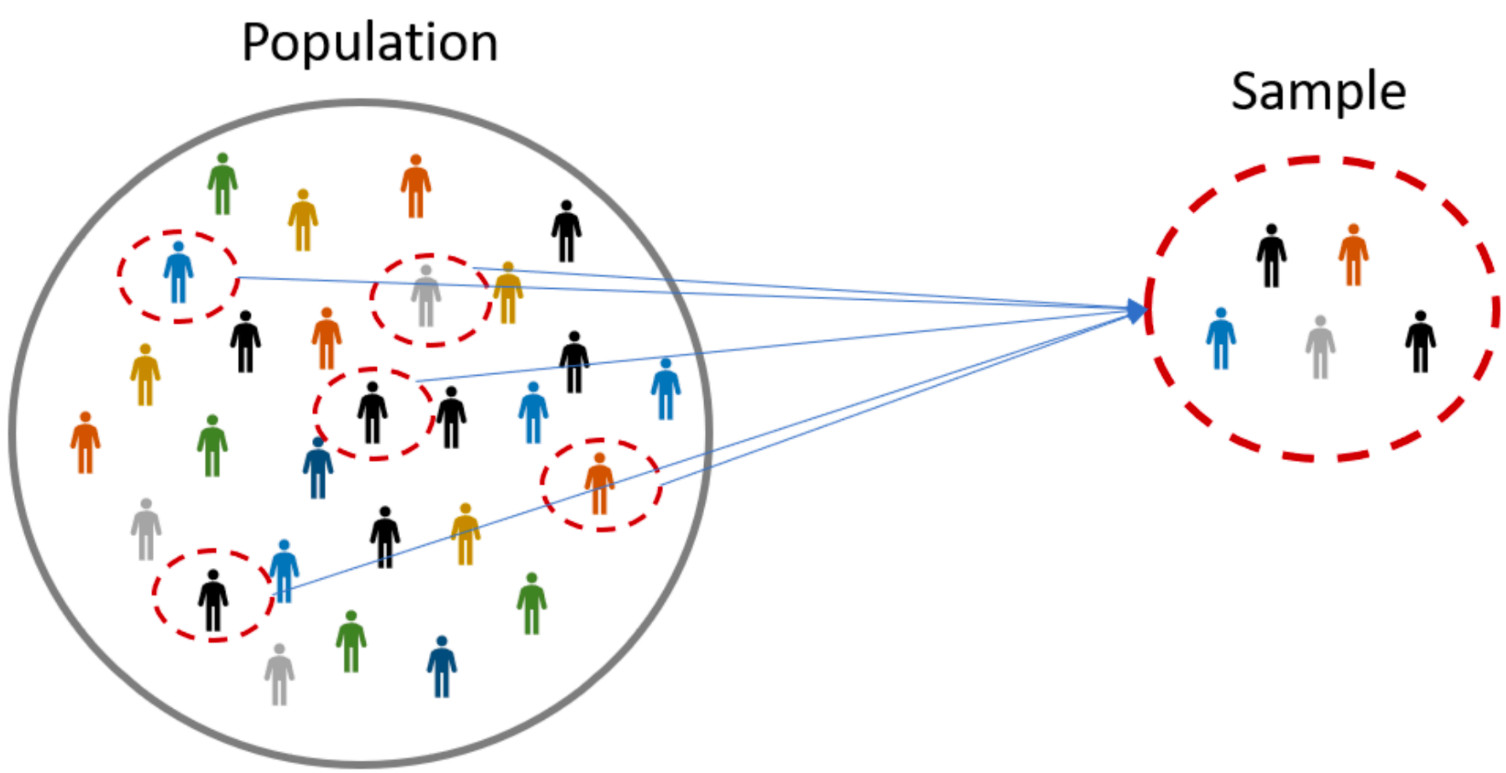
Confidence Intervals and Hypothesis Testing

Contents

Review of Sampling Distributions

- Populations and samples



Mean value

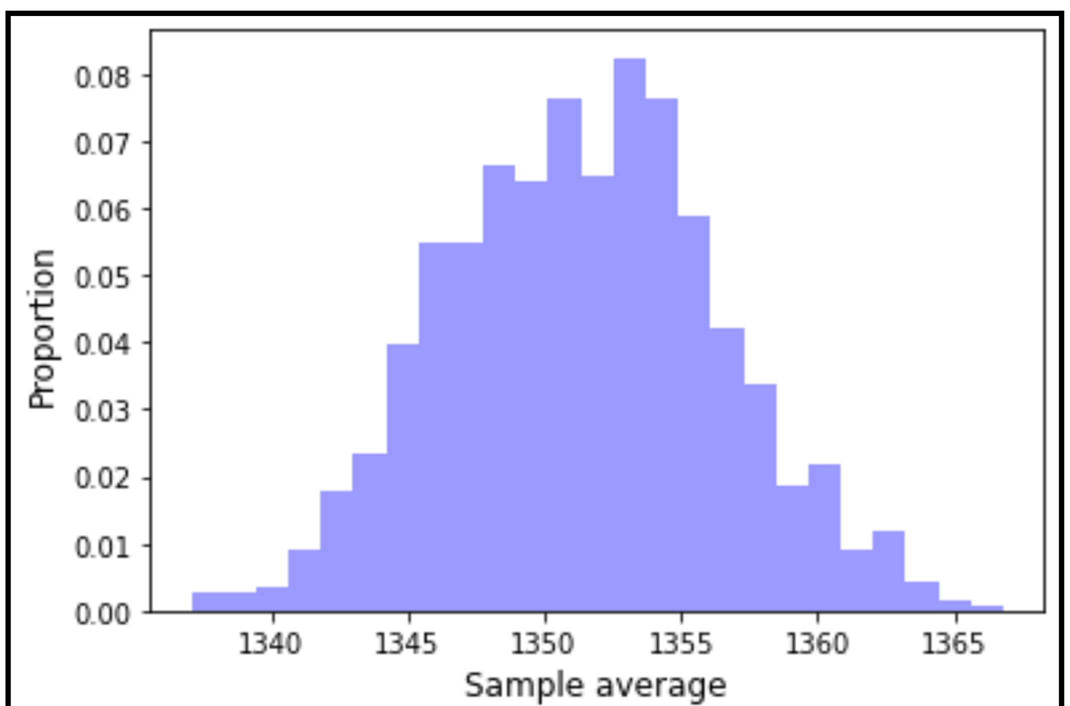
$$\mu = \begin{cases} \sum_{i=1}^{k} x_i p_i \\ \int_{x \in \mathcal{X}} x f(x) dx \end{cases}$$

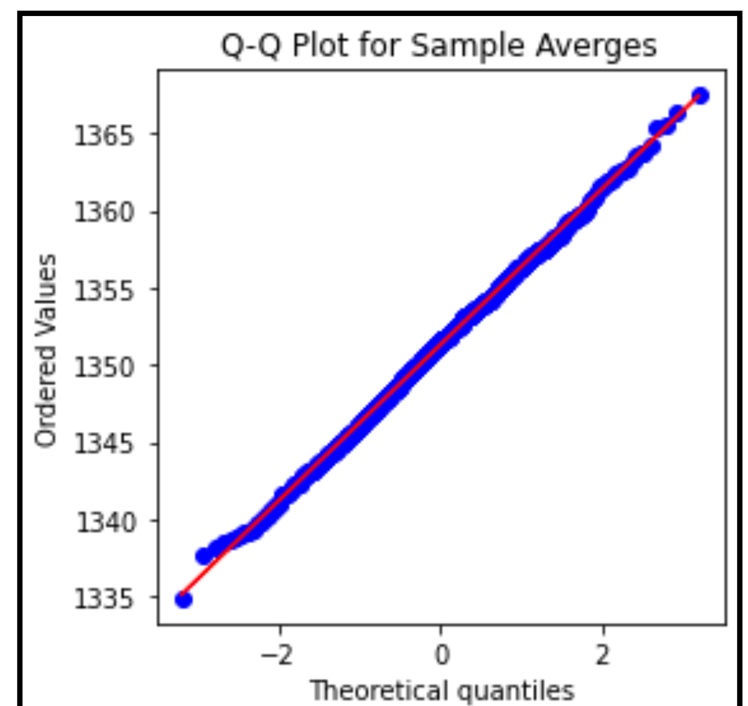
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Review of Sampling Distributions

- Central limit theorem

Notes: Central Limit Theorem (CLT): For a relatively large sample size, the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normally distributed, regardless of the distribution of the population. The approximation becomes better with increased sample size.





Review of Sampling Distributions

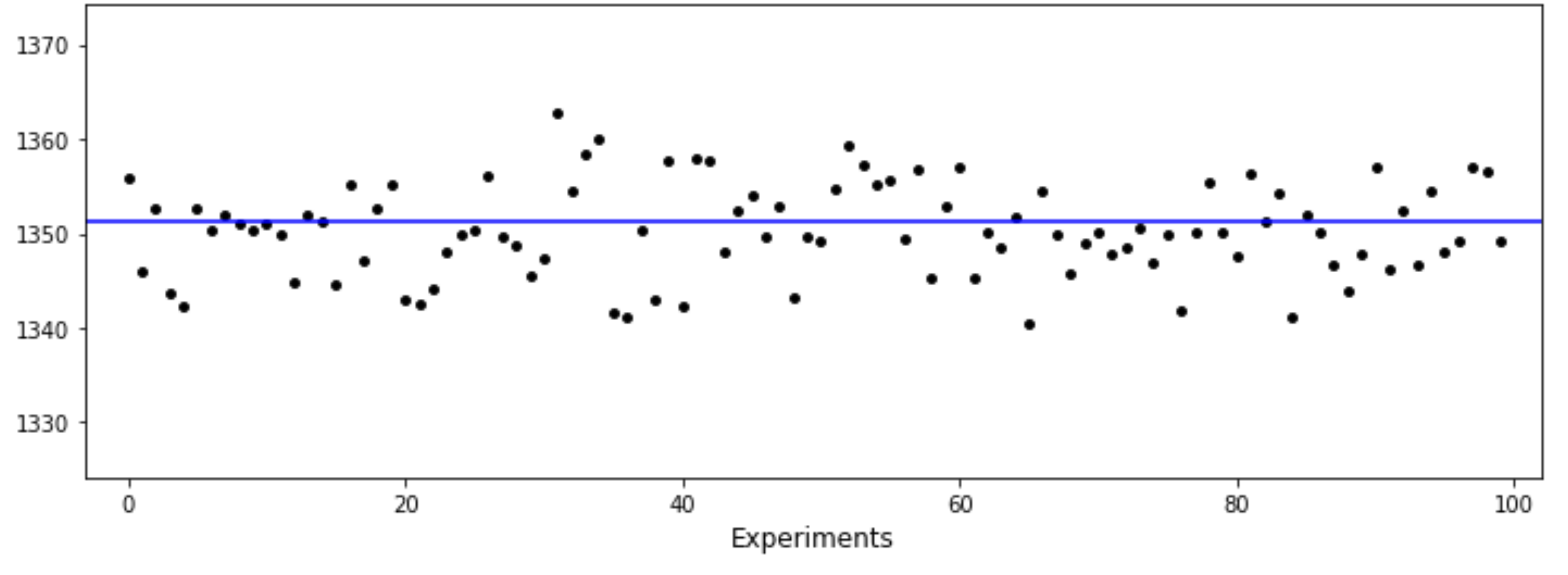
- Expected values and variance of the sample average

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

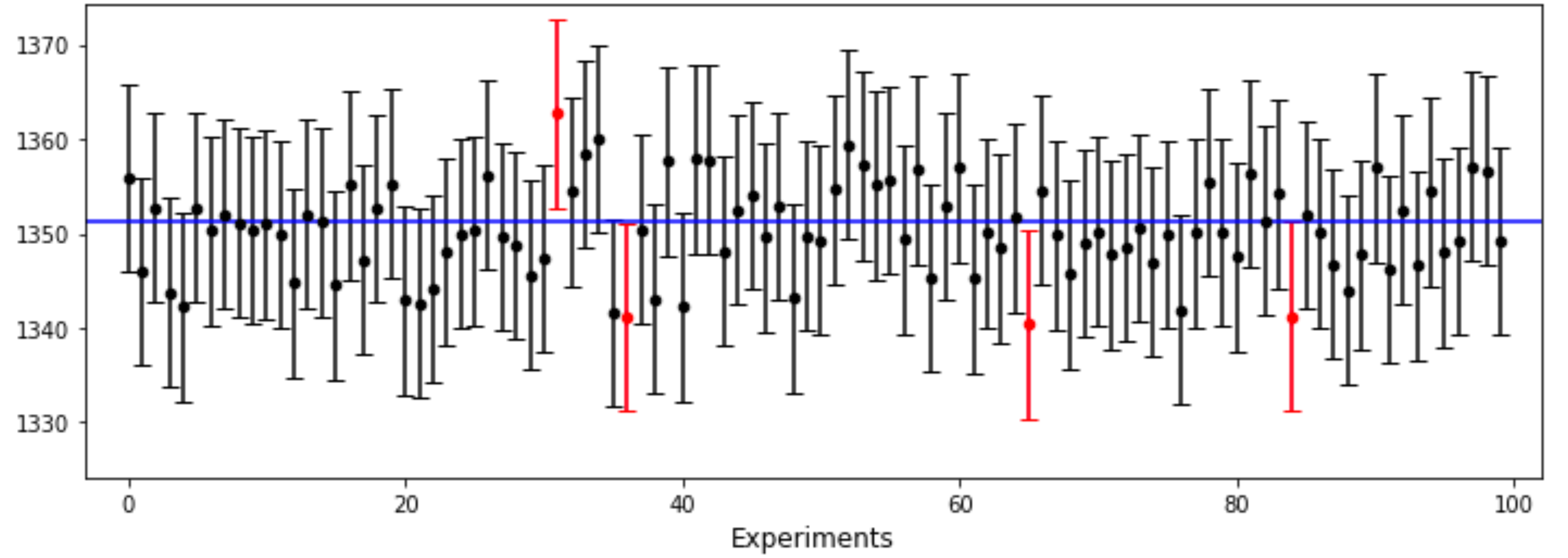
Confidence Intervals

- Confidence intervals for population means
  - General idea
    - ✓ Point estimates of the population mean



Confidence Intervals

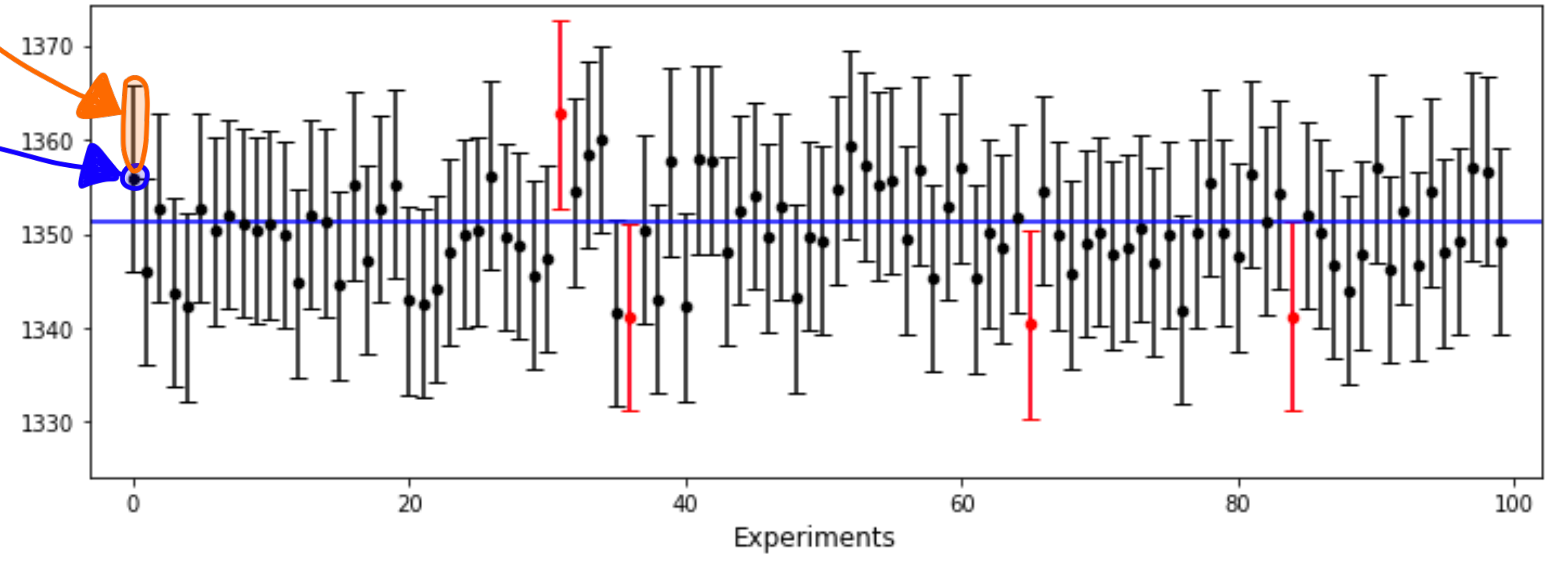
- Confidence intervals for population means
  - General idea
    - ✓ Point estimates of the population mean
    - ✓ A range of plausible values



Confidence Intervals

- Confidence intervals for population means
  - Equations

estimate ± margin of error



Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known
    - ✓ $\bar{X}$ is approximately normally distributed
    - ✓ The mean value of $\bar{X}$ is the population mean μ
    - ✓ The standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$

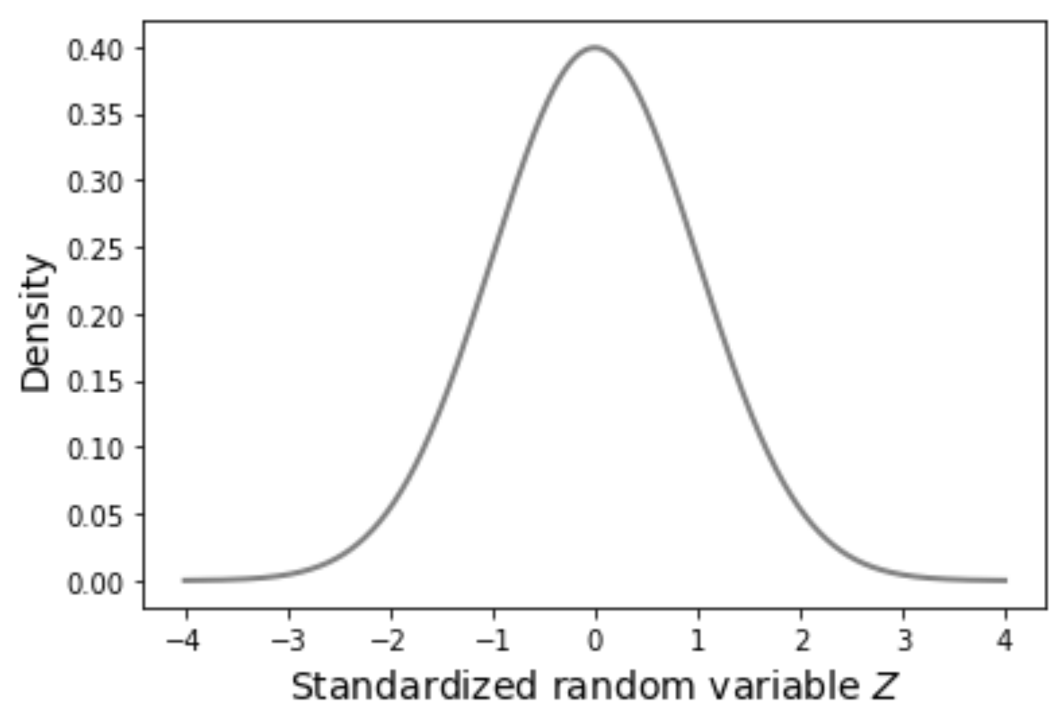
Standard normal distribution

$\bar{X} \sim N(\mu, \sigma^2/n)$ ⇨ z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Confidence Intervals

- Confidence intervals for population means
    - Population standard deviation σ is known

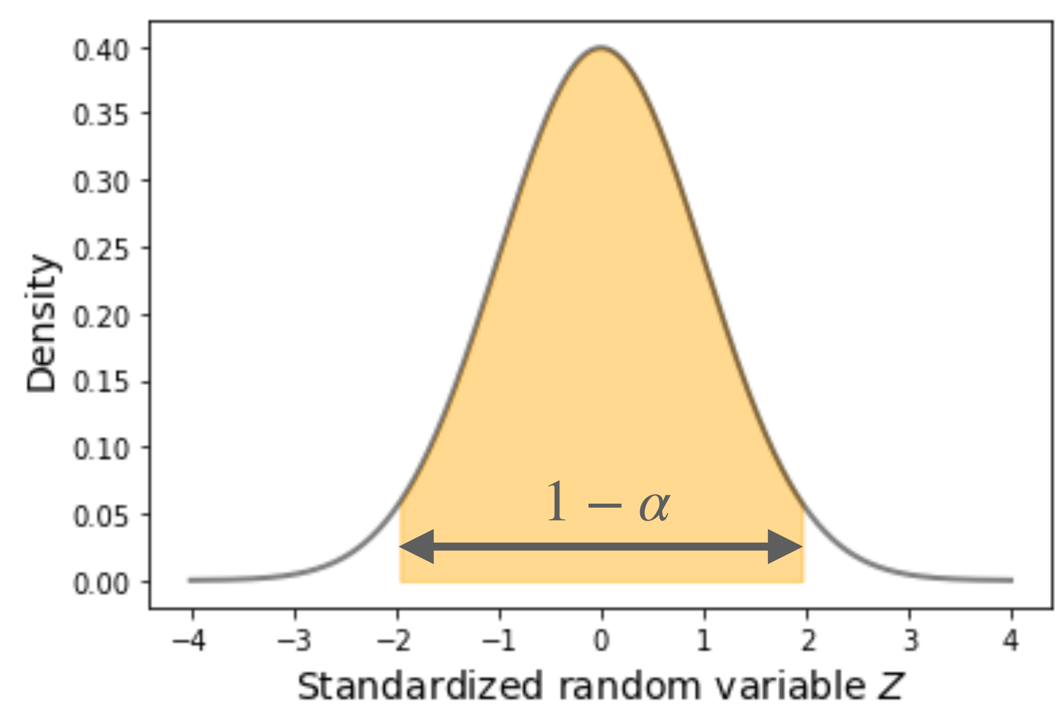
z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$



Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

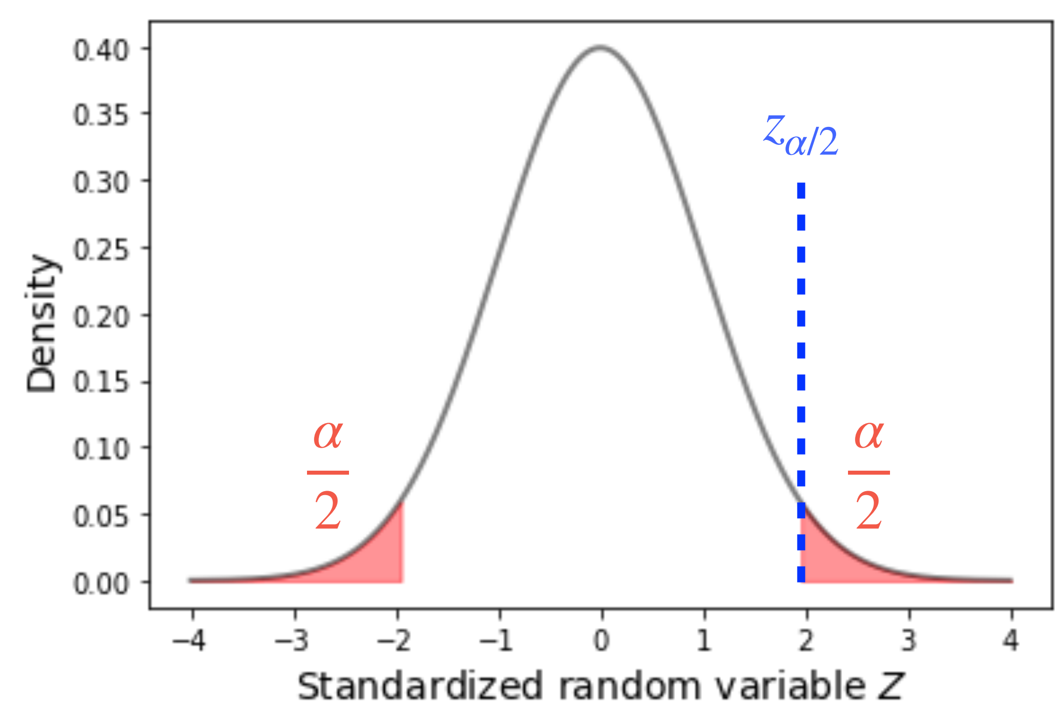


Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

$$\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha/2}$$

Confidence Intervals

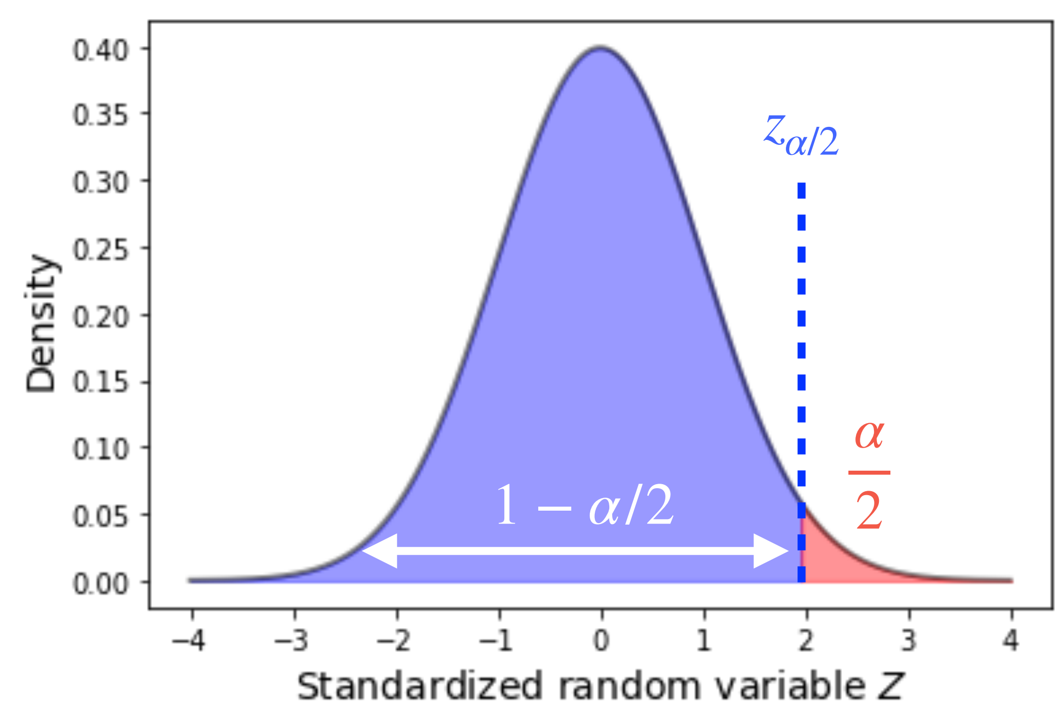
- Confidence intervals for population means
  - Population standard deviation σ is known

z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$P(Z \leq z_{\alpha/2}) = F(z_{\alpha/2}) = 1 - \alpha/2$

$$z_{\alpha/2} = F^{-1}(1 - \alpha/2)$$

Percent point function



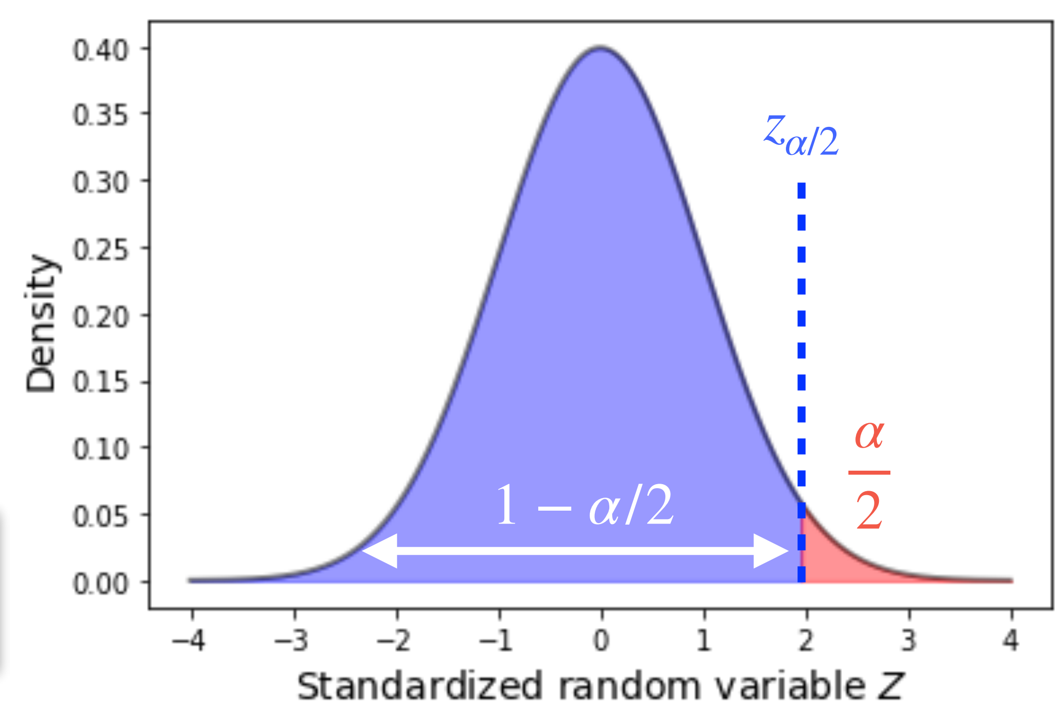
Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

estimate $\pm$ margin of error

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

$$z_{\alpha/2} = F^{-1}(1 - \alpha/2)$$




```
from scipy.stats import norm

z_alpha2 = norm.ppf(1-alpha/2)
```

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

Example 1: Consider the dataset "bulb.csv" as the population, a sample with $n = 25$ records is randomly selected to infer the population mean. Assuming that the population standard deviation σ is known, calculate the confidence interval with the confidence level to be $1 - \alpha = 95\,\%$.

```python
data = pd.read_csv('bulb.csv')
population = data['Lifespan']
sigma = population.values.std()
print(f'The population standard deviation: {sigma}')
```

```
The population standard deviation: 25.437524255752564
```

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

```
n = 25
sample = population.sample(n=25, replace=True)
```

```
estimate = sample.mean()

alpha = 0.05
z_alpha2 = norm.ppf(1-alpha/2)
moe = z_alpha2 * sigma/n**0.5

lower = estimate - moe
upper = estimate + moe

print(f'CI: [{lower}, {upper}]')
```

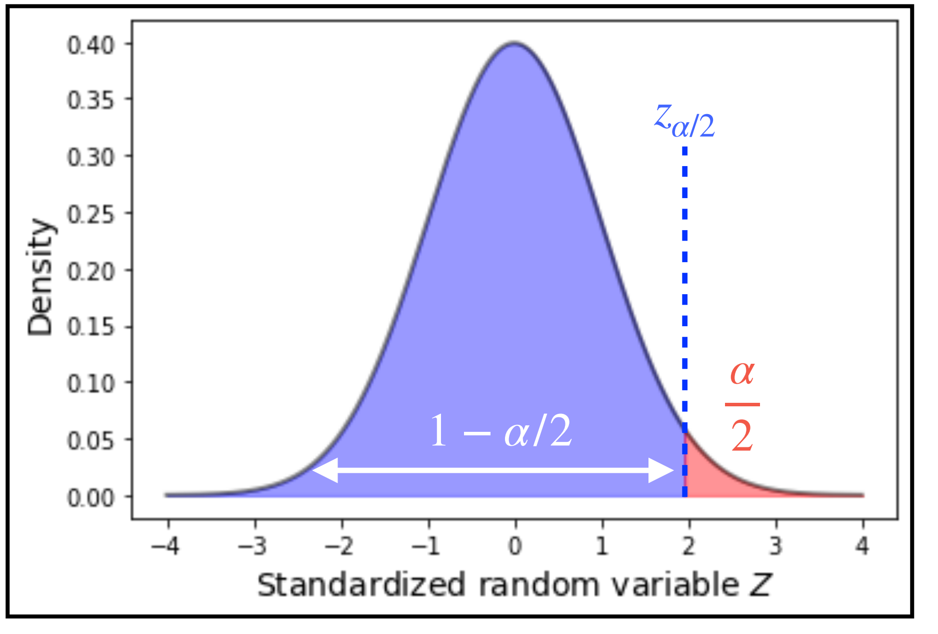
```
CI: [1340.298523893984, 1360.2411764528397]
```

Confidence Intervals

- Confidence intervals for population means
    - Population standard deviation σ is known

```
n = 25
sample = population.sample(n=25, replace=True)
```

```
estimate = sample.mean()

alpha = 0.05
z_alpha2 = norm.ppf(1-alpha/2)
moe = z_alpha2 * sigma/n**0.5

lower = estimate - moe
upper = estimate + moe

print(f'CI: [{lower}, {upper}]')
```




```
CI: [1340.298523893984, 1360.2411764528397]
```

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

```
n = 25
sample = population.sample(n=25, replace=True)
```

```
estimate = sample.mean()

alpha = 0.05
z_alpha2 = norm.ppf(1-alpha/2)
moe = z_alpha2 * sigma/n**0.5

lower = estimate - moe
upper = estimate + moe

print(f'CI: [{lower}, {upper}]')
```

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

```
CI: [1340.298523893984, 1360.2411764528397]
```

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

```
lowers = []
uppers = []
repeats = 1000

alpha=0.05
z_alpha2 = norm.ppf(1-alpha/2)

for i in range(repeats):
    sample = population.sample(n=25, replace=True)
    estimate = sample.mean()
    moe = z_alpha2 * sigma/n**0.5

    lowers.append(estimate - moe)
    uppers.append(estimate + moe)

conf_int = pd.DataFrame({'lower': lowers,'upper': uppers})
```

Calculate the estimate and margin of error

Append the lower and upper bounds

Peng Xiong, DAO, NUS Business School

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

```
lowers = []
uppers = []
repeats = 1000

alpha=0.05
z_alpha2 = norm.ppf(1-alpha/2)

for i in range(repeats):
    sample = population.sample(n=25, replace=True)
    estimate = sample.mean()
    moe = z_alpha2 * sigma/n**0.5

    lowers.append(estimate - moe)
    uppers.append(estimate + moe)

conf_int = pd.DataFrame({'lower': lowers,'upper': uppers})
```

	lower	upper
0	1339.757322	1359.699974
1	1350.411770	1370.354423
2	1336.795506	1356.738158
3	1340.608315	1360.550968
4	1341.931845	1361.874497
...	...	...
995	1340.642709	1360.585362
996	1344.704269	1364.646921
997	1342.893465	1362.836118
998	1350.264806	1370.207458
999	1329.216515	1349.159168

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is known

```
cond1 = mean_pop >= conf_int['lower']
cond2 = mean_pop <= conf_int['upper']
prob = (cond1 & cond2).mean()

print(f'The probability is {prob}')
```

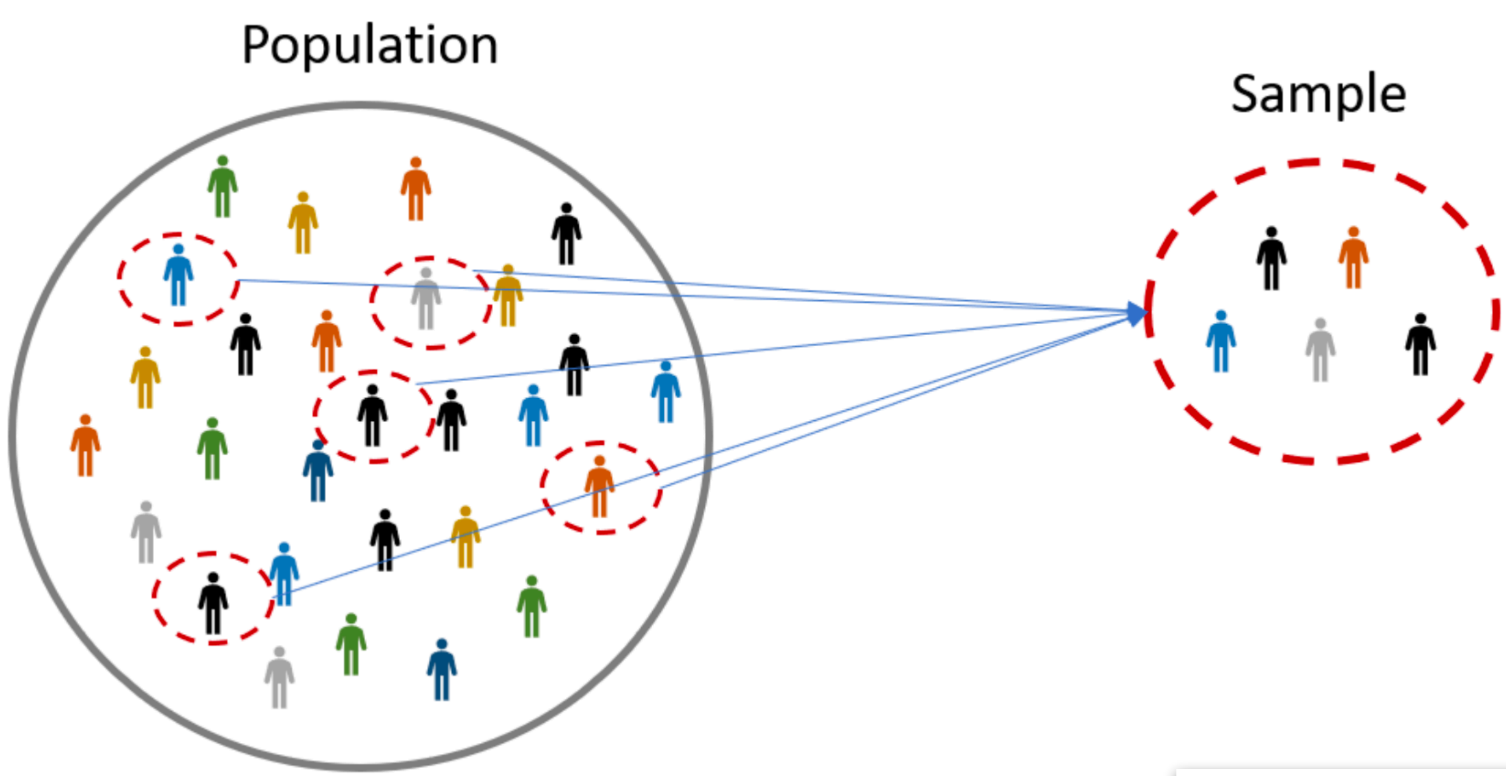
```
The probability is 0.951
```

The probability is approximately the confidence level $1-\alpha$

  - Programming for Business Analytics

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is unknown



Population variance σ^2

Sample variance $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Confidence Intervals

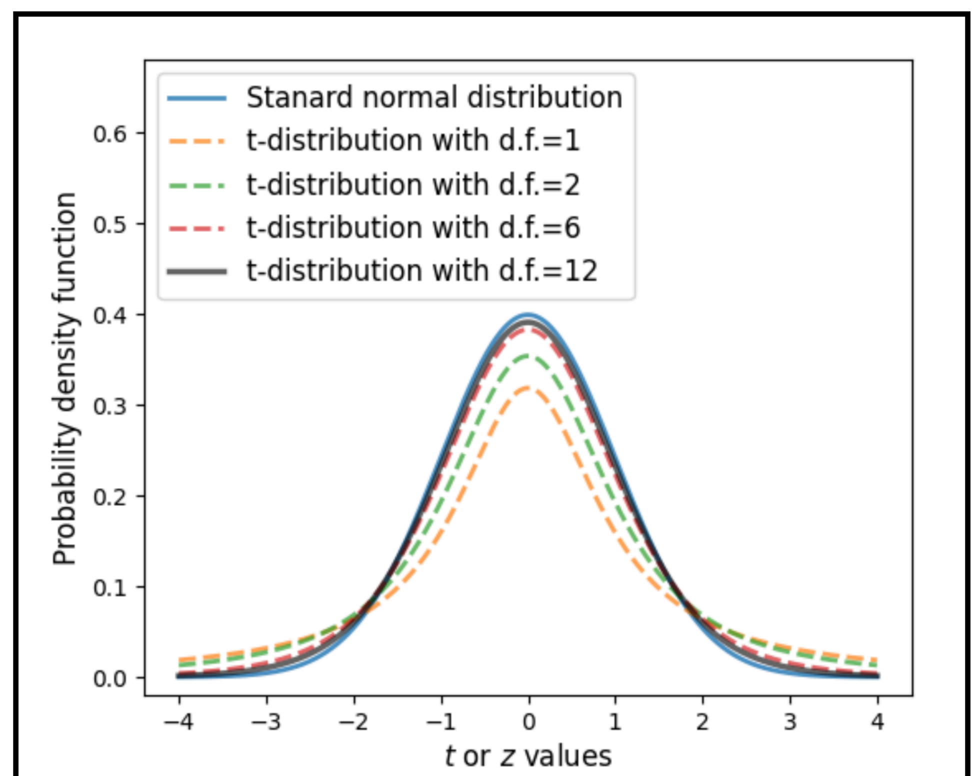
- Confidence intervals for population means
  - Population standard deviation σ is unknown

z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ $\Rightarrow$ t-value: $T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim t$-distribution

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is unknown

z-value: $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ $\Rightarrow$ t-value: $T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}} \sim$ t-distribution

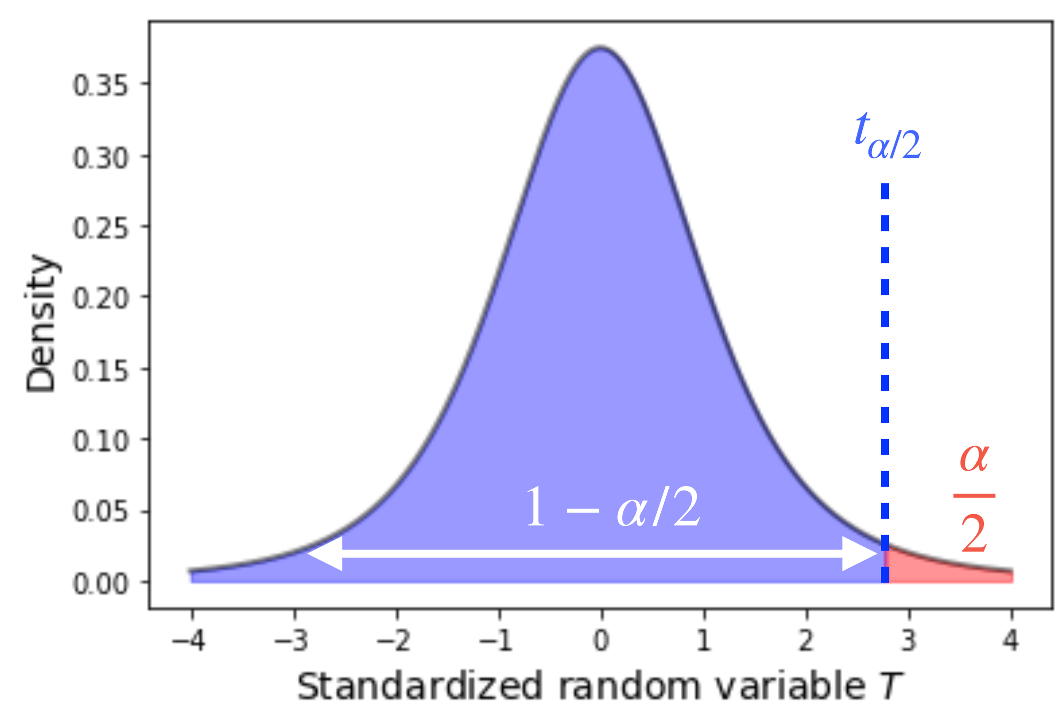


Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is unknown

estimate ± margin of error

$$\bar{X} \pm \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}$$




```
from scipy.stats import t

t_alpha2 = t.ppf(1-alpha/2, n-1)
```

Degree of freedom

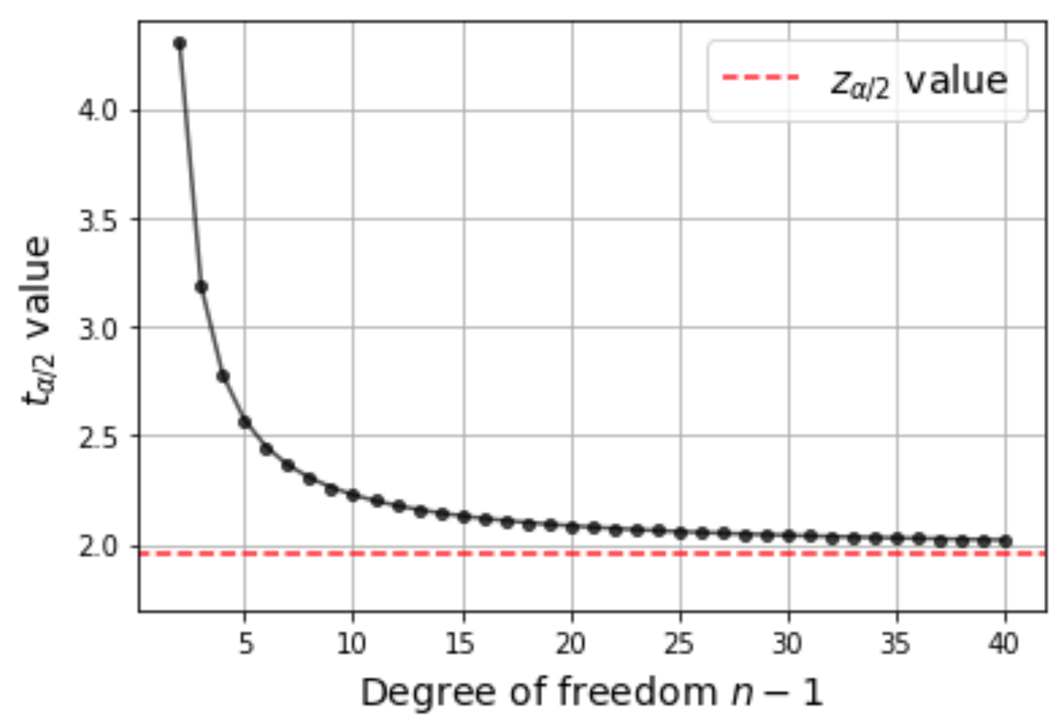
Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is unknown

estimate $\pm$ margin of error

$$\bar{X} \pm \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}$$

$t_{\alpha/2} \approx z_{\alpha/2}$, for large n




```
from scipy.stats import t

t_alpha2 = t.ppf(1-alpha/2, n-1)
```

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is unknown

Example 2: Consider the dataset "bulb.csv" as the population, a sample with $n = 25$ records is randomly selected to infer the population mean. Now the population standard deviation σ is unknown, calculate the confidence interval with the confidence level to be $1 - \alpha = 95\,\%$.

```
n = 25
sample = population.sample(n=25, replace=True)
```

Confidence Intervals

- Confidence intervals for population means
  - Population standard deviation σ is unknown

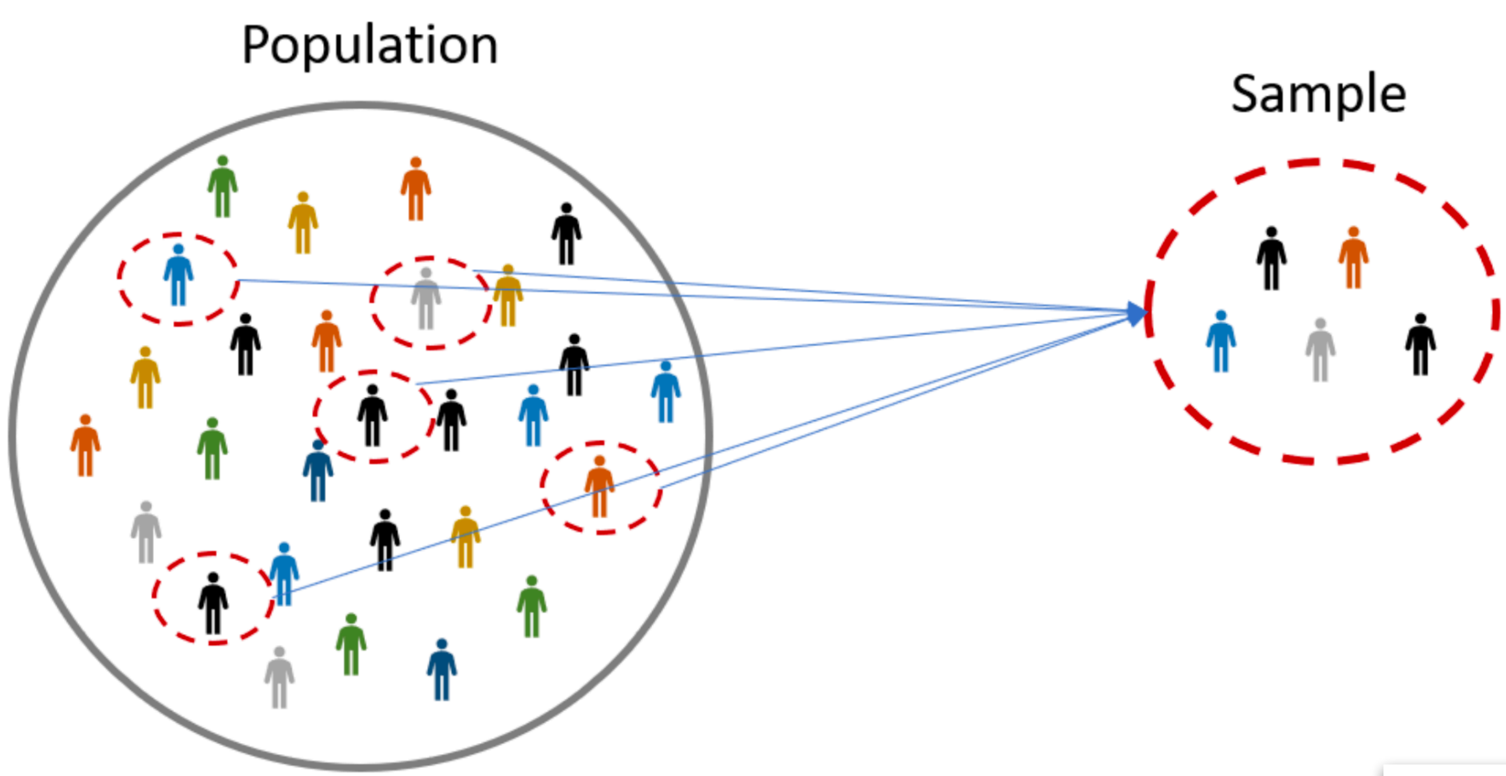
```python
estimate = sample.mean()

alpha = 0.05
t_alpha2 = t.ppf(1-alpha/2, n-1)
s = sample.std()
moe = t_alpha2 * s/n**0.5

lower = estimate - moe
upper = estimate + moe

print(f'CI: [{lower}, {upper}]')
```

$$\bar{X} \pm \frac{s}{\sqrt{n}} \cdot t_{\alpha/2}$$

```
CI: [1345.4908264720336, 1366.3649342996725]
```

Confidence Intervals

- Confidence intervals for population proportions



Population proportion p

Sample proportion $\hat{p} = \dfrac{m}{n}$

Confidence Intervals

- Confidence intervals for population proportions
  - Sampling distributions of sample proportions

Example 3: It is assumed that in Singapore, $p = 73\,\%$ of customers prefer Coke over Pepsi. We are conducting n surveys to investigate customers' preference, and among these surveys, m people choose Coke. Plot the sampling distributions of the **sample proportion** $\hat{p} = m/n$ under different sample sizes $n = 5$, 10, 50, and 100.

Sample proportion $\hat{p} = \frac{m}{n}$ $\longrightarrow$ $m \sim B(n, p)$

Confidence Intervals

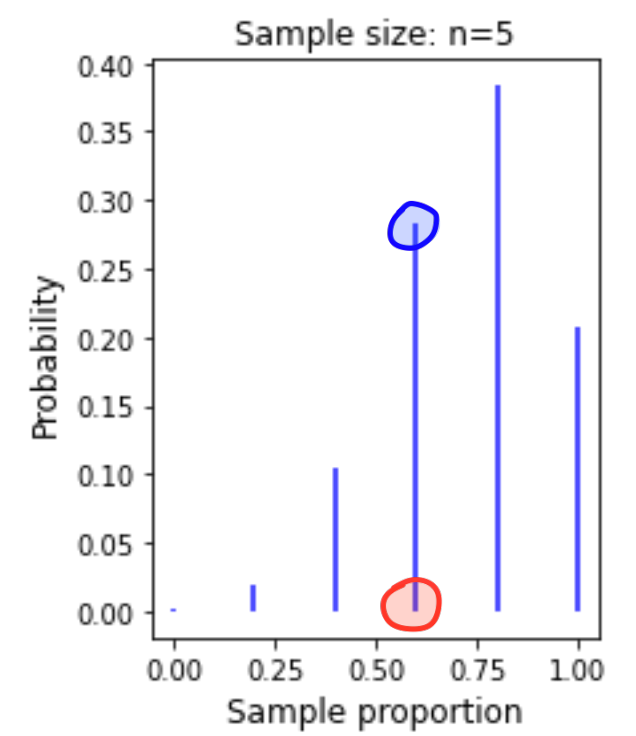
- Confidence intervals for population proportions
    - Sampling distributions of sample proportions

```
n = 5
p = 0.73

m = np.arange(n+1)
pmf = binom.pmf(m, n, p)
estimate = m/n

plt.figure(figsize=(3, 4))
plt.vlines(estimate, ymin=0, ymax=pmf,
           linewidth=2, colors='b', alpha=0.7)
plt.xlabel('Sample proportion', fontsize=12)
plt.ylabel('Probability', fontsize=12)
plt.title(f'Sample size: n={n}')
plt.show()
```

Confidence Intervals

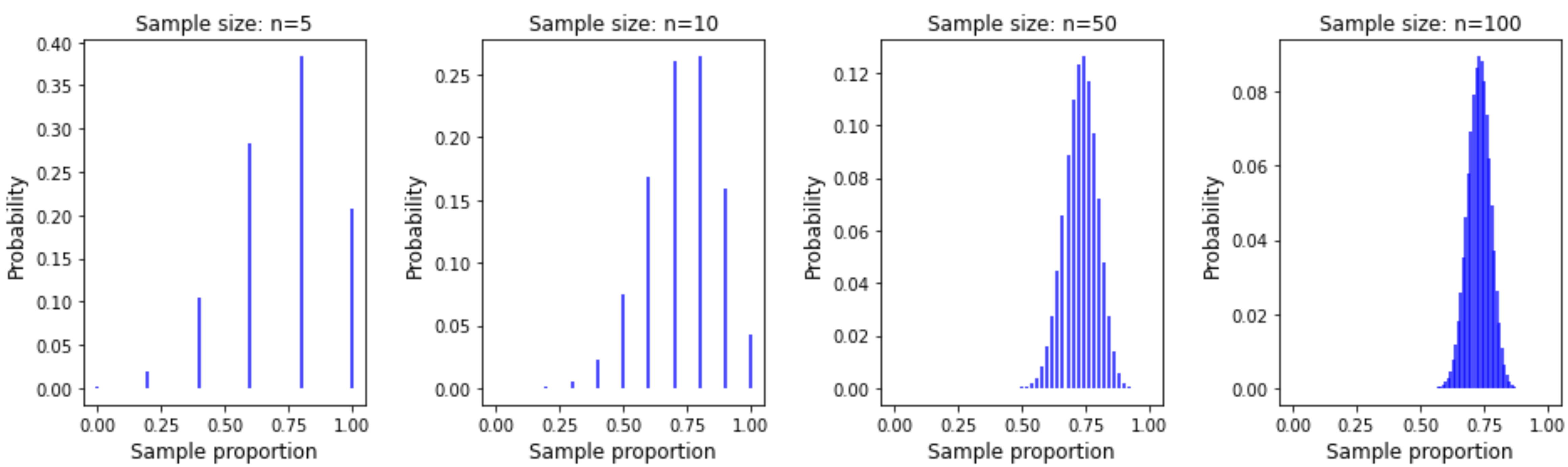
- Confidence intervals for population proportions
  - Sampling distributions of sample proportions

```
n = 5
p = 0.73

m = np.arange(n+1)
pmf = binom.pmf(m, n, p)
estimate = m/n

plt.figure(figsize=(3, 4))
plt.vlines(estimate, ymin=0, ymax=pmf,
           linewidth=2, colors='b', alpha=0.7)
plt.xlabel('Sample proportion', fontsize=12)
plt.ylabel('Probability', fontsize=12)
plt.title(f'Sample size: n={n}')
plt.show()
```



Confidence Intervals

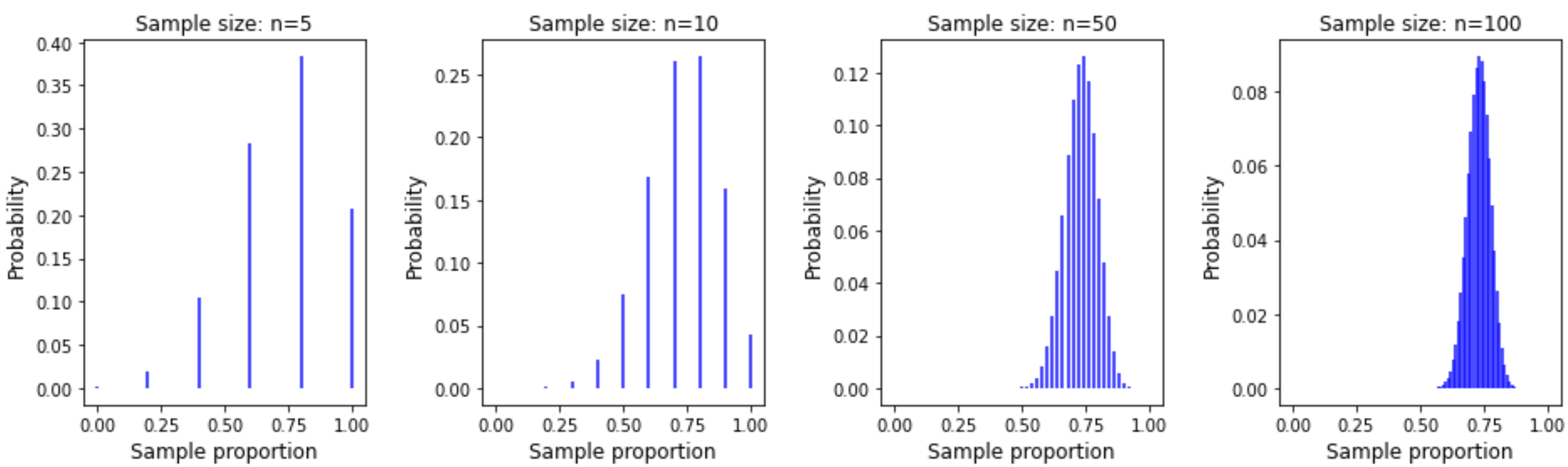
- Confidence intervals for population proportions
    - Sampling distributions of sample proportions



Confidence Intervals

- Confidence intervals for population proportions
  - Sampling distributions of sample proportions
    - ✓ The sample proportion is centered at the population proportion p

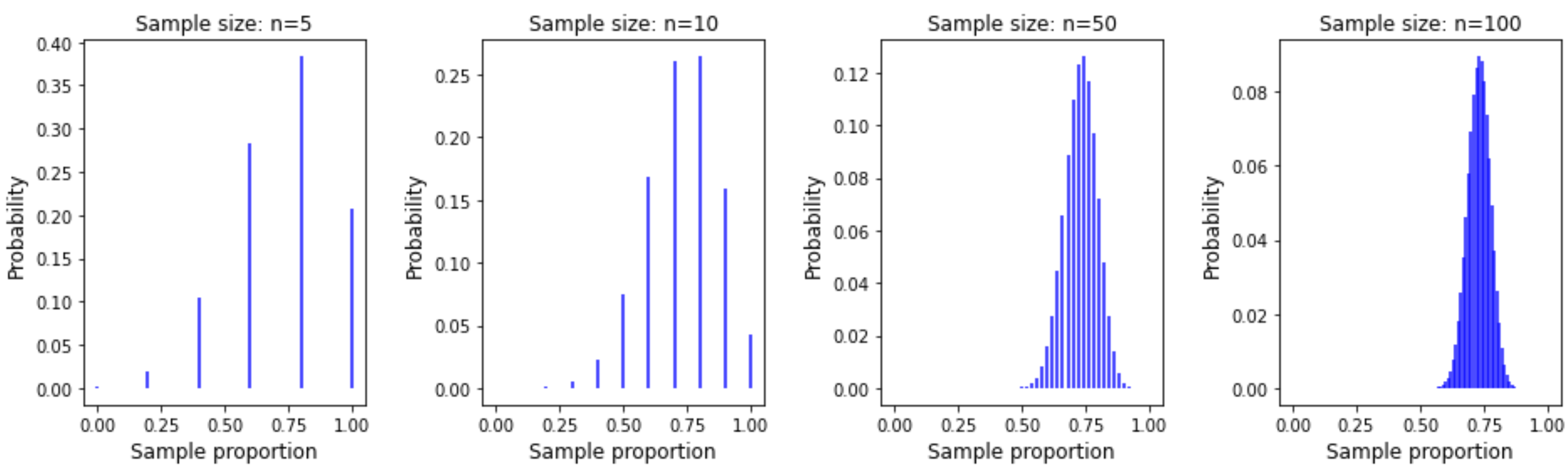
$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{m}{n}\right) = \frac{\mathbb{E}(m)}{n} = \frac{np}{n} = p$$



Confidence Intervals

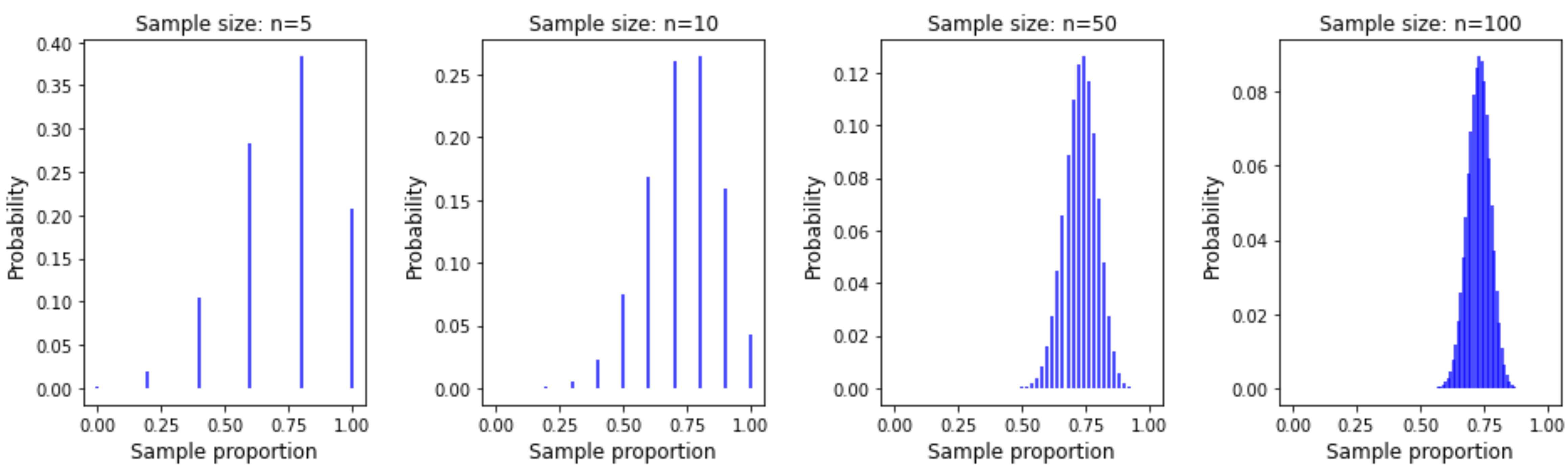
- Confidence intervals for population proportions
  - Sampling distributions of sample proportions
    - ✓ The variance of the sample proportion is decreased as n increases

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{m}{n}\right) = \frac{\text{Var}(m)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$



Confidence Intervals

- Confidence intervals for population proportions
  - Sampling distributions of sample proportions
    - ✓ The shape of the sampling distribution approaches a normal distribution as n increases



Confidence Intervals

- Confidence intervals for population proportions
  - Equation for the confidence interval

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

$$\Rightarrow$$

z-value: $Z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$

$$\Downarrow$$

z-value: $Z = \dfrac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$

Confidence Intervals

- Confidence intervals for population proportions
  - Equation for the confidence interval

z-value: $Z = \dfrac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$$

$$\Downarrow$$

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

$$\Rightarrow$$

estimate ± margin of error

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

Confidence Intervals

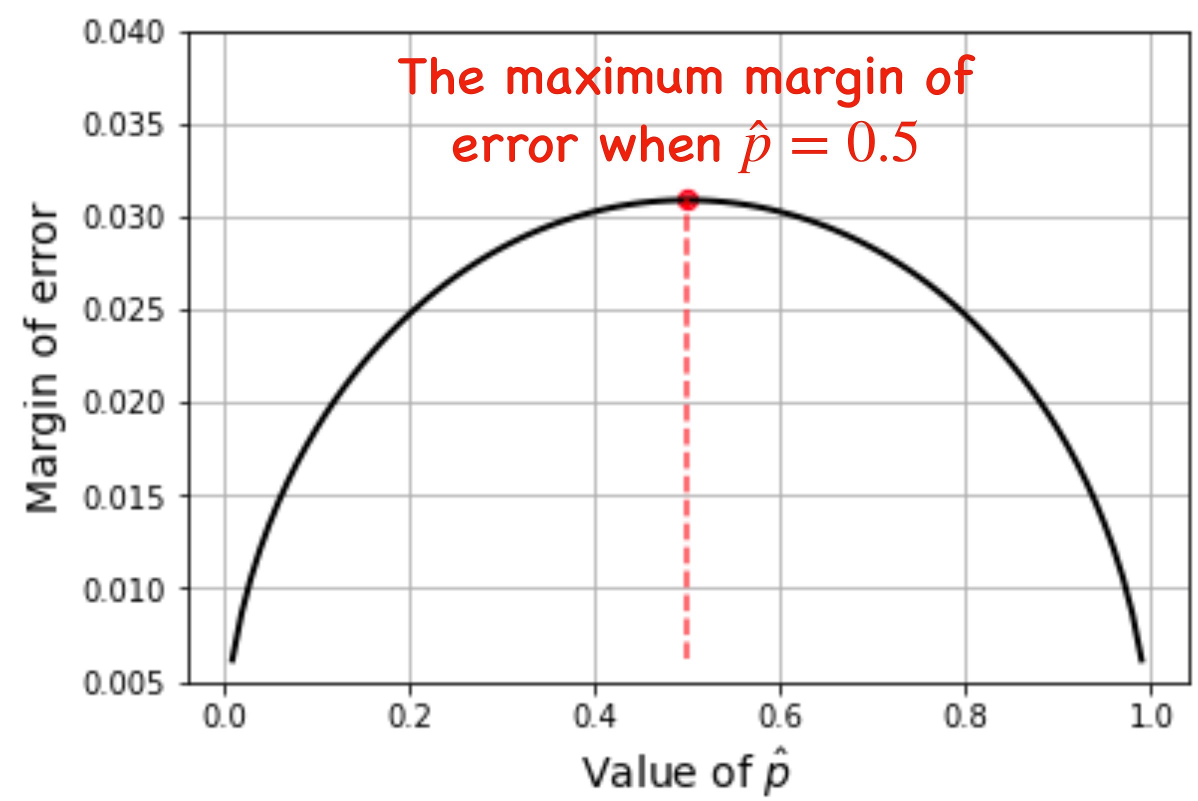
- Confidence intervals for population proportions

Example 4: Political polling is usually used to predict the results of an election. Typically, a poll of $n = 1004$ people can be used to represent hundreds of million of voters across the country. Why such a small sample size is considered sufficient? How can we interpret the results?

margin of error: $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \cdot z_{\alpha/2}$

$\Downarrow$

margin of error $\leq \dfrac{0.5}{\sqrt{n}} \cdot z_{\alpha/2}$



Confidence Intervals

- Confidence intervals for population proportions

```
n = 1004
p_hat = 0.5

alpha = 0.05
z_alpha2 = norm.ppf(1-alpha/2)
moe = z_alpha2 * (p_hat*(1-p_hat)/n)**0.5

print(f'The margin of error: {moe}')
```

The value of $\hat{p}$ that maximizes the margin of error

```
The margin of error: 0.030927957432873788
```

$$\text{margin of error } \leq \frac{0.5}{\sqrt{n}} \cdot z_{\alpha/2} \approx 0.03$$

We have $1-\alpha = 95\,\%$ confidence that the true population proportion is within a $\pm 3\,\%$ interval around the sample proportion $\hat{p}$

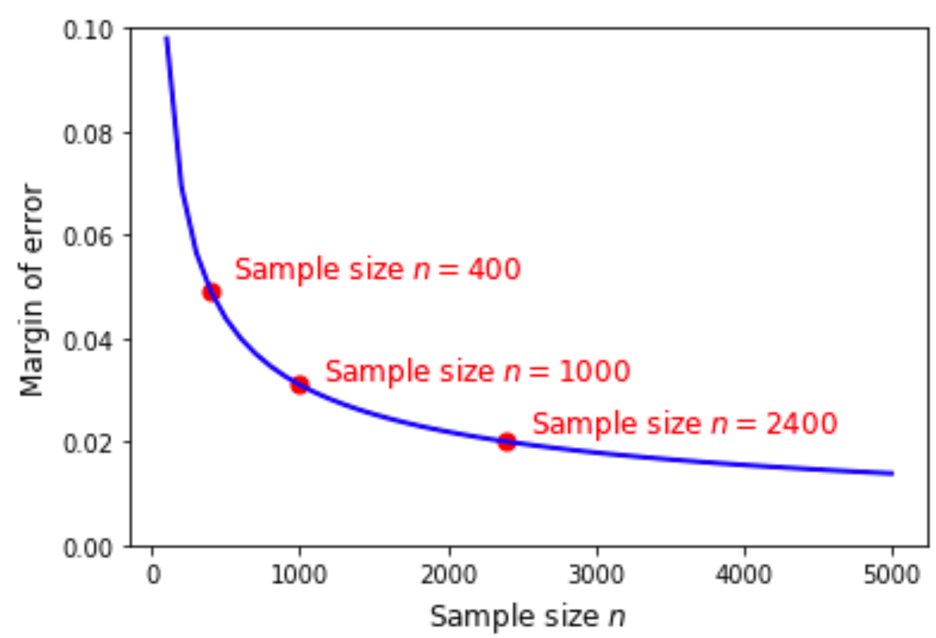
Confidence Intervals

- Confidence intervals for population proportions

```
n = 1004
p_hat = 0.5

alpha = 0.05
z_alpha2 = norm.ppf(1-alpha/2)
moe = z_alpha2 * (p_hat*(1-p_hat)/n)**0.5

print(f'The margin of error: {moe}')
```

```
The margin of error: 0.03092795743287378
```

$$\text{margin of error} \leq \frac{0.5}{\sqrt{n}} \cdot z_{\alpha/2}$$



Confidence Intervals

- Summary
  - Equation

estimate ± margin of error

Parameter	Estimate	Margin of error	Remarks
Mean value μ	Sample average $\bar{X}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ if σ is known $t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ if σ is unknown	$t_{\alpha/2}$ can be replaced by $z_{\alpha/2}$ for very large n.
Proportion p	Sample proportion $\hat{p}$	$z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$	-

  - Programming for Business Analytics

Hypothesis Testing

- Introduction to hypothesis testing

Notes:

- The **null hypothesis** is usually the current thinking, or status quo, denoted by H_0.
- The **alternative (research) hypothesis**, denoted by H_a, is a hypothesis considered to be the alternative to the null hypothesis. It is usually the hypothesis we want to prove, the values of the parameter we prefer, or consider plausible.
- **Hypothesis test**: the problem to decide whether the null hypothesis should be rejected in favor of the alternative hypothesis.

Notes: Basic Logic of Hypothesis Testing: Take a random sample from the population. If the sample data are consistent with the null hypothesis, do not reject the null hypothesis; if the sample data are inconsistent with the null hypothesis and supportive of the alternative hypothesis, reject the null hypothesis in favor of the alternative hypothesis.

Hypothesis Testing

- Steps of hypothesis testing
  - Hypotheses
    - ✓ Types of tests
      - Two-tailed test: $H_a : \mu \neq \mu_0$
      - Left-tailed test: $H_a : \mu < \mu_0$
      - Right-tailed test: $H_a : \mu > \mu_0$

Hypothesis Testing

- Steps of hypothesis testing
  - Sampling distributions
    - ✓ The population mean
      - Assume the null hypothesis is true $(\mu = \mu_0)$
      - Standardization of the sample mean

z-value: $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

The population standard deviation σ is known

t-value: $T = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t$-distribution

The population standard deviation σ is unknown

Hypothesis Testing

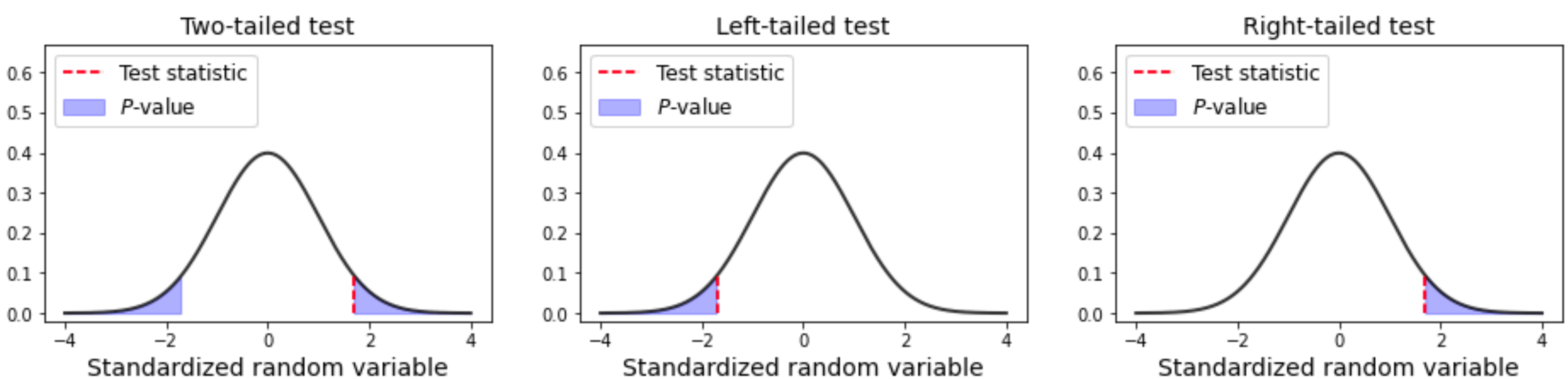
- Steps of hypothesis testing
  - Sampling distributions
    - ✓ The population proportion
      - Assume the null hypothesis is true $(p = p_0)$
      - Standardization of the sample proportion

z-value: $Z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0,1)$

Hypothesis Testing

- Steps of hypothesis testing
  - Calculation of the P-value

Notes: The P-value of a hypothesis test is the probability of getting sample data at least as inconsistent with the null hypothesis (and supportive of the alternative hypothesis) as the sample data actually obtained.



Hypothesis Testing

- Steps of hypothesis testing
  - Conclusion
    - ✓ We reject the null hypothesis H_0 in favor of the alternative hypothesis, if the P-value is **lower** than the selected significance level α;
    - ✓ Otherwise, we do not reject the null hypothesis.

Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

```
data = pd.read_csv('bulb.csv')
population = data['Lifespan']

n = 25
sample = population.sample(n, replace=True)
```

Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\,\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

  - Hypotheses

Null hypothesis: $H_0 : \mu \leq \mu_0 = 1340$

Alternative hypothesis: $H_a : \mu > \mu_0 = 1340$

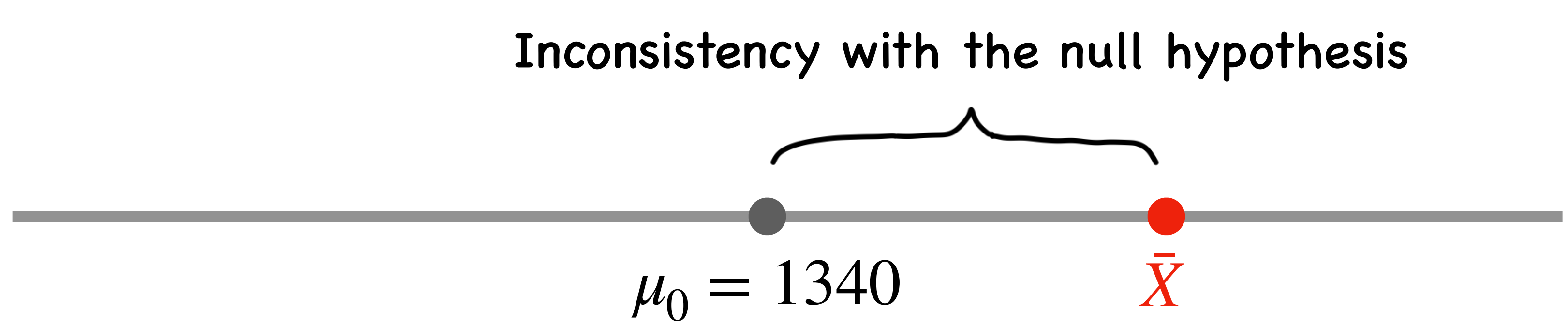
Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\,\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

- Hypotheses

$$\text{Null hypothesis: } H_0 : \mu \leq \mu_0 = 1340$$
$$\text{Alternative hypothesis: } H_a : \mu > \mu_0 = 1340$$

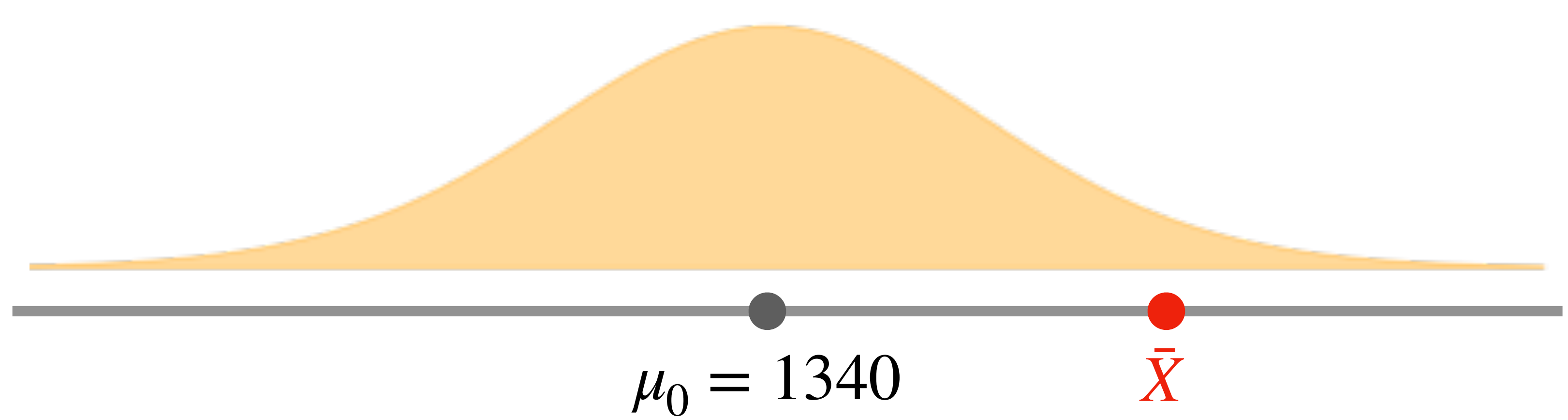


Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\,\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

  - Sampling distributions

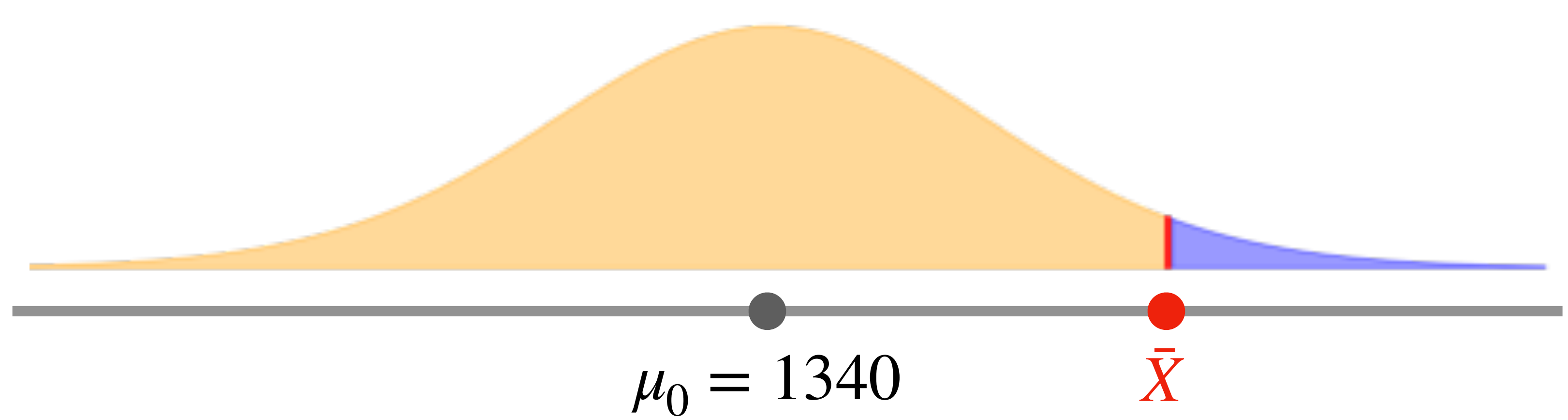


Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\,\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

- Calculation of the P-value



Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\,\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

```
estimate = sample.mean()
s = sample.std()
mu0 = 1340

t_value = (estimate - mu0) / (s/n**0.5)
```

t-value: $T = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim$ t-distribution

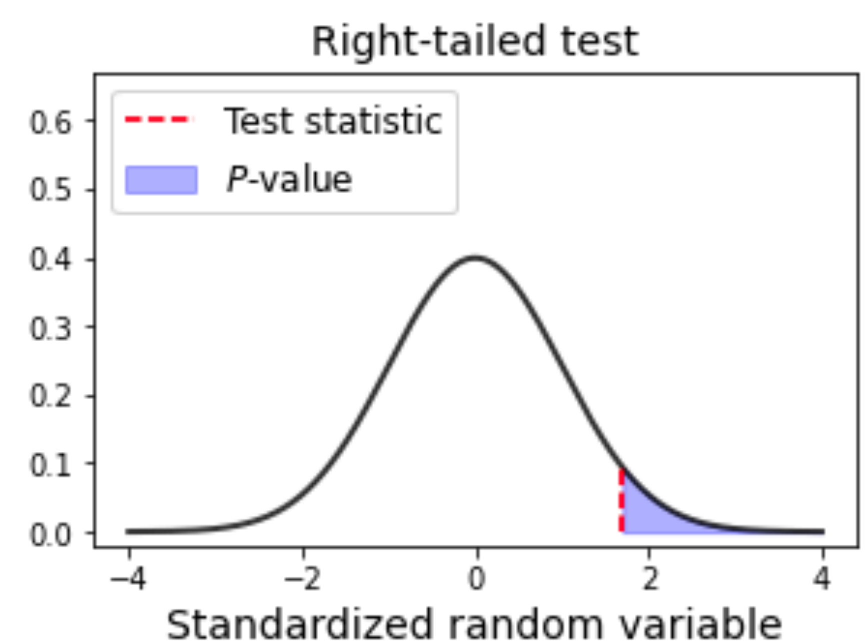
Hypothesis Testing

- Steps of hypothesis testing

Example 5: We randomly select a sample with $n = 25$ records from the "bulb.csv" dataset. Based on the sample data and given the significance level $\alpha = 5\%$, can we conclude that the mean lifespan of all bulbs in this batch is longer than 1340 hours?

```
p_value = 1 - t.cdf(t_value, n-1)
print(f'P-value: {p_value}')
```

```
P-value: 0.021506363623185032
```



Conclusion

- Reject the null hypothesis in favor of the alternative hypothesis.
- The mean lifespan is longer than 1340 hours