UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: **Ivo Dinov,**

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Fred Phoa, Kirsten Johnson, Ming Zheng & Matilda Hsieh

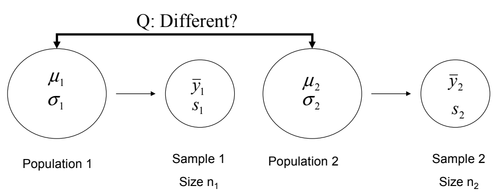
University of California, Los Angeles, Fall 2005

http://www.stat.ucla.edu/~dinov/courses_students.html

Comparison of Two Independent Samples

Comparison of Two Independent Samples

- Many times in the sciences it is useful to compare two groups
  - Male vs. Female
  - Drug vs. Placebo
  - NC vs. Disease



Comparison of Two Independent Samples

- Two Approaches for Comparison
  - Confidence Intervals
    - we already know something about CI's
  - Hypothesis Testing
    - this will be new
- What seems like a reasonable way to compare two groups?
- What parameter are we trying to estimate?

Comparison of Two Independent Samples

- RECALL: The sampling distribution of $\bar{y}$ was centered at μ, and had a standard deviation of $\sigma / \sqrt{n}$
- We'll start by describing the sampling distribution of $\bar{y}_1 - \bar{y}_2$
  - Mean: $\mu_1 - \mu_2$
  - Standard deviation of $\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$
- What seems like appropriate estimates for these quantities?

Standard Error of $\bar{y}_1 - \bar{y}_2$

- We know $\bar{y}_1 - \bar{y}_2$ estimates $\mu_1 - \mu_2$
- What we need to describe next is the precision of our estimate, $SE_{(\bar{y}_1 - \bar{y}_2)}$

$$SE_{(\bar{y}_1 - \bar{y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{SE_1^2 + SE_2^2}$$

Standard Error of $\bar{y}_1 - \bar{y}_2$

Example: A study is conducted to quantify the benefits of a new cholesterol lowering medication. Two groups of subjects are compared, those who took the medication twice a day for 3 years, and those who took a placebo. Assume subjects were randomly assigned to either group and that both groups data are normally distributed. Results from the study are shown below:

	Medication	Placebo
$\bar{y}$	209.8	224.3
n	10	10
s	44.3	46.2
SE	14.0	14.6

Standard Error of $\bar{y}_1 - \bar{y}_2$

Example: Cholesterol medicine (cont')

(e.g., ftp://ftp.nist.gov/pub/dataplot/other/reference/CHOLEST2.DAT)

Calculate an estimate of the true mean difference between treatment groups and this estimate's precision.

- First, denote medication as group 1 and placebo as group 2

$$(\bar{y}_1 - \bar{y}_2) = 209.8 - 224.3 = -14.5$$

	Medication	Placebo
$\bar{y}$	209.8	224.3
n	10	10
s	44.3	46.2
SE	14.0	14.6

$$SE_{(\bar{y}_1 - \bar{y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{44.3^2}{10} + \frac{46.2^2}{10}} = 20.24$$

Pooled vs. Unpooled

- $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is know as an ***unpooled*** version of the standard error
  - there is also a "pooled" SE
- First we describe a "pooled" variance, which can be thought of as a weighted average of s_1^2 and s_2^2

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Pooled vs. Unpooled

- Then we use the pooled variance to calculate the pooled version of the standard error

$$SE_{pooled} = \sqrt{s^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- NOTE: If ($n_1 = n_2$) and ($s_1 = s_2$) the pooled and unpooled will give the same answer for $SE_{(\bar{y}_1 - \bar{y}_2)}$
- It is when $n_1 \neq n_2$ that we need to decide whether to use pooled or unpooled:
  - if $s_1 = s_2$ then use pooled (unpooled will give similar answer)
  - if $s_1 \neq s_2$ then use unpooled (pooled will NOT give similar answer)

Pooled vs. Unpooled

- RESULT: Because both methods are similar when s_1=s_2 and n_1=n_2, and the pooled version is not valid when
- Why all the torture? This will come up again in chapter 11.
- Because the ***df*** increases a great deal when we do pool the variance.

CI for μ_1 - μ_2

- RECALL: We described a CI earlier as:
  the estimate ± (an appropriate multiplier)x(SE)
- *A 100(1- α)% confidence interval for μ_1 - μ_2* (p.227)

$$(\bar{y}_1 - \bar{y}_2) \pm t(df)_{\alpha/2}\left(SE_{\bar{y}_1 - \bar{y}_2}\right)$$

where df = $\dfrac{\left(SE_1^2 + SE_2^2\right)^2}{SE_1^4/(n_1 - 1) + SE_2^4/(n_2 - 1)}$

CI for μ_1 - μ_2

Example: Cholesterol medication (cont')

Calculate a 95% confidence interval for μ_1 - μ_2

We know $\bar{y}_1 - \bar{y}_2$ and $SE_{(\bar{y}_1 - \bar{y}_2)}$ from the previous slides.
Now we need to find the t multiplier

$$df = \frac{\left(14^2 + 14.6^2\right)^2}{\frac{14^4}{(10-1)} + \frac{14.6^4}{(10-1)}} = \frac{167411.9056}{9317.021} = 17.97 \approx 17$$

Round down to be conservative

- NOTE: Calculating that df is not really that fun, a quick rule of thumb for checking your work is:
  $n_1 + n_2 - 2$

CI for μ_1 - μ_2

$$\begin{aligned}
&(\bar{y}_1 - \bar{y}_2) \pm t(df)_{\alpha/2}\left(SE_{\bar{y}_1 - \bar{y}_2}\right) \\
&= -14.5 \pm t(17)_{0.025}(20.24) \\
&= -14.5 \pm 2.110(20.24) \\
&= (-57.21, 28.21)
\end{aligned}$$

CONCLUSION: We are highly confident at the 0.05 level, that the true mean difference in cholesterol between the medication and placebo groups is between -57.02 and 28.02 mg/dL.

Note the change in the conclusion of the parameter that we are estimating. Still looking for the 5 basic parts of a CI conclusion (see slide 38 of lecture set 5).

CI for μ_1 - μ_2

- What's so great about this type of confidence interval?
- In the previous example our CI contained zero
  - This interval isn't telling us much because:
    - the true mean difference could be more than zero (in which case the mean of group 1 is larger than the mean of group 2)
    - or the true mean difference could be less than zero (in which case the mean of group 1 is smaller than the mean of group 2)
    - or the true mean difference could even be zero!
  - The ZERO RULE!
  - Suppose the CI came out to be (5.2, 28.1), would this indicate a true mean difference?

Hypothesis Testing: The independent t test

- The idea of a hypothesis test is to formulate a hypothesis that nothing is going on and then to see if collected data is consistent with this hypothesis (or if the data shows something different)
  - Like innocent until proven guilty
- There are four main parts to a hypothesis test:
  - hypotheses
  - test statistic
  - p-value
  - conclusion

Hypothesis Testing: #1 The Hypotheses

- There are two hypotheses:
  - Null hypothesis (aka the "status quo" hypothesis)
    - denoted by H_o
  - Alternative hypothesis (aka the research hypothesis)
    - denoted by H_a

Hypothesis Testing: #1 The Hypotheses

- If we are comparing two group means nothing going on would imply no difference
  - the means are "the same"
    $(\mu_1 - \mu_2) = 0$
- For the independent t-test the hypotheses are:
  H_o: $(\mu_1 - \mu_2) = 0$
  (no statistical difference in the population means)
  H_a: $(\mu_1 - \mu_2) \neq 0$
  (a statistical difference in the population means)

Hypothesis Testing: #1 The Hypotheses

Example: Cholesterol medication (cont')

Suppose we want to carry out a hypothesis test to see if the data show that there is enough evidence to support a difference in treatment means.

Find the appropriate null and alternative hypotheses.

H_o: $(\mu_1 - \mu_2) = 0$

(no statistical difference the true means of the medication and placebo groups)

H_a: $(\mu_1 - \mu_2) \neq 0$

(a statistical difference in the true means of the medication and placebo groups, medication has an effect on cholesterol)

Hypothesis Testing: #2 Test Statistic

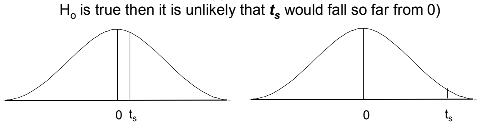
- A test statistic is calculated from the sample data
  - it measures the "disagreement" between the data and the null hypothesis
    - if there is a lot of "disagreement" then we would think that the data provide evidence that the null hypothesis is false
    - if there is little to no "disagreement" then we would think that the data do not provide evidence that the null hypothesis is false

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE_{\bar{y}_1 - \bar{y}_2}}$$

subtract 0 because the null says the difference is zero

Hypothesis Testing: #2 Test Statistic

- On a t distribution t_s could fall anywhere
  - If the test statistic is close to 0, this shows that the data are compatible with H_o (no difference)
    - the deviation can be attributed to chance
  - If the test statistic is far from 0 (in the tails, upper or lower), this shows that the data are <u>in</u>compatible to H_o (there is a difference)
    - deviation does not appear to be attributed to chance (ie. If H_o is true then it is unlikely that $\boldsymbol{t_s}$ would fall so far from 0)



Hypothesis Testing: #2 Test Statistic

Example: Cholesterol medication (cont')

Calculate the test statistic

$$t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE_{\bar{y}_1 - \bar{y}_2}} = \frac{(209.8 - 224.3) - 0}{20.24} = -0.716$$

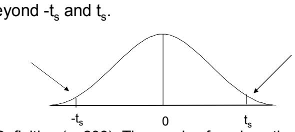
- Great, what does this mean?
  - $\bar{y}_1$ and $\bar{y}_2$ differ by about 0.72 SE's
  - this is because $\boldsymbol{t_s}$ is the measure of difference between the sample means expressed in terms of the SE of the difference

Hypothesis Testing: #2 Test Statistic

- How do we use this information to decide if the data support H_o?
  - Perfect agreement between the means would indicate that $\boldsymbol{t_s = 0}$, but logically we expect the means do differ by at least a little bit.
    - The question is how much difference is statistically significant?
  - If H_o is true, it is unlikely that $\boldsymbol{t_s}$ would fall in either of the far tails
  - If H_o is false it is unlikely that $\boldsymbol{t_s}$ would fall near 0

Hypothesis Testing: #3 P-value

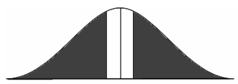
- How far is far?
- For a two tailed test (i.e. H_a: $(\mu_1 - \mu_2) \neq 0$) The p-value of the test is the area under the Student's T distribution in the double tails beyond $-t_s$ and t_s.



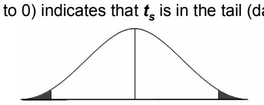
- Definition (p. 238): The p-value for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme or more extreme as the value of the test statistic that was actually obtained [from the data].

Hypothesis Testing: #3 P-value

- What this means is that we can think of the p-value as a measure of compatibility between the data and H_o
  - a large p-value (close to 1) indicates that t_s is near the center (data support H_o)



  - a small p-value (close to 0) indicates that t_s is in the tail (data do not support H_o)



Hypothesis Testing: #3 P-value

- Where do we draw the line?
  - how small is small for a p-value?
- The threshold value on the p-value scale is called the significance level, and is denoted by α
  - The significance level is chosen by whomever is making the decision (BEFORE THE DATA ARE COLLECTED!)
  - Common values for include 0.1, 0.05 and 0.01
- Rules for making a decision:

If $p \le \alpha$ then reject H_o (statistical significance)

α

If $p > \alpha$ then fail to reject H_o (no statistical significance)

Hypothesis Testing: #3 P-value

<u>Example</u>: Cholesterol medication (cont')

Find the p-value that corresponds to the results of the cholesterol lowering medication experiment

We know from the previous slides that t = -0.716 (which is close to 0)

This means that the p-value is the area under the curve beyond $\pm$ 0.716 with 18 df.

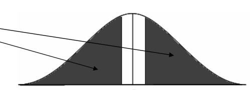
Hypothesis Testing: #3 P-value

Example: Cholesterol medication (cont')

Using SOCR we can find the area under the curve beyond $\pm$ 0.716 with 18 df to be:

p > 2(0.2) = 0.4

NOTE, when H_a is $\neq$, the p-value is the area beyond the test statistic in BOTH tails.



Hypothesis Testing: #4 Conclusion

Example: Cholesterol medication (cont')

Suppose the researchers had set α = 0.05

Our decision would be to fail to reject Ho because p > 0.4 which is > 0.05

(#4) CONCLUSION: Based on this data there is <u>no statistically significant</u> <u>difference between true mean cholesterol</u> of <u>the medication and placebo groups</u> (p > 0.4).

- In other words the cholesterol lowering medication does not seem to have a significant effect on cholesterol.

- Keep in mind, we are saying that we couldn't provide sufficient evidence to show that there is a significant difference between the two *population* means.

Hypothesis Testing Summary

- Important parts of Hypothesis test conclusions:
  1. Decision (significance or no significance)
  2. Parameter of interest
  3. Variable of interest
  4. Population under study
  5. (optional but preferred) P-value
  - 6. Even if the book doesn't ask for it
  - 7. A hypothesis test is a test with 4 parts
  - 8. How can you identify a hypothesis test problem?

Was Cavendish's experiment biased?

A number of famous early experiments of measuring physical constants have later been shown to be biased.

Mean density of the earth

True value = 5.517

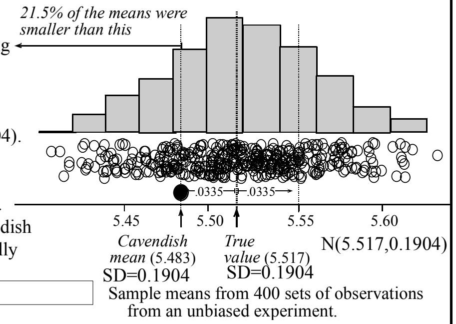
Cavendish's data: (from previous Example 7.2.2)

5.36, 5.29, 5.58, 5.65, 5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.10, 5.27, 5.39, 5.42, 5.47, 5.63, 5.34, 5.46, 5.30, 5.75, 5.68, 5.85

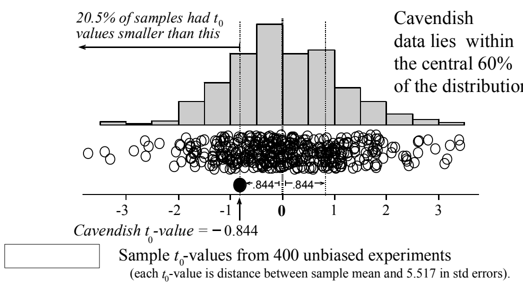
n = 23, sample mean = 5.483, sample SD = 0.1904

Was Cavendish's experiment biased?

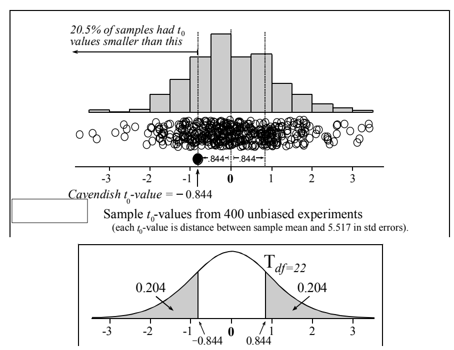
Simulate taking 400 sets of 23 measurements from N(5.517,0.1904). Plotted are the results of the sample means. Are the Cavendish values unusually diff. from true mean?



Cavendish: measuring distances in std errors



Cavendish data lies within the central 60% of the distribution



Measuring the distance between the true-value and the estimate in terms of the SE

- Intuitive criterion: Estimate is credible if it's not far away from its hypothesized true-value!
- But how far is far-away?
- Compute the distance in standard-terms:

$$T = \frac{\text{Estimator} - \text{TrueParameterValue}}{\text{SE}}$$

- Reason is that the distribution of ***T*** is known in some cases (Student's t, or N(0,1)). The estimator (obs-value) is typical/atypical if it is close to the center/tail of the distribution.

Comparing CI's and significance tests

- These are different methods for coping with the uncertainty about the true value of a parameter caused by the sampling variation in estimates.
- **Confidence interval**: A fixed level of confidence is chosen. We determine ***a range of possible values*** for the parameter that are consistent with the data (at the chosen confidence level).
- **Significance test**: *Only one possible value* for the parameter, called the hypothesized value, is tested. We determine the *strength of the evidence* (confidence) provided by the data against the proposition that the hypothesized value is the true value.

Review

- What intuitive criterion did we use to determine whether the hypothesized parameter value (p=0.2 in the ESP Example 9.1.1, and μ = 5.517 in Example 9.1.2) was credible in the light of the data? (Determine if the data-driven parameter estimate is consistent with the pattern of variation we'd expect get if hypothesis was true. If hypothesized value is correct, our estimate should not be far from its hypothesized true value.)
- Why was it that μ = 5.517 was credible in Example 9.1.2, whereas p=0.2 was not credible in Example 9.1.1? (The first estimate is consistent, and the second one is not, with the pattern of variation of the hypothesized true process.)

Review

- What do t_0-values tell us? (Our estimate is typical/atypical, consistent or inconsistent with our hypothesis.)
- What is the essential difference between the information provided by a confidence interval (CI) and by a significance test (ST)? (Both are uncertainty quantifiers. CI's use a fixed level of confidence to determine possible range of values. ST's one possible value is fixed and level of confidence is determined.)

Hypotheses

Guiding principles

We cannot **rule in** a hypothesized value for a parameter, we *can only* determine whether there is evidence *to* **rule out** a hypothesized value.

The ***null hypothesis*** tested is typically a skeptical reaction to a ***research hypothesis***

Comments

- Why can't we (**rule-in**) prove that a hypothesized value of a parameter is exactly true? (Because when constructing estimates based on data, there's always sampling and may be non-sampling errors, which are normal, and will effect the resulting estimate. Even if we do 60,000 ESP tests, as we saw earlier, repeatedly we are likely to get estimates like 0.2 and 0.200001, and 0.199999, etc. – non of which may be exactly the theoretically correct, 0.2.)
- Why use the rule-out principle? (Since, we can't use the rule-in method, we try to find compelling evidence against the observed/data-constructed estimate – to reject it.)
- Why is the null hypothesis & significance testing typically used? (H_o: skeptical reaction to a research hypothesis; ST is used to check if differences or effects seen in the data can be explained simply in terms of sampling variation!)

Comments

- How can researchers try to demonstrate that effects or differences seen in their data are real? (Reject the hypothesis that there are no effects)
- How does the alternative hypothesis typically relate to a belief, hunch, or research hypothesis that initiates a study? (H_1=H_a: specifies the type of departure from the null-hypothesis, H_0 (skeptical reaction), which we are expecting (research hypothesis itself).
- In the Cavendish's mean Earth density data, null hypothesis was $H_0 : \mu =5.517$. We suspected bias, but not bias in any specific direction, hence $H_a:\mu != 5.517$.

Comments

- In the ESP Pratt & Woodruff data, (skeptical reaction) null hypothesis was $H_0 : \mu =0.2$ (pure-guessing). We suspected bias, toward success rate being higher than that, hence the (research hypothesis) $H_a:\mu>0.2$.
- Other commonly encountered situations are:
  - $H_0 : \mu_1 - \mu_2 = 0 \rightarrow H_a : \mu_1 - \mu_2 > 0$
  - $H_0 : \mu_{rest} - \mu_{activation} = 0 \rightarrow H_a : \mu_{rest} - \mu_{activation} != 0$

The t-test

Using $\hat{\theta}$ to test H_0: $\theta = \theta_0$ versus some alternative H_1.

STEP 1 Calculate the ***test statistic***,

$$t_0 = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} = \frac{\text{estimate - hypothesized value}}{\text{standard error}}$$

[This tells us how many standard errors the estimate is above the hypothesized value (t_0 positive) or below the hypothesized value (t_0 negative).]

STEP 2 Calculate the P-value using the following table.

STEP 3 Interpret the P-value in the context of the data.

The t-test

Alternative hypothesis	Evidence against H_0: $\theta > \theta_0$ provided by	P-value
H_1: $\theta > \theta_0$	$\hat{\theta}$ too much bigger than θ_0 (i.e., $\hat{\theta}$ - θ_0 too large)	P = pr($T \geq t_0$)
H_1: $\theta < \theta_0$	$\hat{\theta}$ too much smaller than θ_0 (i.e., $\hat{\theta}$ - θ_0 too negative)	P = pr($T \leq t_0$)
H_1: $\theta \neq \theta_0$	$\hat{\theta}$ too far from θ_0 (i.e., $\lvert\hat{\theta}$ - $\theta_0\rvert$ too large)	P = 2 pr($T \geq \lvert t_0\rvert$)

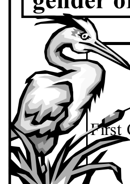
where T ~ Student(df)

Interpretation of the p-value

TABLE 9.3.2 Interpreting the Size of a *P*-Value

Approximate size of *P*-Value		Translation
> 0.12	(12%)	**No** evidence against H_0
0.10	(10%)	**Weak** evidence against H_0
0.05	(5%)	**Some** evidence against H_0
0.01	(1%)	**Strong** evidence against H_0
0.001	(0.1%)	**Very Strong** evidence against H_0

Is a second child gender influenced by the gender of the first child, in families with >1 kid?



First and Second Births by Sex

		Second Child Male	Female	Total
First Child	Male	3,202	2,776	5,978
	Female	2,620	2,792	5,412
	Total	5,822	5,568	11,390

- Research hypothesis needs to be formulated first before collecting/looking/interpreting the data that will be used to address it. Mothers whose 1st child is a girl are more likely to have a girl, as a second child, compared to mothers with boys as 1st children.
- Data: 20 yrs of birth records of 1 Hospital in Auckland, NZ.

Analysis of the birth-gender data – data summary

	Second Child	
Group	**Number of births**	**Number of girls**
1 (Previous child was girl)	5412	2792 (approx. 51.6%)
2 (Previous child was boy)	5978	2776 (approx. 46.4%)

- Let p_1=true proportion of girls in mothers with girl as first child, p_2=true proportion of girls in mothers with boy as first child. <u>Parameter of interest is p_1- p_2</u>.
- H_0: p_1- p_2=0 (skeptical reaction). H_a: p_1- p_2>0 (research hypothesis)

Hypothesis testing as decision making

Decision Making

	Actual situation	
Decision made	H_0 is true	H_0 is false
Accept H_0 as true	OK	Type II error
Reject H_0 as false	Type I error	OK

- Sample sizes: n_1=5412, n_2=5978, Sample proportions (estimates) $\hat{p}_1 = 2792/5412 \approx 0.5159$, $\hat{p}_2 = 2776/5978 \approx 0.4644$,
- H_0: p_1- p_2=0 (skeptical reaction). H_a: p_1- p_2>0 (research hypothesis)

Analysis of the birth-gender data

- Samples are large enough to use Normal-approx. Since the two proportions come from totally diff. mothers they are independent → use formula 8.5.5.a

$$t_0 = \frac{\text{Estimate - HypothesizedValue}}{SE} = 5.49986 =$$

$$\frac{\hat{p}_1 - \hat{p}_2 - 0}{SE\left(\hat{p}_1 - \hat{p}_2\right)} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} =$$

$$P - value = \Pr(T \geq t_0) = 1.9 \times 10^{-8}$$

Analysis of the birth-gender data

- We have strong evidence to reject the H_0, and hence conclude mothers with first child a girl a more likely to have a girl as a second child.
- How much more likely? A 95% CI:

CI (p_1- p_2) =[0.033; 0.070]. And computed by:

$$\text{estimate} \pm z \times \text{SE} = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

$$0.0515 \pm 1.96 \times 0.0093677 = [3\% ; 7\%]$$