β

The Effects of Smoking & The Analysis of Variance copy 2

by Ramesh Hariharan

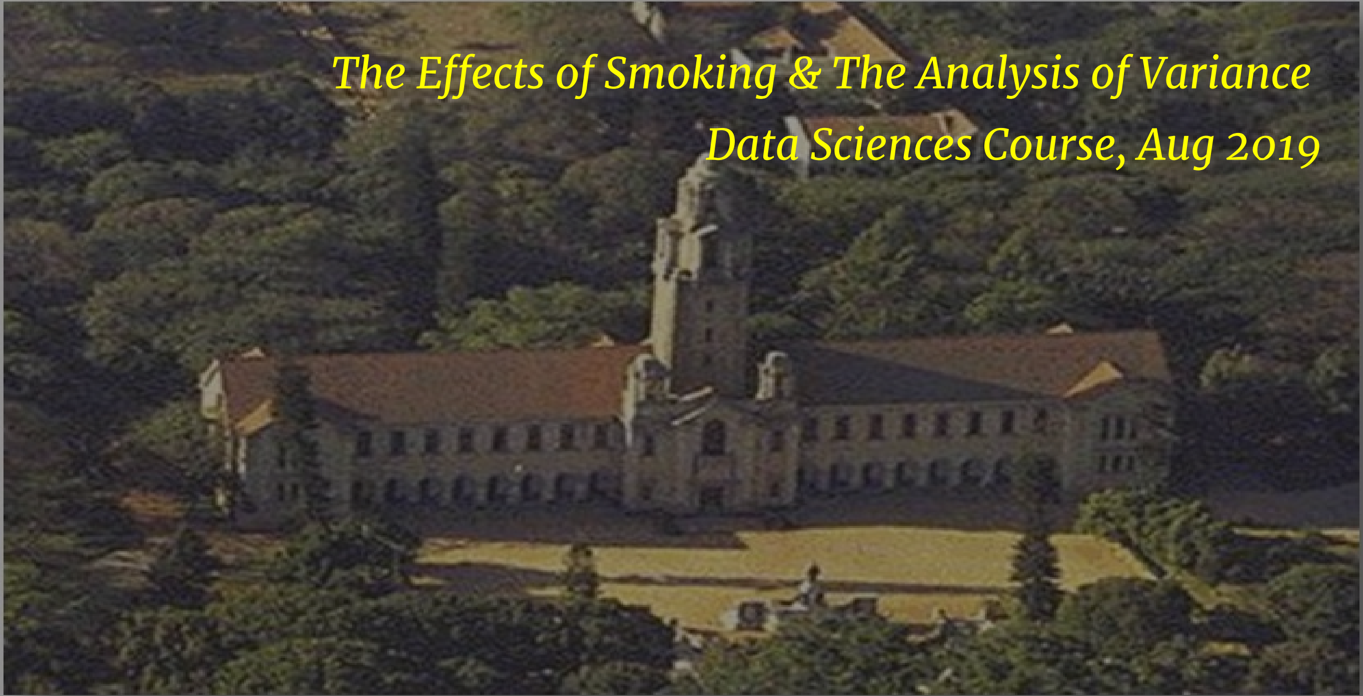
created: 2019-Oct-18

updated: 2019-Oct-18

views: 44

pages: 52

likes: 5

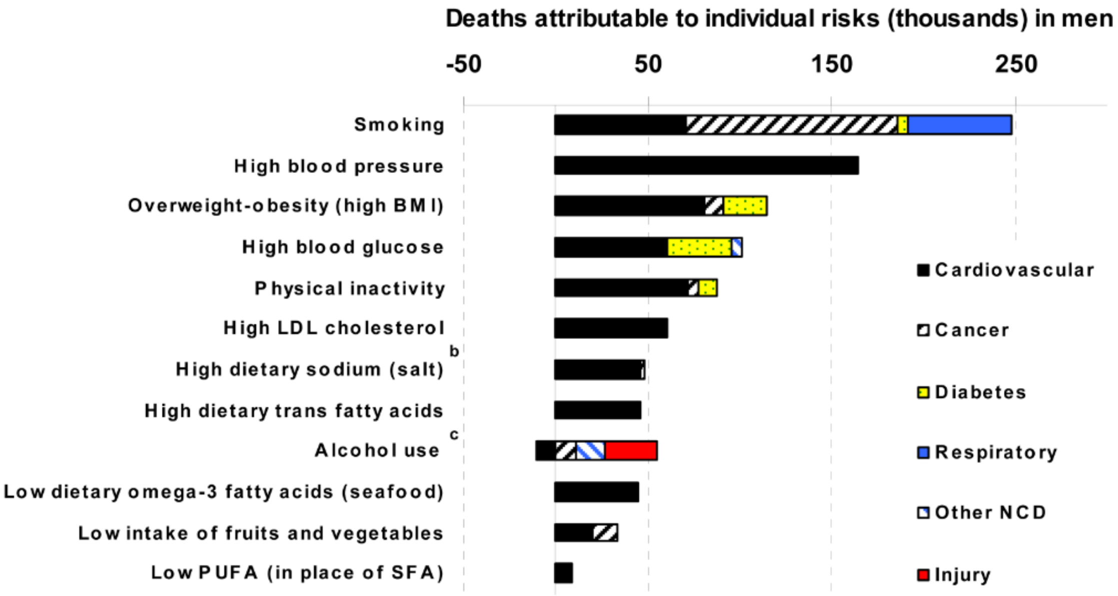


The Effects of Smoking & The Analysis of Variance

Data Sciences Course, Aug 2019

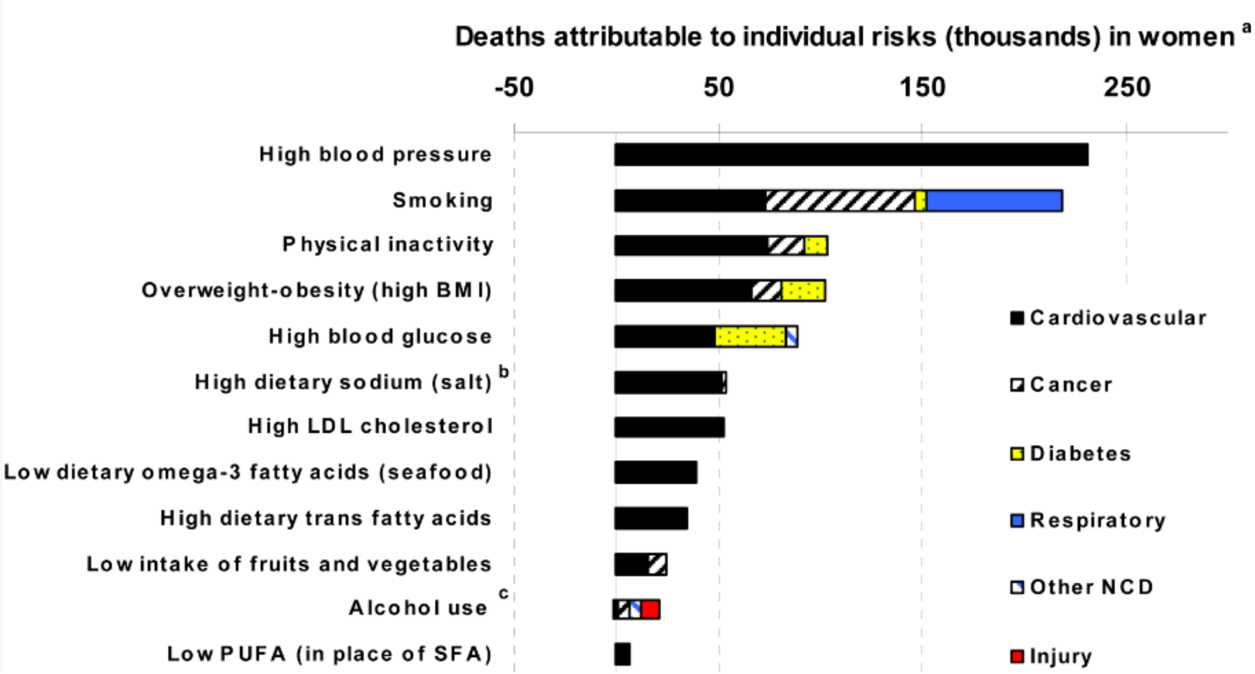
Ramesh Hariharan

Smoking leads Preventable Deaths in the US for Men



Ramesh Hariharan

Smoking leads Preventable Deaths in the US for Women As Well



Ramesh Hariharan

Differential Effects on Women

- Effects of smoking are more serious for women than for men
  - more vulnerable to cigarette smoke-induced respiratory diseases
  - adverse affects on fertility, early menopause, pregnancy complications
  - higher risk of type-2 diabetes
  - higher absolute risk for lung cancer
  - additional hazards such as breast cancer, ovarian cancer, and cancer of the cervix
- On the positive side, women have been found to have higher survival rates regardless of lung cancer type, stage and therapy

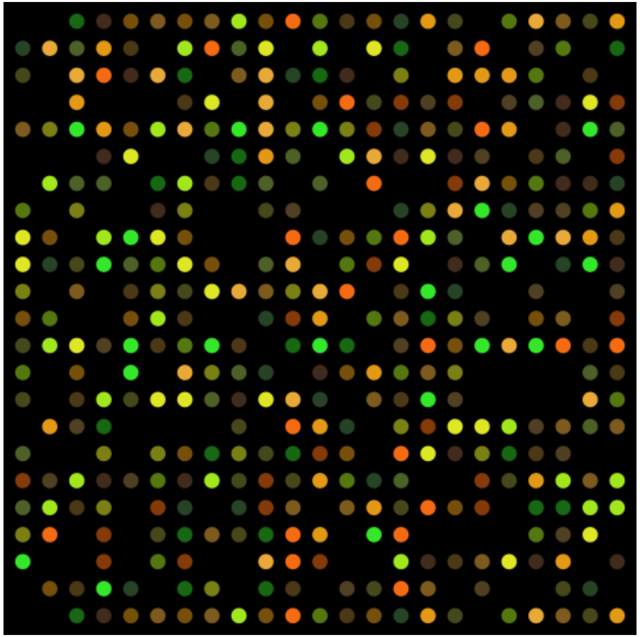
Genes & Gene Expression

- ~20,000 genes in each cell of the body, coded in (largely) read-only DNA
- Genes are transcribed to mRNA, then translated to protein
- Proteins react chemically to drive various biological functions
- The amount of gene-->mRNA transcription (gene expression) is dynamic, as a function of the cell type, the stimulus, age etc
- How does gene expression for the various genes respond to smoke?
- Which gene increase in expression, which decrease?
- Are these different between men and women?
- What biological functions do these differential genes influence?
- Does the disruption in these functions explain observed pathology due to smoking?

Measuring Gene Expression

- A gene is a double stranded sequence of A, C, G, Ts, very stable
- The two strands are complementary
- Could be thousands or tens of thousands of characters long
- mRNA is a single stranded sequence, but highly unstable, meant for temporary purposes
- mRNA can be extracted from a cell and converted to the complementary single stranded DNA (called cDNA)
- A probe is a shorter DNA sequence, ~25-100 characters long, complementary to this cDNA
- Many copies of a probe can be spotted on a glass surface, with different spots carrying probes for cDNA from different genes
- Typically use a few distinct probes per gene, so tens of thousands of spots
- mRNA converted to cDNA from a collection of cells is then poured on the glass slide
- cDNA from each gene gravitates towards its respective spot
- A cDNA molecule hybridized to its probe glows
- The glow at a spot is proportional to the amount of mRNA for that gene

A Microarray Picture

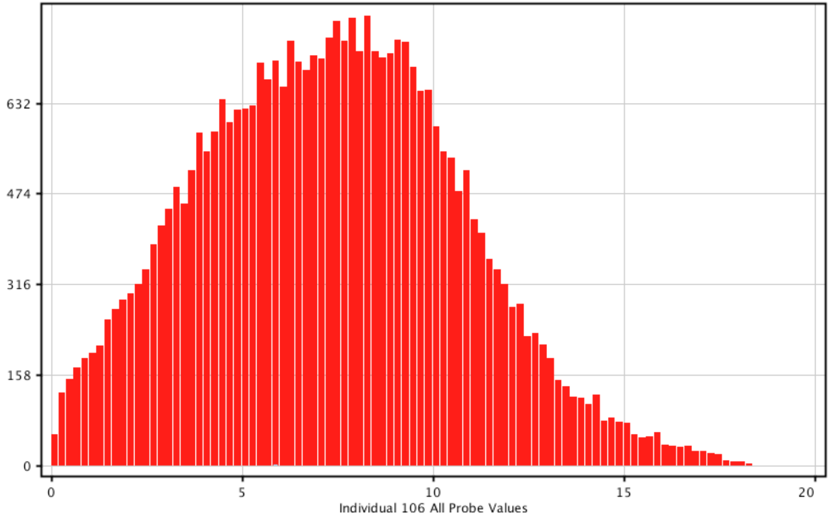


- Image analysis identifies each spot and measures its intensity
- For each probe, we now have the intensity
- And we also know the corresponding gene for each probe
- So, we now have one or more measurements of the expression level of each of the 20,000 genes
- And we can repeat this for multiple humans, some who are smokers, some non-smokers, some males, some females

The Data

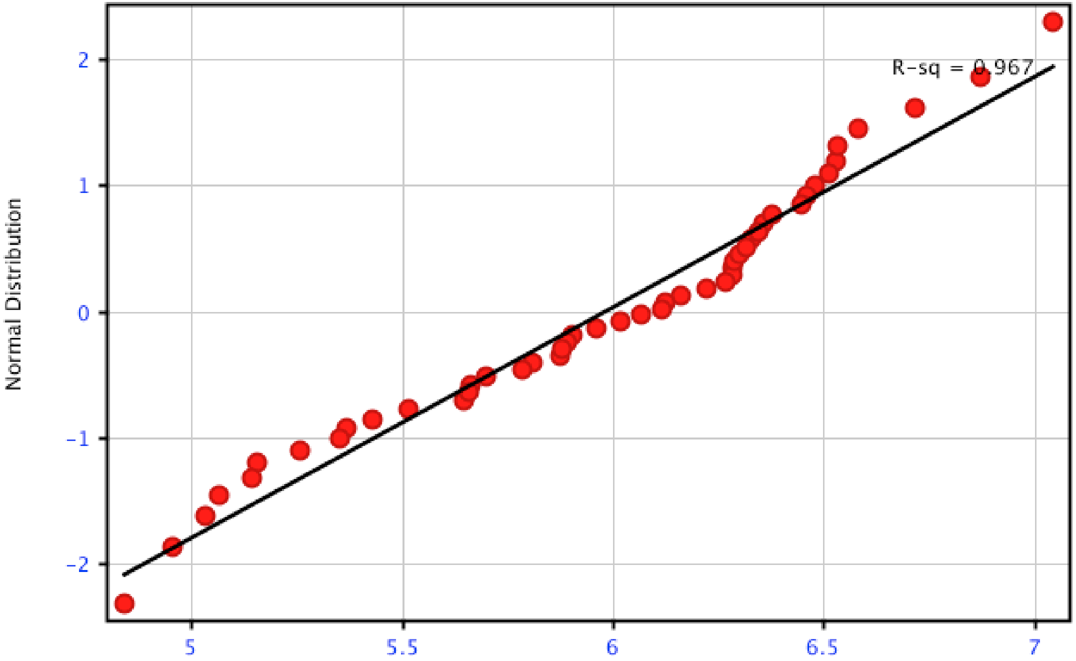
- Data is generated from white blood cells from 48 individuals reference
- A single file with 48 columns of data, plus some auxiliary columns, here
- Auxiliary columns: Probe name, Gene Symbol, Entrez Gene Id, ignore the rest
- A single gene (identified by a Gene Symbol or Entrez Gene Id) could have multiple probes
- Totally 41,094 probes
- Data Columns:
    - 12 Male Non-smokers (106-117)
    - 12 Male Smokers (118-129)
    - 12 Female Non-Smokers (130-141)
    - 12 Female Smokers (142-153)
- Values are logs to the base 2 of the original value
- There are some 0 values as well, due to thresholding low value before taking the log

Distribution for a Single Individual



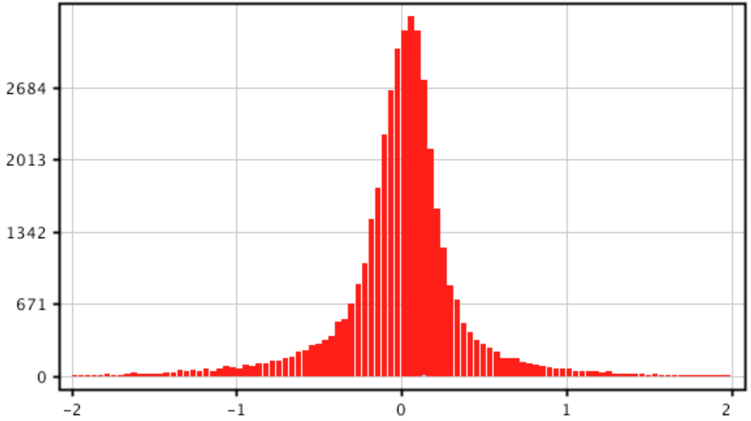
- Values go from roughly 0 to 20 on the log scale, so roughly 0-1,000,000 on the linear scale
- Median roughly 8.6, or 400 on the linear scale
- Distribution not quite Gaussian

Distribution for a Single Probe & Normality



- This Normal Probability Plot displays the 48 data points for probe A_24_P470079 against a corresponding number of (almost) equi-area-distant data points from a Gaussian $N(0,1)$ distribution ($\Phi^{-1}(\frac{i-0.5}{n})$)
- A straight line indicates the data is close to Gaussian
- This will be important later

Are Probe Variances the Same in the Various Groups?

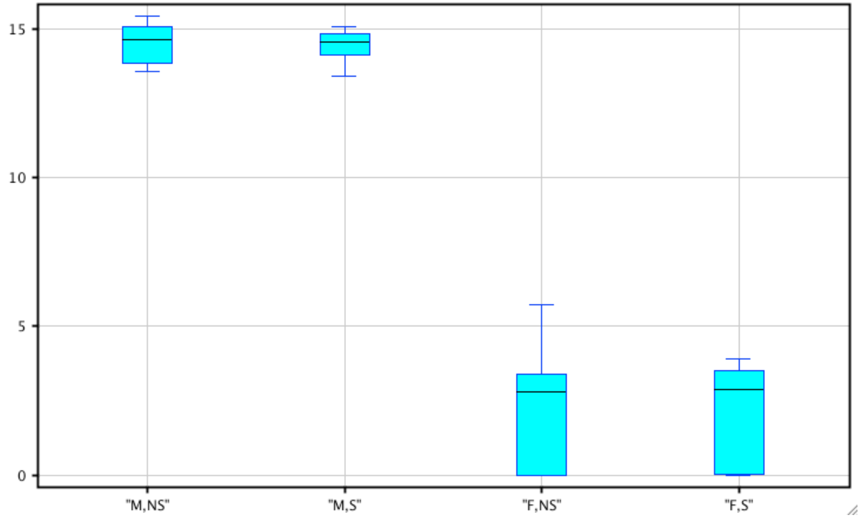


- Take the standard deviation across samples in a group for a single probe
- Take the difference of the above between two groups
- Plot the distribution of this quantity across all probes
- It turns out to be a distribution centered around 0, indicating standard deviations are similar in the various groups
- This will be important as well later

Differential Expression Analysis

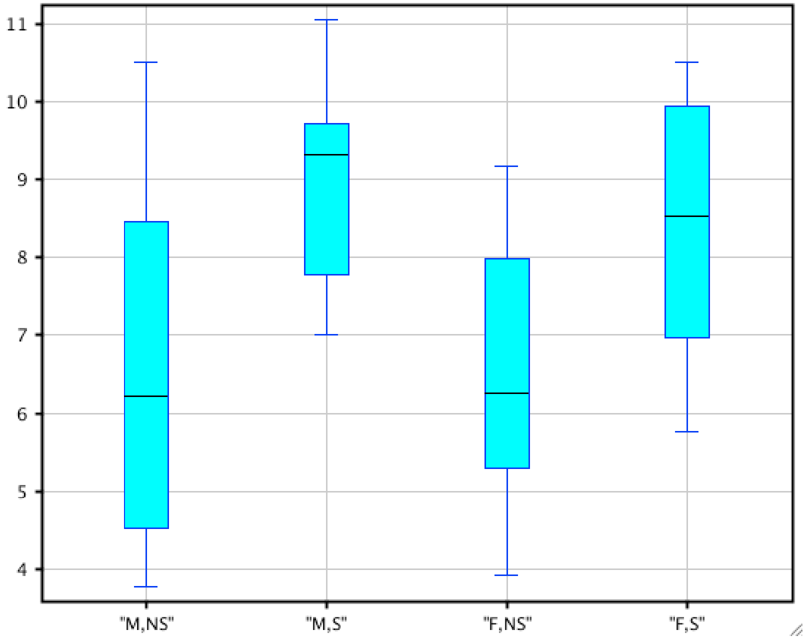
- Which of the ~20,000 genes behaves differently between
  - Males and Females, independent of Smoking Status
  - Smokers and Non-smokers, independent of Gender
  - Male Smokers and Non-smokers, vs Female Smokers and Non-smokers
- What do these genes tell us about our observations on females being more susceptible to smoking-related diseases but more robust to surviving smoking-related cancer?

Male vs Female Differential Response (*RPS4Y2*, A23_P324384)



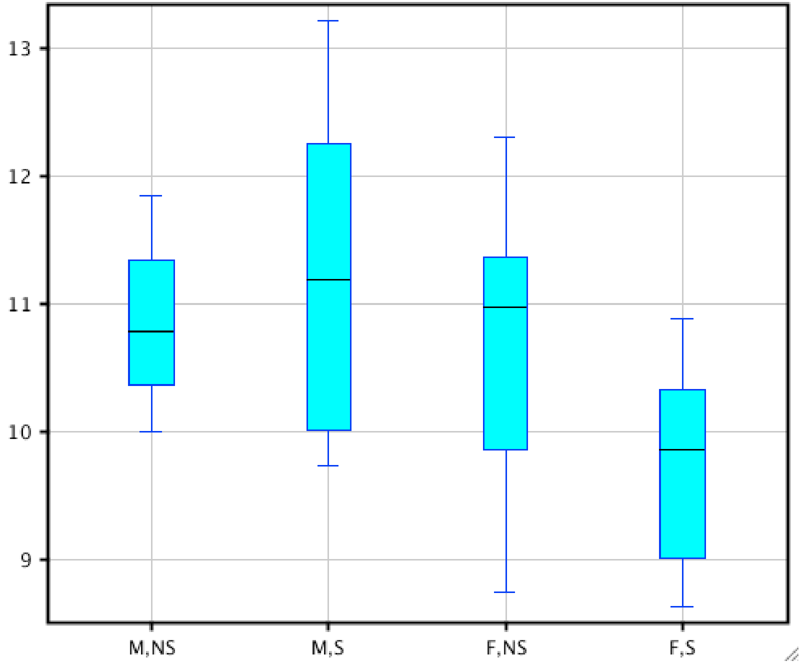
- The *RSPY2* gene is clearly different between males and females, regardless of smoking status
- The gene is located on the Y chromosome, which explains it
- There are 3 paralogs (similar copies) in the genome, one on the X chromosome and 2 on the Y chromosome
- Females express both copies on the X chromosome, while males express one on X and the two on Y, the latter to a lesser degree link

Smoker vs Non-Smoker Diff. Response (*AHR*, A23_P215566)



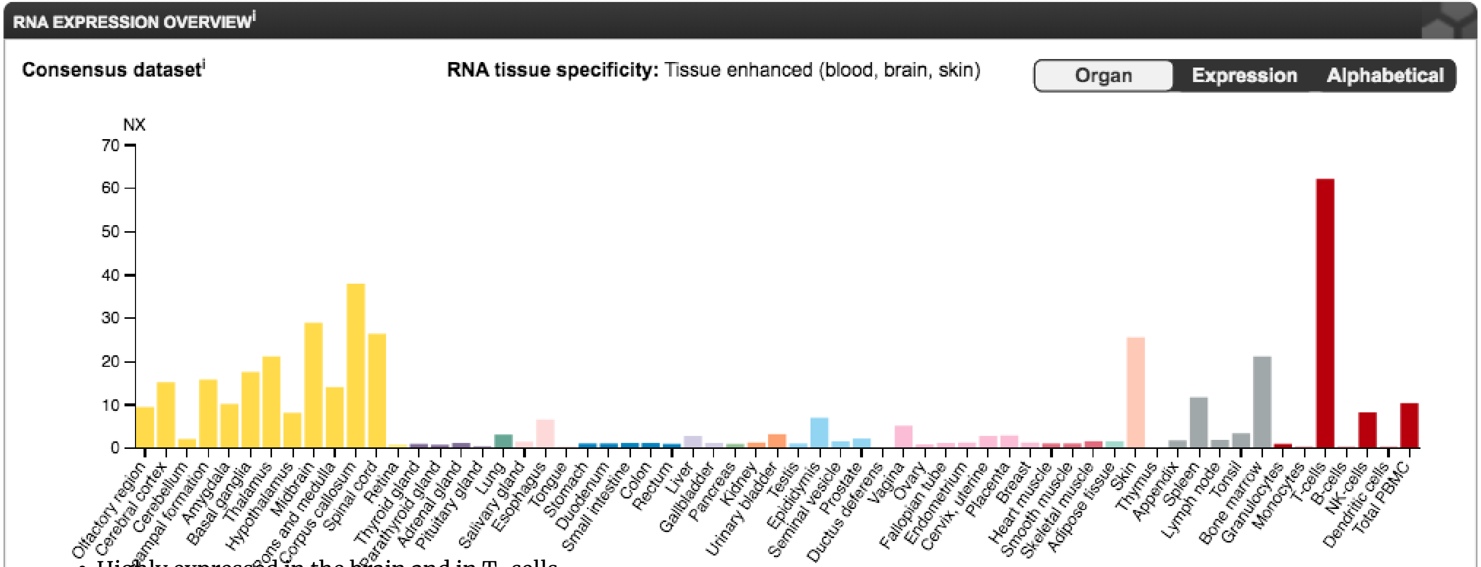
- The Aryl Hydrocarbon receptor (*AHR*) is a sensor of xenobiotic chemicals, such as those found in smoke
- It also causes the expression of other genes which metabolize (break down) these chemicals
- Side-effects of this breakdown include free radicals which cause DNA damage, which is widespread in smokers
- In addition, *AHR* expression is also found to be increased in many cancer cells, indicating a link to cancer

Gender x Smo. Status Diff. Response (*S1PR5*, A23_P107744)



- *S1PR5* is not so different between smokers and non-smokers in men
- Or between men and women non-smokers
- However, its expression appears quite reduced in female smokers relative to both female non-smokers and with males as a whole
- Little is known about the function of this gene

S1PR5, and Immune Response to Smoking in Women



RNA EXPRESSION OVERVIEW[i]

Consensus dataset[i]

RNA tissue specificity: Tissue enhanced (blood, brain, skin)

- Highly expressed in the brain and in T-cells
- Could play a role in immunity
- Indeed, it is known that changes in the immune system of women smokers are more pronounced than in men ref

Summary of Analysis

- Genes associated with cancer and the immune system are altered in both females and male smokers vs non-smokers
- Many more immune function genes were down-regulated in female smokers than in males; these differential changes in immune function in females could explain their greater susceptibility to several diseases
- Many genes associated with DNA repair, xenobiotic metabolism, free radical scavenging and natural killer cells cytotoxicity are down-regulated in female smokers relative to male smokers; these could explain the increased susceptibility of females to smoke-induced cancer
- There may also be some clues as to why females survive cancer better, though a conclusive answer is still not there (e.g., *CYP4F2*, *CYP4F12*)

Switching now to The Hypothesis Testing Problem

- Suppose you have an hypothesis, e.g.
    - Male and female heights have the same distribution, *versus*
    - Male and female heights do NOT have the same distribution
- How would you verify which is the case?
- You sample a few males and a few females *independently* at random, and measure their heights
- From this random sample, you could estimate the mean and variance of the underlying distribution(s) (*how*?)
- And then check if the means of the two distributions are the same or not (*how*?)

Sample Mean and Distribution Mean

- How does the sample mean relate to the mean of the underlying distribution?
- Let X be a random variable denoting the underlying distribution
- Let $E(X) = \mu, Var(X) = \sigma^2$
- Let $X_1, \ldots, X_n$ be the *independent* sample values (say heights of samples males)
- Then $E(X_i) = \mu, \forall i$
- $E(\Sigma X_i/n) = \Sigma_i E(X_i)/n = \Sigma_i \mu/n = \mu$ (note: Linearity of Expectation, regardless of independence)
- So the expected value of the sample mean is μ
- But how close does the sample mean get to μ?

Distribution of the Sample Mean

- The distribution of the sample mean (which is different from the underlying distribution) has mean μ
- What is the variance of this distribution?
- *This is where independence comes in*
- $Var(\Sigma X_i/n) = \Sigma_i Var(X_i)/n^2 = \sigma^2/n$ (note: Linearity of Variance, given independence)
- Variance $\propto \frac{1}{n}$, Std Dev $\propto \frac{1}{\sqrt{n}}$
- In fact, as n increases, the distribution of the sample mean gets closer to $N(\mu, \sigma/\sqrt{n})$ (the Normal distribution, regardless of the underlying distribution)

Estimating the Mean of the Underlying Distribution

- Suppose you want to estimate the underlying distribution mean so with greater than 95% probability you are within 10% of the actual number
  - Choose n so that $N(\mu, \sigma/\sqrt{n})$ has less than 5% outside $\mu(1 \pm 0.1)$
  - Then pick n samples independently at random
  - Take the sample mean as an estimate of the underlying distribution mean
- Catch: You need to know the variance σ^2!!
- How does one estimate the variance of the underlying distribution?

Estimating the Variance of the Underlying Distribution

- Would sample variance estimate the variance of the underlying distribution?
- Sample variance = $\frac{1}{n}\Sigma_i(X_i - \Sigma_j X_j/n)^2$
- $E[\frac{1}{n}\Sigma_i(X_i - \Sigma_j X_j/n)^2]$
- $= \frac{1}{n}E[\Sigma_i X_i^2 + \Sigma_i(\Sigma_j X_j/n)^2 - 2\Sigma_i X_i(\Sigma_j X_j/n)]$
- $= \frac{1}{n}E[\Sigma_i X_i^2 - (\Sigma_j X_j)^2/n] = \frac{1}{n}E[\Sigma_i \frac{n-1}{n} X_i^2 - \frac{2}{n}\Sigma_{i\neq j} X_i X_j]$
- $= \frac{1}{n}[\frac{n-1}{n}\Sigma_i E(X_i^2) - \frac{2}{n}\Sigma_{i\neq j} E(X_i X_j)]$
- $= \frac{1}{n}[(n-1)E(X_1^2) - (n-1)\mu^2]$ (note $E(X_i X_j) = E(X_i)E(X_j)$ by independence)
- $= \frac{n-1}{n}[E(X_1^2) - \mu^2] = \frac{n-1}{n}\sigma^2$ (note: not quite the distribution variance)

An Unbiased Estimator for the Distribution Variance

- The sample variance underestimates the variance of the underlying distribution!
- $E[\frac{1}{n}\Sigma_i(X_i - \Sigma_j X_j/n)^2] = \frac{n-1}{n}\sigma^2$
- Use $\frac{1}{n-1}\Sigma_i(X_i - \Sigma_j X_j/n)^2$ instead!
- In summary
    - the sample mean is an unbiased estimator for the distribution mean
    - $\frac{n-1}{n}$ times the sample variance is an unbiased estimator for the distribution variance
    - pick n so the distribution of the sample mean is tight around the distribution mean (based on the estimated distribution variance), or based on *what you can afford*

Are the Male and Female Distributions *Significantly* Different?

- You take an independent random sample of males, and a separate sample of females; measure heights in each sample and find the sample means $\hat{\mu}_M$ and $\hat{\mu}_F$ (*the hat typically indicates sample as opposed to distribution*)
- Are the two distributions different, i.e., are the sample means significantly different? (Assumption: variances are the same)
- What does signficant mean?
- $|\hat{\mu}_M - \hat{\mu}_F| > \Delta$ for some suitable Δ?
- If σ is large, then even if the two distributions were the same, $\hat{\mu}_M$ and $\hat{\mu}_F$ could be quite different (recall sample mean $\sim N(\mu, \sigma/\sqrt{n})$)
- So perhaps $\frac{|\hat{\mu}_M - \hat{\mu}_F|}{\sqrt{\hat{\sigma}_M^2 + \hat{\sigma}_F^2}} > \Delta$?

Test Statistics and P-Values

- $\hat{t} = \frac{|\hat{\mu}_M - \hat{\mu}_F|}{\sqrt{\hat{\sigma}_M^2 + \hat{\sigma}_F^2}}$ is called a *test statistic*
- This statistic has a certain distribution T (imagine sampling many many times and seeing what values of the statistic you get)
- T depends on the two underlying distribution(s) and the sample size
- The probability that $T \geq \hat{t}$ assuming that the two underlying distributions are the same (i.e., same means) is called the *p-value*
- A small p-value indicates that the two underlying distributions are likely to be different
- The conventional cut-off is 0.05
- To derive a p-value, one must first define a statistic (as a function of the sample values) and then obtain the distribution of this statistic assuming identical underlying distributions

The ANOVA Approach for Heights

- Two models:
    - The gender-specific model: Males $h = \mu_M + \epsilon$, females $h = \mu_F + \epsilon$, where ϵ is random variable with some distribution (to be determined) with mean 0
    - The universal model: $h = \mu + \epsilon$ (we will call this the *null model*)
- Now we sample females with height $\hat{h}_1, \ldots, \hat{h}_n$ and males $\hat{h}_{n+1}, \ldots, \hat{h}_m$
- From this sample, we determine the values of μ_M, μ_F that minimize $\alpha^2 = \mathbf{min}_{\mu_F, \mu_M} [\sum_{i=1}^{n} (\hat{h}_i - \mu_F)^2 + \sum_{i=n+1}^{m} (\hat{h}_i - \mu_M)^2] = \sum_{i=1}^{m} \hat{\epsilon}_i^2$ using standard linear regression
- Next we do the same for the null model: find μ that minimizes $\alpha'^2 = \mathbf{min}_{\mu} [\sum_{i=1}^{m} (\hat{h}_i - \mu)^2] = \sum \hat{\epsilon}_i^2$
- The statistic we use is $F = \frac{\hat{\alpha'^2}}{\hat{\alpha^2}} - 1 = \frac{\hat{\alpha'^2} - \hat{\alpha^2}}{\hat{\alpha^2}}$ is called the F-statistic (with a small modification later)
    - If it is large, then the null model is a less likely candidate from which the data at hand could have been drawn
    - What is the distribution of this statistic under the null model? How do we get a p-value?

The Gender-Specific Model in Matrix Form

- $$\begin{pmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \vdots \\ \hat{h}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_F \\ \mu_M \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} \tag{1}$$

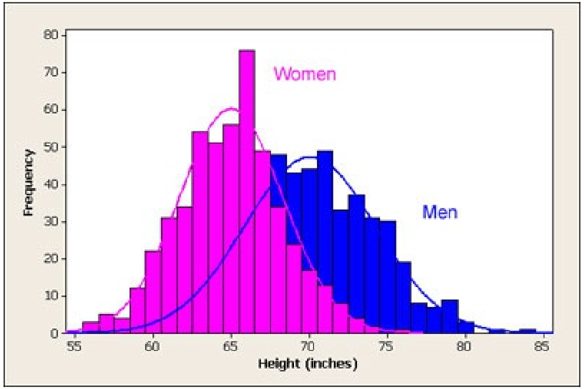
*There are m 1's and $n-m$ 0's in the first column

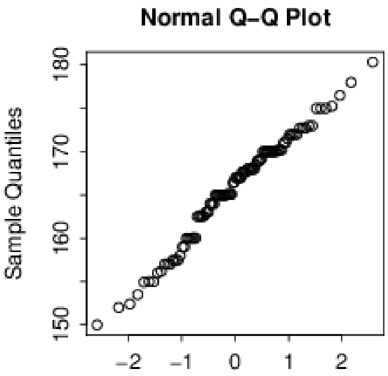
The Universal (Null) Model in Matrix Form

- $$\begin{pmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \vdots \\ \hat{h}_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \end{pmatrix} \left(\mu \right) + \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} \tag{2}$$

- Further, we will need some assumption of the distribution from which $\hat{\epsilon}_i$'s are drawn; typical assumption is $N(0, \sigma^2)$, but needs to be justified

The $N(0, \sigma^2)$ Assumption





- Adult male heights mean 70in, s.d 4in link
- Adult female heights mean 65in, s.d 3.5in link
- Both appear Gaussian as shown by the Q-Q Plot link

Linear Regression to Minimize Sum of Squared Errors

- For each model, A and A', in turn (shown below for A only)
    - $\vec{\hat{h}} = A\vec{x} + \vec{\hat{\epsilon}}$
    - $(\vec{\hat{h}} - A\vec{x})^T(\vec{\hat{h}} - A\vec{x}) = \vec{\hat{\epsilon}}^T\vec{\hat{\epsilon}} = \Sigma_i \hat{\epsilon}_i^2$
    - To minimize the RHS over $\vec{x}$, we need to solve $A^T\vec{\hat{h}} = A^TA\vec{x}$
    - The best $\vec{x} = (A^TA)^\dagger A^T\vec{\hat{h}}$, where † is the pseudoinverse
    - The minimum value of $\Sigma_i \hat{\epsilon}_i^2$ then is $(\vec{\hat{h}} - A(A^TA)^\dagger A^T\vec{\hat{h}})^T(\vec{\hat{h}} - A(A^TA)^\dagger A^T\vec{\hat{h}}) = \vec{\hat{h}}^T(I - A(A^TA)^\dagger A^T)\vec{\hat{h}}$ $\ldots\mathbf{1}$
    - The above follows because $(I - A(A^TA)^\dagger A^T)$ is idempotent (X is idempotent if $XX = X$).

The F-Statistic & P-value

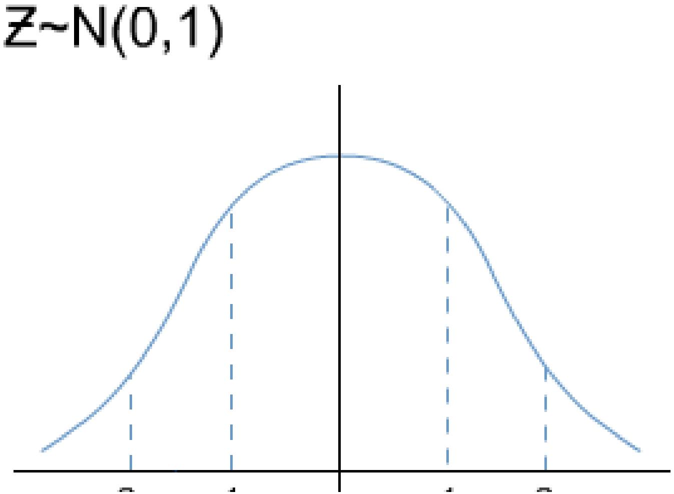
- The two A's of interest are:
- F-Statistic$= \frac{\vec{\hat{h}}^T (I - A'(A'^T A')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - A(A^T A)^\dagger A^T)\vec{\hat{h}}} - 1 = \frac{\vec{\hat{h}}^T (A(A^T A)^\dagger A^T - A(A^T A)^\dagger A^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - A(A^T A)^\dagger A^T)\vec{\hat{h}}}$
- To compute the p-value, we need the distribution of this statistic under the assumption that the universal (null) model holds, i.e., each $\hat{\epsilon}_i$ is drawn independently from $N(0, \sigma^2)$, i.e., $\hat{h}_i$ is drawn from $N(\mu, \sigma^2)$.
- A small p-value means the universal (null) model is unlikely to support the F-statistic derived from the data at hand, so the universal (null) model can be rejected
- Additionally, since we have evidence that heights are distributed as Gaussian and with the same/similar variance for males and females, then a low p-value shows that the gender-specific model is likely

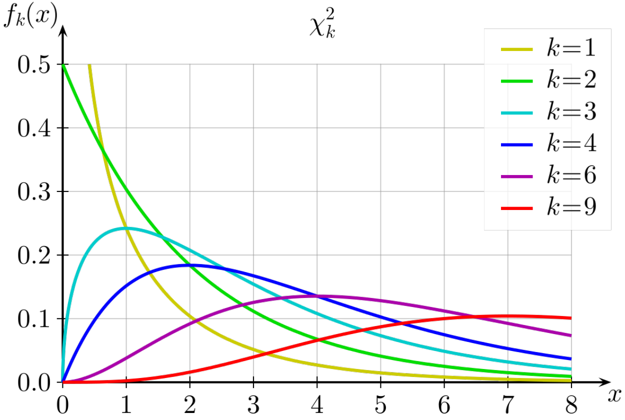
$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, A' = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad (3)$$

Distribution of the Num/Denom under the Null Model

- What is the distribution of $\vec{\hat{h}}^T f(X,Y)\vec{\hat{h}}$, where $f(X,Y) = X(X^TX)^\dagger X^T - Y(Y^TY)^\dagger Y^T$?
- $\vec{\hat{h}}^T f(X,Y)A^T\vec{\hat{h}} = \sigma^2 \times \frac{\vec{\hat{h}}-A'\mu}{\sigma}^T f(X,Y)\frac{\vec{\hat{h}}-A'\mu}{\sigma}$ (because $f(X,Y)A'\mu = \vec{0}$) $\ldots\mathbf{2}$
- $\vec{\hat{y}} = \frac{\vec{\hat{h}}-A'\mu}{\sigma}$ has entries that are independently sampled from a standard normal $N(0,1)$
- Claim: If the column space of Y is a subspace of the column space of X, and ALL eigenvalues of $X(X^TX)^\dagger X^T, Y(Y^TY)^\dagger Y^T$ are either 0 or 1 then $\ldots\ldots\mathbf{3}$
  - Distribution of $\vec{\hat{y}}^T f(X,Y)\vec{\hat{y}}$ is that of $\vec{\hat{y}}^T \Sigma\vec{\hat{y}}$ where Σ is the diagonal eigenvalue matrix of $f(X,Y)$ with $rank(X) - rank(Y)$ 1 eignvalues
  - Which is the sum of squares of $rank(X) - rank(Y)$ independent $N(0,1)$ random variables, or a Chi Square distribution with $rank(X) - rank(Y)$ degrees of freedom
  - $E[\vec{\hat{y}}^T f(X,Y)\vec{\hat{y}}] = rank(X) - rank(Y)$

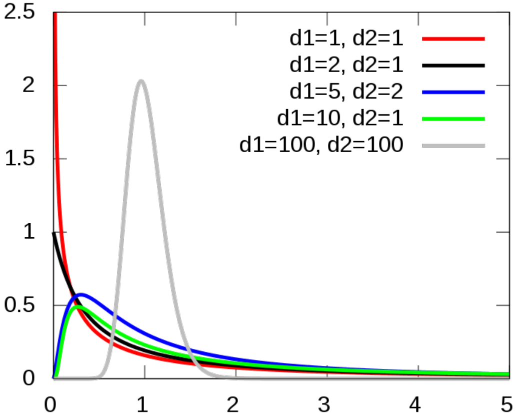
Chi Square Distributions for various Degrees of Freedom





Distribution of the F-Statistic under the Null Model

- Modified F-statistic (note the second term with degrees of freedom) $= \frac{\vec{\hat{h}}^T (A(A^TA)^{\dagger}A^T - A'(A'^TA')^{\dagger}A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I-(A(A^TA)^{\dagger}A^T)\vec{\hat{h}}} * \frac{n-rank(A)}{rank(A')-rank(A)}$
- the numerator is Chi Square distributed with $rank(A) - rank(A')$ degrees of freedom, the denominator is Chi Square distributed with $n - rank(A)$ degrees of freedom, and the two are independent
- the statistic is distributed as the F-distribution, with parameters $rank(A') - rank(A) = 1$ and $n - rank(A) = n - 2$
- Closed-form and calculators are commonly available for the F-distribution. The p-value is the total area to the right of the value calculated from the given data.



Proof of Claims toward the Distribution of the F-Statistic

- $A^T A\vec{x} = \vec{y}$ has a solution for any vector $\vec{y}$ in the column space of A^T, i.e., $\vec{x} = (A^T A)^\dagger \vec{y}$
    - $A^T A\vec{x}$ is clearly in the column space of A^T
    - Therefore, the column space of $A^T A$ is a subspace of the column space of A^T
    - It suffices to show that the column spaces of $A^T A$ and A^T have the same dimension; that would imply equality of the two spaces
    - Since nullspace of $A^T A$ and A are the same, the column spaces of $A^T A$ and A have the same dimension
    - The column space of A has the same dimension as the row space of A
    - It follows that the column space of $A^T A$ has the same dimension as the column space of A^T

Proof of Claims toward the Distribution of the F-Statistic

- $A^T A\vec{x} = \vec{y}$ has a solution for any vector $\vec{y}$ in the column space of A^T, i.e., $\vec{x} = (A^T A)^\dagger \vec{y}$
- $\implies$ $A(A^T A)^\dagger A^T \vec{y} = \vec{y}$ for any vector $\vec{y}$ in the column space of A (i.e., $\vec{y} = A\vec{z}$), and 0 for all vectors $\vec{y}$ orthogonal to this column space. Ditto for A'
- $\implies$ Eigenvalues of $A(A^T A)^\dagger A^T$ are 0,1 with exactly $rank(A)$ 1's. Ditto for A'
- $\implies$ All vectors in the column space of A are eigenvectors of $A(A^T A)^\dagger A^T$. Ditto for A'
- $\implies$ $(I - A(A^T A)^\dagger A^T)(I - A(A^T A)^\dagger A^T) = (I - A(A^T A)^\dagger A^T)$ because $A(A^T A)^\dagger A^T A = A$. Ditto for A'
- $\implies$ $A(A^T A)^\dagger A^T A'\mu = A'(A'^T A')^\dagger A'^T A'\mu$, where $A'\mu$ is the non- gender specific model (because $A'\mu$ is in the column space of A', which in turn is in the column space of A). Ditto for A'

Proof of Claims toward the Distribution of the F-Statistic

- $\implies A(A^TA)^\dagger A^T, A'(A'^TA')^\dagger A'^T$
  - have $rank(A')$ orthonormal eigenvectors in common with eigenvalue 1 (because the column space of A' is a contained in that of A)
  - have $rank(A) - rank(A')$ orthonormal eigenvectors in common, but with eigenvalue 1 for $A(A^TA)^\dagger A^T$ and 0 for $A'(A'^TA')^\dagger A'^T$ (these are in the column space of A but orthogonal to that of A')
  - have $n - rank(A)$ orthonormal eigenvectors in common with eigenvalue 0 (these are orthogonal to the column spaces of both A and A')
- $\implies A(A^TA)^\dagger A^T = V\Sigma V^T, A'(A'^TA')^\dagger A'^T = V\Sigma' V^T$, where Σ, Σ' are the corresponding diagonal eigenvalue matrices and the columns of V are the orthonormal eigenvectors
- $\implies A(A^TA)^\dagger A^T - A'(A'^TA')^\dagger A'^T = V(\Sigma - \Sigma')V^T$, where $\Sigma - \Sigma'$ has only $rank(A) - rank(A')$ 1's

Proof of Claims toward the Distribution of the F-Statistic

- $\implies I - A(A^TA)^\dagger A^T = V(I-\Sigma)V^T$, where $I-\Sigma$ has only $n - rank(A)$ 1's
- $\implies \vec{\hat{y}}^T(A(A^TA)^\dagger A^T - A'(A'^TA')^\dagger A'^T)\vec{\hat{y}} = \vec{\hat{y}}^T V(\Sigma - \Sigma')V^T\vec{\hat{y}}$, where $\vec{\hat{y}}$ is a vector whose entries are chosen independently from $N(0,1)$
- $\implies \vec{\hat{y}}^T(I - A(A^TA)^\dagger A^T)\vec{\hat{y}} = \vec{\hat{y}}^T V(I-\Sigma)V^T\vec{\hat{y}}$, where $\vec{\hat{y}}$ is a vector whose entries are chosen independently from $N(0,1)$
- Since the columns of V are orthonormal, we could rotate the coordinate axes so V becomes I. What happens to $\vec{\hat{y}}$ in the process?
- Because of spherical symmetry, $\vec{\hat{y}}$ remains a vector whose entries are chosen independently from $N(0,1)$!
- $\implies \vec{\hat{y}}^T(A(A^TA)^\dagger A^T - A'(A'^TA')^\dagger A'^T)\vec{\hat{y}}$ has the same distribution as $\vec{\hat{y}}^T(\Sigma - \Sigma')\vec{\hat{y}}$
- $\implies \vec{\hat{y}}^T(I - A(A^TA)^\dagger A^T)\vec{\hat{y}}$ has the same distribution as $\vec{\hat{y}}^T(I-\Sigma)\vec{\hat{y}}$

Proof of Claims toward the Distribution of the F-Statistic

- $\vec{\hat{y}}^T(\Sigma - \Sigma')\vec{\hat{y}}$ is the sum of squares of a subset of the entries of $\vec{\hat{y}}$
- Likewise for $\vec{\hat{y}}^T(I - \Sigma)\vec{\hat{y}}$
- The two subsets are disjoint because the 1s in Σ' are a subset of the 1s in Σ
- $\implies$
  - $\vec{\hat{y}}^T(A(A^TA)^\dagger A^T - A'(A'^TA')^\dagger A'^T)\vec{\hat{y}}$ is distributed as the sum of squares of $rank(A) - rank(A')$ independent $N(0,1)$ random variables
  - $\vec{\hat{y}}^T((I - A(A^TA)^\dagger A^T)\vec{\hat{y}}$ is distributed as the sum of squares of $n - rank(A)$ independent $N(0,1)$ random variables
  - The two distributions are independent

Summarizing ANOVA for Heights

- Sample females with height $\hat{h}_1, \ldots, \hat{h}_n$ and males $\hat{h}_{n+1}, \ldots, \hat{h}_m$
- Calculate F-statistic $\hat{f} = \frac{\vec{\hat{h}}^T (A(A^T A)^\dagger A^T - A'(A'^T A')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - (A(A^T A)^\dagger A^T)\vec{\hat{h}}} * \frac{n - rank(A)}{rank(A) - rank(A')}$
- Calculate the area to the right of $\hat{f}$ in the density plot of the F-distribution with $rank(A) - rank(A')$ and $n - rank(A)$ degrees of freedom; this is the p-value
- Reject the universal (null) model if p-value is small
- Note, the proofs above hinge on the following two facts
    - We have two models A, A'
    - The column space of the null model A' is contained in that for A
    - The $\hat{\epsilon}$'s in the null model are all sampled from $N(0, \sigma^2)$ independently
- Further given there is reason to believe that heights are Gaussian and with equal/similar variance across genders, unlikeliness of the universal model translates to likeliness of the gender-specific model, with distinct means for males and females

Many Groups

- The two A's of interest are:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, A' = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{4}$$

- F-Statistic$= \frac{\vec{\hat{h}}^T (I - A'(A'^T A')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - A(A^T A)^\dagger A^T)\vec{\hat{h}}} - 1 = \frac{\vec{\hat{h}}^T (A(A^T A)^\dagger A^T - A(A^T A)^\dagger A^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I - A(A^T A)^\dagger A^T)\vec{\hat{h}}}$
- The rest of the process is identical to the two-groups case above
- Rejection of the null model suggests that the underlying group means are not all the same

Two-Way ANOVA

- Two dimensions of groups (or *factors*)
    - Status: Smokers vs Non-Smokers
    - Gender: Males vs Females
- There are three questions now
    - Are the means the same for Males and Females?
        - Derive f_{Gender} as above.
    - Are the means the same for Smokers vs Non-Smokers?
        - Derive f_{Status} as above.
    - Are the individual means for each of the 4 Status x Gender groups just a simple additive combination of the Status and the Gender means?
        - What about this?

Status x Gender Model

- $$\begin{pmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \vdots \\ \hat{h}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} g_{S,M} \\ g_{S,F} \\ g_{NS,M} \\ g_{NS,F} \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} \qquad (5)$$

Status + Gender (Null) Model

- $$\begin{pmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \vdots \\ \hat{h}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} g_S \\ g_{NS} \\ g_M \\ g_F \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} \tag{6}$$

F-Statistic for Status x Genter Interaction

- $$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, A' = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \tag{7}$$

- $\boldsymbol{rank(A) = 4, rank(A') = 3}$, the column space of $\boldsymbol{A}$ contains the column space of $\boldsymbol{A'}$

- F-statistic for Status x Gender $f_{Status\hat{},Gender} = \frac{\vec{\hat{h}}^T(A(A^TA)^\dagger A^T - A'(A'^TA')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T(I-(A(A^TA)^\dagger A^T)\vec{\hat{h}}} * \frac{n-rank(A)}{rank(A)-rank(A')}$

The Two-Way ANOVA Process

- Compute F-statistic for Status x Gender & obtain the p-value
- If too low, then reject the additive (null) distribution, i.e., it is unlikely that the data can be supported by the means for the individual groups being a sum of the Status-wise and Gender-wise means
- Otherwise
    - Do the one way ANOVA process to see if the universal (null) model that assumes the same means for all Genders can be rejected
    - Separately, ditto for Status

Onw-Way Repeated Measures

- A single group of individuals measured repeatedly over time
- Does time make a systematic difference in measurement across multiple individuals?
    - E.g., measurements $0, 10, 20$ at timepoint 1 and $2, 12, 22$ at timepoint 2
    - The variation among individuals within each time point is large and the difference between the two timepoints pales in comparison to this variation; so conventional one-way ANOVA will not reject the universal (null) model
    - However, there is indeed a systematic effect of time: measurements increase by 2 for each individual
- Use Individual and Time as two factors
    - Use a Universal (on Time) model as the null (individuals have different underlying means but these do not change with time)
    - Compared to an additive model of Individual and Time effects

F-Statistic for One Way Repeated Measures

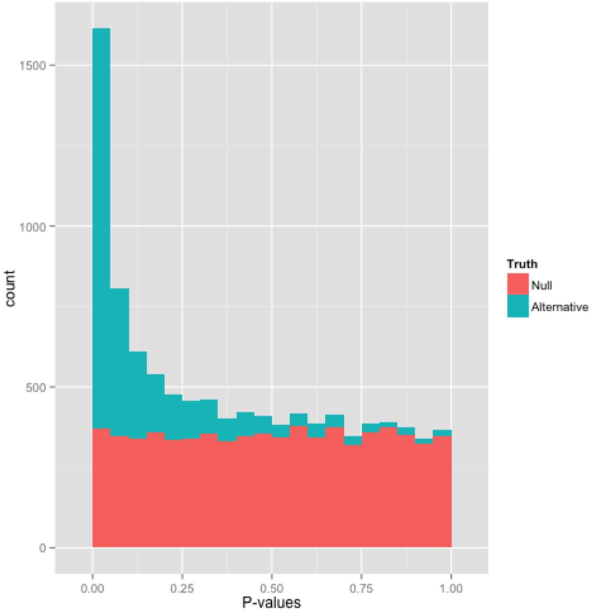
- $$A = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}, A' = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{8}$$
- The first 3 columns of A are individual effects, the last two are time
- A' has only individual effect columns
- $rank(A) = 4, rank(A') = \mathbf{3}$, the column space of A contains the column space of A'
- F-statistic $f_{\widehat{Status,Gender}} = \frac{\vec{\hat{h}}^T (A(A^TA)^\dagger A^T - A'(A'^TA')^\dagger A'^T)\vec{\hat{h}}}{\vec{\hat{h}}^T (I-(A(A^TA)^\dagger A^T)\vec{\hat{h}}} * \frac{n-rank(A)}{rank(A)-rank(A')}$

Correction for Multiple Testing

- Recall, for each gene
  - we independently derive a p-value under the null model
  - this p-value is the probability that the drawing samples for this gene from the null model yields as high an F-statistic as is obtained from the data at hand, thus c
  - note: this p-value has a uniform distribution between 0 and 1
- But, we have $n = 20,000$ genes
- Assume n_0 of these satisfy the null model (n_0 is the majority in practice)
- For these n_0 genes, assume p-values are drawn independently from a uniform distribuion (?)
- So the smallest of these will have an expected value of $\frac{1}{n_0}$, which could be as low as 0.00005!
- Which means that many of these n_0 genes will have the null model rejected, falsely; expected number $n_0 q$ if q is the cut-off.
- The probability that even one of these n_0 genes passes the cut-off is $1 - (1 - q)^{n_0} \;\; n_0 q$. Using q/n_0 instead of q as the cut-off ensures that this probability is less than q

Estimating n_0

- Note, we don't know n_0
- But p-values for these genes can be assumed to be independent and uniformly distributed in 0..1
- The other genes will have p-values biased towards 0
- Draw the histogram of all the p-values (use a suitable bin size, say 0.1)
- If you see bias closer to 0 (greater density near 0 than near 1), use the density closer to 1 to estimate n_0
- Otherwise, use n as a conservative estimate of n_0



link

False Discovery Rate (FDR)

- Sort all the p-values
- Suppose he ith smallest p-value is p_i
- The expected number of false positives from the n_0 genes with p-values smaller than x is $n_0 p_i$
- The fraction of false positives, roughly speaking, is expected to be $\frac{n_0 p_i}{i}$.
- Control this fraction at say cut-off q
- So pick the largest i such that $\frac{n_0 p_i}{i} \leq q$
- Or, in other words, pick the largest i such that $p_i < \frac{qi}{n_0}$
- Estimate n_0 as above
- Among all genes which pass this test, you can show that the expected false positive fraction is at most q

The Asignment

- The data file is here
- Your goal is to identify genes which respond differently to smoke in men vs women (Smoking Status x Gender model vs the Smoking Status + Gender null)
    - Use the above 2-way ANOVA framework to generate p-values for each row
    - Draw the histogram of p-values
    - See if a better (than n) estimate for n_0 is derivable from this histogram; justify your estimate
    - Use an FDR cut-off of 0.05 to shortlist rows
    - Create a shortlist of gene symbols from these rows
    - Intersect with the following gene lists: Xenobiotic metabolism, Free Radical Response, DNA Repair, Natural Killer Cell Cytotoxicity
    - Report intersection counts for each list, split into four groups; going down in women smokers vs non-smokers/going up in women smokers vs non-smokers x ditto for men