

Introductory statistics for biological engineers (+ligation)

Module 1, Lecture 4

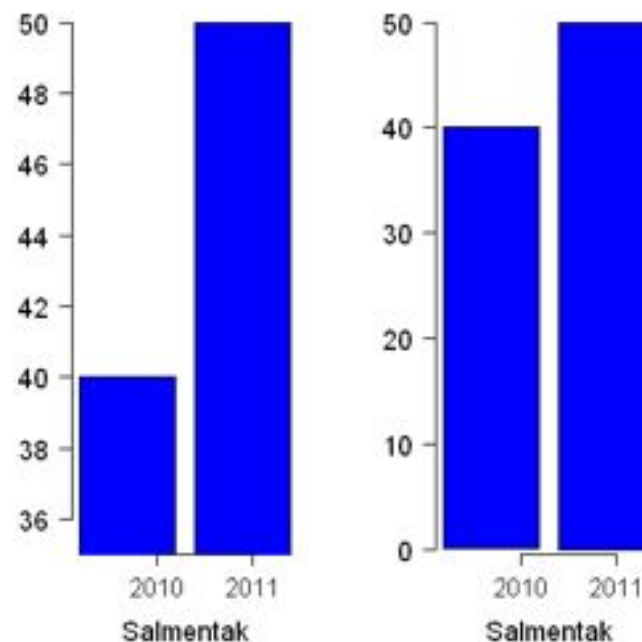
20.109 Fall 2014

Core content developed by: Bevin Engelward
Additional context/framing by: Agi Stachowiak
A few tweaks by: Zac Nagel (Samson lab postdoc)

Statistics: what are they good for?

visualize + synthesize
data concisely

determine legitimacy
significance of
your data



Source: Wikimedia Commons

Author: Joxemai

License: Creative Commons Attribution-
Share Alike 3.0 Unported

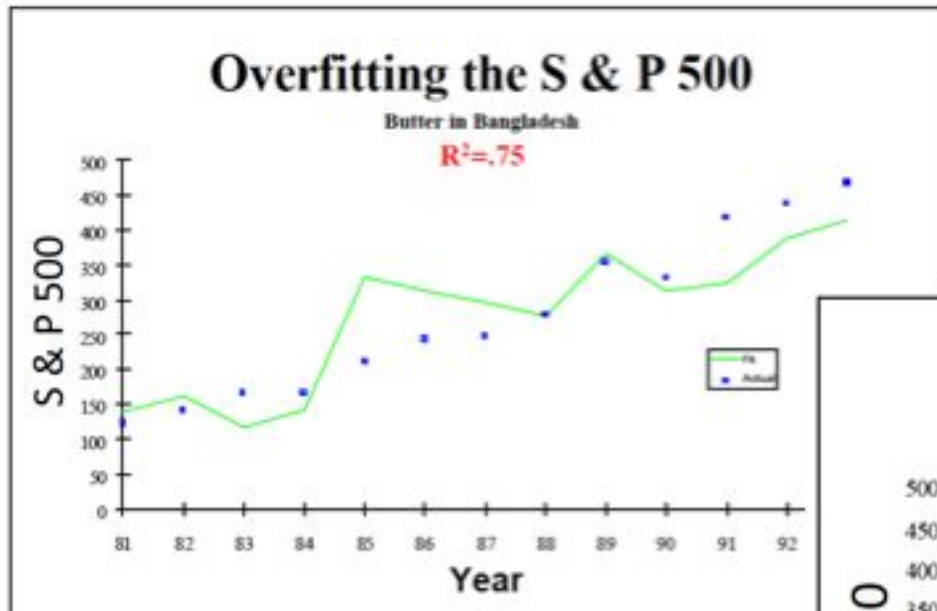
Statistics enhance data analysis and its communication

- A check on **biases** (imperfect!)
- Pick out **subtle** differences from noise
- Provide **common** language
- A **concise** depiction of (certainty of) knowledge

*Did I get an answer? How sure am I of the answer?
Can we all agree on how sure of it we are? -ZN*

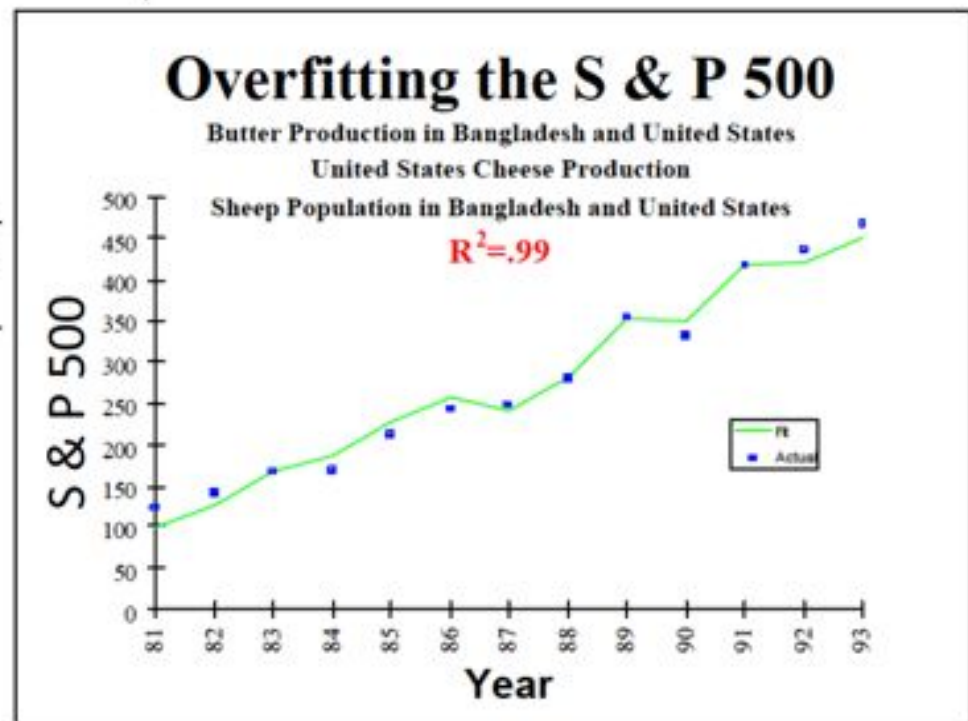
But the science must be sound: nonsense questions

implications for "big data"
publishing expectations



The Journal of Investing
(2007) 16:15-22

h/t ZN



But the science must be sound: biased or leading questions

- The question posed may limit/slant its answer
 - most birds have no phallus
 - male ducks have long corkscrew penises – a puzzle
 - until 2000's, no one asked what female ducks look like!
 - female oviducts wind the other way, serve as a barrier
 - coevolution discovered by Dr. Patricia Brennan

NOT EXACTLY ROCKET SCIENCE: May 6, 2014

Where's All The Animal Vagina Research?

by Ed Yong

National Geographic .com

These stereotypes are pervasive. In [the most cited studies on sexual conflict](#), authors use active words like 'intimidation' and 'coercion' to describe males, but passive words like 'resistance' and 'avoidance' to describe females. More tellingly, males have 'adaptations' and females have 'counter-adaptations'. Males act; females react.

But we know that, as in ducks, females [exert a tremendous amount of control during sex](#). They can store sperm in pouches, expel unwanted sperm, or mate with more males. All of these tricks allow them to reject a male as a father even after having sex with him, and all of them are hard to observe unless you're actually looking. And perhaps, people don't look very hard.

Essential measurement concepts

- Sample size: n # of subjects, # of measurements

- Sample mean: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

- Sample standard deviation: s

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

Essential theoretical concepts

Sample

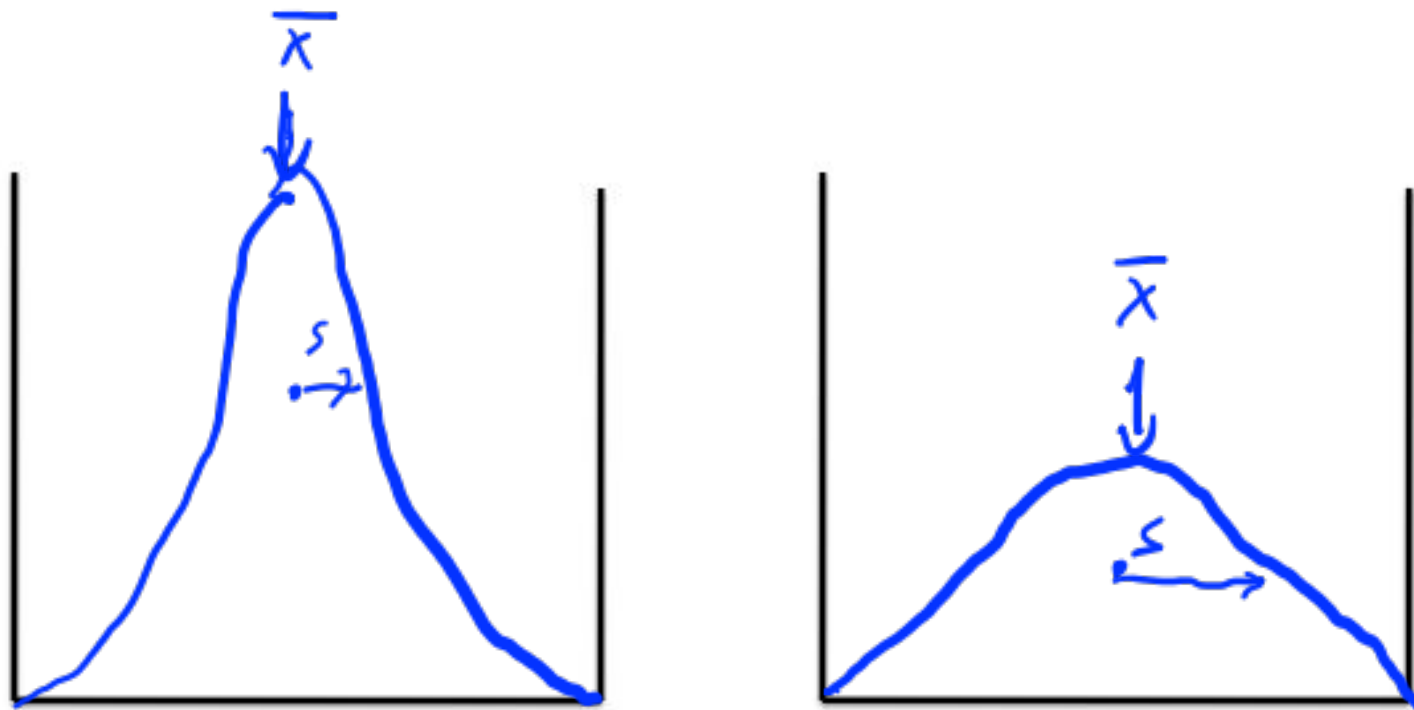
~~Population~~ mean/std dev reflects a limited dataset.

What if we had all the data? -ZN

- True mean: μ
- True std dev: σ
- As n increases, ~~population~~ ^{sample} values \rightarrow true values
we can only approach μ, σ
- Degrees of freedom: DOF ~~±~~ *of unconstrained parameters*
for a single population, DOF = $n - 1$
 \hookrightarrow errors $(x_i - \bar{x}) = 0 \leftarrow$ constraint

Essential concepts illustrated

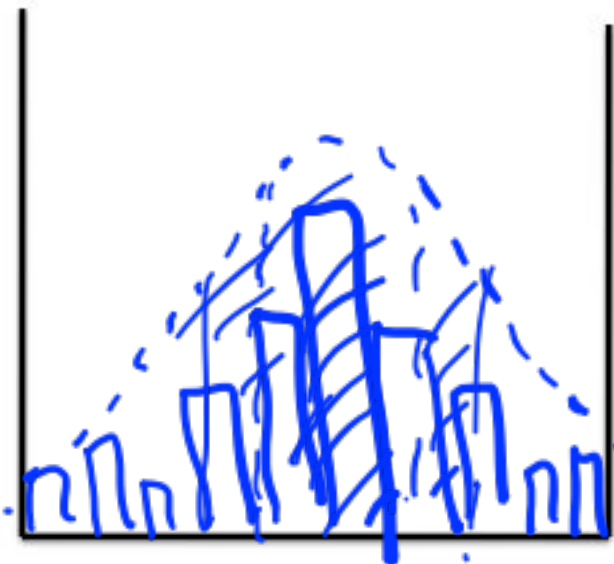
normal (Gaussian) distributions



x-axis: measured value (e.g., intensity)
y-axis: # or fraction of samples with that value

Integrating definitions and illustration

- Distribution: frequency of 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



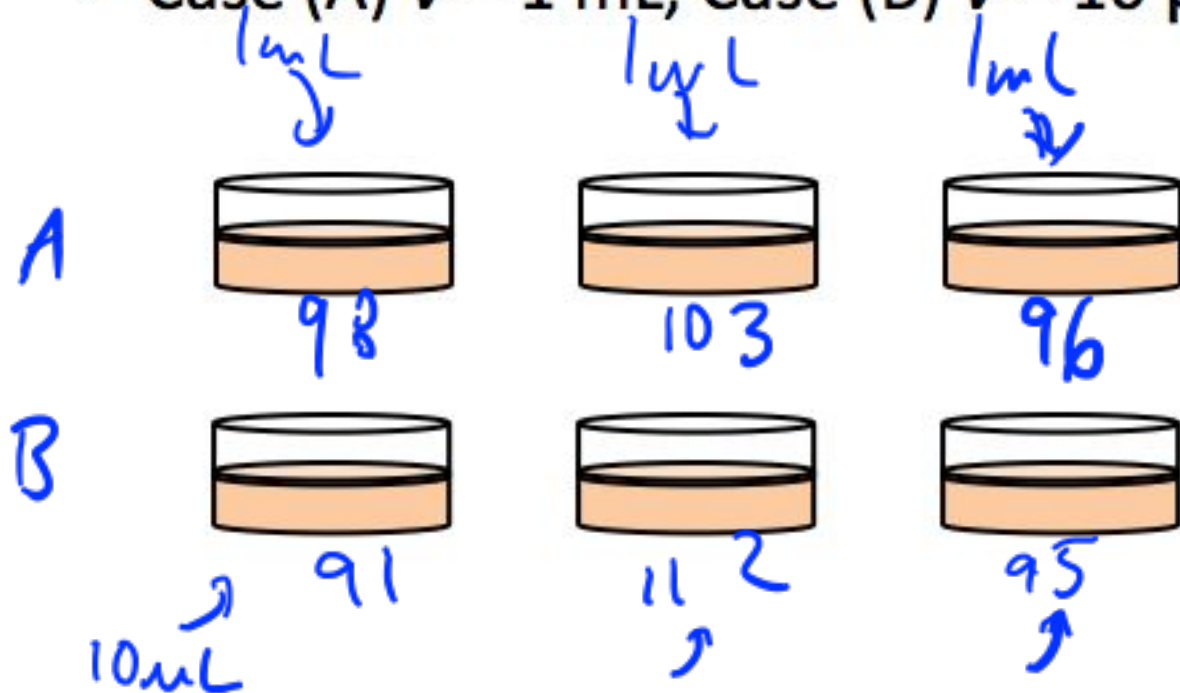
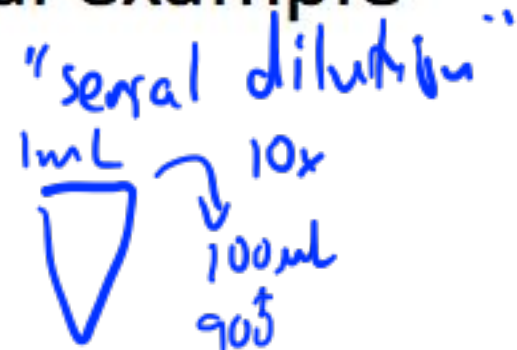
1 s includes ~68 % of data

2 s includes ~95 %

3 s includes ~99.7 %

Essential concepts: biological example

- Solution with 100 cells/unit volume
- Repeatedly plate that volume V
- Case (A) $V = 1 \text{ mL}$; Case (B) $V = 10 \mu\text{L}$



$$\bar{x} = 99$$

$$s = 3.6$$

$$\bar{x} = 99.3$$

$$s = 11$$

Two key statistical questions

- How confident are we about an estimate of a population mean?


sample

⇒ confidence intervals

- How confident are we that two populations are different?

⇒ t-test (confidence level)

Confidence intervals (CI): principle

- Bio example A (1 mL):
 - $\bar{x} = 99$, $s = 3.6$, and 95% CI is ± 5.7
- Bio example B (10 uL):
 - $\bar{x} = 99$, $s = 11$, and 95% CI is ± 28 !
- True: “There is a 95% probability that the confidence interval 99 ± 5.7 contains the true mean, μ .”
- Technically *false*: “There is a 95% probability that the true mean, μ lies in the confidence interval 99 ± 5.7 .”
- Best explanation: contradiction if repeat experiment
- Regardless, CI is a range of “plausible” values for μ 

http://www.psych.utah.edu/gordon/Classes/Psych_6500/Stats%20and%20Params/ConfidenceIntervaloftheMean.pdf

Illustrating CI principle

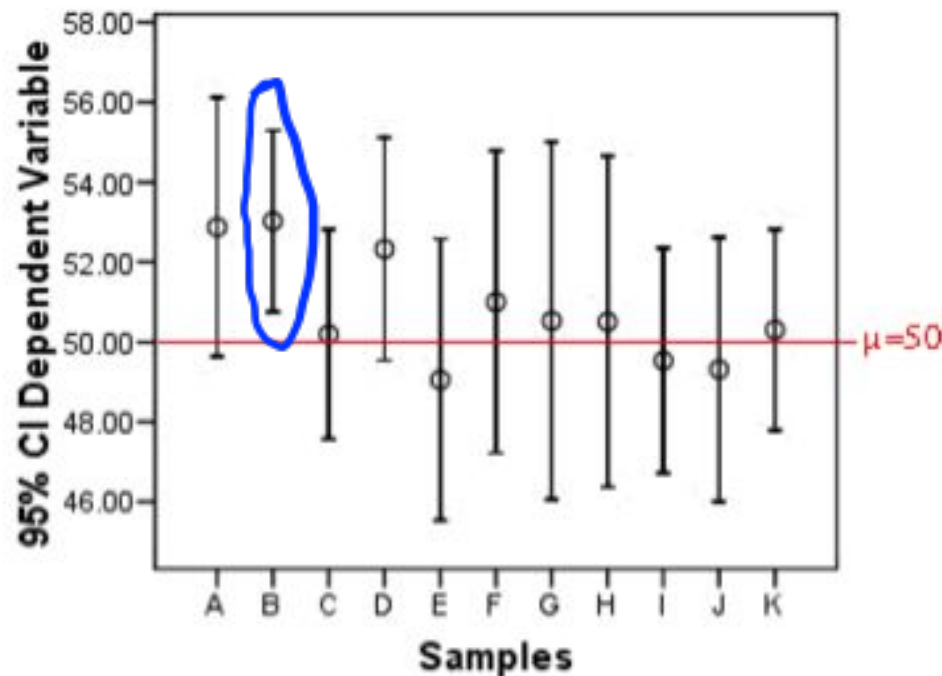


Figure 2: Confidence intervals from 11 samples drawn from a population with a mean of 50.

http://www.psych.utah.edu/gordon/Classes/Psych_6500/Stats%20and%20Params/ConfidenceIntervaloftheMean.pdf

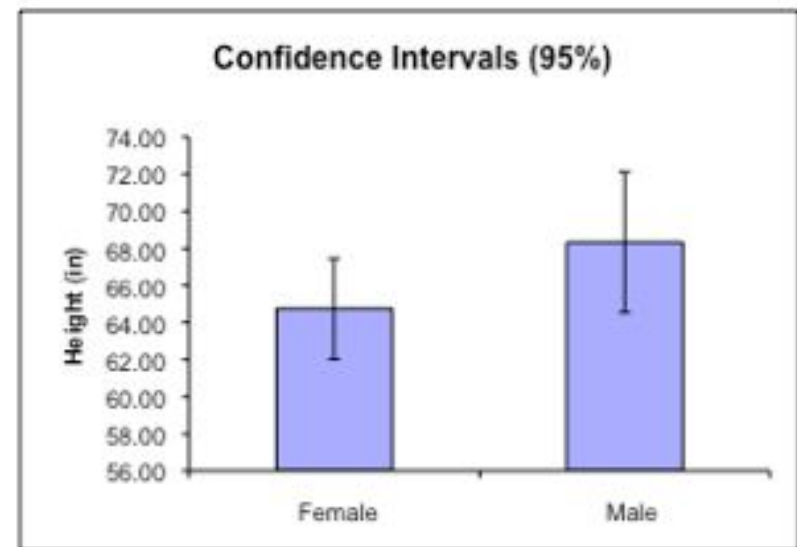
Confidence intervals (CI): intuition

- Consider 90% CI for Exp. A: $\mu = 99 \pm a$
 - Is $a < 5.7$, $a > 5.7$, or $a = 5.7$? Why?
- Thought experiment: extremes
 - What is a when CI = 0%? $a = 0$
 - What is a when CI = 100%? $a = \infty$
- Thought experiment: betting
 - Do you bet more or less \$ as you have to be more and more precise in your answer to be considered right?

★ trade-off b/w precision + confidence ★
- Effect of n for a given CI? *more precise*

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

\downarrow
 $n-1$

- In Excel, us $TINV$ function
 - input p -value = $(100-CI)/100$

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



(need more exps. ultimately)

Calculating significance by t-test

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

DOF $n_1 + n_2 - 2$

t_{table} listed by DOF
vs. confidence level

- If $t_{calc} > t_{table}$ difference is significant at that CL
signal: noise difference in means vs. spread
- In Excel, use *TTEST* function
 - returns p -value \rightarrow confidence level (CL)
if $p = 0.01$, C.L. = 99%

t-table excerpt

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02
df								
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764

<http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf> (San Jose State University)

What is a 1-tailed vs. 2-tailed test?

- 1-tailed
 - a priori hypothesis
 - using half of distribution
 - is x bigger than y ?
- 2-tailed
 - no a priori hypothesis
 - using full distribution
 - is x different than y ?
- The more rigorous choice is: 2-tailed

Context for statistical comparisons

- Every statistical test
 - Asks a specific question
 - Has *assumptions*
 - Requires *human interpretation (AIGO)*
- Some t-test assumptions
 - normal distribution (cf. Mann-Whitney test)
 - equal variances (type 2 in Excel; type 3 = unequal)
 - not appropriate for multiple comparisons (cf. ANOVA)

• Posing a question *are mean male and female heights different at a CL of 95%?*

(ignore in 105)

What can we do to increase our confidence?

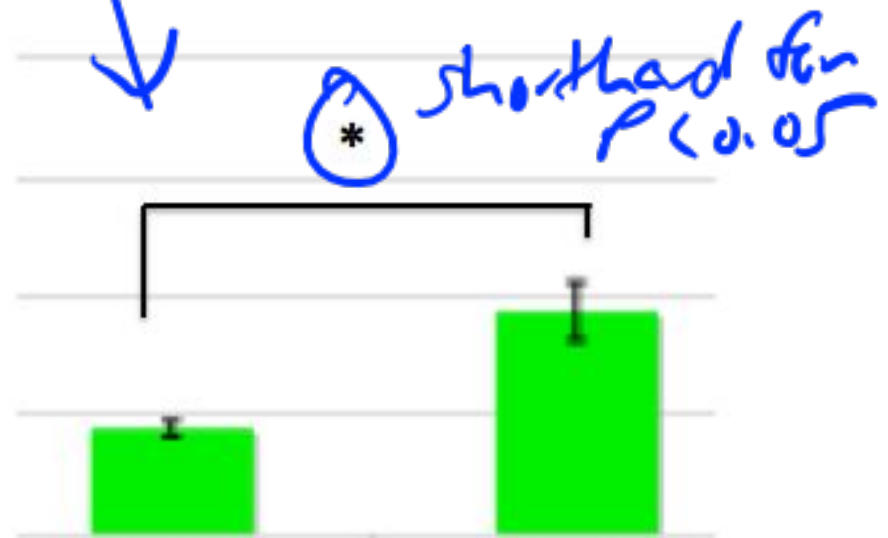
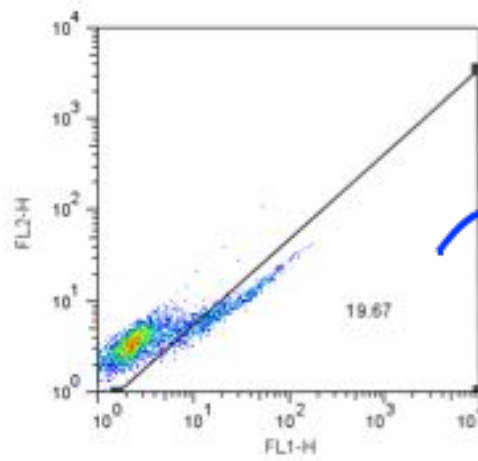
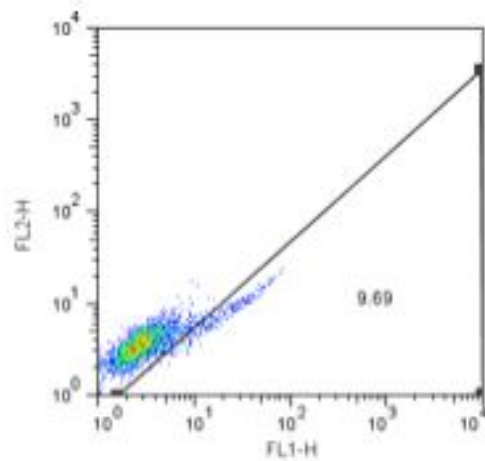
$$t_{calc} = \frac{|\bar{x}_1 - \bar{x}_2|}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

- Seek larger t values by increasing n
 - For a given total number of observations ($n_1 + n_2$), t is maximized when $n_1 = n_2$
- More precise experimental technique
- Avoiding sampling bias, reporting bias
- Perform a second, complementary experiment

A few key points easily forgotten

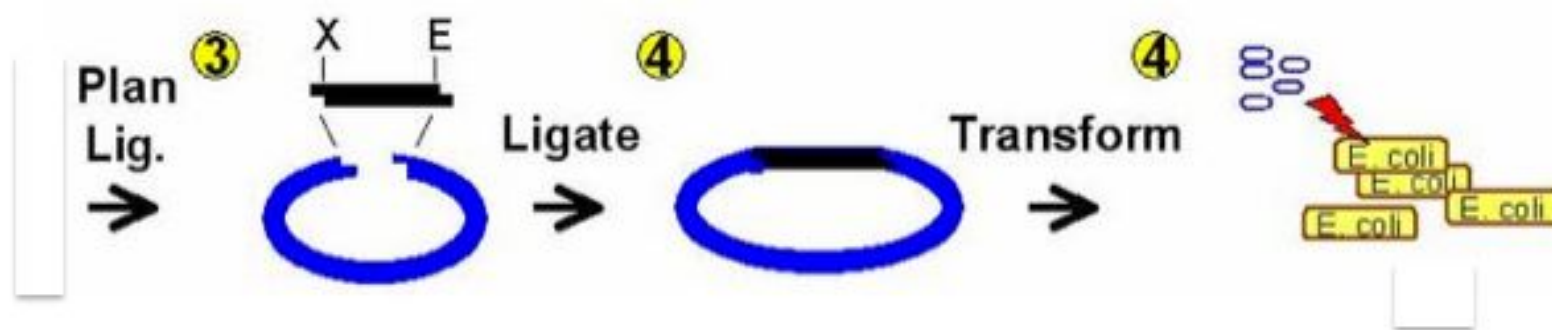
- Confidence interval (CI) is for a *single* population/data set
- A t-test is for comparing *two* populations at a given confidence level (CL)
- Note: you will work on a stats exercise on **M1D6** to practice with both CI and CL
- Excel expects large datasets, and may set $t = 1.96$ ($n = \infty$) by default – so check!

How will we use statistics in Mod 1?



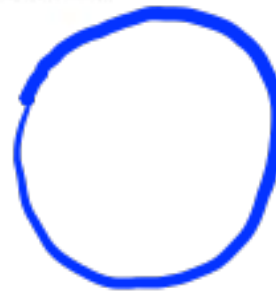
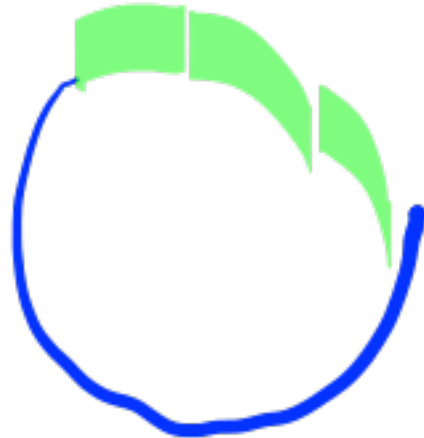
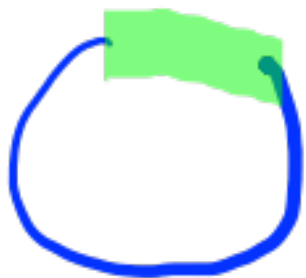
M1D4: Ligation and transformation

- What we'll cover in pre-lab
 - chemistry (briefly)
 - factors affecting yield
 - transformation
 - review controls
- What we'll cover now
 - controls to assess whether ligation was successful



Ligation: what products can be formed?

- Starting reminders
 - reactions are not binary → our shorthand drawings abstract away *populations* of molecules
 - we use antibiotic selection on the **backbone** to isolate and purify any products that are formed



Ligation: What products are likely to be formed? Why? Can be controlled for?

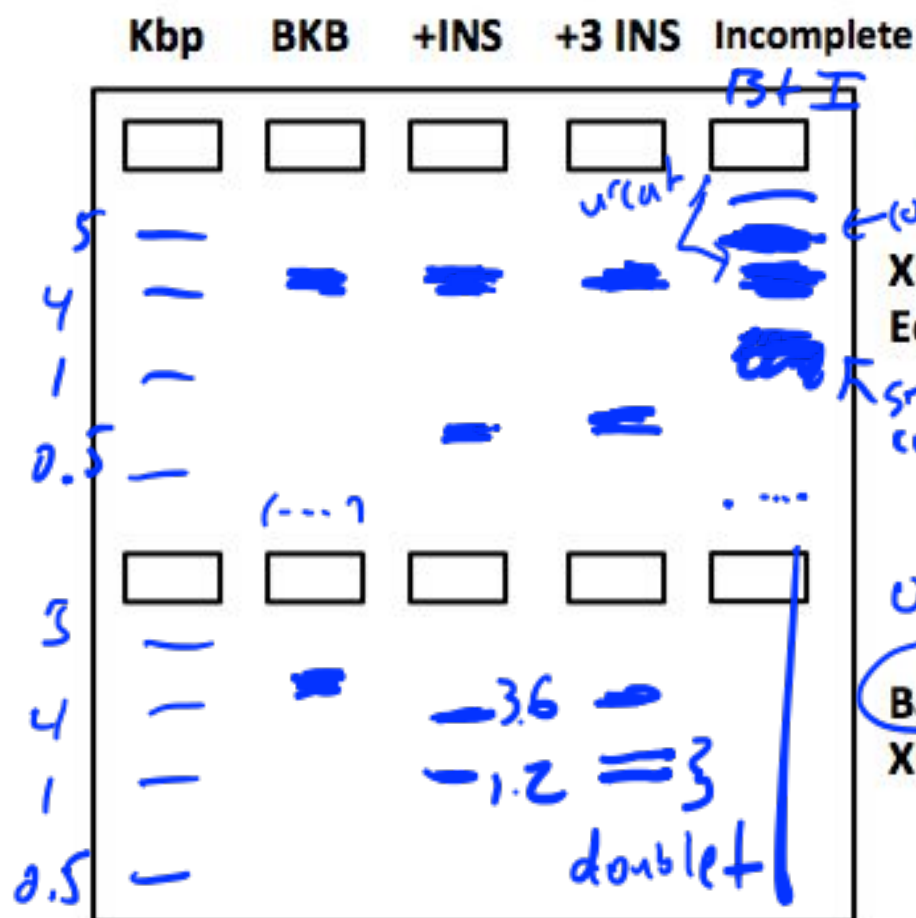
- B + I: desired product, likely w/ correct $[B]$, $[I]$ and I:B ratio
- B + 3I: multi-insert, somewhat likely given I:B > 1
- B + 5I: } not likely
- 2B: }
- B: no insert, likely if poor digestion.

next slide

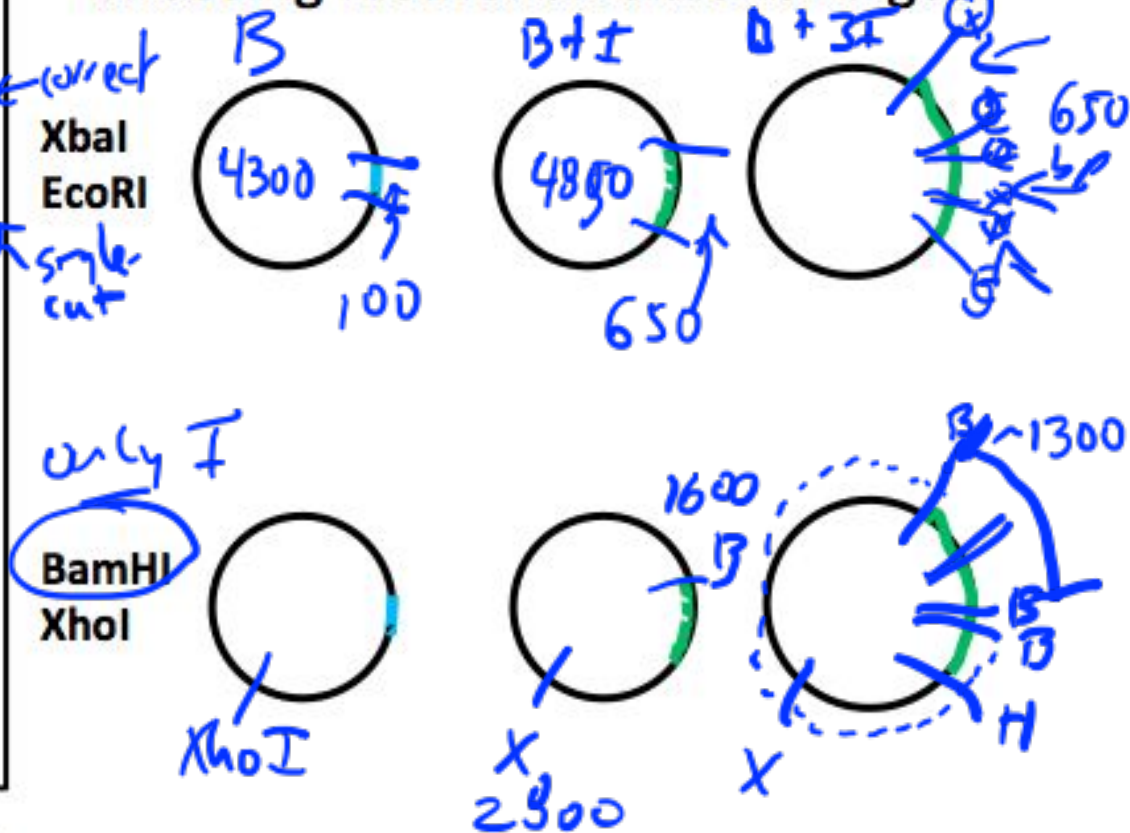
M1D4 ligation controls' purposes

- backbone + insert + ligase
product (+ background) $\times 3$
- backbone + ligase, no insert
control for partial digestion \rightarrow single-cut (+ background?)
reclosure
- backbone (+ insert), no ligase
control for uncut plasmid
* compare yields to determine ratio of product:background

M1D5 diagnostic digest preview



Choosing restriction sites for digest



The only way to identify B + 3I!