Lecture 31
Statistical description of data

- Measures of association based on entropy.
- Linear correlation.

Measures of association based on entropy:

Consider the game of "twenty questions" where by repeated yes/no questions you try to eliminate all but one correct possibility for an unknown. In fact, consider a generalization where multiple choice questions are also allowed, but that are mutually exclusive and exhaustive (like yes/no).

The value to you of an answer increases with the number of possibilities it eliminates. An answer that eliminates all but a fraction p of answers can be assigned a value $-\ln(p)$ (the – sign is so it is a positive number, the log so the numbers can be added, since an answer that eliminates 1/6 of possibilities is equivalent to two that eliminate 1/3 and ½ respectively.

So we determined the value of the answer. What is the value of the question? If there are I possible answers to the question (i=1,…,I) and the fraction of possibilities consistent with the i-th answer is p_i (with the sum of p_i's equal to one), then the value of the question is the expectation value of the answer, denoted H,

$$H = -\sum_{i=1}^{I} p_i \ln p_i$$

The value of H is between 0 and ln(I). It is zero only if one of the p_i's is one and the others zero. In that case the question is useless. It takes the maximum value when all the p_i's are equal. The value H is conventionally called the *entropy* of the question, a term borrowed from statistical physics.

So far we have said nothing about associations among variables. Say we are deciding which question to ask next and we have to choose between two candidates or possibly we want to ask both in a given order or another. Say one question has I possible answers, labeled by i, and the other J possible answers labeled by j. Then the possible outcomes of asking both questions form a contingency table Nij, and normalized with respect to the number of remaining possibilities N give all the information about the p's. To connect to the previously defined measured of correlations we identify,

$$p_{ij} = \frac{N_{ij}}{N}$$

$$p_{i\cdot} = \frac{N_{i\cdot}}{N} \qquad \text{(outcomes of question } x \text{ alone)}$$

$$p_{\cdot j} = \frac{N_{\cdot j}}{N} \qquad \text{(outcomes of question } y \text{ alone)}$$

The entropies of the questions x and y are, respectively,

$$H(x) = -\sum_i p_{i\cdot} \ln p_{i\cdot} \qquad H(y) = -\sum_j p_{\cdot j} \ln p_{\cdot j}$$

And together,

$$H(x,y) = -\sum_{i,j} p_{ij} \ln p_{ij}$$

We can now consider the entropy of the question y given x (that is, if x is asked first). It is what is called a conditional probability and it is denoted with a vertical bar and is defined in terms of the joint probability P(A and B) = P(A) P(B|A). So the entropy of y given x the expectation value over the answers to x of the entropy of the conditional probability of y given x,

$$H(y|x) = -\sum_i p_{i\cdot} \sum_j \frac{p_{ij}}{p_{i\cdot}} \ln \frac{p_{ij}}{p_{i\cdot}} = -\sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_{i\cdot}}$$

And similarly for H(x|y).

We now have everything we need to define a measure of the "dependency" of y on x, that is to say, a measure of association. It is sometimes called the uncertainty coefficient of y,

$$U(y|x) \equiv \frac{H(y) - H(y|x)}{H(y)}$$

This lies between zero (x and y have no association) and 1 (x entirely predicts y). For intermediate values it gives a notion of the fraction of entropy of y that is lost if x is already known (that is which is redundant in the information in x).

```
SUBROUTINE cntab2(nn,ni,nj,h,hx,hy,hygx,hxgy,uygx,uxgy,uxy)
INTEGER ni,nj,nn(ni,nj),MAXI,MAXJ
REAL h,hx,hxgy,hy,hygx,uxgy,uxy,uygx,TINY
PARAMETER (MAXI=100,MAXJ=100,TINY=1.e-30)
```

Given a two-dimensional contingency table in the form of an integer array `nn(i,j)`, where `i` labels the x variable and ranges from 1 to `ni`, `j` labels the y variable and ranges from 1 to `nj`, this routine returns the entropy `h` of the whole table, the entropy `hx` of the x distribution, the entropy `hy` of the y distribution, the entropy `hygx` of y given x, the entropy `hxgy` of x given y, the dependency `uygx` of y on x (eq. 14.4.15), the dependency `uxgy` of x on y (eq. 14.4.16), and the symmetrical dependency `uxy` (eq. 14.4.17).

Parameters: `MAXI` and `MAXJ` define the maximum size of table; `TINY` is a small number.

```
INTEGER i,j
REAL p,sum,sumi(MAXI),sumj(MAXJ)
sum=0
do 12 i=1,ni                                    Get the row totals.
    sumi(i)=0.0
    do 11 j=1,nj
        sumi(i)=sumi(i)+nn(i,j)
        sum=sum+nn(i,j)
    enddo 11
enddo 12
do 14 j=1,nj                                    Get the column totals.
    sumj(j)=0.
    do 13 i=1,ni
        sumj(j)=sumj(j)+nn(i,j)
    enddo 13
enddo 14
hx=0.                                           Entropy of the x distribution,
do 15 i=1,ni
    if(sumi(i).ne.0.)then
        p=sumi(i)/sum
        hx=hx-p*log(p)
    endif
enddo 15
hy=0.                                           and of the y distribution.
do 16 j=1,nj
    if(sumj(j).ne.0.)then
        p=sumj(j)/sum
        hy=hy-p*log(p)
    endif
enddo 16
```

Linear correlation:

When one has sets of data that are ordinal (not “how many events that satisfy condition x” but a variable y that can take values with certain probability) a common measure of correlation between two datasets is the linear correlation,

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

The regression coefficient r is 1 if the data are just proportional to each other (with a positive proportionality constant) and is -1 if they are proportional with a negative proportionality constant. A value near zero indicates the variables are uncorrelated.

If the distributions have enough convergent moments and N is large, then the distribution of r is normal with mean zero and standard deviation $1/N^{1/2}$ and the probability that two uncorrelated data sets yield a value of r is given by,

$$\text{erfc}\left(\frac{|r|\sqrt{N}}{\sqrt{2}}\right)$$

With erfc the complementary error function,

$$\text{erfc}(x) \equiv 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$$

This is pretty basic stuff. Statistics books try to go beyond this, but that inevitably requires making certain assumptions about the data. The most typical one is that the distributions of x and y jointly form a binormal or two dimensional Gaussian distribution around their mean values.

The joint probability density is given by,

$$p(x,y)\,dxdy = \text{const.} \times \exp\left[-\frac{1}{2}(a_{11}x^2 - 2a_{12}xy + a_{22}y^2)\right]\,dxdy$$

And the correlation coefficient is: $$r = -\frac{a_{12}}{\sqrt{a_{11}a_{22}}}$$

Provided your data complies with the assumptions (at least approximately) then one can do several things. First one can consider the common case in which the number of points N is not large. It turns out that the statistic,

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

In the case of no correlation is distributed as a Student t distribution with ν=N-2 degrees of freedom. As N becomes large this becomes asymptotically the erfc function and we recover the expression we had before.

Moreover, if N is only moderately large N>10, we can compare if the difference of two significantly nonzero r's (for instance coming from different experiments), is itself significant. That is, we can quantify if a change in some control variable significantly alters an existing correlation between two other variables. This is done through Fisher's z-transformation which associates each measured r with a variable z

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

Then z is approximately normally distributed with a mean value,

$$\overline{z} = \frac{1}{2} \left[\ln \left(\frac{1+r_{\text{true}}}{1-r_{\text{true}}} \right) + \frac{r_{\text{true}}}{N-1} \right]$$

Where r_{true} is the population value of the correlation coefficient, the r one would compute if one had all the possible values of x an y.

From this one can devise several significant tests. For instance, the significance level at which a measured value of r differs from some hypothesized value r_{true} is given by,

$$\mathrm{erfc}\left(\frac{|z-\overline{z}|\sqrt{N-3}}{\sqrt{2}}\right)$$

Similarly, the significance of a difference between two measured correlation coefficients r_1 and r_2 is,

$$\mathrm{erfc}\left(\frac{|z_1-z_2|}{\sqrt{2}\sqrt{\frac{1}{N_1-3}+\frac{1}{N_2-3}}}\right)$$

But one should always keep in mind that all this becomes meaningless if the statistics are far from Gaussian!

```
      SUBROUTINE pearsn(x,y,n,r,prob,z)
      INTEGER n
      REAL prob,r,z,x(n),y(n),TINY
      PARAMETER (TINY=1.e-20)                    Will regularize the unusual case of com-
C     USES betai                                     plete correlation.
        Given two arrays x(1:n) and y(1:n), this routine computes their correlation coefficient
        r (returned as r), the significance level at which the null hypothesis of zero correlation
        is disproved (prob whose small value indicates a significant correlation), and Fisher's z
        (returned as z), whose value can be used in further statistical tests as described above.
      INTEGER j
      REAL ax,ay,df,sxx,sxy,syy,t,xt,yt,betai
      ax=0.
      ay=0.
      do 11 j=1,n                                Find the means.
        ax=ax+x(j)
        ay=ay+y(j)
      enddo 11
      ax=ax/n
      ay=ay/n
      sxx=0.
      syy=0.
      sxy=0.
      do 12 j=1,n
        xt=x(j)-ax
        yt=y(j)-ay
        sxx=sxx+xt**2
        syy=syy+yt**2
        sxy=sxy+xt*yt
      enddo 12
      r=sxy/(sqrt(sxx*syy)+TINY)
      z=0.5*log(((1.+r)+TINY)/((1.-r)+TINY))
      df=n-2
      t=r*sqrt(df/(((1.-r)+TINY)*((1.+r)+TINY)))
      prob=betai(0.5*df,0.5,df/(df+t**2))
C     prob=erfcc(abs(z*sqrt(n-1.))/1.4142136)
      return
      END
```

Correlation coefficient

Fisher's z transformation

Student probability

Summary

- Ideas of entropy can be used to codify relations between variables.
- One can establish correlations between variables and there exist tests for significance.