n.s

sllobs

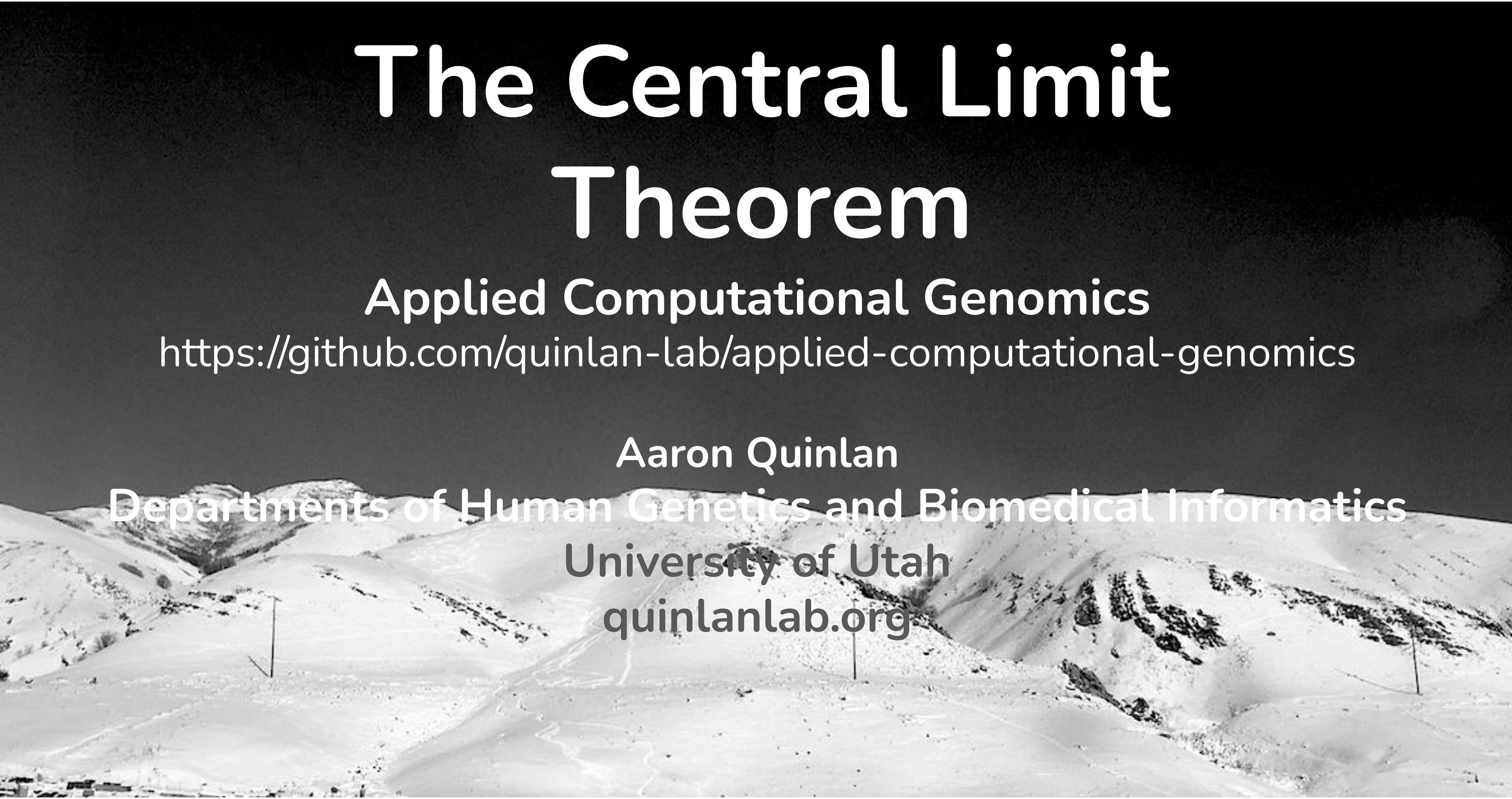
> salt lake learners of biostatistics

Meeting #13. The Central Limit Theorem and Error Bars

Aaron Quinlan

October 14, 2019

bit.ly//sllobs



The Central Limit Theorem

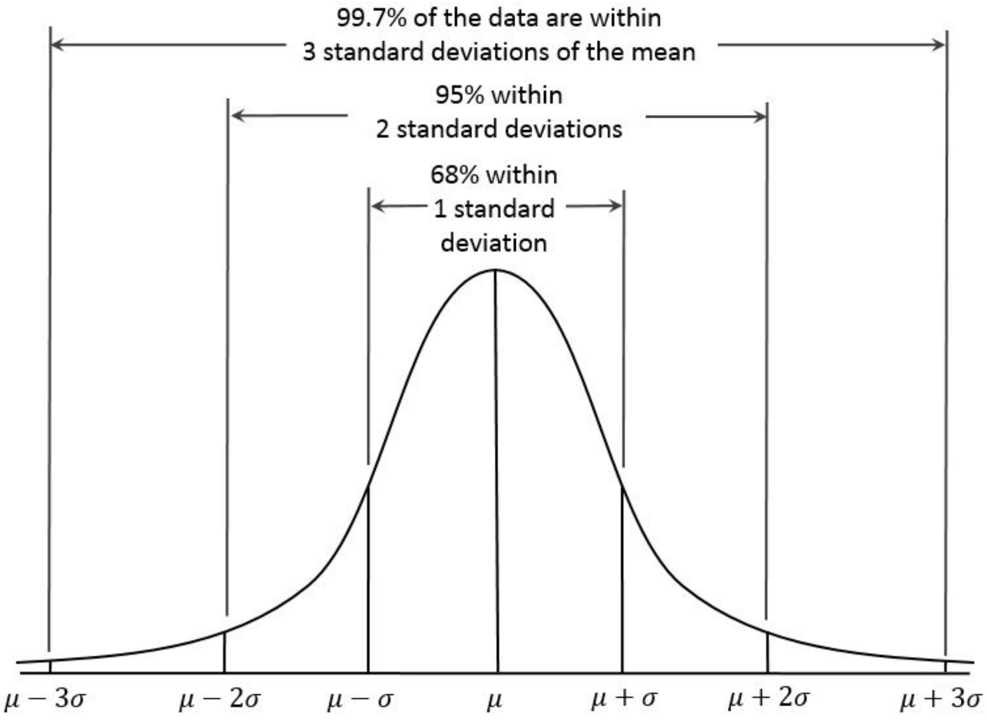
Applied Computational Genomics

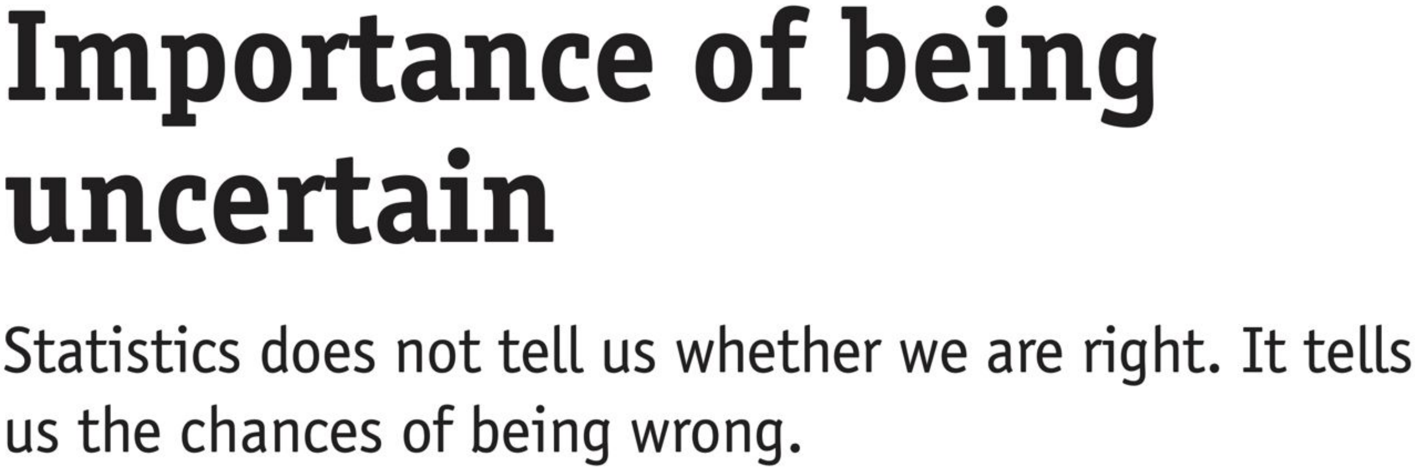
https://github.com/quinlan-lab/applied-computational-genomics

Aaron Quinlan
Departments of Human Genetics and Biomedical Informatics
University of Utah
quinlanlab.org

Recall: Gaussian (Normal) distributions

Two parameters: mean (μ) and standard deviation (σ)

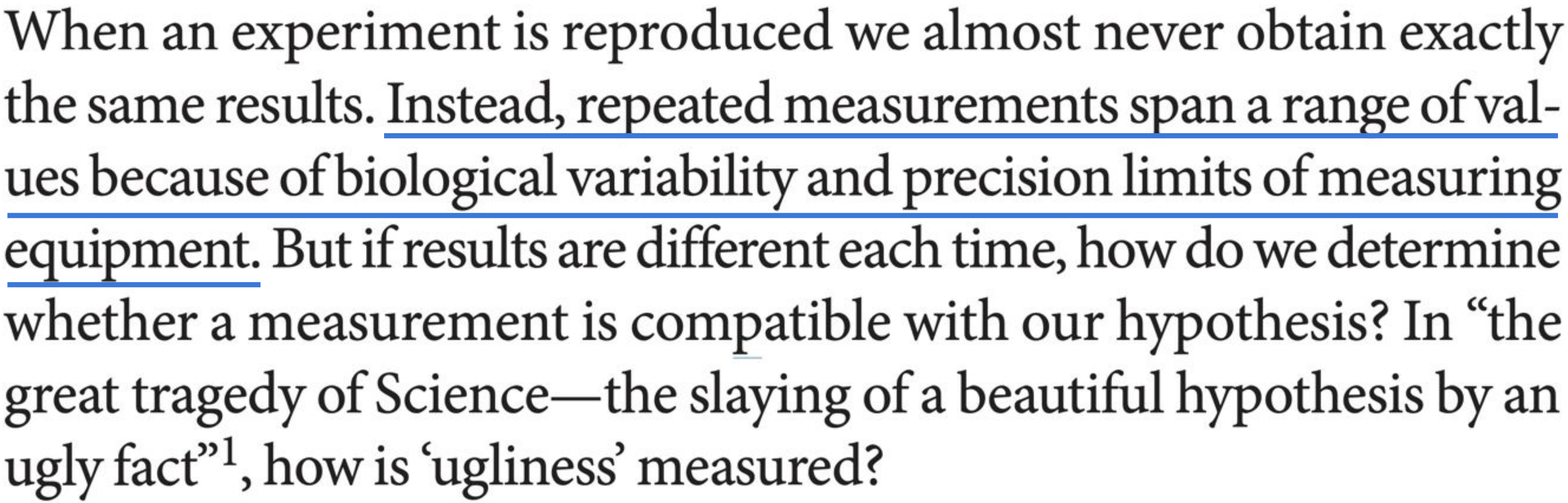




Importance of being uncertain

Statistics does not tell us whether we are right. It tells us the chances of being wrong.

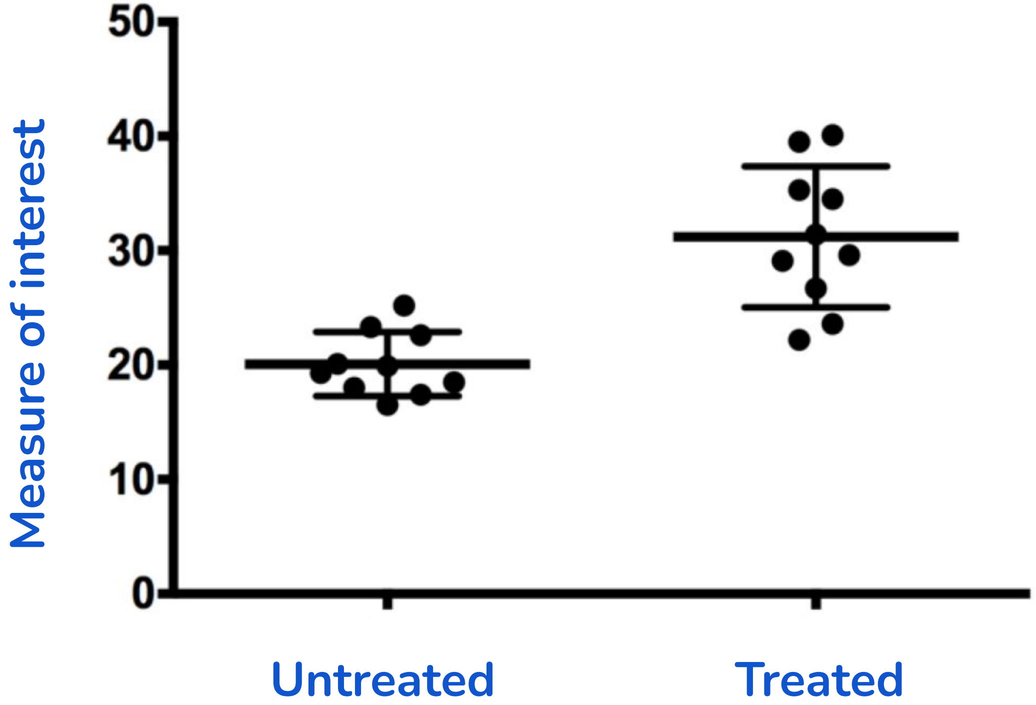
https://www.nature.com/articles/nmeth.2613.pdf



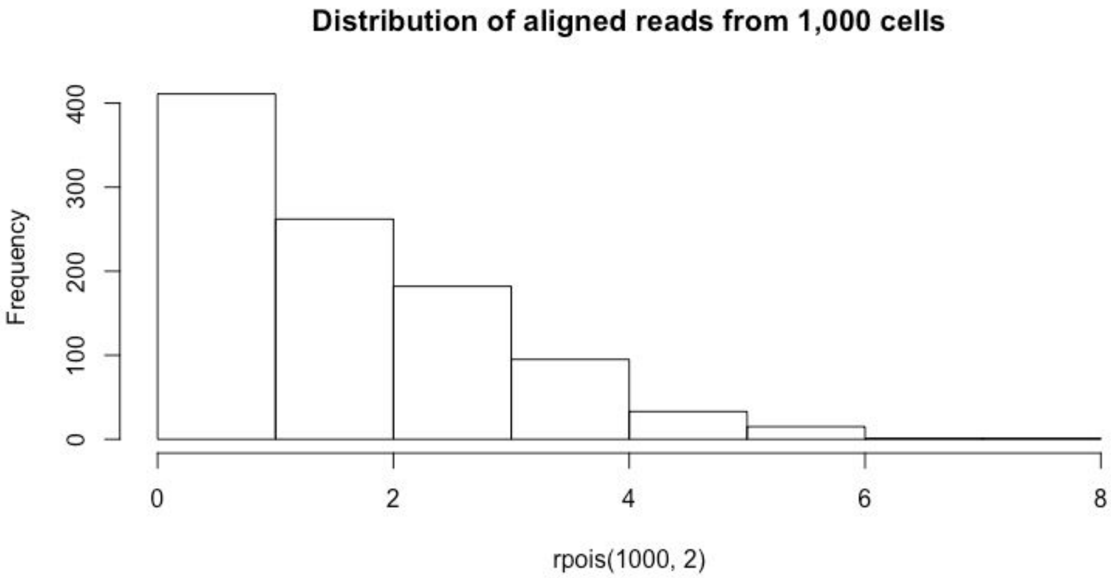
When an experiment is reproduced we almost never obtain exactly the same results. Instead, repeated measurements span a range of values because of biological variability and precision limits of measuring equipment. But if results are different each time, how do we determine whether a measurement is compatible with our hypothesis? In “the great tragedy of Science—the slaying of a beautiful hypothesis by an ugly fact”[1], how is ‘ugliness’ measured?

https://www.nature.com/articles/nmeth.2613.pdf

We are rarely able to observe an entire population. Instead, we take (ideally) random samples.

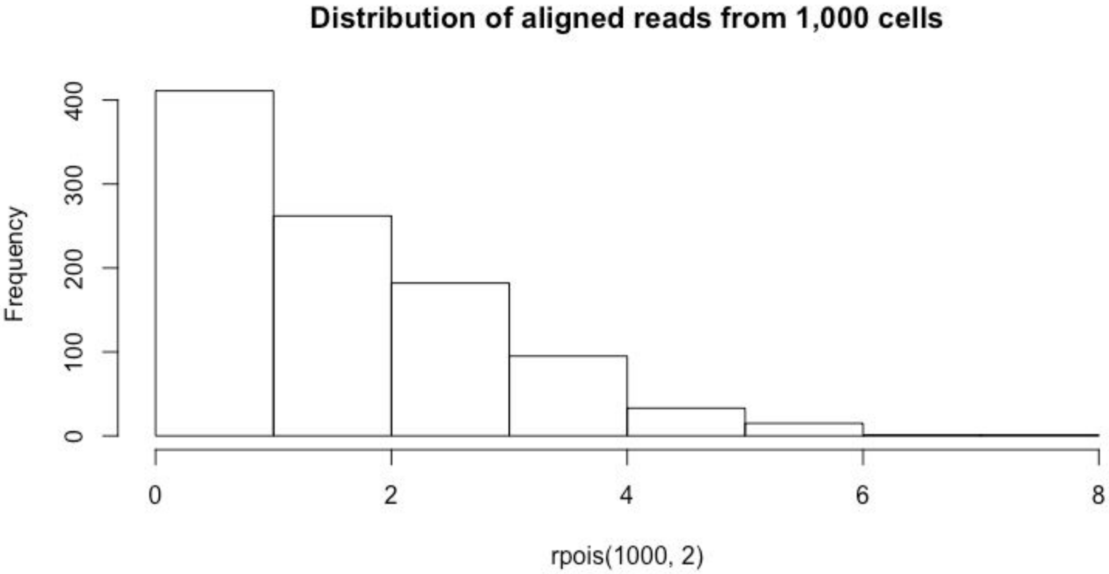


Example: observed read alignments at TP53 from all single cells

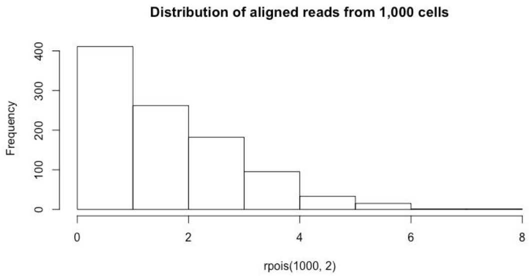


```
hist(rpois(1000,2), main="Distribution of TP53 transcript counts from 1,000 single cells")
```

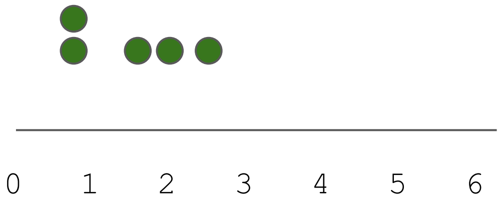
This is a "population distribution" of our random variable: (i.e., TP53 expression in single cells)



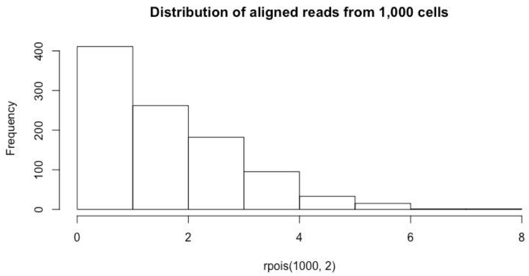
Single cell sequencing is expensive:
What if we can only count reads at TP53 from 4 cells at a time? This is a size 4 sample to estimate the population



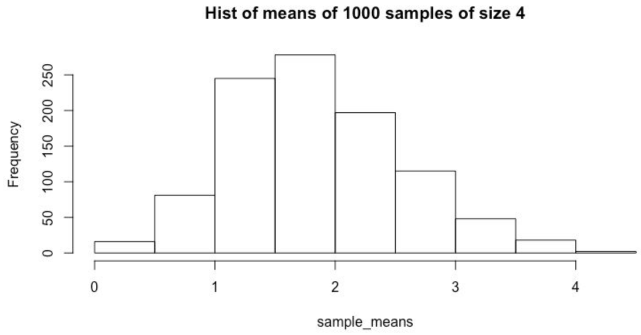
```
# sample 1
> rpois(4,2)               # sample 1 mean
[1] 4 1 4 1      ───→      2.5
# sample 2
> rpois(4,2)               # sample 2 mean
[1] 2 2 1 3      ───→      2.0
# sample 3
> rpois(4,2)               # sample 3 mean
[1] 0 0 3 0      ───→      0.75
# sample 4
> rpois(4,2)               # sample 4 mean
[1] 1 4 1 1      ───→      1.75
# sample 5
> rpois(4,2)               # sample 5 mean
[1] 0 1 0 2      ───→      0.75
```



Let's repeat this 1000 times using a sample of 4 cells.



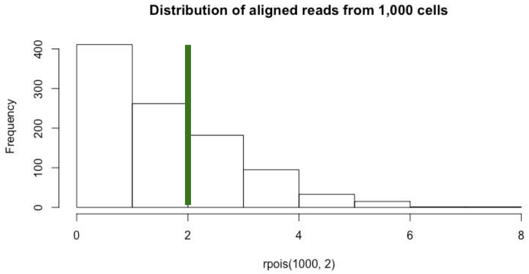
```
# sample 1          →    # sample 1 mean
> rpois(4,2)             1.5
[1] 2 2 2 0
# sample 2          →    # sample 2 mean
> rpois(4,2)             1.0
[1] 2 2 1 3

...

# sample 1000
> rpois(4,2)        →    # sample 1000 mean
[1] 0 1 0 2              1.75
```



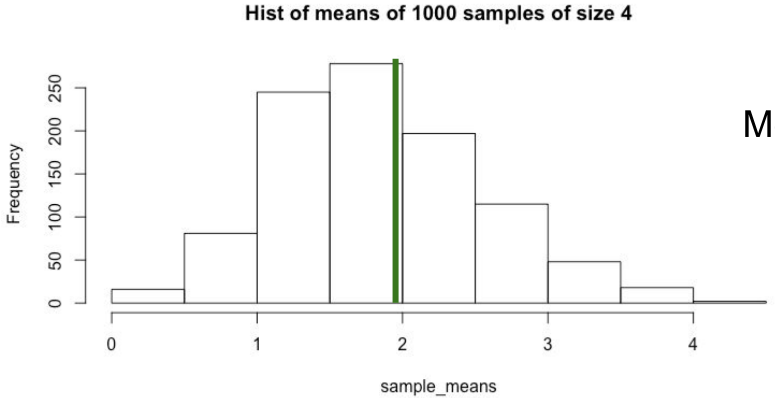
What does this distribution look like?

Central Limit Theorem:

As the sample size increases, the means of samples will become increasingly close to a normal distribution with a mean (***μ***) equal to the mean of the population!



Population mean = 2

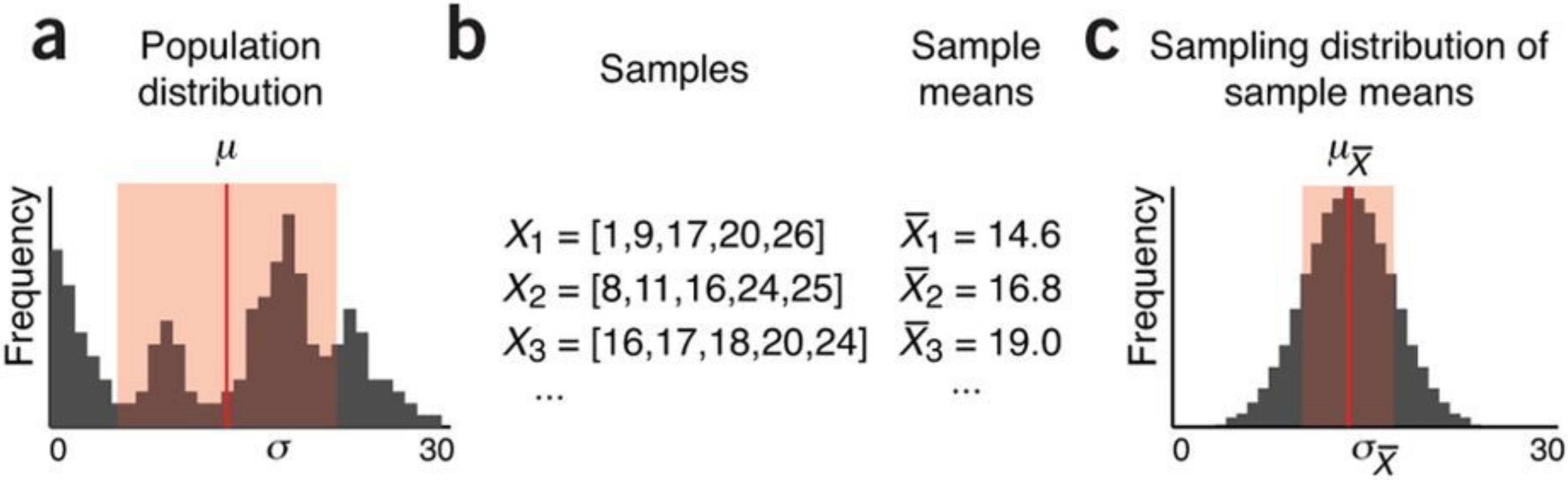


Mean of the sample means = 1.97

Sampling distribution simulator

https://onlinestatbook.com/stat_sim/sampling_dist/

Central Limit Theorem holds true for any distribution



$X_1 = [1,9,17,20,26]$ $\overline{X}_1 = 14.6$

$X_2 = [8,11,16,24,25]$ $\overline{X}_2 = 16.8$

$X_3 = [16,17,18,20,24]$ $\overline{X}_3 = 19.0$

... ...

"Just like the population, the sampling distribution [of sample means] is not directly measurable because we do not have access to all possible samples. However, it turns out to be an extremely useful concept in the process of estimating population statistics."

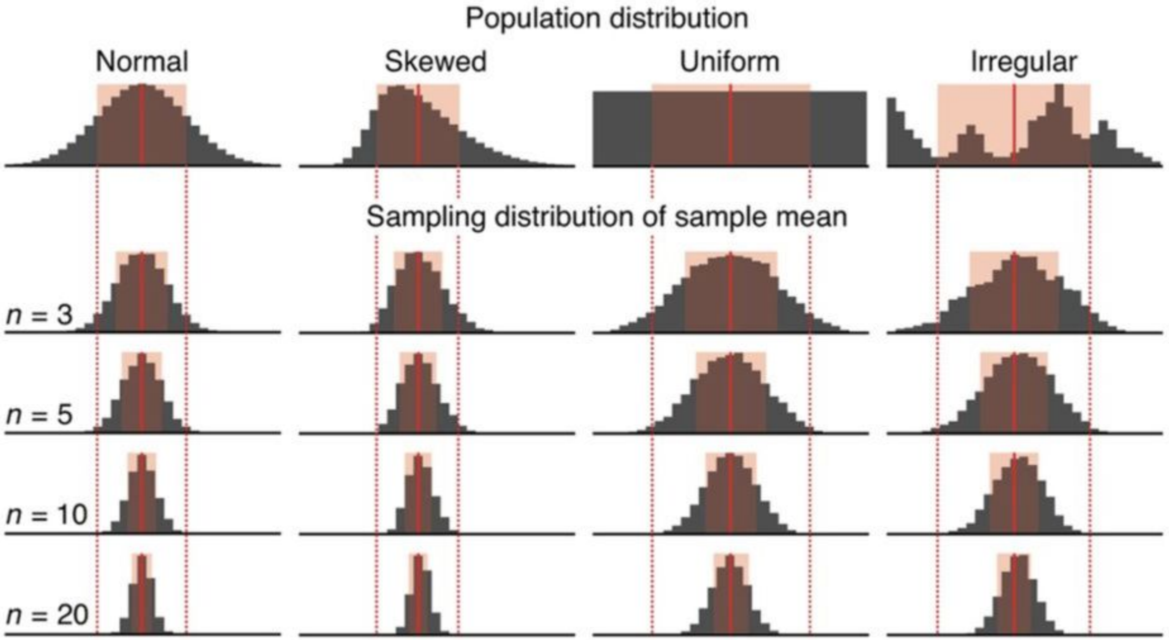
https://www.nature.com/articles/nmeth.2613/figures/2

Why do we care?

- Knowing that the sample means will always be normally- distributed means (ha!) that we don't need to know the properties of underlying population distribution.
- This allows us to compute confidence interval
- Run t-tests between samples of two different populations
- Run ANOVAs between samples of three or more different populations

https://www.nature.com/articles/nmeth.2613/figures/2

With larger sample sizes (*n*), the standard deviation of the sample means decreases



Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in Figure 1.

https://www.nature.com/articles/nmeth.2613/figures/3

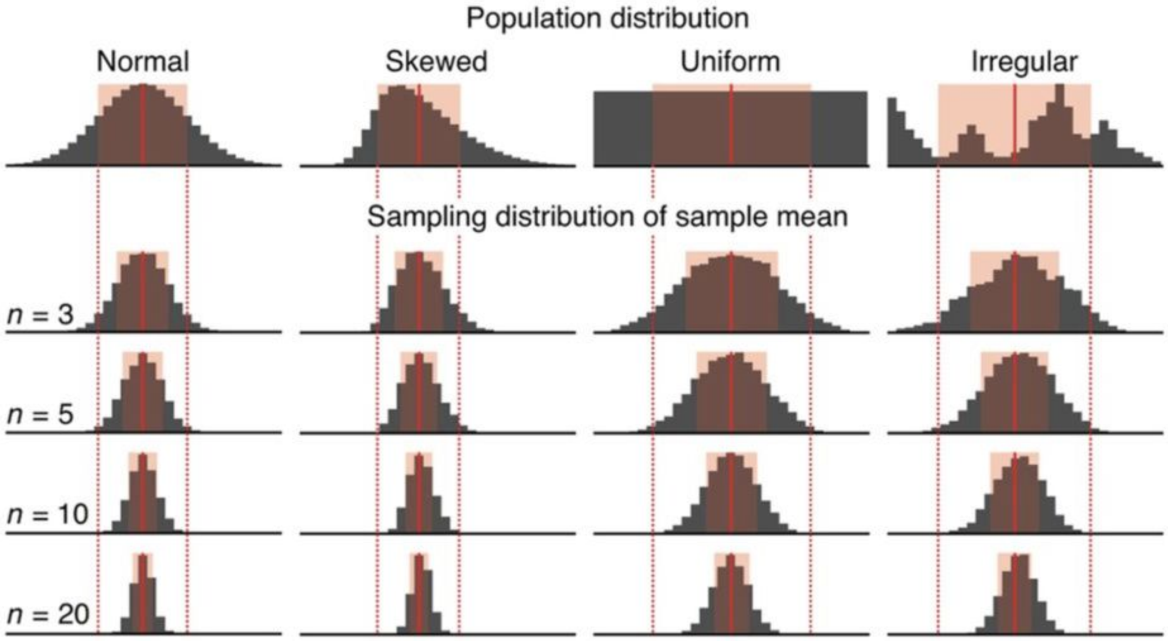
Standard deviation of the sample means

$$\sigma\bar{X} = \sigma/\sqrt{n}$$

Standard deviation of the population

As *n* increases, σx̄ decreases. That is, the samples will have more similar means. Most importantly, when using a larger n, **your sample mean is much more likely to be close to the true population mean. This is important in biology because we typically do one sample of size n (i.e., n replicates).**

Be wary of small sample sizes. They can cause many of the sample means to vary substantially from the true population mean

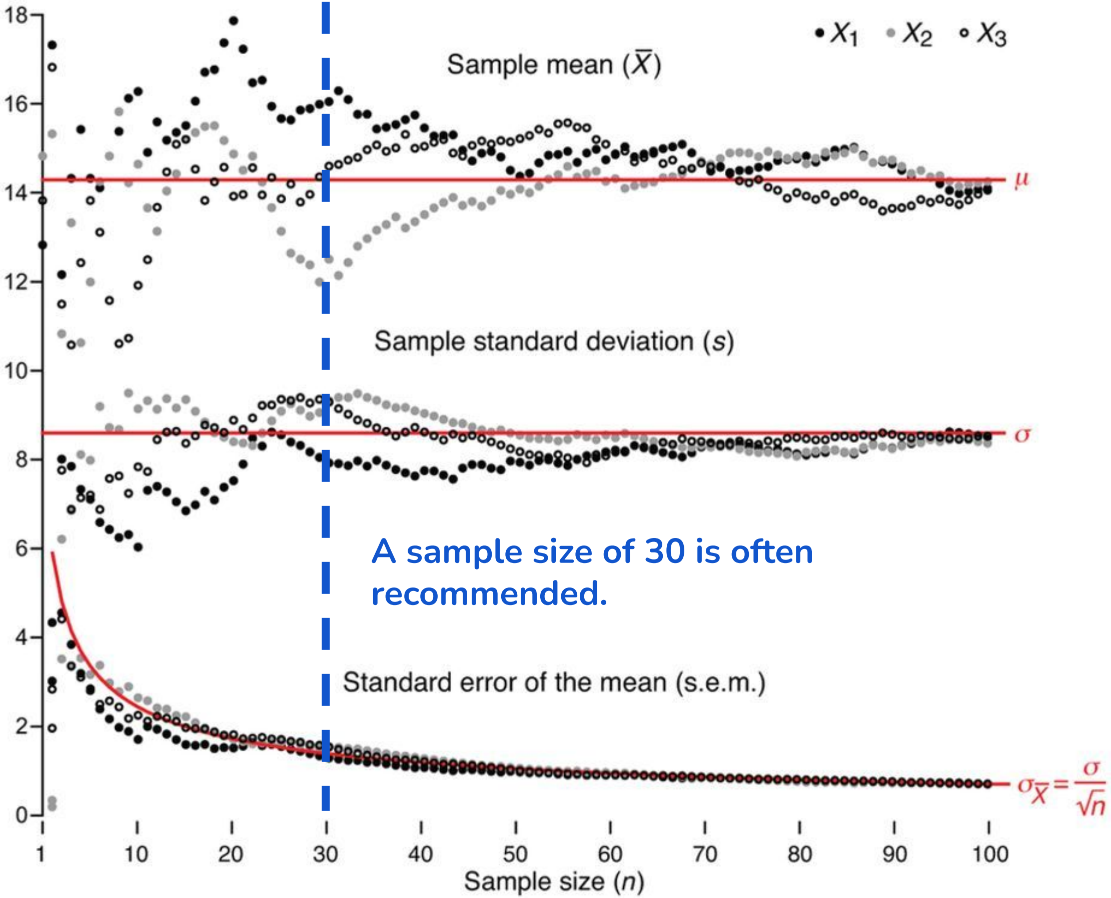


Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in Figure 1.

"It is still possible for a sample mean to fall far from the population mean, especially for small *n*. **For example, in ten iterations of drawing 10,000 samples of size *n* = 3 from the irregular distribution, the number of times the sample mean fell outside μ ± σ (indicated by vertical dotted lines) ranged from 7.6% to 8.6%.**

Thus, use caution when interpreting means of small samples."

Samples better approximate population as *n* increases.

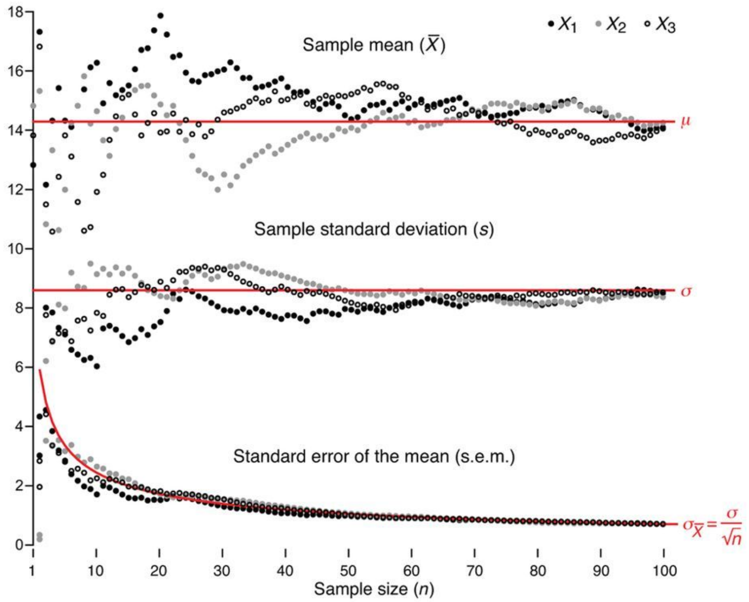


A sample size of 30 is often recommended.

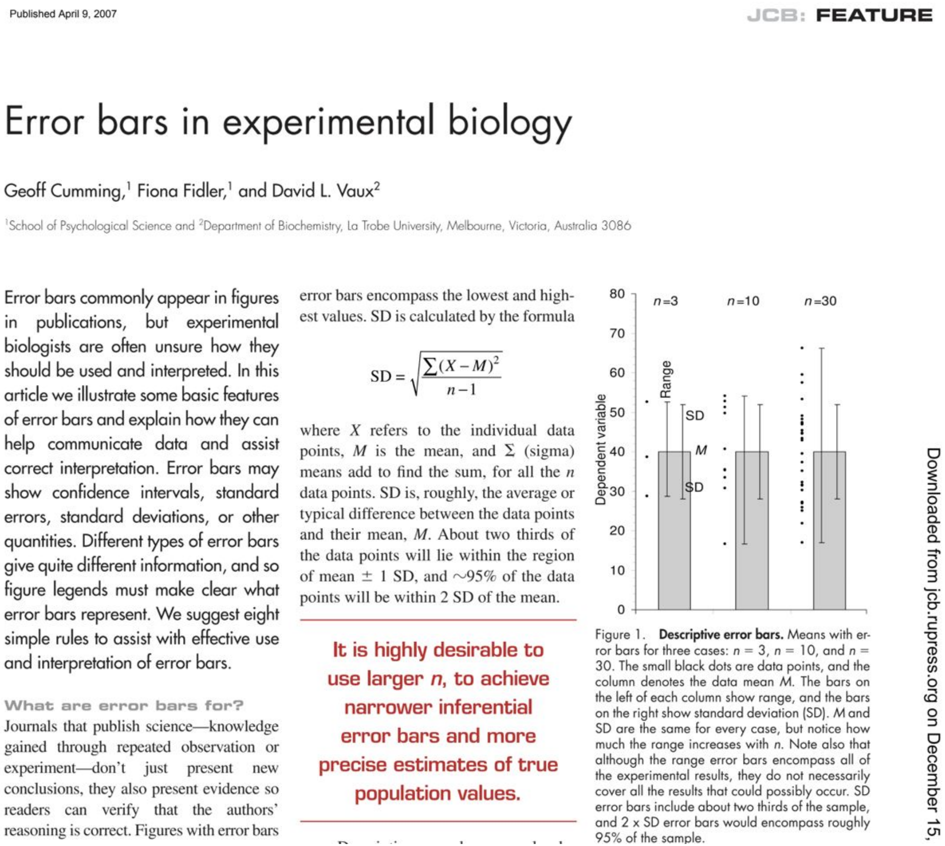
The measured spread of sample means is also known as the standard error of the mean (s.e.m., SE$\bar{X}$) and is used to estimate σ$\bar{X}$, which we cannot know because we cannot collect all possible samples.

Estimates from samples have uncertainty.

How can we quantify the degree to which a (random) sample's mean and standard deviation is a good representation of the true population's mean and standard deviation?



Error bars are commonly used, misused, and misunderstood



Published April 9, 2007

JCB: FEATURE

Error bars in experimental biology

Geoff Cumming,[1] Fiona Fidler,[1] and David L. Vaux[2]

[1]School of Psychological Science and [2]Department of Biochemistry, La Trobe University, Melbourne, Victoria, Australia 3086

Error bars commonly appear in figures in publications, but experimental biologists are often unsure how they should be used and interpreted. In this article we illustrate some basic features of error bars and explain how they can help communicate data and assist correct interpretation. Error bars may show confidence intervals, standard errors, standard deviations, or other quantities. Different types of error bars give quite different information, and so figure legends must make clear what error bars represent. We suggest eight simple rules to assist with effective use and interpretation of error bars.

What are error bars for?

Journals that publish science—knowledge gained through repeated observation or experiment—don't just present new conclusions, they also present evidence so readers can verify that the authors' reasoning is correct. Figures with error bars

error bars encompass the lowest and highest values. SD is calculated by the formula

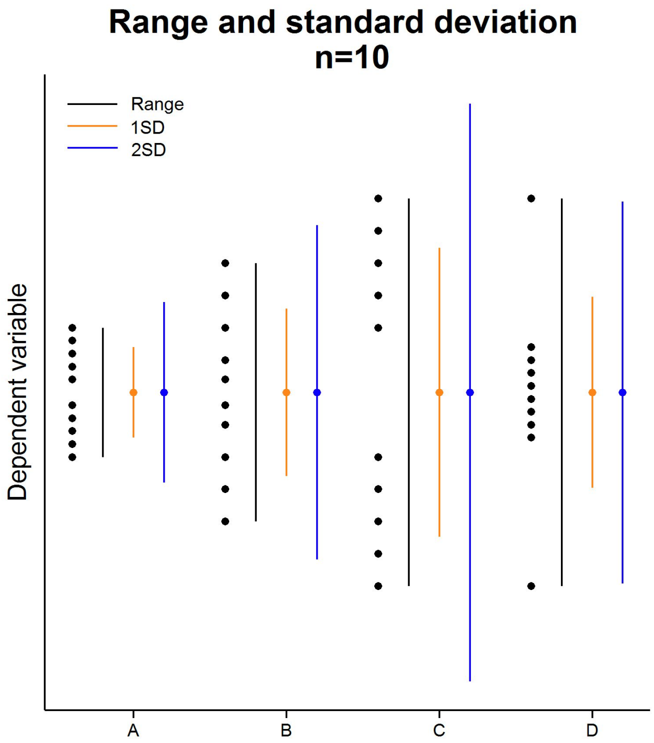
$$\mathrm{SD} = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$$

where X refers to the individual data points, M is the mean, and Σ (sigma) means add to find the sum, for all the n data points. SD is, roughly, the average or typical difference between the data points and their mean, M. About two thirds of the data points will lie within the region of mean ± 1 SD, and ~95% of the data points will be within 2 SD of the mean.

It is highly desirable to use larger *n*, to achieve narrower inferential error bars and more precise estimates of true population values.

Figure 1. **Descriptive error bars.** Means with error bars for three cases: n = 3, n = 10, and n = 30. The small black dots are data points, and the column denotes the data mean M. The bars on the left of each column show range, and the bars on the right show standard deviation (SD). M and SD are the same for every case, but notice how much the range increases with n. Note also that although the range error bars encompass all of the experimental results, they do not necessarily cover all the results that could possibly occur. SD error bars include about two thirds of the sample, and 2 x SD error bars would encompass roughly 95% of the sample.

Two types of error bars: descriptive and inferential



Descriptive error bars

- Meant to give show the “spread” of the data
- Range (min to max value)
- Standard deviation (SD)
- Useful for asking whether a single result fits within a normal range (e.g., cholesterol levels)
- Not useful for comparison of conditions or groups

Two types of error bars: descriptive and inferential

In biology, we typically want to compare samples from different groups or experimental conditions (e.g., wild-type to mutant, experimental versus control). In order to make inferences and convey whether the groups are **significantly different beyond what could be expected by random chance**, we should use **inferential error bars**

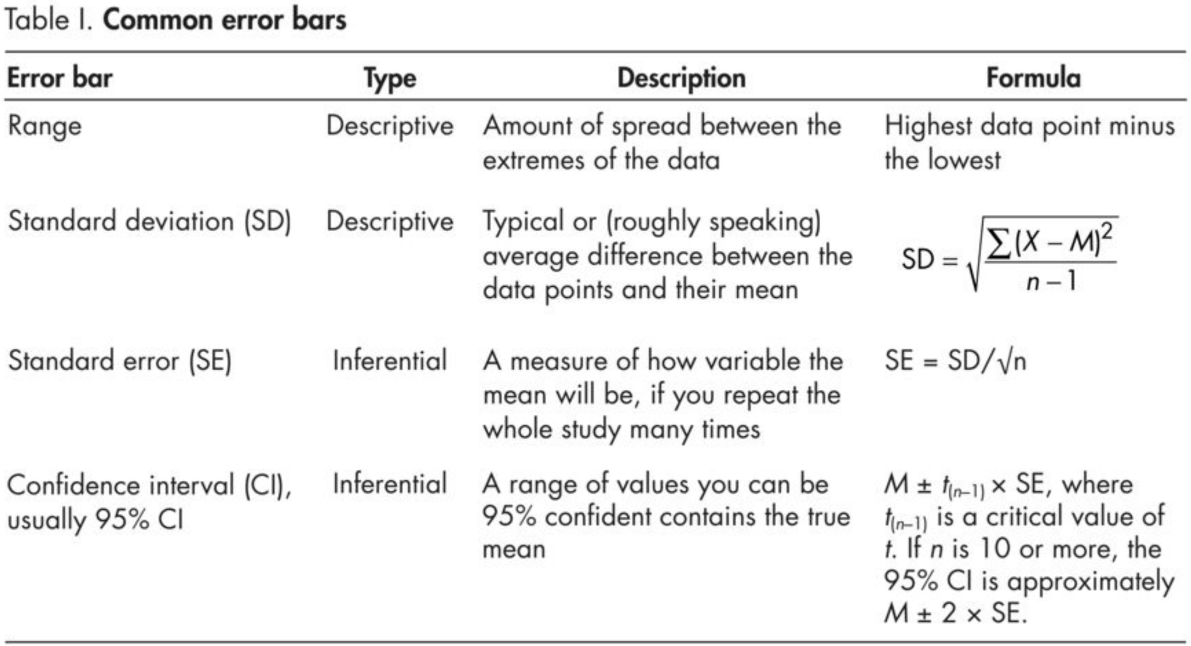


Table I. **Common error bars**

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$\mathrm{SD} = \sqrt{\dfrac{\sum (X - M)^2}{n-1}}$
Standard error (SE)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$\mathrm{SE} = \mathrm{SD}/\sqrt{n}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{(n-1)} \times \mathrm{SE}$, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times \mathrm{SE}$.

SE and CI give a sense of where the mean of the complete population should lie

Rule #1: Legends must state type of error bar

Because error bars can be descriptive or inferential, and could be any of the bars below or even something else, they are meaningless, or misleading, if the legend does not state what kind they are.

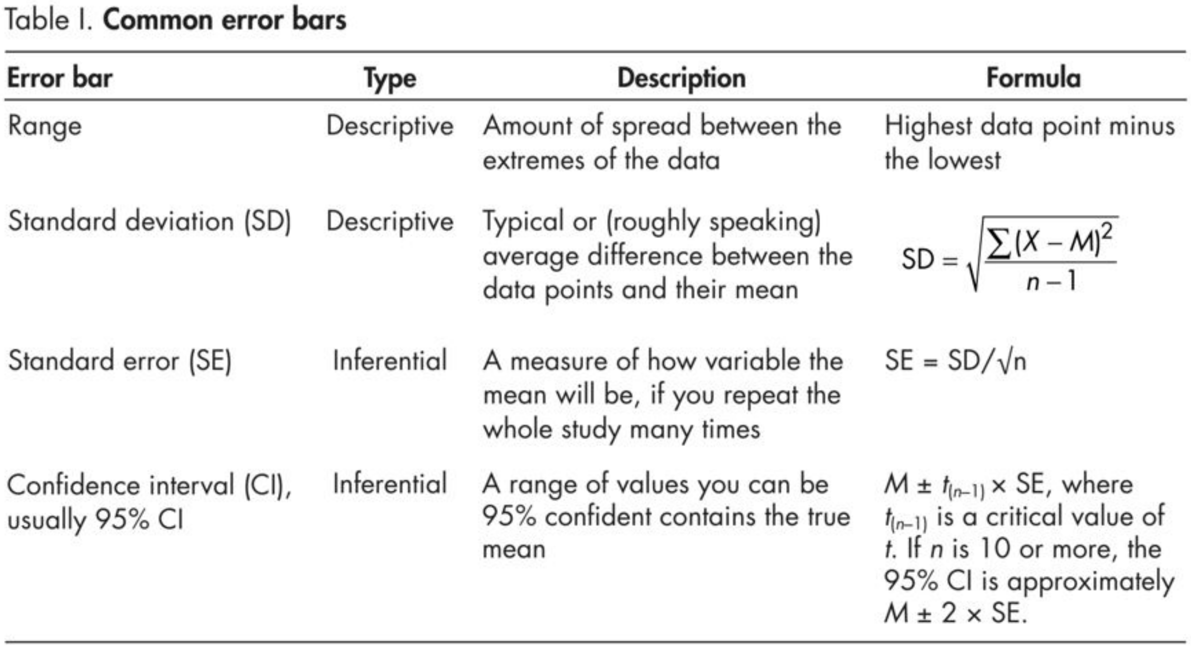


Table I. **Common error bars**

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$\mathrm{SD} = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$
Standard error (SE)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$\mathrm{SE} = \mathrm{SD}/\sqrt{n}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{(n-1)} \times \mathrm{SE}$, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times \mathrm{SE}$.

Confidence Intervals as error bars

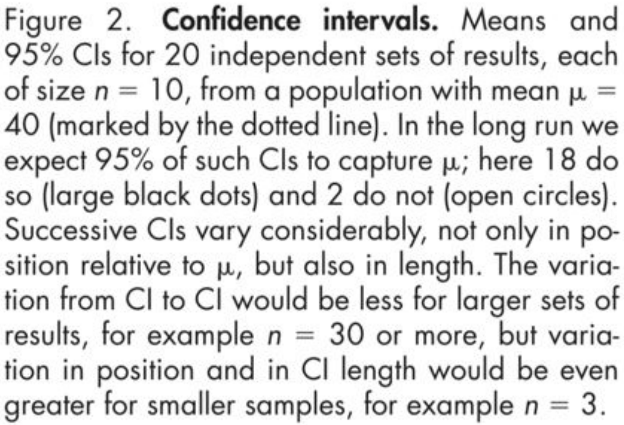
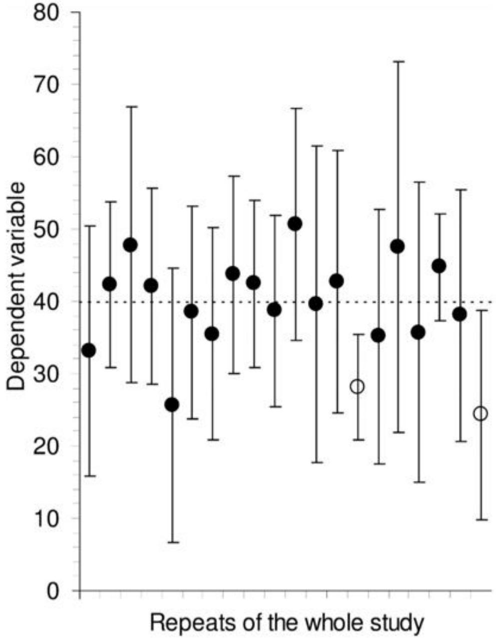


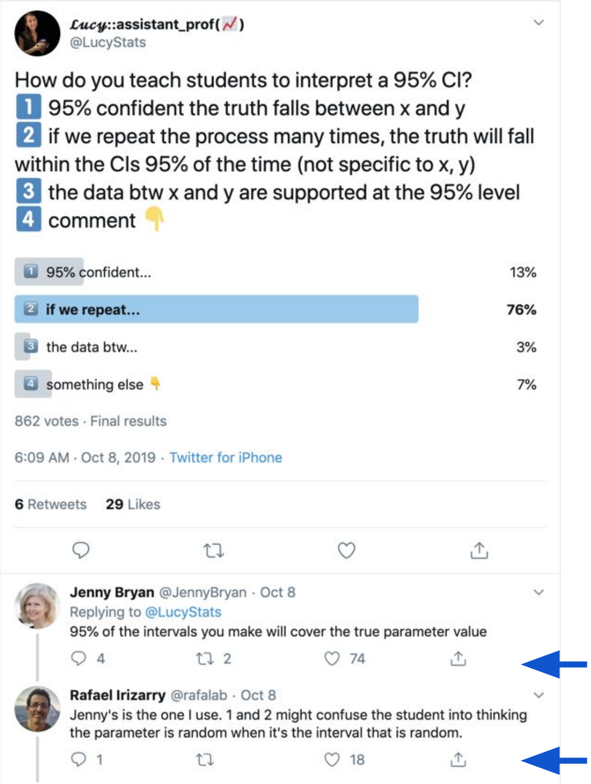
Figure 2. **Confidence intervals.** Means and 95% CIs for 20 independent sets of results, each of size $n = 10$, from a population with mean $\mu = 40$ (marked by the dotted line). In the long run we expect 95% of such CIs to capture μ; here 18 do so (large black dots) and 2 do not (open circles). Successive CIs vary considerably, not only in position relative to μ, but also in length. The variation from CI to CI would be less for larger sets of results, for example $n = 30$ or more, but variation in position and in CI length would be even greater for smaller samples, for example $n = 3$.



In 20 repetitions of the study, the true population mean (dashed line) fell outside of the 95% CI twice. In the long run (that is, if we did this hundreds or thousands of times, it should fall outside of the CI 5% of the time

A big advantage of inferential error bars is that their length gives a graphic signal of how much uncertainty there is in the data: **We are asserting that the 95% of the intervals we make will cover the true population mean (μ)**. Wide inferential bars indicate large error; short inferential bars indicate high precision.

Confidence Intervals as error bars



𝓛𝓾𝓬𝔂::assistant_prof(📈)
@LucyStats

How do you teach students to interpret a 95% CI?
1️⃣ 95% confident the truth falls between x and y
2️⃣ if we repeat the process many times, the truth will fall within the CIs 95% of the time (not specific to x, y)
3️⃣ the data btw x and y are supported at the 95% level
4️⃣ comment 👇

1 95% confident...	13%
2 **if we repeat...**	**76%**
3 the data btw...	3%
4 something else 👇	7%

862 votes · Final results

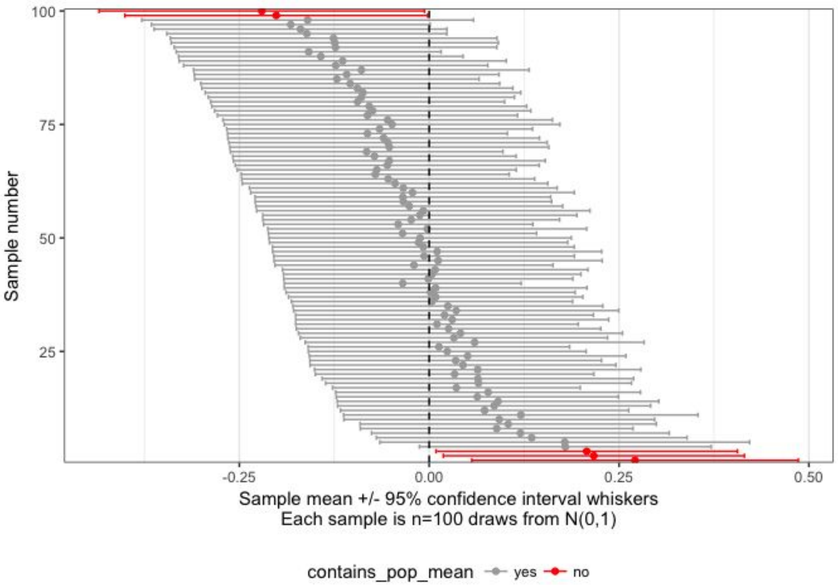
6:09 AM · Oct 8, 2019 · Twitter for iPhone

6 Retweets **29** Likes

Jenny Bryan @JennyBryan · Oct 8
Replying to @LucyStats
95% of the intervals you make will cover the true parameter value
4 2 74

Rafael Irizarry @rafalab · Oct 8
Jenny's is the one I use. 1 and 2 might confuse the student into thinking the parameter is random when it's the interval that is random.
1 18

Code to generate CI simulation:
https://github.com/leonjessen/confidence_intervals_visualised



Hat tip to Brent Pedersen: https://twitter.com/LucyStats/status/1181542102779531264

Rule #2: The value of n (i.e., the sample size, or the number of independently performed experiments) must be stated in the figure legend.

What is n?

- N is the number of independent experiments, not the number of replicates
- Example: you select one mutant and one wild type mouse and perform 10 experiments in replicate on each of their tails.
- The mean and SD of the replicates is not sufficient for a figure, as n=1 for each mouse genotype. This design does not measure natural variation from animal to animal.
- "If an experiment involves triplicate cultures, and is repeated four independent times, then n = 4, not 3 or 12. The variation within each set of triplicates is related to the fidelity with which the replicates were created, and is irrelevant to the hypothesis being tested."

Inferential error bars shrink with larger n

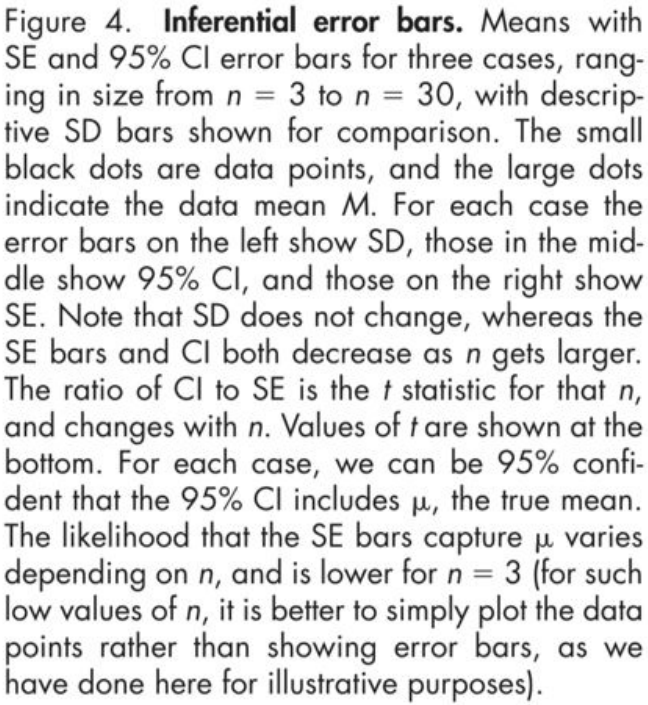
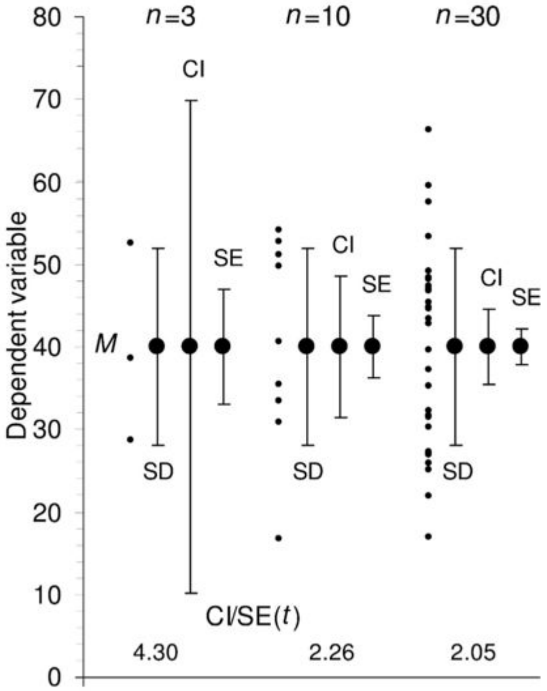


Figure 4. **Inferential error bars.** Means with SE and 95% CI error bars for three cases, ranging in size from $n = 3$ to $n = 30$, with descriptive SD bars shown for comparison. The small black dots are data points, and the large dots indicate the data mean *M*. For each case the error bars on the left show SD, those in the middle show 95% CI, and those on the right show SE. Note that SD does not change, whereas the SE bars and CI both decrease as *n* gets larger. The ratio of CI to SE is the *t* statistic for that *n*, and changes with *n*. Values of *t* are shown at the bottom. For each case, we can be 95% confident that the 95% CI includes μ, the true mean. The likelihood that the SE bars capture μ varies depending on *n*, and is lower for $n = 3$ (for such low values of *n*, it is better to simply plot the data points rather than showing error bars, as we have done here for illustrative purposes).



Rule #3: error bars and statistics should only be shown for independently repeated experiments, and never for replicates. If a "representative" experiment is shown, it should not have error bars or P values, because in such an experiment, n = 1

Estimating significance with <u>inferential</u> SE error bars

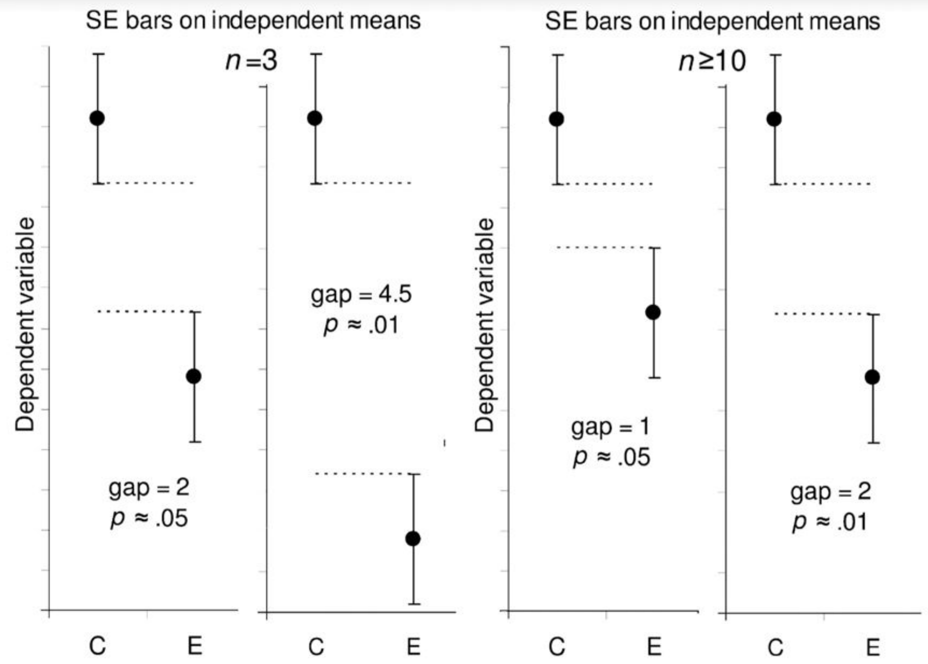


Figure 5. **Estimating statistical significance using the overlap rule for SE bars.** Here, SE bars are shown on two separate means, for control results C and experimental results E, when *n* is 3 (left) or *n* is 10 or more (right). "Gap" refers to the number of error bar arms that would fit between the bottom of the error bars on the controls and the top of the bars on the experimental results; i.e., a gap of 2 means the distance between the C and E error bars is equal to twice the average of the SEs for the two samples. When $n = 3$, and double the length of the SE error bars just touch (i.e., the gap is 2 SEs), P is ~0.05 (we don't recommend using error bars where $n = 3$ or some other very small value, but we include rules to help the reader interpret such figures, which are common in experimental biology).

The "Gap" refers to the number of error bar lengths that separate the two conditions.

When N=3 and Gap between the control and experimental SE error bars is >= 2, then P~0.05.

However, when N>=10, a gap of >=2 yields P~0.01.

Estimating significance with inferential CI error bars

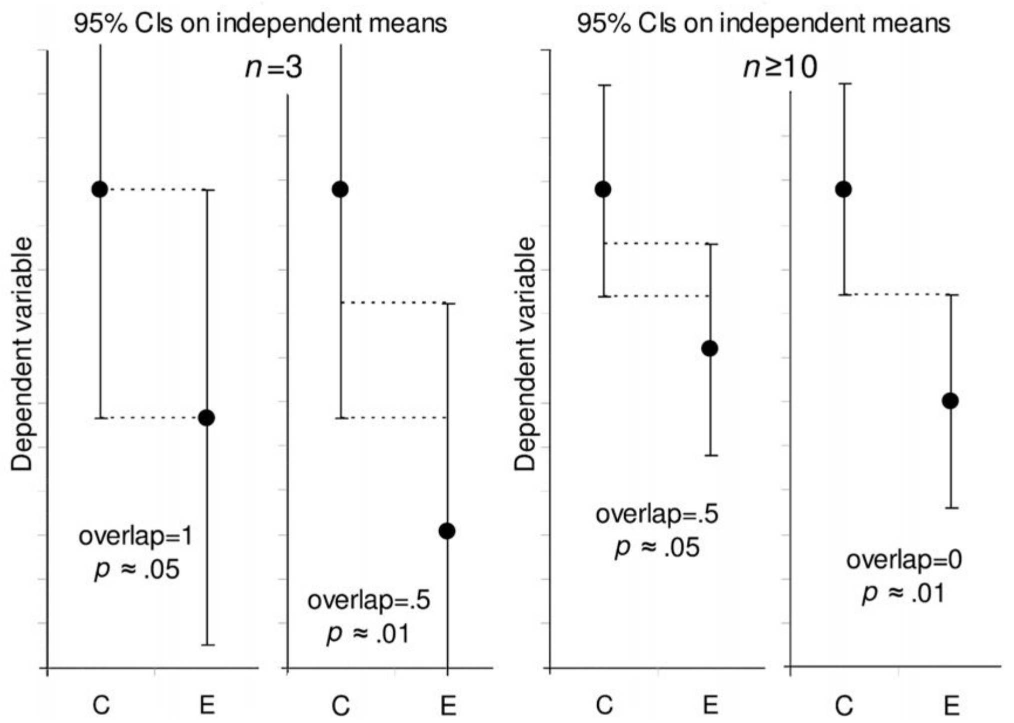


Figure 6. **Estimating statistical significance using the overlap rule for 95% CI bars.** Here, 95% CI bars are shown on two separate means, for control results C and experimental results E, when *n* is 3 (left) or *n* is 10 or more (right). "Overlap" refers to the fraction of the average CI error bar arm, i.e., the average of the control (C) and experimental (E) arms. When $n \geq 10$, if CI error bars overlap by half the average arm length, $P \approx 0.05$. If the tips of the error bars just touch, $P \approx 0.01$.

When n=3 and overlap between the control and experimental SE error bars is 1, then P~0.05.

However, when n>=10, an overlap of 0.5 yields P~0.05.

When using CI and n>=10, error bars that touch (overlap = 0), p~0.01.

Know the type of error bar to know how to interpret the error bar overlap.

Overall recommendation for comparison / inference. Use confidence intervals for error bars.

“Determining CIs requires slightly more calculating by the authors of a paper, but for people reading it, **CIs make things easier to understand, as they mean the same thing regardless of n**. For this reason, in medicine, CIs have been recommended for more than 20 years, and are required by many journals.”