3.051J/20.340J

Lecture 16
Statistical Analysis in Biomaterials Research (Part II)

C. F Distribution

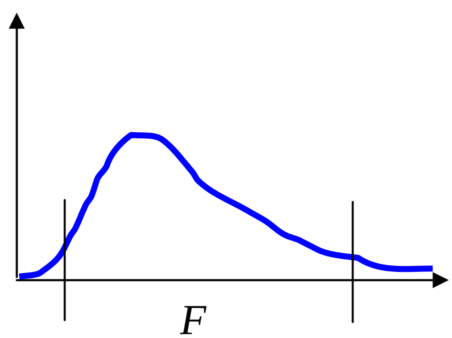
- Allows comparison of variability of behavior between populations using test of hypothesis: $\sigma_x = \sigma_{x'}$

Named for British statistician Sir Ronald A. Fisher.

Define a statistic:

$$\chi_{\nu_1}^{\ 2} = \nu_1 S^2 / \sigma^2$$ where ν_1 is degrees of freedom.

then $$F = \frac{\chi_{\nu_1}^{\ 2} / \nu_1}{\chi_{\nu_2}^{\ 2} / \nu_2}$$



For $\sigma_x = \sigma_{x'}$ $\Rightarrow$ $$F = \frac{S_x^2}{S_{x'}^2}$$

Procedure to test variability hypothesis:

1. Calculate S_x^2 and $S_{x'}^2$ (with $\nu_1 = N–1$ and $\nu_2 = N'-1$, respectively)

2. Compute F

3. Look in F-distribution tables for critical F for ν_1, ν_2, and desired confidence level P

4. For $$F_{\frac{1-P}{2}} < F < F_{\frac{1+P}{2}} \Rightarrow \sigma_x = \sigma_{x'}$$

Case Example: Measurements of C5a production for blood exposure to an extracorpeal filtration device and tubing gave same means, but different variabilities. Are the standard deviations different within 95% confidence?

Control (tubing only): $S_{x'}^2 = 26\ (\mu g/ml)^2$, $\nu_2 = 9$

Filtration device: $S_x^2 = 32\ (\mu g/ml)^2$, $\nu_1 = 7$

1. Calculate S_x^2 and $S_{x'}^2$ (provided)

2. Compute F

$$F = \frac{S_x^2}{S_{x'}^2} = 32/26 = 1.231$$

3. Determine critical F values from F-distribution chart

$\nu_1 = 7$ and $\nu_2 = 9$ (m, n for use with tables)

$$\frac{1-P}{2} = 0.025 \quad \Rightarrow \quad F_{0.025} = 0.207$$

$$\frac{1+P}{2} = 0.975 \quad \Rightarrow \quad F_{0.975} = 4.20$$

For $0.207 \le F \le 4.20 \Rightarrow \sigma_x = \sigma_{x'}$

F=1.231 falls within this interval.

Conclude σ values for two systems are the same!

D. Other distributions of interest

- Radioactive decay$\Rightarrow$ *Poisson* distribution
- Only 2 possible outcomes$\Rightarrow$ *Binomial* distribution

3. Standard deviations of computed values

- If quantity *z* of interest is *a function* of measured parameters

z = f(x,y,...)

What is S_z? We assume: $\langle z \rangle = f\left(\langle x \rangle, \langle y \rangle, \ldots\right)$

Deviations (δz) of *z* from its universal value can be written as:

$$\partial z = \frac{\partial z}{\partial x}\partial x + \frac{\partial z}{\partial y}\partial y + \ldots$$

The standard deviation for z is calculated:

$$S_z = \sqrt{\left(\frac{\partial \langle z \rangle}{\partial \langle x \rangle}\right)^2 S_x^2 + \left(\frac{\partial \langle z \rangle}{\partial \langle y \rangle}\right)^2 S_y^2 + \ldots}$$

Case Example: We measure motility (μ) and persistence (*P*) of a cell and want to know the standard deviation of the speed:

$$\langle \mu \rangle = \frac{\langle S \rangle^2 \langle P \rangle}{2} \quad \text{rearranges to} \quad \langle S \rangle = \sqrt{\frac{2\langle \mu \rangle}{\langle P \rangle}}$$

$$\frac{\partial\langle S\rangle}{\partial\langle P\rangle}=-0.5\sqrt{2\langle\mu\rangle}\langle P\rangle^{-3/2}=\frac{-\langle S\rangle}{2\langle P\rangle}\qquad\qquad \frac{\partial\langle S\rangle}{\partial\langle\mu\rangle}=0.5\frac{\sqrt{2}}{\langle P\rangle\langle\mu\rangle}=\frac{\langle S\rangle}{2\langle\mu\rangle}$$

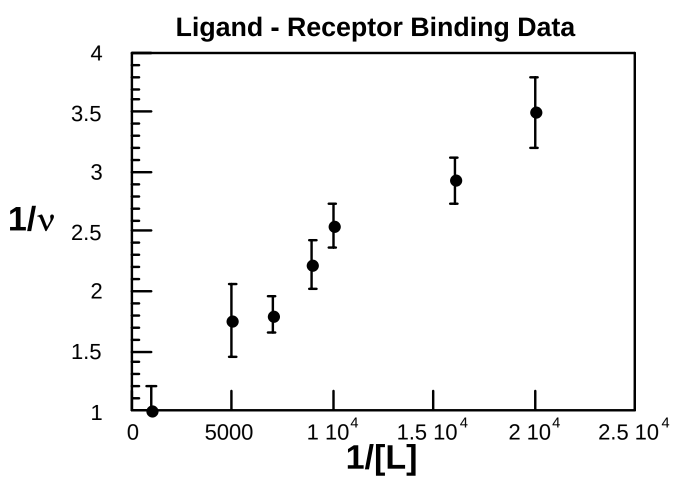
$$S_S=\sqrt{\left(\frac{\partial\langle S\rangle}{\partial\langle P\rangle}\right)^2 S_P^2+\left(\frac{\partial\langle S\rangle}{\partial\langle\mu\rangle}\right)^2 S_\mu^2}==\sqrt{\frac{\langle S\rangle^2}{4\langle P\rangle^2}S_P^2+\frac{\langle S\rangle^2}{4\langle\mu\rangle^2}S_\mu^2}$$

4. Least Squares Analysis of Data (Linear Regression)

- Computing the straight line that best fits data.

Suppose we have some measured data for binding of a ligand to its receptor:

L + R = C



This equilibrium is described by: K = [C]/[L][R]

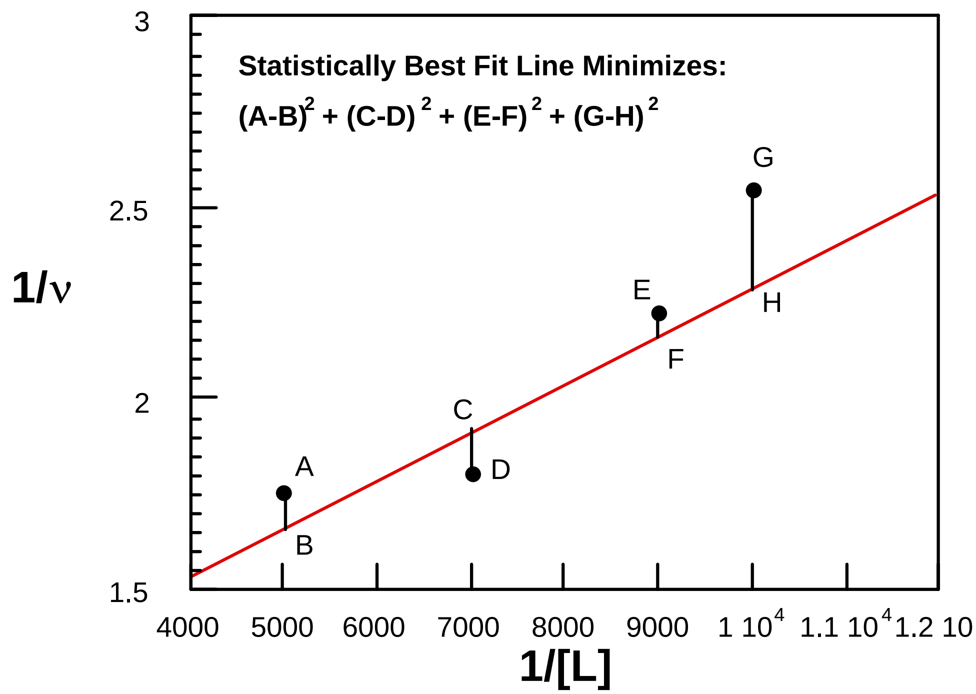
ν = fraction of occupied receptors = [C]/([C] + [R]) = K[L]/(1 + K[L])

1/ν = 1 + 1/K[L]

Question: How can we numerically obtain the linear equation that best represents the data?

Answer: Minimize the *squared deviation* of the line from each point.

NOTE: This is a generic tool in data regression, independent of the fitting function.



The deviation of any given measured point (x_i, y_i) from the line is:

$$y_i - y_{line} = y_i - (mx_i + b)$$

Where *m* and *b* are the slope and intercept of the line.

Our minimization criterion can thus be written:

$$M = \sum_{i=1}^{N} [y_i - (mx_i + b)]^2 = \text{minimum}$$

Mathematically we require: $\dfrac{\partial M}{\partial m} = 0, \dfrac{\partial M}{\partial b} = 0$

Solving these two equations for the two unknowns (best fit m and b for the line), we get:

$$m = \frac{N\sum_{i=1}^{N}(x_i y_i) - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2} \qquad b = \frac{\sum_{i=1}^{N} y_i - m\sum_{i=1}^{N} x_i}{N}$$

➢ Quantifying Error of the Straight-Line Fit

If the error on each y_i is unknown (e.g., a single measurement was made):

The *standard deviation for the regression line* is given by:

$$\sigma = \sqrt{\frac{M}{N-2}}$$

N-2 in denominator since 2 degrees of freedom are taken in calculating m and b (two points make a line, so σ for *N*=2 is meaningless.)

This assumes:

- a normal distribution of data points about the line
- spread of points is of similar magnitude for full data range

"Goodness" of fit can be further characterized by the *correlation coefficient, r* (or *coefficient of determination,* r^2), calculated as:

$$r^2 = \frac{\sum_{i=1}^{N}\left(y_i - \langle y \rangle\right)^2 - M}{\sum_{i=1}^{N}\left(y_i - \langle y \rangle\right)^2}$$

For a "perfect fit"
M=0 $\Rightarrow$ $r^2=1$

For $r^2=0$*, <y> represents data as well as a line*

- Many calculators, spreadsheets & other math tools are programmed to perform linear least-squares fitting, as well as fits to more complex equations following a similar premise.

- Many nonlinear equations can be linearized by taking the log of both sides

e.g.,

$y = bx^m$ becomes $\ln y = m \ln x + \ln b$

or $y' = mx' + b'$

- Multiple regression

In some cases, we wish to fit data dependent on more than one independent variable. The procedure will be exactly analogous to that used above, and solutions can be obtained through matrix algebra.

Here we will consider the simple case of a linear dependence on 2 independent variables.

Our minimization criterion can thus be written:

$$M = \sum_{i=1}^{N} [y_i - (a + bx_i + cz_i)]^2 = \text{minimum}$$

Mathematically we require: $\dfrac{\partial M}{\partial a} = 0, \ \dfrac{\partial M}{\partial b} = 0, \ \dfrac{\partial M}{\partial c} = 0$

which yields the 3 equations:

$$Na + \left(\sum_{i=1}^{N} x_i\right)b + \left(\sum_{i=1}^{N} z_i\right)c = \sum_{i=1}^{N} y_i$$

$$\left(\sum_{i=1}^{N} x_i\right)a + \left(\sum_{i=1}^{N} x_i^2\right)b + \left(\sum_{i=1}^{N} (x_i z_i)\right)c = \sum_{i=1}^{N} (x_i y_i)$$

$$\left(\sum_{i=1}^{N} z_i\right)a + \left(\sum_{i=1}^{N} x_i z_i\right)b + \left(\sum_{i=1}^{N} (z_i^2)\right)c = \sum_{i=1}^{N} (z_i y_i)$$

These equations can be solved to obtain *a, b* and *c*.

References

1) D.C. Baird, *Experimentation: An Introduction to Measurement Theory and Experiment Design*, *2nd Ed*., Prentice Hall, Englewood Cliffs, NJ (1988).
2) D.C. Montgomery, *Design and Analysis of Experiments, 3rd Ed*., John Wiley and Sons, New York, NY (1991).
3) A. Goldstein, *Biostatistics: An Introductory Text*, MacMillan Co., New York, NY (1964).
4) C.I. Bliss, *Statistics in Biology, Volume 2*, McGraw-Hill, Inc. New York, NY (1970).
5) R.J. Larson and M.L. Marx, An Intro. to Mathematical Statistics and it Applications, 2nd ed., Prentice-Hall, Englewood, NJ (1986).
6) A.C. Bajpai, I.M. Calus and J.A. Fairley, *Statistical Methods for Engineers and Scientists: A Students' Course Book*, John Wiley and Sons, Chichester, Great Britain (1978).