CS395T
Computational Statistics with Application to Bioinformatics

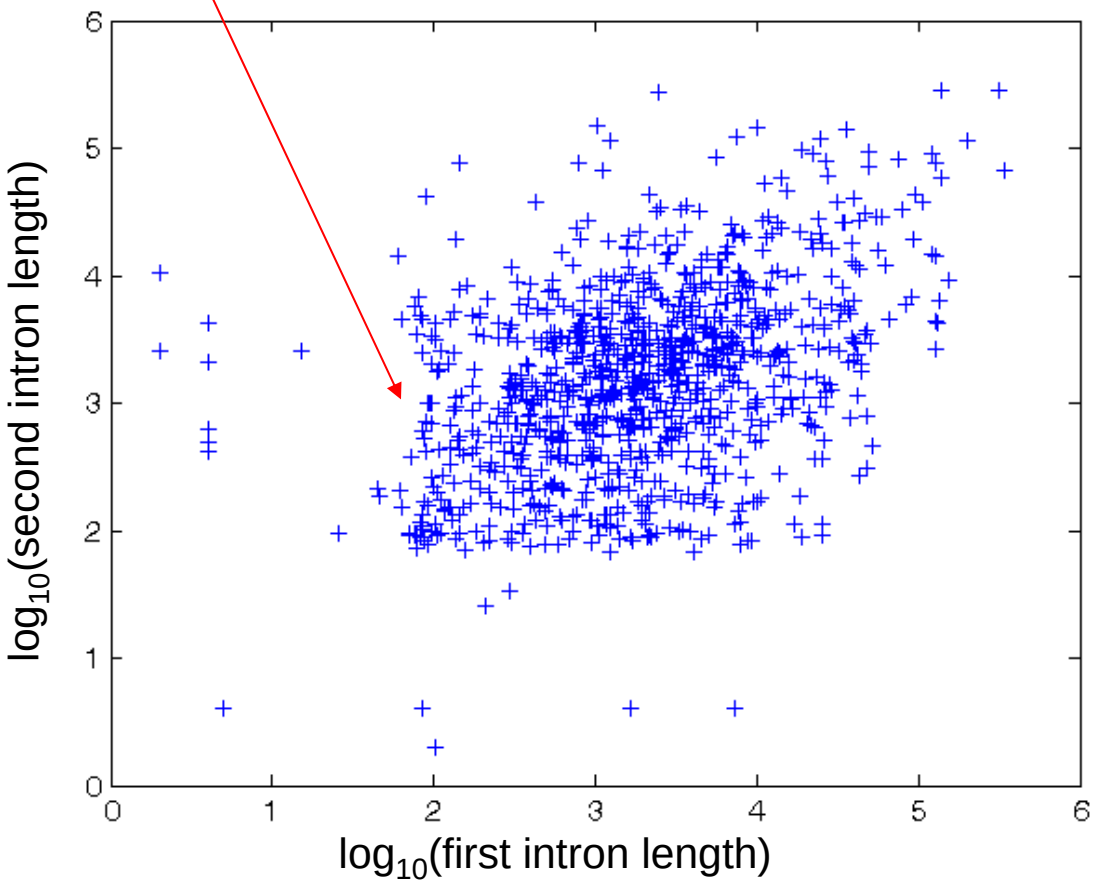
Prof. William H. Press
Spring Term, 2011
The University of Texas at Austin

Lecture 10

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!



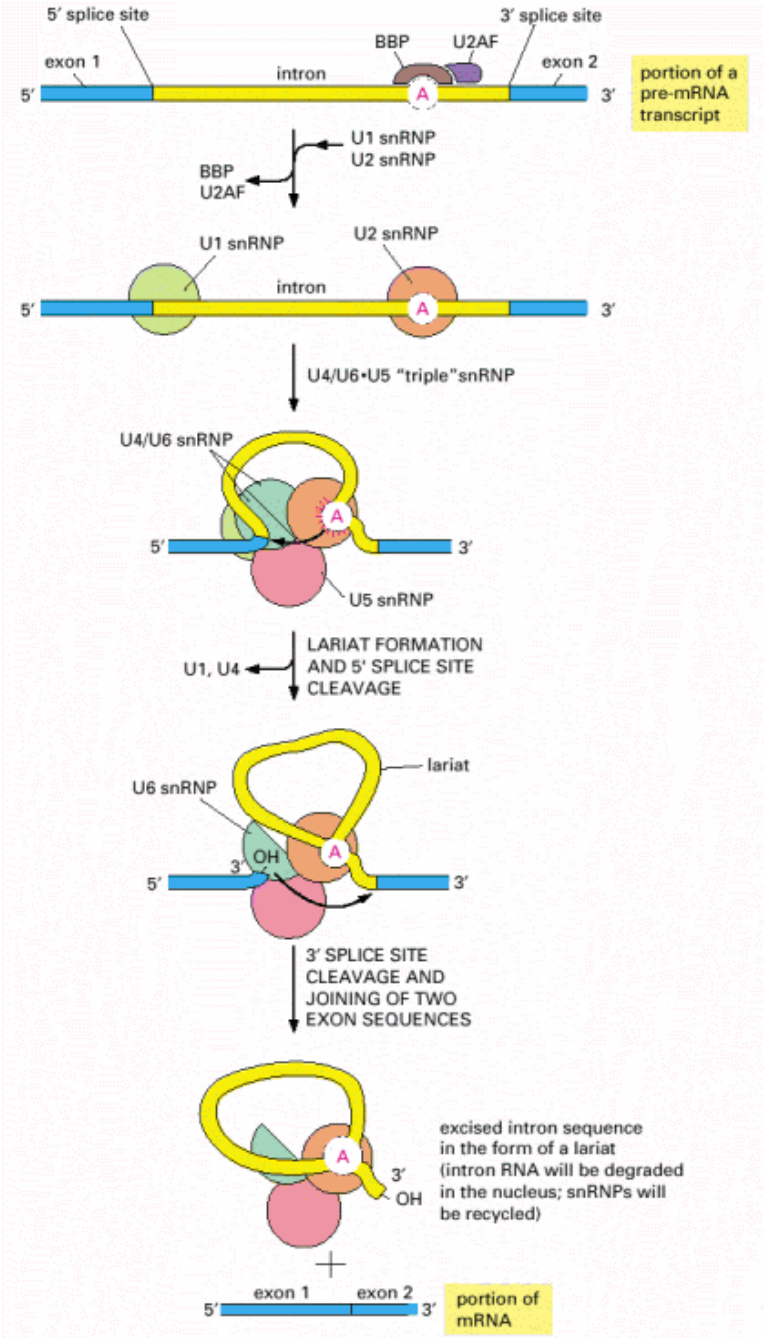
Is there a significant correlation here? If the first intron is long, does the second one also tend to be? Or is our eye being fooled by the non-Gaussian shape?

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?



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 $\mathbf{C}$:

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

This also shows that $\mathbf{C}$ 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

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

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,

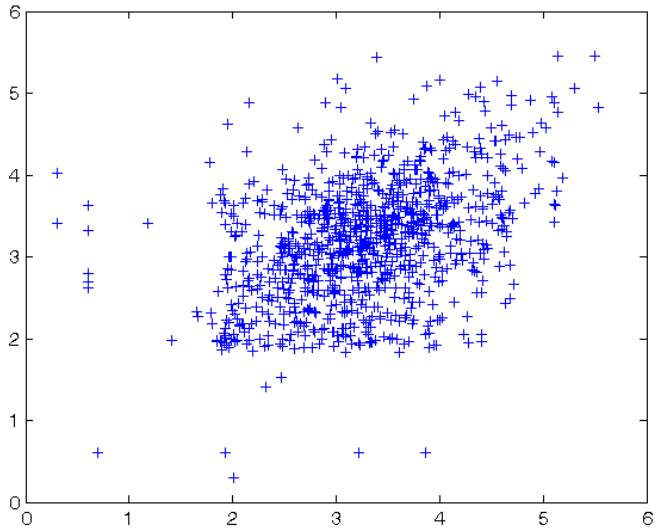
   $r \sim \mathrm{N}(0, N^{-1/2})$

   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!

Let's talk more about **chi-square**.
Recall that a t-value is (by definition) a deviate from $\mathrm{N}(0,1)$

χ^2 is a "statistic" defined as the sum of the squares of n independent t-values.

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

$\mathrm{Chisquare}(\nu)$ is a distribution (special case of Gamma), defined as

$$\chi^2 \sim \mathrm{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)} (\chi^2)^{\frac{1}{2}\nu - 1} \exp\left(-\tfrac{1}{2}\chi^2\right) d\chi^2, \qquad \chi^2 > 0$$

The important theorem is that $\chi 2$ is in fact distributed as Chisquare.

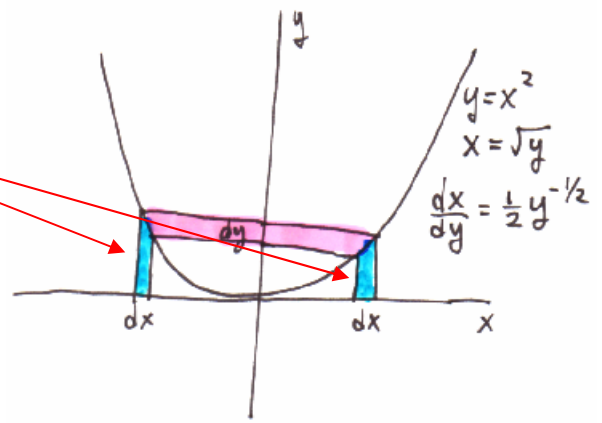
Let's prove it.

Prove first the case of ν=1:

Suppose $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \Rightarrow \quad x \sim \mathrm{N}(0,1)$

and $y = x^2$

$p_Y(y)\,dy = 2p_X(x)\,dx$



$$\text{So, } p_Y(y) = y^{-1/2} p_X(y^{1/2}) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$$

$$\sim \mathrm{Chisquare}(1)$$

To prove the general case for integer ν, compute the characteristic function

$$\chi^2 \sim \text{Chisquare}(\nu), \qquad \nu > 0$$

$$p(\chi^2)d\chi^2 = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu)}(\chi^2)^{\frac{1}{2}\nu-1}\exp\left(-\tfrac{1}{2}\chi^2\right)d\chi^2, \qquad \chi^2 > 0$$

```
In[9]:= pchi2 = (1 / (2^(nu / 2) Gamma[nu / 2])) y^(nu / 2 - 1) Exp[-y / 2]
```

Out[9]=

$$\frac{2^{-\text{nu}/2}\, e^{-y/2}\, y^{-1+\frac{\text{nu}}{2}}}{\text{Gamma}\left[\frac{\text{nu}}{2}\right]}$$

```
In[10]:= Integrate[pchi2, {y, 0, Infinity}, GenerateConditions → False]
```

Out[10]=

```
1
```

```
In[11]:= Integrate[pchi2 Exp[I t y], {y, 0, Infinity},
          GenerateConditions → False]
```

Out[11]=

```
(1 - 2 i t)^-nu/2
```

Since we already proved that ν=1 is the distribution of a single t^2-value, this proves that the general ν case is the sum of ν t^2-values.

Question: What is the generalization of

$$\chi^2 = \sum_i \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \qquad x_i \sim \mathrm{N}(\mu_i, \sigma_i)$$

to the case where the x_i's are normal, **but not independent**?
I.e., **x** comes from a multivariate Normal distribution?

Answer:

$$\chi^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \qquad \mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Proof is one of those Cholesky things,

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T, \quad \mathbf{L}\mathbf{y} = \mathbf{x} - \boldsymbol{\mu},$$

show that $\mathbf{y}$ is product of independent N(0,1)'s, as we did before, and that

$$\chi^2 = \mathbf{y}^T \mathbf{y} = \sum y_i^2$$

Weighted Nonlinear Least Squares Fitting
a.k.a. χ^2 Fitting
a.k.a. Maximum Likelihood Estimation of Parameters (MLE)
a.k.a. Bayesian parameter estimation
(with uniform prior and maybe some other normality assumptions)

these are not all exactly identical, but they're very close!

returned by Google for image search on "least squares fitting"!

$$y_i = y(\mathbf{x}_i|\mathbf{b}) + e_i$$

measured values supposed to be a model, plus an error term

$$e_i \sim N(0, \sigma_i)$$

the errors are Normal, either independently…

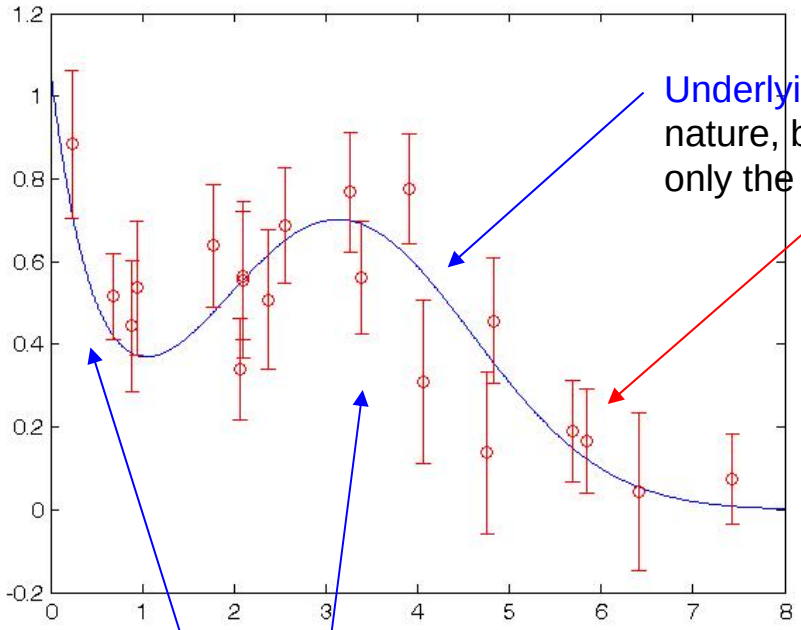
$$\mathbf{e} \sim N(0, \mathbf{\Sigma})$$

… or else with errors correlated in some known way (e.g., multivariate Normal)

We want to find the parameters of the model **b** from the data.

An example might be something like fitting a known functional form to data

$$f(x) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$



Underlying curve is known to nature, but not to us! We see only the red data points.

Fit 5 parameters from 20 irregularly space points, with normal errors of known standard deviations.

Can we do it? How well?

for example, this rise might be an instrumental or noise effect, while this bump might be what you are really interested in

Fitting is usually presented in frequentist, MLE language.
But one can equally well think of it as Bayesian:

$$
\begin{aligned}
P(\mathbf{b}|\{y_i\}) &\propto P(\{y_i\}|\mathbf{b})P(\mathbf{b}) \\
&\propto \prod_i \exp\left[-\tfrac{1}{2}\left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b}) \\
&\propto \exp\left[-\tfrac{1}{2}\sum_i \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i}\right)^2\right] P(\mathbf{b}) \\
&\propto \exp[-\tfrac{1}{2}\chi^2(\mathbf{b})]P(\mathbf{b})
\end{aligned}
$$

Now the idea is: Find (somehow!) the parameter value $\mathbf{b}_0$ that minimizes χ^2 .

For linear models, you can solve linear "normal equations" or, better, use Singular Value Decomposition. See NR3 section 15.4

In the general nonlinear case, you have a general minimization problem, for which there are various algorithms, none perfect.

Those parameters are the MLE. (So it is Bayes with uniform prior.)

By the way, minimum finding in general is as hard as any computationally hard problem!

For example factoring integers:

$$C = ab$$

$$(a,b) = \underset{a,b}{\text{minarg}} \left[(C-ab)^2 + \sin^2(\pi a) + \sin^2(\pi b)\right]$$

has a single global minimum of 0 if C is product of two primes,
multiple minima of 0 if C is product of more than two primes,
global minimum > 0 if C is prime

So, in real life, global minimum finding is only as good as your ability to guess a starting value in the "basin of convergence" of the minimum. Different numerical methods have better or worse basins of convergence.

Methods specialized to $\chi 2$ fitting are often <u>much</u> better for $\chi 2$ problems than general methods.