Confidence Intervals

SML201: Introduction to Data Science, Spring 2019
Michael Guerzhoy
The q family of functions

> qbinom(p = .025, size = 100, prob = 0.5)
[1] 40

What is the largest number of heads (out of 100 tosses, with 50% probability of heads) such that \( pbinom(q = nH, size = 100, prob = 0.5) > 0.025? \)

> pbinom(q = 40, size = 100, prob = 0.5)
[1] 0.02844397
> pbinom(q = 39, size = 100, prob = 0.5)
[1] 0.0176001

> qnorm(p = 0.4, mean = 10, sd = 1)
[1] 9.746653
> pnorm(q = 9.75, mean = 10, sd = 1)
[1] 0.4012937
The q family of functions

• If we are sampling from $N(\mu, \sigma)$, what is the interval within which $\bar{X}$ will fall 95% of the time?

$$[\text{qbinom}(p = 0.025, \text{mean} = \mu, \text{sd} = \sigma / \sqrt{n}), \text{qbinom}(p = 0.975, \text{mean} = \mu, \text{sd} = \sigma / \sqrt{n})]$$

• Idea: the sample mean will fall below the point where $\text{pbinom}$ is smaller than 2.5% just 2.5% of the time (same with 97.5%, mutatis mutandis)
Confidence Intervals: Motivation

• Suppose we collect a sample of observed heights, and compute the sample mean
• If we take the sample mean to be an estimate of the population mean, how far away is our estimate from the true mean?
Confidence Interval

• If we construct a 95% confidence intervals (CI) repeatedly every time we collect a sample from the population, the CI must contain the true mean of the sample at least 95% of the time

• Note: this is not the same as saying that any particular CI contains the true mean with probability 95%
CI: Normal Distribution

• Suppose we know our $n$ samples come from a normal distribution with standard deviation $\sigma$. Then the 95% CI for $\mu$ will be

$$[\text{qnorm}(.025, \text{mean} = \text{mean}(x), \text{sd} = \sigma/\sqrt{n}), \text{qnorm}(.975, \text{mean} = \text{mean}(x), \text{sd} = \sigma/\sqrt{n})]$$

• Computationally: pretend $\bar{X}$ is the true mean, and construct a CI around it

• Note: the CI becomes smaller with larger $n$
CI: binomial distribution

• Use the normal approximation
• The number of heads is distributed according to
  \[ \text{Sum}(X) \sim N(np, np(1 - p)) \]
• The proportion of heads is distributed according to
  \[ X \sim N(p, p(1 - p)/n) \]
• The CI becomes smaller with larger n
• If p is not known, estimate \( p = 0.5 \)
  • Produces the widest CI
• With the approximation, the CI for the proportion \( p \) is
  \[
  \left( \text{qnorm}(0.025, \text{mean} = \text{mean}(x), \text{sd} = 0.5/\sqrt{n}), \text{qnorm}(0.975, \text{mean} = \text{mean}(x), \text{sd} = 0.5/\sqrt{n}) \right)
  \]
CI: binomial distribution

- The width of the CI is
  \[ q\text{norm}(0.975, \text{mean} = \text{mean}(x), \text{sd} = 0.5/\sqrt{n}) - q\text{norm}(0.025, \text{mean} = \text{mean}(x), \text{sd} = 0.5/\sqrt{n}) \]

- Half the width of the 95% is often known as the “margin of error”
**CI: binomial distribution**

- The margin of error for a poll with 100 observations is
  \[ qnorm(p = 0.975, \text{mean} = 0, \text{sd} = 0.5/\sqrt{100}) - qnorm(p = 0.025, \text{mean} = 0, \text{sd} = 0.5/\sqrt{100}) \]
  \[ 0.1959964 \]

- Margin of error: 10%
- A poll with 1000 observations will have a 3% margin of error
- A poll with 10000 observations will have a 1% margin of error
The University of New Hampshire's poll was conducted by speaking to 549 New Hampshire adults over the phone from April 10-18. The margin of error of the overall survey is 4.2 percent and 6.3 percent for the 241 likely Democrats surveyed. Read the full results here. —Brendan Morrow

> (qnorm(p = 0.975, mean = 0, sd = 0.5/sqrt(241))-qnorm(p = 0.025, mean = 0, sd = 0.5/sqrt(241)))/2
[1] 0.06312619
Example #1

• Students’ scores in a large class are normally distributed with $\sigma = 10$. What is the confidence interval for the mean if we have a sample of 50 students, with mean 80?
Example #1

• Students scores in a large class are normally distributed with $\sigma = 10$. What is the 95% confidence interval for the mean if we have a sample of 50 students, with mean 80.0?

```r
> c(qnorm(p = 0.025, mean = 80, sd = 10/sqrt(50)), qnorm(p = 0.975, mean = 80, sd = 10/sqrt(50)))
[1] 77.22819 82.77181
```
Example #2

• We take a poll of 60 students. 82% say yes, and 18% say no. What is the 95% confidence interval for the probability of “yes”? 


Example #2

• We take a poll of 60 students. 82% say yes, and 18% say no. What is the 95% confidence interval for the probability of “yes”?

```r
> c(qnorm(p = 0.025, mean = .82, sd = .5/sqrt(60)), qnorm(p = 0.975, mean = .82, sd = .5/sqrt(60)))
[1] 0.6934849 0.9465151
```
Normal distribution, unknown s.d.

• Students’ scores in a large class are normally distributed with $\sigma = 10$. What is the 95% confidence interval for the mean if we have a sample of 50 students, with mean 80, and sample standard deviation 7?

$$[\bar{x} + \frac{s}{\sqrt{n}} \times pt(0.025, df = 49), \bar{x} + \frac{s}{\sqrt{n}} \times pt(0.975, df = 49)]$$
Example #3

• We take a poll of 60 students. 82% say yes, and 18% say no. What is the confidence interval for the probability of “yes”? 
Example #3

• We take a poll of 60 students. 82% say yes, and 18% say no. What is the confidence interval for the probability of “yes”? 

```r
> c(qnorm(p = 0.05, mean = .82, sd = .5/sqrt(60)), 
  qnorm(p = 0.95, mean = .82, sd = .5/sqrt(60)))
[1] 0.7138252 0.9261748
```
95% CI, 90% CI, 70% CI...

• Higher confidence → wider CI
  • Need a wider CI if we want the true value to be in the CI more often

• For approximately normal distributions: length of the CI “arm” (half of the CI width) for a 95% CI is $1.96sd(\bar{X}) = \frac{1.96\sigma}{\sqrt{n}}$.
  (qnorm(.975, mean = 0, sd = 1)=1.96)

• For binomial distributions: length of the 95% CI “arm” is $1.96 \frac{.5}{\sqrt{n}} \approx 1/\sqrt{n}$
Confidence Intervals and Hypothesis Testing

• If the null hypothesis is true, the CI will contain the null hypothesis mean 95% of the time

• If the 95% CI does not contain the null hypothesis mean, we had something happen that doesn’t happen 95% of the time
  • P-value smaller than 5%

• Can use CIs to compute p-values
CI interpretation

• Not correct to say that the 95% CI contains the true mean with 95% probability
• The 95% CI will contain the true mean 95% of the time, if we collect multiple samples