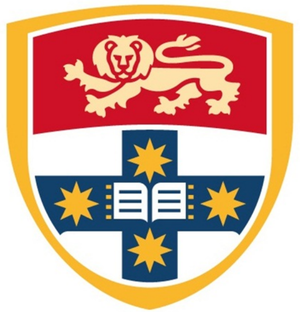
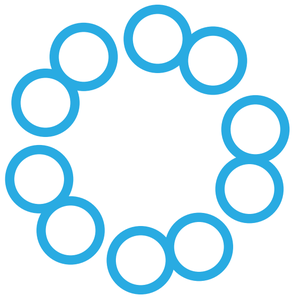


THE UNIVERSITY OF
SYDNEY



CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

# ESSENTIALS OF DATA VISUALIZATION

## THINKING ABOUT DRAWING DATA + COMMUNICATING SCIENCE

# UNCERTAINTY

don’t make errors in error bars

Knowing the limits of your knowledge is very important.

I’m sure that you’ve already encountered concepts of accuracy and precision. In colloquial use these words are often confused.

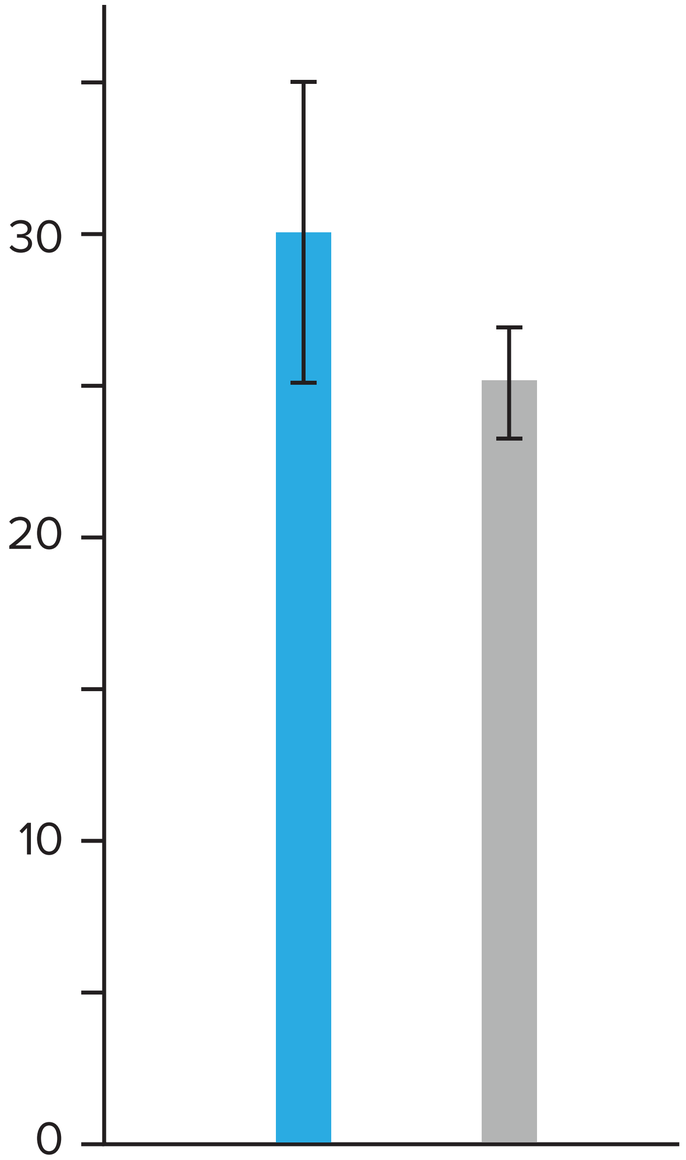
Accuracy measures how close you are to the true value. Unless you know the true value you may actually not know you accuracy at all!

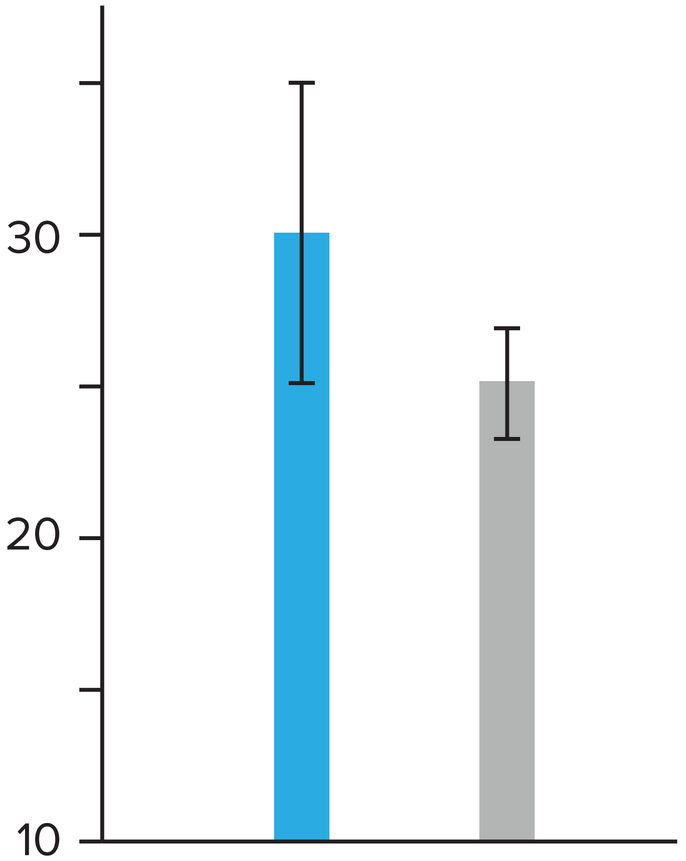
Precision measures the spread in repeated measurements. Experiments should be reproducible and, as such, measurements should have high precision.

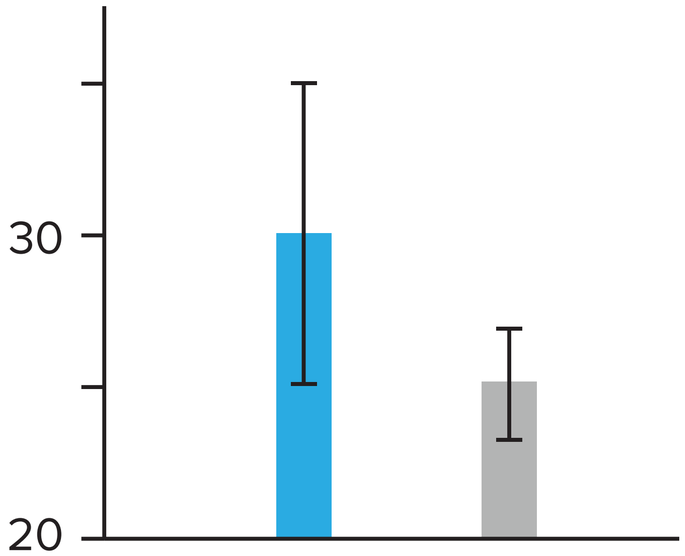
Often the term “variation” is used for this spread and connected closely to the statistical concept of variance. In an experiment you typically have many sources of variation—biological and technical—and it’s important to understand how both impact your measurements.

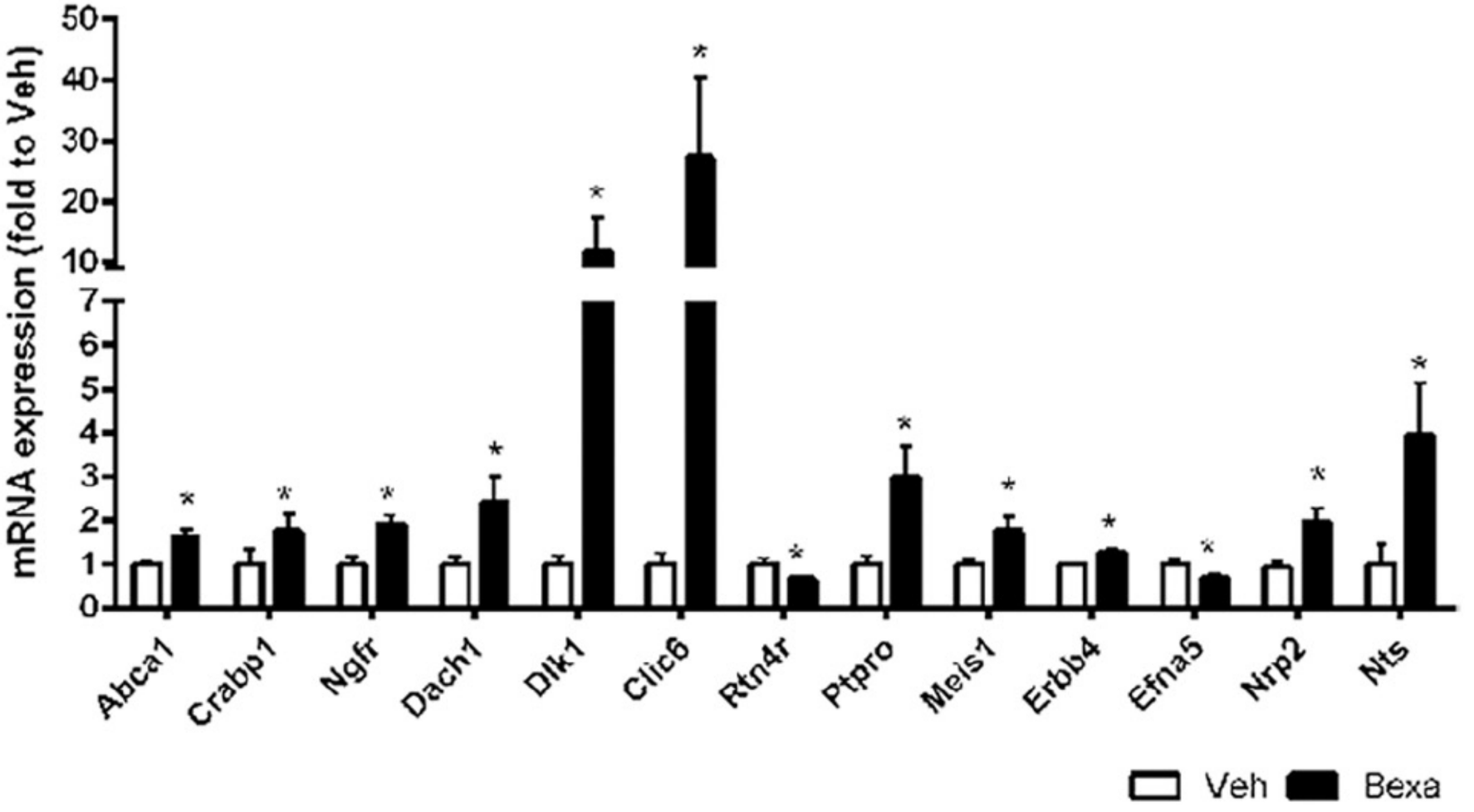
In biology, it’s important to be able to sample the extent of biological variation. And so being able to show this and other forms of variation in measurements or any computed values in visualizations is very important—it addresses reproducibility and your capacity to make statistical inference.

Often this is done with error bars. Ironically, there’s a lot of error associated with the use of and interpretation of error bars.

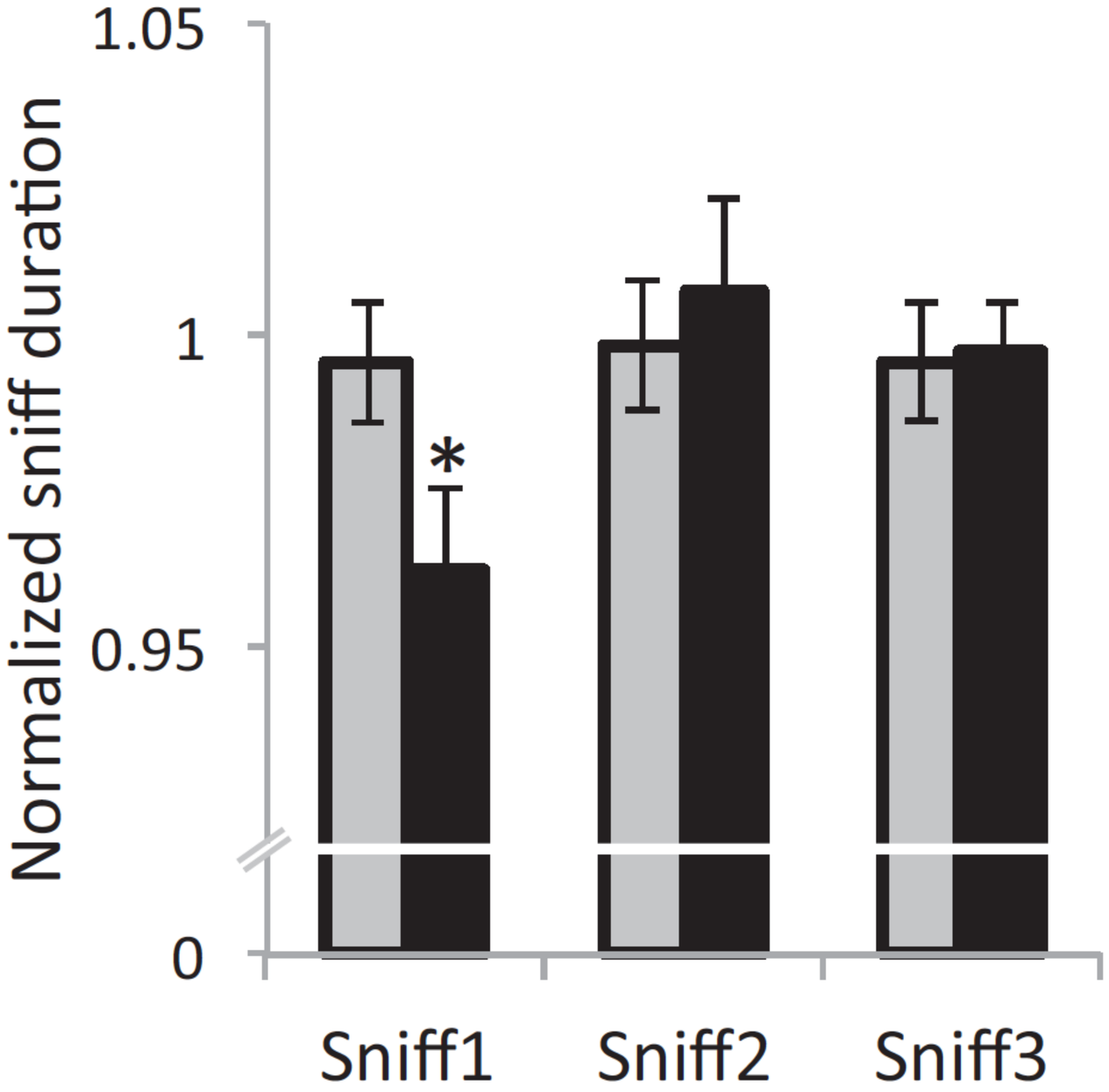

30
20
10
0


30
20
10

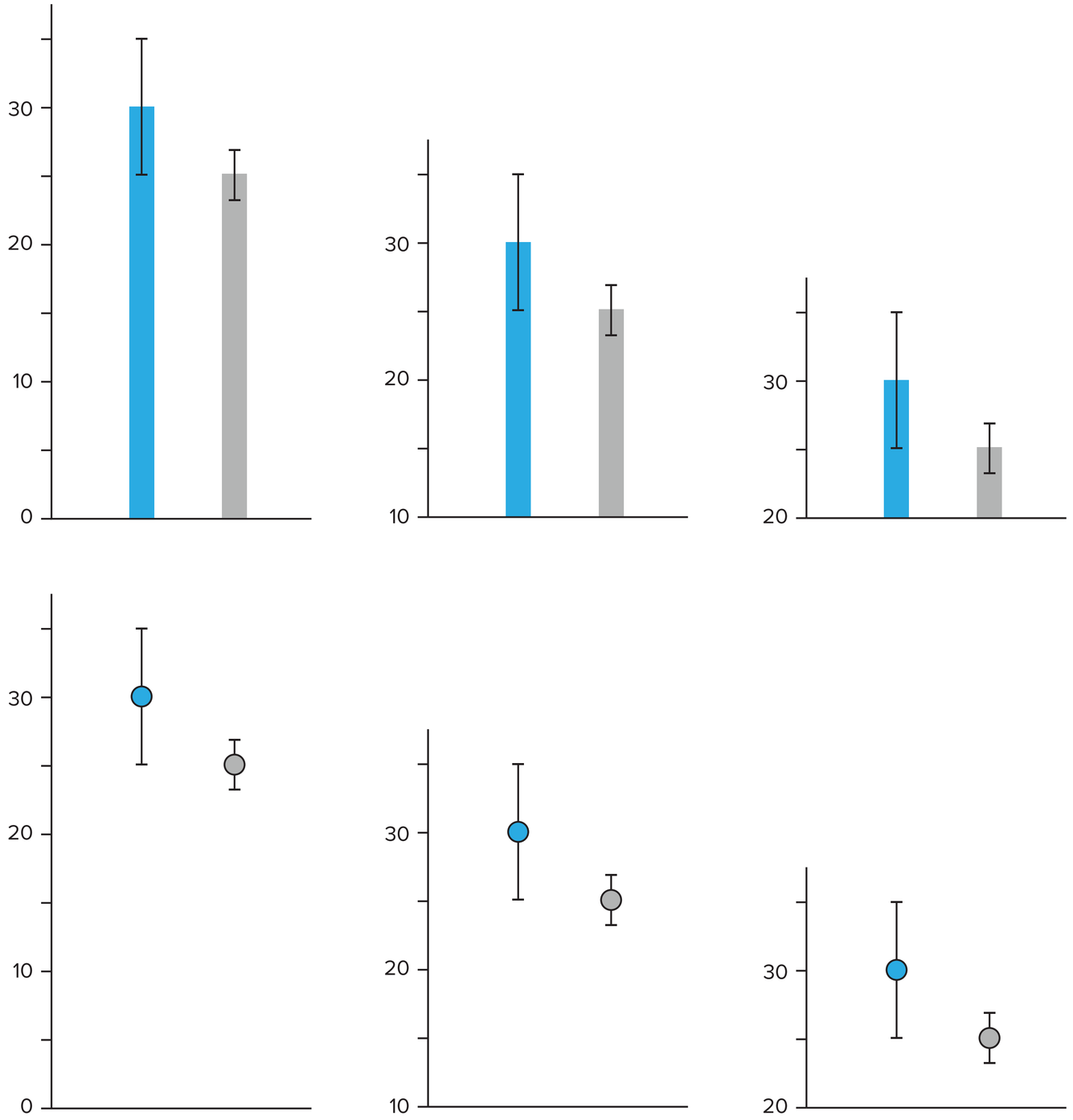

30
20

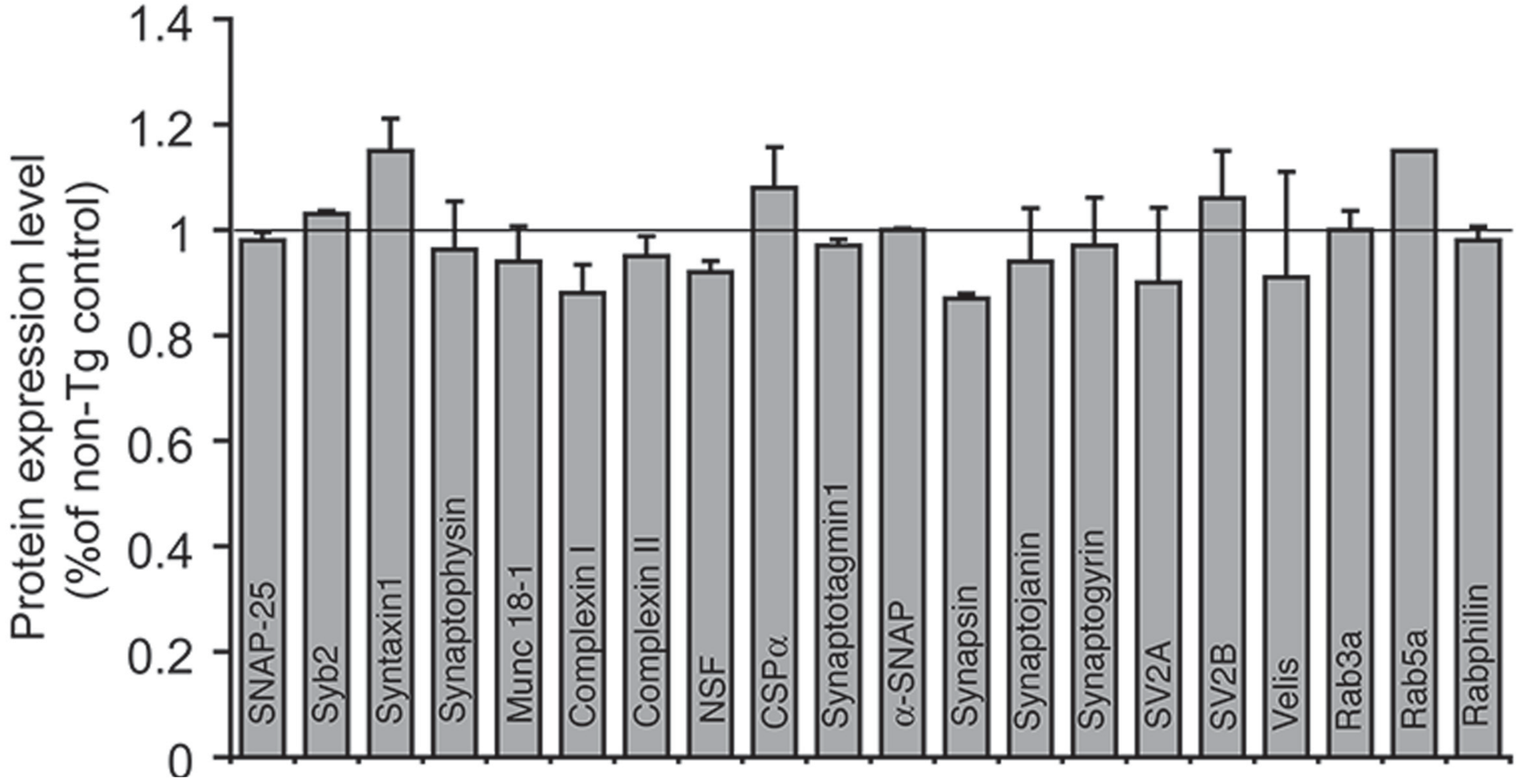



J Neurosci (2015) 35:11862–11876.

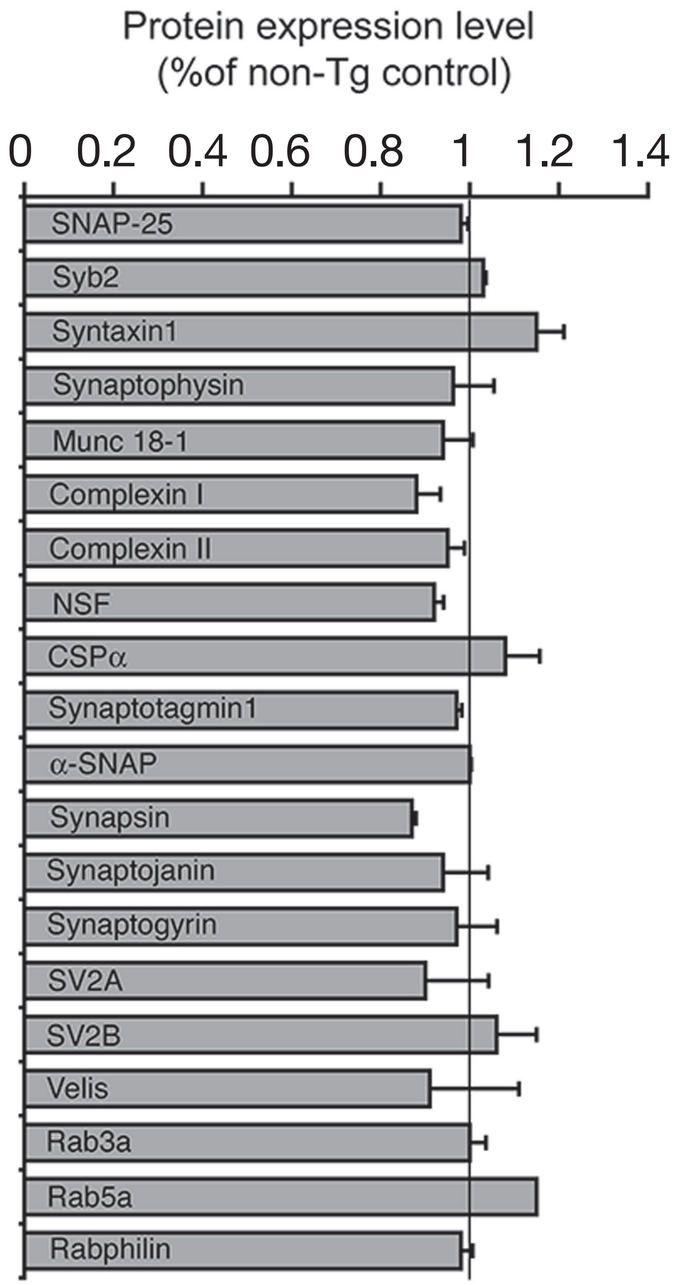


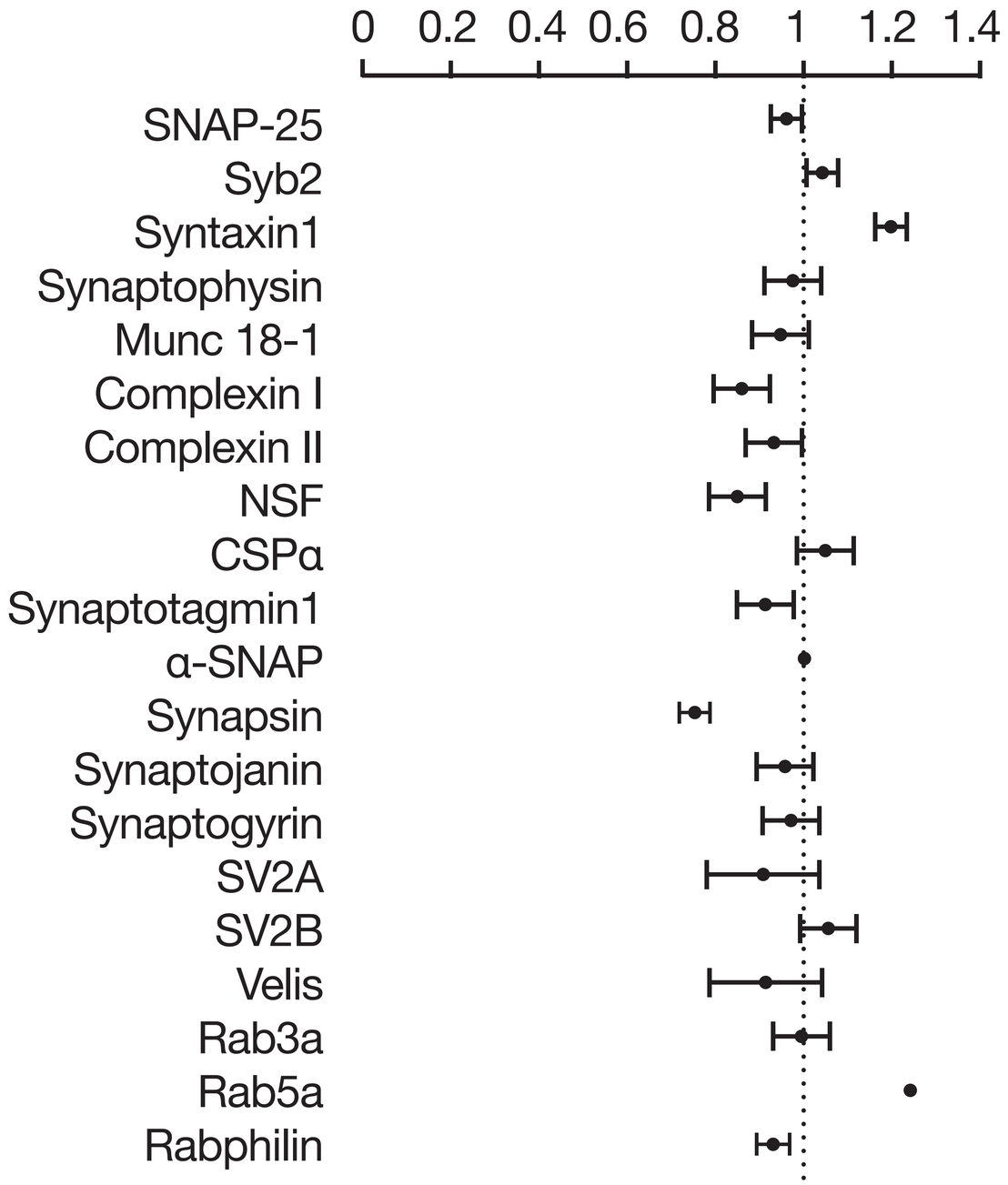

J Neurosci (2014) 34:15382–15393.

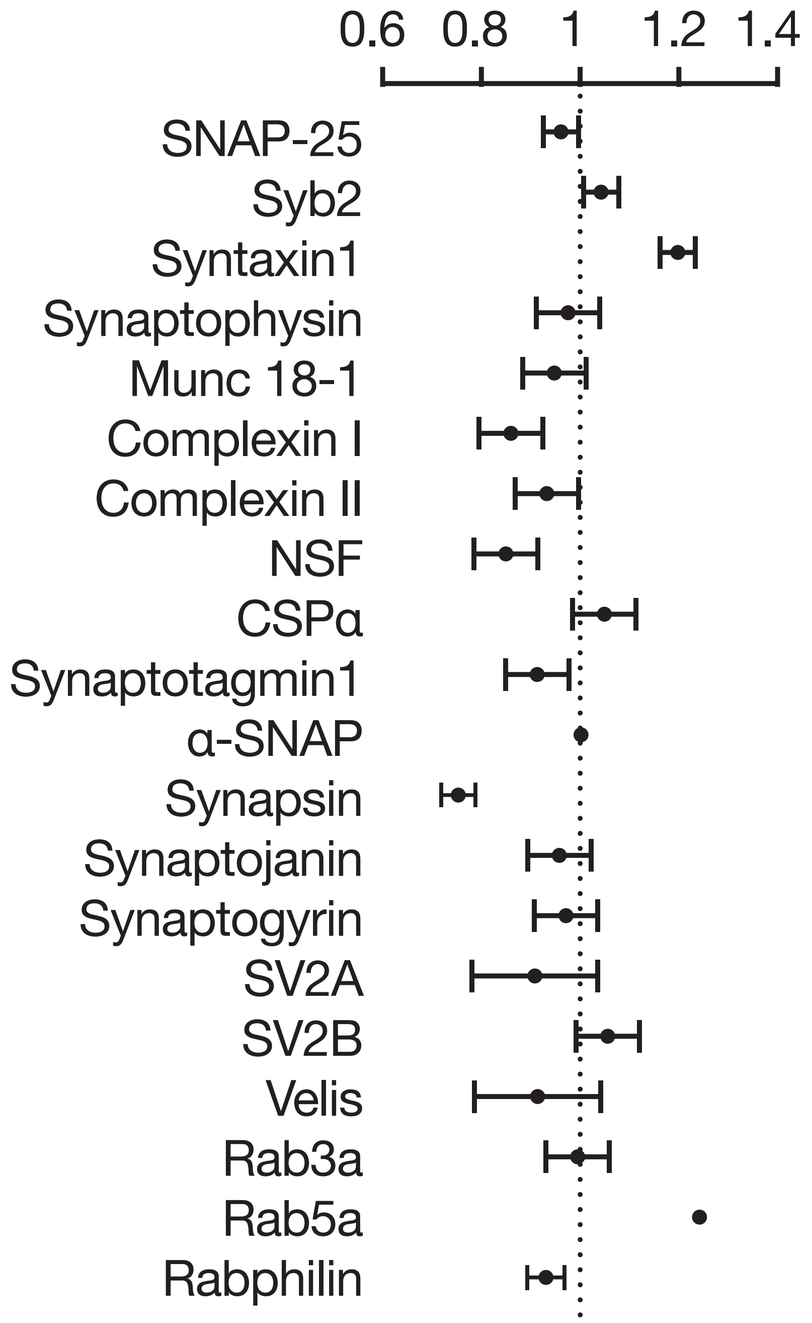

30
20
10
0
30
20
10
30
20
30
20
10
0
30
20
10
30
20




J Neurosci (2015) 35:11514 –11531.


Protein expression level
(%of non-Tg control)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
SNAP-25
Syb2
Syntaxin1
Synaptophysin
Munc 18-1
Complexin I
Complexin II
NSF
CSPα
Synaptotagmin1
α-SNAP
Synapsin
Synaptojanin
Synaptogyrin
SV2A
SV2B
Velis
Rab3a
Rab5a
Rabphilin


0
0.2
0.4
0.6
0.8
1
1.2
1.4
SNAP-25
Syb2
Syntaxin1
Synaptophysin
Munc 18-1
Complexin I
Complexin II
NSF
CSPα
Synaptotagmin1
α-SNAP
Synapsin
Synaptojanin
Synaptogyrin
SV2A
SV2B
Velis
Rab3a
Rab5a
Rabphilin

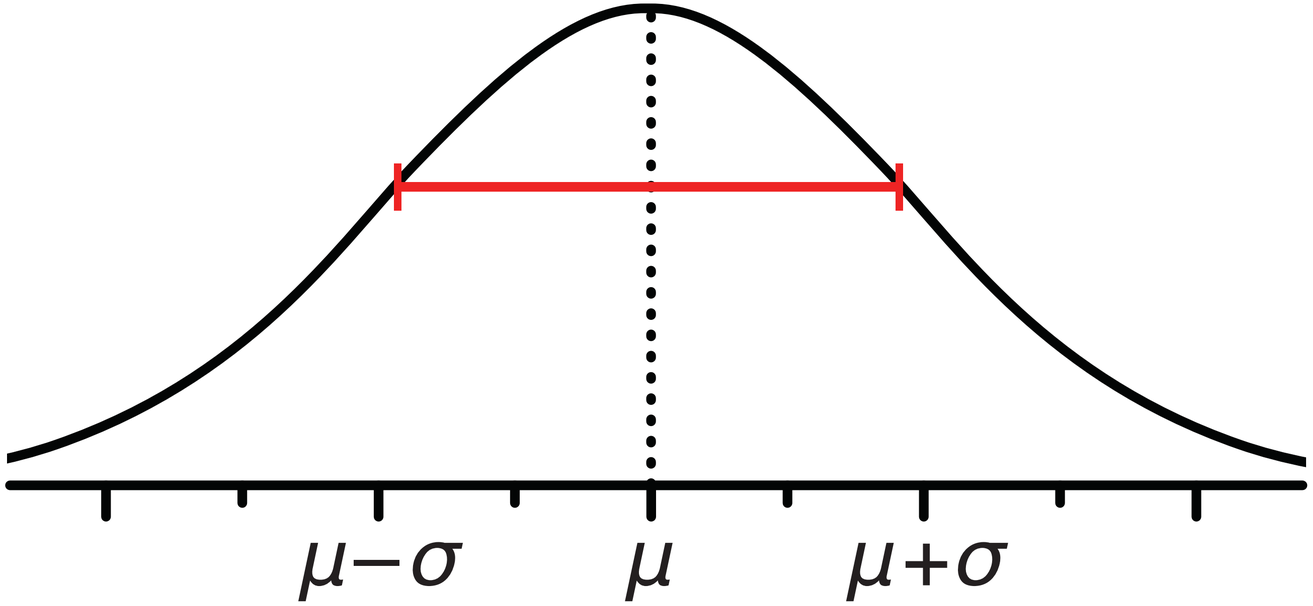

0.6
0.8
1
1.2
1.4
SNAP-25
Syb2
Syntaxin1
Synaptophysin
Munc 18-1
Complexin I
Complexin II
NSF
CSPα
Synaptotagmin1
α-SNAP
Synapsin
Synaptojanin
Synaptogyrin
SV2A
SV2B
Velis
Rab3a
Rab5a
Rabphilin

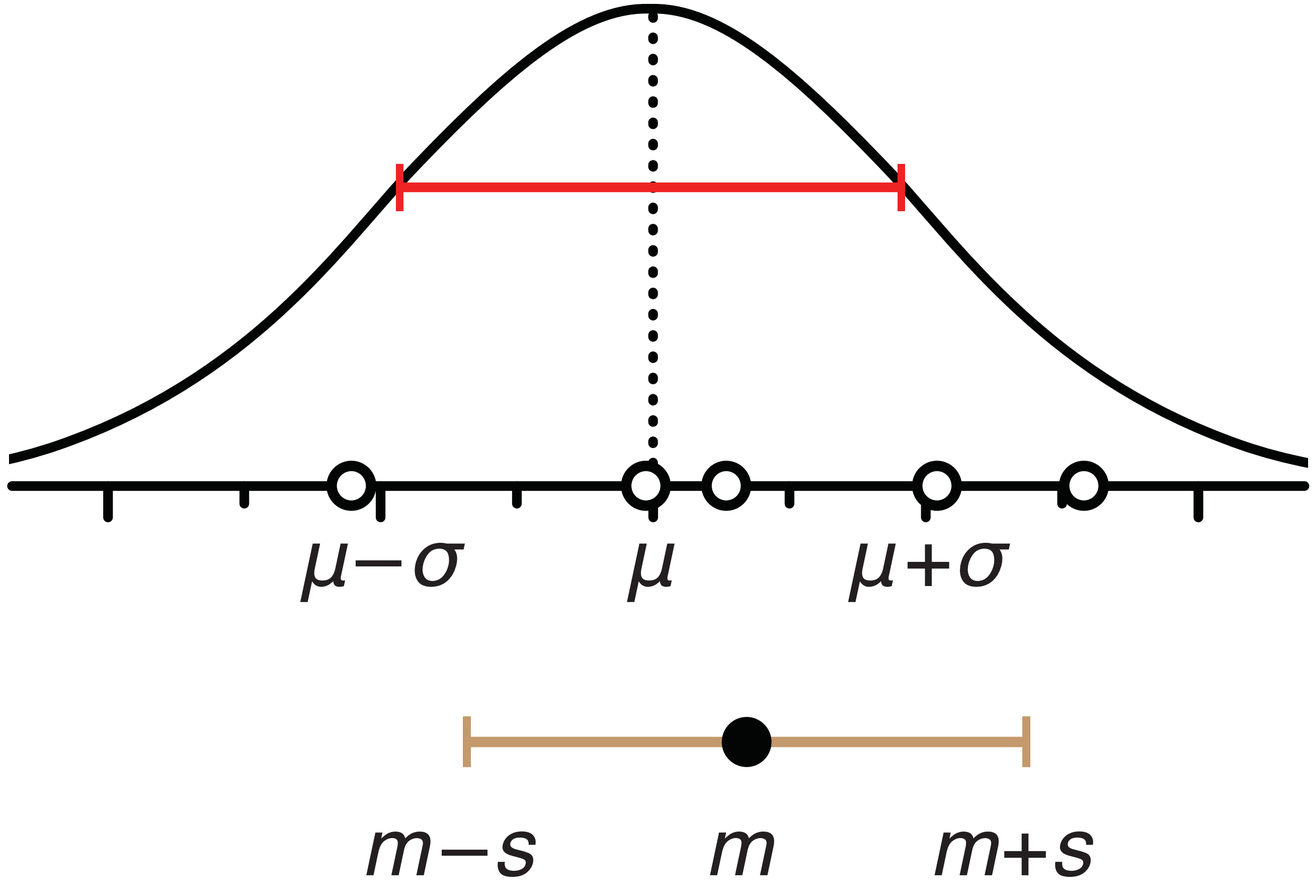
Let’s now look at what the error bars can actually represent. It’s important to realize that their meaning can vary—quite drastically.

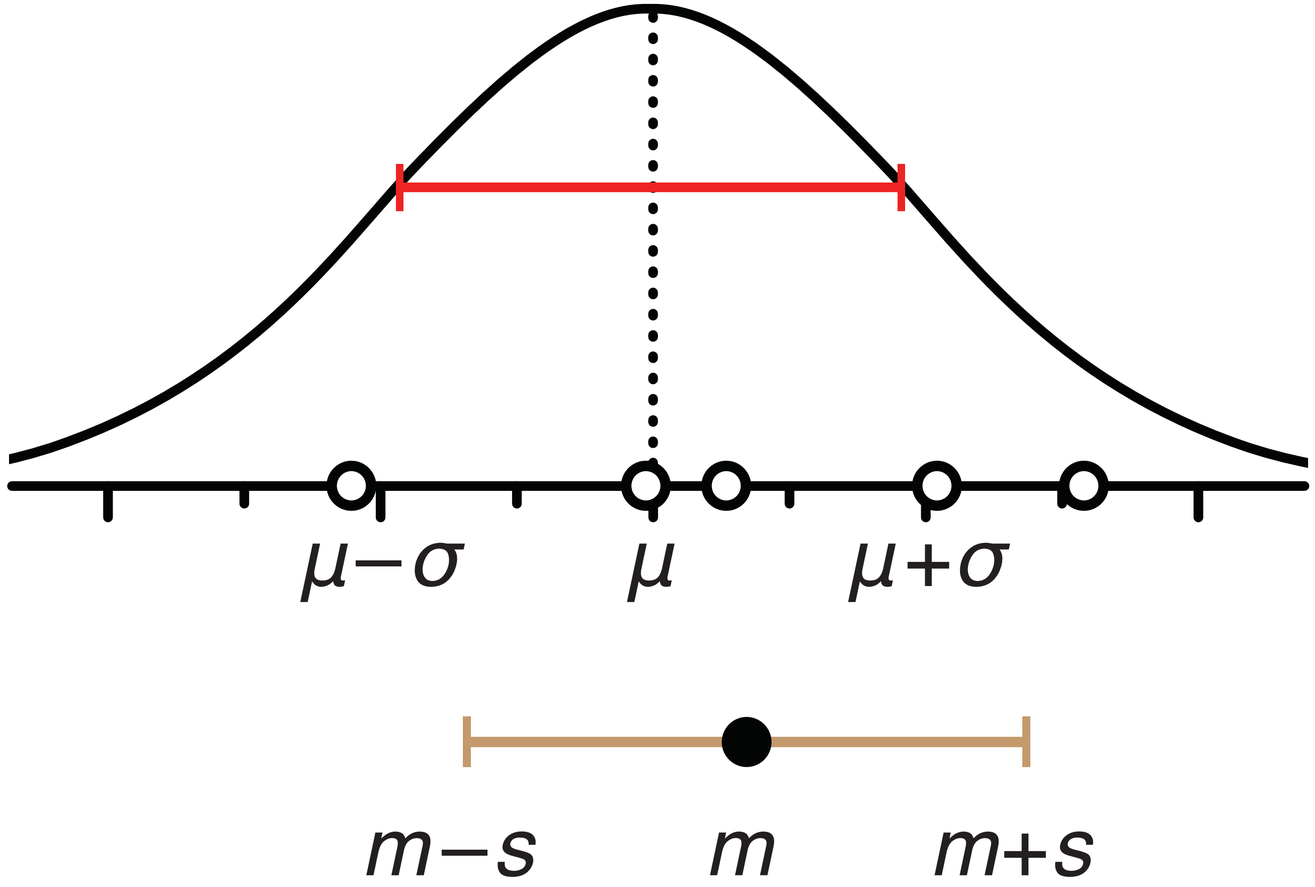
There are three quantities that it is imperative you know how to distinguish. In fact, if you remember one thing from statistics—this should be it.

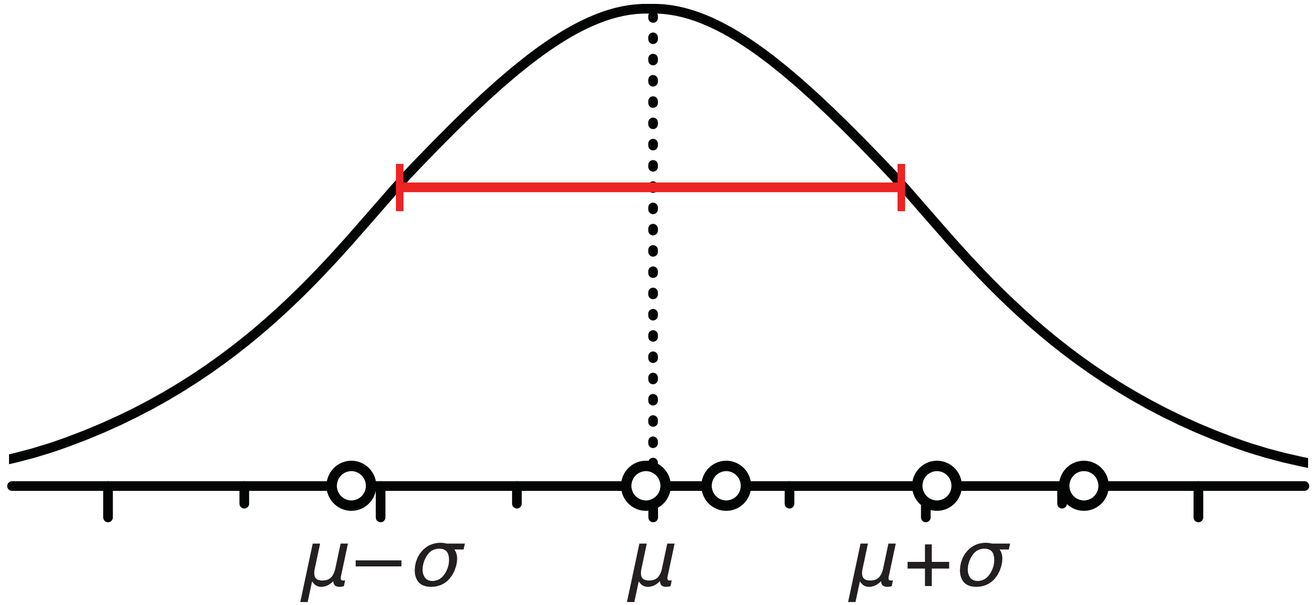
The standard deviation, the standard error of the mean and the 95% confidence interval. Not knowing the difference between these quantities is the new smoking.

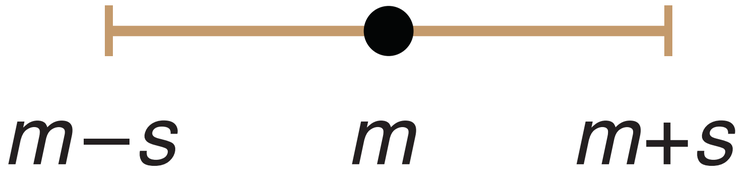
Almost all error bars are one of these three quantities. Mistaking them—especially the first two—can completely alter your perception of the data and any conclusions you make about it.

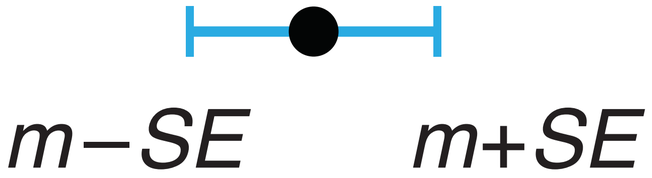

$\mu-\sigma$
$\mu$
$\mu+\sigma$


μ−σ
μ
μ+σ
m−s
m
m+s


μ−σ
μ
μ+σ
m−s
m
m+s


μ−σ
μ
μ+σ

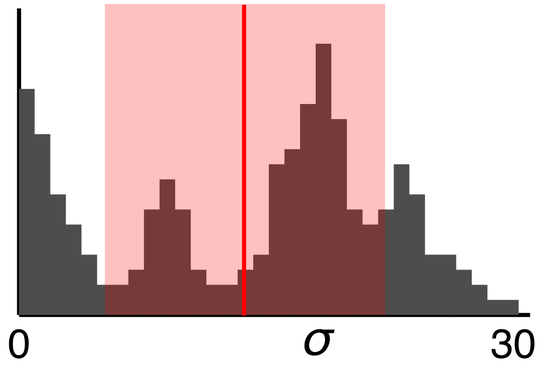

m−s
m
m+s


m−SE
m+SE

A

Population distribution

$\mu$




B

Samples

$X_1$ = [ 1, 9, 17, 20, 26 ]

$X_2$ = [ 8, 11, 16, 24, 25 ]

$X_3$ = [ 16, 17, 18, 20, 24 ]

...

C

Sample means

$\bar{X}_1 = 14.6$

$\bar{X}_2 = 16.8$

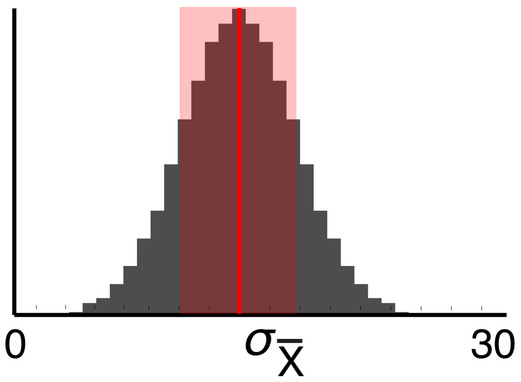
$\bar{X}_3 = 19.0$

...

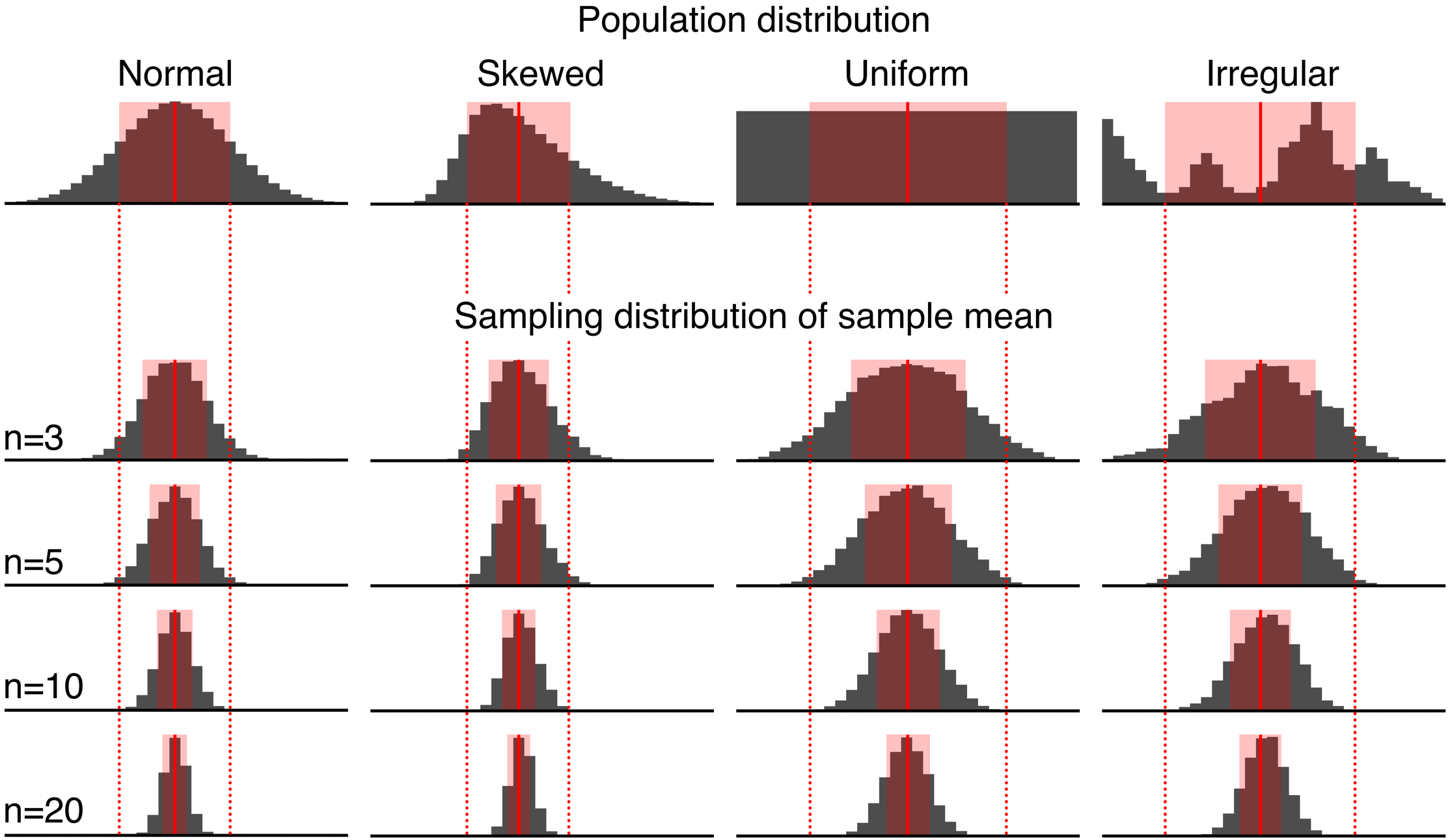
D

Sampling distribution of sample means

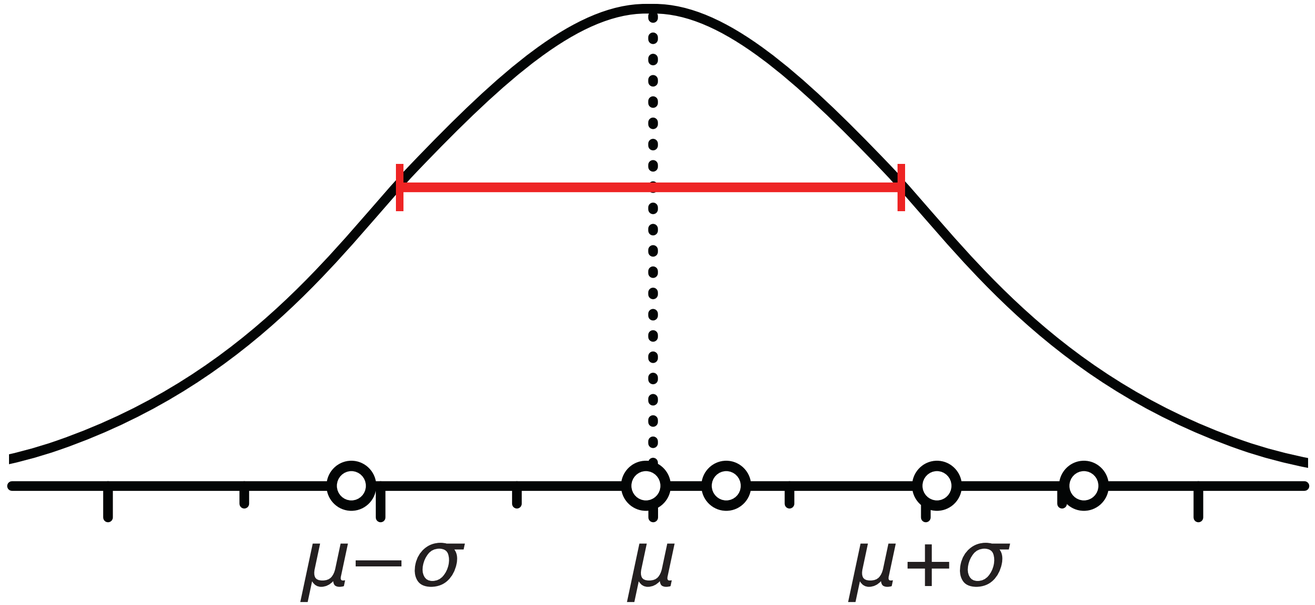
$\mu_{\bar{X}}$

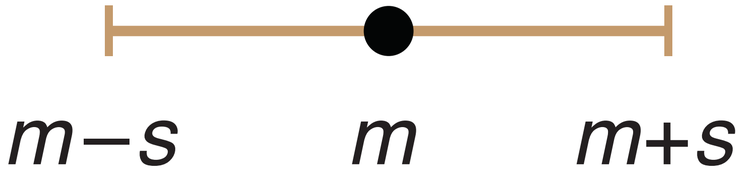


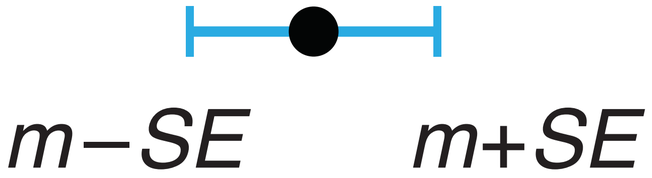

Nat Methods (2013) 10:809–810.

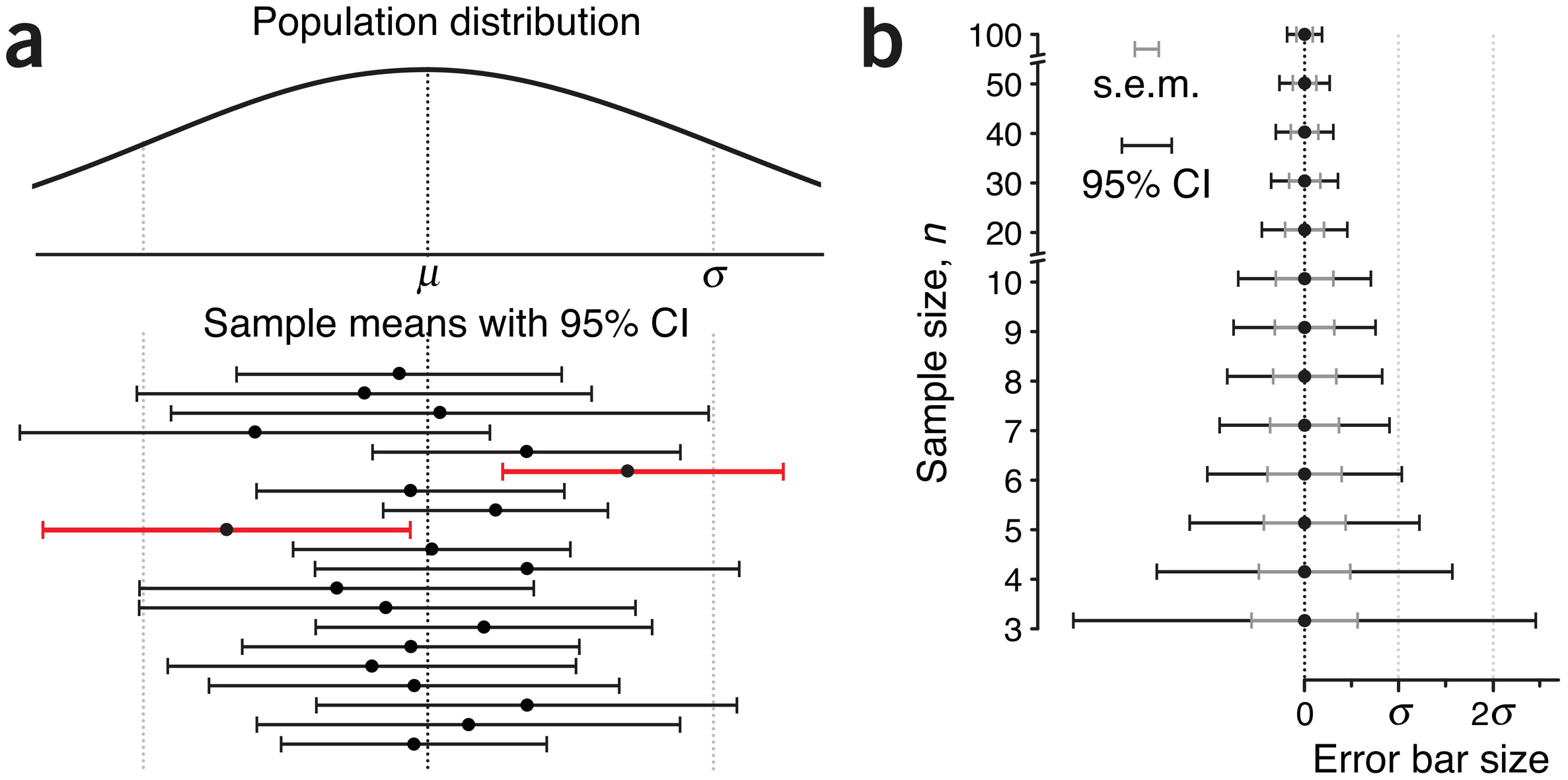



Nat Methods (2013) 10:809–810.

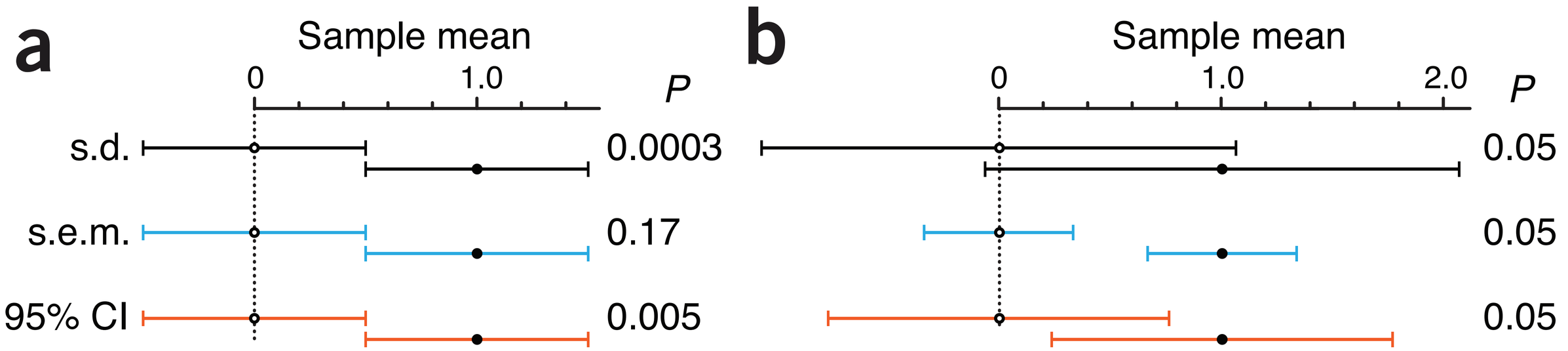

μ−σ
μ
μ+σ


m−s
m
m+s


m−SE
m+SE




Nat Methods (2013) 10:921–922.




Nat Methods (2013) 10:921–922.

Ok, so how exactly are we supposed to be using error bars?

This is a great question.

As you've seen, they're actually quite hard to interpret quantitatively.

If they're standard deviation error bars, all you know is that a good fraction of the time, if you can assume normality, the next sample value will fall within them. This often isn't that useful.

If they're SEM error bars then you have some sense of the precision of estimating the mean of the population. This is made more useful by 95% CI error bars because they incorporate the traditional p-value of 0.05.

But it's still hard to assess exactly how the length of the bar relates to your power to make inferences. Or how differences in error bar length and distance between error bars relate to these inferences.

I've given you some rules of thumb here and I hope been able to clarify ideas about sampling distributions so that you can approach error bars with less confusion. Though still with a healthy dose of scepticism.

created by

Martin Krzywinski, Kim Bell-Anderson & Philip Poronnik

written and designed by

Martin Krzywinski

production

One Ski Digital Media Productions

with financial support by

University of Sydney

filmed at

University of Sydney, Australia

# EXERCISE 1

Read these Points of Significance columns

The Importance of Being Uncertain

http://www.nature.com/nmeth/journal/v10/n9/full/nmeth.2613.html

Error bars

http://www.nature.com/nmeth/journal/v10/n10/full/nmeth.2659.html

Box plots

http://www.nature.com/nmeth/journal/v11/n2/full/nmeth.2813.html

# EXERCISE 2

Pick up a copy of your favourite journal. If you don’t have one yet, get Nature or Science.

Find some figures with error bars.

How many of these figures are bar charts with error bars?

Does the legend specify the type of error bar? How about sample size?