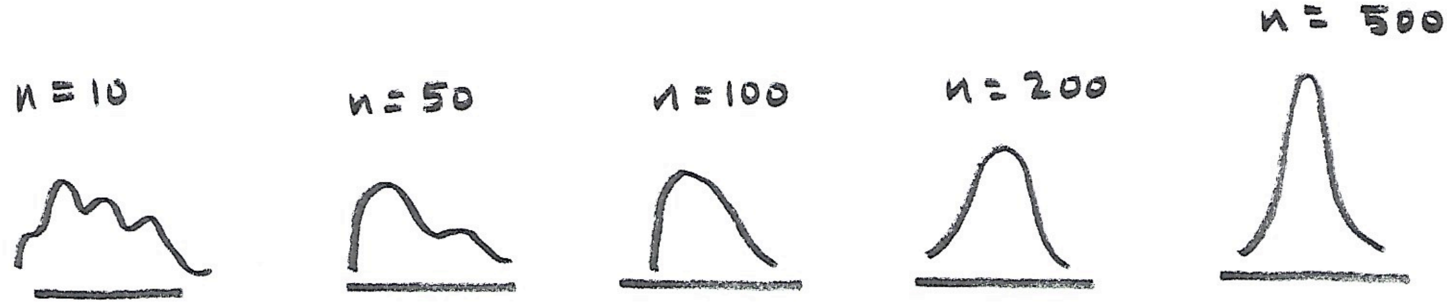
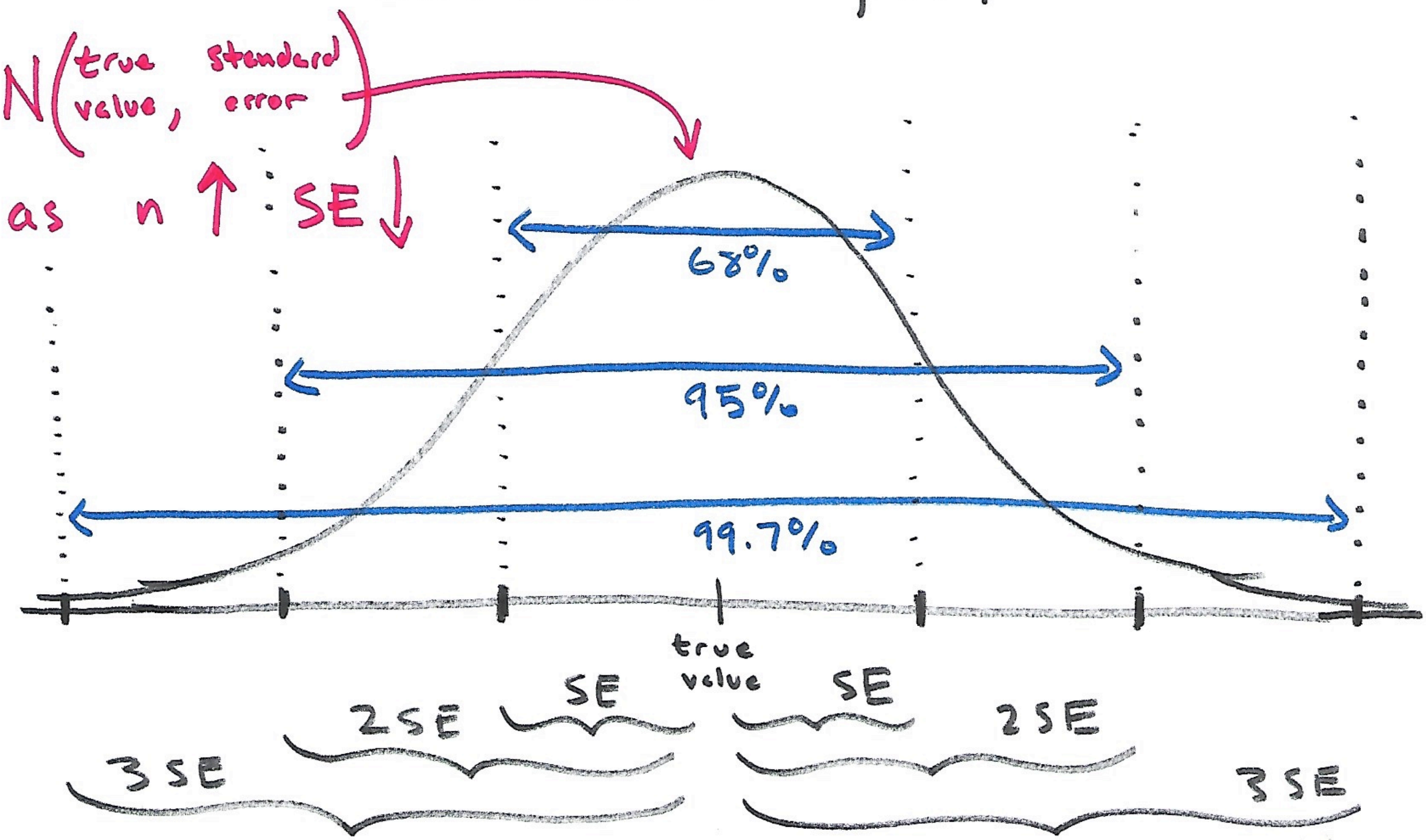
CLT: as sample size grows, sampling distribution of a statistic ...

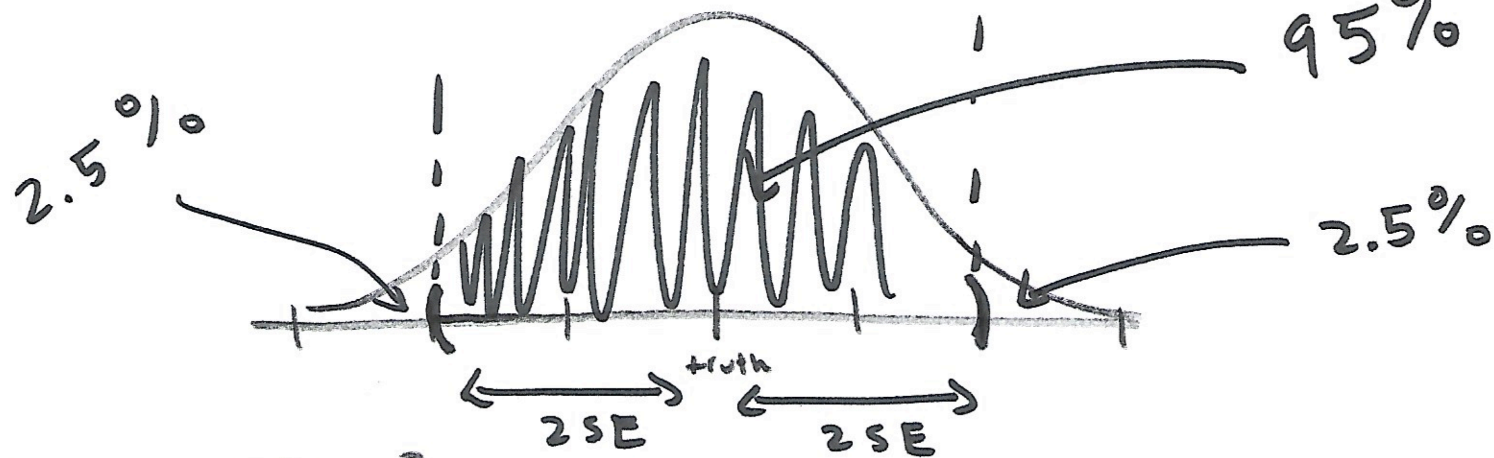
- concentrates around the true value of the population parameter
- looks more and more normal



normal approximation: use CLT to justify approximating the sampling distribution of a statistic with a bell curve instead of with simulation and a histogram:



Idealized confidence interval based on the "true" sampling distribution



Want 95%?

$$L = \text{truth} - 2SE \qquad\qquad U = \text{truth} + 2SE \quad \Longleftrightarrow \quad \text{truth} \pm 2SE$$

Want 68%?

$$L = \text{truth} - SE \qquad\qquad U = \text{truth} + SE \quad \Longleftrightarrow \quad \text{truth} \pm SE$$

Want some other %?

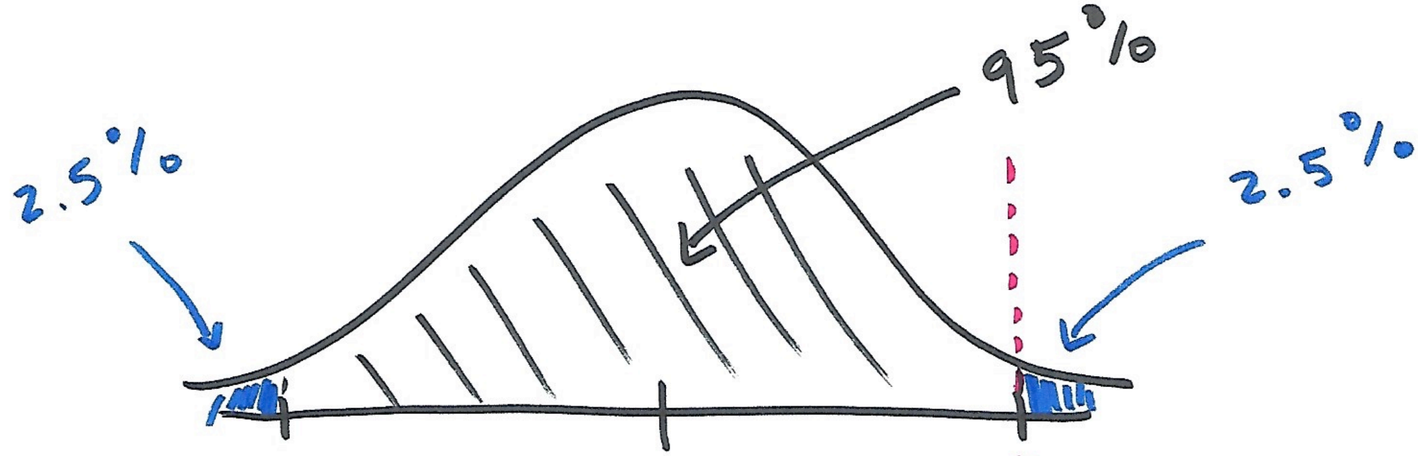
$$\text{truth} \pm z^{*} \times SE$$

"z-score"
quantile of the normal w/ mean 0 and std 1.

WTF is z^*?

It's a quantile of the standard normal bell curve. It tells you how far out from the middle of the curve you have to go so that the bounds of your confidence interval swallow the desired % of the distribution.

Want 95%? The z^* you need is the 0.975 quantile of $N(0,1)$.



z^* is the the point that has 97.5% of the distribution to the left.

Want	z^*
90%	≈ 1.64
95%	≈ 1.96
99%	≈ 2.57
$100 \times (1-\alpha)\%$	qnorm$(1-\frac{\alpha}{2}, 0, 1)$

WTF is standard error (SE)?

It's the special name we give to the standard deviation of a sampling distribution.

The exact formula will depend on context: the data type and the specific statistic you are considering.

sample proportion for binary data...

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

p = true population proportion

n = sample size

sample mean of numerical data...

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ = true population standard deviation of the data

n = sample size

Idealized confidence intervals for different estimation problems

sample proportion $\quad p \pm z^* \times \sqrt{\frac{p(1-p)}{n}}$

sample mean $\quad \mu \pm z^* \times \frac{\sigma}{\sqrt{n}}$

But wait!

We don't know p or μ or σ.
How do use these interval formulas?
Plug in your point estimates

sample proportion $\quad \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

sample mean $\quad \bar{x} \pm z^* \frac{\hat{\sigma}}{\sqrt{n}}$

in general . . . $\quad \text{point estimate} \pm \underbrace{z^* \widehat{SE}}_{\text{margin of error}}$

These are <u>approximate</u> confidence intervals based on the normal distribution. If margin of error is large, uncertainty is high,

Approximate confidence intervals

proportion $\quad \hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

mean $\quad \bar{X} \pm z^* \times \frac{\hat{\sigma}}{\sqrt{n}}$

in general... $\quad \text{point estimate} \pm \underbrace{z^* \times \widehat{SE}}_{\text{margin of error}}$

- if margin of error is high, uncertainty is high, estimate is less reliable
- if margin of error low, uncertainty is low, estimate is more reliable

- all these intervals are centered at the point estimate.