Introductory statistics for biological engineers

Module 1, Lecture 5

20.109 Fall 2013

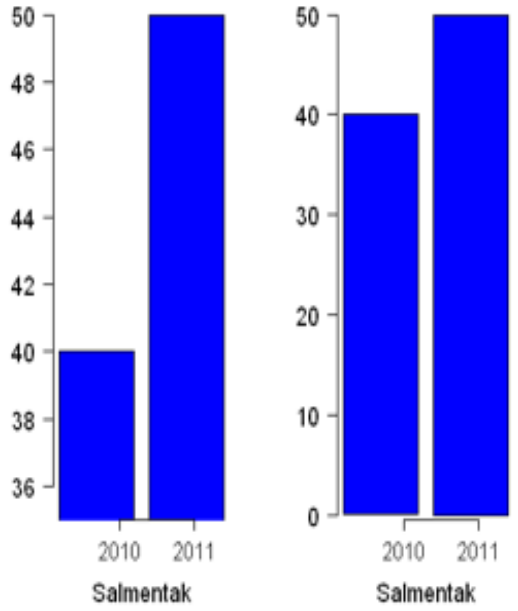
Agi Stachowiak and Bevin Engelward

Statistics: what are they good for?

predictions

determining significance

portray a relationship



Source: Wikimedia Commons
Author: Joxemai
License: Creative Commons Attribution-Share Alike 3.0 Unported

Statistics enhance analysis and its communication

- Check on biases (imperfect!)
- Pick out subtle differences from noise

- Common language
- Concise representation of knowledge

But the science must be sound

- The question posed may limit/slant the answer found
- Political example
  - Do you favor repeal of the death tax, so that many families won't be unfairly burdened with hefty taxes at the time of their grief?
    http://analysights.wordpress.com/2010/07/19/avoiding-biased-survey-questions/
- Science example
  - most birds have no phallus
  - male ducks have long corkscrew penises – a puzzle
  - until 2000's, no one asked what the female ducks look like!
  - female oviducts wind the other way, serve as a barrier
  - coevolution discovered by Dr. Patricia Brennan

Essential concepts defined

- Sample size: n # of subjects, # of measurements
- Degrees of freedom: DOF # of unconstrained parameters
  - for a single pop.
  - DOF = $n-1$
  - $\sum$ errors $(x_i - \bar{x}) = 0$ ← constraint
- Population mean: $\bar{x} = \frac{\sum_i x_i}{n}$
- Population standard deviation: s

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

- True mean: μ
- True std dev: σ

} when $n = \infty$

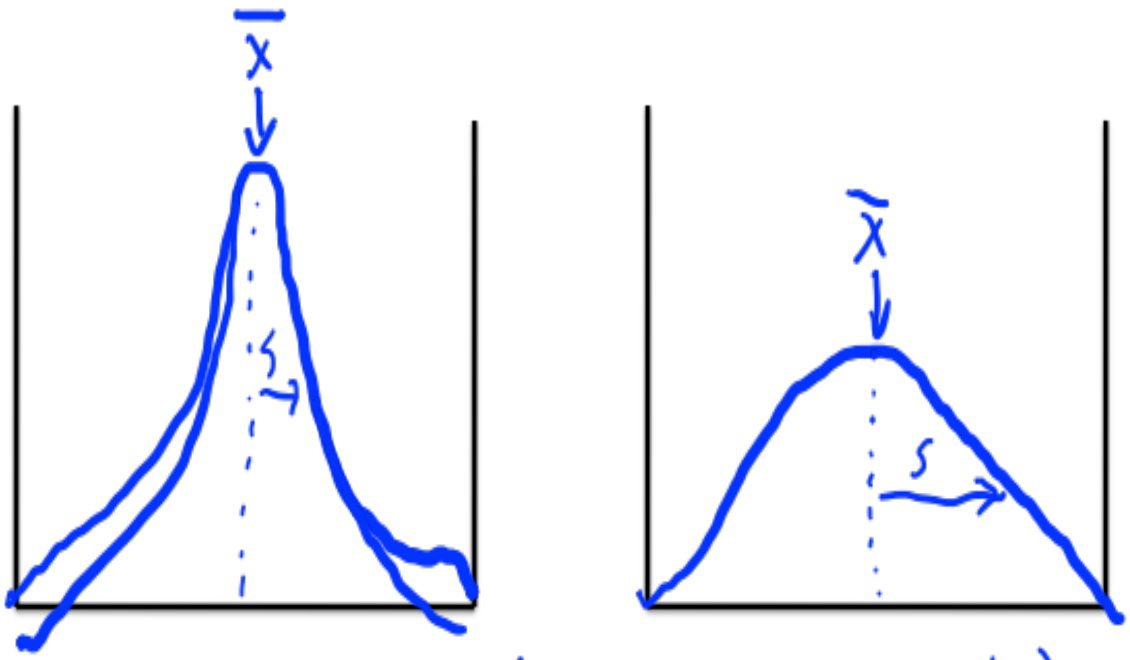
- As n increases, population values → true values

we only ever approach μ, σ

Essential concepts illustrated

sorry for bad drawing :(

normal (Gaussian) distribution

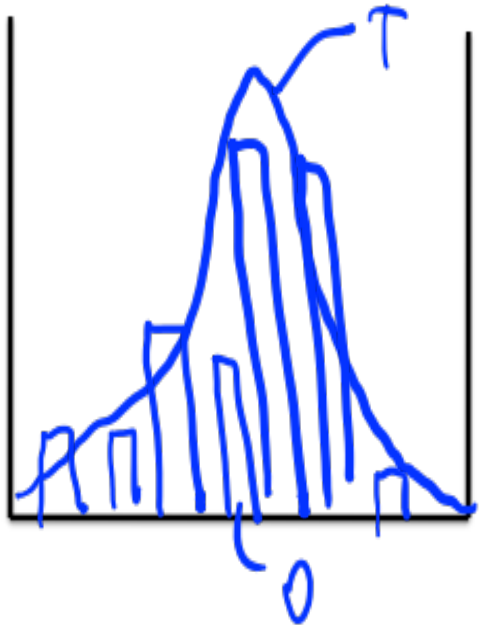


x-axis: measured value (e.g., intensity)

y-axis: # or fraction of samples w/ that value

Integrating definitions and illustration

- Distribution: frequency for measured values → theoretical or observed
- Normal distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$ – equal errors about mean – few outliers
- Each *s* includes a given % of data



1 *s* includes ~68 %

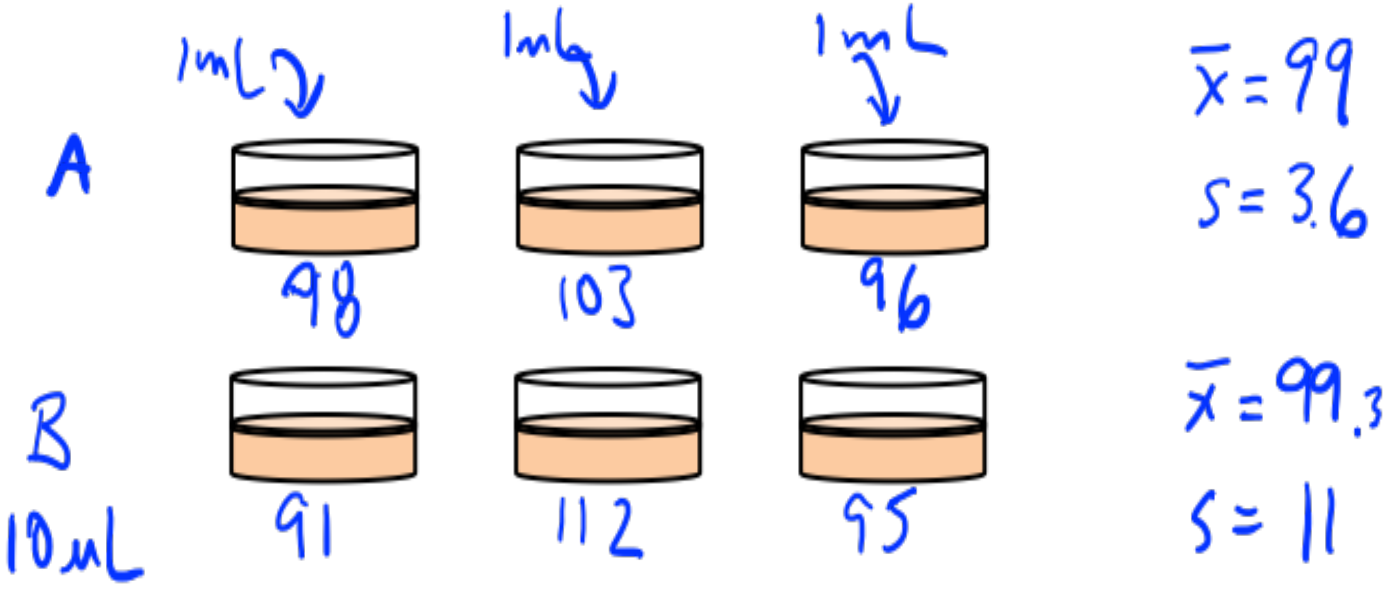
2 *s* includes ~95 %

3 *s* includes 99.7 %

Thanks, Bevin!

Essential concepts: biological example

- Solution with 100 cells/unit volume
- Repeatedly plate that volume *V*
- Case (A) *V* = 1 mL; Case (B) *V* = 10 µL



$\bar{x} = 99$

$s = 3.6$

$\bar{x} = 99.3$

$s = 11$

Confidence intervals (CI): principle

- Example: $\bar{x}$ = 60 and 95% CI is ± 3
- Meaning: 95% of time μ lies in range 60 ± 3
- Consider 90% CI: $\mu = \bar{x} \pm a$
  - Is $a < 3$, $a > 3$, or $a = 3$? Why?

trade-off b/w precision and confidence

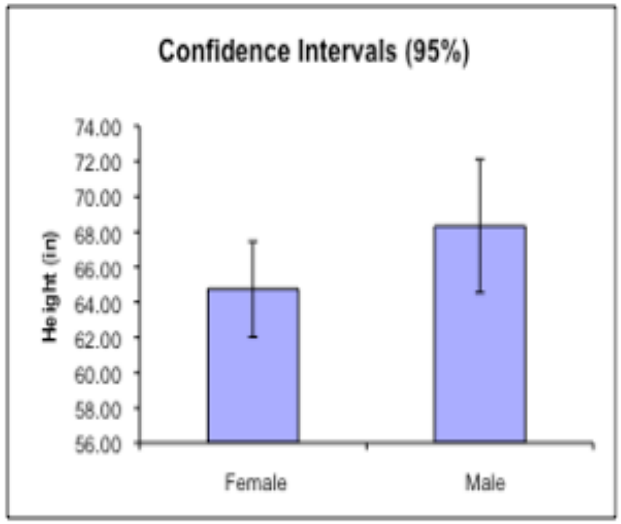
100% CI
0% CI

how confident are you?

- Mnemonics: extremes; betting example
- Effect of *n*? as $n \uparrow$, more precise for a given CI

Calculating CI: old school or software

$$\mu = \bar{x} \pm \frac{t s}{\sqrt{n}}$$



- Can find *t* tabulated by DOF vs CI%
  - look up and plug in
- In Excel, us *TINV* function
  - input *p*-value = (100-CI)/100

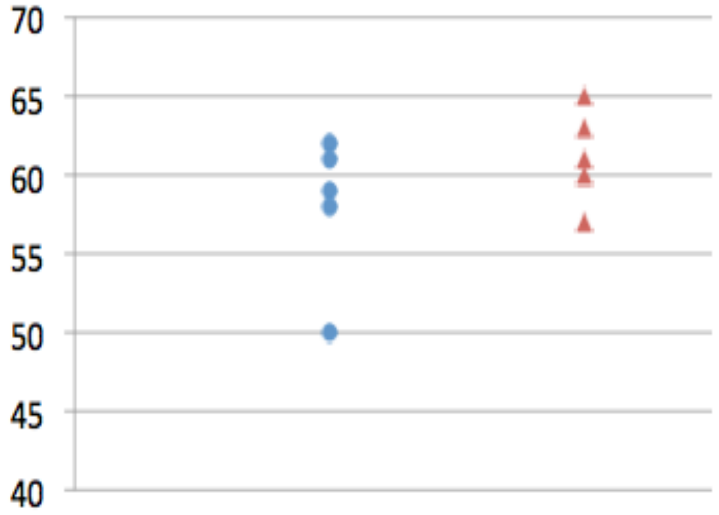
n-1

if CI=95%
→ p = 0.05

Utility of t-test: identifying difference

- Are these two data sets different?
- What are the odds of seeing data sets this different by chance?

62	63
61	60
59	57
58	65
50	61



Calculating significance by t-test

signal / noise

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

DOF $= n_1 + n_2 - 2$

t_{table} listed by DOF vs CL (confidence level)

- If t_{calc} > t_{table} difference *is* significant at that CL
- In Excel, us *TTEST* function
  - returns *p*-value → confidence level (CL)

if $p = 0.01$ → CL = 99%

- 1-tailed vs. 2-tailed test

1- one-sided; hypothesis in advance

2- full distribution; no apriori hypothesis

↑ when in doubt! more rigorous

Context for statistical comparisons

- Every statistical test
  - Has assumptions
  - Asks a specific question
  - Requires human interpretation (GIGO)
- Some t-test assumptions
  - normal distribution (cf. Mann-Whitney test)
  - equal variances (type 2 in Excel; type 3 = unequal)
- Posing a question are mean male and female heights different at a CL of 95%?

A few key points easily missed

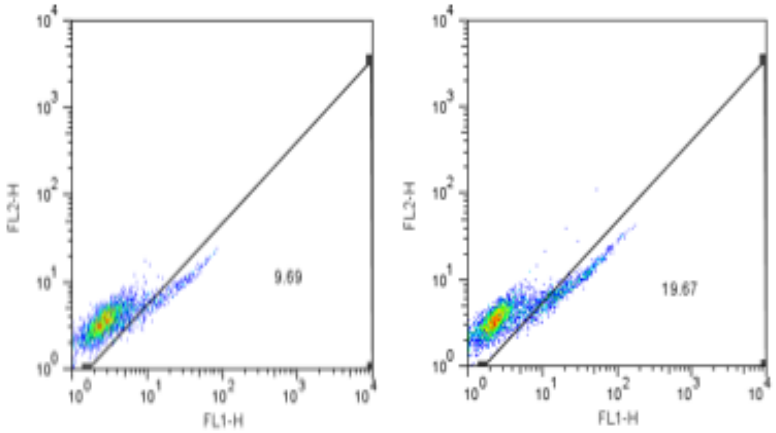
- Confidence interval (CI) is for a *single* population/data set
- A t-test is for comparing *two* populations at a given confidence level (CL)

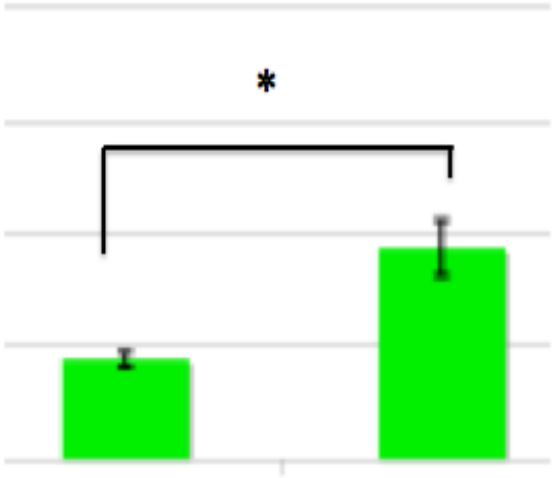
- Excel expects large datasets, and may set $t = 1.96$ ($n = \infty$) by default – so check!

Practice assignment

- Female heights: 64, 72, 67, 68, 68, 65, 71
- Male heights: 75, 74, 76, 70, 76
- Calculate 95% CI for each mean
- Plot means on bar graph with CI error bars
- Apply t-test to the means
  - for multiple comparisons, ANOVA is better
  - comparing many means requires correction
  - remember, $p = 0.05$ means 1 in 20 false positives!
- Excel file available on M1 homepage, *Resources*

How will we use statistics in Mod 1?





Succinct representation of data and analysis