Assignment 6 Submission deadline: 15th May 2025, 11:30 pm

1. **Gas Mileage**: Li Lei and Han Meimei want to investigate if the gas mileages of two brands of gasoline are the same or not. They plan to select cars and perform experiments to compare the mean mileages of Brand A and Brand B.

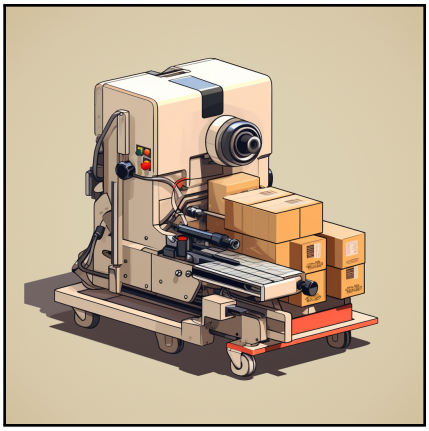
   1.1) **(2½ points)** Li Lei randomly selected many cars, assign half of them to use Brand A and the other half to use Brand B. He needs to use which of the following test for the investigation:

   A. One-sample t-test
   B. Paired t-test
   C. Independent two-sample t-test
   D. Z-test

   1.2) **(2½ points)** Han Meimei decides to use a different strategy. She randomly selects many cars and has each car use Brand A and Brand B. Then she compares the differences in mileage for each car. She needs to use which of the following test for the investigation:

   A. One-sample t-test
   B. Paired t-test
   C. Independent two-sample t-test
   D. Z-test

2. **Comparing Packing Machines**: In a packing plant, a machine packs cartons more quickly and consistently than a human does. It is supposed that a new machine will pack faster on the average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The results, in seconds, are shown in the tables below:



New machine				
42.1	41.3	42.4	43.2	41.8
41.0	41.8	42.8	42.3	42.7

Old machine				
42.7	43.8	42.5	43.1	44.0
43.6	43.3	43.5	41.7	44.1

2.1) **(2 points)** Write the null ($\boldsymbol{H_0}$) and alternative ($\boldsymbol{H_1}$) hypotheses.

2.2) **(3 points)** In you own words, explain what is the meaning of selecting a significance level of $\alpha = 0.05$.

2.3) **(5 points)** Compute the sample statistics: sample means and sample variances of the new and old machines.

2.4) **(2 points)** What statistical test should you use, and what assumptions should you check?

2.5) **(5 points)** Assume all the assumptions are met, and the variances of the two machines are equal. Compute the test statistic and $p-$value.

2.6) **(3 points)** The $p-$value you get from question **2.5)** is p, which of the following statement is true:

A. The null hypothesis is rejected, but it still has a probability of p to be true.

B. At the significance level of $\alpha = 0.05$, the null hypothesis is wrong.

C. The null hypothesis is rejected, but there is a chance that we made a wrong decision. The probability of the null hypothesis is true is p.

D. If the null hypothesis were true, the probability of observing the difference in our current samples or more extreme is small.

3. **Serum cholesterol level:** In the "serum cholesterol level" example during **Lecture 26**, we demonstrated that when taking a sample of size $n = 25$ and setting a significance level $\alpha = 0.05$, to test the null hypothesis ($H_0 : \ \mu \leqslant 180$ mg/100mL), we have a power of 0.702 when $\mu_1 = 200$ mg/100mL.

3.1) **(2½ points)** Now, for the same test, suppose we increase the sample size to $n = 50$ and assume the actual mean $\mu_1 = 200$ mg/100mL. What is the power of the test?

3.2) **(2½ points)** Now, for the same test, suppose the actual mean $\mu_1 = 190$ mg/100mL. What will be the required sample size to reach a significance level $\alpha = 0.05$ and a power of 0.702?

3.3) **(5 points)** Which of the following is true ?

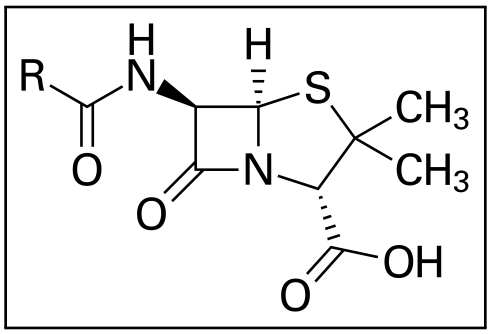
A. The power of the test will decrease when the sample size increases.

B. When the null hypothesis is false and the difference between μ_0 and μ_1 is small, you need larger sample size to reject the null hypothesis.

C. When the difference between μ_0 and μ_1 is small, you need smaller sample size to reach the same significance level and power.

D. With large enough sample size, you can always reject a null hypothesis.

4. **Penicillin in World War II:** Penicillin, the first widely used antibiotic, was discovered by Scottish biologist Alexander Fleming in 1928. He noticed that a mould called *Penicillium rubens* inhibited the growth of bacteria, leading to the development of this revolutionary medicine. During World War II, bacterial infections were rampant on the battlefield, often leading to severe complications and even death. The successful use of penicillin during World War II revolutionised medical treatment and saved countless lives. It marked the beginning of the antibiotic era, shaping modern medicine and paving the way for the development of many other life-saving antibiotics.



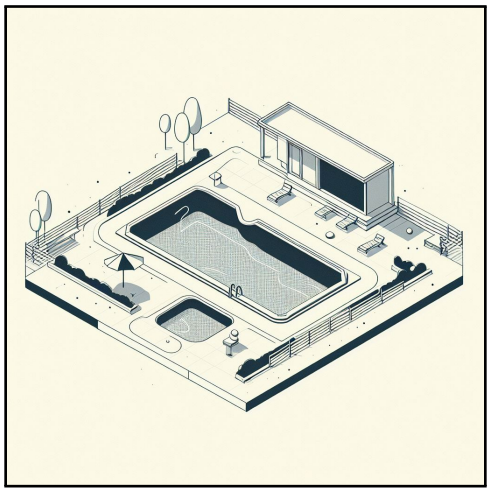
Referring to historical records and medical reports, you collect data on soldiers with bacterial infections who received penicillin treatment and soldiers who did not receive penicillin (control group). For the 2,000 soldiers in the penicillin group, the mortality rate was 12%; for the 1,800 soldiers in the control group, the mortality rate was 50%. From this data, you want to see if the effect of penicillin is statistically significant.

4.1) **(5 points)** State the null and alternative hypothesis.

4.2) **(5 points)** Compute the p-value of the statistical test.

4.3) **(10 points)** The p-value is at the scale of 100^{-151}, which can only be calculated using a computer program. To give you some context, the estimated total number of atoms in the observable universe is about 10^{80}. In your own words, describe how you should interpret this extremely small p-value.

5. **Type I & II errors**: A Type I error is when we reject a true null hypothesis. Lower values of α makes it harder to reject the null hypothesis, so choosing lower values for α can reduce the probability of a Type I error. The consequence here is that if the null hypothesis is false, it may be more difficult to reject using a low value for α. So using lower values of α can increase the probability of a Type II error. A Type II error is when we fail to reject a false null hypothesis. Higher values of α makes it easier to reject the null hypothesis, so choosing higher values for α can reduce the probability of a Type II error. The consequence here is that if the null hypothesis is true, increasing α makes it more likely that we commit a Type I error (rejecting a true null hypothesis). Now consider the following examples: employees at a health club do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment. We can state the hypotheses for their test as $\{H_0$: The water quality is acceptable$\}$ *v.s.* $\{H_1$: The water quality is not acceptable$\}$.



5.1) **(2½ points)** What would be the consequence of a Type I error in this setting?

A. The club closes the pool when it needs to be closed.
B. The club closes the pool when it doesn't need to be closed.
C. The club doesn't close the pool when it needs to be closed.
D. None of the above.

5.2) **(2½ points)** What would be the consequence of a Type II error in this setting?

A. The club closes the pool when it needs to be closed.
B. The club closes the pool when it doesn't need to be closed.
C. The club doesn't close the pool when it needs to be closed.
D. None of the above.

5.3) **(2½ points)** In terms of safety, which error has the more dangerous consequences in this setting?

A. Type I error
B. Type II error

5.4) **(2½ points)** What significance level should they use to reduce the probability of the more dangerous error?

A. 0.01
B. 0.025
C. 0.05
D. 0.10

6. **Treatment of smallpox in the 18th century England:** Smallpox was extremely common before vaccination, infecting many Londoners by adulthood in the 18th century. It was one of the most devastating diseases of the time, with multiple strains including hemorrhagic forms with high mortality rate. 18th century London experienced regular smallpox outbreaks due to poor sanitation and overcrowding. The disease disproportionately impacted the poor.

   6.1) **(2½ points)** According to the **humoral theory** back in that time, mercury can extract the infectious "morbid matter". Therefore, doctors think mercury treatment was effective in shortening the course of infection. In a hypothetical scenario where you want to test if mercury is effective, write the null and the alternative hypotheses.

   6.2) **(5 points)** However, mercury treatments had harsh side effects such as organ failures. What are the type I and type II errors in this context? Which one has a more dangerous consequences?

   In the 18th century England, standard treatment for smallpox involved mercury dosing. However, a doctor was skeptical about the effect of mercury. After a major smallpox outbreak in 1760, he decided to conduct trials comparing treatments[1].

   6.3) **(5 points)** He enrolled 50 patients, randomly assigning 25 to receive an herbal remedy and 25 to a standard mercury-based treatment. Recovery times were recorded in days. In the herbal group, the mean recovery time was 12 days, with a standard deviation of 3 days; in the mercury group, the mean recovery time was 14 days, with a standard deviation of 4 days. Use the 8-step procedure for NHST that we learnt during the lecture to test if there is a difference in recovery time between the herbal and mercury treatment groups. Assume all assumptions are satisfied.

   6.4) **(5 points)** Encouraged by initial results, he repeated the trial on 100 additional patients the following year, with 50 in each group. In the new experiment, the mean recovery time was 11 days with a standard deviation of 2 days for the herbal

[1]The experiment and data in this question are all made up. For the real data and experiment conducted by the English doctor William Watson, you can read this NEJM article Clinical Investigation of Smallpox in 1767.

treatment group; the mean recovery time was 13 days with a standard deviation of 3 days for the mercury treatment group. Use the 8-step procedure for NHST that we learnt during the lecture to test if there is a difference in recovery time between the herbal and mercury treatment groups.

6.5) **(2½ points)** Compare the p-values from **3.3**) and **3.4**). What do you observe?

7. **A new treatment for diabetes:** Dr. Smith is a researcher studying a new drug called "DIDEILE" aimed at controlling blood glucose levels in patients with type 2 diabetes. In her latest clinical trial, she recruited 50 patients with poorly controlled diabetes and randomly assigned them to either receive the new drug or a placebo daily for 8 weeks. At the beginning of the study and again at the end of the 8 weeks, fasting blood glucose (FBG) levels were measured for all patients. Dr. Smith is now interested in analysing the results to see if the new drug was effective at lowering FBG levels compared to the placebo. Assume the FBG levels follow normal distributions and the drug does not change the dispersions of FBG levels.

7.1) **(2½ points)** In this context, write the null and alternative hypotheses. What is a type I error? What is a type II error? Which one has a more dangerous consequences?

7.2) **(5 points)** At the start of the clinical trial, the mean FBG level for the 25 patients in the drug group was 160 mg/dL with a standard deviation of 20 mg/dL. For the 25 patients in the placebo group, the mean FBG level was 165 mg/dL with a standard deviation of 18 mg/dL. Use the 8-step procedure for NHST that we learnt during the lecture to test if the starting FBG levels were different between the two groups. Was Dr. Smith's randomisation successful?

7.3) **(5 points)** At the end of the 8-week study, the mean FBG level of the patients in the drug group was 130 mg/dL with a standard deviation of 15 mg/dL. For the placebo group, the mean was 160 mg/dL with a standard deviation of 17 mg/dL. Use the 8-step procedure for NHST that we learnt during the lecture to test if the new drug was effective at reducing FBG levels compared to the placebo.

7.4) **(5 points)** Apparently, Dr. Smith was using a cross-sectional experimental design. What other experimental design can she use to test if the drug is effective in lowering the FBG levels, and how is she going to perform the experiment and statistical test?

7.5) **(2½ points)** Provide a clinical interpretation of your analysis from the previous two questions. Did the results support the effectiveness of the new diabetes drug? How could Dr. Smith's findings be useful to other researchers or clinicians studying new treatments for diabetes?

8. **Pulmonary Disease:** A possible important environmental determinant of lung function in children is the amount of cigarette smoking in the home. The forced expiratory volume (FEV) is a metric used to evaluate the lung function. Suppose this question is studied by selecting two groups: Group 1 consists of 23 nonsmoking children 5-9 years of age, both of whose parents smoke, who have a mean FEV of 2.1 litre and a standard deviation of 0.7 litre; group 2 consists of 20 nonsmoking children of comparable age, neither of whose parents smoke, who have a mean FEV of 2.3 litre and a standard deviation of 0.4 litre. Assuming FEV follows a normal distribution in the population and smoking does not change the dispersion of the distribution. Work out the following questions:

8.1) **(5 points)** Dr. Li Lei wants to test if the mean FEVs of group 1 and group 2 are the same or not. Use the 8-step procedure for NHST that we learnt during the lecture to help him with that.

8.2) **(5 points)** Dr. Han Meimei wants to test if the mean FEV of group 1 is lower than that of group 2. Use the 8-step procedure for NHST that we learnt during the lecture to help her with that.

The above data is regarded as the pilot study. Assuming the estimates of the population parameters in the pilot study are correct, then:

8.3) **(5 points)** At a significance level of $\alpha = 0.05$, how much power does Li Lei have? How much power does Han Meimei have?

8.4) **(5 points)** How many children are needed in each group (assuming equal numbers in each group) to have a 95% chance of detecting a significant difference with $\alpha = 0.05$? How many does Li Lei need? How many does Han Meimei need?