Confidence Intervals

STA 198: Introduction to Health Data Science

Yue Jiang

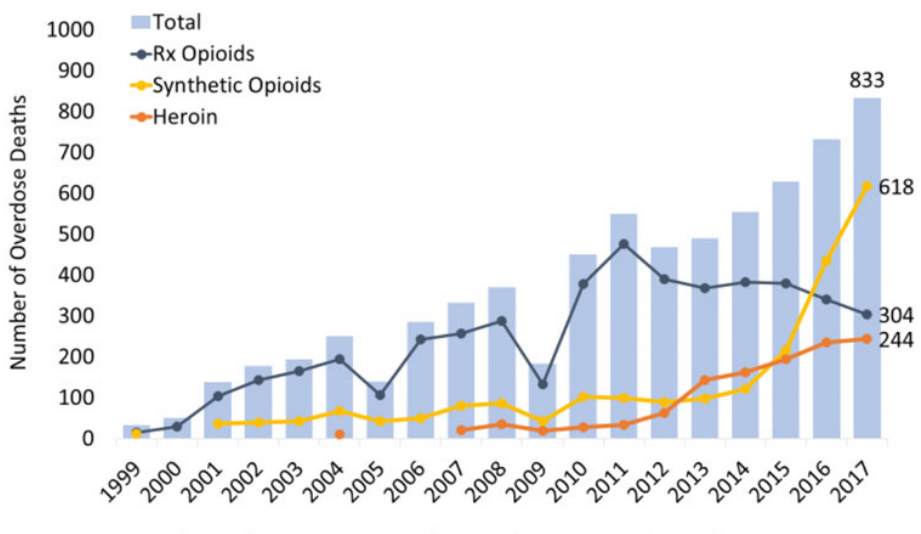
June 12, 2020

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

National Institute on Drug Abuse
Advancing Addiction Science

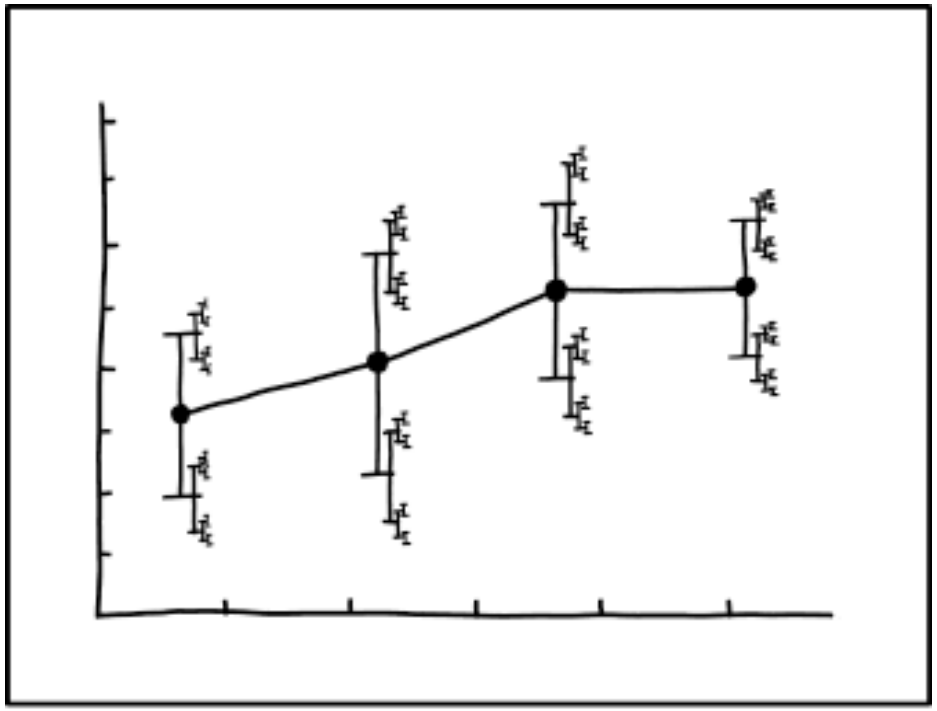


West Virginia has the highest age-adjusted rate of drug overdose deaths involving opioids

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 populuation 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?



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 σ (this very rarely ever happens, since it is a population parameter)

Two-sided confidence intervals

Given a random variable X with mean μ and standard deviation σ, the CLT tells us that

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

where Z has a standard normal distrsibution 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

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

(how did we get this?)

Generic form of confidence intervals

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

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

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. Approximately 95% of these intervals would cover the true μ 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 μ 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 μ is a parameter, it's either in our confidence interval or not

Other coverage probabilities

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

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

Although 95% CIs are the most common, we can easily generate intervals with other coverage probabilities by adjusting the confidence multiplier

Other coverage probabilities

The confidence multiplier, $z^\star_{1-\alpha/2}$, 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 $z^\star$ is the 97.5^{th} quantile of the standard normal distribution (calculated using software packages)

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

When can we use this CI?

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

Remember, this is only ok to use when σ is known, and:

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

What can we do if σ isn't known?

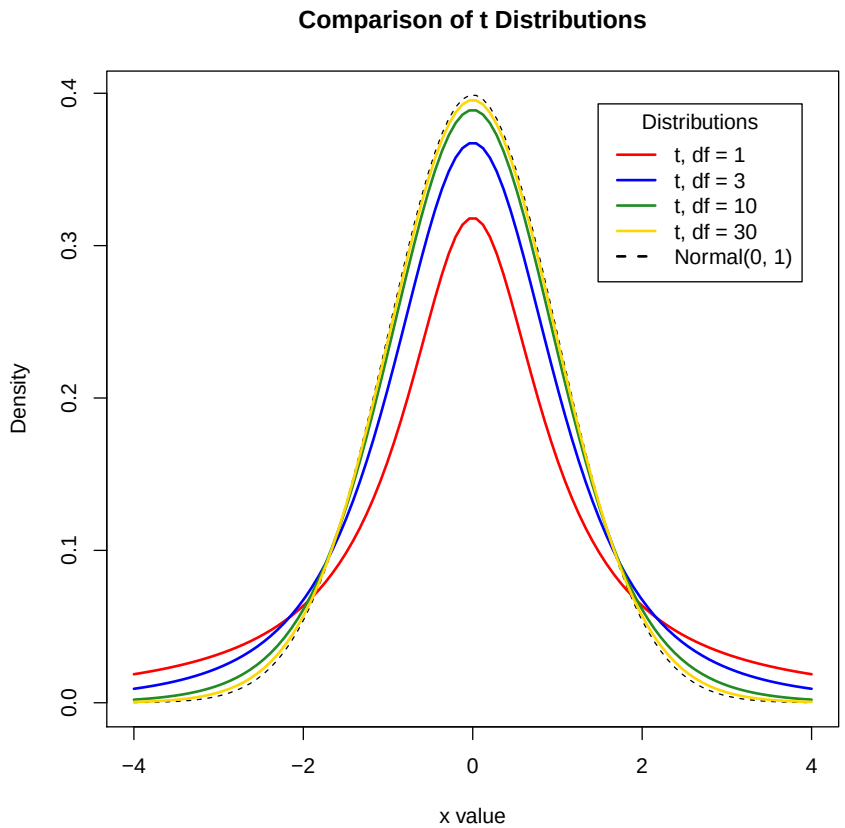
What can we do if σ isn't known?

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).

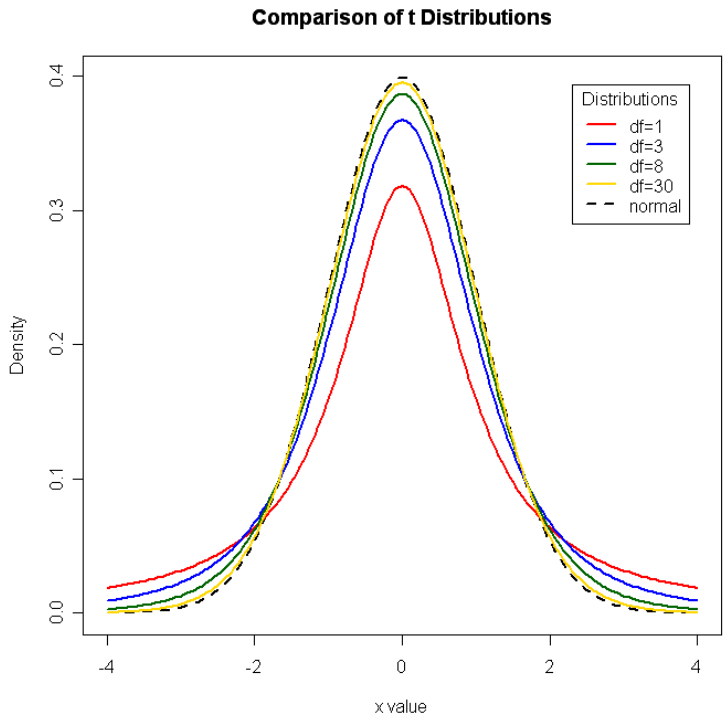
Student's t distribution

- Gosset used his new distribution to determine how large a sample should be for taste testing beer
- The t distribution is appropriate for constructing a confidence interval for the mean when σ is unknown



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 σ by using s
- As the sample size increases, s is a better and better estimate of σ, 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 σ 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)

How does this compare to the standard normal distribution?

Two-sided interval with unknown σ

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

What about one-sided intervals?

The symmetric two-sided confidence intervals we have dealt with so far give the shortest intervals with the desired coverage probability (for symmetric distributions)

Occasionally, we may only want an upper limit or a lower limit for the population mean, for instance in a non-inferiority clinical trial for a generic drug (we don't expect the generic to work better, but we do expect it to be "not worse")

How might we construct these types of intervals?