# Problem set 1

In this problem set, we will explore statistical and computational concepts that underlie various topics in computational biology and bioinformatics. Some questions will ask you to calculate the answer, while some will ask you to provide explanation/justification for your ideas.

### Differential expression analysis

You used microarray to study the effects of a new drug on the transcriptome of human cell lines and obtained the following log-expression data for gene A.

| **Control (untreated)** | | | | | **Drug-treated cells** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4.32 | 3.74 | 3.94 | 4.28 | 3.81 | 5.20 | 4.83 | 4.97 | 5.01 | 4.86 |

You want to test whether the expression of gene A is significantly up-regulated after drug treatment. Because you have learned that log-expression data from microarray follow normal distribution, you plan to use *t*-test.

**Q1**: Will you use paired or un-paired *t*-test? Why?

**Q2**: Perform the selected *t*-test on any software of your choice. Explain your work and report the resulting p-value.

**Q3**: If we don't want to assume that the data is normally distributed and want to use non-parametric tests which test would you use? Why?

**Q4**: Perform the selected non-parametric test on any software of your choice. Explain your work and report the resulting p-value.

### Correction for multiple testing

**Q5**: If your microarray dataset from above contains data for 3000 genes, what would be your adjusted p-value for the results in **Q2** and **Q4**? Show how you derive the answer.

Consider the following *t*-test results on 5 genes:

| **Gene** | **A** | **B** | **C** | **D** | **E** |
|---|---|---|---|---|---|
| P-value | 0.046 | 0.0023 | 0.015 | 0.083 | 0.026 |

**Q6**: Using Bonferroni method, which gene would pass the test at p-value cutoff of 0.05

**Q7**: Using Benjamini-Hochberg method, which gene would pass the test at false discovery rate (FDR) cutoff of 0.05

## Functional enrichment analysis

After analyzing the above microarray data, you obtained a list of 500 genes whose expression significantly differ between control and drug-treated cells. You suspected that this drug affects the ribosome biogenesis pathway. Among 20,000 human genes, 70 are related to ribosome biogenesis.

**Q8**: If there is no relationship between the drug and ribosome biogenesis pathway, what is the expected number of ribosome biogenesis genes among your list of 500 genes?

**Q9**: After checking the annotation, you identified 30 ribosome biogenesis-related genes in your list of 500 genes. What is the probability of observing this result by chance? *You can leave the expression in the form of* $\binom{N}{k}$ *without calculating them.*

**Q10**: How would you calculate the p-value of this observation. Here, the definition of "the same or more extreme" observation would be "the same or higher number of ribosome biogenesis genes among the list of 500 genes". *Again, you don't need to calculate the actual value.*

## Bayes' rule

The basis of Bayesian statistics is the following identity

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(B \mid A)} \times \mathrm{P(A)}}{\mathrm{P(B)}}$$

which allows us to calculate the conditional probability $\mathrm{P(A \mid B)}$ that might be difficult to directly interpret, in terms of $\mathrm{P(B \mid A)}$ that is easier to handle.

**Q11**: Use Bayes' rule to express $\mathrm{P(Uninfected\ by\ COVID \mid Positive\ ATK\ test)}$.

The manufacturer of ATK reported that (i) if a person is uninfected, then the ATK would turn positive only 1% of the time, and that (ii) if a person is infected, then the ATK would turn positive 99% of the time. A survey of COVID-19 infection in Bangkok revealed that an estimated 20% of the population is infected.

**Q12**: Calculate $\mathrm{P(Uninfected\ by\ COVID \mid Positive\ ATK\ test)}$

**Permutation test**

You want to show that drug D can suppress cytokine production in immune cells in a dose-dependent manner. So, you performed an experiment to measure cytokine level after administering various doses of drug D to cell cultures.

| **Control** | **Dose = 0.1 M** | **Dose = 0.5M** | **Dose = 1M** |
|---|---|---|---|
| Cytokine level = 1.36 | 1.32 | 1.06 | 0.99 |
| Cytokine level = 1.42 | 1.28 | 1.19 | 1.03 |
| Cytokine level = 1.38 | 1.25 | 1.15 | 0.92 |

You may be tempted to perform multiple *t*-tests to show that cytokine levels at higher doses are lower than cytokine levels in the control or at lower doses. However, in this problem set, we will explore a different way to tackle this problem.

**Q13**: How would you quantify the **rate of reduction in cytokine level per unit dose of D**? Note that we are looking for a **numerical score**. *Hint: Try plotting this data.*

Because we have no way to model the distribution of the **rate of reduction in cytokine level per unit dose of D**, we will propose a **null hypothesis** and show that the p-value is low instead.

**Q14**: Formulate an appropriate **null hypothesis**. *Hint: If drug D does not affect cytokine level in a dose-dependent manner, what would the* ***rate of reduction in cytokine level per unit dose of D*** *be?*

**Q15**: Describe how permutation test can be used to calculate the p-value for your null hypothesis in **Q14**.

**Information theory**

In class, we learned that Entropy is a measurement of randomness of a population. If the population is pure (i.e., contain only one type of objects), then the Entropy is minimized at zero. If the population is uniformly mixed (i.e., equal number of everything), then the Entropy is maximized.

Let's consider a population of people and let $p$ be the proportion of cancer patients in this population. Hence, the Entropy is $-p\log_2(p) - (1-p)\log_2(1-p)$

**Q16**: Visualize the graph of Entropy as a function of $p$ on the range [0, 1].

**Q17**: Gini impurity is defined as $p(1-p)$. Visualize the graph of Gini impurity and discuss whether this measure can also be used to describe the randomness of a population.

You developed two tests for diagnosing cancer, A and B, with the performance shown below. You plan to classify people who passed the test as potentially having cancer and people who failed the test as normal. The objective is to decide which test is better.

| | **Test = Passed** | | **Test = Failed** | |
|---|---|---|---|---|
| **Test** | **Cancer** | **Healthy** | **Cancer** | **Healthy** |
| **A** | 46 | 22 | 4 | 28 |
| **B** | 36 | 5 | 14 | 45 |

A good test is one that can split the population into a pure group of cancer patients and a pure group of healthy people (i.e., low Entropy and low Gini impurity scores). Hence, the sum of Gini impurity scores of the Passed group and the Failed group can be used to compare the two tests and identify the better one.

**Q18**: Calculate the Gini impurity score for tests A and B. Which one would you pick? Why?

**Q19**: Repeat the analysis with Entropy. Will you pick the same test as in **Q18**?

**Q20**: Can you propose another way to pick between tests A and B? Provide your reason(s).

**Droplet Digital PCR** (see Wikipedia for more details) (**This is a challenging problem**)

In droplet digital PCR experiment (ddPCR), DNA molecules permeate into oil droplets at random and the polymerization reactions then occur inside those droplets. Droplets undergoing reactions will emit fluorescence signal while empty droplets will not. To interpret the result of droplet digital PCR, we count the number of "positive" droplets. The higher the number of "positive" droplets, the higher the number of DNA molecules.

ddPCR can be used to determine the frequency of a particular allele of interest by designing PCR probes that are specific to that allele. Therefore, given the same starting DNA concentration, samples with higher frequencies of the allele of interest would result in higher number of positive droplets.

The first step toward developing a statistical testing framework for ddPCR result is to model the permeation of individual DNA molecules into individual droplets using Poisson distribution.

**Q1***: Explain why Poisson distribution can describe this physical process. *Hint: Study the descriptive definition of Poisson distribution (not the mathematical formulation) and try to fit it to the characteristics of this process.*

**Q2***: Given that there are $M$ DNA molecules of the allele targeted by our probe and $N$ total droplets, what should be set as the value of $\lambda$, the expected number of "events", for the Poisson distribution? Express $\lambda$ in the form of $M$ and $N$. *Hint: Poisson distribution is described by a single parameter $\lambda$ which specifies the mean value of the distribution.*

**Q3***: Write the probability that a droplet will be "positive" in the form of $M$ and $N$? *Hint: It might be easier to calculate the probability that a droplet will be "negative". What must happen for the droplet to be "negative" (assuming that there is no defect in the assay)?*

Using the probability that a droplet will be "positive" derived in Q15, we can model the probability of observing exactly $k$ "positive" droplets using a Binomial distribution.

**Q4***: Explain why Binomial is appropriate for modeling this event. *Hint: Study the descriptive definition of Poisson distribution (not the mathematical formulation) and try to fit it to the characteristics of this event.*

**Q5***: Write the probability of observing exactly $k$ positive droplets in the form of $M$, $N$, and $k$. *Hint: Binomial distribution is described by two parameters, the number of trials and the probability of success for each trial. Can you express those two parameters in the form of $M$ and $N$. If you cannot solve **Q15**, you may assume that the answer is $p$ and use that here.*

Finally, let's assume that there are 1,000 droplets in the assay and the total number of DNA molecules (all alleles) is 10,000. A fraction $f$ of DNA molecules corresponds to the target allele. A ddPCR experiment yielded 200 positive droplets.

**Q6***: Write the likelihood of this observation in the form of $f$.

**Q7***: Using the maximum likelihood idea, how would you estimate the fraction $f$ of target allele?

**Q8***: Solve for the maximum likelihood estimate for $f$.