Opinionated Lessons in Statistics

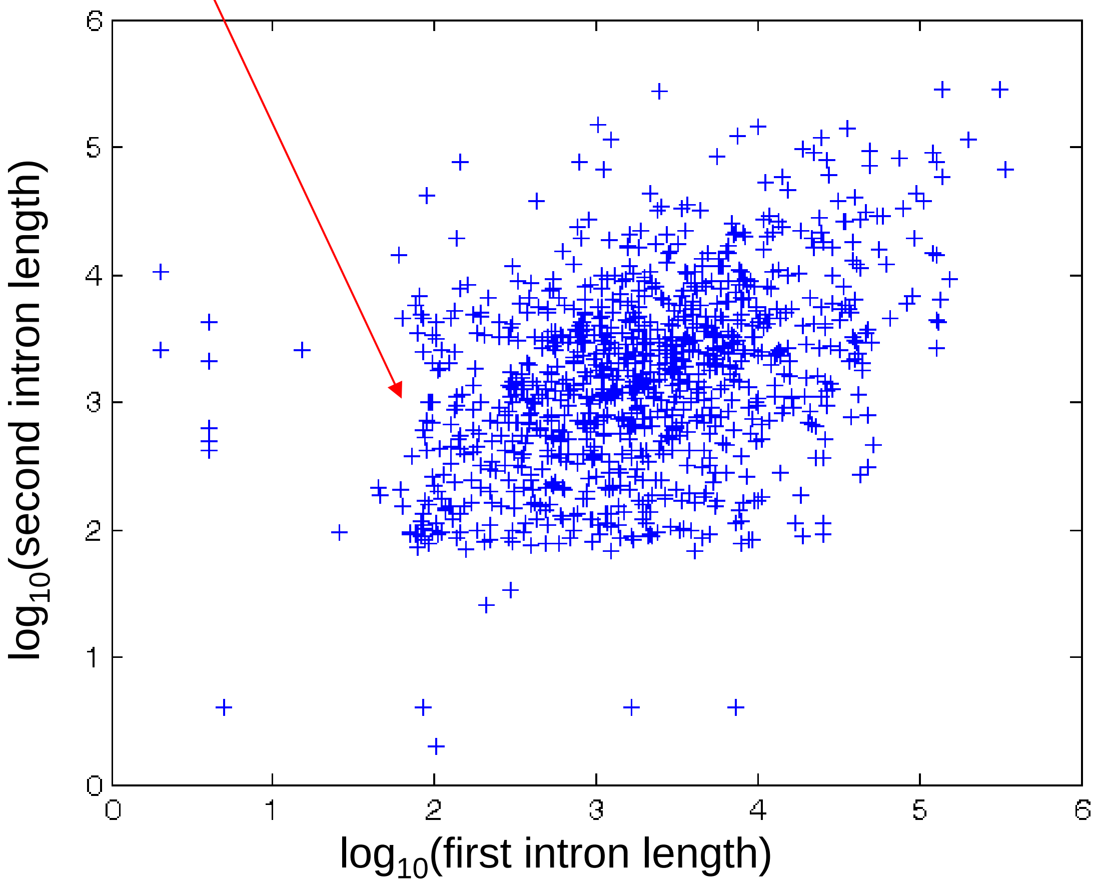
by Bill Press

#18 The Correlation Matrix

Log_{10} of size of 1st and 2nd introns for 1000 genes:

This is kind of fun, because it's not just the usual featureless scatter plot

notice the "hard edges"
this is biology!



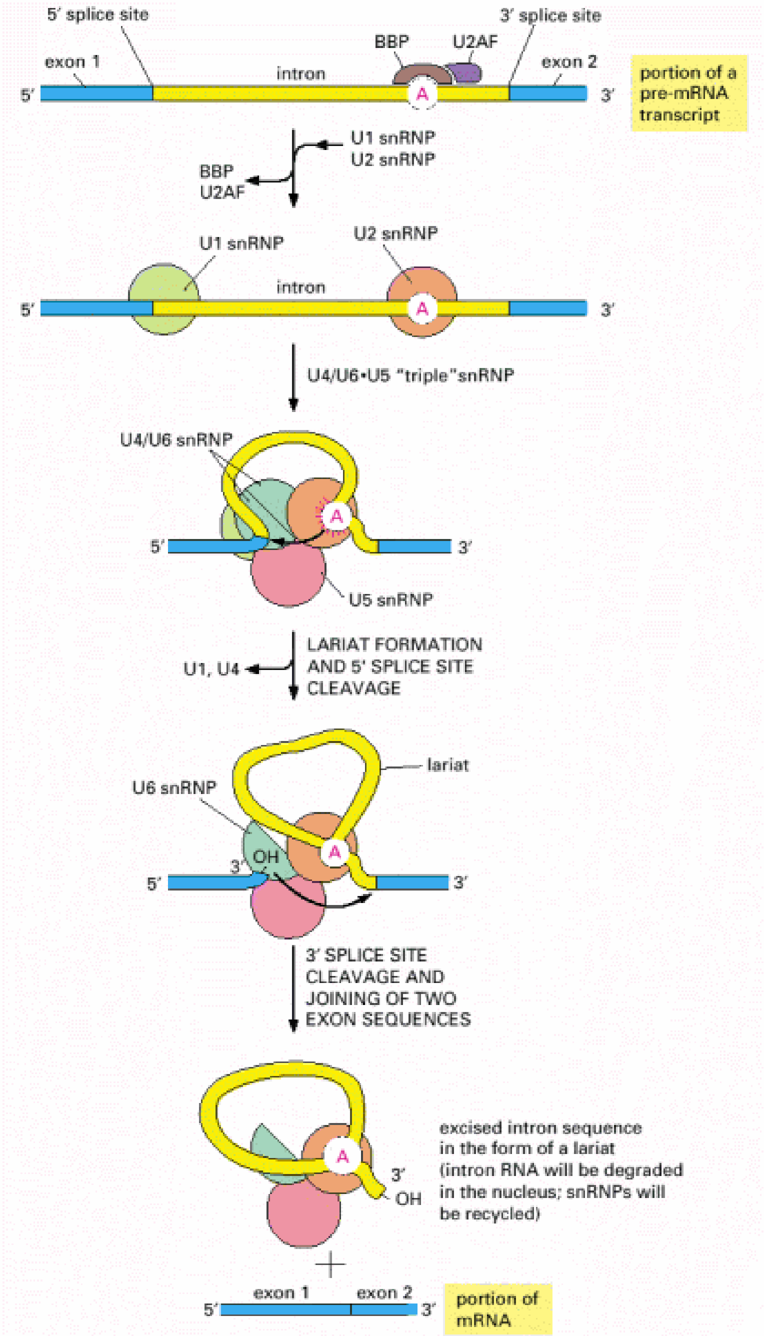
Is there a significant correlation here? If the first intron is long, does the second one also tend to be? Or is our eye being fooled by the non-Gaussian shape?

Biology:

The hard lower bounds on intron length are because the intron has to fit around the “big” spliceosome machinery!

It’s all carefully arranged to allow exons of any length, even quite small.

Why? Could the spliceosome have evolved to require a minimum exon length, too? Are we seeing chance early history, or selection?


5′ splice site
3′ splice site
BBP
U2AF
exon 1
intron
exon 2
5′
A
3′
portion of a pre-mRNA transcript
U1 snRNP
U2 snRNP
BBP
U2AF
U1 snRNP
U2 snRNP
intron
5′
A
3′
U4/U6•U5 "triple"snRNP
U4/U6 snRNP
5′
A
3′
U5 snRNP
U1, U4
LARIAT FORMATION AND 5′ SPLICE SITE CLEAVAGE
U6 snRNP
lariat
3′
OH
A
5′
3′
3′ SPLICE SITE CLEAVAGE AND JOINING OF TWO EXON SEQUENCES
A
3′
OH
excised intron sequence in the form of a lariat (intron RNA will be degraded in the nucleus; snRNPs will be recycled)
+
exon 1
exon 2
5′
3′
portion of mRNA

credit: Alberts et al.
Molecular Biology of the Cell

The covariance matrix is a more general idea than just for multivariate Normal.
You can compute the covariances of any set of random variables.
It's the generalizaton to M-dimensions of the (centered) second moment Var.

$$\mathrm{Cov}\,(x, y) = \langle (x - \overline{x})(y - \overline{y}) \rangle$$

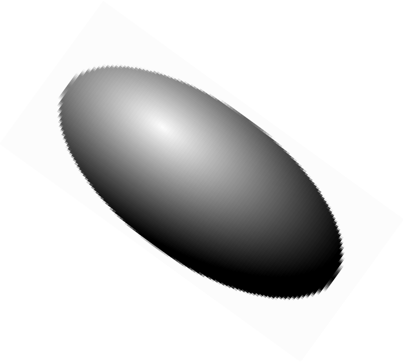
For multiple r.v.'s, all the possible covariances form a (symmetric) matrix:

$$\mathbf{C} = C_{ij} = \mathrm{Cov}\,(x_i, x_j) = \langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \rangle$$

Notice that the diagonal elements are the variances of the individual variables.

The variance of any linear combination of r.v.'s is a quadratic form in **C** :

$$\begin{aligned} \mathrm{Var}\,\left(\sum \alpha_i x_i\right) &= \left\langle \sum_i \alpha_i (x_i - \overline{x_i}) \sum_j \alpha_j (x_j - \overline{x_j}) \right\rangle \\ &= \sum_{ij} \alpha_i \left\langle (x_i - \overline{x_i})(x_j - \overline{x_j}) \right\rangle \alpha_j \\ &= \boldsymbol{\alpha}^T \mathbf{C}\, \boldsymbol{\alpha} \end{aligned}$$



This also shows that **C** is positive definite, so it can still be visualized as an ellipsoid in the space of the r.v.'s., where the directions are the different linear combinations.

The covariance matrix is closely related to the linear correlation matrix.

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$$

more often seen written out as

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

When the null hypothesis is that X and Y are independent r.v.'s, then r is useful as a p-value statistic ("test for correlation"), because

1. For large numbers of data points *N*, it is normally distributed,

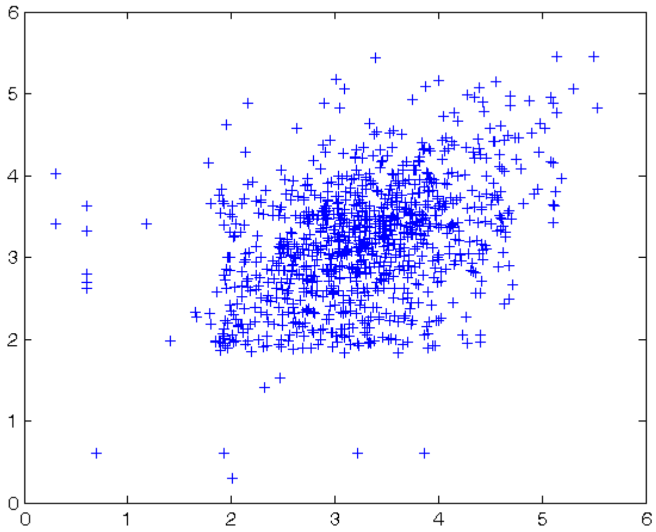
   $r \sim \mathrm{N}(0, N^{-1/2})$

   so $r\sqrt{N}$ is a normal t-value

2. With small numbers of data points, if the underlying distribution is multivariate normal, there is a simple form for the p-value (comes from a Student t distribution).

3. If you substitute ranks for values, there is a universal distribution related to Student t. This is Spearman correlation.

For the exon length data, we can easily now show that the correlation is highly significant.



```
r = sig ./ sqrt(diag(sig) * diag(sig)')
tval = sqrt(numel(len1))*r
r =
    1.0000    0.3843
    0.3843    1.0000
tval =
   31.6228   12.1511
   12.1511   31.6228
```

statistical significance of the correlation in standard deviations (but note: uses CLT)

```
[rr p] = corrcoef(i1llen,i2llen)
rr =
    1.0000    0.3843
    0.3843    1.0000
p =
    1.0000    0.0000
    0.0000    1.0000
```

Matlab has built-ins

not clear why Matlab reports 1 on the diagonals. I’d call it 0!