# Sta 111 - Summer II 2017
# Probability and Statistical Inference

## 7. Confidence intervals

Lu Wang

Duke University, Department of Statistical Science

July 11, 2017

# Outline

1. Use confidence intervals to estimate population parameters
2. Confidence level, critical value
3. Calculate the sample size to achieve desired margin of error
4. Common misconceptions about confidence intervals

# Confidence intervals vs point estimates

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

  - We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

## Constructing a confidence interval

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is defined as

$$\textit{point estimate} \pm 2 \times \textit{SE}$$

$$\textit{SE} = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$\begin{aligned} \bar{x} \pm 2 \times \textit{SE} &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \\ &= (2.7, 3.7) \end{aligned}$$

## Statistical inference methods based on the CLT depend on the same conditions as the CLT

### What are conditions for CLT to apply?

1. *Independence:* Observations in the sample must be independent
2. *Sample size / skew:* $n \geq 30$ and population distribution should not be extremely skewed

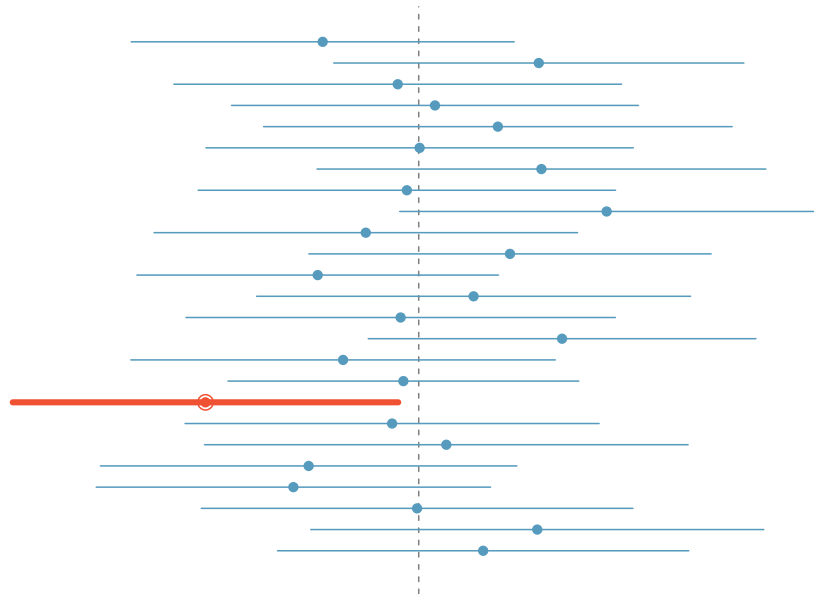
Always check these in context of the data and the research question!

*Note:* We will discuss working with samples where $n < 30$ in the next chapter.

# What does "95% confident" mean?

- Suppose we took many samples and built a confidence interval from each sample using the expression *point estimate* $\pm 2 \times$ *SE*.
- Then about 95% of those intervals would contain the true population mean ($\mu$).

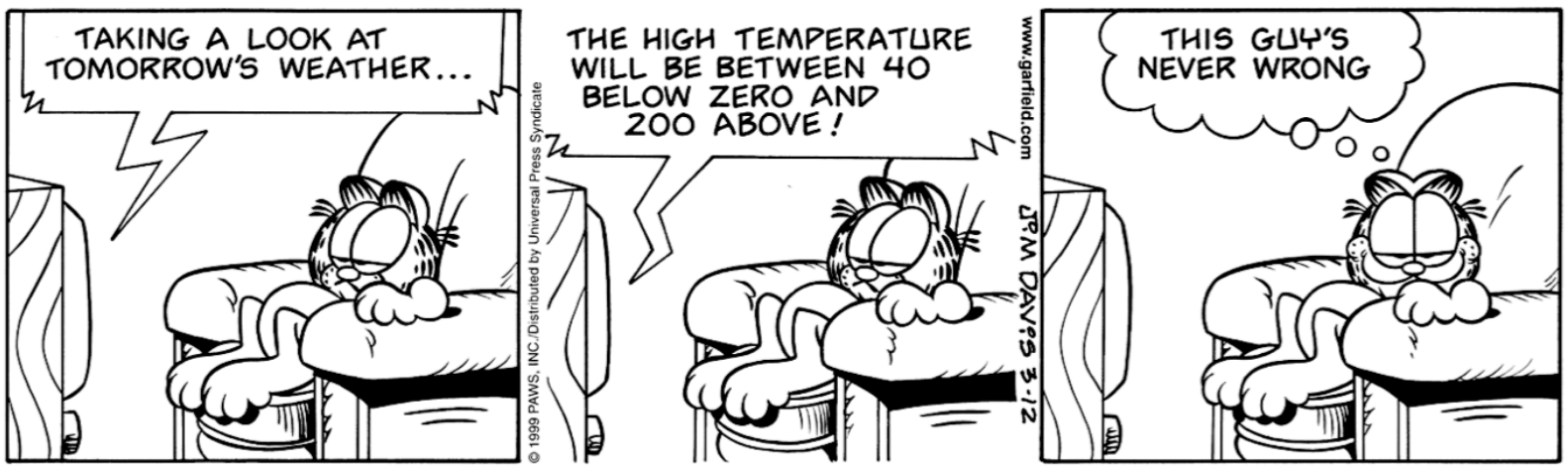
- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



# Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?



## Confidence interval, a general formula

$$\textit{point estimate} \pm z^\star \times \textit{SE}$$

- If the parameter of interest is the population mean, and the point estimate is the sample mean, CI becomes $\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$.
- The *critical value* $z^\star$ in the confidence interval depends on the confidence level. For a 95% confidence interval, $z^\star = 1.96$.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%. In order to change the confidence level we need to adjust $z^\star$ in the above formula.

# Find critical value given confidence level

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

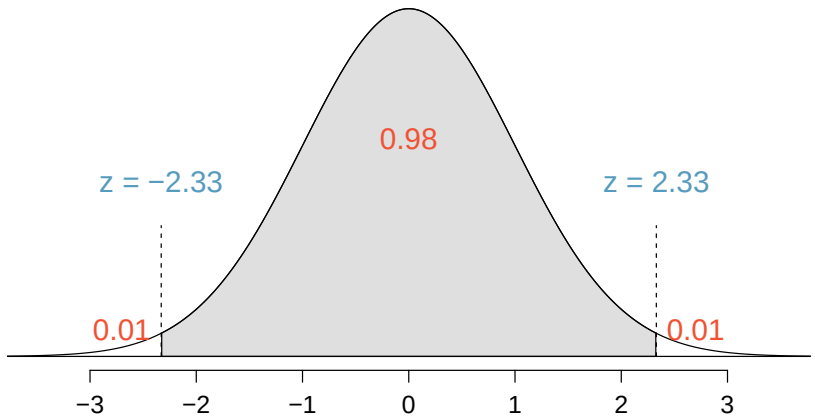
(a) $Z = 2.05$

(b) $Z = 1.96$

(c) $Z = 2.33$

(d) $Z = -2.33$

(e) $Z = -1.65$



## Margin of error

In a confidence interval, $z^\star \times SE$ is called the *margin of error* ($ME$).

- So if we know the desired $ME$, and confidence level (and hence $z^\star$), and the sample standard deviation $s$, we can solve for $n$.

**Practice:** A question on the survey of 1,154 US residents is "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?". The average time spent relaxing was 3.68 hours, with a standard deviation of 2.6 hours.

Calculate the sample size necessary to obtain a 90% confidence interval with a margin of error of 0.06 hours.

$$1.65 \times \frac{2.6}{\sqrt{n}} \leq 0.06 \rightarrow n \geq \left(\frac{1.65 \times 2.6}{0.06}\right)^2 = 5112.25$$

At least 5,113 people.

# Common misconceptions about confidence intervals

1. The confidence level of a confidence interval is the probability that the true population parameter is in the confidence interval you construct for a single sample.
   *The confidence level is equal to the proportion of random samples that result in confidence intervals containing the true population parameter.*

2. A narrower confidence interval is always better.
   *This is incorrect since it is possible to make very precise statements with very little confidence.*

3. A wider interval means less confidence.
   *This is incorrect since the width is a function of both the confidence level and the standard error.*