STAT 621 Lecture Notes
Independence and Correlation

Next we will look at a selection of distribution-free methods for determining if and how two random variables are related to each other. We will first discuss two methods for estimating and testing the significance of the correlation between two numeric random variables. After that we will consider measuring and testing for association between two categorical variables.

Spearman's Rank Correlation

Let $(X_1, Y_1), \ldots (X_n, Y_n)$ be a sample of independent pairs of continuous random variables X and Y, with joint cdf F_{xy} and marginal cdfs F_x and F_y. Interest is in measuring the population correlation, often represented by ρ. First let's recall what the correlation tells us.

A standard parametric estimator of ρ was given by Pearson,

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

This is the default value computed with the R function `cor`. An interesting result is that if (X, Y) is bivariate normal,

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

follows a t-distribution with $n - 2$ degrees of freedom. A little discussion about this estimator.

Spearman's correlation statistic is a nonparametric version of r, where this sample correlation is computed on ranks instead of the actual data values. That is,

$$\widehat{\rho}_s = \frac{\sum(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum(R_i - \overline{R})^2 \sum(S_i - \overline{S})^2}}$$

where

- $R_i =$

- $\overline{R}=$

- $S_i=$

- $\overline{S}=$

One can show that this may be rewritten as

$$\widehat{\rho}_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)} \qquad \text{where} \qquad D_i = S_i - R_i$$

Example: The data below are measurements of collagen and proline in liver samples from 7 men suffering from cirrhosis.

Collagen (X)	7.0	7.1	7.2	8.3	9.4	10.5	11.4
Proline (Y)	2.8	2.9	2.7	2.6	3.5	4.6	5.0

Compute Spearman's correlation.

In R, you can compute $\widehat{\rho}_s$ with the `cor` function, specifying `method="spearman"`.

```
> collagen=c(7.0, 7.1 , 7.2 , 8.3,  9.4  ,10.5 , 11.4)
> proline=c(2.8 , 2.9 , 2.7 , 2.6 , 3.5 , 4.6 , 5.0)

> cor(collagen,proline,method="spearman")
     [1] 0.6785714
```

In the event of ties either within or between the two samples. it is recommended to replace ties with average ranks. R, and most other statistical software, follow this rule.

Hypothesis Test: A hypothesis test for the population correlation ρ based on Spearman's rank statistic tests the null

$$H_0 : \rho = 0$$

against any of the usual alternatives. The test statistic used by R is

$$S = \sum D_i^2.$$

So in terms of $\widehat{\rho}_s$, that would be what? Find S for the liver data.

It's a bit unclear how R derives the p-value for this test. The null distribution of S may be found exactly for small samples, likely using a randomization test. For larger samples, approximate methods are likely used.

Carry out the test in R with the `cor.test` function.

```
> cor.test(collagen, proline, method="spearman")

Spearman's rank correlation rho

data:  collagen and proline
S = 18, p-value = 0.1095
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6785714
```

Additional Notes: Ties, large sample approximation....

Kendall's τ

A second statistic that measures the correlation between two continuous random variables was proposed by Kendall. Again suppose that X and Y have joint cdf F_{xy} and marginal cdfs F_x and F_y. Kendall's population correlation coefficient is defined below. Discuss.

$$\tau = 2P[(Y_2 - Y_1)(X_2 - X_1) > 0] - 1$$

Kendall's correlation statistic (Kendall's-τ) is based on the numbers of *concordant* and *discordant* pairs. Consider any two pairs (X_i, Y_i) and (X_j, Y_j). How many such pairs are possible?

Two pairs are said to be concordant if either of the following are true.

a) $X_i < X_j$ and $Y_i < Y_j$

b) $X_i > X_j$ and $Y_i > Y_j$

They are said to be discordant pairs if either

a) $X_i < X_j$ and $Y_i > Y_j$

b) $X_i > X_j$ and $Y_i < Y_j$

If there are no ties, Kendall's τ is computed as the difference in the number of concordant and discordant pairs, divided by the number of possible pairs. Formally this can be written as

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i),\ (X_j, Y_j)]$$

where

$$Q[(a,b),\ (c,d)] = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0 \\ -1 & \text{if } (d-b)(c-a) < 0 \end{cases}$$

Some discussion.

Example: (Hollander, Wolfe and Chicken) A seafood marketing company studied factors contributing to quality in canned tuna. A measure of lightness was found on each of 9 lots of canned tuna. Average consumer preference score was also obtained on each lot.

Lightness	44.4	45.9	41.9	53.3	44.7	44.1	50.7	45.2	60.1
Score	2.6	3.1	2.5	5.0	3.6	4.0	5.2	2.8	3.8

I've filled in most of the values for Q in the table below. Find the remaining ones and compute τ.

$j \backslash i$	1	2	3	4	5	6	7	8
2	___							
3	1	___						
4	1	1	1					
5	1	___	1	1				
6	-1	-1	1	1	-1			
7	1	1	1	___	1	1		
8	1	1	1	1	-1	-1	1	
9	1	1	1	-1	1	-1	-1	1

Hypothesis Test: A hypothesis test for independence is based on Kendall's statistic. The null hypothesis is

$$H_0 : \tau = 0$$

What does it imply if H_0 is true?

The test typically uses the test statistic

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i),\ (X_j, Y_j)]$$

(although it seems that the test statistic reported by R is simply the number of *concordant* pairs). Here's the test in R with the tuna data. Here the alternative is $H_A : \tau > 0$.

```
lightness=c(44.4,45.9,41.9,53.3,44.7,44.1,50.7,45.2,60.1)
score=c(2.6,3.1,2.5,5.0,3.6,4.0,5.2,2.8,3.8)

cor(lightness, score, method="kendall")
   [1] 0.4444444

cor.test(lightness, score, method="kendall", alternative="greater")

       Kendall's rank correlation tau

     data:  lightness and score
     T = 26, p-value = 0.05972
     alternative hypothesis: true tau is greater than 0
     sample estimates:
           tau
      0.4444444
```

Ties and the τ_b Estimator: In the case of ties, Kendall's τ estimator is refined as follows. Let

$$Q[(a, b),\ (c, d)] = \begin{cases} 1 & \text{if } (d-b)(c-a) > 0 \\ 0 & \text{if } (d-b)(c-a) = 0 \\ -1 & \text{if } (d-b)(c-a) < 0 \end{cases}$$

Also let M be the number of pairs with no tied values. Let X_0 be the number of pairs tied only on the X-value and let Y_0 be the number of pairs tied only on the y-value. Then,

$$\tau_b = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i),\ (X_j, Y_j)]}{\sqrt{(M + X_0)(M + Y_0)}}$$

Some software programs will compute τ_b in the case of ties; it appears that R does not. But writing you own code wouldn't be too hard!