3.051J/20.340J

Lecture 15
Statistical Analysis in Biomaterials Research

1. Rationale: Why is statistical analysis needed in biomaterials research?

❖ **Many error sources** in measurements on living systems!

Examples of Measured Values and Error in Biomaterials Data

Graphs removed for copyright reasons.

Case Example: cell sedimentation assay of % of cells adhering to a biomaterial vs. control (the simplest cell assay)

Sources of data variation:

- Contamination of surface ⇒ cytotoxicity or modified surface chemistry
- Variations in the number of cells seeded on each surface
- Variations in biomaterial synthesis (reagent amounts, T, time, etc.)
- Variations in cells (different cell passages)
- Variations in media (e.g., different concentrations, proteins agglomerated)
- Variations in sterilization procedure
- Researcher error (e.g., sneezing on samples)

- Behavior is best characterized by a ***population*** or ***distibution*** of values

Case Example: Titration of HCl solution with equimolar NaOH solution.

Graphs removed for copyright reasons.

Data converge to a population as $N \to \infty$:

Graph removed for copyright reasons.

❖ **Our Goal:** to make enough measurements to accurately characterize the *distribution of behaviors,* via

➢ The ***mean*** of the data distribution = <x> from *N* measurements

$$\langle x \rangle = \frac{\sum_{i=1}^{N} x_i}{N}$$

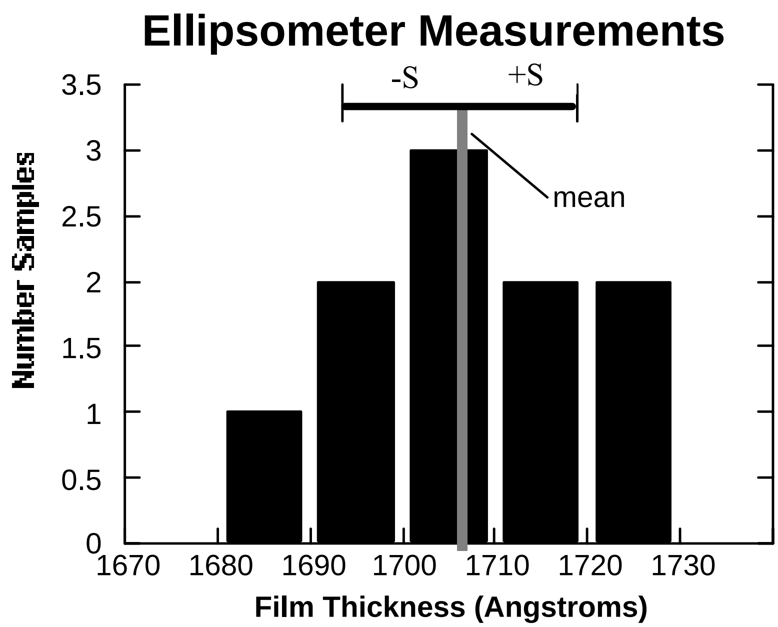
➢ The width of the distribution, or **standard deviation, *S***:

$$S = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2}{N-1}}$$

Note that the standard deviation, *S*, is divided by *N*-1 to avoid biasing, since we do not know the "true" mean, μ.. The universal standard deviation, σ, is divided by *N*.

Equivalently the **variance, S^2**, is defined by:

$$S^2 = \frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N(N-1)}$$



2. Important Distribution Functions

A. Gaussian (or normal) distribution

- describes processes dominated by diffusive forces
  Examples: cell migration

- processes subject to random error or fluctuation: (+ or – equally likely)
  Example: cell number seeded onto surface

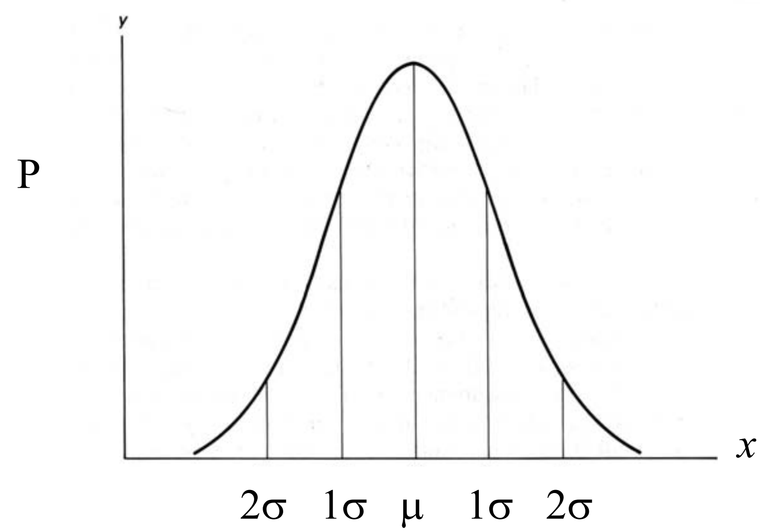
For a Gaussian distribution, the probability P of measuring a value x is given by:

$$P(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

where μ is the true (universal) mean of the distribution and σ is the universal standard deviation.

68% of values fall within μ±σ

95% of values fall within μ±2σ



Real Datasets vs. Distribution

- involve a finite set of points
- approach the theoretical distribution only as $N \to \infty$
- must account for standard deviation of *measured* mean (<x>) (<x> → μ as N → ∞)

⇒ ***standard deviation of the mean, S_m*** (also known as the ***standard error***):

$$S_m = \frac{S}{\sqrt{N}}$$

68% of measured <x> values fall within $\mu \pm S_m$

95% of measured <x> values fall within $\mu \pm 2S_m$ } Use as "confidence interval" for datasets with large N

B. Student's t Distribution

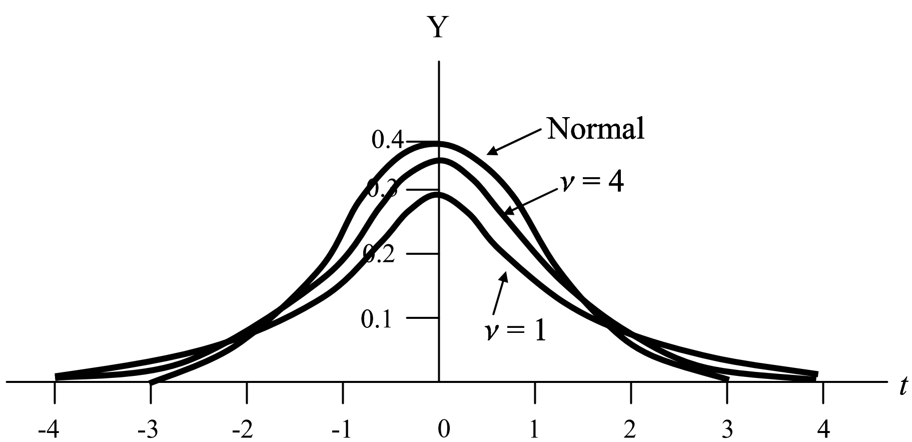
> Named for W.S. Gossett, who published work on statistics under pseudonym "Student" in early 1900s.

- Appropriate for small datasets ($N < 30$) (can be applied to any size)

$$P = \frac{P_0}{(1+\frac{t^2}{N-1})^{N/2}}$$

where P_0 is a constant chosen so that the area under the probability curve ≡ 1 (obtained by integrating P) and t is the statistic:

$$t = \frac{<x> - \mu}{S}\sqrt{N-1}$$



Student's t distribution in various values of ν.
Figure by MIT OCW.

- **Uses of the *t* Distribution**

i) Calculate *confidence interval* for a mean derived from small datasets:

$$\left(\langle x\rangle - t_{\frac{1+P}{2}} S_m\right) < \mu < \left(\langle x\rangle + t_{\frac{1+P}{2}} S_m\right)$$

Interval in x over which you are P% confident of finding the "universal" mean μ.

$t_{\frac{1+P}{2}}$ is the critical t value for a given *confidence level P* (e.g., 90%, 95%, 99%) and a dataset with *degrees of freedom* $\nu = N$-1

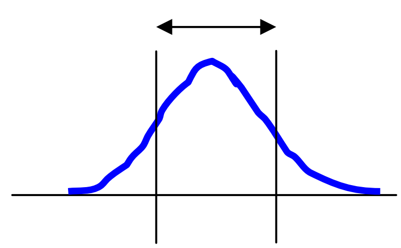
$t_{\frac{1+P}{2}}$ is obtained from t_p Distribution Percentile chart (see handout)

Case Example: 11 measurements of migration speed of cells are made. What is the 95% confidence interval for the mean migration speed?

measurement #	migration speed (μm/min)
1	62
2	52
3	68
4	23
5	34
6	45
7	27
8	42
9	83
10	56
11	40

ν=10

$t_{.5(1+.95)} = t_{.975} = 2.23$ (from table)



1. Calculate <x>

$$\langle x \rangle = \frac{\sum_{i=1}^{N} x_i}{N} = 48.4$$

2. Calculate S_m

$$S = \sqrt{\frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N(N-1)}} = \sqrt{\frac{11(29,000)-(532)^2}{11(10)}} = 18.1$$

$$S_m = \frac{S}{\sqrt{N}} = 18.1/(11)^{0.5} = 5.46$$

The true mean (μ) migration speed is 95% likely to be found in the interval:

$$\left(\langle x\rangle - t_{\frac{1+P}{2}} S_m\right) < \mu < \left(\langle x\rangle + t_{\frac{1+P}{2}} S_m\right)$$

$$48.4-2.23(5.46) < \mu < 48.4+2.23(5.46)$$

$$36.2 < \mu < 60.6$$

ii) Evaluate whether sample populations are statistically different

For a ***two-sample t-test:***

$$t = \frac{\langle x\rangle - \langle x'\rangle - (\mu_x - \mu_{x'})}{\sigma_p\sqrt{\frac{1}{N}+\frac{1}{N'}}}$$

test that $\mu_x \neq \mu_{x'}$ to some confidence level where σ_p = standard deviation of pooled population

$$\sigma_p^{\ 2} = \frac{\sum_{i=1}^{N}(<x> - x_i)^2 + \sum_{j=1}^{N'}(<x'> - x_j')^2}{(N-1)+(N'-1)} = \frac{(N-1)S_x^{\ 2} + (N'-1)S_{x'}^2}{N+N'-2}$$

Case Example: Does a surface modification change the % cells adhered?

fraction of cells adhered to control surface	fraction of cells adhered to modified surface
0.225	0.209
0.262	0.205
0.217	0.196
0.240	0.210
0.230	0.202
0.229	0.207
0.235	0.224
0.217	0.223
	0.220
	0.201

1. Calculate $<x>$, $<x'>$, S_x, $S_{x'}$

$$<x> = \frac{\sum_{i=1}^{N} x_i}{N} = 0.232 \qquad <x'> = \frac{\sum_{i=1}^{N} x_i'}{N} = 0.210$$

$$S_x = \sqrt{\frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N(N-1)}} = \sqrt{\frac{8(0.4316)-(1.8552)^2}{8(7)}} = 0.0145$$

$$S_{x'} = \sqrt{\frac{N\sum_{i=1}^{N} x_i'^2 - \left(\sum_{i=1}^{N} x_i'\right)^2}{N'(N'-1)}} = \sqrt{\frac{10(0.4406)-(2.097)^2}{10(9)}} = 0.00977$$

2. Calculate σ_p

$$\sigma_p = \sqrt{\frac{(N-1)S_x^{\,2} + (N'-1)S_{x'}^2}{N+N'-2}} = \sqrt{\frac{7(0.0145)^2 + 9(0.00977)^2}{8+10-2}} = 0.012$$

3. Calculate two-sample t value for $\mu_x = \mu_{x'}$

$$t = \frac{\langle x\rangle - \langle x'\rangle - (\mu_x - \mu_{x'})}{\sigma_p\sqrt{\frac{1}{N}+\frac{1}{N'}}}$$

$$t = \frac{<x> - <x'>}{\sigma_p\sqrt{\frac{1}{N}+\frac{1}{N'}}} = \frac{0.232-0.210}{0.012\sqrt{\frac{1}{8}+\frac{1}{10}}} = 3.86$$

4. Test $\mu_x \neq \mu_{x'}$ at 99% confidence level (P<0.01):

for $-t_{\frac{1+P}{2}} < t < t_{\frac{1+P}{2}} \quad \Rightarrow \mu_x = \mu_{x'}$

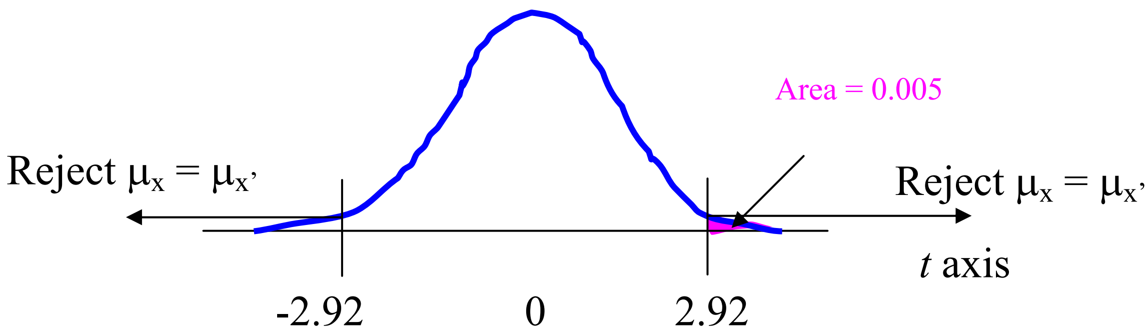
$\nu = N + N' - 2 = 16$

$t_{.5(1+.99)} = t_{.995} = 2.92$ (from table)

$-2.92 \leq t \leq 2.92$

t=3.86 does not fall within this interval.

The means are statistically different with P<0.01, so a change in % cells adhered was observed.



Statistical Significance

In experiments involving biological systems, $P < 0.05$ is generally accepted to indicate measured change is not attributable to *random* error

(still may not be practically meaningful!)

"significant at the 5% level"

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).