Chapter 6: Confidence Intervals

DSCC 462

Computational Introduction to Statistics

Anson Kahng

Fall 2022

Inference

- Goal: Describe population based on a sample
- *Point estimation*: use a single number to estimate the population parameter
  - E.g., sample mean $\bar{x}$ is a point estimate of the population mean μ
- Different samples will produce different estimates, so there is some uncertainty involved with a point estimates
- *Interval estimation*: Range of reasonable values that are intended to contain the parameter of interest with a certain degree of confidence
  - *Confidence intervals*

Confidence Intervals: Example

- 37% of all quokkas are actually as happy as they look
- The margin of error is ±4%, 19 times out of 20
- This means that we are 95% sure that the percentage of all quokkas that are actually as happy as they look is captured by the interval (33%, 41%)

Confidence Intervals: Unknown Mean, Known Variance

- Suppose we want to construct a confidence interval for μ
- We use $\bar{x}$ as our point estimate for μ
- Drawing upon the sampling distribution of this mean, we can construct our confidence interval around $\bar{x}$
- Recall that the CLT tells us that for a random variable X with mean μ and variance σ^2,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

given that n is large enough or X is normally distributed

Confidence Intervals: Unknown Mean, Known Variance

- For the standard normal distribution $N(0,1)$, recall that 95% of all observations lie between -1.96 and 1.96
  - $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$
- Going from a standard normal distribution to any normal distribution:

$$\Pr\left(-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

Confidence Intervals: Unknown Mean, Known Variance

- Rearranging terms, we get

$$\Pr\left(-1.96 \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

$$\Pr\left(-1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\Pr\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- Or, we are 95% confident that the interval $\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ contains the true population mean μ

    - If we sample 100 different confidence intervals for μ, approximately 95 of these intervals will contain the true population mean and 5 will not

Confidence Intervals: Illustration

Confidence Intervals: Example

- Setup:
  - Let X be the amount of money spent on concert tickets in the past year
  - Assume $X \sim N(\mu, \overset{\text{stdev}}{80})$ how do we know stdev: sometimes you can use sample sd to estimate sigma
  - Suppose we take a sample of 100 concert-goers and determine how much each person spent on concert tickets in the past year
  - The average amount spent for this sample is 220
- Q: Based on this sample, what is a 95% confidence interval for μ?

Confidence Intervals: Example

- Often, we look at 95% confidence intervals, but this choice is fairly arbitrary. We can consider other intervals (e.g., 90% or 99% or…)
- Also, these intervals so far have been *two-sided*, which means that in the case of a 95% confidence interval, we want a 2.5% probability of falling above our upper limit and a 2.5% probability of falling below our lower limit
- In general, for a two-sided $100 \cdot (1-\alpha)$% confidence interval, we want $\alpha/2$% probability of falling above the upper limit and $\alpha/2$% probability of falling below the lower limit

Two-Sided Confidence Intervals

- Let $z_{\alpha/2}$ (resp. $-z_{\alpha/2}$) be the value that cuts off an area of $\alpha/2$ in the upper tail (resp. lower tail) of the standard normal distribution

Confidence	α	**R code**	$z_{\alpha/2}$
90%	0.10	`qnorm(1-0.10/2)`	1.645
95%	0.05	`qnorm(1-0.05/2)`	1.96
99%	0.01	`qnorm(1-0.01/2)`	2.576

Two-Sided Confidence Intervals

- Under this generic framework, we have that a $100\% \cdot (1-\alpha)$ confidence interval for μ is $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$
- Interpretation: We are $100\% \cdot (1-\alpha)$ confident that this interval covers μ

Narrower Confidence Intervals

- Suppose that we want to make a confidence interval narrower without reducing the confidence level
  - Recall that given a confidence level $100\% \cdot (1 - \alpha)$, we have the confidence interval $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$
- What are the parameters we can change?
- We need a larger sample size; as $n \uparrow$, the standard error $\sigma/\sqrt{n} \downarrow$, resulting in a narrower confidence interval

Effect of Sample Size

n	95% Confidence Limits	Length of Interval
1	$\overline{X} \pm 1.96\sigma$	3.92σ
10	$\overline{X} \pm 0.620\sigma$	1.24σ
100	$\overline{X} \pm 0.196\sigma$	0.392σ
1000	$\overline{X} \pm 0.062\sigma$	0.124σ

Margin of Error

- Recall: confidence level $100\% \cdot (1 - \alpha)$, interval $\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$
- We call $m = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ the **margin of error**
- The length of the confidence interval is $2 \cdot m = 2 \cdot z_{\alpha/2} \cdot \sigma/\sqrt{n}$

Margin of Error and Sample Size

- Suppose that we want a confidence interval to be a certain length (m = half of this length)
    - E.g., if we are testing the effect of a new drug, we may want the treatment mean to be estimated within some given margin of error
- Given a fixed margin of error, how many samples do we need?
    - We know that $m = z_{\alpha/2} \cdot \sigma / \sqrt{n}$
    - Therefore, $n = \left\lceil \frac{z_{\alpha/2}^2 \cdot \sigma^2}{m^2} \right\rceil$ (always round up!)

Sample Size Example

- Setup:
  - Let X be the amount of money spent on concert tickets in the past year
  - Assume $X \sim N(\mu, 80)$
  - Suppose we take a sample of 100 concert-goers and determine how much each person spent on concert tickets in the past year
  - The average amount spent for this sample is 220
- Q: How large of a sample size do we need in order to create a 95% confidence interval of length 40?

Sample Size Example

- Setup:
  - Let X be the amount of money spent on concert tickets in the past year
  - Assume $X \sim N(\mu, 80)$
  - Suppose we take a sample of 100 concert-goers and determine how much each person spent on concert tickets in the past year
  - The average amount spent for this sample is 220
- Q: How large of a sample size do we need in order to create a 95% confidence interval of length 40?

Margin of error: $m = 40/2 = 20$

$\dfrac{z_{\alpha/2}^2 \cdot \sigma^2}{m^2} = \dfrac{1.96^2 \cdot 80^2}{20^2} = 61.47$, so we need a sample of $n = 62$ concert goers to get a 95% CI of length 40

One-Sided Confidence Intervals

- Most often, we are interested in two-sided confidence intervals
- In some scenarios, we may only be concerned with an upper limit or a lower limit, but not both
- When this is the case, we can create a *one-sided confidence interval*

One-Sided Confidence Intervals

- Consider a new cholesterol drug. We are interested in seeing if this new drug helps lower cholesterol
- Suppose cholesterol for people on this new drug has a distribution with unknown mean μ and standard deviation $\sigma = 30$ mg/dL
- People on this new medicine tend to have lower cholesterol than those who are not on the medicine
- We are interested in finding an *upper bound* for μ
  - What is the highest that we would expect this mean to be? Is this still lower than the mean cholesterol level for people who are not on the drug?

One-Sided Confidence Intervals

- To construct a one-sided confidence interval, we consider only the area in one tail of the standard normal distribution
- Since we are concerned with the upper limit, we use $\bar{X} + z_{\alpha}\dfrac{\sigma}{\sqrt{n}}$
- Note that we have z_{α} instead of $z_{\alpha/2}$ because we are only considering one tail

One-Sided Confidence Intervals: Illustration

One-Sided Confidence Intervals: Example

- Consider the same cholesterol drug as before. We take a sample of 100 people on the new medicine, and find that their mean cholesterol is 184 mg/dL. Recall that $\sigma = 30$ mg/dL for this population

- Calculate the one-sided 95% upper-bound confidence interval

One-Sided Confidence Intervals

- For an upper limit confidence interval, we use $\bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$
  - $\left(-\infty, \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \right)$
- For a lower limit confidence interval, we use $\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$
  - $\left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty \right)$

Confidence Intervals: Unknown Mean, Unknown Variance

- So far, we have assumed that the population variance σ^2 is known and only the mean μ is unknown

- In reality, we often do not know what the variance σ^2 is either

- However, we can estimate σ^2 with the sample variance s^2

  - But we have to be a bit more careful here, because the sampling distribution for $\overline{X}$ is more variable and the value of s^2 is likely to differ from sample to sample

- We can use something called the **Student's t distribution**

Student's t Distribution

- Recall that, assuming n is sufficiently large, $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- If we replace σ with s, we get $t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, which is not a standard normal
- Instead, t has a Student's t distribution with $n - 1$ degrees of freedom, denoted t_{n-1}

the amount of information that you can use to estimate

Student's t Distribution: Properties

- Intuition: Student's t distribution is similar to a standard normal distribution but has thicker tails

- The t distribution is unimodal and symmetric about 0

- It has thicker tails, meaning that extreme values are more likely to occur

- The shape of the t distribution reflects the extra variability introduced by estimating the variance

- The *degrees of freedom (df)* measure the amount of information available in the data that can be used to estimate σ^2

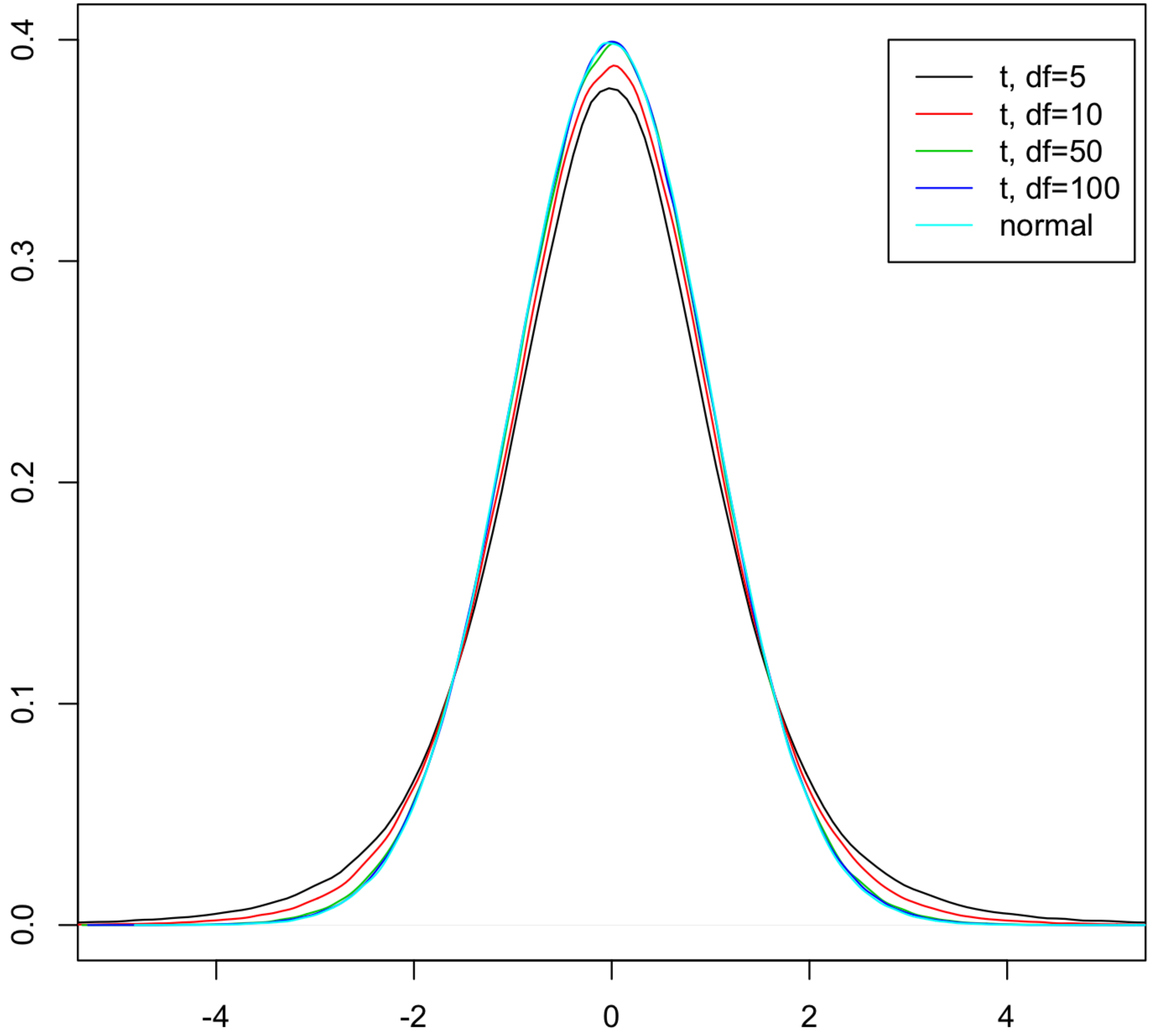
  - Because we lose one degree of freedom in estimating the mean (in order to estimate variance), we are left with $n - 1$ df to estimate σ^2

Student's t Distribution: Degrees of Freedom

- For each possible value of the degrees of freedom, there is a different t distribution
- When the degrees of freedom are low, the distribution is more spread out with heavier tails (worse estimate means more variability)
- As the degrees of freedom approach infinity, the t distribution approaches the normal distribution
    - Intuition: if n is very large, our estimate of s^2 is essentially the same as knowing σ^2

Student's t Distribution: Visualization

as df increases, t distribution and normal distribution become more similar



Student's t Distribution: R

- We can use R to calculate probabilities and quantiles, similar to the normal distribution
- Let $T = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$ be a Student's t distribution with $n - 1$ df
- Calculate $\Pr(T \leq t)$ using `pt(t, n-1)`
- Calculate $\Pr(T \geq t)$ using `1 - pt(t, n-1)`
- Calculate $\Pr(t_1 \leq T \leq t_2)$ using `pt(t₂, n-1) - pt(t₁, n-1)`
- Calculate t such that $\Pr(T \leq t) = q$ (quantile) using `qt(q, n-1)`

Student’s t Distribution: Example

- Consider our concert spending example, but now suppose that we do not know the population variance σ^2

- Suppose we sample $n = 64$ people and get a sample mean of $\bar{x} = 200$ and a sample standard deviation of $s = 80$

- Calculate a 95% confidence interval of the mean

x bar +- t0.025, 63 x s/sqrt(n) = 200 +- qt(0.975 or 0.025, df = 63) * 10