Assignment # 2
Resampling based confidence intervals and hypothesis testing

Responsible: Mats Gustafsson

Contents

1 Goals

After completing this homework you are expected have acquired the following knowledge and competences.

Regarding Confidence Intervals:

1. Explain and apply the potential of so-called "Bootstrap Percentile Intervals", as an alternative to conventional analytically derived confidence intervals, for model free quantification of uncertainty about the true values of interest.
2. Explain and apply about how to use a collected dataset to obtain bootstrap intervals for (i) the unknown mean value of a pdf and for (ii) a particular proportion in a discrete (categorical) population.
3. For grades 4/5: Explain about and apply the idea behind the so-called "Basic Bootstrap Interval".

Regarding Hypothesis Testing:

1. Explain and apply the potential of performing hypothesis testing using resampling without distributional assumptions, as an alternative to the conventional model based approach.
2. Explain and apply about how to use a collected dataset to perform resampling based 2-group hypothesis testing.
3. Explain and apply about how to use a collected dataset to perform resampling based hypothesis testing when there are more than two hypotheses/groups.
4. Explain and apply resampling based hypothesis testing for the general multi-variate case.

2 Part A: Resampling based confidence intervals

2.1 Percentile bootstrap intervals for mean

2.1.1 Exercise A:1 - Percentile Bootstrap Intervals for the mean and the standard deviation

(a) Generate a dataset D consisting of $N = 20$ examples drawn from $N(\mu = 5, \sigma = 2.7)$.

(b) Determine the sample mean value m and sample standard deviation s.

(c) Generate $B = 1000$ bootstrap datasets D_b^* from D and for each such dataset, determine the corresponding mean value m_b and standard deviation s_b.

(d) Determine a 95% Percentile Bootstrap interval $[a_{20}, b_{20}]$ for μ where a_{20} is the 2.5% percentile of the values among the bootstrap mean values $\{m_b\}$, and b_{20} is the corresponding 97.5% percentile. Hint: The values should become $a_{20} \approx 4$ and $b_{20} \approx 6$ (with quite some variation if you run the code multiple times).

(e) Repeat a-d using $N = 2000$ instead, and compare the resulting bootstrap interval $[a_{2000}, b_{2000}]$ with $[a_{20}, b_{20}]$. Why is it shorter?

Remark.
Only conventional 95% confidence intervals are guaranteed to cover the true unknown value with probability (frequency) 95%. You should recall from previous courses that for $N = 20$, such an intervals is defined as

$$[a, b] = [m - t_{19}(0.975) \cdot s/\sqrt{N}, m + t_{19}(0.975) \cdot s/\sqrt{N}$$

where $t_{19}(0.975)$ is the 97.5% percentile of the t-distribution having 19 degrees of freedom.

2.2 Percentile bootstrap intervals for correlation

2.2.1 Exercise A:2 - Percentile Bootstrap Intervals for Correlations

You can choose any of the two problem formulations below.

Industrial Analytics formulation:
Assume you are jointly measuring the levels of two industrial variables X and Y across a set of 10 different machines (or factories/companies), and that the measured values are the following values

levels for_variable_X: $[13.2, 8.2, 10.9, 14.3, 10.7, 6.6, 9.5, 10.8, 8.8, 13.3]$
levels for_variable_Y: $[14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3, 9.3, 13.6]$

This means the first pair (13.2, 14.0) of measurement values comes from the first machine, the second pair (8.2, 8.8) comes from the second machine, etc. A natural question associated with this kind of data is the following. *Question:* Do the two variables X and Y seem to be dependent (co-vary, correlate)?

Molecular biotechnology formulation:
Assume you are jointly measuring the mRNA gene expression levels and of two different genes X and Y in a set of 10 different human cell cultures and that the measured values are the following:

gene_expressions_for_gene_X: $[13.2, 8.2, 10.9, 14.3, 10.7, 6.6, 9.5, 10.8, 8.8, 13.3]$
gene_expressions_for_gene_Y:$[14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3, 9.3, 13.6]$

This means the first pair (13.2, 14.0) of measurement values comes from the first cell culture, the second pair (8.2, 8.8) comes from the second cell culture, etc. A natural question associated with this kind of data is the following: Do the genes seem to be co-regulated?

TASKS (same for both alternatives above)
(a) Make suitable scatter plots to visualize the collected data. What is your spontaneous answer to the question above, based on this plot? Do the variables X and Y seem related?

(b) Determine the (Pearson) correlation coefficient for the levels for X and Y. Is if far from zero? What is your spontaneous answer to the questions above, based on this result? Do the variables X and Y seem related?

(c) Determine a Bootstrap percentile confidence interval (resampling based measure of uncertainty) for the correlation coefficient as follows:

1. Draw with replacement 10 pairs from the data set collected. This means that the same pair, for example the pair (8.2, 8.8), might be chosen multiple times.
2. Determine the correlation coefficient using the new bootstrap data set generated above.
3. Perform the two steps above repeatedly 1000 times to get 1000 correlation coefficient values.
4. Make a histogram using these 1000 values.
5. Determine the 2.5% and 97.5% percentiles a and b for these 1000 values to get the uncertainty interval $[a, b]$. Does this interval contain the origin (zero)? Based on this results, what is now your answer to the questions above? Do the variables X and Y seem related?

2.3 Percentile bootstrap intervals for a fraction (ratio)

2.3.1 Exercise A:3 - Bootstrap confidence interval for an unknown fraction

Industrial Analytics formulation: Assume you have discovered a new randomly occurring crack/error in your industrial product and you therefore start investigating a sample consisting of 60 additional examples of the same product. It turns out that 14 of them have the same kind of error. Report an approximate 95% bootstrap confidence interval for the fraction of sold products that have this kind of error by studying the variability around the observed value $14/60 \approx 0.233$.

Molecular biotechnology formulation: Assume you have discovered a new oncogene and you therefore start investigating a sample consisting of 60 tumor samples from colorectal cancer. It turns out that 14 of them have a mutation in this gene. Report an approximate 95% bootstrap confidence interval for the fraction of colorectal cancers that have this mutation by studying the variability around the observed value $14/60 \approx 0.233$.

Solve this problem by means of the following procedure:

(i) Write Python that makes you simulate drawing 60 samples with replacement from an urn containing 14 red and 46 white balls. NOTE: The data your observed data exactly reflects the observed distribution.

(ii) Run the code created in (i) 10^4 times to create 10^4 simulated data sets consisting of 60 samples each.

(iii) For each data set created, determine the proportion (fraction) of mutations/errors.

(iv) Make a histogram showing the distribution of the 10^4 proportions obtained in (iii).

(iv) Determine a 95% bootstrap confidence interval $[a, b]$by determining the 2.5% and 97.5% percentiles $a$and b for the set of 10^4 proportions obtained in (iii).

Expected results (here using 20 bins in the histogram): See histogram in Fig. 1 and the interval $[a, b] = [0.13, 0.35]$.

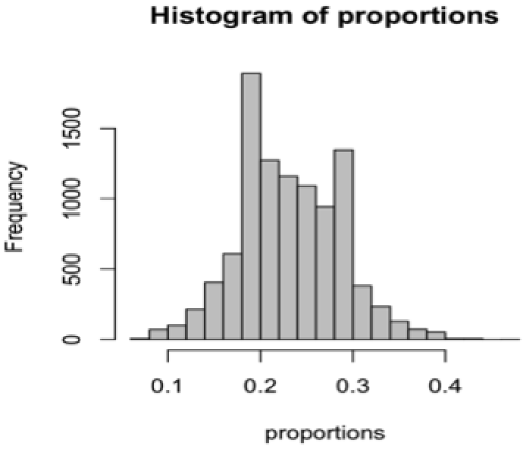

Figure 1: Histogram based on bootstrap estimates of proportions.

2.3.2 Exercise A:4 - Bootstrap confidence interval for an unknown robot success rate

Assume you are testing a pipetting robot by making it transfer liquids to all wells on a 96-well plate and that the number of successful transfers (wells) equals 89. Determine a 95% bootstrap confidence interval for the estimated success rate $89/96 \approx 0.93$ using 1000 or 10000 bootstrap sets. Tip: You should expect to get a bootstrap interval very similar to $[a, b] = [0.88, 0.98]$.

3 Part B: Resampling based hypothesis testing

Hypothesis testing using resampling methods is relatively straightforward. Here is the general algorithm:

1. First an urn model is created that reflects what is valid the null hypothesis of interest and a test statistic of interest is defined. The urn model can contain the elements in a dataset already collected, or it can contain elements that you select yourself in order to reflect the null hypothesis of interest.
2. Then samples are drawn from the urn model in order to simulate new data sets and thereby calculate a distribution of test statistics obtained assuming the null hypothesis.
3. Finally a p-value is determined that equals the probability of obtaining the observed value of the test statistic used, or something more extreme, assuming the null hypothesis is true.
4. If the p-value is smaller than your pre-defined threshold, then the null hypothesis is rejected.

In the following, a set of hypothesis testing problems are presented and for many of them a suggested solution using resampling is sketched.

One-sided vs two-sided p-value: "More extreme" may mean a one-sided or two-sided outlier. This depends on what is the alternative hypothesis H_1. **By extreme, we mean an observation that starts to suggest that the null hypothesis cannot be true**. If the alternative hypothesis is for example $H_1 : \mu_1 > \mu_2$ then it is not an extreme observation under the null hypothesis $H_0 : \mu_1 = \mu_2$ to observe the $m_2 >> m_1$. This is because such an observation gives essentially no support for the alternative hypothesis H_1, and therefore also no support for the possibility that the null hypothesis H_0 is not true under this particular hypothesis testing problem.

Important - how to report p-values: If your observed value is more extreme than all N values simulated under the null hypothesis, then you should not report that the p-value is zero: Instead you should report that the p-values is smaller than $1/N$, written as "$p < 1/N$".

Example - How to report p-values:
If $N = 1000$ and the observed value is more extreme than all the 1000 values simulated, then the corresponding p-value should not be reported to be zero but should be reported as "$p < 1/1000 = 0.001$". This reporting makes a lot of sense, because if we would have used for example $N = 10000$ instead of $N = 1000$ simulations, then perhaps 2 of the simulated values would actually be more extreme than the observed one. Then we would have $p = 2/10000 = 0.0002$, which is indeed smaller than 0.001 as claimed above based on $N = 1000$ simulations.

Remark
If the p-value truly is zero, it would mean that the observation you have obtained is impossible (zero probability) to observe under the null hypothesis. However, if you have so far only seen N observations, you cannot conclude that $p = 0$.

3.1 Hypothesis testing for proportions (fractions, ratios)

In this subsection, the problems are about testing hypotheses related to proportions (fractions, ratios).

3.1.1 Exercise B:1 - Test if a classifier performs better than random chance

Assume you have built a neural network classifier to discriminate between two subtypes of a disease and you test it on 20 new independent test examples. The result is that you make in total 7 errors. Can you reject the possibility (null hypothesis H_0 that the classifier is just guessing with probability 50% for both classes? Consider the alternative hypothesis H_1 that the classifier performs better than 50%. Since we observed 7 errors, there is at least some indication that the classifier performs better than chance. However, how (im)possible is it to get 7 errors or something even better by random chance under H_0 (meaning H_0 is true)?

Tip: Solve the problem by first creating an urn that contains one red and one white ball, thereby reflecting the null hypothesis (guessing). Then draw with replacement 20 balls from the urn and determine the fraction of red balls. Determine the p-value for the observation 7 errors by determining the probability of getting 7 or less red balls. What is your conclusion - should you reject the hypothesis that the classifier is guessing?

3.1.2 Exercise B:2 - Test claim that 88% in a population have a particular property

(a) Industrial Analytics formulation: A battery manufacturer reports (claims) that 88% of a particular battery produced are still operational after 100 hours of use. You suspect that the true value is lower so you randomly buy and test 1200 batteries, which results in 900 positive tests (meaning a fraction of 75%). Should you reject the manufacturer´s claim or not?

Molecular Biotechnology formulation: A report claims that 88% in a population have antibodies against Epstein-Barr Virus. You suspect that the true value is lower so you randomly collect 1200 individuals for testing resulting in 900 positive tests (meaning a fraction of 75%). Should you reject the reported claim or not?

Solve the problem by creating an appropriate urn model with red and white balls for the null hypothesis case and draw 1200 samples to simulate your testing procedure. Perform this at least 1000 times and determine for each simulation the fraction of positive tests. Finally determine the p-value for the event to get 900 or fewer positive tests under the null hypothesis model. (Meaning the alternative hypothesis H_1 is that the true fraction is lower than 88%).

(b) Assume that you would instead had bought and tested only 12 batteries and found that 9 were positive (again meaning a fraction of 75%). In this case, should you have rejected the reported claim or not?

3.2 Hypothesis testing for difference between two mean values

In this subsection, the problems are concerned with difference in mean values between two distributions.

3.2.1 Exercise B:3 - Hypothesis testing of sex difference of a feature/property

Industrial Analytics formulation: You have discovered a so far unexplored customer feature created/derived from quantitative interview (for example age or income), which you are suspecting is involved in the customer purchase behaviour. As a next step you would like to study if the value of this feature is different between men and women. In order to do this, you contact randomly selected men and women and you are able to complete 12 interviews with women and 11 with men. Based on your structured and quantitative interviews, you calculate the value of the unexplored customer feature and obtain the following values:
women =[60,62,73,81,27,72,61,59,42,39,71,74]
men =[74,79,80,89,71,72,92,67,91,83,43]

Question: Is there any statistical reason to believe there is a difference between men and women?

Molecular Biotechnology formulation: You have discovered a so far unknown hormone, which you are suspecting is involved in the control of blood glucose. Therefore you would like to study if the level of this hormone is different between men and women. In order to do this, you visit the blood central to pick up randomly selected blood samples from men and women and you are able to obtain 12 samples from women and 11 from men as follows (unit: pg/L)
women =[60,62,73,81,27,72,61,59,42,39,71,74]
men =[74,79,80,89,71,72,92,67,91,83,43]

Question: Is there any statistical reason to believe there is a difference between men and women?

Tip: Solve the problem as follows.

1. Calculate the difference between the mean values for the two groups $d_{\text{obs}} = m_{\text{women}} - m_{\text{men}}$.
2. Under the null hypothesis that there is no difference between the two groups, then all data have been drawn from the same underlying distribution. Therefore one may simulate the procedure performed as the result of random selection with replacement from an urn that contain all the 23 observed values. This means you should create an urn that contains all 23 values and draw with replacement 12 values to simulate the outcomes among 12 randomly chosen women and then draw 11 values with replacement to simulate the outcome among 11 randomly chosen men.

   Remark: If there is no difference between men and women, then the 23 values have been drawn from a common underlying distribution. Therefore we simulate draws from this unknown distribution by sampling with replacement from a ”bucket” (set) containing all our 23 values.
3. Calculate the mean difference d_{sim} for the simulated dataset in (ii) and perform this procedure in total 10000 times to get a set of 10000 simulated differences under the null hypothesis.
4. Determine a two-sided p-value for the observed difference obtained in (i) and compare it with the significance level you choose, for example 5%.

Based on this procedure and the resulting p-value, what is your answer to the question posed above? Why does it make sense to use a two-sided test in this case?

3.3 Testing for difference in the spreads between two distributions

In this subsection, we consider the problem of detecting the existence of a difference in the spreads of two distributions.

3.3.1 Exercise B:4 - Detection of difference in IQR

Industrial Analytics formulation: You are expecting that the variability in customer satisfaction X is larger in one market than in another. Therefore you collect data from both markets and the inter quartile range (IQR) for each of them. The data you obtained were the following values.

Molecular Biotechnology formulation: You are expecting that the variability in mRNA gene expression for a gene X is larger in one animal species than in another. Therefore you collect data from both species and calculate the inter quartile range (IQR) for each of them. The data you obtained were the following values.

values_of_group_A = [13.2, 8.2, 10.9, 14.3, 10.7, 6.6, 9.5, 10.8, 8.8, 13.3]
values_of_group_B = [14.0, 8.8, 11.2, 14.2, 11.8, 6.4, 9.8, 11.3, 9.3, 13.6]

(i) Determine the differences $d_{\mathrm{IQR}} = IQR_1 - IQR_2$ for these data.

(ii) Under the null hypotheses that there is no difference at all between the underlying distributions, merge the two data sets into one urn and draw 10 values with replacement twice to simulate two new datasets per species.

(iii) Calculate d_{IQR} for the simulated data set in (ii) and perform this procedure in total 10^4 times so that you get 10^4 simulated values of d_{IQR} under the null hypothesis.

(iv) Make histograms to visualize how the distributions of simulated values look.

(v) Determine a two sided p-value for the observed differences in (i) and (ii). Based on these results, what is your conclusion - is there any difference in spread between the two groups?

4 DEADLINE AND EXAMINATION

This examination will consists of the following parts:

(1) Answers and Results: Upload you Colab code to Studium and prepare for discussing your code during the mandatory seminar. Deadline: You must **submit before the first seminar** that belong to this assignment. Thus if there are 3 seminars where this assignment will be discussed, you must submit before the first one, even if you join a later session. This is in order to be fair and not having everyone coming to the last session.

NOTE: If you are working in pairs, each person has to submit their own document, even if it is almost identical. You are of course always responsible for your own submission and must be able to explain all of it during seminars and similar examinations.

(2) Python code as one single Colab notebook: The notebooks should be ready to run in Google Colab and it should be submitted after running it so that results (printouts and figures) are visible when opening the notebook. If your code consists of more than one notebook (and/or some additional code called from inside the notebook) your instructions how to run the code should be provided as a text file called `readme.txt`. Make sure that all files you are submitting have your first name and surname as part of the file name, like:

```
Assignment_1_Mats_Gustafsson.ipynb
```

(3) Personal reflections and feedback as pdf: An uploaded pdf document to Studium providing personal reflection and feedback:

```
Assignment_1_reflections_Mats_Gustafsson.pdf
```

This pdf document should contain reflective answers to the following questions:
(a) Did you have sufficient background knowledge to perform this exercise? If not, what background was missing?
(b) Make sure you understand that a bootstrap based confidence interval comes with no guarantees in terms of coverage (covering the true value), which is the case for conventional confidence intervals. Also make sure you understand that a conventional confidence interval may also totally miss the true value of interest (for a 95% interval, this happens in 5% of all cases).
(c) Make sure you understand and remember that a bootstrap data set is always having the same size as the original dataset.
(d) Explain why it is easy to determine a bootstrap confidence interval for the correlation coefficient while it would be very difficult (analytically as well as in terms of assumptions) to determine a corresponding conventional confidence interval.
(e) What did you personally find interesting and/or useful an/or informative with the assignment, and why? If you actually did not find it enlightening, in that case we would like to know about why.
(f) What did you find frustrating and/or difficult, and/or unnecessary, and why?
(g) What would you suggest as an improvement of this assignment for next year, and why?
(h) Did you spend more than 20 hours to complete this exercise? Why? NOTE: If you are working in pairs, each person has to send in a personal reflection.

(4) Mandatory Seminar: You discussing/presenting your results during the associated mandatory seminar. You do NOT have to prepare any separate slide presentation, it will be fine to show your Colab notebook if/when you are requested to present something.

Enjoy and Good Luck!