

# Confidence Intervals

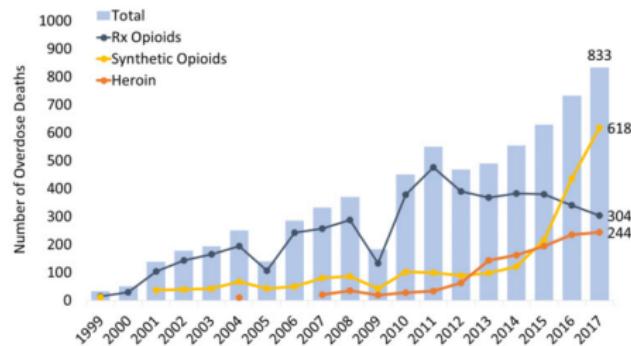
## STA 102: Introduction to Biostatistics

Yue Jiang

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

# The opioid crisis

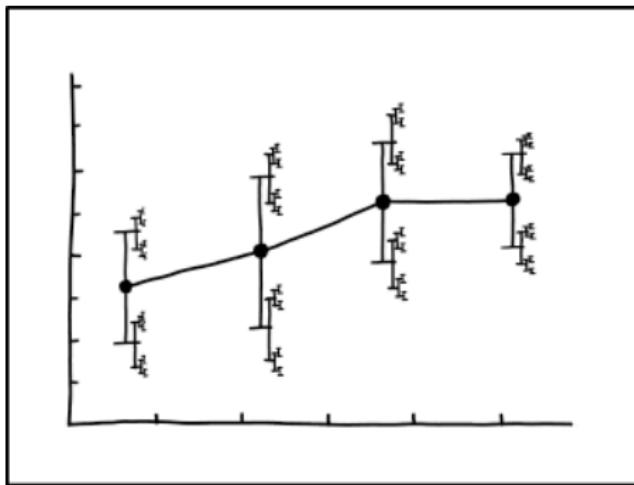


West Virginia has the highest age-adjusted rate of drug overdose deaths involving opioids

# Statistical inference

- ▶ **Point estimation**: estimating an unknown parameter using a single number calculated from the sample
  - ▶ We estimate the one-year death rate due to opioid-related overdoses in WV to be 49.6 per 100,000
- ▶ **Interval estimation**: estimating an unknown parameter using a range of values that is likely to contain the true parameter
  - ▶ We estimate that the one-year death rate due to opioid-related overdoses in WV is between 45 and 55 per 100,000
- ▶ **Hypothesis testing**: evaluating whether our observed sample data provides evidence against some population claim
  - ▶ We evaluate the hypothesis that the opioid overdose death rate is the same in WV and NC. In a random sample of death certificates from the two states, the rate was considerably higher in WV, providing evidence against this hypothesis

# Why should we care about interval estimation?



I DON'T KNOW HOW TO PROPAGATE  
ERROR CORRECTLY, SO I JUST PUT  
ERROR BARS ON ALL MY ERROR BARS.

Randall Munroe, xkcd

# What is a confidence interval, anyway?



A **confidence interval** gives a range of values that is intended to cover the parameter of interest to a certain degree of “confidence”

Confidence interval = **point estimate**  $\pm$  **margin of error**

# How do you interpret a confidence interval?

## Primary endpoint

Annual asthma exacerbation rate over  
48 weeks\*

Number of patients analysed 267

Rate estimate (95% CI) 1.33  
(1.12-1.58)

Researchers conducted a clinical trial of a drug intended for severe asthma patients. Their primary endpoint was evaluating whether the mean rate of asthma exacerbation over 48 weeks was different between placebo and treatment arms. Above is the 95% confidence interval for the mean rate among the placebo patients.

How do you interpret this interval? (more on this very soon)

## Brief caveat

For now, let's assume that we know  $\sigma$  (this very rarely ever happens, since it is a population parameter)

## Two-sided confidence intervals

Given a random variable  $X$  with mean  $\mu$  and standard deviation  $\sigma$ , the CLT tells us that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

where  $Z$  has a standard normal distribution if  $X$  is normal, and  $Z$  is approximately normal if  $X$  is not normal, but  $n$  is large enough

## Deriving the two-sided interval

For a standard normal random variable, 95% of the observations lie between -1.96 and 1.96 for  $Z \sim N(0, 1)$ , so

$$0.95 = P(-1.96 \leq Z \leq 1.96)$$

So, a 95% CI is given by

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

## Generic form of confidence intervals for the mean

$$\left( \bar{X} - z_{1-\alpha/2}^* \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}^* \frac{\sigma}{\sqrt{n}} \right)$$

Point estimate  $\pm$   $\underbrace{\text{confidence multiplier} \times \text{standard error}}_{\text{Margin of error}}$

## Other coverage probabilities

The **confidence multiplier**,  $z_{1-\alpha/2}^*$ , is the z-score that cuts off the upper  $100\% \times \alpha/2$  of the distribution (the  $1 - \alpha/2$  percentile)

For  $\alpha = 0.05$ , we have  $1 - \alpha/2 = 0.975$ , and so  $z^*$  is the 97.5<sup>th</sup> quantile of the standard normal distribution (calculated using software packages)

Compromising on confidence level to obtain narrower CIs is...highly frowned upon

# CI interpretation

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

Suppose we select  $M$  different random samples from the population of size  $n$ , and use them to calculate  $M$  different 95% CIs in the same way as above. We expect 95% of these intervals would cover the true  $\mu$  and 5% do not

Interactive activity

# CI interpretation

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

~~"There is a 95% chance that  $\mu$  lies in the interval"~~

## CI interpretation

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

**Important:** we do not know whether any particular interval is in the 95% of them that cover the mean or the 5% that don't

Since  $\mu$  is a parameter, it's either in our confidence interval or not

# When can we use this CI?

$$\left( \bar{X} - z^* \frac{\sigma}{\sqrt{n}}, \bar{X} + z^* \frac{\sigma}{\sqrt{n}} \right)$$

Remember, this is only ok to use when  $\sigma$  is known, and:

- ▶  $X$  is normal
- ▶  $X$  is non-normal, but  $n$  is sufficiently large

# What can we do if $\sigma$ isn't known?



# What can we do if $\sigma$ isn't known?



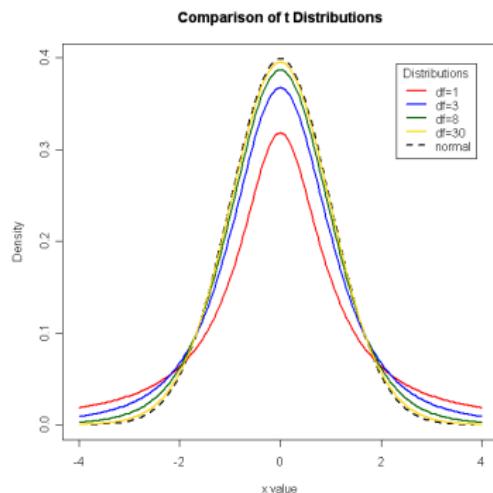
As a Guinness Brewery employee, William Sealy Gossett published a paper on the  $t$  distribution, which became known as Student's  $t$  (the brewery didn't allow him to use his own name)

A. Student. *The probable error of a mean* (1908).

The  $t$ -distribution is used to construct a confidence interval for the mean when  $\sigma$  is unknown and must be estimated from the data

# Student's $t$ distribution

- ▶ The  $t$  distribution looks like the normal distribution except it has fatter tails, leading to wider CIs
- ▶ This is due to the uncertainty involved in estimating  $\sigma$  by using  $s$
- ▶ As the sample size increases,  $s$  is a better and better estimate of  $\sigma$ , and so the  $t$  distribution looks more and more like the normal distribution



# Degrees of freedom

The **degrees of freedom** of a  $t$  distribution tells us how much information is “available” for estimating  $\sigma$  using  $s$ . The random variable

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a  $t$  distribution with  $n - 1$  degrees of freedom ( $df$ ), which we denote by  $t_{n-1}$  (we lose one  $df$  by estimating the sample mean using  $\bar{X}$ ). The  $t$  distribution only has one parameter ( $df$ )

## Two-sided interval with unknown $\sigma$

$$\left( \bar{X} - t_{n-1;1-\alpha/2}^* \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1;1-\alpha/2}^* \frac{s}{\sqrt{n}} \right)$$