters printed before the printhead fails. Suppose the printer manufacturer tests $n = 15$ randomly selected printheads and records the number of characters printed until failure for each. These 15 measurements (in millions of characters) are listed in Table 5.4.

a. Form a 99% confidence interval for the mean number of characters printed before the printhead fails. Interpret the result.

b. What assumption is required for the interval, part **a,** to be valid? Is it reasonably satisfied?

TABLE 5.4 Number of Characters (in Millions) for $n = 15$ Printhead Tests

1.13	1.55	1.43	.92	1.25
1.36	1.32	.85	1.07	1.48
1.20	1.33	1.18	1.22	1.29

⇨ see file PRINTHEAD.xls

Example

For this small sample ($n = 15$), we use the t-statistic to form the confidence interval.

$$\bar{x} \pm t_{.005}\left(\frac{s}{\sqrt{n}}\right)$$

$t_{\alpha/2} = t_{.005} = 2.977$ for df=15-1=14

$\bar{x} = 1.239$

$s = .193$

$$\bar{x} \pm t_{.005}\left(\frac{s}{\sqrt{n}}\right) = 1.239 \pm 2.977\left(\frac{.193}{\sqrt{15}}\right)$$

$$= 1.239 \pm .148 \text{ or } (1.091, 1.387)$$

Confidence interval for population proportion

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Standard error

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

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

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

Sampling error
SE

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

Large-sample confidence interval for population proportion

Large-Sample Confidence Interval for *p*

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} \approx \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p} = \frac{x}{n}$ and $\hat{q} = 1 - \hat{p}$

Note: When n is large, $\hat{p}$ can approximate the value of p in the formula for $\sigma_{\hat{p}}$.

Conditions required for a Valid Large-Sample Confidence Interval for p:

1. The sample is random
2. The sample is large enough $n \geq 100$ - rule of thumb

Example

Problem Many public polling agencies conduct surveys to determine the current consumer sentiment concerning the state of the economy. For example, the Bureau of Economic and Business Research (BEBR) at the University of Florida conducts quarterly surveys to gauge consumer sentiment in the Sunshine State. Suppose that BEBR randomly samples 484 consumers and finds that 257 are optimistic about the state of the economy. Use a 90% confidence interval to estimate the proportion of all consumers in Florida who are optimistic about the state of the economy. Based on the confidence interval, can BEBR infer that the majority of Florida consumers are optimistic about the economy?

Example

Solution The number, x, of the 484 sampled consumers who are optimistic about the Florida economy is a binomial random variable if we can assume that the sample was randomly selected from the population of Florida consumers and that the poll was conducted identically for each sampled consumer.

The point estimate of the proportion of Florida consumers who are optimistic about the economy is

$$\hat{p} = \frac{x}{n} = \frac{257}{484} = .531$$

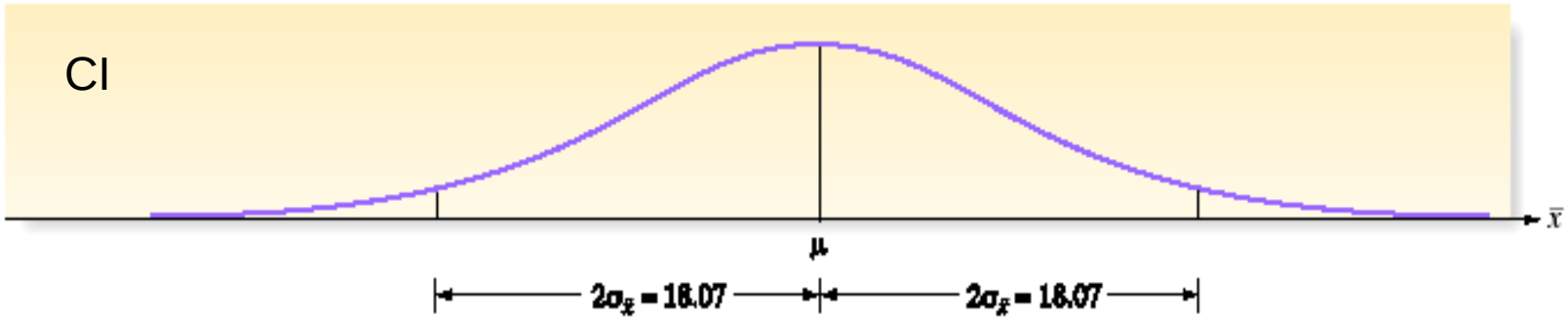
$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} \approx \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= .531 \pm 1.645\sqrt{\frac{(.531)(.469)}{484}} = .531 \pm .037 = (.494, .568)$$

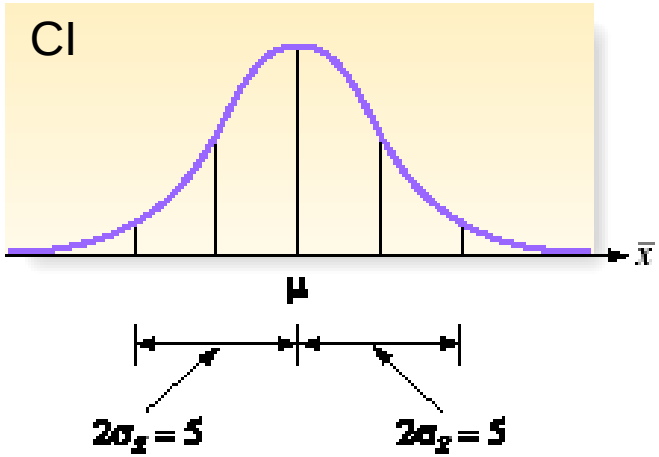
Determining the sample size

Determining the sample size

n=100



n=1306



Question: How many people should we ask to receive the results with sufficiently narrow confidence interval?

Sample size – Mean

Sample Size Determination for 100(1 − α)% Confidence Interval for μ

In order to estimate μ with a sampling error SE and with 100(1 − α)% confidence, the required sample size is found as follows:

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = SE$$

The solution for n is given by the equation

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{(SE)^2}$$

Note: The value of σ is usually unknown. It can be estimated by the standard deviation, s, from a prior sample. Alternatively, we may approximate the range R of observations in the population, and (conservatively) estimate $\sigma \approx R/4$. In any case, you should round the value of n obtained *upward* to ensure that the sample size will be sufficient to achieve the specified reliability.

Sample size - Proportion

Sample Size Determination for 100(1 − α) % Confidence Interval for *p*

In order to estimate a binomial probability p with sampling error SE and with $100(1 - \alpha)$ % confidence, the required sample size is found by solving the following equation for n:

$$z_{\alpha/2}\sqrt{\frac{pq}{n}} = \text{SE}$$

The solution for n can be written as follows:

$$n = \frac{(z_{\alpha/2})^2(pq)}{(\text{SE})^2}$$

Note: Since the value of the product pq is unknown, it can be estimated by using the sample fraction of successes, $\hat{p}$, from a prior sample. Remember (Table 5.6) that the value of pq is at its maximum when p equals .5, so you can obtain conservatively large values of n by approximating p by .5 or values close to .5. In any case, you should round the value of n obtained *upward* to ensure that the sample size will be sufficient to achieve the specified reliability.

Example

Problem A cellular telephone manufacturer that entered the postregulation market too quickly has an initial problem with excessive customer complaints and consequent returns of the cell phones for repair or replacement. The manufacturer wants to determine the magnitude of the problem in order to estimate its warranty liability. How many cellular telephones should the company randomly sample from its warehouse and check in order to estimate the fraction defective, p, to within .01 with 90% confidence?

Example

Problem A cellular telephone manufacturer that entered the postregulation market too quickly has an initial problem with excessive customer complaints and consequent returns of the cell phones for repair or replacement. The manufacturer wants to determine the magnitude of the problem in order to estimate its warranty liability. How many cellular telephones should the company randomly sample from its warehouse and check in order to estimate the fraction defective, p, to within .01 with 90% confidence?

Solution:

$$n = \frac{(z_{\alpha/2})^2(pq)}{(\text{SE})^2}$$

The equation for the sample size n requires an estimate of the product pq. We could most conservatively estimate $pq = .25$ (i.e., use $p = .5$), but this may be overly conservative when estimating a fraction defective. A value of .1, corresponding to 10% defective, will probably be conservatively large for this application. The solution is therefore

$$n = \frac{(z_{\alpha/2})^2(pq)}{(\text{SE})^2} = \frac{(1.645)^2(.1)(.9)}{(.01)^2} = 2{,}435.4 \approx 2{,}436$$

Thus, the manufacturer should sample 2,436 cellular telephones in order to estimate the fraction defective, p, to within .01 with 90% confidence.