CSCI 3022

intro to data science with probability & statistics

October 17, 2018

Introduction to Statistical Inference & Confidence Intervals

- HW Minopoly Q.

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

Extra credit opportunity! Purely attendance-based.

People respond to incentives and we **know attendance correlates with course grade**

- Consider this carefully if the exams/homeworks are not going as well as you'd like
- Download **Arkaive** -- attendance app (or go to https://arkaive.com/login)
- Enroll in this course with enrollment code: **D0XK** (1 PM w/ Tony)
  XTPL (3 PM w/ Dan)
- During first 15 minutes of each class, check in on Arkaive. You can get credit for attending either section, but not both.
- Remainder of course will be worth **2% pts extra credit** (*before* factored into any final curve)

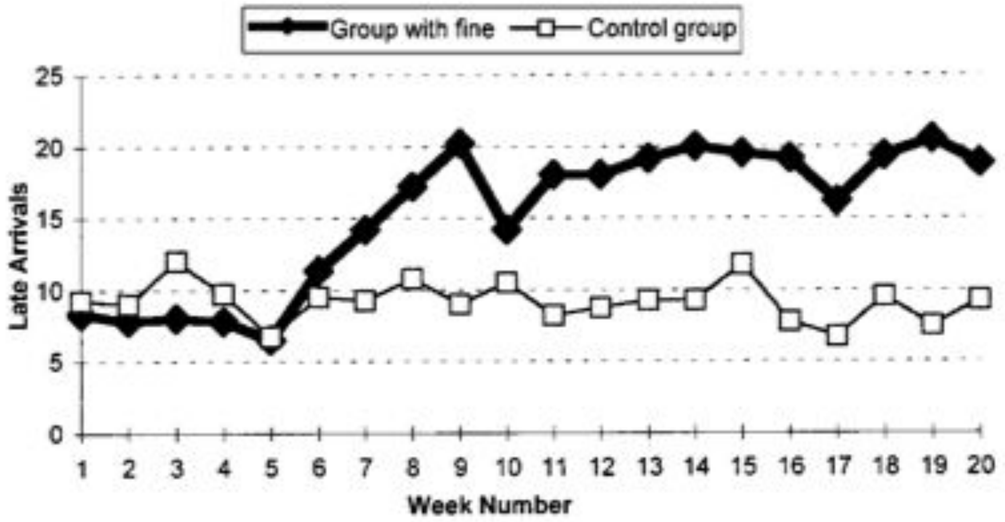


FIGURE 1.—Average number of late-coming parents, per week

Last time on CSCI 3022

- **The Central Limit Theorem**: Let $X_1, X_2, \ldots, X_n$ be i.i.d. draws from some distribution. Then as n becomes large

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- **Box-Muller**: $z = \frac{x - \mu}{\sigma}$ $\qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

- **The Normal CDF**:

$$F(z) = \int_{-\infty}^{z} f(x)\,dx = \Phi(z)$$

stats.norm.cdf(z).

Statistical Inference

- **Goal**: we want to learn the properties of an underlying population by analyzing sampled data.

Questions:

- Is sample mean $\bar{x}$ a good approximation of the population mean μ ?
- Is sample proportion $\hat{p}$ a good approximation of the population proportion p?
- Is there a statistically significant difference between the mean of two samples?
- If the answer is **yes**, *how sure are we*?
- How much data do we need in order to be **confident** in our conclusion?

Confidence Intervals

- The Central Limit Theorem tells us that as the sample size n increases, the sample mean of X is close to *normally* distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$ (variance is $\frac{\sigma^2}{n}$)
- "Standardizing" the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable.

$Z = \frac{X-\mu}{\sigma}$ in general. So, if $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, then $Z = \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})}$

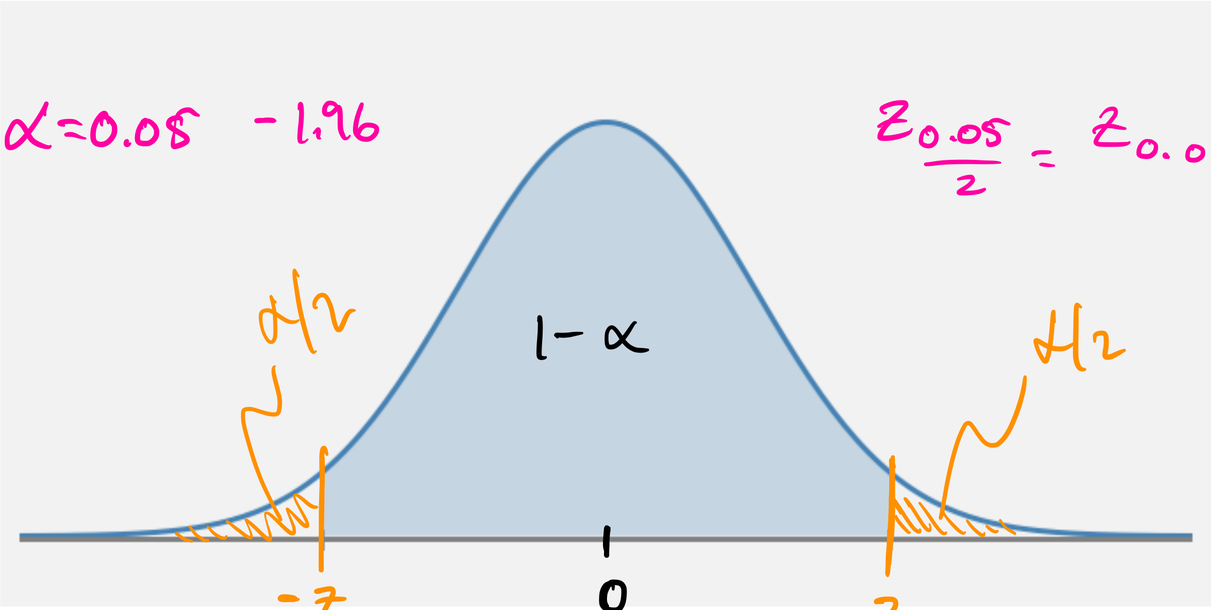
Question: how big does our sample need to be

- …if the variable of interest is normally distributed?
- …if the variable of interest is not normally distributed?

Confidence Intervals

$\alpha = 0.05$

- We saw a while ago that the **95% of the area** under the standard normal curve falls **between –1.96 and +1.96**, so we know that



- This is equivalent to:

Random Variable

$$P(-1.96 \le z \le 1.96) = 0.95 = P\left(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Confidence Intervals

- The **95% confidence interval** for the mean is then given by:

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

middle

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = \bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

pm

- Question: which things in this expression are random variables and which are fixed??

$\bar{X}$, σ, n, $z_{\alpha/2}$

Confidence Intervals

- The 95% CI is centered at $\bar{X}$ and extends $1.96\frac{\sigma}{\sqrt{n}}$ to each side of $\bar{X}$

- The 95% CI's width is $2 \times 1.96\sigma/\sqrt{n}$ which is **not** random; only the location of the interval's midpoint $\bar{X}$ is random.

- We often write the CI $\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}}\right]$ as $\bar{X} \pm 1.96\sigma/\sqrt{n}$.

stats.norm.ppf(1-0.025)

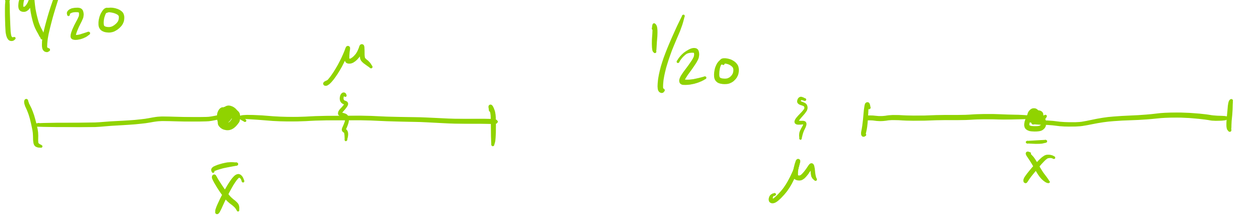
$z_{\alpha/2} = z_{\frac{0.05}{2}} = z_{0.025}$

Interpreting the Confidence Interval

- **Statement**: We are 95% confident that the true population mean is in this interval. $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

- **Correct Interpretation**: In repeated sampling, 95% of all CIs obtained from sampling will actually contain the true population mean. The other 5% of CIs will not.

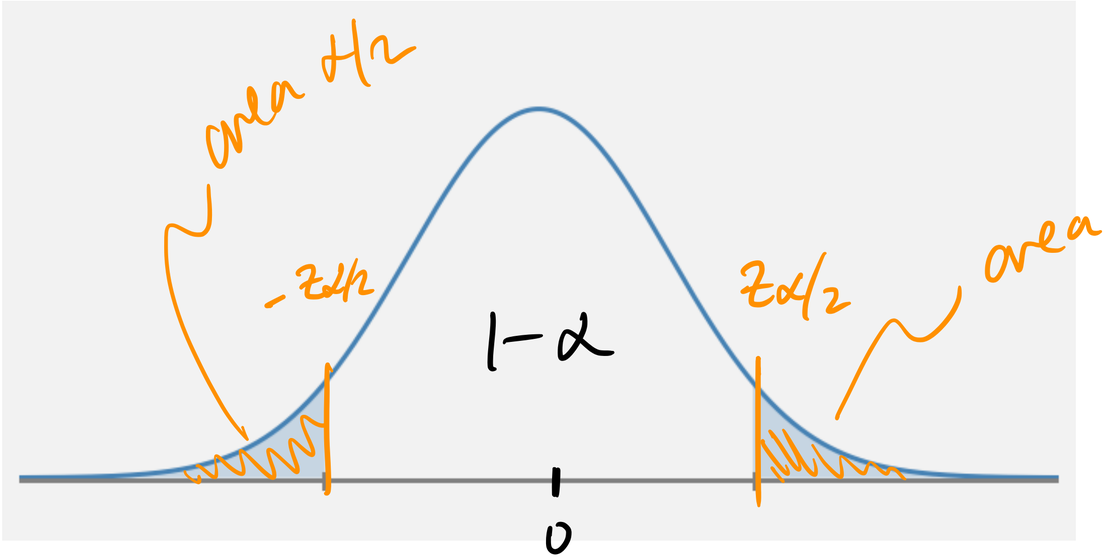
In the long run:



- The confidence level is not a statement about any one particular interval. Instead it describes what would happen if a very large number of CIs were computed using the same CI formula.

Other Levels of Confidence

- A probability of $1-\alpha$ is achieved by using $z_{\alpha/2}$ in place of $z_{0.05/2} = z_{0.025} = 1.96$



- A $100(1-\alpha)\%$ confidence interval for the mean μ when the value of σ is known is given by:

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

compute data mean

known or from data

stats.norm.ppf

number of points

len(data)

CI Example

- The General Social Survey is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day. **Find a 90% confidence interval for the amount of relaxation hours per day.**

$\bar{x}$ n σ

CI: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\longrightarrow$ $3.6 \pm 1.645 \cdot \frac{2}{\sqrt{1000}}$

what is α?

$90\% = 100(1-\alpha)\%$ $= [3.496, 3.704]$

$\alpha = 0.1$ $z_{0.05} = \text{stats.norm.ppf}(1-0.05) = 1.645$

CI Example

- The General Social Survey is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day. **Find a 95% confidence interval for the amount of relaxation hours per day.**

$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ same → see prev slide, but substitute new $z_{\alpha/2}$

↑ $\alpha = 0.05$, $z_{\alpha/2} = 1.96$

$[3.48, 3.72]$ 95%

$[3.50, 3.70]$ 90%

- **Q**: what are the advantages/disadvantages of a wider confidence interval?

balance between true information and useful info.

Test your understanding!

- **Concept Check**: In the previous example we found a 95% CI for relaxation time to be [3.48, 3.72]. Which of the following statements are true?

nope!

A. 95% of Americans spend 3.48 to 3.72 hours per day relaxing after work.

yep!

B. 95% of random samples of 1000 residents will yield CIs that contain the true average number of hours that Americans spend relaxing after work each day.

nope!

C. 95% of the time the true average number of hours an American spends relaxing after work is between 3.48 and 3.72 hours per day.

nope

D. We are 95% sure that Americans in this sample spend 3.48 to 3.72 hours per day relaxing after work.

Computing required sample size

- **Example**: For the GSS data, how large would n have to be to get a 95% CI with width at most 0.1?

width of CI ... $= 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

CI: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

2 $\quad$ 1.96 $\quad$ 2

width $\leq 0.1 \Rightarrow$ $\quad 2 \cdot 1.96 \cdot 2 / \sqrt{n} \leq 0.1$

$2 \cdot 1.96 \cdot 2 \cdot 10 \leq \sqrt{n}$

$2^2 \cdot 1.96^2 \cdot 2^2 \cdot 100 \leq n$ $\quad$ 6400 ish $\leq n$

Confidence IRL…?

- In the previous example we assumed that we knew the population standard deviation.
- **Question**: how often does this happen in real life?

Confidence IRL…?

- In the previous example we assumed that we knew the population standard deviation.
- **Question**: how often does this happen in real life? **never**
- **Solution**: If *n* is large we use the sample variance instead

$$CI_\alpha = \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

replace σ with $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$n \geq 30$ rule of thumb for this class.

- **Solution**: If *n* is small we have to do something else (more on this later)

Confidence intervals for proportions

- Let p denote the proportion of "successes" in a population (e.g. individuals who graduated from college, compute nodes that didn't fail on a given day)
- A random sample of n individuals is selected, and X is the number of successes in the sample
- Then X can be modeled as a Binomial random variable with:

$$E[X] = np$$

$$Var(X) = np(1-p)$$

n: # of flips

$p(1-p)$: var(Ber(p))

Confidence intervals for proportions

- The estimator for p is given by: $\hat{p} = \frac{X}{n}$ $\quad \frac{\#\text{heads}}{\#\text{flips}}$

- The estimator is approximately normally distributed with:

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}np = \boxed{p} \qquad \text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{Var}(X) = \frac{1}{n^2}np(1-p) = \boxed{\frac{p(1-p)}{n}}$$

- Standardizing the estimate yields:

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

- This gives us a confidence interval of:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

cf: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Confidence intervals for proportions

- Example: The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels above 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

n

$\hat{p} = \frac{127}{200}$

$\alpha = 0.01$

99% CI: $\hat{p} \pm z_{0.005}\sqrt{\frac{p(1-p)}{n}}$

$\frac{127}{200} \pm 2.57\sqrt{\frac{\frac{127}{200}\left(1-\frac{127}{200}\right)}{200}}$

$0.635 \pm 2.57\sqrt{\frac{0.635(1-0.635)}{200}}$

$[0.548, 0.722]$