STATISTICAL MODELS

Course 2024-2025

Resampling Practice

(These problems are provided as homework to prepare for Labs 3 and 4)

1 Basic Problems

PROBLEM 1.- The following data

$$\mathbf{y} = \{1;\ 3;\ 2;\ 7;\ 5;\ 6\}$$

corresponds to a random sample from a random variable Y for which we wish to estimate the population variance, σ^2. From the above data, we can compute the sample variance: $s^2 = 5.6$

a) Generate $R = 5$ bootstrap resamples from $\mathbf{y}$.

b) Use your bootstrap resamples to estimate the bias of s^2

c) Use your bootstrap resamples to estimate a 95% confidence interval for σ^2

d) Discuss how accurate you believe are your estimates in (b) and (c). Justify.

PROBLEM 2.- Table 1 shows the marks for $n = 5$ students from 2 different Universities (A and B), on two aspects: science proficiency score and communication proficiency score. We wish to answer the following research question: *on average, do students from University A perform differently than students from University B, in terms of those scores?*

For this problem, we are only interested in the Science proficiency scores and we want to assess whether there is a difference in the **median** science scores of students when comparing both universities. For this we will use bootstrap. To avoid long calculations, you will assume that there is a large number (e.g. $B = 9,999$) of bootstrap rounds, but you will compute explicitly only the first ten (10) rounds. For these 10 rounds, however, you must provide as much detail as you can[1], for each of the following tasks:

a) Use bootstrap to compute a 95% confidence interval for the difference of the median science scores of University A $-$ University B.

b) Use bootstrap to test the relevant null hypothesis for this problem.

PROBLEM 3.- Repeat Problem 2 but now use the computer to fully run the bootstrap algorithms for $R = 999$ replications.

[1] You are expected to perform all the numeric calculations that you are able to.

Students from University A					
Science score	10	9	8	10	8
Communication score	8	6	6	4	6
Students from University B					
Science score	9	6	5	5	5
Communication score	8	8	10	8	6

Table 1: Data for Problem 2

PROBLEM 4.- Given the following data sample:

$$\mathbf{X} = \begin{bmatrix} -6 & -5 & -2 & 3 & 5 & 7 & 6 & 2 & -4 & -7 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}$$

we wish to use bootstrapping to estimate a 90% confidence interval of the correlation coefficient between the two dimensions.

a) Generate $B = 3$ bootstrap resamples that could be used for the task.

b) Now imagine that you have $B = 999$ bootstrap resamples, similar to those generated in (a). Explain in detail how you would use the bootstrap resamples to estimate the requested confidence interval. You do not need to perform the actual calculations, but **you must provide all the necessary formulas**.

c) Use the computer to fully run the bootstrap algorithms for $B = 999$ replications.

PROBLEM 5.- The following data is a random sample of the amounts spent by 15 consecutive shoppers at a supermarket;

$$\begin{aligned} \mathbf{x} = \{&0.82;\ 80.76;\ 18.85;\ 8.41;\ 2.69; \\ &32.49;\ 1.89;\ 24.21;\ 37.80;\ 6.79; \\ &118.74;\ 17.88;\ 7.21;\ 5.64;\ 40.97\} \end{aligned}$$

a) Inspect the data to determine if we could assume it to be normally distributed.

b) Even if the data is not normally distributed, the sample mean should be approximately normally distributed for sufficiently large N, but 15 samples do not seem sufficient. Use bootstrapping to assess if the distribution of the sample mean can be assumed normal.

c) Based on the bootstrap from (b), estimate the standard error and the bias of your estimator.

d) Estimate a 95% confidence interval for the mean using student's t. Based on your analysis from (b), discuss whether this is likely to be accurate or not.

e) Estimate a 95% confidence interval to the mean using bootstrapping[2] and compare with (d).

[2]There exist several methods to estimate confidence intervals based on bootstrapping. Unless one of them is explicitly indicated, you are free to choose any bootstrap-based method, always keeping in mind that it must be well suited for the data you are analyzing.

PROBLEM 6.- File `SM22_Lab_3_Supermarket.xlsx` contains more samples from the same supermarket data as Problem 5. In this case, however, there are $n = 100$ samples, and therefore we expect that the sample mean will have a distribution much closer to normal than in Problem 1. Repeat points (a) to (e) from Problem 1 but using the data provided in the file.

PROBLEM 7.- Using the data from Problem 5:

a) Use bootstrapping to assess if the distribution of the sample **median** can be assumed normal.

b) Based on the bootstrap from (a), estimate the standard error and the bias of your estimator.

c) Estimate a 95% confidence interval for the **median** using student's t (hint[3]). Based on your analysis from (b), discuss whether this is likely to be accurate or not.

d) Estimate a 95% confidence interval for the **median** using bootstrapping and compare with (c).

PROBLEM 8.- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The sample skewness can be estimated as[4]:

$$\beta = \sqrt{n}\frac{\sum_{j=1}^{n}(x_j - \overline{x})^3}{\left(\sum_{j=1}^{n}(x_j - \overline{x})^2\right)^{3/2}}$$

where x_j are the observed data samples, with $1 \leq j \leq n$ and $\overline{x}$ is the sample mean. Using the data from Problem 5:

a) Use bootstrapping to assess if the distribution of the sample **skewness** can be assumed normal.

b) Based on the bootstrap from (a), estimate the standard error and the bias of your estimator.

c) Estimate a 95% confidence interval for the **skewness** using student's t (without bootstrapping). Based on your analysis from (b), discuss whether this is likely to be accurate or not.

d) Estimate a 95% confidence interval for the **skewness** using bootstrapping and compare with (c).

PROBLEM 9.- The following data

$$\mathbf{y} = \{3;\ 5;\ 7;\ 18;\ 43;\ 85;\ 91;\ 98;\ 100;\ 130;\ 230;\ 487\}$$

corresponds to a random sample of the time between failures of some equipment, for which we wish to estimate the underlying mean (this is of interest because its reciprocal is the failure rate).

a) Compute a 95% confidence interval for the mean time between failures based on the sample mean (without bootstrapping) as you would do if the distribution of your sample was normal. Check whether such assumption seems to be supported by the data.

b) Use bootstrap to estimate the standard error of the sample mean and use it to construct a 95% confidence interval under the assumption that the distribution of the sample mean is normal. Indicate whether such assumption seems to be supported by the data.

c) Compute a 95% confidence interval using the bootstrap-t method, i.e. without assuming that the distribution of your statistic follows a normal distribution.

[3] You may want to use bootstrapping to estimate the standard error of your statistic.

[4] Keep in mind that there exist several other ways to estimate skewness.

d) Compute a 95% confidence interval using the percentiles of the bootstrap distribution. Here, again, you are not assuming Gaussianity.

PROBLEM 10.- Let $\mathbf{x_1}$ and $\mathbf{x_2}$ be two random samples, as follows:

$$\mathbf{x_1} = \begin{bmatrix} 82 & 79 & 81 & 79 & 77 & 79 & 79 & 78 & 79 & 82 & 76 & 73 & 64 \end{bmatrix}$$
$$\mathbf{x_2} = \begin{bmatrix} 84 & 86 & 85 & 82 & 77 & 76 & 77 & 80 & 83 & 81 & 78 & 78 & 78 \end{bmatrix}$$

We want to test if $\mu_1 = E[\mathbf{x_1}]$ is smaller than $\mu_2 = E[\mathbf{x_2}]$ at significance level $\alpha = 0.05$ (where $E[\cdot]$ denotes expected value) and estimate the corresponding p-value.

a) State the null and alternative hypotheses for this problem.

b) Use bootstrapping to generate $R = 1000$ simulated resamples of $\mathbf{x_1}$ and $\mathbf{x_2}$. Use them to estimate the p-value of interest. What do you conclude about $\mu_1 < \mu_2$?

c) Repeat (b), making sure that simulated resamples are not exactly the same as before (most random number generators do this by default unless you specify the same *seed*). Does your result coincide with (b)? - Explain why.[5]

d) Estimate a 95% confidence interval for your p-value. How does this affect your conclusion from (b) about $\mu_1 < \mu_2$?

e) Do you consider the results so far conclusive? If not, indicate what else could be done to obtain a more accurate p-value and report your revised results.

PROBLEM 11.- Consider again the data from the previous problem, but now as a bivariate sample (i.e. the two dimensions of a bivariate population $\mathbf{X}$):

$$\mathbf{x} = \begin{bmatrix} 82 & 79 & 81 & 79 & 77 & 79 & 79 & 78 & 79 & 82 & 76 & 73 & 64 \\ 84 & 86 & 85 & 82 & 77 & 76 & 77 & 80 & 83 & 81 & 78 & 78 & 78 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$$

We wish to test whether the two dimensions of out hypothetical population $\mathbf{X}$ are correlated, at significance level $\alpha = 0.95$. One possible approach to do so consists in computing a confidence interval for the correlation coefficient. We will use bootstrap for this:

a) State the null and alternative hypotheses for this problem.

b) Use bootstrapping to generate $R = 1000$ simulated resamples of $\mathbf{x_1}$ and $\mathbf{x_2}$. Use them to estimate the standard error and bias of the sample correlation.

c) Based on (b), estimate a 95% confidence interval for the correlation coefficient.

d) State your conclusions about the hypotheses from (a).

PROBLEM 12.- Consider again data from Problem 10, and the same research question: are both dimensions correlated? Another way to address this question is by directly evaluating the null hypothesis (as opposed to the *indirect* way through the confidence intervals that we used in Problem 10).

a) State the null and alternative hypotheses for this problem.

[5] Hint: Results from (b) and (c) should NOT coincide exactly.

b) Use bootstrapping to generate $R = 999$ simulated resamples of $\mathbf{x}$ **adequately modified so that the sampling process corresponds to the case in which the null hypothesis is true.**

c) Based on (b), estimate a p-value for the hypotheses in (a) and state your conclusion.

d) Based on (b), we may also estimate a 95% confidence interval for the correlation coefficient **under the null hypothesis**. Explain where the interval is centered and what value shall we check against this interval to decide about the null.

PROBLEM 13.- Let $\mathbf{x_1}$ and $\mathbf{x_2}$ be two independent random samples, as follows:

$$\mathbf{x_1} = \begin{bmatrix} 5 & 7 & 8 & 9 & 10 & 11 \end{bmatrix}$$
$$\mathbf{x_2} = \begin{bmatrix} 3 & 4 & 6 & 12 & 13 & 14 \end{bmatrix}$$

We want to test if $\mu_1 = E[\mathbf{x_1}]$ is smaller than $\mu_2 = E[\mathbf{x_2}]$ at significance level $\alpha = 0.05$ (where $E[\cdot]$ denotes expected value) and estimate the corresponding p-value. To do so, we want to use bootstrapping:

a) State the null and alternative hypotheses for this problem.

b) Describe the bootstrap procedure that you would use; provide all the calculations and details for the first 2 bootstrap replications, including the appropriate choice of the bootstrap samples.

c) Now assume that you have $B = 999$ replications as the ones described in (b). Detail how you would use them to test the hypothesis in (a).

2 Deliverable-like Practice Problems

PROBLEM 14.- File `SM22_Lab_3_Height_and_Weight.xlsx` contains samples of height and weight from a set of individuals. Based on them, it is possible to estimate the Body Mass Index (BMI), which will be our variable of interest. BMI can be computed as follows:

$$BMI = \frac{Weight}{(Height)^2}$$

where $Weight$ must be expressed in kilograms and $Height$ in meters. If we compute BMI in this way, the distribution of our sample shows a rather large kurtosis, which is above 5. However, we wish to know how much we can trust the kurtosis value estimated from this sample, and thus we wish to estimate a confidence interval for it.

a) Use bootstrapping to assess if the distribution of the sample **kurtosis**([6]) can be assumed normal. Clearly indicate your conclusion.

b) Based on the bootstrap from (a), estimate the standard error and the bias of your estimator.

c) Estimate a 95% confidence interval for the **kurtosis** using student's t. Based on your analysis from (b), discuss whether this is likely to be accurate or not.

d) Compute a 95% confidence interval for the **kurtosis** using the percentiles of the bootstrap distribution.

[6] You can look for the formula of the sample kurtosis or use existing functions to compute it.

PROBLEM 15.- File `SM22_Lab_3_Marks.xlsx` shows the marks from 32 students from statistical models in 2021 for labs 1 and 2. We wish to investigate whether students have, on average, improved their marks in Lab 2 with respect to Lab 1.

a) State the null and alternative hypotheses for this problem.

b) Use bootstrapping to test the hypotheses defined in (a) at significance level $\alpha = 0.05$. Report your estimated p-value and state your conclusion answering the research question.[7]

c) Use bootstrapping to compute a 95% confidence interval of the average difference between the marks of labs 1 and 2. Analyze whether the conclusion from this interval confirms or contradicts your conclusion from (b).

PROBLEM 16.- File `SM22_Lab_3_Baseball.xlsx` contains the salaries and batting averages of 50 randomly selected baseball players. We want to assess it there is a linear relationship (correlation) between the player salary and his performance (in this case, in terms of batting average).

a) State the null and alternative hypotheses for this problem.

b) Use bootstrapping to test the hypotheses defined in (a) at significance level $\alpha = 0.05$. Report your estimated p-value and state your conclusion answering the research question.

c) Use bootstrapping to compute a 95% confidence interval of the correlation coefficient between salaries and batting averages. Analyze whether the conclusion from this interval confirms or contradicts your conclusion from (b).

PROBLEM 17.- Incumbent local exchange carriers (ILECs) install and maintain local telephone lines, lease capacity, and perform repairs for the competing local exchange carriers (CLECs).
File `SM22_Lab_3_ILEC_1.xlsx` provides the random sample $\mathbf{x_2}$ corresponding to the repair times (in hours) of 1664 service requests from customers of an ILEC and file `SM22_Lab_3_ILEC_2.xlsx` provides $\mathbf{x_2}$ corresponding to 23 requests from customers of a CLEC during the same time period.

a) Assess whether these two random samples seem to come from normal distributions.

b) Use bootstrapping to generate a simulated population of random resamples that allows you to inspect the distribution of the statistic $t = \overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2}$.

c) Estimate the bias of $t = \overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2}$.

d) Compute a 95% confidence interval for t based on the percentiles of the bootstrap distribution. Based on the analysis from (a) and (b) discuss whether this interval seems appropriate.

e) Based on your answers to (b) and (d), what can you conclude about the statement following statement: *"repair times for ILEC customers are significantly smaller than repair times for CLEC customers"*.

PROBLEM 18.- File `SM22_Lab_3_3D.xlsx` contains a random sample from some tri-dimensional data. We are interested in assessing how dominant is the first principal component of the data, for which we choose the statistic t to be the ratio between the largest and second-largest eigenvalues of the

[7] Hint: use at least B = 10,000 resamples.

covariance matrix of our random sample. That is, if $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the eigenvalues of the sample covariance matrix, we want a 95% confidence interval for

$$t = \frac{\lambda_1}{\lambda_2}$$

Analyze the data carefully and justify the method chosen to estimate the requested interval.

PROBLEM 19.- Last year, in Lab 1, we analyzed the concentration of 5 contaminants in salmon samples from different origins. For simplicity, file `SM22_Lab_1_Salmon2.xlsx` is provided again here with the same data. When we analyzed salmon from Scotland, there was large consensus from a majority of students that we could not reject univariate normality for any of the dimensions (contaminants). However, the conclusions about bivariate normality were less clear, given that the chi-square plots for some pairs of dimensions showed deviations from the expected behavior for normality from which it was not trivial to decide whether we shall or shall not reject normality.

In this problem we will re-analyze the samples of salmon from Scotland and will assess bi-variate normality quantitatively with the help of parametric bootstrap. To this end, we will use two statistics to evaluate how much a chi-square plot deviates from bivariate normality: $i)$ the *correlation* coefficient (similarly to what we do for qq-plots) and $ii)$ the *slope* of the resulting plot[8].

a) Use parametric bootstrapping to generate B replicates[9] of the two statistics indicated above, under the null hypothesis H_0: *the population from which we sample our bootstraps follows a bivariate normal distribution*. Plot the resulting replicates together[10] as the two dimensions of a single 2-dimensional statistic $\hat{\theta}$. Keep in mind that these are replicates intended to assess bivariate normality of our sample from Scotish salmon.

b) Assess whether the distribution of the replicates from $\hat{\theta}$ generated in (a) can be considered multivariate normal. Clearly indicate your conclusion.

c) Since our statistic $\hat{\theta}$ is bi-dimensional, we may compute simultaneous confidence intervals that should be fulfilled by a given chi-square plot if it comes from a bivariate normal distribution (at a given significance level). Compute such intervals for significance level $\alpha = 0.05$. Hint: analyze carefully what would be the expected behavior under normality to determine whether each of the intervals shall be one- or two-sided. Justify adequately.

d) Use the intervals obtained in (c) to analyze bivariate normality of all possible pairs of contaminants in the dataset of salmon from Scotland. Clearly indicate your conclusions.

[8] The slope of a linear approximation of our resulting plot.

[9] Choose an adequately large value for B.

[10] For example, you can use a scatter plot or a density estimate.

3 Answers to Selected Problems

The answer below correspond to **one possible outcome** of the bootstrap simulations. To facilitate the possibility to reproduce the same numbers (which is not necessary), the random number generator was reset at the beginning of each problem, by running the command `rng( 0 );`

PROBLEM 5.-

a) Not normal.

b) After bootstrap with $B = 1000$ replications the estimated distribution of the statistic is not normal (correlation of the qqplot $\simeq 0.9934$).

c) Bias $\simeq -0.1714$; Standard error $\simeq 8.2667$ (using the same bootstrap resamples generated in (b)).

d) Fully theoretical interval: $8.6942 \leq \mu \leq 45.3258$
Student's-t interval using the bootstrap standard error and bias (from (c)): $9.4511 \leq \mu \leq 44.9117$

e) Using the method of percentiles: $10.6193 \leq \mu \leq 41.3900$

PROBLEM 6.-

a) Not normal.

b) After bootstrap with $B = 1000$ replications we cannot reject that the estimated distribution of the statistic is normal (correlation of the qqplot $\simeq 0.9992$).

c) Bias $\simeq +0.0353$; Standard error $\simeq 1.9694$

d) Fully theoretical interval: $15.9358 \leq \mu \leq 23.6981$
Student's-t interval w/ the bootstrap standard error and bias (from (c)): $15.8739 \leq \mu \leq 23.6894$

e) Using the method of percentiles: $15.8630 \leq \mu \leq 23.4371$

PROBLEM 7.-

a) After bootstrap with $B = 1000$ replications we can clearly reject normality.

b) Bias $\simeq -1.9032$; Standard error $\simeq 8.5581$

c) Student's-t interval w/ the bootstrap standard error and bias (from (b)): $1.4280 \leq \nu \leq 38.1385$ where ν is the median of the population.

d) Using the method of percentiles: $-2.0400 \leq \nu \leq 30.1200$
Using the Bootstrap-t method, with a second round of bootstrap with $B_2 = 500$ to estimate the standard error of each bootstrap resample: $3.4026 \leq \nu \leq 47.8752$

PROBLEM 8.-

a) After bootstrap with $B = 1000$ replications we observe deviations from normality (especially asymmetry); this can be confirmed also numerically (we can reject normality).

b) Bias $\simeq -0.3060$; Standard error $\simeq 0.6257$

c) Student's-t interval w/ the bootstrap standard error and bias (from (b)): $0.7000 \leq \beta \leq 3.38425$

d) Using the method of percentiles: $0.7609 \leq \beta \leq 3.2212$
Using the Bootstrap-t method, with a second round of bootstrap with $B_2 = 500$ to estimate the standard error of each bootstrap resample: $1.0217 \leq \beta \leq 3.8139$

PROBLEM 9.-

a) Fully theoretical interval: $21.5256 \leq \mu \leq 194.6411$
After bootstrap with $B = 1000$ replications we conclude that for this data we cannot assume normality of the sample mean.

b) Bias $\simeq -0.6717$; Standard error $\simeq 38.2944$

c) Using a second round of bootstrap with $B_2 = 500$ to estimate the standard error of each bootstrap resample: $42.7549 \leq \mu \leq 299.6842$

d) $23.2500 \leq \mu \leq 170.4167$

PROBLEM 11.-

b) Bias $\simeq 0.0426$; standard error $\simeq 0.1343$

c) We use the percentiles method and obtain: $0.1521 \leq \rho \leq 0.6919$ - We may also use Stuedent's-t intervals since the bootstrap distribution looks normal.