# Distributions

1. **What is a probability density function (PDF) vs. a cumulative distribution function (CDF)? Describe at a conceptual level how you can obtain p-values using the density functions.**
2. **What is the central limit theorem? How is it useful for the analyses of biological data?**
3. **How are the exponential and Poisson distributions related conceptually? Give an example of two related questions that can be answered using these two distributions.**

The Poisson is a discrete distribution describing the number of events per unit time, and the exponential is a continuous distribution describing the length of time between events. For a Poisson process, then, if events happen at a rate of λ per unit time on average, an average of λt events will occur per t unit of time.

For example, the exponential survivorship function describes situations where the probability of mortality is the same for all individuals in a population, since the rate of events is always the same for a Poisson process. In other words, the chance of dying is independent of age! This is not true for mammals, but it is true of some birds, rodents, lizards, and sea animals.

1. **Why is the negative binomial a better model than the Poisson for RNA-seq data?**

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.

# Hypothesis Testing

1. **A. What kinds of errors can arise in comparing two samples? B. What’s the tradeoff between error and power? Be sure to define error types and power in your answer.**

Type I error: False Positive (alpha) – erroneously accept HA when H0 is true

Type II error: False Negative (beta) – erroneously reject HA when it is true

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.

Power depends on the significance threshold (alpha), the effect size (E), the sample size (n), and the population variance (sigma2).

Power = True positive rate = 1 – beta = E\*alpha/(sigma/sqrt(n))

Decreasing alpha (stringency of the test, e.g. from 5% to 1%) will decrease false positives at the expense of false negatives, and vice versa. Power is proportional to alpha, so increasing alpha will give more power to detect true positives, but also more false positives.

1. **(a) Define a p-value. (b) What is the main shortcoming of p-values? (c) Can a result be significant but not meaningful? Explain.**

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.

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.

A result can be significant but not meaningful 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.)

**4) (a) Define a confidence interval. What, specifically, does a 95% CI mean? (b) Why and how are confidence intervals useful? (c) Why do confidence intervals provide an important complement to p-values?**

a. 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. 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.

c. Since p-value can vary a lot from experiment to experiment (particularly if the sample sizes are small), they may not be representative of the true population parameters. P-values also use arbitrary cutoffs. The CI gives an idea of how reliable we think our estimate would be in repeated experiments.

1. **(a) What is a two-sample t-test used for? (b) What assumptions need to be met in order to apply a t-test? (c) Define in words how the test statistic is calculated. (d) How is the confidence interval related to the t-distribution?**

a. To test 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. The main assumptions are that the:

* Sample data are normally distributed.
* Sample data are random and independent.
* Parent populations have equal variances.

c. The t-statistic is the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error ( where SE = SD / sqrt(sample size) ). If the t-statistic is greater in magnitude than the critical value, then the difference is considered to be significant.

d. The confidence interval is determined using the standard error and the critical value for the t-distribution for the available degrees of freedom (for two-sample tests, this is the distribution of mean differences). 95%CI = mean estimate +/- t(crit)\*SE(y). For alpha=0.05, t(crit) ~ 2.78.

**5) (a) When is a non-parametric (rank-based) test preferred over a parametric test?**

**(b) Which nonp-parametric test compares samples from two independent populations to find if the means are different?**

**(c) Describe the basic idea behind such a test, and how the test statistic is calculated.**

**(d) What test statistic is used to determine significance, and what kind of distribution does it follow?**

* 1. Parametric tests use standard models of statistical distributions to estimate significance and confidence intervals and assume that the populations from which samples are taken follow a specific distribution. Non-parametric tests make no such assumptions and so are preferred when one or more assumptions required for a parametric test is violated.
  2. Wilcoxon Rank Sum Test.
  3. Combine the data and rank them. If there is no difference between groups, the sums of ranks (T) should be about the same.
  4. The W statistic is the smaller of T – T(min) between the two groups. W approximately follows a normal distribution with known mean and SD, which is used to determine significance.

1. **(a) Why might you choose resampling to test for significant differences instead of a standard statistical test?**
2. **How significance determined using resampling methods? Describe the steps you would use to obtain a bootstrap p-value and CI for the mean difference between two samples.**
   1. 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.
   2. Resampling mehtods empirically determine significance and confidence intervals by sampling different subsets of the data many times. Significance is determined as the proportion of data that fall at the extremes of the sampled data (e.g. 5% of the data).

For example, let’s say we have 10 measurements for each of two different samples, groupA and groupB. We can sample 10 values from each group ( 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 a p-value for the difference between the groups.

1. **(a) Why is multiple hypothesis testing important for high-dimensional data?**

**(b) What are two popular methods of p-value adjustment, and what is the essential difference between them?**

a. 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. Bonferroni = p-value / total # of tests. It controls the FWER, i.e. the probability that the null hypothesis is false for at least one test.

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.

1. **(a) What kind of data would you use a contingency table for? Give an example.**

**(b) What is Fisher’s exact test? How is the test statistic calculated, and what distribution does the test statistic follow?**

**(c) What are the advantages and limitations of Fisher’s test vs. a Chi-squared test?**

**8) (a) What kind of data would you use a contingency table for? Give an example.**

**(b) What is a Chi-square test? How is the test statistic calculated, and what distribution does the test statistic follow?**

**(c) Under what conditions is a Chi-squared test not recommended?**

* 1. When 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.
  2. Fisher’s exact test is a special case of the hypergeometric distribution, which gives the probability of x successes when sampling without replacement from a finite population. 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 (neutral expectation) than the observed table.

The Chi-square test uses a normal (continuous) approximation of a binomial (discrete) distribution to compute significance (p-value). The Chi-squared test statistic compares individual proportions in each group to the expected proportion based on the population mean estimate and is calculated as the sum of squared differences between observed and expected values over the expected value across all cells in the table.

Text, letter

Description automatically generated

* 1. For a 2x2 contingency table, a Fisher’s Exact Test is preferred when the number of sampled items is small (typically, when any of the cells has a value less than 5). Since it is based on the hypergeometric distribution, its calculation requires factorials and so is computationally more intensive. For larger tables, R can compute Fisher’s when total counts are not too large; otherwise, Chi-squared is preferred.

For a Chi-squared test, the sampling distribution of the test statistic is well approximated by an ideal χ2 distribution when the measurements are independent and the number of items in each sample is sufficiently large. When these conditions are not met, the test is not reliable and it is recommended to use Fisher’s exact test instead. General rules of thumb are:

* Expected frequencies for all categories are at least 1 or greater.
* Expected frequencies should be <5 for no more than 1/5 of categories.

# Statistical Modeling

1. **(a) In the formula Y ~ X, what is another name for Y and for X?**

**(b) Describe a hypothetical experimental scenario when an interaction term might be useful. How would you write a formula that includes the interaction term?**

**(c) How would you determine whether an interaction term should be included in your model or not?**

* 1. Y is the response or dependent variable; X is the predictor or independent variable.
  2. 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.

y ~ x1 + x2 + x1\*x2 OR y ~ x1\*x2

* 1. Compare simpler and more complex models to see if adding an interaction term is significant and improves the model.

**9) (a) What is the difference between planned and unplanned experimental designs?**

**(b) Give an example of each type of experiment.**

**(c) What test is used to determine the effect size in each type of experiment?**

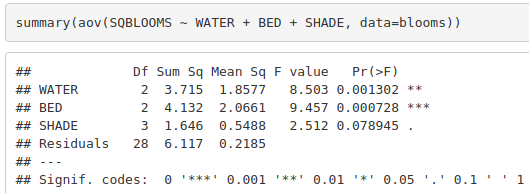
a. A planned experiment has a single reference, or control, group. In an unplanned experiment, we are interested in comparing everything to everything.

b. …

c. Planned experiment: apply Dunnett’s Test to an aov() model to look at the differences between each group and the control.

Unplanned experiment: apply Tukey’s Honest Significant Differences (Tukey’s HSD) Test to an aov() model to look at all pairwise differences between the groups.

1. **ANOVA models**
   1. What does ANOVA test, and what types of values (continuous, discrete, or categorical) are the Response and Predictor variables?
   2. For the image of an ANOVA result below:
      1. What is the Df column describing?
      2. What is the Sum Sq column describing?
      3. How is the F distribution created? How is the F-statistic calculated? Either write a formula or describe the general idea behind it.
      4. What is the null hypothesis of the F-test?
      5. What factor(s) are significant in this model?



a. ANOVA tests whether there is a significant difference between more than two groups. Response is continuous, predictor is categorical (though, can run ANOVA on linear models with continuous predictors).

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

ii. 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. 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. 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.

v. WATER and BED are significant.

1. **(a) What assumptions should be met in order to perform linear regression?**

**(b) What is the basic conceptual difference between correlation and regression?**

**(c) What general approach is used to find the best-fit line in linear regression? In logistic regression?**

**(d) In the output of a linear model in R, what does the Estimate mean? Describe the basic idea of how the estimate is calculated.**

* + - * 1. L.I.N.E.
        2. Correlation: Amount of variation explained by association of two variables relative to their individual variation

Regression: amount of variation in Response variable Y that is explained by Predictor variable(s) Xi.

* + - * 1. Least squares; maximum likelihood
        2. …

**11) (a) What are the main differences between ANOVA, linear, and logistic regression?**

**(d) In looking at the results of a linear model, what does R2 represent? What kind of information is used to calculate it?**

**(c) What is the null hypothesis of the test that provides the p-value for the predictor in a regression model?**

**(d) In logistic regression, what function is used as the Response variable in the linear model?**

ANOVA tests whether there is a significant difference between groups.

Linear regression models the magnitude of response between independent and dependent variables and provides predictions for new data.

Logistic models qualitative outcomes using quantitative predictors. It is used when the outcome to be modeled is the probability of two alternative outcomes, the predictors are quantitative (or can be encoded as binary), and there is some correlation between the predictors and response variables.

|  |  |  |
| --- | --- | --- |
| **Method** | **Predictor Variables** | **Response Variables** |
| **ANOVA** | quantitative / categorical | quantitative |
| **Linear regression** | quantitative / categorical (ordinal / interval) | quantitative |
| **Logistic regression** | quantitative or discrete (categorical can be encoded if binary) | qualitative (binary, mutually exclusive) |

* 1. R2 tells us how much of the variance can be explained by the model. It is essentially the Sum of Squares of the model divided by the total sum of squares. The total sum of squares is the difference between the points and the mean of all points squared. The Sum of Squares of the linear model is the difference between the points and the linear model squared.
  2. …
  3. Log odds (logit)

1. **(a) What is the fundamental difference between Bayesian and classical statistics?**
2. **What are known and known or unknown conditional probabilities called in a Bayesian model, and how are they related?**
3. **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.**

**13) (a) Describe the difference between the “frequentist” and Bayesian worldviews.**

**(b) What are the three main components of a Bayesian model called, and what do they mean?**

**(c) Using an example from class, outline the basic framework for a Bayesian analysis and describe why this is a useful paradigm.**

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

d. 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.

1. **Dimensionality Reduction**
   1. Why is it useful to use dimensional reduction methods like PCA, t-SNE, and UMAP?
   2. What's the basic idea behind PCA, and how are principal components identified?
   3. 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? How many t-SNE/UMAP dimensions would you calculate for the same data?
   4. What is the biological meaning of principal components and t-SNE/UMAP dimensions?

a. 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.

b. 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.

c. 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. For t-SNE and UMAP, you would normally calculate 2 or 3 dimensions.

d. Strictly speaking, the axes of PCA, t-SNE or UMAP have no biological meaning, as they are just linear combinations of all variables multiplied by certain coefficients. However, they may correlate with certain biological features, which is useful for the interpretation of the plots.

1. **a. Describe the steps for K-means clustering.**
2. **How does K-means differ from hierarchical clustering?**
3. **Describe one method that can be used to judge the quality of clusters and how this can help you find the optimal number of clusters in your data.**
   * + - 1. K-means is a top-down approach where we first decided the number of groups we want and then we start by randomly assigning the centroids of these k groups. We then assign genes to the cluster whose centroid is closest to them and recalculate the centroid (middle point) for each group of genes. This process is repeated until the centroids are no longer moving.
         2. Hierarchical clustering is a bottom-up approach where we make a pairwise distance matrix between all genes or samples and group the pair(s) that are closest to each other. Once a group is made, we recalculate all the distances from the group to the other genes and groups of genes, using some linkage method (single, complete, average, centroid). This process is repeated until all items are linked.
         3. E.g. Silhouette

**14) What is the difference between Euclidean vs. Correlation Distance? When might you prefer one over the other?**

1. What are three major differences between the mechanics of Hierarchical and K-means clustering?
2. What are advantages and disadvantages of these two methods?
   1. 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.

Pearson correlation normalizes the covariance to produce correlation values ranging from -1 to 1.

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. Hierarchical clustering is a bottom up approach where we first start looking at genes or samples that are closest to each other and start grouping them. Once the group is made, we recalculate all the distances from the group to the other genes and groups of genes. K-means is a top-down approach where we first decided the number of groups we want and then we start by randomly assigning the centroids of these k groups. Then we assign the genes to the cluster whose centroid is closest to them. Then we move the centroid to be in the center of the genes. Repeat till the centroids are no longer moving.
  2. Advantages of Hierarchical clustering is that it if you repeat the analysis you will get the same results. However due to the randomness of k-means clustering algorithm, it is not necessary for the results to be the same. Another disadvantage of k-means clustering is that it uses Euclidean distance whereas Hierarchical clustering can handle others. For Hierarchical clustering you don’t need to know the number of groups you want until after the analysis is done, however for K-means you need to know before.