LECTURE 3
Hypothesis testing

The framework of hypothesis testing is used in computational biology extensively.

We will look at hypothesis testing in both the classical and Bayesian framework as well as multiple hypothesis testing.

The Bayesian and classical frameworks ask two different questions:

- Bayesian: "What is the probability of the hypothesis being true given data ?" - $\Pr(H = t|D)$, posterior.
- Classical: "Assume the hypothesis is true, what is the probability of the data?" - $\Pr(D|H = t)$, likelihood.

The first question is more natural but requires a prior on hypotheses.

3.0.11. Classical hypothesis testing

The classical hypothesis testing formulation is called the Neyman-Pearson paradigm. It is a formal means of distinguishing between probability distributions based upon random variables generated from one of the distributions.

The two distributions are designated as:

- the null hypothesis: H_0
- the alternative hypothesis: H_A.

There are two types of hypothesis, simple and composite:

- simple hypothesis: all aspects of the distribution are specified. For example, $H_0 : X \sim N(\mu_1, \sigma^2)$ is simple since the distribution is fully specified. Similarly $H_A : X \sim N(\mu_2, \sigma^2)$ is also simple.
- composite hypotheses: the hypothesis does not specify the distribution. For example, $H_A : X \sim \text{Bernoulli}(n, p > /25)$, is composite since $p > .25$ in the Bernoulli distribution so the distribution is not specified.

In general two types of the alternative hypothesis are one and two sided:

- one sided: $H_A : X \sim \text{Binomial}(n, p > /25)$
- two sided: $H_A : X \sim \text{Binomial}(n, p \neq /25)$.

In this paradigm we first need a test statistic $t(\mathbf{X})$ which can be computed from the data $\mathbf{X} = (x_1, ..., x_n)$. We have a decision problem in that given $\mathbf{X}$ we compute $t(\mathbf{X})$ and decide whether we reject H_0, a positive event, or accept H_0 the negative

event. The sets of values of t for which we accept H_0 is the acceptance region and the sets for which we reject H_0 are the rejection region. In this paradigm the following four events written in a contingency table can happen. Two of the events are errors, B and C.

	$H_0 = T$	$H_A = T$
Accept H_0	A	B
Reject H_0	C	D

C is called a type I error and measure the probability of a false positive,

$$\alpha = \Pr(\text{null is rejected when true}).$$

The reason why it is called a false positive is that rejecting the null is a positive since one in general in an experiment is looking to reject the null since this corresponds to finding something different from the lack of an effect. In general in the hypothesis testing framework we will control the α value explicitly (this will be our knob) and is called the significance level.

B is called the type II error and measure the probability of false negatives,

$$\beta = \Pr(\text{null is accepted when it is false}).$$

The power of a test is

$$1 - \beta = \Pr(\text{null is rejected when it is false}).$$

Ideally we would like a test with $\alpha = \beta = 0$. Like most of life this ideal is impossible. For a fixed sample size increasing α will in general decrease β and vice versa. So the general prescription in designing hypothesis tests is to fix α of a small number and design a test that will give as small a β as possible, the most powerful test.

Example. *We return to the DNA sequence matching problem where we get a string of* $2n$ *letters corresponding to two strands of DNA ask about the significance of the number of observed matches. Our null hypothesis is that the nucleotides* A, C, T, G *are equally likely and independent. Another way of saying this is*

$$H_0 : X \sim \textit{Binomial}(p = .25, n)$$

the alternative hypothesis is

$$H_A : X \sim \textit{Binomial}(p > .25, n).$$

Assume we observe $Y = 32, 33$ *matches out of* 100 *according to the distribution under the null hypothesis.*

$$\Pr(Y \geq 32 | p = .25, n = 100) = .069,$$
$$\Pr(Y \geq 33 | p = .25, n = 100) = .044.$$

Therefore, to achieve an α *level of* .05 *we would need a significance point (critical value) of* 33*.*

The p-value is the smallest α for which the null will be rejected. It is also called the achieved significance level. Another way of stating this is the p-value is the probability of obtaining an observed value under the distribution given by the

null hypothesis that is greater (more extreme) than the statistic computed on the data $t(\mathbf{X})$.

Example. *We return to the matching problem.*

$$H_0 : X \sim Binomial(p = .25, n)$$

the alternative hypothesis is

$$H_A : X \sim Binomial(p > .25, n).$$

We find 11 *matches out of* 26

$$\Pr(Y \geq 11 | p = .25, n = 26) = .04,$$

so the p-value is .04

We find 278 *matches out of* 100

$$\Pr(Y \geq 278 | p = .25, n = 100) \approx .022,$$

and by the Normal approximation of the Binomial

$$Y \sim N(250, 187.5).$$

We now look at an example that introduces a classic null distribution, the t-statistic and the t-distribution.

Example. *We have two cell types cancer A and B. we measure the expression of one protein from the two cells.*

$$\begin{aligned} A &: X_1, ..., X_m \sim N(\mu_1, \sigma^2) \\ B &: Y_1, ..., Y_n \sim N(\mu_2, \sigma^2) \end{aligned}$$

so we draw m observations from cell type A and n from cell type B.

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu \\ H_A &: \mu_1 \neq \mu_2. \end{aligned}$$

The statistic we use is the t-statistic

$$t(\mathbf{X}) = \frac{(\bar{X} - \bar{Y})\sqrt{mn}}{\hat{S}\sqrt{m+n}},$$

where

$$\hat{S}^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n-2}.$$

The distribution under the null hypothesis of the above is the t-distribution with $d = m + n - 2$ *degrees of freedom*

$$p(t) = \frac{\Gamma[(d+1)/2]}{\sqrt{d\pi}\Gamma[d/2](1 + \frac{t^2}{d})^{(d+1)/2}}.$$

The above example is the most classical example of a hypothesis test and statistic. It is also a parametric test in that we have a parametric assumption for the null hypothesis. Here the assumption is that the samples are normally distributed.

There are a class of hypothesis tests that are called nonparametric in that the parametric assumptions on the null hypothesis are weak. Typically these tests are

called rank statistics in that the ranks of the observations are used to compute the statistic. We will look at two such statistics: the Mann-Whitney (MW) and Kolmogorov-Smirnov statistics. We first define the MW statistic and state its property. We then look more carefully at the KS statistic and use it to illustrate why rank based statistics are nonparametric.

Example (Mann-Whitney statistic)**.** *We have two cell types cancer A and B. we measure the expression of one protein from the two cells.*

$$\begin{aligned} A &: X_1, ..., X_m \sim F_A \\ B &: Y_1, ..., Y_n \sim F_B, \end{aligned}$$

where F_A and F_B are continuous distributions. This is why the test is nonparametric.

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu \\ H_A &: \mu_1 > \mu_2. \end{aligned}$$

We first combine the lists

$$Z = \{X_1, ..., X_m, Y_1, ..., Y_n\},$$

we then rank order Z

$$Z_{(r)} = \{Z_{(1)}, ..., Z_{(m+n)}\}.$$

Given the rank ordered list $Z_{(r)}$ we can compute two statistics

$$\begin{aligned} R_1 &= \textit{sum of ranks of samples in } A \textit{ in } Z_{(r)} \\ R_2 &= \textit{sum of ranks of samples in } B \textit{ in } Z_{(r)} \end{aligned}$$

Given R_1 and R_2 we compute the following statistics

$$\begin{aligned} U_1 &= mn + \frac{(m+1)m}{2} - R_1 \\ U_1 &= mn + \frac{(n+1)n}{2} - R_2, \end{aligned}$$

$U = \min(U_1, U_2)$. *The statistic*

$$\hat{z} = \frac{\left|U - \frac{mn}{2}\right|}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0,1).$$

This test is called nonparametric in the the distributional assumptions (in terms of parameters) on the null hypothesis are extremely week.

Example (Kolmogorov-Smirnov statistic)**.** *We have two cell types cancer A and B. we measure the expression of one protein from the two cells.*

$$\begin{aligned} A &: X_1, ..., X_m \sim F_A \\ B &: Y_1, ..., Y_n \sim F_B, \end{aligned}$$

where F_A and F_B are continuous distributions. This is why the test is nonparametric.

$$\begin{aligned} H_0 &: F_A = F_B \\ H_A &: F_A \neq F_B. \end{aligned}$$

We first construct empirical distribution functions for the two sets of data $X = \{X_1, ..., X_m\}$, $Y = \{Y_1, ..., Y_n\}$

$$\begin{aligned} F_m(x) &= \frac{\#[\{X\} \leq x]}{m} \\ F_n(x) &= \frac{\#[\{Y\} \leq x]}{n}, \end{aligned}$$

where $\#[\{X\} \leq x]$ *indicates the number of elements in* X *that are smaller than* x*, similarly for* $\#[\{Y\} \leq x]$*. Note that the above quantities are basically rank quantities.*

The first result is one by Smirnov.

Theorem. *Given the statistic*

$$D_{mn} = \sup_x |F_n(x) - F_m(x)|,$$

with $X_1, ...X_m, Y_1, ..., Y_n \sim F(x)$*. The distribution*

$$\Phi_{mn}(\lambda) = \Pr\left(D_{mn} \leq \lambda\sqrt{\frac{mn}{m+n}}\right),$$

is independent of $F(x)$*.*

The above theorem states that the distribution under the null hypothesis for the Kolmogorov-Smirnov statistic is independent of $F(x)$*.*

We now sketch why this is true.

We first use the idea of symmetrization which we first encountered in the symmetrization proposition on page 16. For simplicity we assume that $n = m$ *in this context the symmetrization lemma on page 16 stated that: Given two draws from a distribution* $x_1, ..., x_n$ *and* $x'_1, ..., x'_n$

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(x'_i)\right| \geq \epsilon\right) \leq 2\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(x)\right| \geq \epsilon/2\right).$$

A result similar to the above was used to Kolmogorov and Smirnov to show that

$$\Pr\left(D_{mn} \leq \lambda\sqrt{\frac{mn}{m+n}}\right) \approx \Pr\left(D_n \leq \lambda\sqrt{2n}\right),$$

where

$$D_n = \sup_x |F_n(x) - F(x)|,$$

with $X_1, ...X_n \sim F(x)$*.*

In a paper that appeared in the Italian Journal of the Actuarial Institute in 1933 Kolmogorov proved the convergence of the empirical distribution function to the distribution function. This result was used by Smirnov in 1939 to derive the KS test result (Smirnov was a student of Kolmogorov).

Kolmogorov showed that:

Theorem. *Given the statistic*

$$D_n = \sup_x |F_n(x) - F(x)|,$$

with $X_1, ...X_n, Y_1$*. The distribution*

$$\Phi_n(\lambda) = \Pr\left(D_n \leq \lambda\sqrt{n}\right),$$

is independent of $F(x)$. In addition the limiting distribution is

$$\lim_{n\to\infty} \Phi_n(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}.$$

A key idea in the proof of the result was the following lemma which is at the heart of the reason why rank statistics are nonparametric and independent of the distribution $F(x)$. The essence of this lemma is that looking at the difference between $F_n(x) - F(x)$ is equivalent to looking at the difference between $U_n(x) - U(x)$ where $U(x)$ is the distribution function for the uniform distribution in the interval $[0, 1]$ and $U_n(x)$ is the empirical distribution function for n observations drawn iid from $U[0, 1]$.

Lemma. *The distribution $\Phi_n(\lambda)$ is independent of $F(x)$ if $F(x)$ is continuous.*

Proof. *Let X be a random variable with continuous distribution function $F(X)$, the random variable $Y = F(X)$ has the following distribution $F^{(0)}(x)$*

$$\begin{aligned} F^{(0)}(x) &= 0, \quad x \le 0; \\ F^{(0)}(x) &= x, \quad 0 \le x \le 1; \\ F^{(0)}(x) &= 0, \quad 1 \le x. \end{aligned}$$

The above can be restated as Y is distributed as the uniform distribution on the interval $[0, 1]$. Given that $F_n(x)$ and $F_n^{(0)}(x)$ represent the empirical distribution functions for X and Y after n observations the following hold:

$$\begin{aligned} F_n(x) - F(x) &= F_n^{(0)}[F(x)] - F^{(0)}[F(x)], \\ &= F_n^{(0)}(y) - F^{(0)}(y) \\ \sup_x |F_n(x) - F(x)| &= \sup_x |F_n^{(0)}(y) - F^{(0)}(y)|. \ \square \end{aligned}$$

The implication of the above lemma is that to study the distribution under the null hypothesis for the difference of distribution functions it suffices to study the uniform distribution.

The above lemma can be used to analyze the Mann-Whitney statistic since the difference in the average ranks of

$$\begin{aligned} A &: X_1, ..., X_m \sim F_A \\ B &: Y_1, ..., Y_n \sim F_A, \end{aligned}$$

can be written as

$$\bar{R}_A - \bar{R}_A = \frac{1}{m}\sum_{i=1}^{m} F_m(x_i) - \frac{1}{n}\sum_{i=1}^{n} F_n(x_i),$$

and the above lemma holds for this case as well. Note, the Mann-Whitney statistic can be rewritten in terms of the difference in average ranks of the two samples A and B.

Both the Mann-Whitney and KS statistics are nonparametric and so in the context of adaptability these are good tests. The general question of what is a good hypothesis test or is test A better than test B has still not been addressed. This question is typically addressed via the likelihood ratio testing framework and the Neyman-Pearson Lemma.

We start with the case of two simple hypotheses. Again these hypotheses are simple since the densities are completely specified under the null and alternative hypotheses.

$$\begin{aligned} H_0 &: X \sim p(X|H_0 = T) \\ H_A &: X \sim p(X|H_A = T). \end{aligned}$$

Given the sample $X = \{X_1, ..., X_n\}$ we can write down the likelihood ratio

$$\Lambda = \frac{p(X|H_0)}{p(X|H_A)}.$$

It would seem reasonable to reject H_0 for small values of Λ.

Lemma (Neyman-Pearson)**.** *Suppose the likelihood ration test rejects H_0 whenever $\frac{p(X|H_0)}{p(X|H_A)} < c$ has significance level α*

$$\Pr\left(\frac{p(X|H_0)}{p(X|H_A)} < c\right) = \alpha.$$

Then any other test which has significance level $\alpha^* \leq \alpha$ has power less than or equal to the likelihood ratio test.

This lemma address the issue of optimality for simple hypotheses.

Example. *We draw n observations iid from a normal distribution,*

$$X_1, ..., X_n \sim N(\mu, \sigma^2).$$

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu = \mu_A. \end{aligned}$$

$$\Lambda = \frac{p(X|\mu_0, \sigma^2)}{p(X|\mu_A, \sigma^2)} = \frac{\exp(-\sum_{i=1}^n (x_i - \mu_0)^2/2\sigma^2)}{\exp(-\sum_{i=1}^n (x_i - \mu_A)^2/2\sigma^2)}.$$

$$\begin{aligned} \log(\Lambda) &= \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_A)^2, \\ &= \sum_i x_i^2 - 2\mu_0 \sum_i x_i + n\mu_0^2 - \sum_i x_i^2 + 2\mu_A \sum_i x_i - n\mu_A^2 \\ &= 2n\bar{X}(\mu_0 - \mu_A) + n\mu_A^2 - n\mu_0^2. \end{aligned}$$

So we can use a statistic $t(X)$

$$t = 2n\bar{X}(\mu_0 - \mu_A) + n\mu_A^2 - n\mu_0^2,$$

as our statistic to reject the null. Under the null hypothesis

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right),$$

so for the case where $\mu_0 - \mu_A < 0$ the likelihood ratio test is a function of $\bar{X}$ and is small when $\bar{X}$ is small. In addition

$$\Pr(\bar{X} \geq x_0) = \Pr\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{x_0 - \mu_0}{\sigma/\sqrt{n}}\right),$$

so we can solve

$$\frac{x_0 - \mu_0}{\sigma/\sqrt{n}} = z(\alpha).$$

Which is the most powerful test under this model.

The problem with the likelihood ratio test framework is that in general the alternative hypothesis is not simple but composite. A typical situation would be

$$\begin{aligned} H_0 &: X \sim N(\mu_0, \sigma^2) \\ H_A &: X \sim N(\mu \neq \mu_0, \sigma^2), \end{aligned}$$

where the distribution under the alternative is not specified but forms a family of distributions. In this setting likelihood ration tests can be generalized to the concept of generalized likelihood ration tests which also have optimality conditions but these conditions are more subtle and we will not study these conditions.

Given the sample $X = \{X_1, ..., X_n\}$ drawn from a density $p(X|\theta)$ with the hypotheses

$$\begin{aligned} H_0 &: \theta \in \omega_0 \\ H_A &: \theta \in \omega_A, \end{aligned}$$

where $\omega_0 \cap \omega_A = \emptyset$ and $\omega_0 \cup \omega_A = \Omega$. The generalized likelihood ratio is defined as

$$\Lambda^* = \frac{\max_{\theta \in \omega_0} p(X|H_0(\theta))}{\max_{\theta \in \omega_A} p(X|H_A(\theta))}.$$

For technical reasons we will work with a slight variation of the above likelihood ration

$$\Lambda = \frac{\max_{\theta \in \omega_0} p(X|H_0(\theta))}{\max_{\theta \in \Omega} p(X|H_A(\theta))}.$$

Note that $\Lambda = \min(\Lambda^*, 1)$ so small values of Λ^* correspond to small values of Λ so in the rejection region using either one is equivalent.

Example. *We draw n observations iid from a normal distribution,*

$$X_1, ..., X_n \sim N(\mu, \sigma^2).$$

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0. \end{aligned}$$

The generalized likelihood is

$$\Lambda = \frac{\frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2)}{\max_\mu \frac{1}{(\sigma\sqrt{2\pi})^n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2)},$$

the denominator is maximized at $\mu = \bar{X}$. So we can write

$$\begin{aligned} -2\log\Lambda &= \frac{1}{\sigma^2}\left(\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{X})^2\right) \\ &= n(\bar{X} - \mu_0)^2/\sigma^2. \end{aligned}$$

We computed previously that if $X_i \sim N(\mu_0, \sigma^2)$ that $\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$. The distribution of the square of a normal random variable is the chi-square distribution so

$$t = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} \sim \chi^2_1,$$

where χ^2_1 is the chi-square distribution with one degree of freedom and we would reject the null hypothesis when

$$\frac{n(\bar{X}-\mu_0)^2}{\sigma^2} > \chi^2_1(\alpha).$$

We will apply the generalized likelihood ration to two problems in classical genetics. However, the relevant statistical model involved in both problems requires the multinomial distribution which we now introduce.

There are n independent trials where for each trial one of r possibilities can occur each with probability $p_1, ..., p_r \geq 0$ where $\sum_{i=1}^r p_r = 1$. The actual counts from the n independent trials are $n_1, ..., n_r$ where $\sum_{i=1}^r n_i = n$. The (joint) distribution on the above counts is

$$\Pr(n_1, ..., n_r) = \frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i^{n_i}.$$

We will parameterize the multinomial as $M(p, n, r)$ where $p = \{p_1, ..., p_r\}$.

Given observations $X = \{n_1, ..., n_r\}$ from the multinomial distribution with r possibilities and

$$\begin{aligned} H_0 &: X \sim M(\theta, n, r) \text{ with } \theta \in \omega_0 \\ H_A &: X \sim M(\theta, n, r) \text{ with } \theta \in \omega_A. \end{aligned}$$

We write down the likelihood ratio as

$$\Gamma = \frac{\max_{\theta\in\omega_0} \frac{n!}{n_1!\cdots n_r!} \prod_{i=1}^r p_i(\theta)^{n_i}}{\max_{\theta\in\Omega} \frac{n!}{n_1!\cdots n_r!} \prod_{i=1}^r p_i(\theta)^{n_i}}.$$

The maximization in the denominator is unconstrained so the unconstrained maximal likelihood estimate results in

$$\hat{p}_i = \frac{n_i}{n}.$$

The maximization in the numerator is constrained and we denote the estimate as

$$\hat{\theta} = \arg\max_{\theta\in\omega_0} \left[\frac{n!}{n_1! \cdots n_r!} \prod_{i=1}^r p_i(\theta)^{n_i} \right].$$

If we plug the above back into the likelihood ratio

$$\begin{aligned} \Gamma &= \frac{\frac{n!}{n_1!\cdots n_r!} \prod_{i=1}^r p_i(\hat{\theta})^{n_i}}{\frac{n!}{n_1!\cdots n_r!} \prod_{i=1}^r \hat{p}_i^{n_i}} \\ &= \prod_{i=1}^r \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{n_i}, \\ &= \prod_{i=1}^r \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{n\hat{p}_i}, \end{aligned}$$

the last line comes from $n_i = \hat{p}_i n$. Taking the log

$$\begin{aligned} -2\log\Gamma &= -2n \sum_{i=1}^r \hat{p}_i \log\left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right) \\ &= 2\sum_{i=1}^r O_i \log\left(\frac{O_i}{E_i} \right), \end{aligned}$$

where $O_i = n\hat{p}_i$ and $E_i = np_i(\hat{\theta})$ are the observed and expected counts, respectively. If the dimensionality of the model space ω_0 is k then under the null hypothesis

$$t = 2\sum_{i=1}^{r} O_i \log\left(\frac{O_i}{E_i}\right) \sim \chi^2_{r-k-1},$$

where χ^2_{r-k-1} is the chi-square distribution with $r-k-1$ degrees of freedom.

Example. *Under Hardy-Weinberg equilibrium the genotypes AA, Aa, and aa occur in a population with frequencies* $(1-\tau)^2$, $2\tau(1-\tau)$, *and* τ^2. *In a sample from the Chinese population in Hong Kong in 1937 blood types occurred with the frequencies given in the table below. Given the observed numbers and the Hardy-Weinberg equilibrium model we can estimate* τ *using maximum likelihood,* $\hat{\tau} = .4247$. *This allows us to compute the expected counts under our model which is given in the table as well.*

	M	*MN*	*N*
Observed	*342*	*500*	*187*
Expected	*340.6*	*502.8*	*185.6*

Given the above data our hypotheses are

$$\begin{aligned} H_0 &: X \sim M(\{(1-\tau^2), 2\tau(1-\tau), \tau^2\}, n = 1029, r = 3) \\ H_A &: X \sim M(\theta, n = 1029, r = 3) \text{ with } \theta \neq \{(1-\tau^2), 2\tau(1-\tau), \tau^2\}, \end{aligned}$$

so the null and alternate are both multinomial however the null assumes the Hardy-Weinberg model. For this case

$$-2\log\Gamma = 2\sum_{i=1}^{3} O_i \log\left(\frac{O_i}{E_i}\right) = .032,$$

the likelihood is .98 *and* $r-k-1=1$ *since* $r=1$ *and* $k=1$ *so the p-value is* .86. *There is no good reason to reject the Hardy-Weinberg model.*

Example (Fisher's reexamination of Mendel's data)**.** *Mendel liked to cross peas. He crossed 556 smooth, yellow male peas with wrinkled, green female peas. According to the genetic theory he developed the frequency of the baby peas should be*

Smooth yellow	$\frac{9}{16}$
Smooth green	$\frac{3}{16}$
Wrinkled yellow	$\frac{3}{16}$
Wrinkled green	$\frac{1}{16}$

The observed and expected counts are given in the following table.

Type	*Observed count*	*Expected count*
Smooth yellow	*315*	*312.75*
Smooth green	*108*	*104.25*
Wrinkled yellow	*102*	*104.25*
Wrinkled green	*31*	*34.75*

Given the above data our hypotheses are

$$\begin{aligned} H_0 &: X \sim M\left(\left\{\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right\}, n = 556, r = 4\right) \\ H_A &: X \sim M(\theta, n = 556, r = 4) \text{ with } \theta \neq \left\{\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right\} \end{aligned}$$

so the null and alternate are both multinomial however the null assumes what became Mendel's law. For this case

$$-2\log\Gamma = 2\sum_{i=1}^{4} O_i \log\left(\frac{O_i}{E_i}\right) = .618,$$

the likelihood is .73 *and* $r - k - 1 = 3$ *since* $r = 4$ *and* $k = 1$ *so the p-value is* .9*. There is no good reason to reject the Mendel's law.*

Mendel did this experiment many times and Fisher pooled the results in the following way. Two independent experiments give chi-square statistics with p and r degrees of freedom respectively. Under the null hypothesis that the models were correct the sum of the test statistic would follow a chi-square distribution with $p + r$ *degrees of freedom. So Fisher added the chi-square statistics for all the independent experiments that Mendel conducted and found a p-value of* .99996 *!*

3.0.12. Multiple hypothesis testing

In the various "omics" the issue of multiple hypothesis testing arises very often. We start with an example to illustrate the problem.

Example. *The expression level of* $12,000$ *genes can be measured with the technology available in one current platform. We measure the expression level of these* $12,000$ *genes for* 30 *breast cancer patients of which* C_1 *are those with ductal invasion and* C_2 *are those with no ductal invasion. There are* 17 *patients in* C_1 *and* 13 *in* C_2*. We consider a matrix* x_{ij} *with* $i = 1, ..., 12,000$ *indexing the genes and* $j = 1, ..., 30$ *indexing the patients. Assume that for each gene we use a t-test to determine if that gene is differentially expressed under the two conditions of for each* $i = 1, ..., 12,000$ *we assume* $X_{ij} \sim N(\mu, sigma^2)$

$$\begin{aligned} H_0 &: \mu_{C_1} = \mu_{C_2} \\ H_A &: \mu_{C_1} \neq \mu_{C_2}, \end{aligned}$$

We set the significance level of each gene to $\alpha = .01$ *and we find that we reject the null hypothesis for* 250 *genes. At this point we need to stop and ask ourselves a basic question.*

Assume H_0 *is true for* $i = 1, ..., m = 12,000$ *and we observe* $n = 30$ *observations that are* $N(\mu, \sigma^2)$ *split into groups of* 13 *and* 17*. We can ask about the distribution of the following two random variables*

$$\begin{aligned} \xi = &\ \#\ \text{rejects at } \alpha = .01 \quad | \quad H_0 = T\ \forall i = 1, ...m \\ \xi = &\ \#\ \text{times } t \geq t_{df}(.01) \quad | \quad H_0 = T\ \forall i = 1, ...m \end{aligned}$$

where t_{df} *is the t-distribution with degree of freedom df. We can also ask whether* $\mathbb{E}\xi \approx 120$ *or* $\mathbb{E}\xi \gg 120$*. This is the question addressed by multiple hypothesis testing correction.*

The following contingency table of outcomes will be used ad nauseum in understanding multiple hypothesis testing.

	Accept null	Reject null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

The main quantity that is controlled or worried about in multiple hypothesis testing is V the number of false positives or false discoveries. The other possible error is the type II error or the number of false negatives or missed discoveries. The two main ideas are the Family-wise error rate (FWER) and the False Discovery Rate (FDR). We start with the FWER.

3.0.12.1. *FWER.* The family-wise error rate consists of finding a cutoff level or α value for the individual hypothesis tests such that we can control

$$\alpha_{_{FWER}} = \Pr(V \geq 1),$$

this means we control the probability of obtaining any false positives. Our objective will be to find an α-level or cutoff for the test statistic of the individual tests such that we can control $\alpha_{_{FWER}}$.

The simplest correction to control for the FWER is called the Bonferroni correction. We derive this correction

$$\begin{aligned}
\alpha_{_{FWER}} &= \Pr(V \geq 1), \\
&= \Pr\left(\{T_1 > c\} \cup \{T_2 > c\}, ..., \cup \{T_m > c\} | H_0\right) \\
&\leq \sum_{i=1}^{m} \Pr\left(\{T_i > c\}\right) \\
&\leq m \Pr\left(\{T > c\}\right) \\
&\leq m\alpha,
\end{aligned}$$

so if we want to control $\alpha_{_{FWER}}$ at for example at .05 we can find a cutoff or select for $\alpha = \frac{.05}{m}$ for the individual hypotheses. There is a huge problem with this correction the inequality between steps 2 and 3 can be large if the hypothesis are not disjoint (or the union bound sometimes sucks).

Let us make this concrete by using the example from the beginning of this section.

Example. *In the previous example we use a t-test to determine if a gene is differentially expressed under the two conditions, for each* $i = 1, ..., 12,000$ *we assume* $X_{ij} \sim N(\mu, sigma^2)$

$$\begin{aligned}
H_0 &: \mu_{_{C_1}} = \mu_{_{C_2}} \\
H_A &: \mu_{_{C_1}} \neq \mu_{_{C_2}},
\end{aligned}$$

We set the significance level of each gene to $\alpha = .01$ *this gives us the Bonferroni correction of* $\alpha_{_{FWER}} = .01 \times 12,000 = 120$*, which is a joke. In addition, the assumption that the hypotheses, genes, are independent is ludicrous.*

We now use an alternative approach based upon a computational approach that falls under the class of permutation procedures. This lets us avoid the union

bound and control the FWER more accurately. We will also deal with the issue that t-distribution assumes normality and our data may not be normal. We will develop this procedure in the context of the previous example.

Example. *We will use a t-statistic to determine if a gene is differentially expressed under the two conditions. However, for each* $j = 1, ..., 12,000$ *we do not make distributional assumptions about the two distributions*

$$\begin{aligned} H_0 &: C_1 \text{ and } C_2 \text{ are exchangeable} \\ H_A &: C_1 \text{ and } C_2 \text{ are not exchangeable}, \end{aligned}$$

by exchangeable we mean loosely $\Pr(x, y|y \in C_1) = \Pr(x, y|y \in C_2)$.

For each gene we can compute the t-statistic: t_i. *We then repeat the following procedure* $\pi = 1, ..., \Pi$ *times for each gene*

(1) *permute labels*
(2) *compute* t_i^π.

For each gene i *we can get a p-value by looking at where* t_i *falls in the ecdf generated from the sequence* $\{t_i^\pi\}_{i=1}^{\Pi}$,

$$p_i = \hat{\Pr}(\xi > t_i | \{t_i^\pi\}_{\pi=1}^{\Pi}).$$

The above procedure gives us a p-value for the individual genes without an assumption of normality but how does it help regarding the FWER.

The following observation provides us with the key idea

$$\begin{aligned} \alpha_{FWER} &= \Pr(V \geq 1), \\ &= \Pr\left(\{T_1 > c\} \cup \{T_2 > c\}, ..., \cup \{T_m > c\} | H_0\right) \\ &= \Pr\left(\max_{i=1,...,m} \{T_i > c\} | H_0\right). \end{aligned}$$

This suggests the following permutation procedure For each gene we can compute the t-statistic: $\{t_i\}_{i=1}^m$, *where* $m = 12,000$. *Then repeat the following procedure* $\pi = 1, ..., \Pi$

(1) *permute labels*
(2) *compute* t_i^π *for each gene*
(3) *compute* $t^\pi = \max_{i=1,...,m} t_i^\pi$,

for each gene we can compute the FWER as

$$p_{i,FWER} = \hat{\Pr}(\xi > t_i | \{t^\pi\}_{\pi=1}^{\Pi}).$$

One very important aspect of the permutation procedure is we did not assume that the genes were independent and in some sense modeled the dependencies.

For many problems even this approach with a more accurate estimation of the FWER p-value does not give us significant hypotheses because the quantity we are trying to control $\Pr(V \geq 1)$ is too stringent.

3.0.12.2. *FDR.* The false discovery rate consists of finding a cutoff level or α value for the individual hypothesis tests such that we can control

$$q_{FDR} = \mathbb{E}\left[\frac{V}{R}\right],$$

this is controlling the proportion of false positives among the hypotheses we reject. For the reason that the above is not well defined when $R = 0$ we adjust the statistic so that we condition on there being rejects

$$q_{pFDR} = \mathbb{E}\left[\frac{V}{R}|R > 0\right].$$

We first illustrate the procedure with the permutation procedure we introduced previously as an example and then we will look at the more classical parametric case.

Example. *We will use a t-statistic to determine if a gene is differentially expressed under the two conditions. However, for each $j = 1, ..., 12,000$ we do not make distributional assumptions about the two distributions*

$$\begin{aligned} H_0 &: C_1 \text{ and } C_2 \text{ are exchangeable} \\ H_A &: C_1 \text{ and } C_2 \text{ are not exchangeable,} \end{aligned}$$

by exchangeable we mean loosely $\Pr(x, y|y \in C_1) = \Pr(x, y|y \in C_2)$.

For each gene we can compute the t-statistic: $\{t_i\}_{i=1}^m$, where $m = 12,000$. Then repeat the following procedure $\pi = 1, ..., \Pi$

(1) *permute labels*
(2) *compute t_i^π for each gene*

note that the statistics $\{t_i^\pi\}_{i,\pi}$ were all drawn under the assumption that the null is true. So if we reject any of them they would be a false positive. The statistics $\{t_i\}_{i=1}^m$, are drawn from a combination of hypotheses for which the null hypothesis holds true and those for which the alternative is true. Define the ranked list of the statistics under the null hypothesis as

$$\{Null_{(i)}\}_{i=1}^{\Pi \times m} = ranked\{t_i^\pi\} \quad for\ i = 1, ..., m, \pi = 1, ..., \Pi.$$

Similarly we can define the ranked list of the statistics coming from the set of the alternative and the null as

$$\{Tot_{(i)}\}_{i=1}^m = ranked\{t_i\} \quad for\ i = 1, ..., m.$$

We now look at a series of cutoffs corresponding to

$$c = Tot_{(1)}, Tot_{(2)}, ..., Tot_{(k)}.$$

For each value of c we can compute the following two quantities

$$\begin{aligned} \%R(c) &= \frac{\#\{Null_{(i)}\} \leq c}{m \times \Pi}, \\ \%V(c) &= \frac{\#\{Tot_{(i)}\} \leq c}{m}, \end{aligned}$$

from which we can compute the pFDR for the given cutoff c

$$q_{pFDR}(c) = \frac{\%V(c)}{\%R(c)}.$$