Probability & Statistics for DS & AI

Confidence Intervals

Michele Guindani

Summer

How good is our estimate?

- An estimator $\widehat{\Theta}$ is a function of the samples $X_1, \ldots, X_N$:

$$\widehat{\Theta} = g\left(X_1, \ldots, X_N\right)$$

 By construction, $\widehat{\Theta}$ is a random variable because it is a function of the random samples.

- $\widehat{\Theta}$ is a random variable. It has a PDF, CDF, mean, variance, etc. It has a distribution!!

How good is our estimate?

- An estimator $\widehat{\Theta}$ is a function of the samples $X_1, \ldots, X_N$:

$$\widehat{\Theta} = g\left(X_1, \ldots, X_N\right)$$

  By construction, $\widehat{\Theta}$ is a random variable because it is a function of the random samples.

- $\widehat{\Theta}$ is a random variable. It has a PDF, CDF, mean, variance, etc. It has a distribution!!

- Since $\widehat{\Theta}$ is a random variable, we should report both the value the estimator has assumed in the sample we have and the some information about the variability of this estimator (confidence) when reporting it.

How good is our estimate?

- An estimator $\widehat{\Theta}$ is a function of the samples $X_1, \ldots, X_N$:

$$\widehat{\Theta} = g\left(X_1, \ldots, X_N\right)$$

  By construction, $\widehat{\Theta}$ is a random variable because it is a function of the random samples.

- $\widehat{\Theta}$ is a random variable. It has a PDF, CDF, mean, variance, etc. It has a distribution!!
- Since $\widehat{\Theta}$ is a random variable, we should report both the value the estimator has assumed in the sample we have and the some information about the variability of this estimator (confidence) when reporting it.
- The confidence measures the quality of $\widehat{\Theta}$ when compared to the true parameter θ.
- If $\widehat{\Theta}$ fluctuates a great deal (has high variance) we may not be confident of our estimates.

Example

A class of 1000 students took a test. The distribution of the score is roughly a Gaussian with mean 50 and standard deviation 20. A teaching assistant was too lazy to calculate the true population mean. Instead, he sampled a subset of 5 scores listed as follows:

Student ID	1	2	3	4	5
Scores	11	97	1	78	82

He calculated the average, which is 53.8. This is a very good estimate of the class mean (which is 50). What is wrong with their procedure?

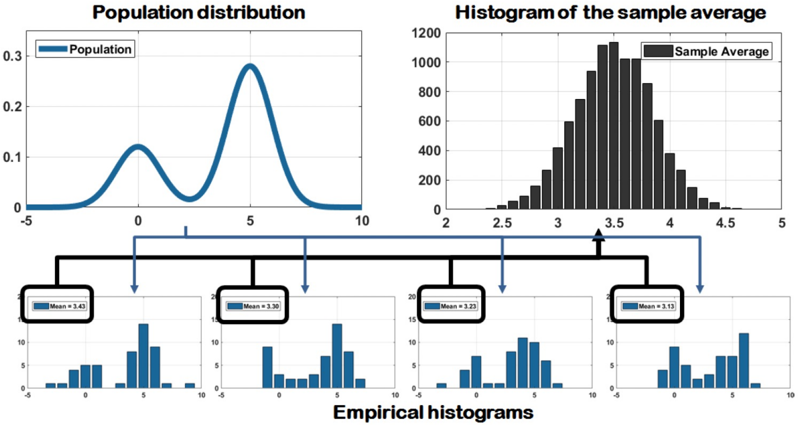
Example

Solution He was just lucky. It quite possible that if he sampled another 5 scores, he would get something very different. For example, if he looks at the 11 to 15 student scores, he could get:

Student ID	11	12	13	14	15
Scores	44	29	19	27	15

In this case the average is 26.8.

⚠ Both 53.8 and 26.8 are legitimate estimates, but they are the random realizations of a random variable $\widehat{\Theta}$.
This $\widehat{\Theta}$ has a PDF, CDF, mean, variance, etc. It may be misleading to simply report the estimated value from a particular instant, so the confidence of the estimator must be specified.



Pictorial illustration of the randomness of the estimator $\widehat{\Theta}$. Given a population, our datasets are usually a subset of the population. Computing the sample average from these finite-sample distributions introduces the randomness to $\widehat{\Theta}$. If we plot the histogram of the sample averages, we will obtain a distribution. The mean of this distribution is the population mean, but there is a nontrivial amount of fluctuation. The purpose of the concept of confidence interval is to quantify this fluctuation.

Confidence Interval

- The confidence interval is a probabilistic statement about the variability of the estimator $\widehat{\Theta}$.

Confidence Interval

- The confidence interval is a probabilistic statement about the variability of the estimator $\widehat{\Theta}$.
- Let's consider the random interval:

$$\mathcal{I} = [\widehat{\Theta} - \epsilon, \quad \widehat{\Theta} + \epsilon]$$

  for some ϵ to be determined
- **Why random?** This interval will assume different values in different samples, because we will have different estimates $\widehat{\theta}$ in different samples.

Confidence Interval

- The confidence interval is a probabilistic statement about the variability of the estimator $\widehat{\Theta}$.
- Let's consider the random interval:

$$\mathcal{I} = [\widehat{\Theta} - \epsilon, \quad \widehat{\Theta} + \epsilon]$$

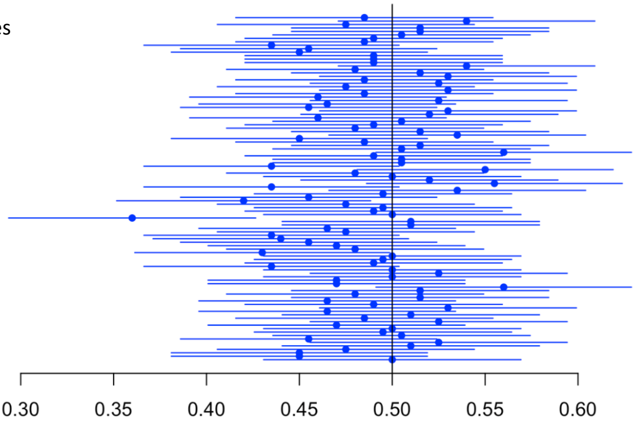
for some ϵ to be determined

- **Why random?** This interval will assume different values in different samples, because we will have different estimates $\widehat{\theta}$ in different samples.

Experiment: throw a coin 200 times ($n=200$)
and count the number of heads, y.

In each experiment,
the probability of head can be estimated by **y/n (blue dot)**

Repeat the experiment 100 times



Sometimes, the intervals we build include the true value
of the parameter of interest (probability of success); Other times, they don't.

95% Confidence Interval

95% Confidence Interval

If we assume the distribution of the estimator $\widehat{\Theta}$ to be available,
a 95% confidence interval fixes the value of ϵ in the previous interval, such that the probability

$$\mathbb{P}[\widehat{\Theta} - \epsilon \leq \theta \leq \widehat{\Theta} + \epsilon]$$

95% Confidence Interval

If we assume the distribution of the estimator $\widehat{\Theta}$ to be available,
a 95% confidence interval fixes the value of ϵ in the previous interval, such that the probability

$$\mathbb{P}[\widehat{\Theta} - \epsilon \leq \theta \leq \widehat{\Theta} + \epsilon] = \mathbb{P}[\theta - \epsilon \leq \widehat{\Theta} \leq \theta + \epsilon]$$

95% Confidence Interval

If we assume the distribution of the estimator $\widehat{\Theta}$ to be available,
a 95% confidence interval fixes the value of ϵ in the previous interval, such that the probability

$$\mathbb{P}[\widehat{\Theta} - \epsilon \leq \theta \leq \widehat{\Theta} + \epsilon] = \mathbb{P}[\theta - \epsilon \leq \widehat{\Theta} \leq \theta + \epsilon] \geq 0.95$$

95% Confidence Interval

If we assume the distribution of the estimator $\widehat{\Theta}$ to be available,
a 95% confidence interval fixes the value of ϵ in the previous interval, such that the probability

$$\mathbb{P}[\widehat{\Theta} - \epsilon \leq \theta \leq \widehat{\Theta} + \epsilon] = \mathbb{P}[\theta - \epsilon \leq \widehat{\Theta} \leq \theta + \epsilon] \geq 0.95$$

Remark: Of course the choice of 95% is completely arbitrary (although it is common. In general we could define a $(1-\alpha)\%$ confidence interval for any $(1-\alpha)$ confidence level, with $\alpha \in (0,1)$.

95% Confidence Interval

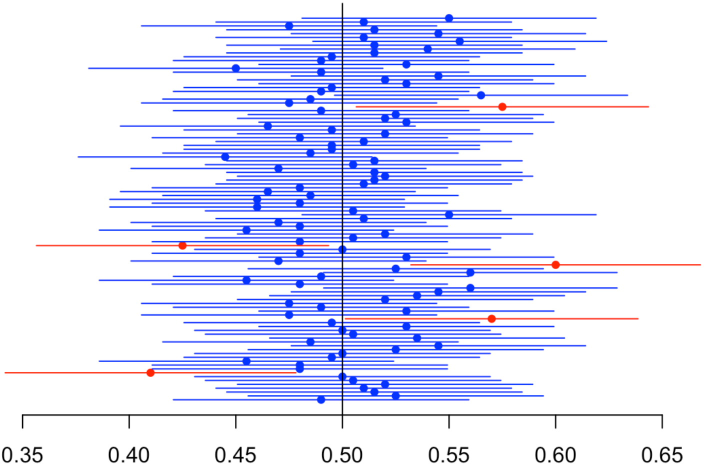
If we assume the distribution of the estimator $\widehat{\Theta}$ to be available, a 95% confidence interval fixes the value of ϵ in the previous interval, such that the probability

$$\mathbb{P}[\widehat{\Theta} - \epsilon \leq \theta \leq \widehat{\Theta} + \epsilon] = \mathbb{P}[\theta - \epsilon \leq \widehat{\Theta} \leq \theta + \epsilon] \geq 0.95$$

Remark: Of course the choice of 95% is completely arbitrary (although it is common. In general we could define a $(1-\alpha)\%$ confidence interval for any $(1-\alpha)$ confidence level, with $\alpha \in (0,1)$.

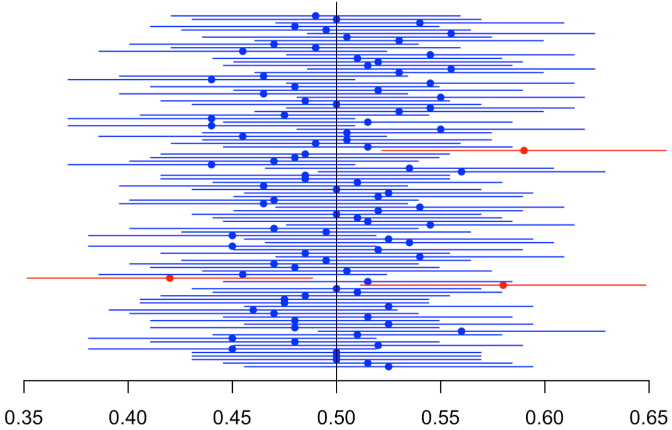
- So, a 95% confidence interval is a random interval $[\widehat{\Theta} - \epsilon, \widehat{\Theta} + \epsilon]$ such that there is 95% probability for it to include the true parameter θ.

Interpretation



On average, we expect that the deterministic interval build from a specific sample will contain the true value of the parameter 95% of the times.
Note that in reality, we don't know if the deterministic interval actually contains p)

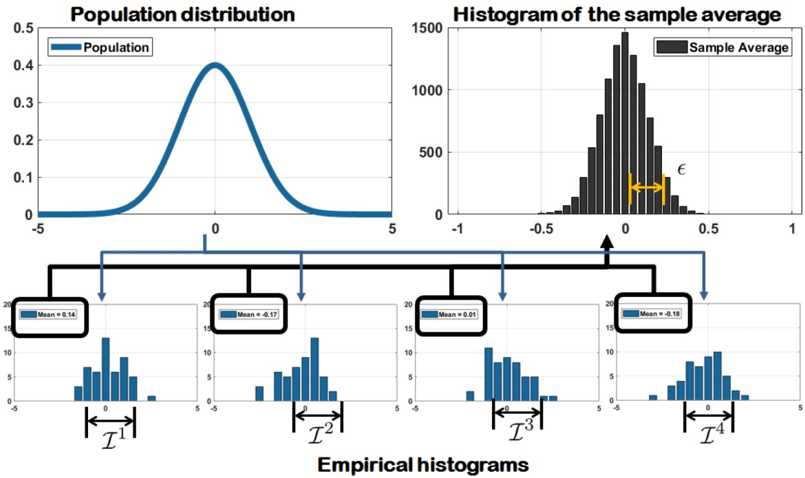
Interpretation



On average, we expect that the deterministic interval build from a specific sample will contain the true value of the parameter 95% of the times.

Note that in reality, we don't know if the deterministic interval actually contains p)

How to build a confidence interval



Conceptual illustration of how to construct a confidence interval. Starting with the population, we draw random subsets. Each random subset gives us an estimator, and correspondingly an interval.

Example (95% CI for a Normal - **known** variance)

- Suppose that we have a set of i.i.d. observations $X_1, \ldots, X_N$ from a Normal with unknown mean θ and **known** variance σ^2.

- We consider the maximum-likelihood estimator, which is the sample average:

$$\widehat{\Theta} = \frac{1}{N} \sum_{n=1}^{N} X_n$$

Example (95% CI for a Normal - **known** variance)

- Suppose that we have a set of i.i.d. observations $X_1, \ldots, X_N$ from a Normal with unknown mean θ and **known** variance σ^2.

- We consider the maximum-likelihood estimator, which is the sample average:

$$\widehat{\Theta} = \frac{1}{N} \sum_{n=1}^{N} X_n$$

- Since the data are normally distributed, we could prove that

$$\widehat{\Theta} \sim N(\mu, \frac{\sigma^2}{N})$$

Example (95% CI for a Normal - **known** variance)

- Suppose that we have a set of i.i.d. observations $X_1, \ldots, X_N$ from a Normal with unknown mean θ and **known** variance σ^2.

- We consider the maximum-likelihood estimator, which is the sample average:

$$\widehat{\Theta} = \frac{1}{N} \sum_{n=1}^{N} X_n$$

- Since the data are normally distributed, we could prove that

$$\widehat{\Theta} \sim N(\mu, \frac{\sigma^2}{N})$$

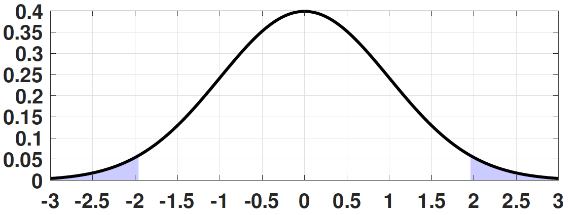
- Recall that we can always standardize a gaussian random variable

$$\widehat{Z} = \frac{\widehat{\Theta} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0, 1)$$

Example (95% CI for a Normal - **known** variance (ctd))

- So, if we look at a standard normal, the 95% CI determines ϵ such that

$$\underbrace{\mathbb{P}[|\widehat{Z}| \leq \epsilon]}_{\text{two tails of a standard Gaussian}} \geq 1 - \alpha$$



PDF of the random variable $\widehat{Z} = (\widehat{\Theta} - \mathbb{E}[\widehat{\Theta}])/\sqrt{\mathrm{Var}[\widehat{\Theta}]}$. The shaded area denotes the $\alpha = 0.05$ confidence level.

S Since $\mathbb{P}[\widehat{Z} \leq \epsilon]$ is the CDF of a Gaussian, it follows that

$$\begin{aligned}\mathbb{P}[|\widehat{Z}| \leq \epsilon] &= \mathbb{P}[-\epsilon \leq \widehat{Z} \leq \epsilon] \\ &= \mathbb{P}[\widehat{Z} \leq \epsilon] - \mathbb{P}[\widehat{Z} \leq -\epsilon] \\ &= \Phi(\epsilon) - \Phi(-\epsilon)\end{aligned}$$

- Using the symmetry of the Gaussian, it follows that $\Phi(-\epsilon) = 1 - \Phi(\epsilon)$ and hence

$$\mathbb{P}[|\widehat{Z}| \leq \epsilon] = 2\Phi(\epsilon) - 1$$

- If we ask $\mathbb{P}[|\widehat{Z}| \leq \epsilon] = 1 - \alpha$ then

$$\epsilon = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

```
# Python code to compute the width of the confidence interval
import scipy.stats as stats
alph = 0.05;
mu = 0; sigma = 1; # Standard Gaussian
epsilon = stats.norm.ppf(1-alph/2, mu, sigma)
print(epsilon)
#1.959963984540054   approx =1.96
```

```
# Python code to compute the width of the confidence interval
import scipy.stats as stats
alph = 0.05;
mu = 0; sigma = 1; # Standard Gaussian
epsilon = stats.norm.ppf(1-alph/2, mu, sigma)
print(epsilon)
#1.959963984540054   approx =1.96
```

$$
\begin{aligned}
0.95 &\approx \mathbb{P}[-1.96 \le \widehat{Z} \le 1.96], \quad (\widehat{Z} \text{ is within 1.96 std from the mean of } \widehat{Z}) \\
&= \mathbb{P}\left[-1.96 \le \frac{\widehat{\Theta} - \theta}{\sigma/\sqrt{N}} \le 1.96\right] \\
&= \mathbb{P}\left[\theta - 1.96\frac{\sigma}{\sqrt{N}} \le \widehat{\Theta} \le \theta + 1.96\frac{\sigma}{\sqrt{N}}\right]
\end{aligned}
$$

Example

Suppose that the number of photos a Facebook user uploads per day is a random variable with $\sigma = 2$. In a set of 341 users, the sample average is 2.9. Find the 95% confidence interval of the population mean.

Example

Suppose that the number of photos a Facebook user uploads per day is a random variable with $\sigma = 2$. In a set of 341 users, the sample average is 2.9. Find the 95% confidence interval of the population mean.

The 50% confidence interval is then

$$\left[\widehat{\theta} - 1.96\frac{2}{\sqrt{341}}, \quad \widehat{\theta} + 1.96\frac{2}{\sqrt{341}}\right] = [2.69, 3.11]$$

Therefore, with 95% confidence (not probability, like the book says), the interval $[2.69, 3.11]$ includes the population mean.

```python
import numpy as np
import scipy.stats as stats
N=341
Theta_hat=2.9
S_hat=2 #standard deviation
alpha = 0.05
z=1.96
CI_L = Theta_hat-z*S_hat/np.sqrt(N)
CI_U = Theta_hat+z*S_hat/np.sqrt(N)
print(CI_L, CI_U)

#Alternatively,
stats.norm.interval(0.95, loc=2.9, scale=2/np.sqrt(341))
```

```
2.687720098459038 3.112279901540962

(2.687723999152044, 3.112276000847956)
```

Example

Suppose that the number of photos a Facebook user uploads per day is a random variable with $\sigma = 2$. In a set of 341 users, the sample average is 2.9. Find the 90% confidence interval of the population mean.

Example

Suppose that the number of photos a Facebook user uploads per day is a random variable with $\sigma = 2$. In a set of 341 users, the sample average is 2.9. Find the 90% confidence interval of the population mean.

The 90% confidence interval is then

$$\left[\widehat{\theta} - 1.64\frac{2}{\sqrt{341}}, \quad \widehat{\theta} + 1.64\frac{2}{\sqrt{341}}\right] = [2.72, 3.08]$$

Therefore, with 90% confidence (not probability, like the book says), the interval $[2.72, 3.08]$ includes the population mean.

What if the variance is not known?

What if the variance is not known?

- We not to use also the variance estimator $\widehat{S}$, which can be defined as

$$\widehat{S}^2 \stackrel{\text{def}}{=} \frac{1}{N-1} \sum_{n=1}^{N} \left(X_n - \widehat{\Theta} \right)^2$$

What if the variance is not known?

- We not to use also the variance estimator $\widehat{S}$, which can be defined as

$$\widehat{S}^2 \stackrel{\text{def}}{=} \frac{1}{N-1} \sum_{n=1}^{N} \left(X_n - \widehat{\Theta} \right)^2$$

⚠ The distribution of $T \stackrel{\text{def}}{=} \frac{\widehat{\Theta}-\theta}{\widehat{S}/\sqrt{N}}$ is NOT a Gaussian anymore

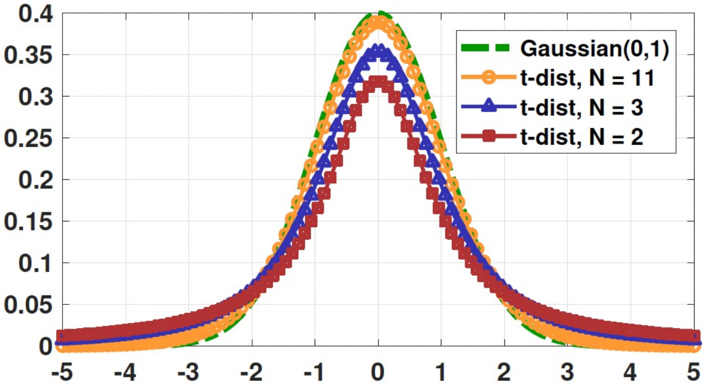
What if the variance is not known?

- We not to use also the variance estimator $\widehat{S}$, which can be defined as

$$\widehat{S}^2 \stackrel{\text{def}}{=} \frac{1}{N-1} \sum_{n=1}^{N} \left(X_n - \widehat{\Theta} \right)^2$$

⚠ The distribution of $T \stackrel{\text{def}}{=} \frac{\widehat{\Theta}-\theta}{\widehat{S}/\sqrt{N}}$ is NOT a Gaussian anymore

⇨ It's a Student's t-distribution with N-1 "degrees of freedom"



The PDF of Student's t-distribution with $\nu = N - 1$ degrees of freedom.

- If we want $\mathbb{P}\left[|T| \leq z_\alpha\right] = 1 - \alpha$, it follows that

$$z_\alpha = \Psi_\nu^{-1}\left(1 - \frac{\alpha}{2}\right)$$

- If we want $\mathbb{P}\left[|T| \leq z_\alpha\right] = 1 - \alpha$, it follows that

$$z_\alpha = \Psi_\nu^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Example

A class of 10 students took a midterm exam. Their scores are given in the following table.

Student	1	2	3	4	5	6	7	8	9	10
Score	72	69	75	58	67	70	60	71	59	65

Find the 95% confidence interval.

```python
# Python code to generate a confidence interval
import numpy as np
import scipy.stats as stats
x = np.array([72, 69, 75, 58, 67, 70, 60, 71, 59, 65])
N = x.size
Theta_hat = np.mean(x) # Sample mean
S_hat = np.std(x) # Sample standard deviation
nu = x.size-1 # degrees of freedom
alpha = 0.05 # confidence level
z = stats.t.ppf(1-alph/2, nu)
CI_L = Theta_hat-z*S_hat/np.sqrt(N)
CI_U = Theta_hat+z*S_hat/np.sqrt(N)
print(CI_L, CI_U)


#Alternatively
stats.t.interval(0.95, loc=np.mean(x),
                 scale=np.std(x)/np.sqrt(len(x)), df=len(x)-1)
```

```
62.588894889062544 70.61110511093744

(62.588894889062544, 70.61110511093744)
```

Real Estate Evaluation data

Real Estate Evaluation data

- We assumed that the data were Gaussian and found that $\hat{\mu} = 37.98$ and $\hat{\sigma} = 13.59$
- We had lots of data ($N = 414$) house prices

Real Estate Evaluation data

- We assumed that the data were Gaussian and found that $\hat{\mu} = 37.98$ and $\hat{\sigma} = 13.59$
- We had lots of data ($N = 414$) house prices ⇨ Student's t $\approx$ Gaussian
- More precisely, we assume

$$\bar{X} \sim N(\mu, \sigma^2/N)$$

Real Estate Evaluation data

- We assumed that the data were Gaussian and found that $\hat{\mu} = 37.98$ and $\hat{\sigma} = 13.59$
- We had lots of data ($N = 414$) house prices ⇨ Student's t ≈ Gaussian
- More precisely, we assume

$$\bar{X} \sim N(\mu, \sigma^2/N)$$

```python
import numpy as np
import scipy.stats as stats
N=414
Theta_hat=37.98
S_hat=13.59 #standard deviation
alpha = 0.05
z=1.96
#z = stats.norm.ppf(1-alpha/2)
CI_L = Theta_hat-z*S_hat/np.sqrt(N)
CI_U = Theta_hat+z*S_hat/np.sqrt(N)
print(CI_L, CI_U)

stats.norm.interval(0.95, loc=Theta_hat,
                    scale=S_hat/np.sqrt(N))
```

```
36.6708923596919 39.289107640308096

(36.67091641485199, 39.289083585148006)
```

Supermarket Data sales

- Here, we were assuming that the number of product sold followed a Poisson(λ).
- We estimated $\hat{\lambda} = 5.47$ items per customer as average rate of sale

Supermarket Data sales

- Here, we were assuming that the number of product sold followed a Poisson(λ).
- We estimated $\hat{\lambda} = 5.47$ items per customer as average rate of sale
- To compute the confidence interval, since we have data on $N = 340$ customer purchases we can use the CLT:

$$\bar{X} \sim N(\mu, \sigma^2/N)$$

Where σ^2 is estimated as $\widehat{V(X)} = \hat{\lambda}$ due to the properties of the Poisson distribution

Supermarket Data sales

- Here, we were assuming that the number of product sold followed a Poisson(λ).
- We estimated $\hat{\lambda} = 5.47$ items per customer as average rate of sale
- To compute the confidence interval, since we have data on $N = 340$ customer purchases we can use the CLT:

$$\bar{X} \sim N(\mu, \sigma^2/N)$$

Where σ^2 is estimated as $\widehat{V(X)} = \hat{\lambda}$ due to the properties of the Poisson distribution

```
import numpy as np
import scipy.stats as stats
N=340
Theta_hat=5.47
S_hat=math.sqrt(5.47)

stats.norm.interval(0.95, loc=Theta_hat,
                    scale=S_hat/np.sqrt(N))
```

```
(5.221399329486039, 5.718600670513961)
```

Transplant Survival Data

- Here, we had considered an exponential distribution and estimated $\hat{\lambda} = 0.0044$.
- Since $N = 45$, we can possibly use again the CLT, using the result that if $X \sim \text{Exp}(\lambda)$, then $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$.

Transplant Survival Data

- Here, we had considered an exponential distribution and estimated $\hat{\lambda} = 0.0044$.
- Since $N = 45$, we can possibly use again the CLT, using the result that if $X \sim \text{Exp}(\lambda)$, then $E(X) = 1/\lambda$ and $V(X) = 1/\lambda^2$.

```
import numpy as np
import scipy.stats as st
import statsmodels.datasets
import matplotlib.pyplot as plt
%matplotlib inline
data = statsmodels.datasets.heart.load_pandas().data
data = data[data.censors == 1]
survival = data.survival
N=len(survival) #45
smean = survival.mean()
rate = 1. / smean
svar = (1/rate)**2
S_hat=np.sqrt(svar)


stats.norm.interval(0.95, loc=smean,
                    scale=S_hat/np.sqrt(N))
```

```
(158.04964083175193, 288.5281369460258)
```

Probability & Statistics for DS & AI

Bootstrapped confidence intervals

Michele Guindani

Summer

What if we are not interested in the sample mean?

What if we are not interested in the sample mean?

- In many cases, the mean is not the characteristic we are interested in a population
- For example, in studying income, we are interested in the median income not the mean income
- Similarly, in studying survival, we are typically interested in the median survival, not the mean survival

What if we are not interested in the sample mean?

- In many cases, the mean is not the characteristic we are interested in a population
- For example, in studying income, we are interested in the median income not the mean income
- Similarly, in studying survival, we are typically interested in the median survival, not the mean survival

😱 😱 😱

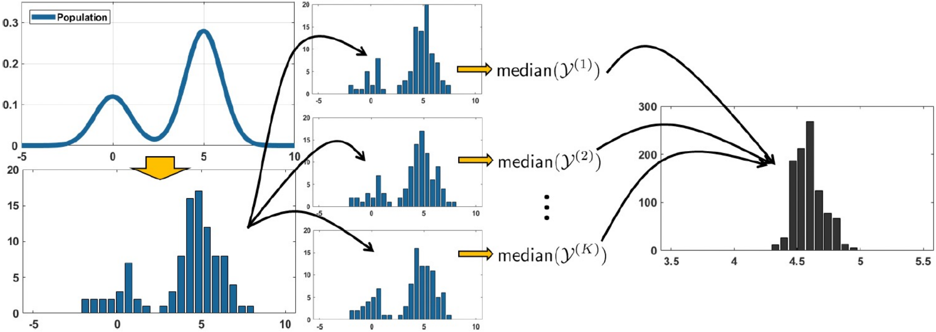
What if we are not interested in the sample mean?

- In many cases, the mean is not the characteristic we are interested in a population
- For example, in studying income, we are interested in the median income not the mean income
- Similarly, in studying survival, we are typically interested in the median survival, not the mean survival

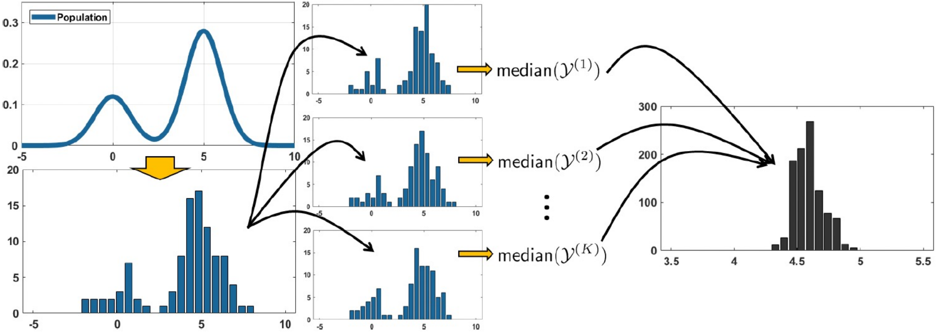
😱 😱 😱

⇨ We cannot use the CLT!!!!

😭 😭 😭

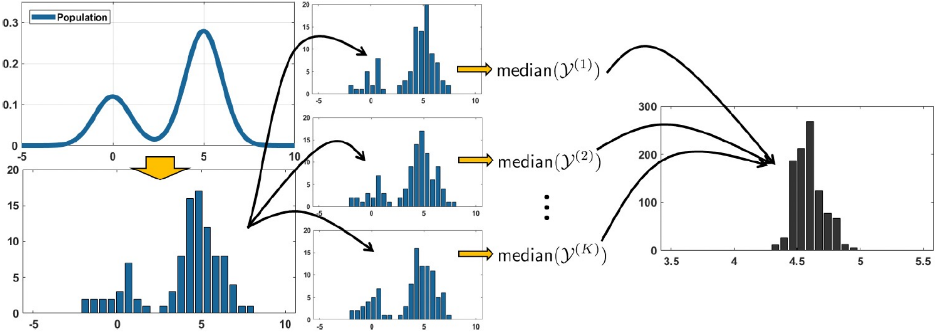


A conceptual illustration of bootstrapping. Given the observed dataset $\mathcal{X}$, we synthetically construct K bootstrapped datasets (colored in yellow) by sampling with replacement from $\mathcal{X}$. We then compute the estimators, e.g., computing the median, for every bootstrapped dataset. Finally, we construct the estimator's histogram (in blue) to compute the bootstrapped mean and variance.



A conceptual illustration of bootstrapping. Given the observed dataset $\mathcal{X}$, we synthetically construct K bootstrapped datasets (colored in yellow) by sampling with replacement from $\mathcal{X}$. We then compute the estimators, e.g., computing the median, for every bootstrapped dataset. Finally, we construct the estimator's histogram (in blue) to compute the bootstrapped mean and variance.

Bootstrapped standard errror: $\widehat{\text{se}}_{\text{boot}} = \sqrt{\mathbb{V}_{\text{boot}}(\widehat{\Theta})}.$



A conceptual illustration of bootstrapping. Given the observed dataset $\mathcal{X}$, we synthetically construct K bootstrapped datasets (colored in yellow) by sampling with replacement from $\mathcal{X}$. We then compute the estimators, e.g., computing the median, for every bootstrapped dataset. Finally, we construct the estimator's histogram (in blue) to compute the bootstrapped mean and variance.

Bootstrapped Confidence Interval: $\mathcal{I} = \left[\widehat{\Theta} - z_\alpha \widehat{\mathsf{se}}_{\mathrm{boot}},\ \ \widehat{\Theta} + z_\alpha \widehat{\mathsf{se}}_{\mathrm{boot}}\right]$.

```python
# Python code to estimate a bootstrapped variance
import numpy as np
np.random.seed(1234)
X = np.array([72, 69, 75, 58, 67, 70, 60, 71, 59, 65])
N = X.size
K = 1000
Thetahat = np.zeros(K)
for i in range(K):
    idx = np.random.randint(N, size=N)
    Y = X[idx]
    Thetahat[i] = np.median(Y)
M = np.mean(Thetahat)
V = np.var(Thetahat)
print(M)
print(V)
stats.norm.interval(0.95, loc=M,
                    scale=np.sqrt(V/N))
```

```
67.3875
7.666093750000002

(65.67142938792702, 69.10357061207299)
```