**Grading:**

1. **Kris**
2. **Kris**
3. **Kris**
4. **Kris**
5. **Kris**
6. **Kris**
7. **Kris**
8. **Kris**
9. **Manny**
10. **Manny**
11. **Manny**
12. **Manny**
13. **Manny**
14. **Manny**
15. **Manny**
16. **Manny**

# **Distributions**

1. The binomial and Gaussian are 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. ***KRIS (5)***

The binomial distribution is created by counting the number of successes in an experiment that produces two outcomes (success and failure) in multiple trials. The Gaussian distribution, also called the normal distribution, is a continuous distribution that approximates the binomial when the number of values is very large.

The binomial distribution can be used to calculate the probability of a mother having 2 sons.

The normal distribution can be used to calculate the probability that an individual is from a certain population using their height.

1. What is a log-normal distribution and when is it sometimes useful? Give an example. ***KRIS (5)***

A log-normal distribution is when you take the log of the values and it gives you 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 pression.

1. How is the Poisson related to the exponential distribution? To what kind of processes / situations would the apply? Give an example of how the Poisson is relevant for single-cell experiments that use microfluidics. ***KRIS (5)***

­ The Poisson is a discrete distribution that describes (rare or infrequent) stochastic events that occur with a fixed average rate in either time or space, where the number of events in successive intervals is statistically independent. The single parameter λ is the rate constant; the mean and variance are both equal to λ.

The exponential is a continuous distribution that describes the time (or distance) interval between events generated by a Poisson process. It is also determined by a single parameter λ, where the mean and variance are both 1/λ. The CDF is F(*t*) = 1 - e-*t*λ.

The *survivorship* function of the Poisson PDF is 1 – the CDF of the exponential. That is, the probability of *no* events in time *t*, or the probability the time T to the *first* occurrence *exceeds* *t*, is: P(X = 0) = P(T > *t*) = e-*t*λ = 1 - (1 - e-*t*λ) = 1 - CDF(Exp), which is the total probability that an event *does* occur in time t.

A non-science example is trains arriving at a station: the number of events per unit time follows a Poisson distribution, and the time between arrivals follows an exponential distribution. Radioactive decay is also a Poisson process: particles decay stochastically at a low rate, and the decay curve is an exponential. Similarly, the number of mutations per kb after mutagenesis with e.g. EMS follows a Poisson; the distance between mutations follows an exponential.

For single-cell genomics, the number of cells per droplet follows a Poisson distribution.

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”.) ***KRIS (5)***

## 

Count data for technical replicates can be modeled by a Poisson distribution, where the mean and the variance are given by the same parameter, λ. Thus, the variance is proportional to the mean.

When considering biological replicates, it turns out that the variation in the counts increases with the number of counts per feature (expression level); it is *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. In this case, 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. In terms of the power analysis, what kinds of errors can arise in comparing two samples? What’s the tradeoff between error and power? Be sure to define error types and power in your answer. ***KRIS (5)***

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 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. P-values and confidence intervals ***KRIS*** 
   1. What is a p-value, and what are some of its shortcomings? **(3)**

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.

The p-value resulting from repeated experiments may vary widely due to small sample sizes, which 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 are also arbitrary cutoffs.

* 1. How can a test result be significant but not important? Give an example. **(3)**

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’s 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. What is a confidence interval? How do CIs complement p-values? Describe what a 95% CI means specifically. **(3)**

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

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).

1. Parametric and non-parametric tests ***KRIS***
   1. What is a t-test for? What are some assumptions about that data that need to be true in order for someone to use the t-test? **(4)**

There are two main applications of T-test. One is to see if the sample collected is from the 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.

The main assumptions are that the:

* 1. Sample data are normally distributed.
  2. Sample data are random and independent.
  3. Variance in the samples and population is the same.
  4. What does a significant p-value of such a t-test mean? **(2)**

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

* 1. List and describe two different alternative hypotheses for two-sample comparisons. **(4)**

The null hypothesis is always that the two samples come from the same parent distribution.

For a one-tailed test, one alternative hypothesis is that is that the mean of the second sample is greater than that of the first sample. A second HA is that the mean of the second sample is smaller.

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

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 also calculating the sd of the samples together

* 1. What is the difference between parametric and non-parametric tests? When would you choose a rank-based test over a parametric one? Give an example. **(4)**

Parametric tests use standard models of statistical distributions to estimate significance and confidence intervals. They assume the data are distributed approximately normally (either in their raw state or after transformation using e.g. log or exponential distributions). Parametric tests depend on estimates of sample means and variation based on normal or t-distributions.

Non-parametric tests make no such assumptions (do not require data to be normally distributed).

Rank-based tests compare the …

Alternatively, resampling methods can be used to empirically determine significance and confidence intervals. Different subsets of data are sampled many times and significance is determined by the proportion of data that fall at the extremes of the sampled data (e.g. 5% of the data).

1. Multiple Hypothesis Testing ***KRIS***
   1. Why is multiple hypothesis testing important for high-dimensional data? **(2)**

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.

* 1. What's the essential difference between the Bonferroni correction and FDR? **(2)**

Bonferroni is the most conservative family-wise error rate (FWER) correction for multiple hypothesis testing, which seeks to limit the chance of getting at least one false positive in a large dataset. The procedure simply divides each p-value from a t-test by the total number of tests performed. If m measurements were taken, then the adjusted p-value is padj = p/m. Bonferroni is generally considered to be overly conservative for many genomics studies.

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, sort the p-values form smallest to largest. Compare the p-value to the BH q-value (rank/number of samples). All the p-values that are less than the q-value are significant. *(from biostathandbook.com )*

# **Statistical Modeling**

1. ANOVA and regression
   1. What types of values (continuous, discrete, or categorical) are the Response and Predictor variables for ANOVA, linear, and logistic regression? Consider making a table to answer this question. ***Manny* (3)**

|  |  |  |
| --- | --- | --- |
| **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. How would you decide between ANOVA, linear regression, and logistic regression? Give an example scenario in which you would apply each of these tests. ***Manny* (3)**

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.

* 1. Why and when is it useful to consider interaction terms instead of just marginal effects? Give an example. ***MANNY* (3)**

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. In regression models, what does the R2 value represent? Describe the concept behind this measure. ***MANNY* (3)**

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. Bayesian Models ***Manny*** 
   1. What is the fundamental conceptual difference between Bayesian statistics and "frequentist" statistics? **(3)**

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.

Bayesian statistics also allow for the inclusion of priors, i.e. information about past experiences (or best guesses) that can be used to update the estimates, or posterior probabilities.

* 1. 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. **(2)**

The example we gave in class was that of estimating the chances that offspring will manifest a genetic disease (from Aho).

# **Descriptive Statistics**

1. Distance and Clustering ***MANNY***
   1. What’s the essential difference between Euclidian and Correlation-based distance? When is it more appropriate to use one or the other to cluster genes? Why? **(3)**

*Ported from 2017:*

*Euclidean distance is appropriate when you want to compare the magnitude of values rather than their “profile” or covariance. This is because it is simply looking at how far the values are from each other rather than where they are in relation to other points. It is possible to scale and center your values which in turn will give you same results as pearson correlation.*

* 1. What are three major differences between the mechanics of Hierarchical and K-means clustering? **(3)**

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.

* 1. What are advantages and disadvantages of each clustering method? **(3)**

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.

1. Principal Components Analysis ***Manny***
2. Why is it useful to use dimensional reduction methods like PCA? **(3)**

Sometimes there are very 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. What's the basic idea behind PCA, and how are principal components identified? **(3)**

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.

(cont’d next page)

1. Give an example of an application for PCA and what you would gain from it. **(2)**

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.

*( we can add more examples … )*

1. How can you decide on the right number of principal components to use for your analysis? **(2)**

A common method is to use an “elbow plot” that illustrates the amount of variation explained by each component. The point at which the **rate** of change decreases noticeably is the “elbow” in the plot and one choice is to use that number of PC’s.

Another method is to keep the number of PCs that explain a certain proportion of the total variation observed, e.g. 90%.

# Tabular Statistics

1. Describe a simple scenario in which you would use a contingency table. ***Manny* (3)**

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. When is it preferable to use Fisher’s Exact Test vs. a Chi-Square test? Discuss the underlying probability distributions these two approaches are based on, and any important assumptions that must be considered. ***Manny* (3)**

The Chi-square test uses a normal (continuous) approximation of a binomial (discrete) distribution to compute significance. This approximation is valid when the measurements are independent and number of items in each sample is sufficiently large. Chi-squared can also be used for contingency tables that are larger than 2x2.

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.

# Resampling methods

1. Why may someone want use resampling instead of a parametric test (e.g. t-test, ANOVA, Chi-squared)? ***Manny* (3)**

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 occurrence of a GO-term is enriched in a list of genes using the resampling method? Describe the steps in detail. ***MANNY* (5)**

Create a matrix containing two different columns: GeneName and Goterm.

The GeneName columns should have all the different gene names and the Goterm columns should have information whether the corresponding gene has that given go-term.

Let’s call our list of genes L.

First find out the number of genes in L that have the goterm. Let’s call this LG.

Now select random genes ( R ). the same size of L. Count the number of R genes that have the go-term G.

Do this 10,000 times.

The p-values of L with G is the number of times R has Go-term greater than or equal to LG divided by 10,000.

***EXTRA SPACE IF NEEDED***

***EXTRA SPACE IF NEEDED***

***EXTRA SPACE IF NEEDED***