The Effects of Smoking

strand

PRECISION MEDICINE SOLUTIONS

FOR BETTER HEALTHCARE DECISIONS

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:
    a. 12 Male Non-smokers (106-117)
    b. 12 Male Smokers (118-129)
    c. 12 Female Non-Smokers (130-141)
    d. 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

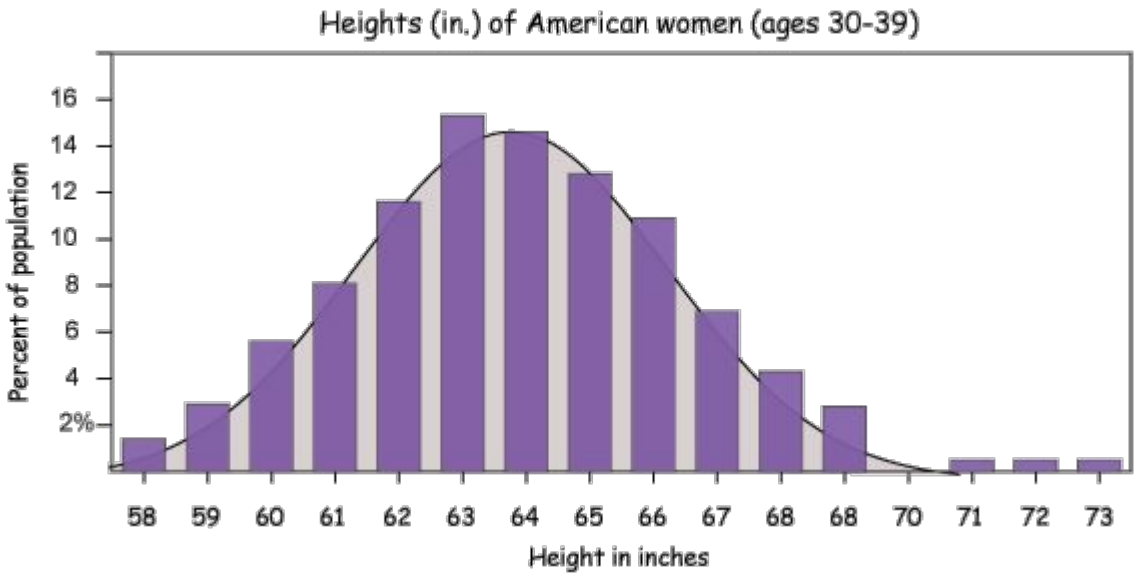
Which probes are different between the various groups?

A Simpler Setting: Heights

- Hypothesis: Male and female height distributions have the same mean
- Sample some males and some females (how many?)
- Compute the sample means
- If the sample means are different enough, then reject the hypothesis
- How different should the sample means be for us to reject the above hypothesis?
- If the hypothesis is true, what are the chances of rejecting it? [p-value]
- If the hypothesis is not true by a mile, what are the chances of not rejecting it

How do we answer the above questions?

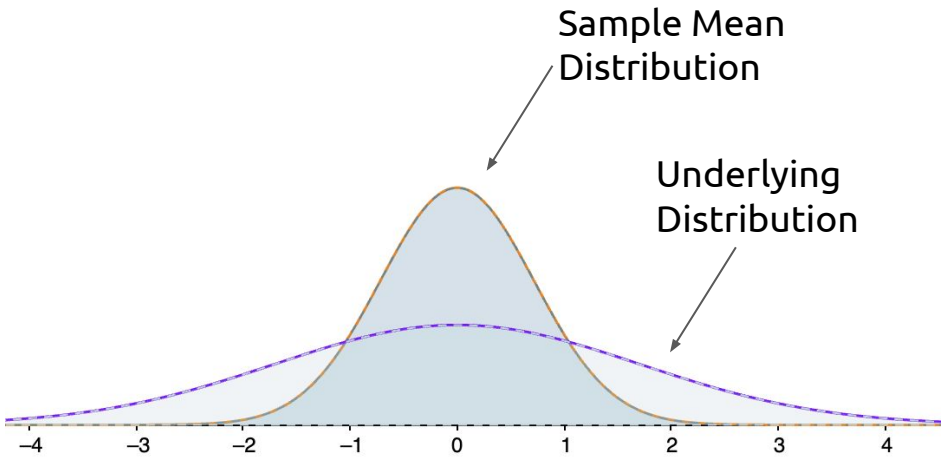
Distributions



Y axis is a percentage or a fraction between 0 and 1; total area under the curve is 1; also recall mean, median, mode

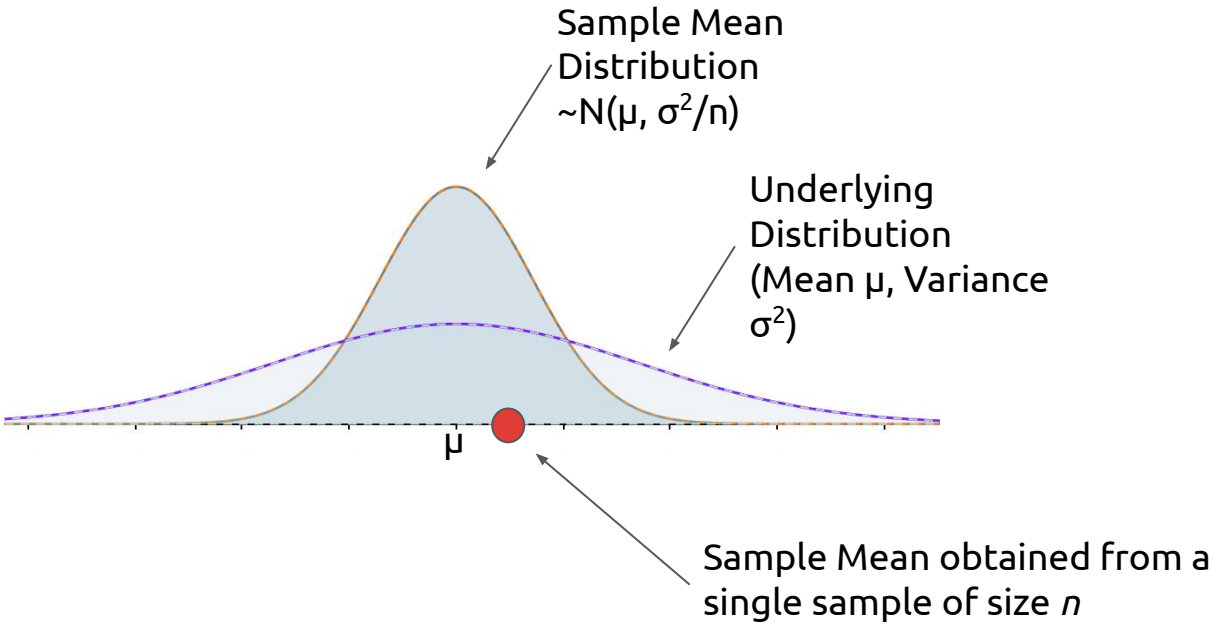
Sample Mean Distribution vs Underlying Distribution

- The *expected value* of the sample mean is equal to the population mean (mean of the underlying distribution)
- The *distribution* of the sample mean gets tighter and tighter around its expected value as *n* increases



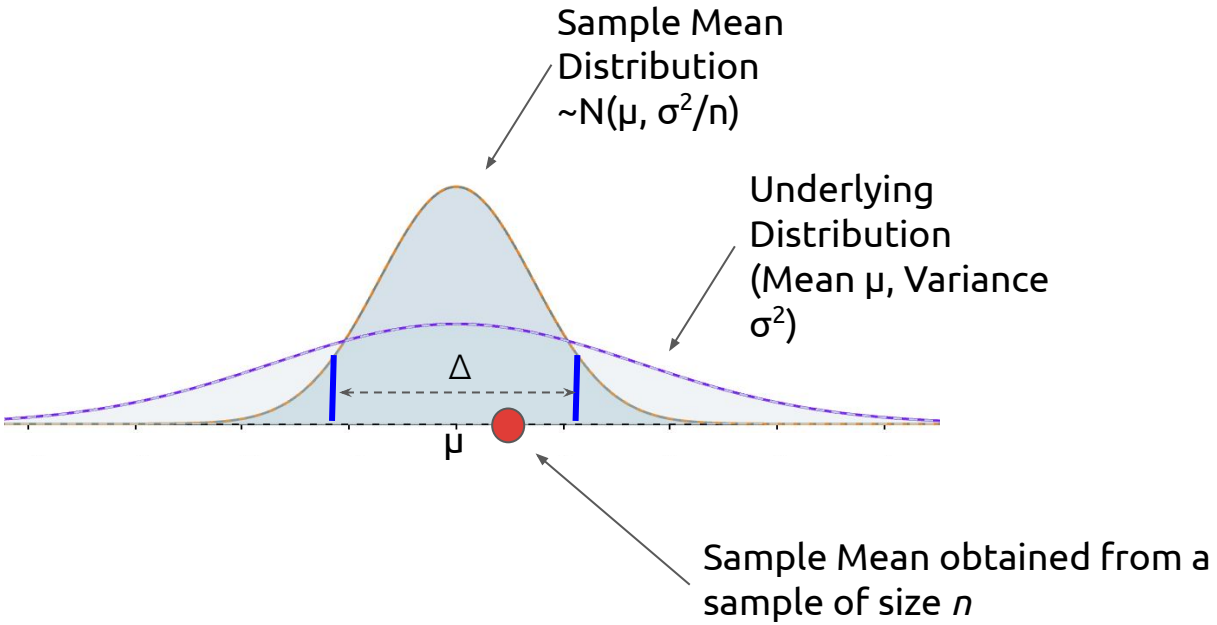
The distribution of the sample mean also gets closer and closer to Normal as *n* increases (n>30 suffices)

Sample Mean Distribution is ~Normal



For large enough n, the sample mean is a single draw from ~N(μ, σ²/n)

Estimating μ



Assuming large enough n, and assuming we know σ^2, we can calculate Δ (range between blue lines) such that the red dot will be in this range with prob say 95%.

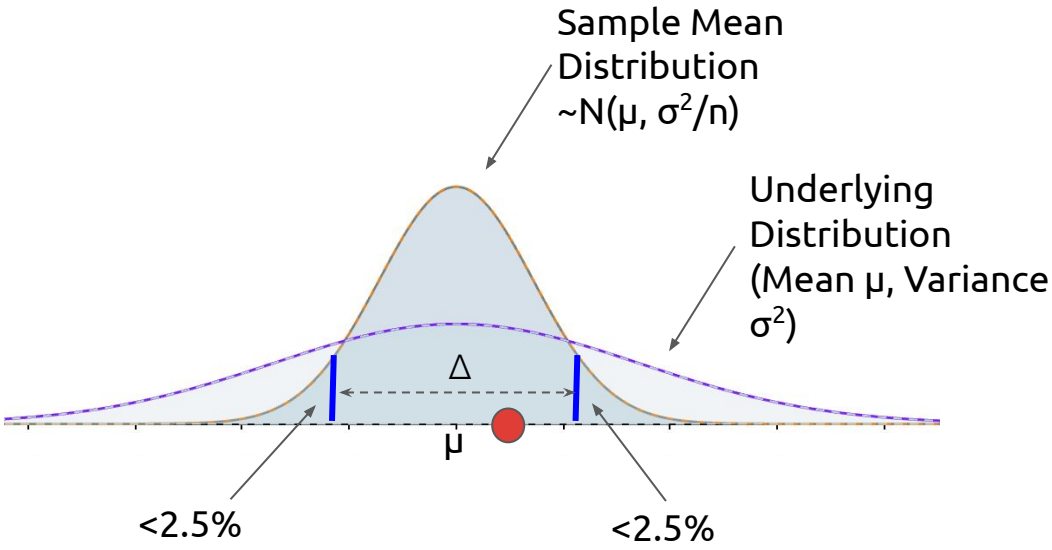
Estimating σ^2

- *Y* estimates σ^2
- Note the *n-1* and not *n;* this is necessary to show that *E(Y)*=σ^2
- However Y will have its own distribution, and a single sample from this distribution could well underestimate σ^2
- This distribution is more advanced (the various terms in the summation are not independent anymore)

$$Y = \frac{1}{n-1}\sum_i (X_i - \sum_j \frac{X_j}{n})^2$$

Use an overestimate for σ depending on background knowledge. E.g., could use 10ft for height!

Choosing n



With the overestimate for σ^2 , choose a Δ, say 0.1in, and pick *n* large enough so the probability mass of $N(\mu, \sigma^2/n)$ within the Δ range is >95%.

Coming Back to Heights

- Hypothesis: Male and female height distributions have the same mean
- Sample n males and n females, independently at random
- n is chosen large enough so that 95% of the probability mass of $N(0,v^2/n)$ is within ±0.1in
- And v^2 is chosen to safely overestimate σ^2 for both male and female distributions
- If |male sample mean minus the female sample mean| > 0.2in, then reject the hypothesis
- If the hypothesis is true, <10% chance it will be rejected
- If the hypothesis is false by a mile (say they are diff by 0.4in), <10% chance it won't be rejected

Can tweak the 10% downwards if needed using larger n. Note: no assumptions on the underlying distributions

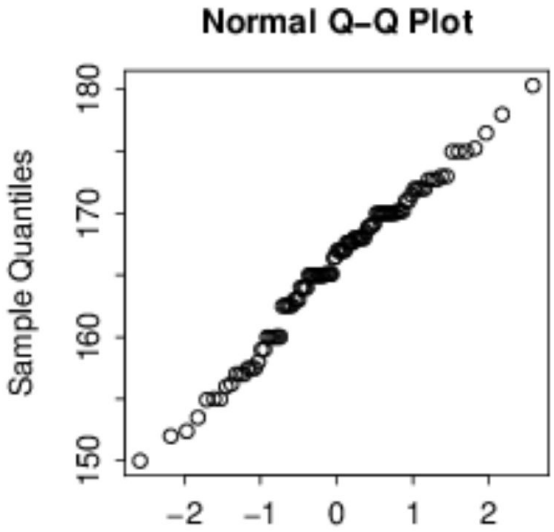
The Challenge

- n is limited by cost in many experimental situations
- What can you do with a given n?
- So the sample mean cannot be assumed to be normally distributed
- Cannot afford to liberally overestimate σ^2
- Need to better characterize sample mean and sample variance distributions
- Can be done if we make certain assumptions on the underlying distribution

Suppose we assume that the male and female underlying distributions are normally distributed and have the same/similar variances

Heights: Normality, Variance

- 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



Underlying distributions appear Normal, and variances in the two groups are “similar”

Even Simpler Setting

- Hypothesis: Underlying distribution has mean 0
- If $nx(\text{sample mean})^2/\text{sample variance} > \Delta$ then reject
- If hypothesis is indeed true, what is the probability of rejection?
- If hypothesis is not true by a mile, what is the probability of non-rejection?
- What is the distribution of $nx(\text{sample mean})^2/\text{sample variance}$?
- What is the threshold for not true by a mile?
- What is the Δ threshold?

Under the assumption that the underlying distribution is $N(\mu, \sigma^2)$, for unknown μ, σ^2; in fact, without loss of generality $\sigma^2=1$ (why?)

Some Facts & Observations

- If $X,Y \sim N(\mu, \sigma^2)$ and $N(\mu', \sigma'^2)$ resp and independent, $X+Y \sim N(u+\mu', \sigma^2+\sigma'^2)$
- If $X \sim N(\mu, \sigma^2)$, then $kX \sim N(ku, k^2\sigma^2)$
- If $X \sim N(0,1)$ is $X^2 \sim$ ChiSquare(1)
- If $X_1 \ldots X_k \sim N(0,1)$ and independent, $X^2_1 + \cdots + X^2_k \sim$ ChiSquare(k)
- Assuming the hypothesis on the previous slide and the underlying distribution being N(0, 1), what is the distribution of nx(sample mean)^2/sample variance?
- It is [ChiSquare(1)/1]/[ChiSquare(n-1)/(n-1)], where n is the sample size. This needs a proof
- The numerator and denominator are independent. This needs a proof

The ratio of independent ChiSquare(a)/a and ChiSquare(b)/b distributions is the well known F-Distribution, F(a,b); calculators are available easily

Back to the Even Simpler Setting

- Hypothesis: Underlying distribution has mean 0
- Sample n items from the distribution
- Compute the F-Statistic as on the right
- Use F(1,n-1) to find Δ such that prob that this F-Statistic>Δ < 5%
- If hypothesis is indeed true, what is the probability of rejection? 5%
- If hypothesis is not true by a mile, what is the probability of non-rejection? Quantification needs further assumptions, but qualitatively in the right direction

$$\frac{n * (\sum_i \frac{X_i}{n})^2}{\sum_i (X_i - \sum_j \frac{X_j}{n})^2/(n-1)}$$

For testing each of our 40K rows in our dataset as above, we will leave the last item as qualitative and only seek to minimize false positives but as tightly as possible

Proof

- (n-1) x Sample Variance= $(AX)^T AX$
- A is real symmetric, AA=A
- So diagonalize it as LDL^T, where $L^T L=I$, D is diagonal with values 0,1,1,...,1
- $(AX)^T AX=X^T A^T AX=X^T(LDL^T)^T(LDL^T)X=X^T LDL^T X$
- This is the sum of squares of projections of X on to all but the first of the basis vectors in the orthonormal basis formed by the columns of L

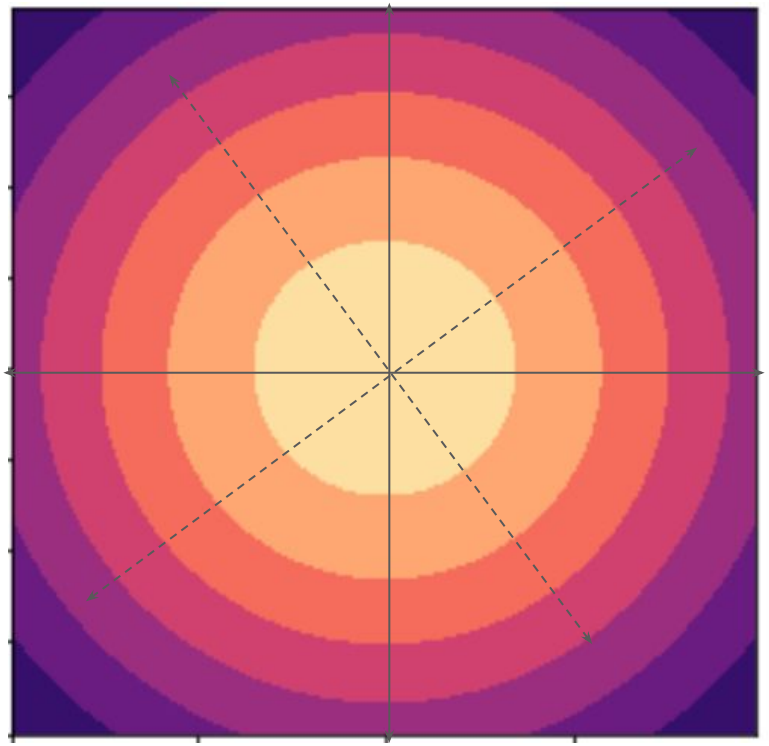
$$A = \begin{pmatrix} \begin{bmatrix} 1-\frac{1}{n} & \frac{-1}{n} & \cdots & \frac{-1}{n} \\ \frac{-1}{n} & 1-\frac{1}{n} & \cdots & \frac{-1}{n} \\ \vdots & \ddots & & \\ \frac{-1}{n} & \frac{-1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix} \end{pmatrix}$$

$$X = \begin{bmatrix} X_1 \\ X2 \\ \vdots \\ X_n \end{bmatrix}$$

Note, if D were I, then $X^T LIL^T X=X^T X \sim$ ChiSquare(n); but the first entry of D is 0; does that make the distribution of $X^T LDL^T X \sim$ ChiiSquare(n-1)?

The Magic

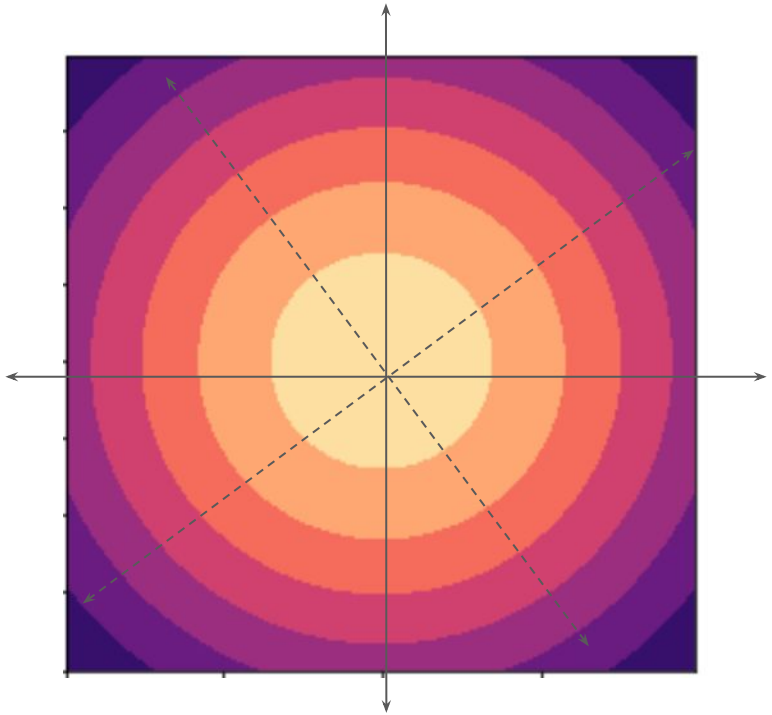
- The components of L^TX are all N(0,1)
- Are they independent?
- The distribution of $X^TX=X^TLL^TX$ is spherically symmetric
- This is of course immune to change of basis
- So the components of L^TX must be independent N(0,1)



The distribution of the sum of squares of projection of X on to all but the first of the basis vectors in the orthonormal basis L is thus ~ ChiSquare(n-1)

Independence of Numerator and Denominator

- Recall: the components of $L^T X$ must be independent N(0,1)
- The first basis vector in L is the all $1/\sqrt{n}$'s vector
- The projection of X on to this vector is $\sum X_i/\sqrt{n}$, the square of which is exactly nx(sample mean)^2



The distribution of the sum of squares of projection of X on to all but the first of the basis vectors in the orthonormal basis L is ChiSquare(n-1) and independent of nx(sample mean)^2

Rewriting the F-Statistic

- The F-Statistic can be rewritten as follows

$$\frac{n * (\sum_i \frac{X_i}{n})^2}{\sum_i (X_i - \sum_j \frac{X_j}{n})^2/(n-1)} = \frac{1}{1/(n-1)} * \left(\frac{\sum_i X_i^2}{\sum_i (X_i - \sum_j \frac{X_j}{n})^2} - 1 \right)$$

- Numerator: Sum of squares relative to our starting hypothesis (also called the null hypothesis)
- Denominator: Sum of squares if we don't assume the null hypothesis but instead plug in our best estimate for the underlying mean
- Denominator and numerator are both divided by the resp degrees of freedom (# of basis vectors over which projections are squared and added)

This will help generalize the F-Statistic to other situations

Back to Heights

- Two populations A and B, N(μ, σ^2) and N(μ', σ^2) resp (note: common variance)
- Null hypothesis: μ=μ'
- Sample n and m items resp from the two populations
- Compute

$$\frac{1}{1/(n+m-2)} * \left(\frac{\sum_{i=1}^{n+m}(X_i - \sum_{j=1}^{n+m}\frac{X_j}{n+m})^2}{\sum_{i=1}^{n}(X_i - \sum_{j=1}^{n}\frac{X_j}{n})^2 + \sum_{i=n+1}^{n+m}(X_i - \sum_{j=n+1}^{n+m}\frac{X_j}{m})^2} - 1 \right)$$

- Numerator: Sum of squares relative to best estimate of the common underlying mean (recall null hypothesis says means are the same)
- Denominator: Sum of squares relative to best estimates of groupwise means (disregarding the hypothesis)
- Distribution is F(1,n+m-2); Use it to find Δ such that prob>Δ < 95%, and reject the hypothesis accordingly

Easily generalizes to more than one group (say, male heights in TN, AP, Kar, Kerala)

Exercise

- Write the numerator and denominator in matrix form $(AX)^T AX$; show AA=I

$$\frac{1}{1/(n+m-2)} * \left(\frac{\sum_{i=1}^{n+m} (X_i - \sum_{j=1}^{n+m} \frac{X_j}{n+m})^2}{\sum_{i=1}^{n} (X_i - \sum_{j=1}^{n} \frac{X_j}{n})^2 + \sum_{i=n+1}^{n+m} (X_i - \sum_{j=n+1}^{n+m} \frac{X_j}{m})^2} - 1 \right)$$

- More advanced: Show that both have a common diagonalized basis; what are the basis vectors corresp to eigenvalue 0 for the two cases?
- Describe how this generalizes to multiple groups (say, male heights in TN, AP, Kar, Kerala), assuming the underlying group variances are all the same

How about more complex situations, like M/F and Smo/NSmo combined?

The 2D Situation

- What is the appropriate contrast so we can define the numerator and denominator accordingly?
- Null Hypothesis (numerator): The SmokingxGender interaction is purely additive, i.e., there exist numbers m,f,s,ns, such that the means of the four underlying distributions are m+s, m+ns, f+s, f+ns respectively
- The Alternative hypothesis (denominator): The SmokingxGender interaction is arbitrary, the 4 underlying distributions could have arbitrary means m_s, m_ns, f_s, f_ns respectively
- Assumption: the 4 distributions have same/similar variances

How do we make the get the best estimates for m,f,s,ns and for m_s, m_ns, f_s, f_ns?

Estimating m,f,s,ns using Linear Regression

- Find y so as to minimize $(X-Ny)^T(X-Ny)$
- The best y is $(N^TN)^\dagger N^TX$
- The sum of squares relative to this best y is $(X-N(N^TN)^\dagger N^TX)^T(X-N(N^TN)^\dagger N^TN)$
- This is $X^T(I-N(N^TN)^\dagger N^T)^T(I-N(N^TN)^\dagger N^T)X$
- Which in turn is $X^T(I-N(N^TN)^\dagger N^T)X$ (try proving this)

X N y

$$\begin{vmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{vmatrix} \sim \begin{vmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{vmatrix} \begin{vmatrix} m \\ f \\ s \\ ns \end{vmatrix}$$

Note that rank(N)=3
(or in general if there are a “genders” and b “smoking statuses”, a+b-1)

Estimating m_s, m_ns, f_s, f_n

- Find y so as to minimize $(X-Dy)^T(X-Dy)$
- The best y is $(D^TD)^\dagger D^TX$
- The sum of squares relative to this best y is $(X-D(D^TD)^\dagger D^TX)^T(X-D(D^TD)^\dagger D^TX)$
- This is $X^T(I-D(D^TD)^\dagger D^T)^T(I-D(D^TD)^\dagger D^T)X$
- Which in turn is $X^T(I-D(D^TD)^\dagger D^T)X$

X D y

$$\begin{vmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{vmatrix} \sim \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} m_s \\ m_n s \\ f_s \\ f_n s \end{vmatrix}$$

Note that rank(D)=4
(or in general if there are a “genders” and b “smoking statuses”, aXb)

The F-Statistic for the 2D CASE

$$\frac{1/(rank(D) - rank(N))}{1/(n - rank(D))} \times \left(\frac{X^T(I - N(N^TN)^\dagger N^T)X}{X^T(I - D(D^TD)^\dagger D^T)X} - 1 \right)$$

- Again the numerator is sum of squares of rank(D)-rank(N) independent N(0,1)'s
- Denominator is sum of squares of n-rank(D) independent N(0,1)'s
- Numerator and denominator are independent
- Proof is based on diagonalization with a common basis:
  - Any vector in the column space of D is an eigenvector of $D(D^TD)^\dagger D^T$, with associated eigenvalue 1, and any vector orthogonal to this column space is an eigenvector with eigenvalue 0
  - Similarly for N
  - The column space of N is a subspace of the column space of D
  - Hence the eigenspaces of $I-D(D^TD)^\dagger D^T$ and $D(D^TD)^\dagger D^T-N(N^TN)^\dagger N^T$ are disjoint

Use the F(rank(D)-rank(N), n-rank(D) distribution to determine the cut-off for rejection of the null hypothesis, so as to provide a guarantee of say <5% chance that the null hypothesis will be rejected even if it is true

Summary

- For each of the ~40K rows, in turn
- Compute the F-Stat (f)
- Compute the probability mass for the F(rank(D)-rank(N), n-rank(D)) distribution that lies to the right of f; this is called the p-value for this row
- Set a cut-off on the p-value, say 5%; in other words, chances of rejecting the null hypothesis for this row even if it were true is <5%
- Take the rows for which the null hypothesis is rejected; those are the interesting rows

How many false positives?

The Multiple Testing Problem

- Suppose all the rows satisfy the null hypothesis
- Assume data in the various rows are drawn independently of each other
- Then the p-values are uniformly distributed between 0 and 1
- So expected number of rows with p values below 5% will be 5%
- I.e., it is expected that 5% of all rows will pass as false positives
- Could make the p value cut off more stringent to say 100/n %
- How many false positives now? And why is this not a good solution?

Key point: Rows that do not satisfy the null hypothesis will have p-values distributed not uniformly but pushed closer to 0

Correction for Multiple Testing (FDR)

- Suppose there are bad rows, i.e., those that satisfy the null hypothesis, and good rows, i.e., those that don't
- Assume data in the various rows are drawn independently of each other
- At any given p-value cut off p, say you have k rows with p-value below p
- Expected number of bad rows amongst these k is p x n
- So (p x n)/k is an estimate of the number of false positives: the False Discovery Rate (FDR)

So pick the p-value cut off p to allow for the largest k with an acceptable FDR, say <0.05 or <0.1 or <0.2

Multiple Testing Correction

1. Use the p-value cut-offs more intelligently
2. E.g., if you use a cut-off of <1/n, then the expected number of false positives will be <1; but this may be too strict and will lead to lots of false negatives
3. E.g., if you use a cut-off of <0.05/n, then the prob there is even one false positive will be <0.05; but this is even stricter
4. An alternative approach is to control the expected fraction and not the expected number of false positives at 5% (False Discovery Rate)
    - Sort rows in increasing order of p-value
    - Identify the largest i such that $p_i < 0.05\ i/n$
    - Reject the null hypothesis for the first i rows

Note: the p-value for a row that satisfies the null hypothesis is uniformly distributed between 0 and 1. FDR assumes independence of rows. The expected number of false positives with p-value<=p_i is at most $np_i < 0.05i$.

THANKS!

ramesh@strandls.com