Homework 5

Here we present solutions to the problems posted in the fifth homework assignment. The solutions and related commentary are put in italics. Remember that problems can have several different but correct ways of solving them.

Multiple choice questions

1. Statistical inference

   A is about trying to guess how data would look if collected anew;

   B is about producing data on a computer that would look similar to real data;

   C is about inferring from the model how data should look like;

   D is about drawing proper graphs that illustrate raw data;

   ***E** is about inferring from the data about parameters that describe an assumed model for the data.

   Solution: *In statistics, a model for the mechanism that has produced data is assumed. The model is characterized by some parameters that are unknown. Having data, statistics tries to infer from them some information about these parameters.*

2. Estimation

   A is a form of statistical inference in which we use computer to evaluate a rough approximation of a certain value that we can not compute exactly;

   B uses a model to predict values of data;

   ***C** uses data to propose reasonable values for parameters of a model;

   D is a procedure of decision making about how to collect data;

   E is another word for testing statistical hypotheses.

   Solution: *Estimation is a form of statistical inference that looks for functions of the data that when evaluated provide with reasonable values for the unknown parameters of a proposed model for the data.*

3. Confidence intervals

   A tell us what are chances for the sample mean that is computed from collected data to fall into these intervals;

   B always include the true value of a parameter of interest;

   C express our confidence that the proposed model for data is correct;

   D are longer when more data are collected;

   ***E** are intervals that are computed from the data and have a good chance to include the true value of a parameter of interest.

Solution: *Confidence intervals at a chosen confidence level, say 95%, are evaluated from the data and have 95% chance to include the true mean, i.e. if their computation is repeated for different data sets the proportion of times that they include the true mean will be approximately 95%.*

4. When testing statistical hypotheses

   A we find methods that allow us to find out the definite truth about the hypotheses at hand;

   B we never make Type I Errors;

   C we tend to reject null hypothesis more often than not because rejecting null hypothesis is the desireable conclusion of testing;

   ***D** the focus is on not to falsely reject the null hypothesis;

   E when we reject null hypothesis we know that we are right.

   Solution: *The significance of a test that is typically a small percentage such as 5% or 1% is set prior to application of the testing procedure to the data, is representing the percentage of times the null hypothesis is falsely rejected (Type I Error). In this way, the test are protect us from falesly rejecting the null hypothesis.*

5. Which of the following is correct?

   A increasing the level of confidence associated with a confidence interval gives a narrower interval;

   B increasing the sample size associated with a 95% confidence interval increases the confidence level of the interval;

   C the closer to 50% is the percentage of failures, the narrower is a confidence interval for the percentage calculated from data sample from the process;

   D 95% of all confidence intervals for a process mean are within 2 standard errors of X-bar;

   ***E** an alternative way to implement a significance test is to calculate the corresponding confidence interval and check whether the null hypothesis value is included or not.

   Solution: *If the confidence level of a confidence interval is, say, 95%, than the chances that the true value will not be in this interval are 5%. If we reject the null hypothesis when the null hypothesis value is not in this interval while it is the true value (the null hypothesis is true), then chances that we make an error is 5%, i.e. we end up with a significance test on the significance level equal to one minus the confidence level. See also the textbook, page and lecture slides for more discussion.*

6. The difference between $\bar{X}$ and μ is

   A the contents of the null hypothesis;

   ***B** that one is a sample statistic and the other is a model parameter;

   C that one is a process characteristic and the other is a population parameter;

   D proportional to the standard error;

   E the standard deviation of the sampling distribution.

Solution: *$\bar{X}$ is a function evaluate on the data (their average) and as a such is called sample statistic, while μ is an unknown parameter of the model we assume for the data.*

Problems

1. In the texbook, pages 146-147 (discussed also in the lecture), we have seen that for the clips data after arrival of a new roll of steel material, the clip gaps have been noticeable smaller. In the lecture we have computed a 95% confidence interval for the mean of clip gaps before arrival of this new batch of material. The computed interval was $[72.2, 75.4]$ and was based on the sample mean of 73.8 and standard deviation 7.3. After the arrival of the new batch, the estimate of the mean of clip gaps is 66.75 which was computed based on the sample size of 40. Compute the new confidence interval assuming that standard deviation has not changed, i.e. that it is still 7.3.

   Solution: *The formula for the confidence interval based on the normal model is: $[\bar{X} - 1.96 \times \sigma/\sqrt{n}, \bar{X} + 1.96 \times \sigma/\sqrt{n}]$. In this particular case, it yields*

$$[66.75 - 1.96 * 7.3/\sqrt{40}, 66.75 + 1.96 * 7.3/\sqrt{40}] = [64.49, 69.01].$$

   We note that the confidence intervals are disjoint.

2. The marketing director of a bank would like to make a 'special offer' ot personal customers whose accounts turn over more than 100,000EUR in a year. He guesses that such customers account for no more than 10% of all the banks customers, and costs the 'special offer' accordingly. The financial controller believes that there are over 20% of such customers and suggests that the cost will be too high. To resolve the issue, the counts department is asked to estimate the percentage of large customers. What sample size is needed to find the interval estimate with the halflength of (i) 1%; (ii) 2%, with (a) 90% confidence; (b) 95% confidence.

   Solution: *For the estimation of percentages the 90% confidence interval is defined by $[\hat{p} - 1.645\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.645\sqrt{\hat{p}(1-\hat{p})/n}]$ and for its halflength to be 1% we need*

$$1.645\sqrt{\hat{p}(1-\hat{p})/n} \leq 0.01.$$

   The largest possible value of $\sqrt{\hat{p}(1-\hat{p})}$ is 1/2 (achieved at $\hat{p} = 0.5$) which we can use since $\hat{p}$ is not known. This leads to the inequality from which we can get n:

$$n \geq \frac{1.645^2}{4 * (0.01^2)} = 6765.06.$$

   Thus the smallest sample size to achieve the desired accuracy is $n = 6766$. For the halflength of 2% we get

$$n \geq \frac{1.645}{4 * (0.02^2)} = 1691.27$$

   and consequently the sample size of 1692 can be chosen. For the 95% confidence intervals we need to replace 1.645 by 1.96 which leads to $n = 9604$ for 1% halflength and $n = 2401$ for 2% halflength.

3. Following analysis of the 52 cash variances observed over the year, it was concluded that three of the values were exceptional and that the remaining values represented a process in statistical control, with standard deviation 4.15; see page 124, Chapter 4. The estimated mean value, based on the in-control variances, was 0.6. Based on these data, test the null hypothesis that the cash variance process has mean zero. Make a report of the test and its result referring to

   - the null hypothesis;
   - proposed significance level;
   - the standardized test statistic;
   - p-value (observed significance level);
   - the conclusion.

   Solution: *Let μ denote the theoretical value of the mean cash variances. We choose as the null hypothesis $\mu = 0$ and the alternative hypothesis that $\mu \neq 0$. We decide for the most popular significance level of 5%. The standardized statistics for this test is*

$$Z = \frac{\bar{X}}{\sigma/\sqrt{n}} = \frac{0.6}{4.15/\sqrt{49}} = 1.0120.$$

   The test rejects the null hypothesis if $|Z|$ is greater than 1.96 *- the standardized critical value. The p-value can be checked in the normal distribution table for the frequency corresponding to value* 1.0120*. We check that this is* 0.8442 *and compute 2*(1-0.8442)=0.311. The p-value (the observed significance level) is 0.311, fairly large. We conclude that we do not have any suport in the data to reject the null hypothesis.*

4. A company with a large number of debtors claims (for the purpose of negotiating a bank loan) that the average amount owed is at least 100EUR. An evaluator acting on behalf of the bank looked at a sample of 250 of the company's outstanding accounts and found an average amount owed of 95.53EUR. She found the standard deviation to be 24EUR. Formally test the statistical significance of the claim made by the company. Indicate what you take to be the *null hypothesis*, the *significance level*, the *test statistics*, the *critical value*, and explain the basis for your test in terms of the *sampling distribution* of the test statistic. Calculate the p-value (the observed significance level) for the test and relate its value to the formal conclusion of the test.

   Solution: *Let μ represent the theoretical value of the average owed to the bank. We want to test the null hypothesis that corresponds to the claim by the bank: $\mu \geq 100EUR$ against the alternative hypothesis $\mu < 100EUR$. As usually, we set the significance level to 5%. The standardized test statistics here is*

$$Z = \frac{95.53 - 100}{24/\sqrt{250}} = -2.945.$$

   Since we deal with a relatively large sample size we may assume the normal model for the sample mean. In this case, we reject the null hypothesis when Z falls below the critical value of -1.645 *taken from the table of the normal model. We see that the test statistics falls below the critical level. We may compute the p value using tables and symmetry of the normal distribution: 1-0.9984=0.0016. We conclude that we can reject (strongly) the null hypothesis.*

5. The catering manager in a hotel suspects that the weight of loaves of bread delivered daily by a bakery is consistently below the nominal weight of 800g. To test this, 10 loaves chosen at random from a day's deliveries are weighed. The mean and standard deviation of the ten weights are 792g and 25g, respectively.

   - Carry out a formal significance test.
   - List the steps involved in this test, from null hypothesis specification to conclusion.
   - Calculate a 95% confidence interval for the average weight of loaves produced by the baker.
   - Comment on the correspondence between the interval, as calculated, and the result of the test.

   Solution: *If μ represents the theoretical mean weight of loaves, it is natural to set the null hypothesis as $\mu = 800mg$ and the alternative $\mu < 800g$ as the manager looks for a strong evidence for the latter. Significance level is set to be 5% and the test statistic is*

$$T = \frac{\bar{X} - 800}{s/\sqrt{n}}.$$

   Due to small sample size we should use Student-t distribution with 9 degrees of freedom rather than the normal distribution. The critical value can be thus taken from the tables of Student-t distribution which, for example, is posted on our schedule of classes as a handout. Here is its relevant portion

PERCENTAGE POINTS OF THE T DISTRIBUTION

Tail Probabilities									
One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
Two Tails		0.20	0.10	0.05	0.02	0.01	0.002	0.001	
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637	1
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6	2
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92	3
R	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610	4
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869	5
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959	6
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041	8
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781	9
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587	10
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437	11
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318	12
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221	13

   We see that the critical value (due to the symmetry) is -1.833 and any value of the standardized statistics below this critical value would lead to the rejection of the null hypohtesis. The value of the standardized statistics in our case is

$$T = \frac{792 - 800}{25/\sqrt{10}} = -1.011929.$$

Since it is not below the critical value, we must that there is not enough evidence to claim that the loaves are lighter than 800g.

The 95% confidence interval computed based on the Student-t distribution is

$$\begin{aligned}[\bar{X} - 2.262s/\sqrt{n}, \bar{X} + 2.262s/\sqrt{n}] &= [792 - 2.262 * 25/\sqrt{10}, 792 + 2.262 * 25/\sqrt{10}] \\ &= [774.11, 809.88].\end{aligned}$$

We see that it contains 800 which is consistent with the conclusion of the test that based on the data we can not exclude 800 as a possible value for the theoretical mean value μ.