L10

# Basic Statistics for Economists

Spring 2020

Department of Statistics

Stockholm University

# Today

**NCT sections 7.4 + 8.1-8.3**

- Continue on **confidence intervals** (**CI**)
  - Interpretation of the concept of confidence
  - CI for a proportion $\boldsymbol{P}$ with the estimate $\widehat{\boldsymbol{p}}$
- **Compare differences, changes**
  - CI for the average of **pairwise** differences $\boldsymbol{D_i = X_i - Y_i}$ and $\overline{\boldsymbol{D}}$
  - CI for differences between independent averages: $\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}}$
  - CI for differences between independent proportions: $\widehat{\boldsymbol{p}}_{\boldsymbol{x}} - \widehat{\boldsymbol{p}}_{\boldsymbol{y}}$

# Summary from F9

## Estimators

- Functions of the values of the samples ⇒ random variables
- **Properties**: unbiasedness, consistency, efficiency
- **Point estimates** of the population mean $\mu$
- **Interval estimation** / **confidence interval** of the mean $\mu$
  - Interval of uncertainty around the point estimate
  - Point estimate ± uncertainty
- **Two cases**
  1. Known variance
  2. Unknown variance

# Confidence interval of the mean $\mu$

$X_i$ iid $N(\mu, \sigma^2)$ where $\sigma^2$ is **known**

- Use table 2 to find the value of $z_{\alpha/2}$
- All that is calculated from the sample is $\bar{x}$

$$\boxed{\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

$X_i$ iid $N(\mu, \sigma^2)$ where $\sigma^2$ is **unknown**

- Find the **degrees of freedom** $\nu = n - 1$
- Use table 3 to find the value of $t_{n-1;\alpha/2}$
- Using the sample, we calculate both $\bar{x}$ and $s^2$

$$\boxed{\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}}$$

It is easy to see that $t_{n-1;\alpha/2} > z_{\alpha/2}$ which makes CI wider



Stockholm University

# Terminology

- Point estimate, $(\boldsymbol{PE})$
- Margin of error, $\boldsymbol{ME}$
- Standard error, $\boldsymbol{SE}$

$$\underbrace{\bar{x}}_{\text{Point estimate}} \pm \underbrace{z_{\alpha/2} \cdot \overbrace{\frac{\sigma_X}{\sqrt{n}}}^{\text{Standard error}}}_{\text{Margin of error for given } \alpha} \qquad \bar{x} \pm t_{n-1;\alpha/2} \cdot \frac{s_x}{\sqrt{n}}$$

- $\boldsymbol{LCL}$ = lower confidence limit = $\boldsymbol{PE} - \boldsymbol{ME}$
- $\boldsymbol{UCL}$ = upper confidence limit = $\boldsymbol{PE} + \boldsymbol{ME}$

$$\boxed{\boldsymbol{CI} = (\boldsymbol{LCL}, \boldsymbol{UCL})}$$

- Total length of the CI: $(\boldsymbol{UCL} - \boldsymbol{LCL}) = 2 \cdot \boldsymbol{ME}$
- Point estimate = midpoint: $(\boldsymbol{UCL} + \boldsymbol{LCL})/2 = \boldsymbol{PE}$



Stockholm University

# $t$-distribution

- Similar to the standardized normal distribution $\boldsymbol{N(0,1)}$:
  - symmetrical, bell shaped, expected value $= 0$, variance $\to 1$
  - ”heavier tails” i.e. extreme values (left and right) are more likely compared to the distribution $\boldsymbol{N(0,1)}$.
- **The degrees of freedom $\boldsymbol{\nu}$** determine how ”heavy” the tails are:
  - Few degrees of freedom, heavier tails
  - Many degrees of freedom, lighter tails

See L9 p. 29

- for confidence interval of $\boldsymbol{\mu}$ of a sample, the following holds:

$$\boxed{\textbf{degrees of freedom} = \boldsymbol{\nu} = \boldsymbol{n} - \mathbf{1}}$$

- (*sv. frihetsgrader*)

# Exercise 1

Assumptions: $X_i$ iid $N(\mu, \sigma^2)$ where $\boxed{\sigma^2 = 4}$

We have the following data from a sample of size $n = 6$:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X_i$ | 7.4 | 5.2 | 4.4 | 6.7 | 1.1 | 3.4 |

Make a point estimate and a 95 % CI for the mean $\mu$.

- The sample mean is $\bar{x} = 4.7$. This is our point estimate.
- $95\ \%$ confidence $\longrightarrow$ $\alpha/2 = 0.025$ and table 2 $\longrightarrow$ $z_{0.025} = 1.96$
- Calculation: $\boxed{\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}} \longrightarrow 4.7 \pm 1.96 \cdot \frac{2}{\sqrt{6}} \longrightarrow \boxed{4.7 \pm 1.60}$

$CI = (3.1, 6.3)$



Stockholm University

# Exercise 2

Assumption: $X_i$ iid $N(\mu, \sigma^2)$

…but pretend $\boxed{\sigma^2 \text{ is unknown}}$ and must be **estimated with $s^2$**.

Use the same data as in exercise 1 and give a point estimate and a 95% CI for the mean $\mu$.

- The sample mean was calculated as $\bar{x} = 4.7$
- The sample variance is $s^2 = 5.256$ and $s = \sqrt{5.256} = 2.293$
- The sample size $n = 6$ means $\nu = n - 1 = 5$ **degrees of freedom**
- $95\ \%$ confidence $\longrightarrow$ $\alpha/2 = 0.025$; table 3 $\longrightarrow$ $t_{5;0.025} = 2.571$
- Calculate: $\boxed{\bar{x} \pm t_{5;0.025} \dfrac{s}{\sqrt{n}}} \longrightarrow 4.7 \pm 2.571 \cdot \dfrac{2.293}{\sqrt{6}} \longrightarrow \boxed{4.7 \pm 2.41}$

$CI = (2.29, 7.11)$

# <u>Not</u> normal distribution, small sample

Two cases: known or unknown variance:

- **Not covered in this course!**

# Summary – which distribution?

- When should you use $z$ and when should you use $t$?

| Sample size $n$ | distribution | Variance $\sigma^2$ | |
|---|---|---|---|
| | | known | unknown, use $s^2$ |
| large | normal | $z_{\alpha/2}$ | $z_{\alpha/2}$ |
| | not normal | **CLT** $\Rightarrow z_{\alpha/2}$ | **CLT** $\Rightarrow z_{\alpha/2}$ |
| small | normal | $z_{\alpha/2}$ | $t_{n-1,\alpha/2}$ |
| | not normal | **Not covered** | **Not covered** |

- In the cases marked blue, large sample and unknown variance, one can argue that $t$ should be used instead of $z$ – this renders more conservative estimates, i.e. slightly wider CI = greater margin of error; one then accounts for the increased uncertainty which stems from estimation of the variance.



Stockholm University

# Interpretation of confidence interval

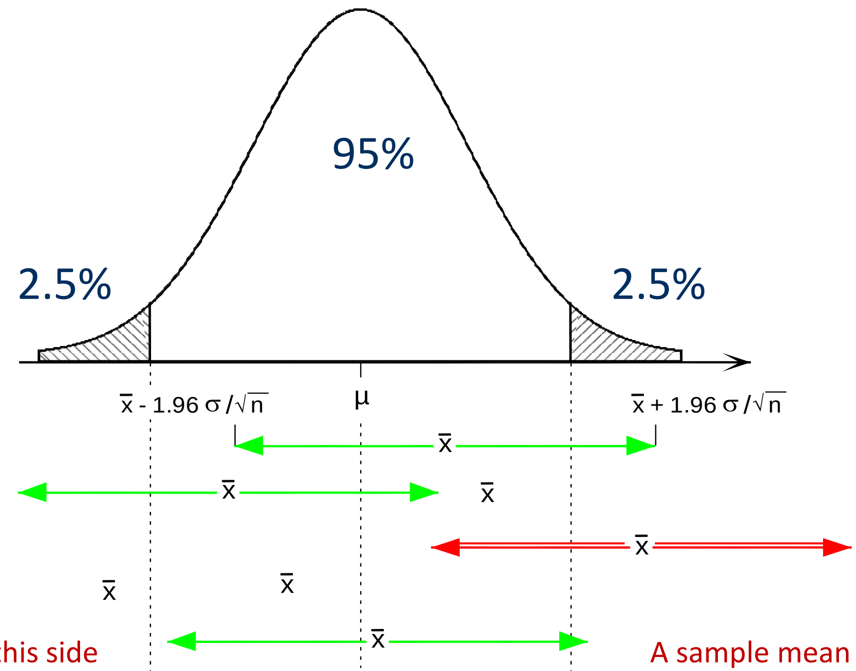
- Suppose that the procedure/experiment could be **repeated many times**. Then the proportion of CI's (different each time) include the true parameter value would approach 95 %.
  - **Confidence level describes a property of multiple experiments, not just one**.
  - There is a 95 % probability that a CI calculated from a sample **drawn in the future** will capture the true parameter value.
  - This is a statement about a **future CI** (that is random), not about the parameter (the parameter is a constant, not random).
  - With the future experiment interpretation, we avoid the idea of repeated experiments, something that often is impossible (i.e. a particular football game this Saturday).



Stockholm University

# Interpretation of confidence interval, cont.

- 95% of all (future) intervals cover $\mu$; 5% do not.




$$\boxed{\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}}$$

$$\bar{x} \pm t_{n-1;0.025} \cdot s/\sqrt{n}$$

A sample mean on this side will yield a CI that <u>doesn't</u> include the true value of μ

A sample mean here will yield a CI that includes the true value of μ

A sample mean on this side will yield a CI that <u>doesn't</u> include the true value of μ

Stockholm University

# Interpretation of confidence interval, cont.

- The confidence level 95 % **cannot** be interpreted as “the true parameter lies within the interval with 95 % probability.” Once the experiment is completed, the parameter value either lies within the CI, or not; it is **no longer** a **random** event with probability, **only uncertainty**.

- Sometimes people think that a 95 % CI captures 95% of the population, on average. This is **not true**!

- Sometimes CI is interpreted to mean that the true parameter value lies in this interval. This is **not** correct! The interval is an extension of a point estimate to a whole set of estimates of the parameter (and we loosely say that we are 95 % confident in these estimates).

# Bernoulli distribution and proportions $P$

Assume

$$X_i = \begin{cases} 0, & 1-P \\ 1, & P \end{cases} \qquad X_i \sim Bernoulli(P) \qquad X_i \sim Bin(1, P)$$

$P$ is an unknown parameter

$$E(X_i) = \mu_X = P \qquad Var(X_i) = \sigma_X^2 = P(1-P)$$

A sample of size $n$; define: $\hat{p} = [\text{proportion of } 1's] = \frac{\sum X_i}{n} = \bar{X}$

- Expected value: $E(\hat{p}) = E(\bar{X}) = E(X_i) = \mu_X = P$ Unbiased estimator for $P$
- Variance: $Var(\hat{p}) = Var(\bar{X}) = \frac{Var(X_i)}{n} = \frac{\sigma_X^2}{n} = \frac{P(1-P)}{n}$



Stockholm University

# CI for proportion $P$

If $X_i$ **iid** $\cancel{N}(\mu, \sigma^2)$ it still follows from **CLT** that $\bar{X} \to N\left(\mu_X, \frac{\sigma_x^2}{n}\right)$ as $n \to \infty$.

- $X_i$ is Bernoulli (not normal); $\hat{p} = \sum X_i / n = \bar{X}$ is the sample mean and it is an estimator of $P$, the true proportion.
- Expected value of $\hat{p}$ is $P$ and variance of $\hat{p}$ is $P(1 - P)/n$.
- According to **CLT**, $\hat{p} \to N\left(P, \frac{P(1-P)}{n}\right)$
- <u>Problem</u>: in order to know the variance $P(1 - P)/n$ we need $P$.
- <u>Solution</u>: use $\hat{p}$ as an approximation.

# CI for a proportion $P$, cont.

- According to **CLT**, we get an **approximate** $\mathbf{100(1-\alpha)}$ % CI of the population proportion $\boldsymbol{P}$ from

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

- This is approximate in the sense that the variance is estimated and CLT, and hence the true confidence level is unknown.

- The earlier rule of thumb $\boldsymbol{nP(1-P) > 5}$ does not work since $\boldsymbol{P}$ is unknown. **New rule of thumb:** $\boldsymbol{n \geq 30}$. Always strive for large samples! Perhaps even $\boldsymbol{n \geq 100}$.



Stockholm University

# Exercise 3

Suppose you want to estimate the proportion of invoices which are not paid before the due date. You draw a random sample of $\boldsymbol{n = 144}$ invoices and find that $\mathbf{54}$ of these were not paid on time. Estimate the probability $\boldsymbol{P}$ that a randomly chosen invoice is not paid on time and calculate a 95 % CI.

- $\boldsymbol{n = 144}$ should be large enough (CLT), normal approximation!
- Point estimate:

$$\hat{p} = \frac{54}{144} = 0.375 \qquad (37.5\%)$$

**Rule of thumb:**
$\boldsymbol{n = 144 > 30}$

- 95 % CI:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.375 \pm 1.96\sqrt{\frac{0.375 \cdot 0.625}{144}} = 0.375 \pm 0.079$$

$$(29.6\% \,;\, 45.4\%)$$



Stockholm University

# Repeated measurements, same object

- It is common to study objects before and after a “treatment”.
  - may be two different treatments, but on the same objects.
- You have **two measurements per object**.
- These **pairwise measurements** are generally **not independent**.
- But the objects are often independent of each other.
- The question is if there are differences before and after, or between treatments. What should we do?

# Pairwise differences

- Matched pairs of observations: $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$ for object $\boldsymbol{i}$
- The sample consists of $\boldsymbol{n}$ **pairs**: $(\boldsymbol{X_1}, \boldsymbol{Y_1}), (\boldsymbol{X_2}, \boldsymbol{Y_2}), \dots, (\boldsymbol{X_n}, \boldsymbol{Y_n})$
- Calculate the **difference** within each pair $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$

$$\boldsymbol{D_i = X_i - Y_i}$$

- If we can assume that each $\boldsymbol{D_i}$ is $\mathbf{iid}\ \boldsymbol{N\left(\mu_D, \sigma_D^2\right)}$, we can calculate a confidence interval for the **average difference** $\boldsymbol{\mu_D}$.
- Often the sample is small and the variance unknown:

$\Rightarrow$ $\boldsymbol{t}$**-distribution!**

# CI for the mean of pairwise differences

- Same construction as for the mean of $X_i$ iid $N(\mu, \sigma^2)$ when $\sigma^2$ is unknown
- Now:

  $\boxed{X_i - Y_i = D_i}$ and $\boxed{D_i \text{ iid } N\left(\mu_D, \sigma_D^2\right)}$ where $\sigma_D^2$ is unknown

- A $100(1-\alpha)\%$ CI for $\mu_D$ is given by

$$\bar{d} \pm t_{n-1;\alpha/2} \frac{s_d}{\sqrt{n}}$$

  where $\bar{d}$ and $s_d$ are the sample mean and sample variance of the observed differences $d_i = x_i - y_i$



Stockholm University

# Exercise 3

- A government agency has two departments, $x$ and $y$, for processing cases. You want to investigate if there is a difference in the case processing times between the two departments. To that end, you sent the same eight cases, denoted 1, 2,..., 8 to each department. The processing time was measured for each case and department, $(x_1, y_1), \ldots, (x_8, y_8)$. The results are shown in the table below.

| Case | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Dept. x | 10 | 18 | 9 | 27 | 8 | 10 | 8 | 12 |
| Dept. y | 9 | 20 | 9 | 29 | 7 | 12 | 7 | 12 |

- Construct a 95 % confidence interval for the average difference in processing time between the two departments.

# Exercise 3, solution

- Calculate the difference $\boldsymbol{d_i = x_i - y_i}$ for each case $\boldsymbol{i}$ and then the sample mean and variance:

| Dept. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Dept. x | 10 | 18 | 9 | 27 | 8 | 10 | 8 | 12 |
| Dept. y | 9 | 20 | 9 | 29 | 7 | 12 | 7 | 12 |
| $\boldsymbol{d_i}$ | **1** | **-2** | **0** | **-2** | **1** | **-2** | **1** | **0** |

$\boldsymbol{\bar{d} = -0.375}$ $\boldsymbol{s_d^2 = 1.9821}$ $\boldsymbol{s_d = 1.408}$

Degrees of freedom $\boldsymbol{= n - 1 = 7}$ $\boldsymbol{t_{7;0.025}} = [\text{Table 3}] = \boldsymbol{2.365}$

- Calculate: $$\boldsymbol{\bar{d} \pm t_{n-1;\alpha/2} \frac{s_d}{\sqrt{n}} = -0.375 \pm 2.365 \cdot \frac{1.408}{\sqrt{8}} = -0.375 \pm 1.177}$$

$$\boldsymbol{(-1.55\,;\ 0.80)}$$

# To compare groups

- Suppose that we have two well-defined groups (populations) which we can call $\boldsymbol{x}$ and $\boldsymbol{y}$.
  - EU compared to BRIC
  - Can also be two different times, the population of Sweden on some issue 2017 compared to 2016.

- We want to compare the properties of $\boldsymbol{x}$ and $\boldsymbol{y}$.
  - means $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ or proportions $\boldsymbol{P_X}$ and $\boldsymbol{P_Y}$.

- We want to estimate the difference $\boldsymbol{\mu_x - \mu_y}$ or $\boldsymbol{P_x - P_y}$ and form a confidence interval for this difference.

# To compare groups, cont.

- **Two samples**, one from each population $\boldsymbol{x}$ and $\boldsymbol{y}$ of sizes $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$ respectively. We estimate $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ using $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$.
- There is **independence** between the samples and **iid** within samples
- We estimate the difference $\boldsymbol{\mu_x - \mu_y}$ with the estimates: $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$
- The difference is a **linear combination** of two r.v.:
  - Expected value: $\boldsymbol{E(\bar{X} - \bar{Y}) = \mu_x - \mu_y}$ **Unbiased estimator**
  - Variance: $\boldsymbol{Var(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$ **Independent – so no covariance terms**
  - Distribution: depends!


Stockholm University

# To compare groups, cont.

- If each of the populations are **normally distributed**, then the difference $\bar{X} - \bar{Y}$ is **normally distributed** (linear combination)
- If the populations are <u>not</u> normally distributed, we use **CLT**;
  if each of the sample sizes are big enough, it follows that $\bar{X} - \bar{Y}$ is **approximatively normally distributed**.

Just as in the single variable case, we need to consider:

- whether the variances are known or unknown
- whether the samples are large or small



Stockholm University

# Different cases of CI for $\mu_x - \mu_y$

- When should you use $z$ and when should you use $t$?

| Sample size $n_x$, $n_y$ | Distribution | Variances $\sigma_x^2$ and $\sigma_y^2$ | |
|---|---|---|---|
| | | known | unknown; $s_x^2$, $s_y^2$ |
| large | normal | $z_{\alpha/2}$ | **approx.** $z_{\alpha/2}$ |
| | not normal | **CLT** $\Rightarrow z_{\alpha/2}$ | **CLT** $\Rightarrow z_{\alpha/2}$ |
| small | normal | $z_{\alpha/2}$ | **Two cases, of which one is not covered** |
| | not normal | **Not covered** | **Not covered** |

← !!

- In the cases marked blue, large sample and unknown variance, one can argue that $t$ should be used instead of $z$ – this renders more conservative estimates, i.e. slightly wider CI = greater margin of error; one then accounts for the increased uncertainty which stems from estimation of the variance.



Stockholm University

# Comparing two proportions

- **Two samples**, one from each of two populations $\boldsymbol{x}$ and $\boldsymbol{y}$ of sizes $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$, respectively. We estimate $\boldsymbol{P_x}$ and $\boldsymbol{P_y}$ using $\boldsymbol{\hat{p}_x}$ and $\boldsymbol{\hat{p}_y}$.

- The samples need to be **independent of each other** and **iid** within.

- We estimate the difference $\boldsymbol{P_x - P_y}$ with the difference $\boldsymbol{\hat{p}_x - \hat{p}_y}$

- The difference is a **linear combination** of two independent r.v.
  - Expected value: $\boldsymbol{E(\hat{p}_x - \hat{p}_y) = P_x - P_y}$ **Unbiased estimator**
  - Variance: $\boldsymbol{Var(\hat{p}_x - \hat{p}_y) = \frac{P_x(1-P_x)}{n_x} + \frac{P_y(1-P_y)}{n_y}}$ **Independent – so no covariance terms**
  - Distribution: If the samples are large enough - **CLT**



Stockholm University

# CI for difference between proportions

- Same problem as before, to know the variances, we need to know $\boldsymbol{P_x}$ and $\boldsymbol{P_y}$. Same solution, use $\boldsymbol{\hat{p}_x}$ and $\boldsymbol{\hat{p}_y}$.

- According to **CLT** a $\mathbf{100(1-\alpha)}\ \%$ CI of the difference in population proportions $\boldsymbol{P_x - P_y}$ is **approximately** given by

$$\left(\hat{p}_x - \hat{p}_y\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

- This is approximate in the sense that one cannot be sure of the true confidence level.

# Exercise 4

A group of physicians studies famine related disease in two famine stricken areas of Africa. Two samples of preschool aged children, one from area A and one from area B, rendered the following results: $\mathbf{520}$ out of $\boldsymbol{n_A} = \mathbf{1300}$ children in area A and $\mathbf{385}$ out of $\boldsymbol{n_B} = \mathbf{1100}$ children in area B were found to suffer from chronical malnutrition.

a) Give a 90 % confidence interval for the proportion $\boldsymbol{P_A}$ that suffers from chronical malnutrition in area A. State your assumptions clearly.

b) Give a 90 % confidence interval of the **difference** between the proportions, i.e. $\boldsymbol{P_A - P_B}$, between the two areas. State your assumptions clearly.

# Exercise 4, solution

a) The point estimate of $p_A$ is $\hat{p}_A = 520/1300 = 0.4$. Since $n_A = 1300$ is large enough the distribution of the point estimate $\hat{p}_A$ be approximated with a normal distribution, according to **CLT**. Observations are **iid**.

Confidence level $90\% \Rightarrow \alpha = 0.10$ and $\alpha/2 = 0.05$ and $z_{\alpha/2} = 1.6449$

Rule of thumb: $n\hat{p}_A(1 - \hat{p}_A) = 312 > 5$; **Ok!**

A $90\%$ confidence interval for $p_A$ is given by:

$$\hat{p}_A \pm z_{0,05}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n}} = 0.4 \pm 1.6449 \cdot \sqrt{\frac{0.4 \cdot 0.6}{1300}} = 0.4 \pm 0.022$$

$$= 40\% \pm 2.2\%$$

$$(37.8\% \,;\, 42.2\%)$$



Stockholm University

# Exercise 4, solution

b) The point estimate of $P_A$ and $P_B$ are $\hat{p}_A = 0.4$ and $\hat{p}_B = 0.35$. Both $n_A$ and $n_B$ are large enough that we may approximate the distribution of the difference $\hat{p}_A - \hat{p}_B$ using a normal distribution, according to **CLT**.

A $90\%$ confidence interval for $p_A - p_B$ is given by

$$(\hat{p}_A - \hat{p}_B) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

$$= (0.40 - 0.35) \pm 1.6449\sqrt{\frac{0.4 \cdot 0.6}{1300} + \frac{0.35 \cdot 0.65}{1100}}$$

$$= 0.05 \pm 1.6449 \cdot 0.019785 \quad = \quad 0.05 \pm 0.033 \qquad (1,7\% \,;\, 8,3\%)$$