# Distributions

1. The shape of the normal distribution is defined by two parameters, mean and standard deviation. They control the center and spread of the normal curve. What parameters control the shape of the following distributions? How do the distributions change as these parameters change?
   1. Poisson distribution
   2. Binominal distribution
   3. Negative binomial distribution
   4. Geometric distribution
   5. Hypergeometric distribution
2. 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. 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. 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.

1. What is probability density function (PDF) and cumulative distribution function (CDF)? When PDF and CDF are plotted, what do the Y and X axes represent in each case? Plotting a PDF is a very common way to visualize distributions – but in which cases may it be more useful to examine the CDF?
2. What is the central limit theorem? How is it useful for the analyses of biological data?

# Hypothesis Testing

1. Hypotheses, Error and Power
   1. What is a "null hypothesis"? What “alternative” hypotheses can be tested?
   2. Define Type I and Type II errors and clearly explain the difference between them.
   3. What is power, and what’s the tradeoff between error and power?

a. 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. 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. 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. Parametric vs Nonparametric
   1. What is the main difference between parametric and nonparametric tests?
   2. What are the advantages of a nonparametric test? What are the advantages of a parametric test?
   3. Which nonparametric test compares values between two independent populations to find if one is greater than the other? What is the test statistic for this nonparametric test? How is it calculated (either write a formula or describe the general idea)?

c. To perform the test, we first compute the W-statistic for both groups:

* Combine the data.
* Assign ranks from smallest (top-ranked) to largest (lowest-ranked).
* Assign ties the midrank (the average of the ranks).
* Compute the sum of ranks T1 for Sample 1 and T2 for Sample 2.
* Calculate W using the formulas below.

W1=T1−n1(n1+1)/2

W2=T2−n2(n2+1)/2

where n1 and n2 are the sizes of the two groups.

1. P-values
   1. What is a p-value?
   2. What are the shortcomings of p-values?
   3. Is it possible for something to be significant but not important? Explain.

a. 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. 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. 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
   1. What is a confidence interval? What, specifically, does a 95% CI mean?
   2. Why and how are confidence intervals useful? In particular, how do confidence intervals complement 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.

1. T-tests
   1. What is the purpose of the t-test?
   2. What are some assumptions about that data that need to be true in order for someone to use the t-test?
   3. What is the formal definition of the t-statistic? Either write a formula or describe the general idea behind it.
   4. What is the difference between a one-sided and two-sided t-test? What are the null and alternative hypotheses for each?
   5. What does a significant p-value of such a t-test mean?
   6. How are confidence intervals for t-tests determined for two-sample comparisons?

a. 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. The main assumptions are that the:

* Sample data are normally distributed.
* Sample data are random and independent.
* Variance in the samples and population is the same.

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.

d. 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.

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

f. 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
   1. Why is multiple hypothesis testing important for high-dimensional data?
   2. Name two popular methods of p-value adjustment.
   3. How does controlling for False Discovery Rate (FDR) work? Outline the general framework for controlling the FDR to 5%.

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. FDR, Bonferroni.

c. 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. Gene Ontology Enrichment
   1. What is Gene Ontology (GO) and why is testing enrichment of Gene Ontologies in a subset of genes often useful?
   2. Explain which statistical test is most frequently used for GO enrichment testing and why.

# Statistical Modeling

1. Model Formulae
   1. In the formula Y ~ X, what is another name for Y and for X?

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

* 1. How do you write a formula if you are interested in an interaction term?

y ~ x1 + x2 + x1\*x2

y ~ x1\*x2

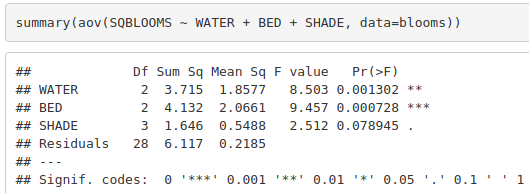
* 1. How would you determine whether an interaction term should be included in your model or not?

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.

* 1. Describe a hypothetical experimental scenario when an interaction term might be significant.

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.

1. ANOVA models
   1. What does ANOVA test?
   2. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables?
   3. Why is it useful to consider interaction terms instead of just marginal effects?
   4. 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?



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

b. Response is continuous, predictor is continuous or categorical

c. 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.

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.

1. Regression models
   1. When performing a linear regression, what type of values (continuous, discrete, categorical) are the Response variable and the Predictor variable?
   2. Why would someone want to create a regression model?
   3. What is the relationship between correlation and regression?
   4. When looking at the results, what does the R2 value represent? Provide a definition for R2 and describe the concept behind this measure.

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 essentially the difference between the points and the mean of all points squared. The Sum of Sqaures of the linear model is the difference between the points and the linear model squared.

* 1. What does the Estimate (in R output) mean? Describe the basic idea of how the estimate is calculated.
  2. What is the null hypothesis of the test that provides the p-value for the predictor?

1. Planned vs Unplanned Experimental design
   1. What is the difference between planned and unplanned experimental designs? Give an example of each.
   2. How does one perform an ANOVA analysis in R of a planned experiment? How do you determine the effect size?
   3. How does one perform an ANOVA analysis in R of an unplanned experiment? How do you determine the effect size?

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

b. If we have a planned experiment, in which we have a control, we can apply Dunnett’s Test to an aov() model to look at the differences between each group and the control.

c. If we have an unplanned experiment, in which we do not have a control, we can apply Tukey’s Honest Significant Differences (Tukey’s HSD) Test to an aov() model to look at all pairwise differences between the groups.

1. Logistic Regression
   1. For what kind of question is logistic regression used? Give an example.
   2. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables?
   3. How would you decide between ANOVA, linear regression, and logistic regression?
2. Bayesian Models
   1. What is the fundamental conceptual difference between Bayesian statistics and "frequentist" statistics?
   2. Outline the basic framework for Bayesian analysis.
   3. What is a prior?
   4. 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.

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.

# Descriptive Statistics

1. Distance
   1. Both distance and covariance can be used to describe relationship between biological samples. What is the conceptual difference between them? Give an example for when one or the other metric may be preferred.
   2. Explain the relationship between covariance and Pearson correlation. What are the similarities? What are the differences? How does Pearson correlation coefficient differ from R2?
   3. What are the assumptions of Pearson correlation? What can you do if they are violated?
2. 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. Clustering
   1. What are the steps for Hierarchical clustering?
   2. What are the steps for K-means clustering?
   3. What are the advantages and disadvantages of each method?
      * + 1. 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.
          2. 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.
          3. An advantage of Hierarchical clustering is that it if you repeat the analysis you will get the same results. However since the initial points are chosen at random in k-means clustering, the results will not be the same each time the analysis is run. For K-means the number of clusters needs to be decided beforehand, whereas the best number of clusters can be chosen for hierarchical clustering after the tree is made (on the other hand, K-means can be run multiple times with different k and evaluated using other criteria using silhouettes). Another disadvantage of k-means clustering is that it uses Euclidean distance whereas Hierarchical clustering can handle other distance metrics (such as correlation distance).

# Tabular Statistics

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?

Text, letter

Description automatically generated

1. When is it NOT OK to use the Chi-Square test?

In practice, the sampling distribution of the test statistic is well approximated by an ideal χ2 distribution only under certain conditions. 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.

When these conditions are not met, the test is not reliable and it is recommended to use Fisher’s exact test instead (we will discuss this soon).

1. What distribution is the Fisher’s Exact Test based on? Is there a model design for which an alternative test might be preferred?

Fisher’s exact test uses the hypergeometric distribution (sampling without replacement from a finite population). Fisher’s exact p-value can be used for datasets larger than 2x2, but it becomes computationally impractical as the number of possibilities increases.

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

By computing the cumulative probability of all possible scenarios that differ as much or more from a neutral expectation than the observed values.

# Resampling methods

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.

1. Explain what the bootstrap is and why it is often useful in practice.

Let’s say we have sampled two different groups, groupA and groupB, and we have 10 measurements for each. 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.

This simulates sampling distributions for the groups which can be useful when we have limited data.