SWARTHMORE

STAT011 Statistical Methods I

Lecture 11 Confidence Interval

Lu Chen
Swarthmore College
2/26/2019

Review

Central Limit Theorem

- Population distribution is Normal, $X \sim N(\mu, \sigma)$,

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Population distribution is not Normal, $\mu_X = \mu$, $\sigma_X = \sigma$,

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

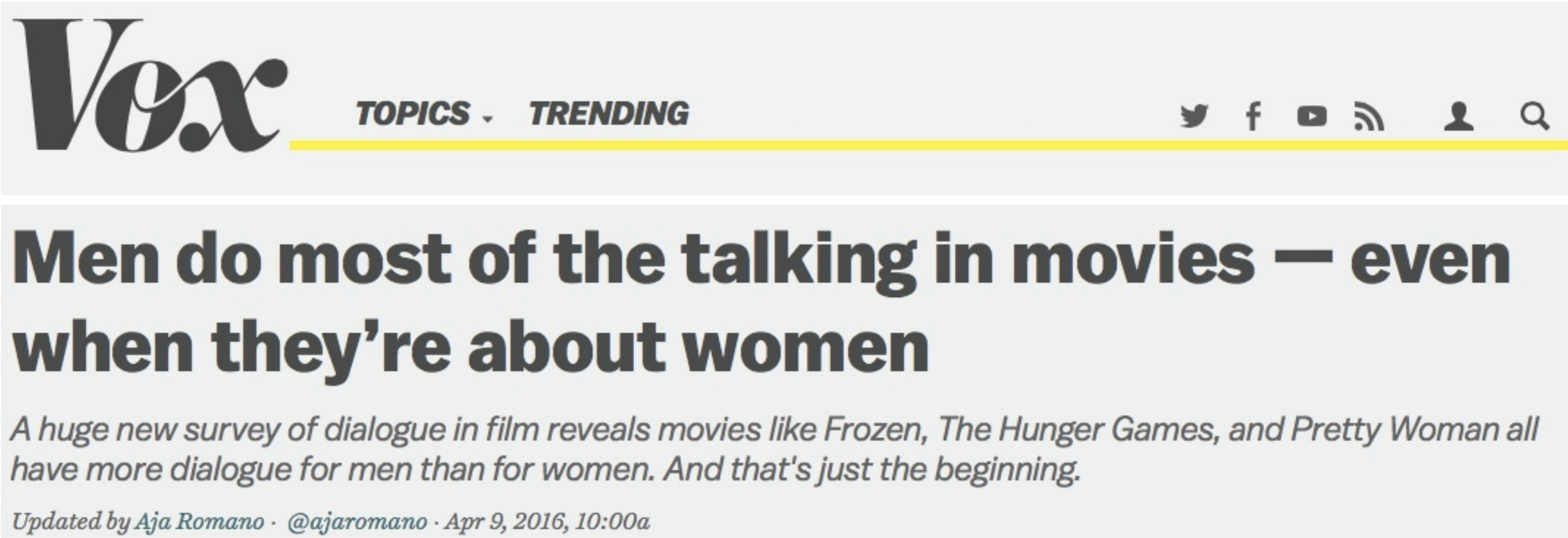
  - Population distribution is Bernoulli, $X \sim Bernoulli(p)$, $\mu_X = p$, $\sigma_X = \sqrt{p(1-p)}$,

$$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Outline

- Data example
- Statistical inference
- Confidence interval
  - A simple simulation study
  - Margin of error and critical points
  - Confidence interval for a population mean
- Calculating confidence intervals for
  - A population mean
  - A population proportion

Data example



Vox

TOPICS TRENDING

Men do most of the talking in movies — even when they're about women

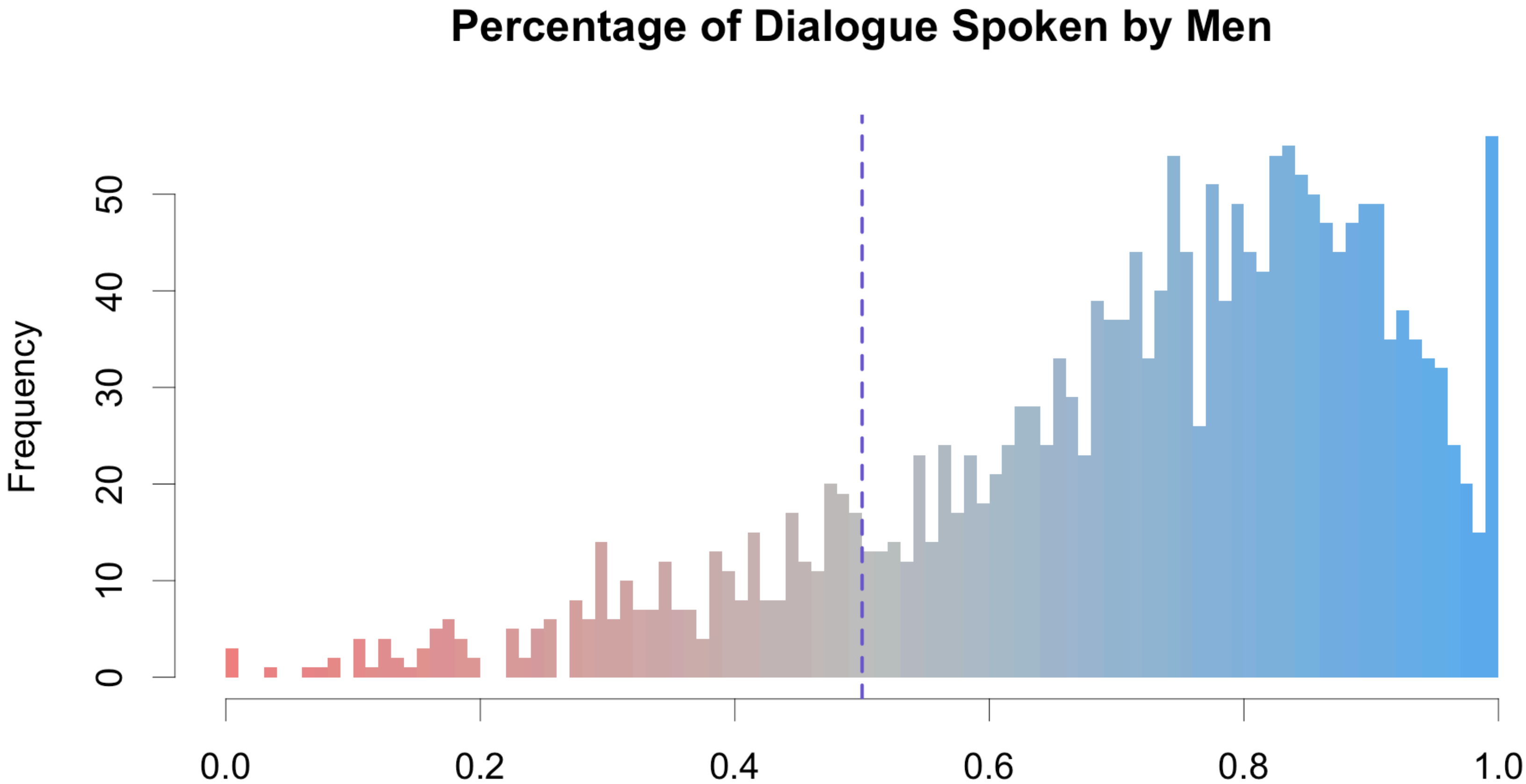
A huge new survey of dialogue in film reveals movies like Frozen, The Hunger Games, and Pretty Woman all have more dialogue for men than for women. And that's just the beginning.

Updated by Aja Romano · @ajaromano · Apr 9, 2016, 10:00a

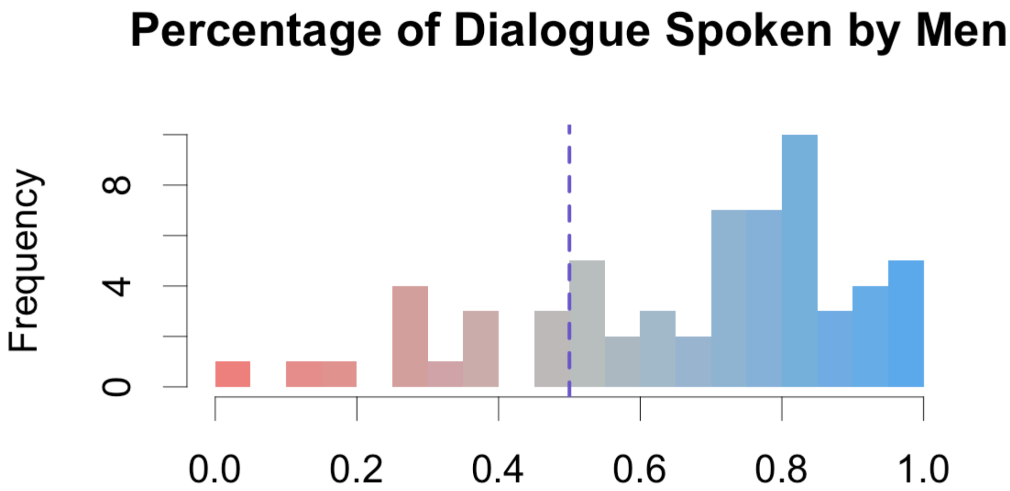
> "we compiled the number of words spoken by male and female characters across roughly 2,000 films, arguably the largest undertaking of script analysis, ever."

https://pudding.cool/2017/03/film-dialogue/

Data example - 2000 screenplays

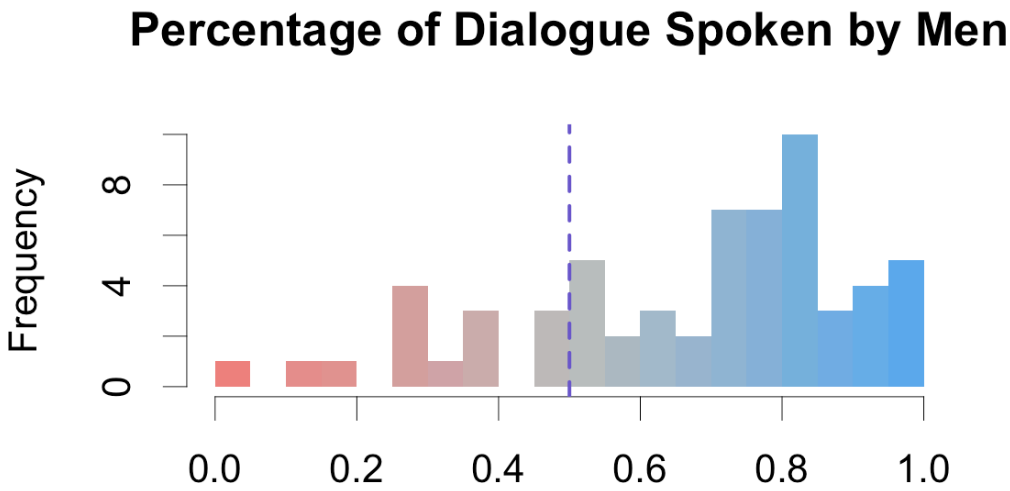


Data example - 62 screenplays in 2015



Star Wars: Episode VII - The Force Awakens
Inside Out
Minions
The Martian
The Big Short
Joy
Sicario
Pan
The Boy Next Door
Spotlight
Woman in Gold
Brooklyn
Ex Machina
Steve Jobs
...

Data example - 62 screenplays in 2015



- X: percentage of dialogue spoken by men.
- In the sample of 2015 movies ($n = 62$), mean percentage of dialogue spoken by men is $\bar{x} = 0.668$.
- **Question**: Statistically, do men truly speak more dialogue than women in 2015 movies?
- **Equivalent question**: What does $\bar{x}$ tell us about the population mean μ?

Data example - 62 screenplays in 2015

- Suppose variable X follows an unknown population distribution with mean μ (unknown) and standard deviation σ (let's assume σ is known *for now*). A sample of size n is generated from the population.
- $\bar{x}$ is the mean of the sample and the estimate of the population mean μ.
- By CLT,

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  - $\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
  - $\bar{x}$ *follows* an approximately Normal distribution with mean μ and SD $\frac{\sigma}{\sqrt{n}}$.
  - $\bar{x}$ *comes from* an approximately Normal distribution with mean μ and SD $\frac{\sigma}{\sqrt{n}}$.

Statistical inference

- What does $\bar{x}$ tell us about the population mean μ?

We answer this question using **statistical inference** methods. There are **two types of inference**:

- **Confidence interval**: assesses how well the sample statistic estimates the population parameter.
  - Lecture 11 (texbook Chapter 6.1)
- **Hypothesis testing**: assesses the evidence provided by the data in favor of some claim about the population parameter.
  - Lecture 12 (texbook Chapter 6.2)

Confidence interval

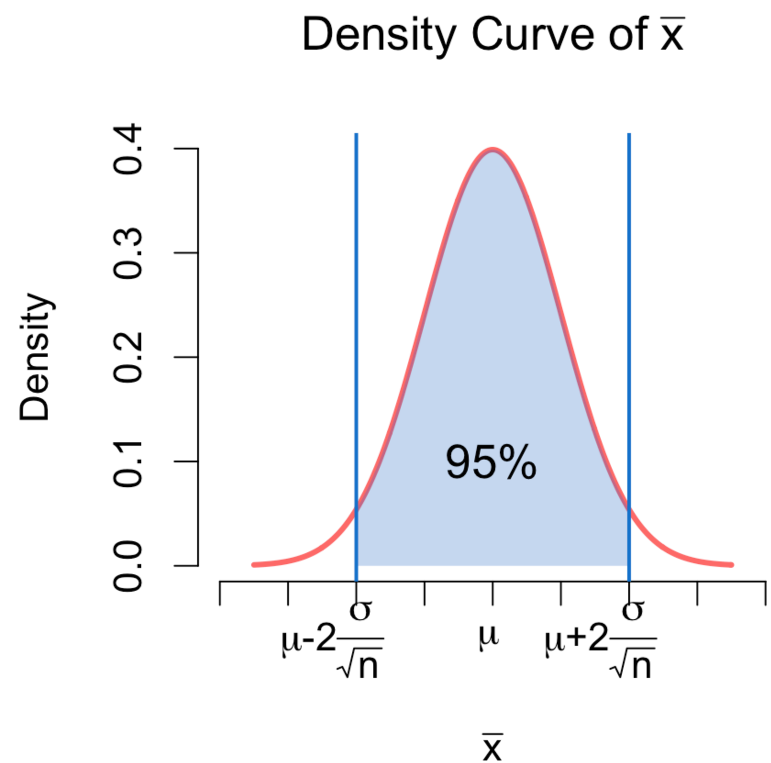
Population: mean μ (unknown), SD σ (known)

Sample: $\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

- By the 95 part of the 68-95-99.7 rule for Normal distribution, we have

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

- The following two statements are equivalent:
  - 95% of the sample mean $\bar{x}$ fall between $2\frac{\sigma}{\sqrt{n}}$ of μ.
  - The probability that $\bar{x}$ is greater than $\mu - 2\frac{\sigma}{\sqrt{n}}$ and less than $\mu + 2\frac{\sigma}{\sqrt{n}}$ is 0.95.



Confidence interval

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

- μ is unknown; $\sigma = 0.197, n = 62, \bar{x} = 0.668$
- $$P\left(\mu - 2 \times \frac{0.197}{\sqrt{62}} < 0.668 < \mu + 2 \times \frac{0.197}{\sqrt{62}}\right) \approx 0.95$$
- Since $\mu - 2 \times \frac{0.197}{\sqrt{62}} < 0.668 \iff \mu < 0.668 + 2 \times \frac{0.197}{\sqrt{62}}$ and $0.668 < \mu + 2 \times \frac{0.197}{\sqrt{62}} \iff \mu > 0.668 - 2 \times \frac{0.197}{\sqrt{62}}$
- $$P\left(0.668 - 2 \times \frac{0.197}{\sqrt{62}} < \mu < 0.668 + 2 \times \frac{0.197}{\sqrt{62}}\right) \approx 0.95$$
- $P(0.618 < \mu < 0.718) \approx 0.95$ What does $\bar{x}$ tell us about the population mean μ?

Confidence interval

For

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

μ is unknown; σ, n and $\bar{x}$ are known. Then

$$P\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

We call this interval $\left[\bar{\boldsymbol{x}} - \mathbf{2}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}, \bar{\boldsymbol{x}} + \mathbf{2}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right]$ the 95% **confidence interval** for the unknown population mean μ.

- What does $\bar{x}$ tell us about the population mean μ?

Confidence interval

95% Confidence interval for population mean μ

$$P\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Wrong interpretation**:
  - 95% of μ fall between $\bar{x} - 2\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + 2\frac{\sigma}{\sqrt{n}}$.
  - The probability that μ is greater than $\bar{x} - 2\frac{\sigma}{\sqrt{n}}$ and less than $\bar{x} + 2\frac{\sigma}{\sqrt{n}}$ is 0.95.
- The **key problem** of the interpretation: it implies that μ is changing while $\bar{x}$ is fixed; however, the population parameter μ is **fixed** but the sample mean $\bar{x}$ **changes** from sample to sample.

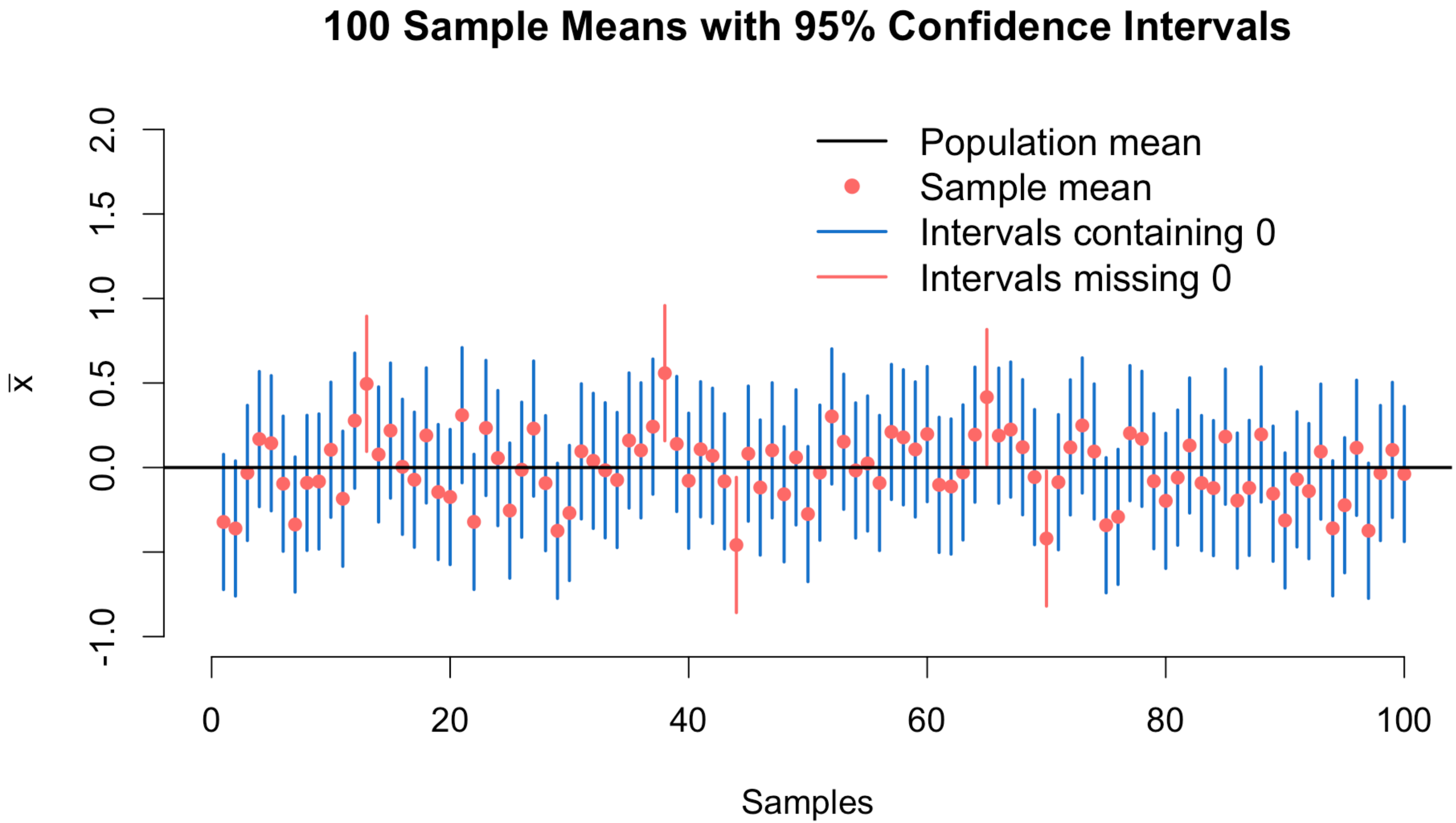
A simple simulation study

1. Generate 100 samples of size 25 from a population with mean 0 and SD 1.
   - $\mu = 0, \sigma = 1, n = 25$
2. Calculate the mean $\bar{x}$ for each sample and $\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N(0, 0.2)$
3. Compute the 95% confidence intervals for each sample.
   - $\bar{x} - 2\frac{\sigma}{\sqrt{n}} = \bar{x} - 0.4$
   - $\bar{x} + 2\frac{\sigma}{\sqrt{n}} = \bar{x} + 0.4$
4. Count how many intervals in step 3 contain $\mu = 0$.

```r
set.seed(10); n <- 25; mean_x <- NULL
for(i in 1:100){
  x <- rnorm(n) # population dist N(0, 1)
  mean_x[i] <- mean(x)
}
sum(mean_x - 0.4 < 0 & mean_x + 0.4 > 0)
```

```
## [1] 95
```

A simple simulation study



A simple simulation study

- 95 out of the 100 intervals $[\bar{x} - 0.4, \bar{x} + 0.4]$ contain $\mu = 0$.
- Guess how many of the 100 intervals $[\bar{x} - 0.2, \bar{x} + 0.2]$ or $[\bar{x} - 0.6, \bar{x} + 0.6]$ will contain $\mu = 0$?
- $P(\bar{x} - 0.2 < \mu < \bar{x} + 0.2) = 0.68$
  About 68 out of the 100 intervals will cover μ (this simulation has 70)
- $P(\bar{x} - 0.4 < \mu < \bar{x} + 0.4) = 0.95$
  About 95 out of the 100 intervals will cover μ (this simulation has 95)
- $P(\bar{x} - 0.6 < \mu < \bar{x} + 0.6) = 0.997$
  About 99.7 out of the 100 intervals will cover μ (this simulation has 100)
- **In practice**, we only have **one sample**. How could this single confidence interval help us?

A simple simulation study

- For example, in the simulation, the first sample has $\bar{x} = -0.32$ and 95% confidence interval $[-0.72, 0.08]$
- By theory and simulation, **the method** we used to compute this interval should produce intervals that contain μ 95% (most) of the time.
- Therefore, we are **pretty confident** that this specific interval $[-0.72, 0.08]$ does contain μ.
- Our confidence is about 95%.
- **Correct interpretation**:
  - We are 95% confident (about the method) that the interval $[-0.72, 0.08]$ will contain the true population mean μ.
  - Note: our confidence is NOT about whether μ is in the interval or not; our confidence is about the method that will produce an interval that contains μ.

Confidence interval

A **level C confidence interval** for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.

- It is an interval **for the population parameter**.
- We usually want a relatively large C value.
- C cannot be too small, otherwise the interval will be too narrow to contain a true population parameter (we have very high chance of missing the parameter).
- C cannot too large, otherwise the interval will be too wide that it tells a little about the true population paramter (we will not miss the parameter but the parameter could be anywhere).
- C usually takes values 0.9, 0.95 and 0.99.

Confidence interval

$$P\left(\bar{x} - ?\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + ?\frac{\sigma}{\sqrt{n}}\right) \approx C$$

- When $C = 0.95$, $? = 2$. How to find the value of "?" when $C = 0.9$ or 0.99?

```
qnorm(1-(1-0.95)/2) # C = 0.95
```

```
## [1] 1.959964
```

```
qnorm(1-(1-0.9)/2) # C = 0.9
```

```
## [1] 1.644854
```

```
qnorm(1-(1-0.99)/2) # C = 0.99
```

```
## [1] 2.575829
```

- Here the "?" value is denoted as z^* and called the **critical point**.
- And $z^* \frac{\sigma}{\sqrt{n}}$ is denoted as m and called **margin of error**.

Confidence interval

Confidence level C	0.9	0.95	0.99
Critical point z^*	1.645	1.960	2.576

For the simple simulation, $\sigma = 1, n = 25$

- $C = 0.9$, margin of error $m = 1.645\frac{\sigma}{\sqrt{n}} = 1.645 \times \frac{1}{\sqrt{25}} = 0.33$
- $C = 0.95$, margin of error $m = 1.960\frac{\sigma}{\sqrt{n}} = 0.39$
  - Sometimes we simply use $z^* = 2$ and then $m = 0.4$
- $C = 0.99$, margin of error $m = 2.576\frac{\sigma}{\sqrt{n}} = 0.52$

The **larger** C, the larger z^* and m, the **wider** the confidence interval.

Confidence interval for a population mean

Choose an SRS of size n from a population having unknown mean μ and known standard deviation σ. The margin of error for a level C confidence interval for μ is

$$m = z^* \frac{\sigma}{\sqrt{n}}.$$

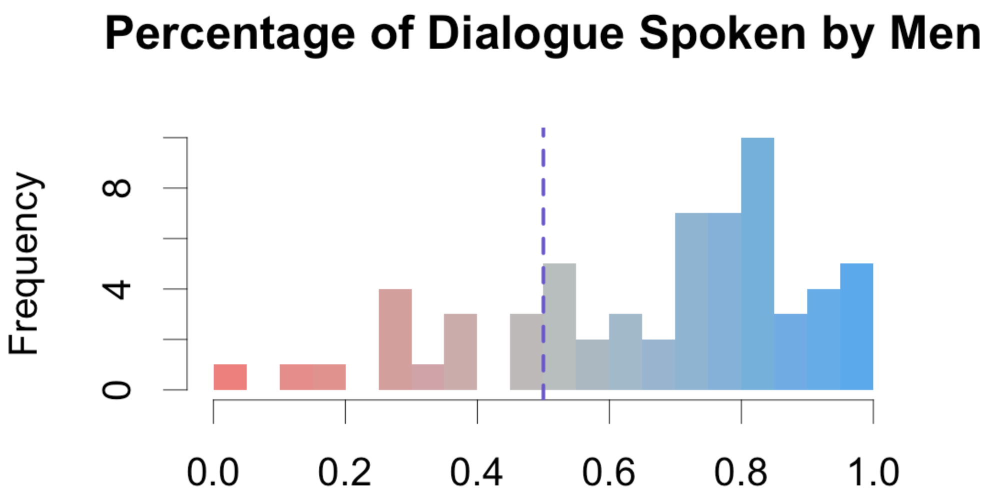
Here, z^* is the value on the standard Normal curve with area C between the critical points $-z^*$ and z^*. The **level C confidence interval for μ** is

$$\bar{x} \pm m = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

The confidence level of this interval is exactly C when the population distribution is Normal and is approximately C when n is large in other cases.

Data example - 62 screenplays in 2015

Question: Statistically, do men truly speak more dialogue than women in 2015 movies?



- This means that 0.5 is extremely unlikely to be the true population mean. Therefore, statistically, men speak more in 2015 movies.

- $\bar{x} = 0.668, n = 62, \sigma = 0.197$
- $C = 0.95, z^* = 1.96$
- $m = z^* \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{0.197}{62} = 0.049$
- The 95% confidence interval for μ is 0.668 ± 0.049 or $[0.619, 0.717]$.
- We are 95% confident about the method that the interval $[0.619, 0.717]$ will contain the true population average of percentage of dialogue spoken by men in 2015 movies.
- 99% CI: $[0.604, 0.732]$.

Data example - Female height

The average height of the 58 female students in the 2019 STAT 11 class is 64.9 inches. Suppose the population standard deviation of female height is 3 inches. Calculate the 90% and 95% confidence intervals for the population mean of female height.

- $\bar{x} = 64.9, n = 58, \sigma = 3.$
- $z^* = 1.645$ for $C = 0.9$ and $z^* = 1.96$ for $C = 0.95.$
- The 90% CI is $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 64.9 \pm 1.645 \times \frac{3}{\sqrt{58}} = 64.9 \pm 0.6$ or $[64.3, 65.5]$.
- The 95% CI is $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 64.9 \pm 1.96 \times \frac{3}{\sqrt{58}} = 64.9 \pm 0.8$ or $[64.1, 65.7]$.
- We are 90% (95%) confident about the method that the interval $[64.3, 65.5]$ ($[64.1, 65.7]$) will contain the true population mean of female height in the STAT 11 class.

Data example - Coin toss

Suppose a student tossed a coin 20 times and got 7 heads. Is it a fair coin?

- $\hat{p} = \frac{7}{20} = 0.35, n = 20, \sigma = \sqrt{p(1-p)} = ?$
- To calculate the CIs for proportions, when the population proportion is unknown, we use the sample proportion to compute the stardard deviation. Therefore, $\sigma = \sqrt{\hat{p}(1-\hat{p})} = \sqrt{0.35 \times (1-0.35)} = 0.48.$
- 95% CI for p is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.35 \pm 1.96 \times \sqrt{\frac{0.35\times(1-0.35)}{20}} = 0.35 \pm 0.21$or $[0.14, 0.56]$.
- We are 95% confident about the method that the interval $[0.14, 0.56]$ will contain the true population proportion of head when tossing this coin. Since this interval does contain 0.5, it is very likely a fair coin.

Summary

- Data example
- Statistical inference: *Confidence interval* and *hypothesis testing*
- Confidence interval
  - A simple simulation study
  - Margin of error $m = z^* \frac{\sigma}{\sqrt{n}}$ and critical points z^* and $-z^*$
  - Confidence interval for a population mean
- Calculating confidence intervals for
  - A population mean $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
  - A population proportion $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$