CSCI 3022

intro to data science with probability & statistics

Lecture 14
March 5, 2018

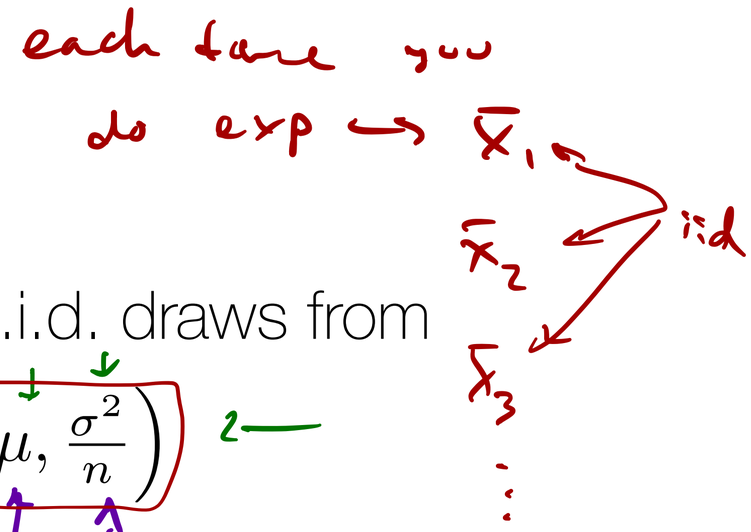
Introduction to Statistical Inference & Confidence Intervals

TONY Wed OH are cancelled

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Dan Larremore

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

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

Any old distribution

μ

Var $\sim \frac{1}{n}$

Examples:

Population? $p = 0.5$

Sample? $n = 50$

- **Example 2**: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?
- Note: this is a little different because we're estimating a *proportion*. What changes?

Population: $p = 0.5$
Sample: $n = 50$

$P(\bar{X} \geq 0.75) = 1 - P(X \leq 0.75)$

$\hat{p} = \frac{X}{n}$

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

$= \frac{1}{n^2} n p(1-p)$

Proportions are different. $\bar{X} = \hat{p} = \frac{Bin(n,p)}{n}$

What is variance of $\hat{p}$? $Var(\hat{p}) = Var\left(\frac{Bin(n,p)}{n}\right) = \frac{1}{n^2} Var(Bin(n,p))$

Last time, on CSCI 3022…

- **Example 2**: Suppose you have a jar of lemon and banana jelly beans where it is known that the true proportion of lemon jelly beans is 0.5. You try to estimate the proportion of lemon beans by reaching in and drawing 50 jelly beans and testing them (by eating them). What is the probability that your sample is 75% or more lemon jelly beans?

$P(p \geq 0.75)$

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?

next few weeks!

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 $\sigma/\sqrt{n}$ of our bag of $\bar{x}$
- "Standardizing" the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable.

$\frac{X-\mu}{\sigma}$ for CLT $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ Box-Muller Z

Question: how big does our sample need to be

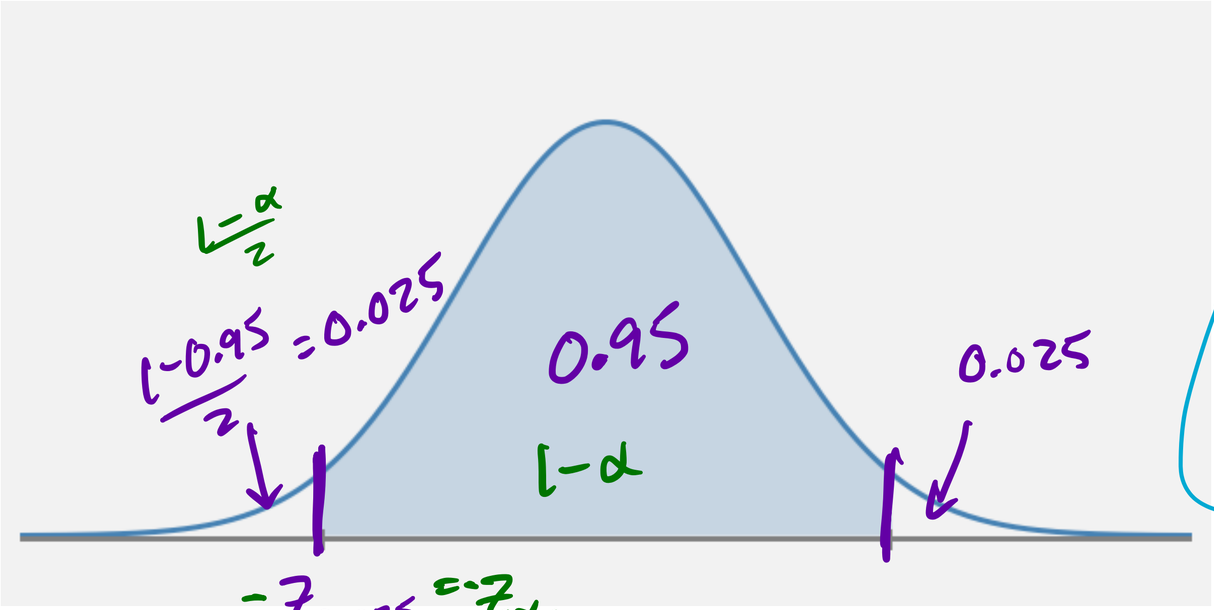
- …if the variable of interest is normally distributed?
- …if the variable of interest is not normally distributed?

Confidence Intervals

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

$\alpha = 0.05$

- 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



we don't know μ

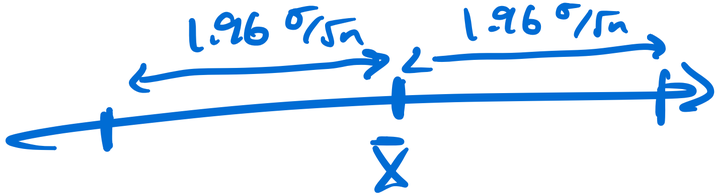
$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Random variable

- This is equivalent to:

$$0.95 = P\left(\underbrace{-1.96}_{-z_{0.025}} \le Z \le 1.96\right) = P\left(-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96\right)$$

Confidence Intervals



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

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

write as an interval:

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

middle?

LaTeX: \pm

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

RANDOM: $\bar{X}$

FIXED: n, σ, $\alpha \rightarrow z_{\alpha/2}$

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 $2 \times 1.96\frac{\sigma}{\sqrt{n}}$ which is **not** random; only the location of the interval's midpoint $\bar{X}$ is random.

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

$1.96 = z_{0.05/2}$

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

Interpreting the Confidence Interval

95%

$\bar{x} \rightarrow CI$

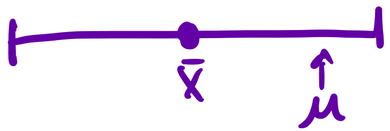
- **Statement**: We are 95% confident that the true population mean is in this interval.

- **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.

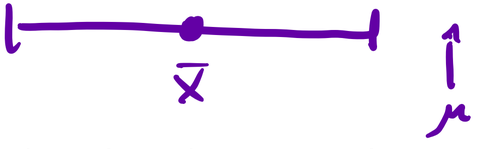
20 experiments:

IN THE LONG RUN:

$\frac{19}{20}$:



$\frac{1}{20}$:

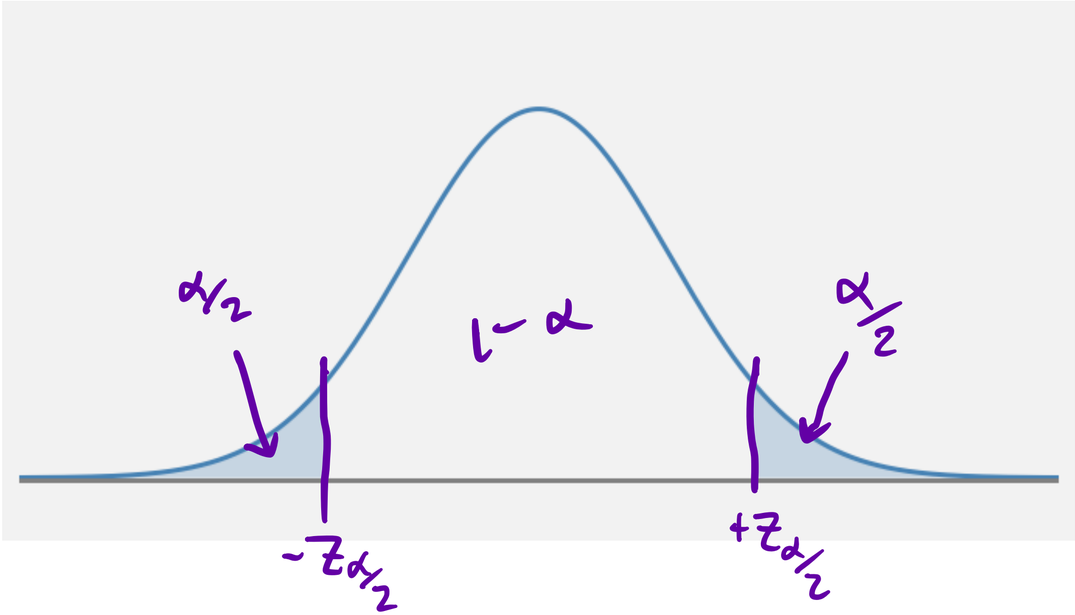


- The confidence level is not a statement about any one particular interval. Instead it describes what would happen if a very large number of CIs were computed using the same CI formula.

Other Levels of Confidence

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

$\alpha = 0.05$ (for a 95% CI)



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

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

#'s we measured; α handed to you; know; data

CI Example

$z_{0.05} = 1.645$

$[3.496, 3.704]$

- 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.**

$\bar{X}$ → 3.6 hours per day

σ → 2 hours per day

n = 1000

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

$\alpha = 0.1$ b/c 90% CI

$\alpha = 1 - .90$

what is $z_{\frac{0.1}{2}} = z_{0.05} = 1.645$

90% CI is $3.6 \pm 1.645 \cdot \frac{2}{\sqrt{1000}}$

90% CI = $[3.496, 3.704]$

CI Example

HERE

$z_{0.025} = 1.96$

[3.48, 3.72]

- 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.**

$\bar{x} \pm z_{0.05/2} \cdot \frac{\sigma}{\sqrt{1000}}$ → ... → [3.48, 3.72] = 95% CI

[3.50, 3.70] = 90% CI

$z_{0.05/2} = 1.96$

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

balance between True & useful info.

Test your understanding!

[3.48, 3.72] 95% CI

- **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?

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

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.

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?

$z_{\alpha/2} = 1.96$

CI: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\longrightarrow$ width $= 2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq 0.1$

$\downarrow$ rearrange for n

$$\sqrt{n}^2 \geq \left(2 \cdot \frac{1.96 \cdot 2}{0.1} \cdot 10\right)^2$$

$$\Rightarrow n \geq 4 \cdot 1.96^2 \cdot 4 \cdot 100$$

estimating

$$\approx 4 \cdot 4 \cdot 4 \cdot 100$$

$$\Rightarrow \boxed{n \geq 6400}$$

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

$\sigma \rightarrow s = \sqrt{\text{var}[x]}$

$= \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$

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

Just like we use data to est. μ using $\bar{x}$, use data to est. σ using s

- **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

Bernoulli r.v. ($Var[x_i] = p(1-p)$)

- Then X can be modeled as a Binomial random variable with:

$E[X] = np$

$Var[X] = n\,p(1-p)$

n trials

Bernoulli variance

Confidence intervals for proportions

"was Newton a ___"

- The estimator for p is given by:

$$\hat{p} = \frac{X}{n}$$

- The estimator is approximately normally distributed with:

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}np = p$$

$$Var[\hat{p}] = Var\left[\frac{X}{n}\right] = \frac{1}{n^2}Var[X]$$

$$= \frac{1}{n^2}np(1-p)$$

$$Var[\hat{p}] = \frac{p(1-p)}{n}$$

- Standardizing the estimate yields:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad \left(\frac{\hat{p} - E[\hat{p}]}{SD[\hat{p}]}\right)$$

- This gives us a confidence interval of:

central est. ↓ ↓ SD

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

Confidence intervals for proportions

$z_{0.005} = 2.57$

$\frac{127}{200} = 0.635$

Sample: n, $\hat{p}$

Population: ?? want p

- 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.

200 = ✓

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

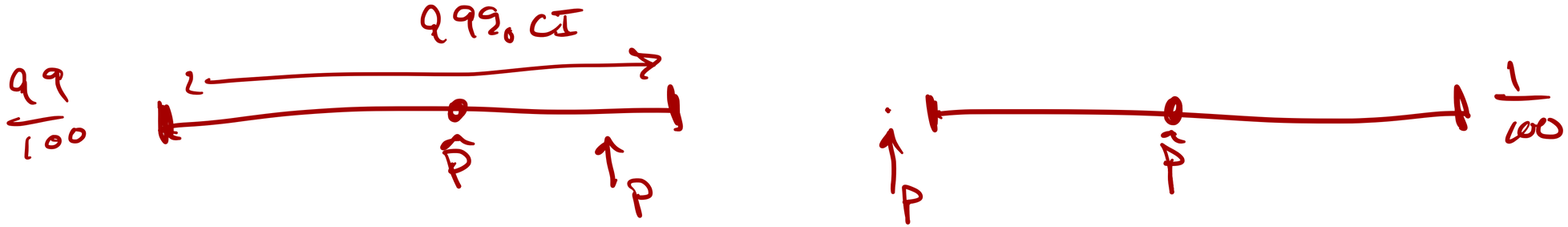
$= 0.635$

$z_{\frac{0.01}{2}}$

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

est. unknown p ($\hat{p}$ (our best guess)) vs.

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



Confidence intervals for proportions

Sample: $n = 200$ #"heads" $= 127$ Population: Nada!

- 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.

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

$\alpha = .01$

$z_{\alpha/2} = z_{\frac{0.01}{2}} = 2.57$ via $\text{qnorm}(0.995)$

$$\hat{p} = \frac{127}{200} = 0.635$$

$n = 200$

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

99% CI $\longrightarrow$ $[0.548, 0.722]$

Tie back to prob statement, last slide