Statistics: Home Exercises for Exam Preparation

Theophanis Tsandilas

January 11, 2022

Imagine that a researcher runs an experiment with 14 healthy university students to study the effect of noise distraction on math performance. More precisely, all the students are asked to answer a multiplication question under two different environmental conditions:

1. **Low Noise.** Environmental noise is low, between 20 and 40 decibel.
2. **High Noise.** Environmental noise is high, between 60 and 80 decibel.

In case of an incorrect answer, the students are required to retry until the correct answer is given. To evaluate performance, the researcher measures the time it takes to correctly answer a question. The following table presents this time (in seconds) for each participant (P1 - P14) and for each condition (Low vs. High Noise).

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**
Low Noise	2.1	6.3	1.8	2.5	3.8	2.4	2.8	2.6	3.1	3.0	4.3	4.1	3.5	2.8
High Noise	2.8	5.6	3.0	2.7	3.6	3.5	2.6	3.8	4.1	3.6	4.9	4.4	3.3	4.1

1 Populations, Sampling, and Experimental Design

Please, answer the following questions as concisely as possible:

1. What are the populations of interest in the above scenario?
2. To what extent are the above samples representative of the populations of interest?
3. Is it reasonable for the above experiment to make the assumption that time observations for different participants are independent? Justify your answer.
4. Is it reasonable for the above experiment to make the assumption that observations under high-noise conditions are independent from observations under low-noise conditions? Justify your answer.
5. Suppose the researcher considers the following experimental procedure: All the participants first complete a first set of tasks with Low Noise. Then, they all complete a second set of tasks with High Noise. What are the problems of this experimental procedure? Could you propose a better solution that avoids or minimizes these problems?

6. Suppose a researcher modifies the experimental design as follows. For each condition, each participant is given to answer five mathematical questions (instead of a single question). His or her time performance is then assessed by taking the average time across all five questions. What are the benefits of this approach?

2 Descriptive Statistics

Please, answer the following questions as concisely as possible:

1. Calculate the mean, the median, and the standard deviation of time performances for the low- and the high-noise samples.
2. Give an estimate (point estimate) for the *population mean* and the *population standard deviation* of each condition.
3. Give an estimate (point estimate) for the *population* mean time difference ($\Delta T = T_{high} - T_{low}$) between the two conditions.
4. Based on the above estimates, can you *safely conclude* that math performance becomes worse under the presence of noise? Justify your answer.
5. Suppose we repeat the same experiment with a different set of participants. What do you expect? Will the mean time performance for each noise condition will be the same or different?

3 Probability Distributions

Please, answer the following questions as concisely as possible:

1. Suppose we know that the *population* difference ΔT in time performance follows a normal distribution, where the mean value is $\mu_{\Delta T} = 0.50$ sec and the standard deviation is $\sigma_{\Delta T} = 0.60$ sec. Given these known parameters, roughly draw the distribution of time differences (probability density function) in the available space of Figure 1. (Note: You can ignore the actual probability values on the y axis.)
2. What does the area under this distribution represent for negative time differences $\Delta T < 0$?
3. Suppose a student is randomly drawn from the above population (i.e., as described by the above distribution) and participates in the experiment. What is the probability (roughly) that the difference ΔT in time performance of this student will be between -0.1 sec and 1.1 sec?
4. Consider a Monte Carlo simulation that draws a large number (> 100000) of random samples of size $N = 14$ from the above probability distribution. The simulation calculates the mean of each sample and produces the distribution of *mean differences* in time performance. What is the form of this new distribution? Calculate its parameters, i.e., its mean and its standard deviation.

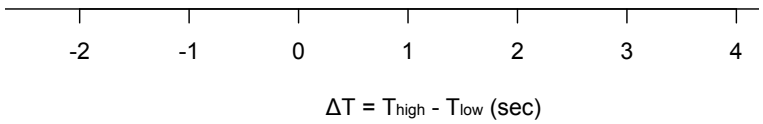

Figure 1: The probability distribution of differences in time performance between and high and low environmental noise

5. Suppose another researcher studies the same problem by following a different experimental design. Again, 14 students are asked to answer a simple multiplication question under the same two environmental conditions. However, the students are now only given 3 seconds to answer the question, and the researcher measures the number of successful answers. Results are as follows:

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**
Low Noise	1	0	1	1	1	0	0	1	0	1	0	0	0	1
High Noise	1	0	0	0	0	0	0	1	0	0	1	0	0	0

where 1 denotes a successful answer and 0 denotes a failure. What distribution would you use to model the number of successful answers for each condition (Low and High Noise)? Indicate the parameters of your distributions.

4 Standard Errors and Confidence Intervals

Please, answer the following questions as concisely as possible:

1. For the low-noise condition, provide an estimate of the *standard error of the mean* (SEM) based on the available sample. What does this SEM estimate represent?

2. Let's assume normal population distributions. Relying on your answer for Question 1, construct the 95% confidence interval (CI) of the mean for the low-noise condition. Take into consideration that the 97.5 and the 2.5 percentiles of the t distribution with $\nu = 13$ degrees of freedom are $t_{13,.975} = -t_{13,.025} = 2.16$.

3. The researcher uses the same approach to construct the 95% CI of the mean performance for the high-noise condition and finds it to be 95% CI = $[3.22 \text{ sec}, 4.21 \text{ sec}]$. Given this interval, could someone conclude that the mean time performance under high noise is *statistically significantly* higher than 3 sec for a significance level of $\alpha = .05$?

4. Given the above 95% CI (Question 3), the researcher concludes that *"there is a* 95% *probability that the true population mean of time performance under high noise is between 3.22 and 4.21 seconds."* Is this conclusion correct? Briefly justify your answer.

5. The researcher wants to find the 95% CI of the mean performance difference between the high- and the low-noise condition. Briefly present the steps that you would take to construct this confidence interval. Again, you can assume that the population probability distributions are normal.

6. Suppose the true mean difference between the two noise conditions is $\mu_{\Delta T} = 0.50$ sec. A large number of independent research teams runs an experiment to assess this difference and constructs a 95% CI to estimate it (as you described above). What percentage of such confidence intervals (CIs) will fail to include the value $\mu_{\Delta T} = 0.50$ sec?

7. A reviewer of the results of the experiment argues that time measurements are positively *skewed* and suggests using a logarithmic transformation to correct for the problem. The researcher responds that given the *Central Limit Theorem*, this is unnecessary. Who do you think is right and why?

5 Significance Tests

Please, answer the following questions as concisely as possible:

1. Formulate (describe in words) a Null Hypothesis H_0 and an Alternative Hypothesis H_1 for the experiment.

2. Assuming normal distributions, which significance test would you recommend using to test the above Null Hypothesis? Does this test make any additional assumptions (e.g., independent samples and equal variances)?

3. Suppose the result of a two-sided significance test is as follows: $t = 2.8802, df = 13, p - value = .01289$. How would you report this result? Briefly justify your answer.

4. Someone interprets the above p-value as follows: *"The probability that* H_0 *is true is* 1.289%*"*. Do you agree with this interpretation? Briefly explain your reasoning.

5. Imagine that a different researcher repeats the same experiment with 8 participants and finds a p-value equal to $p = .062$. What is the conclusion now? If now the conclusion is different, what are the possible reasons for this difference? In your opinion, which of these two experiments has a higher *statistical power*, i.e., a lower chance to result in a Type II error?

6. Imagine that the researcher in Question 5 decides to add four additional participants and then re-runs the significance test. The researcher now finds a p-value equal to $p = .049$ and is happy. Why is this methodology problematic?

6 Covariance, Correlation, and Regression

In order to further investigate the effect of noise on math performance, the researcher conducts a second experiment where 10 different participants are exposed to different levels of noise from 20 to 80 decibels. The results are as follows:

Participant)	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
Noise (db)	20	30	35	40	40	45	50	65	70	80
Time (sec)	3.9	4.5	3.5	5.5	4.4	5.9	4.8	6.4	7.5	5.7

Please, answer the following questions as concisely as possible:

1. From above sample, a researcher estimates the population covariance and finds $\hat{\sigma}_{Noise,Time} = 17.69$. How would you interpret the value that you found? Could you conclude that the two variables are independent? (If yes, under which assumptions?)

2. Do you think that the relationship between environmental noise and performance time is *causal* or not? Please, very briefly explain your answer.

3. The researcher calculates the Pearson correlation of the two variables and constructs a 95% CI by assuming a normal sampling distribution: $r = .77$, 95% CI $[.31, 1.22]$. What problem do you observe in this interval estimate? What is the reason for this problem?

4. The researcher decides to conduct a linear regression to describe the relationship between performance time and environmental noise. Which variable is the *predictor* (independent variable) and which variable is the the *outcome* (dependent variable)?

5. The plot in Figure 2 below shows the individual data points. Roughly draw the line that seems to best fit the sample based on the least squares criterion. Also show the residuals (errors) for any two different data points.

6. By taking into account Question 3 (and without making any calculations with the actual data), assess the R^2 of the linear regression model. How would you interpret the value that you found?

7. Consider again the original experiment (see Page 1), where 14 participants are tested under a low- and a high-noise condition. If you estimate the Pearson correlation between the low- and the high-noise condition, you can find that $r = .83$, 95% CI $[.54, .95]$. How would you explain this result?

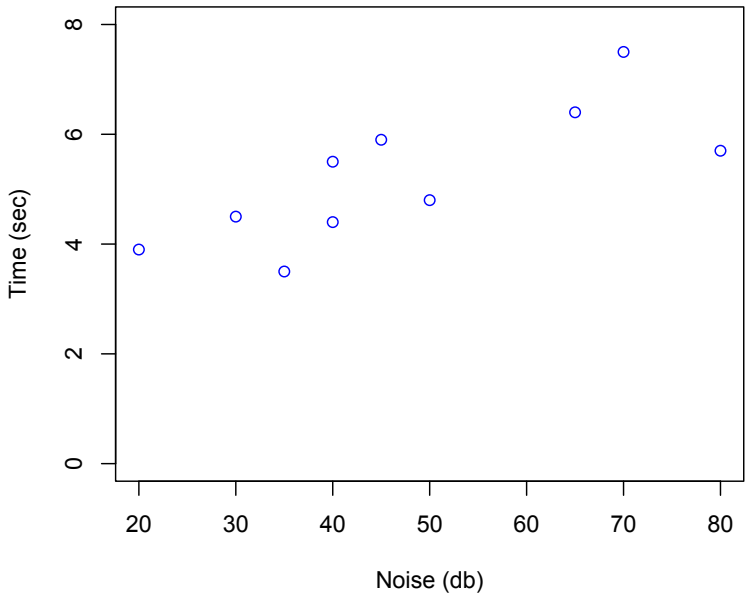

Figure 2: A scatter plot showing the data points for the observed data