**PART-1**

1. **The binomial and negative binomial are discrete distributions that are related in some way. Describe the difference between these, and outline in broad terms an illustrative case study (in biology) where each would be applied.**

The binomial describes the probability of obtaining a particular number of “successes” from a series of independent Bernoulli trials. The negative binomial (NB) describes the number of “failures” prior to obtaining a given number of “successes”, i.e., the waiting time before the last desired success.

The binomial CDF is the “survival” function of the NB, and vice versa. If the CDF is F(X) = P(xX), then the survivor function is P(x>X) = 1 – F(X).

Examples may include:

* for Binomial, we discussed in class the probability of getting x transformants out of y bacterial colonies.
* for NB, an example would be how many colonies you have to pick in order to get 3 “good” clones.

1. **What is a log-normal distribution and when is it sometimes useful? Give an example.**

A log-normal distribution is when taking the log of the values produces a normal distribution. This is helpful because it allows us to use some very important statistical tests such as T-tests and ANOVA, that are based on the assumption that the data is normal. Some examples are gene expression and blood pressure.

1. **Why is the negative binomial a better model than the Poisson for RNA-seq data? (This relates to noise in gene expression studies as a function of gene expression levels and something called “overdispersion”)**

It is appropriate to model data with a Poisson distribution, where the mean and the variance are given by the same parameter, λ and the variance is proportional to the mean. This is true for count data for technical replicates. When considering biological replicates, however, it turns out that the variation in RNA-Seq counts increases with the number of counts per feature (expression level) and is therefore ‘overdispersed’. In such cases, the Poisson is no longer the best model for the data.

Instead, the NB is used to model the uncertainty in the variance as the variation is proportional to the mean, with an added term to account for the dispersion:

σ2 =μ+αμ2 where α is the dispersion parameter. For α > 1, the dispersion is greater than the mean; as α goes to 0, the NB converges on a Poisson distribution.

**PART-2**

1. **Hypotheses, Error and Power  
   a. What is a "null hypothesis"? What “alternative” hypotheses can be tested?**

A null hypothesis is a specific claim about the value of a population parameter that is made for the purpose of argument. Often Ho states that there is no change or difference between two samples. The alternative hypothesis includes all other feasible values of the population parameter other than the ones stated in the null hypothesis.

**b. Define Type I and Type II errors and clearly explain the difference between them**Type I error (also known as false positive or alpha) occurs when the alternative hypothesis is erroneously accepted despite the null hypothesis being true. A Type II error (also known as false negative or beta) on the other hand occurs when an alternative hypothesis is erroneously rejected despite the null hypothesis being false.

**c. What is power, and what’s the tradeoff between error and power?**

Power is the probability that a test will correctly reject the null hypothesis when the alternative hypothesis is true — i.e., the probability of avoiding a type II error. It can therefore also be thought of as the ‘True positive rate’ or (1 – beta).

Power depends on the significance threshold (alpha), the effect size (E), the sample size (n), and the population variance (sigma2). Power is proportional to alpha, so increasing alpha gives more power to detect true positives but also produces more false positives. Decreasing alpha on the other hand decreases false positives at the expense of false negatives.

1. **P-values  
   a. What is a p-value?**

A p-value is a measure of significance in a statistical result. It represents the probability that a value at least as extreme as the value observed would be obtained purely by chance, given the null hypothesis Ho is true.

**b. What are the shortcomings of p-values?**The p-value resulting from repeated experiments may vary widely if the sample sizes are small and may not be representative of the true population parameters. So, for example, drawing random 5 samples from a population may give a mean that is significantly different from the control population, but a different 5 samples may not produce a significant p value. P-values also use arbitrary cutoffs.

**c. Is it possible for something to be significant but not important? Explain**

This can happen when there is a small effect size. If a small value is observed reliably, i.e., it is highly repeatable and with small variance, it may be statistically significant. However, the magnitude of the difference may not be sufficient to warrant action. Example:

* If therapeutic efficacy of a drug is real but very small, then it is probably not worth it to introduce into the market.
* In association studies, a significant association with a marker will usually not represent a causal link and so might not be directly significant (correlation is not causation. But this is not the answer I was looking for.)

1. **Confidence Intervals**  
   **a. What is a confidence interval? What, specifically, does a 95% CI mean?**

A confidence interval provides a range estimate for a population parameter. A 95% CI indicates that 95% of the time, when taking samples of the same size, the true statistic (mean) is expected to be contained in the interval (95/100 samples will contain the expected value).

**b. Why and how are confidence intervals useful? In particular, how do confidence intervals complement p-values?**

The CI represents the uncertainty in the estimate (precision): a smaller CI represents higher confidence in the parameter estimate. In addition, the CI also provides an estimate for the magnitude and direction of the effect.

1. **T-tests**  
   **a. What is the purpose of the t-test?**

There are two main applications of a T-test. One is to see if the sample collected is from a given population, in which case the mean of the population is provided and sd is often estimated using sample sd. The second case is to see if two different samples are obtained from the same distribution, in which case the difference in the means of the two samples should be close to 0.

**b. What are some assumptions about that data that need to be true in order for someone to use the t-test?**

The main assumptions are that the:

a. Sample data are normally distributed.

b. Sample data are random and independent.

c. Variance in the samples and population is the same.

**c. What is the difference between a one-sided and two-sided t-test? What are the null and alternative hypotheses for each?**

The null hypothesis of both one-sided and two-sided t-tests is that the two samples come from the same parent distribution.

For a one-tailed test, one alternative hypothesis could be that the mean of the first sample is greater than that of the second sample while another alternative hypothesis could be that the mean of the first sample is smaller than the second.

A two-tailed test asks if the mean of the second sample is either more OR less than that of the first and is more appropriate if we are interested in change regardless of direction.

**d. What does a significant p-value of such a t-test mean?**

A significant p-value means the null hypothesis should be rejected and the alternative hypothesis should be accepted.

**e. How are confidence intervals for t-tests determined for two-sample comparisons?**

If the t-statistic is greater in magnitude than the critical value, then the difference is considered to be significant. The confidence interval of two sample t-test is determined using the t-distribution of the difference of the means and by calculating the sd of the samples together

1. **Multiple Hypothesis Testing**  
   **a. Why is multiple hypothesis testing important for high-dimensional data?**

For studies like genome-wide gene expression, where we are performing tens of thousands of tests in one dataset, the likelihood of obtaining false positives by chance is greatly increased. For example, for 20,000 t-tests of differential expression at a significance threshold of 5%, 1000 genes will always be considered as “differentially expressed” whether or not this is really the case.

**b. How does controlling for False Discovery Rate (FDR) work? Outline the general framework for controlling the FDR to 5%**

The FDR is the false discovery rate, a.k.a. Benjamini-Hochberg correction. It specifies the rate of false positives you are willing to accept within a set of statistically significant results. In genomics, typical FDR values are 5% or 10%. The FDR uses the q-value as a cutoff rather than the p-value. To compute the FDR, the p-values are sorted form smallest to largest and compared with the BH q-value (rank/number of samples). All the p-values that are less than the q-value are significant.

**PART-3**

1. **ANOVA models  
   a. What does ANOVA test?**

ANOVA tests whether there is a significant difference between more than two groups.

**b. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables?**

Response is continuous, predictor is continuous or categorical

**c. Why is it useful to consider interaction terms instead of just marginal effects?**

Interaction terms are helpful in modeling the combinatorial effect of two different factors. For example, a specific fertilizer may work differently for two different varieties of the same plant. In one variety the growth of the plant may be twice as much as the other. If one models the marginal effect only then both varieties will show the positive growth, however, to see the different in growth you have to look at the interaction term.

**d. For the image of an ANOVA result below:  
i. What is the Df column describing?**

The Df column displays the degrees of freedom for the groups and the residuals

**ii. What is the Sum Sq column describing?**The Sum Sq column displays the sum of squared deviations for the groups i.e., the total variation between the group means and the overall mean and the sum of squared deviations for the residuals i.e., the variation between the group means and the data.

**iii. How is the F distribution created? How is the F-statistic calculated?**

The f-distribution is created by generating all possible values of the f-statistic. The f-statistic is calculated by first computing the sum of squared deviations of each group around the grand mean and the sum of squared deviations of the data around group means. These values are then divided by the appropriate degrees of freedom to calculate the mean square groups and mean square error. The f-statistic is the ratio of the mean square groups and the mean square error.

**iv. What is the null hypothesis of the F-test?**

The null hypothesis of the F-test is that the ratio of the mean square groups and the mean square error is close to 1 i.e., the variance of the groups being compared is equal.

1. **Bayesian Models**

**a. What is the fundamental conceptual difference between Bayesian statistics and "frequentist" statistics?**

Frequentist statistics views the population parameters as a ground truth that is being estimated by the data measured. In the Bayesian worldview, the ground truth is considered as a distribution of probabilities itself. Thus, the values being estimated by the response variables do not represent single, fixed outcomes but a family of possibilities, each with some degree of probability.

**b. Outline the basic framework for Bayesian analysis.**

A typical Bayesian analysis uses priors (known probabilities or beliefs based on past experiences e.g., the known rate of a disease in a population) and likelihoods (known conditional probabilities e.g., the false positive/false negative or true positive/true negative rates of a diagnostic test) to compute posteriors or unknown conditional probabilities using the Bayes theorem (such as the probability of a test correctly diagnosing a patient with the disease).

**c. What is a prior?**

In Bayesian statistics priors signify information about past experiences (or best guesses) that can be used to update the estimates, or posterior probabilities.

**d. Give an example (e.g. from class) to which you could apply a Bayesian model and discuss how your estimates might change with more data.**

An example discussed in class was estimating the chances that a random Down’s syndrome test would produce a positive result given that the fetus actually has the disease. As more data is collected, the prior probabilities may be updated and therefore produce different results.

Maybe state the priors and conditional probabilities explicitly?

**PART-4**

1. **Distance  
   a. Explain the difference between Euclidean distance and Manhattan distance.**

Euclidean distance measures the shortest distance between two points using the Pythagorean theorem. Manhattan distance on the other hand uses a ‘city-block’ or ‘grid-like’ approach to find the sum of the absolute distances across all coordinates between two points.

**b. Explain the relationship between covariance and Pearson correlation. What are the similarities? What are the differences? How does Pearson correlation differ from R2?**Both covariance and Pearson correlation are measures of association between variables. While covariance depends on the scale of the variables being investigated, Pearson correlation normalizes the covariance to produce correlation values ranging from -1 to 1. R2 is the square of the correlation coefficient and represents the proportion of variation in the dependent variable explained by variation in the independent variable. Unlike Pearson correlation, its values range from 0 to 1.

**c. When is it more appropriate to use Euclidean Distance vs. Correlation to cluster genes (and vice versa)? Why?**

Euclidean distance would cluster together genes that show similar magnitude of expression while correlation would cluster together genes that exhibit the same pattern of expression across different samples or timepoints. The correlation is independent of scale because it measures changes relative to the total variation of each variable. Whether Euclidean distance or correlation are more appropriate for clustering therefore depends on whether the magnitude or patterns of expression changes are more important for the question being investigated.

Correlation is more often used for hierarchical clustering, whereas k-means clustering uses average Euclidian distance to compute squared deviations of data points from the centroids. This operation may be performed on data that has already been standardized.

1. **Principal Components Analysis**
2. **What's the basic idea behind PCA, and how are principal components identified?**

Principal components analysis (PCA) uses linear combinations of predictors to identify a new coordinate system that explains most of the variation in the original data. The first PC explains the largest proportion of variation, the 2nd PC explains the second most variation, etc. It is then possible to identify the minimal number of dimensions required to explain most of the variation in the data. This enables visualization and analysis of major factors contributing to observed results.

1. **Why is it useful to use dimensional reduction methods like PCA?**

Sometimes there are many measured variables that can be used as predictors, but some of them may be correlated and thus do not offer much new / independent information for prediction. Dimensional reduction allows identification of a smaller number of predictive variables, resulting in a simpler model.

1. **How many principal components can be calculated for FACS data that has 2500 observed cells and eight features (six fluorescent data channels, side scatter, and forward scatter)? Why?**

Since there are 8 variables, there are 8 principal components, which are linear combinations of the 8 variables that together can explain all the variation among the 2500 cells.

1. **Give an example of an application for PCA and what you would gain from it.**

In single-cell RNA-seq, PCA helps identify a subset of genes that are highly varying across the samples, enabling the identification of marker genes associated with specific cell types.

**PART-5**

1. **Describe a simple scenario in which you would use a contingency table.**

You have two groups in which some proportion of each displays a certain characteristic, and you want to determine whether the proportions are the same or different between the two groups.

An example from class asked whether there is an association between the incidence of breast cancer among women who first gave birth below or above the age of 30.

1. **How do you calculate the Chi-Square test?**

To perform the Chi-square test, all observations in the data are converted into categorical variables (if needed) and a contingency table is prepared. The Chi-squared Test statistic, X2, which compares individual proportions in each group to the expected proportion based on the population mean estimate, is then calculated. For each cell, the square of the difference between the Observed and Expected values is divide by the Expected value and summed.

A p-value associated with this statistic is calculated based on the Chi-square distribution which uses a normal (continuous) approximation of a binomial (discrete) distribution.

1. **How do you calculate the p-value for Fisher’s Exact test?**

(Part of the answer is from last year’s key). Fisher’s exact test is a special case of the hypergeometric distribution, which gives the probability of x successes when sampling without replacement. Fisher’s test follows this model because, with fixed row margins, changing the value of one cell in a contingency table necessarily changes the values in the others. The p-value of the Fisher’s Exact test is calculated by summing the probabilities of all possible contingency tables that represent equal or greater deviation from independence than the observed table.

**PART-6**

1. **Why might someone want to use resampling instead of a t-test?**

When you have reason to believe that data may not be normally distributed. Or you just want to get empirical estimates using the data itself rather than making any assumptions about them.

1. **How can someone determine if the difference of the means from two samples is significant using the resampling method? Describe the steps in detail.**

There are two different groups, groupA and groupB and let’s pretend each have 10 values.

We can sample 10 values from groupA (with replacement) and again sample 10 values from groupB (with replacement) and record the difference of their means. If we do this 1000 times, we will be able to determine the distributions of the means which will provide the standard error and the confidence intervals. This can be used to calculate the p-value.