

Outline

- We did an introduction to estimation problems:
 - Point estimation
 - Biased vs unbiased estimator
 - Most efficient estimator among unbiased: Smallest variance
 - \bar{X} is an unbiased estimator for the mean.

Today

- Point estimation of variance: Biased vs unbiased estimator
- Confidence intervals for the mean

Unbiased Estimator of Variance

The plugin estimator of variance $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ is **biased**.

Unbiased Estimator of Variance

The plugin estimator of variance $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ is **biased**.

We won't do a formal proof, instead we will do a simulation to verify this.

Simulation:

Do the following for different n (say $n = 2, n = 50$):

Generate 100000 random samples, each of size n , from Z .

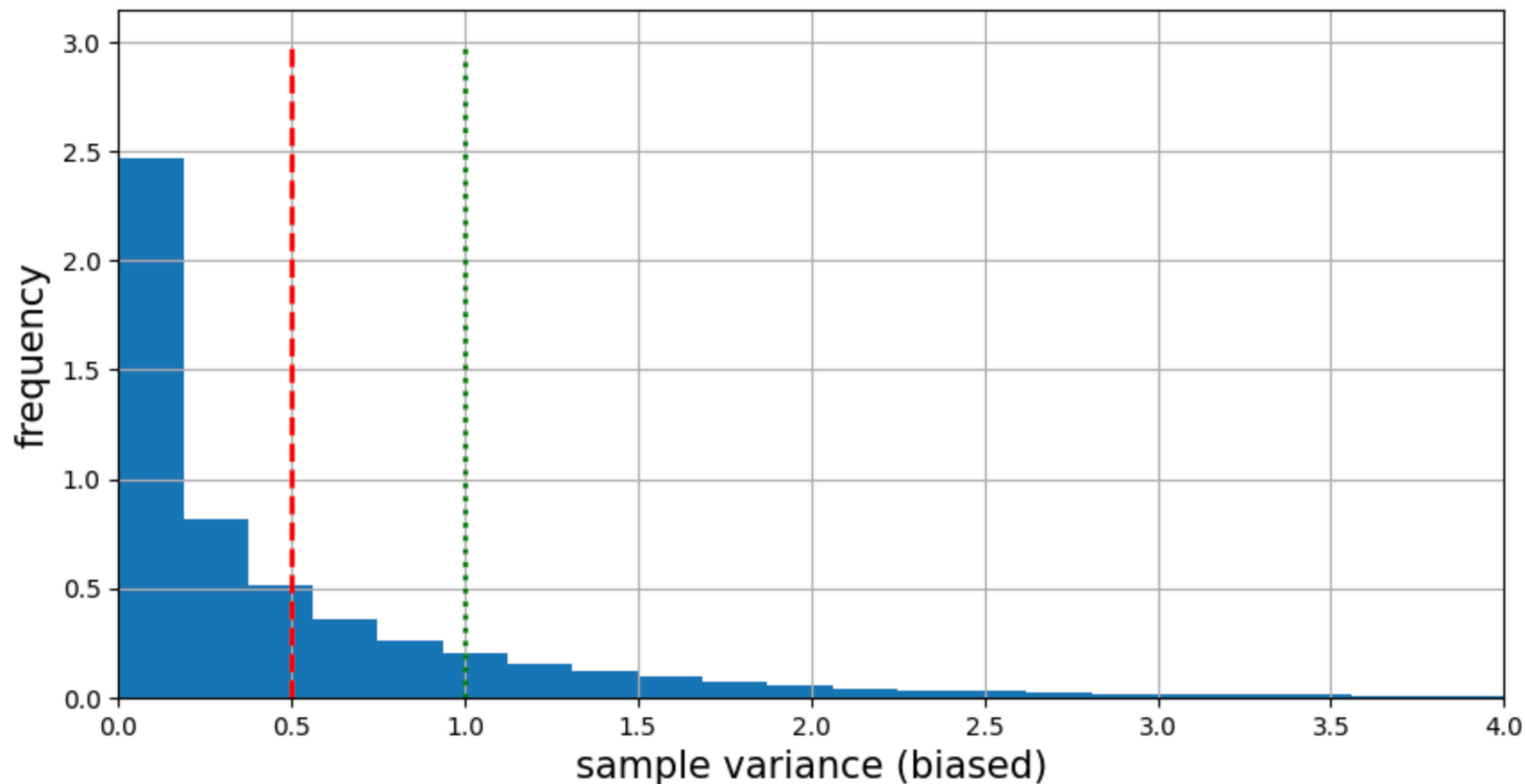
Compute the value of estimator for each random sample.

Plot histogram of all values and show mean value of estimator.

Unbiased Estimator of Variance

The plugin estimator of variance $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ is **biased**.

```
n=2
s = 100000
X = np.random.normal(0,1,[n,s])
# ddof is 0(1) for dividing by n (n-1)
svar_b = np.var(X,axis=0,ddof=0)
mean_svar_b = np.mean(svar_b)
```

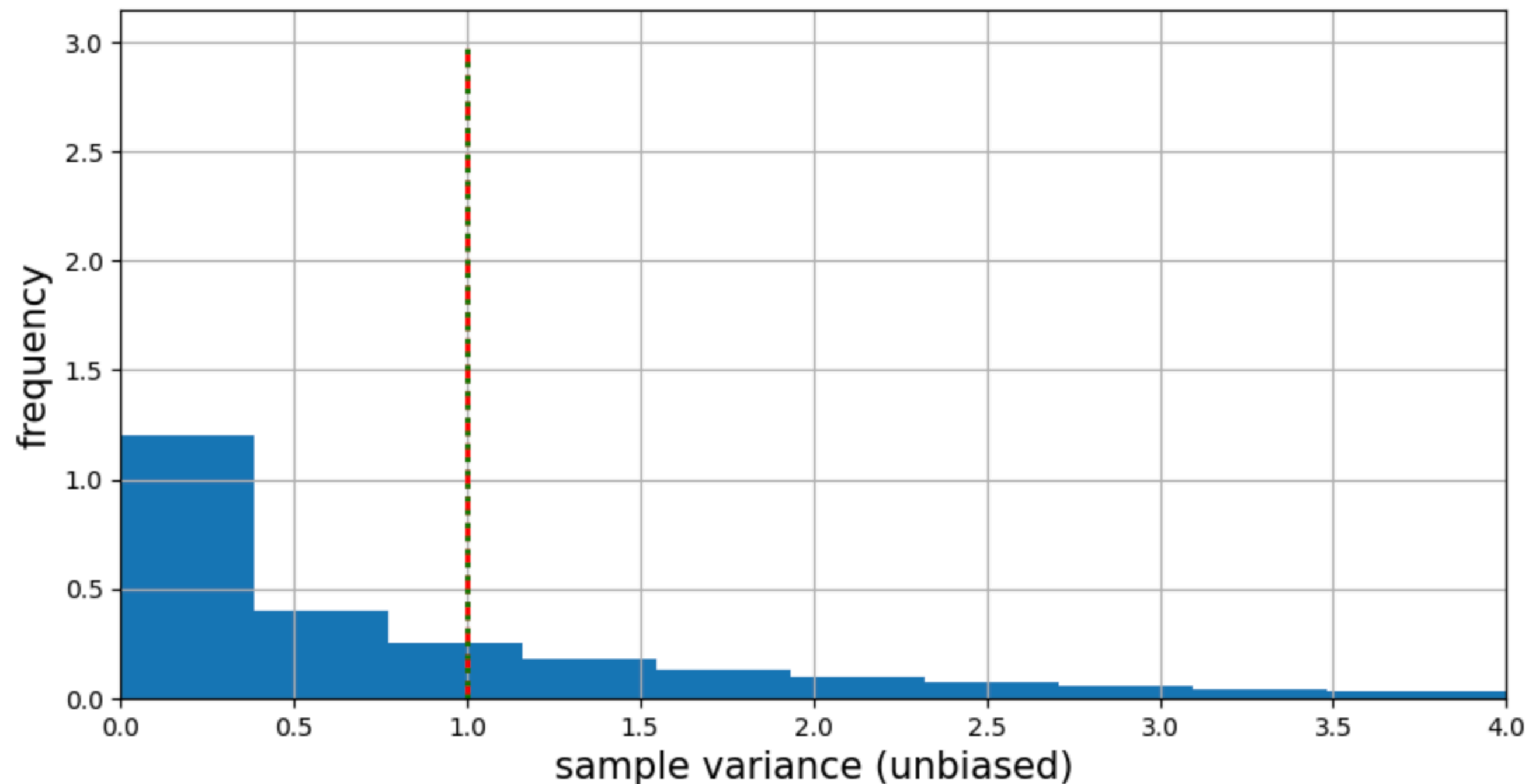


Unbiased Estimator of Variance

The estimator of variance $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is **unbiased**.

```
n=2
s = 100000
X = np.random.normal(0,1,[n,s])
# ddof is 0(1) for dividing by n (n-1)
svar_b = np.var(X,axis=0,ddof=1)
mean_svar_b = np.mean(svar_b)
```

Formal proof in textbook.

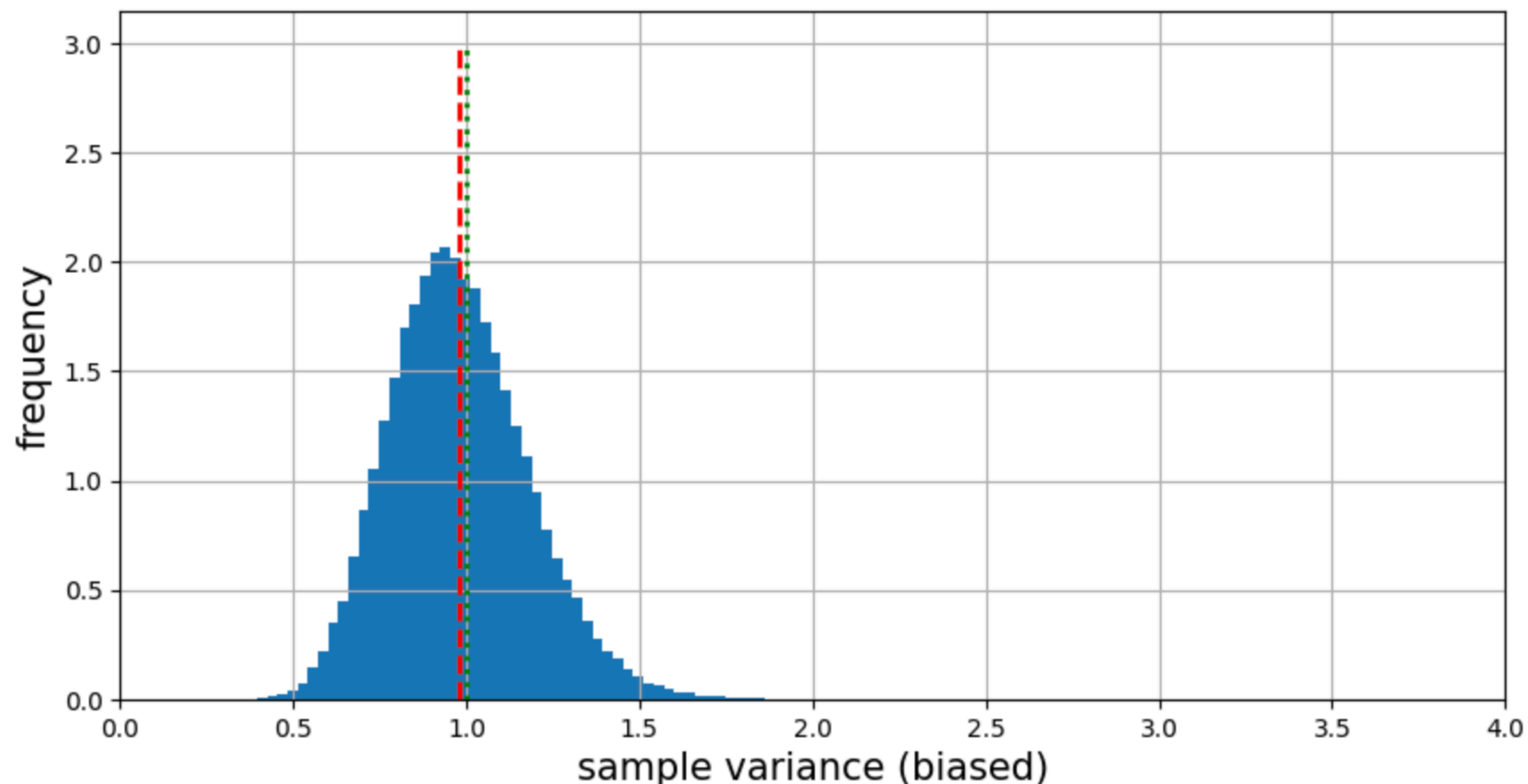


Unbiased Estimator of Variance

Dividing by n vs $n - 1$ has a large effect when n is small.

For large n , they are almost the same!

```
n=50
s = 100000
X = np.random.normal(0,1,[n,s])
# ddof is 0(1) for dividing by n (n-1)
svar_b = np.var(X,axis=0,ddof=0)
mean_svar_b = np.mean(svar_b)
```



More Details About S^2

Observation from Previous Plots:

Distribution of S^2 is skewed for small *dof* and bell-curved for large *dof*.

More Details About S^2

Observation from Previous Plots:

Distribution of S^2 is skewed for small *dof* and bell-curved for large *dof*.

Not a coincidence!

Going back to Section 8.5 of the textbook:

Theorem 8.4: If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2 , then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

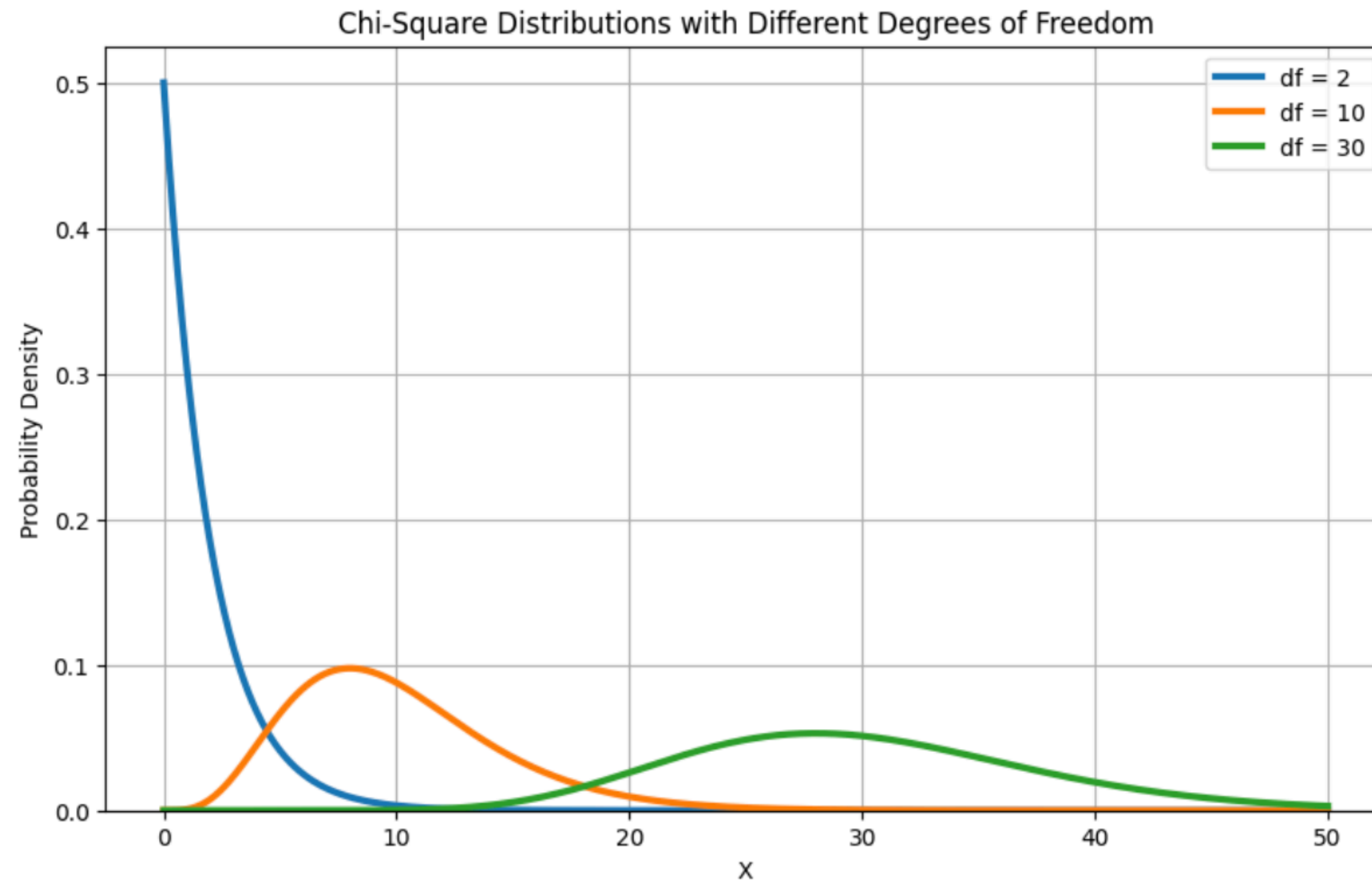
has a chi-squared distribution with $v = n - 1$ degrees of freedom.

More Details About S^2

Observation from Previous Plots:

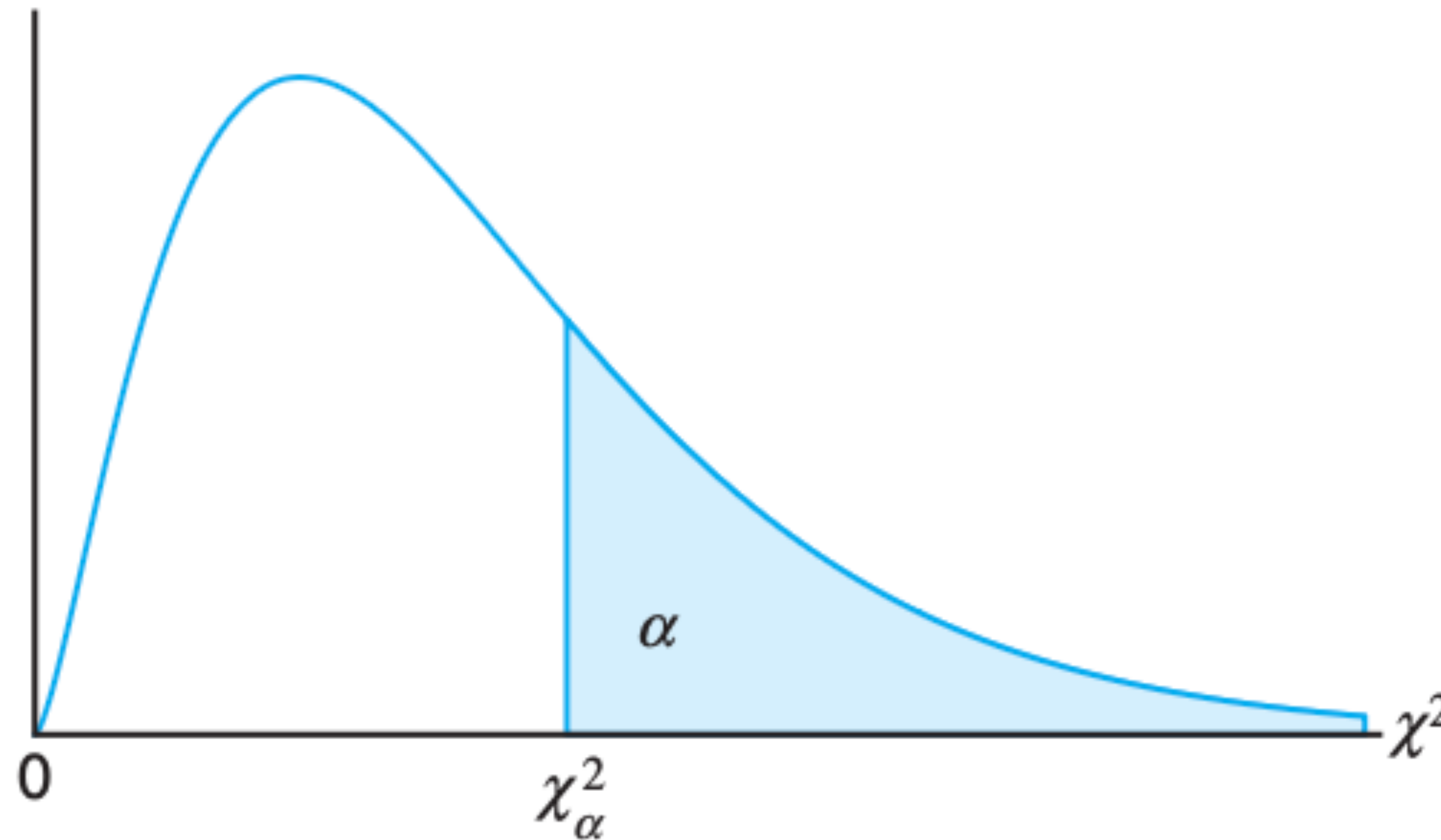
Distribution of S^2 is skewed for small *dof* and bell-curved for large *dof*.

Not a coincidence!



χ^2

More About χ^2



χ^2 value above which the area is α

Note: Given α it can be found by `chi2.ppf(1 - α , ν)` using `scipy.stats`, where ν denotes *dof*.

More About χ^2

8.37 For a chi-squared distribution, find

(a) $\chi_{0.025}^2$ when $v = 15$;

Solution:

```
from scipy.stats import chi2
print(chi2.ppf(0.975, 15))
```

```
27.488392863442975
```

Confidence Intervals

We can apply them to any parameter but most common is mean μ .

We have seen: \bar{X} is a sensible point estimator for mean μ .

Precise estimate, like $\bar{X} = 4$

but how much **confidence** do we have that $\mu = 4$?

Confidence interval:

Make statements like:

with 95 % confidence $\mu \in (\bar{X} - m, \bar{X} + m)$

Confidence Intervals

In general, for interval estimate of a parameter θ , if we find $\widehat{\Theta}_L$, $\widehat{\Theta}_U$ with:

$$P(\widehat{\Theta}_L < \theta < \widehat{\Theta}_U) = 1 - \alpha, \quad \text{for } 0 < \alpha < 1$$

means we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ .

Confidence Intervals

In general, for interval estimate of a parameter θ , if we find $\hat{\Theta}_L$, $\hat{\Theta}_U$ with:

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha, \quad \text{for } 0 < \alpha < 1$$

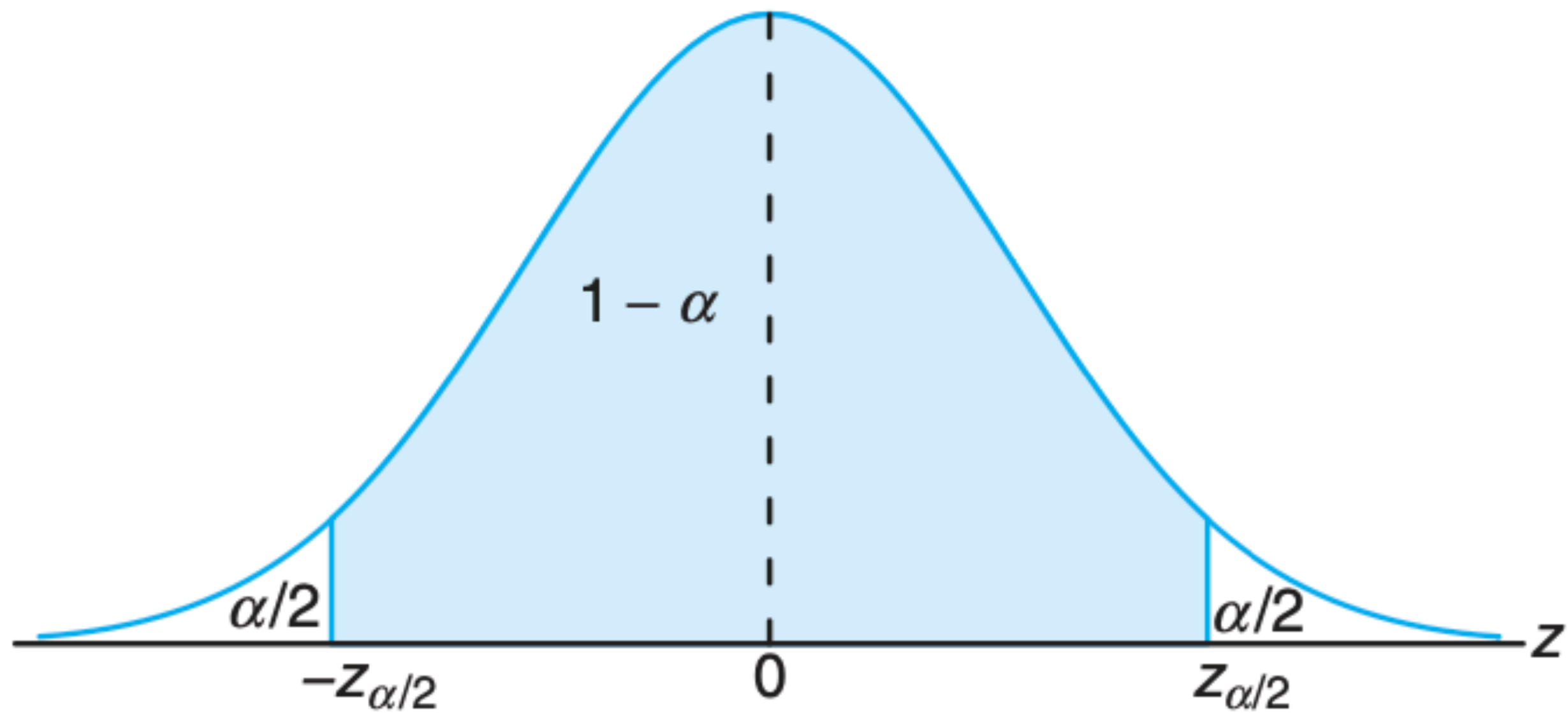
means we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ .

$\hat{\theta}_L < \theta < \hat{\theta}_U$ computed from the selected sample and is called a **$100(1 - \alpha)\%$ confidence interval.**

confidence coefficient (degree of confidence).

$\hat{\theta}_L$, $\hat{\theta}_U$ are called lower and upper **confidence limits.**

Confidence Interval for the Mean

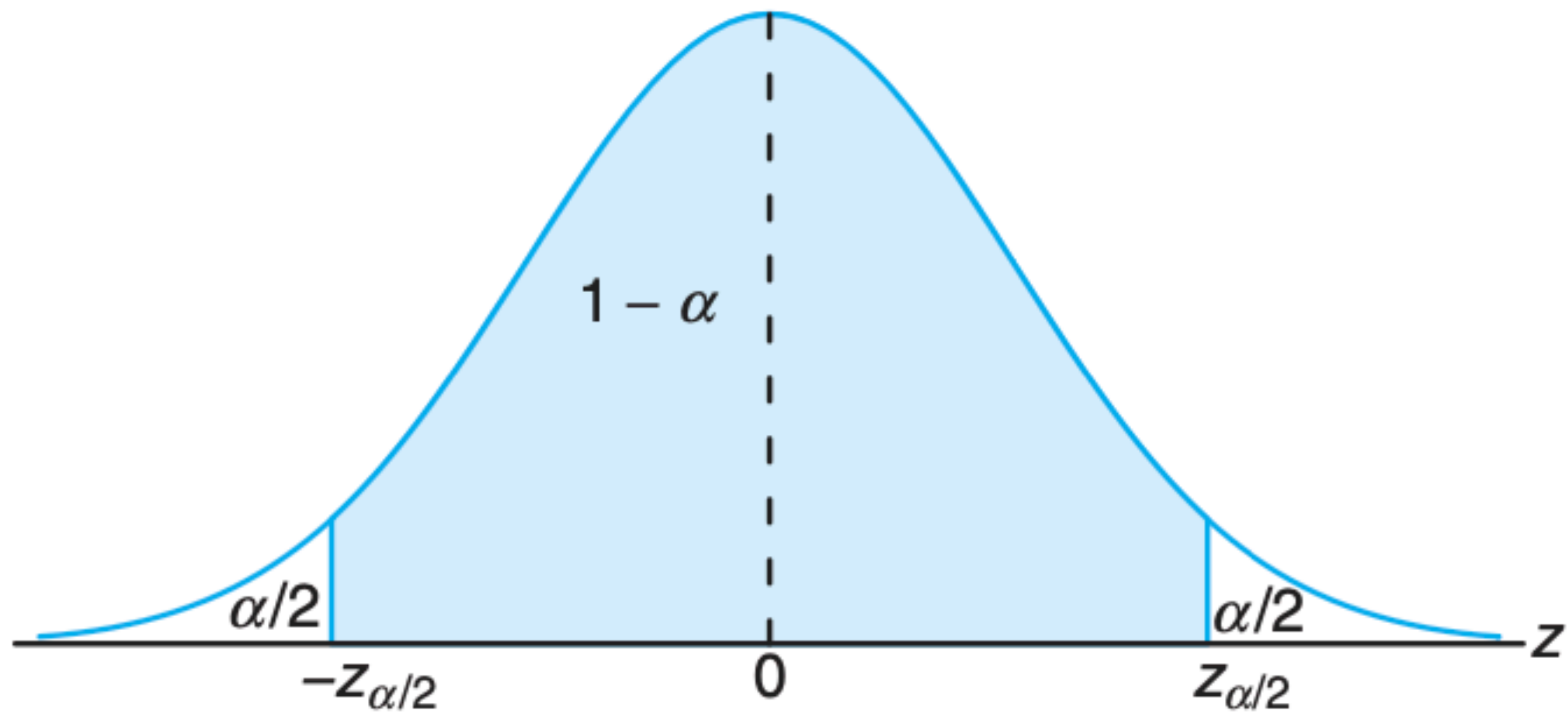


$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

\bar{X} : mean of a large random sample.

With CLT we have $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$.

Confidence Interval for the Mean



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

\bar{X} : mean of a large random sample.

With CLT we have $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$.

Doing some algebra we get:

Confidence
Interval on μ , σ^2
Known

If \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.

Confidence Interval for the Mean

Different samples will yield different interval estimates.

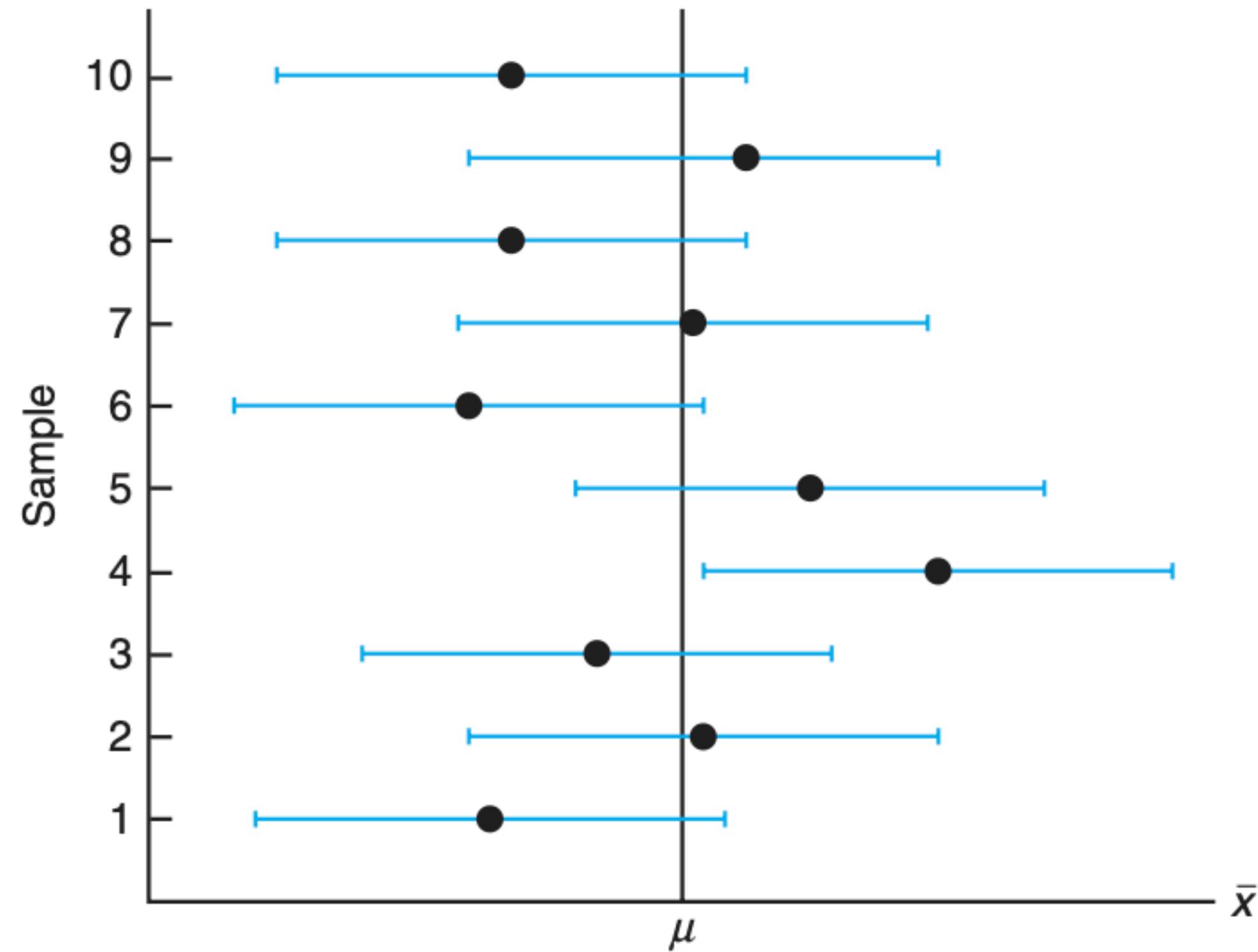
Confidence
Interval on μ , σ^2
Known

If \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.

Confidence Interval for the Mean



Confidence
Interval on μ , σ^2
Known

If \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.

Confidence Interval for the Mean

(Standard deviation of \bar{X}) Standard error of \bar{X}

Confidence
Interval on μ , σ^2
Known

If \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.

Confidence Interval for the Mean

Confidence limit presented equivalently: $\bar{x} \pm z_{\alpha/2} s . e . (\bar{x})$

(Standard deviation of \bar{X}) Standard error of \bar{X}

Confidence
Interval on μ, σ^2
Known

If \bar{x} is the mean of a random sample of size n from a population with known variance σ^2 , a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.

Confidence Interval for the Mean

Example 9.2: The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

Use $z_{0.025} = 1.96$ and $z_{0.005} = 2.575$

Confidence Interval for the Mean

Example: Number of citations of a random paper is a random variable with $\sigma = 5$. In a sample of 100 papers sample mean \bar{X} is 7.5. What is the 95 % confidence interval for mean μ ? Use $z_{0.025} = 1.96$ or $z_{0.05} = 1.645$

Choose which one to use.