Chapter 8: Statistical Analysis of Simulated Data With Confidence Intervals for the Variance

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Outline

8.1 The Sample Mean and Sample Variance

8.2 Interval Estimates of a Population Mean

8.2½ Confidence Intervals for the Variance

8.3 The Bootstrapping Technique for Estimating the Mean Square Error

Homework

8.1 The Sample Mean and Sample Vairince

Suppose we have $X_1,\ldots,X_n$ independent and identically distributed all from $f(X)$. Let $\theta=E[X_i]$ and $\sigma^2=\text{var}[X_i]$, i.e. same mean and variance.

With the arithmetic mean being $\bar{X}=\sum_{i=1}^{n}\frac{X_i}{n}$,

we know that
$$\begin{aligned} E\left[\bar{X}\right] &= E\left[\sum_{i=1}^{n}\frac{X_i}{n}\right] \\ &= \sum_{i=1}^{n}\frac{E\left[X_i\right]}{n} \\ &= \frac{n\theta}{n}=\theta \quad . \end{aligned}$$

8.1 The Sample Mean and Sample Variance

If the expected value of a statistic is equal to the parameter it is estimating, it is said to be an unbiased estimator.

To determine the "worth" of $\bar{X}$ an estimator for θ,
We look at expected squared difference.

$$
\begin{aligned}
E\left[(\bar{X}-\theta)^2\right] &= \operatorname{var}(\bar{X}) \\
&= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var}\left(X_i\right) \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

8.1 The Sample Mean and Sample Vairince

By Chebyshev's inequality

$$P\left\{|\bar{X}-\theta|>\frac{c\sigma}{\sqrt{n}}\right\}\leq\frac{1}{c^2}$$

.

But using the Central Limit Theorem when n is large,

$$P\left\{|\bar{X}-\theta|>\frac{c\sigma}{\sqrt{n}}\right\}=P\{|Z|>c\}=2[1-\Phi(c)]$$

where Φ is the cumulative distribution function of the standard normal distribution.

$$c=1.96$$

$$P\{\}=\frac{1}{(1.96)^2}=.2603$$

$$=0.05$$

8.1 The Sample Mean and Sample Vairince

If we define S^2 to be

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

we know that it is unbiased because

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) ,$$

and because the mean of a χ^2 is $df=(n\text{-}1)$, therefore

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \longrightarrow \frac{(n-1)}{\sigma^2}E\left[S^2\right] = n-1 \longrightarrow E\left[S^2\right] = \sigma^2 .$$

8.1 The Sample Mean and Sample Vairince

In a simulation, we often generate an extremely large number of random variates (i.e. 10^6).
It would be great if we knew when we had enough.

Assume that we are interested in estimating the value of $\theta = E[X_i]$.
One stopping rule is to specify a standard deviation d for $\bar{X}$.

Then, continue generating random variates until $S/\sqrt{n} < d$.
When n is small the following is recommended.

8.1 The Sample Mean and Sample Variance

Method for Determining When to Stop Generating New Data

1. Choose an acceptable value of d for the standard deviation of the estimator.
2. Generate at least 100 data values.
3. Continue to generate additional data values, stopping when you have generated k values and $S/\sqrt{k}<d$, where S is the sample standard deviation based on those k values.
4. The estimate of θ is given by $\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$.

8.2 Interval Estimates of a Population Mean

Assume we have $X_1,\ldots,X_n$ iid all from the same distribution $f(X)$.
We use $\bar{X}$ as a "point" estimator for the population mean θ.

We can also generate an "interval" estimator for θ.

We know that $E[\bar{X}]=\theta$ and $\mathrm{Var}[\bar{X}]=\dfrac{\sigma^2}{n}$.

We use the fact that when n is large, $\bar{X}$ has an approximate normal distribution, i.e. $\bar{X} \dot{\sim} N(\theta,\sigma^2/n)$.

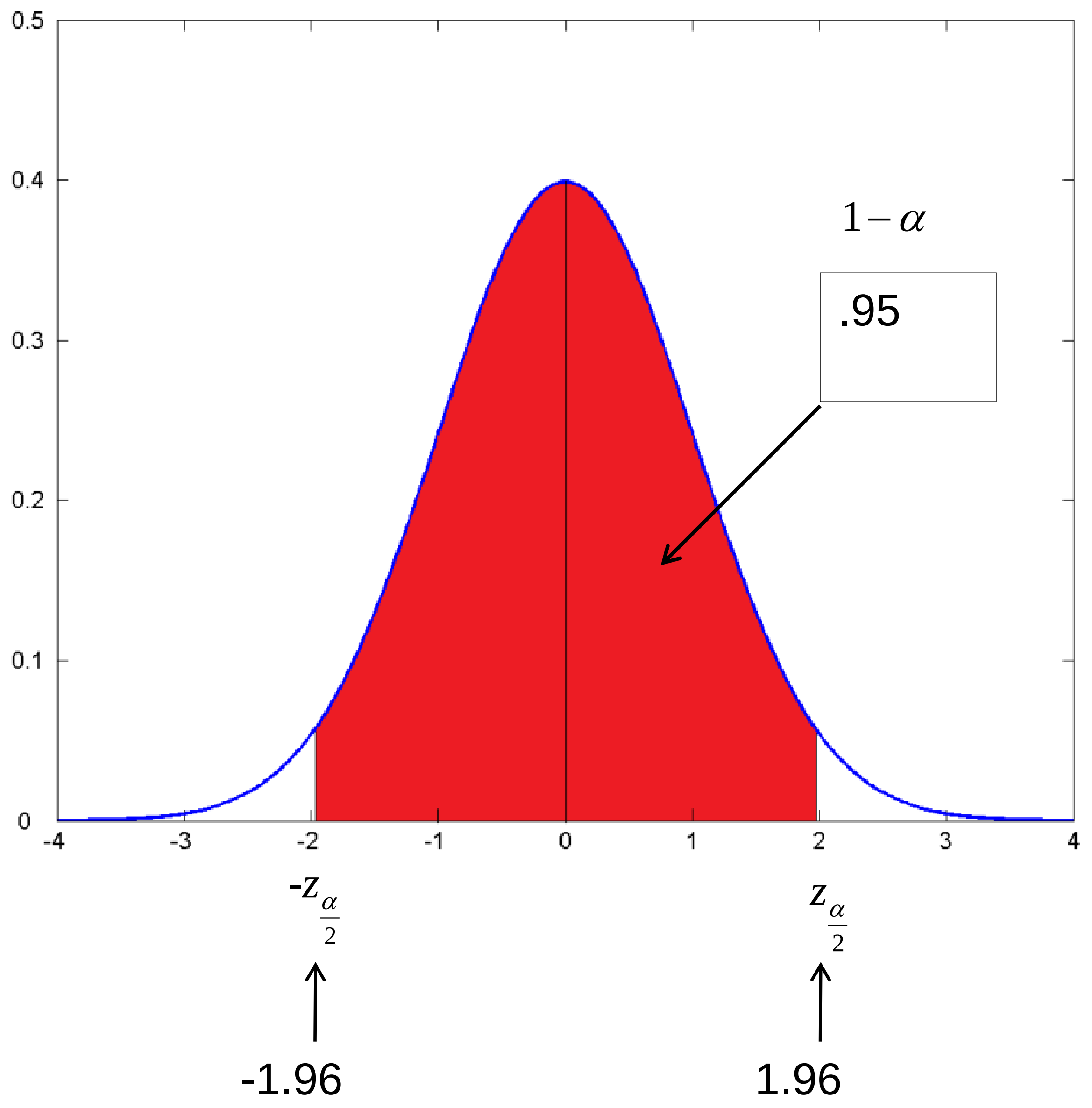
8.2 Interval Estimates of a Population Mean

What this implies is that $z = \dfrac{\bar{X} - \theta}{\sigma / \sqrt{n}}$

has an approximate standard deviation!

$P(-1.96 < z < 1.96) = 0.95$

or more generally, $P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) = 1 - \alpha$.



8.2 Interval Estimates of a Population Mean

The inequality

$$P\left(\boxed{-z_{\frac{\alpha}{2}} < z} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\begin{aligned}
-z_{\frac{\alpha}{2}} &< z \\
-z_{\frac{\alpha}{2}} &< \frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} \\
-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &< \bar{X}-\mu \\
-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}-\bar{X} &< -\mu \\
\bar{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} &> \mu
\end{aligned}$$

8.2 Interval Estimates of a Population Mean

The inequality

$$P(-z_{\frac{\alpha}{2}} \boxed{< z < z_{\frac{\alpha}{2}}}) = 1-\alpha$$

$$\begin{array}{ccc}
z & < & z_{\frac{\alpha}{2}} \\
\dfrac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} & < & z_{\frac{\alpha}{2}} \\
\bar{X}-\mu & < & z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} \\
-\mu & < & z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}-\bar{X} \\
\mu & > & \bar{X}-z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}
\end{array}$$

8.2 Interval Estimates of a Population Mean

We can see the equivalency of these statements

$$P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) = 1-\alpha \;\longrightarrow\; P\left\{\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Thus a $(1\text{-}\alpha)\times 100\%$ confidence interval for θ is

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

which if α=0.05, a 95% confidence interval for θ is

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\quad.$$

8.2 Interval Estimates of a Population Mean

Using similar logic, it is also true that when σ is unknown,

a $(1\text{-}\alpha) \times 100\%$ confidence interval for θ is

$$\bar{X} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$$

$t =$

$$t = \frac{\bar{X} - \theta}{s / \sqrt{n}} \to z = \frac{\bar{X} - \theta}{\sigma / \sqrt{n}}$$

As n increases,
s converges to σ,
and z converges to z.

and if n is large,

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

which if $\alpha=0.05$, a 95% confidence interval for θ is

$$\bar{X} - 1.96 \frac{s}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{s}{\sqrt{n}} \, .$$

8.2 Interval Estimates of a Population Mean

For Bernoulli random variates, where

$$X_i = \begin{cases} 1 \text{ with probability } & p \\ 0 \text{ with probability } & 1-p \end{cases}$$

we have the same scenario. Using similar logic,
a $(1\text{-}\alpha) \times 100\%$ confidence interval for p is

$$P\left\{ \bar{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} < p < \bar{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right\} = 1-\alpha$$

when n is large.

Central Limit Theorem

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

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

$\mu = np \quad \sigma^2 = np(1-p)$.

$\bar{X}$ is the average of Bernoulli random variables

8.2½ Confidence Intervals for the Variance

We know that if $x_1,..,x_n$ are iid $N(\mu,\sigma^2)$ then the distribution of $\dfrac{(n-1)s^2}{\sigma^2}$ is a χ^2 with n-1 degrees of freedom.

A χ^2 distribution with n-1 degrees of freedom has a mean of n-1 and a variance of $2(n$-$1)$.

This means that the mean and variance of s^2 are σ^2 and $\dfrac{2\sigma^4}{(n-1)}$!

$$E(s^2)=\sigma^2 \qquad \mathrm{var}(s^2)=\frac{2\sigma^4}{(n-1)}$$

8.2½ Confidence Intervals for the Variance

Following the general $PE \pm CV \times SE(PE)$ procedure,
the confidence interval for the variance should be

$$s^2 \pm \chi^2(\frac{\alpha}{2})\sqrt{\frac{2\sigma^4}{n-1}} \quad ?$$

This is what you were taught in your first stats class.

Is this correct even though the χ^2 distribution is not symmetric?

The answer is no.

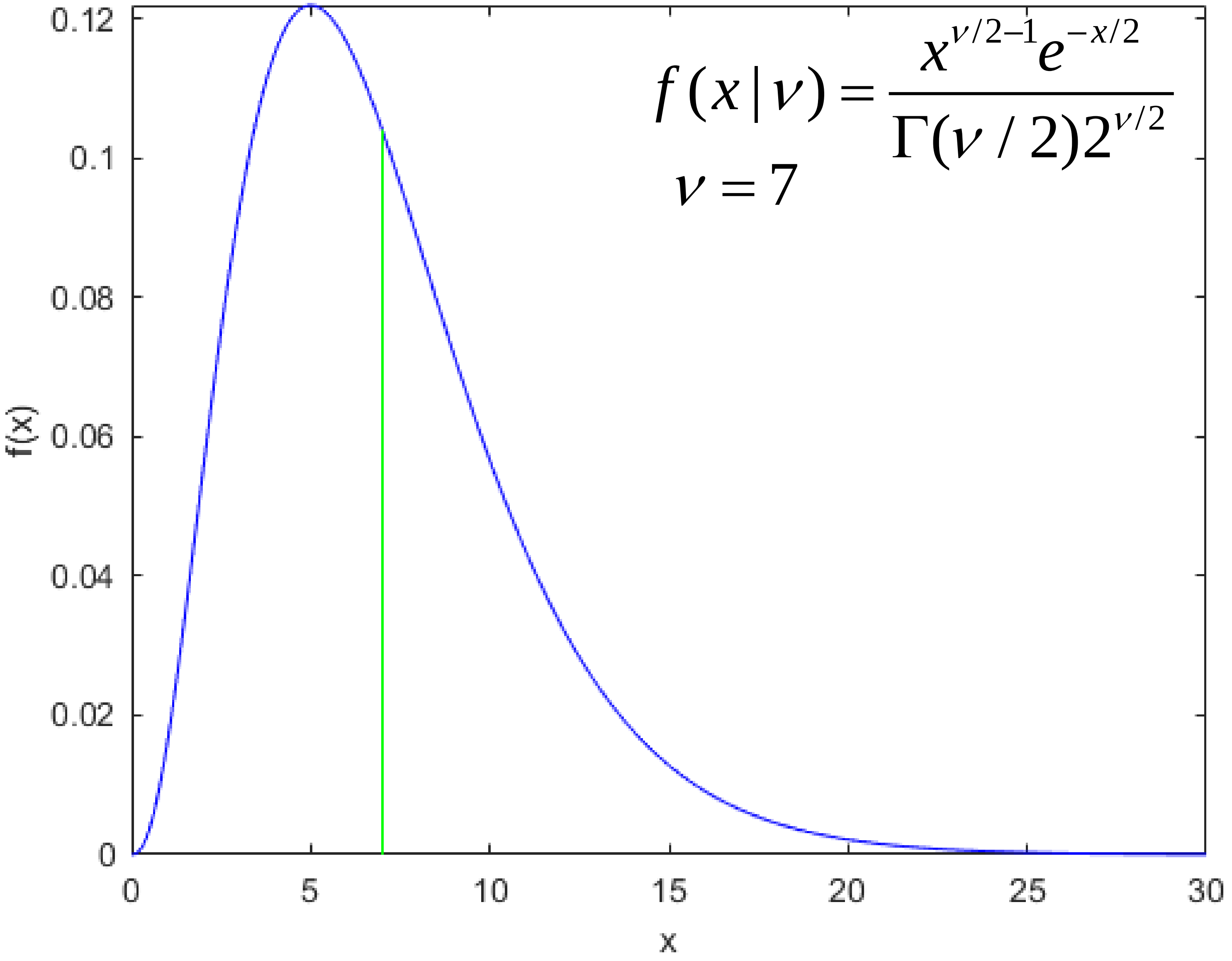
8.2½ Confidence Intervals for the Variance

We know $x = \dfrac{(n-1)s^2}{\sigma^2}$ has a chi-square PDF with $(n\text{-}1)$ degrees of freedom

$$f(x|\nu) = \frac{x^{\nu/2-1}e^{-x/2}}{\Gamma(\nu/2)2^{\nu/2}}$$

$$E(x|\nu) = \nu$$

$$\mathrm{var}(x|\nu) = 2\nu$$



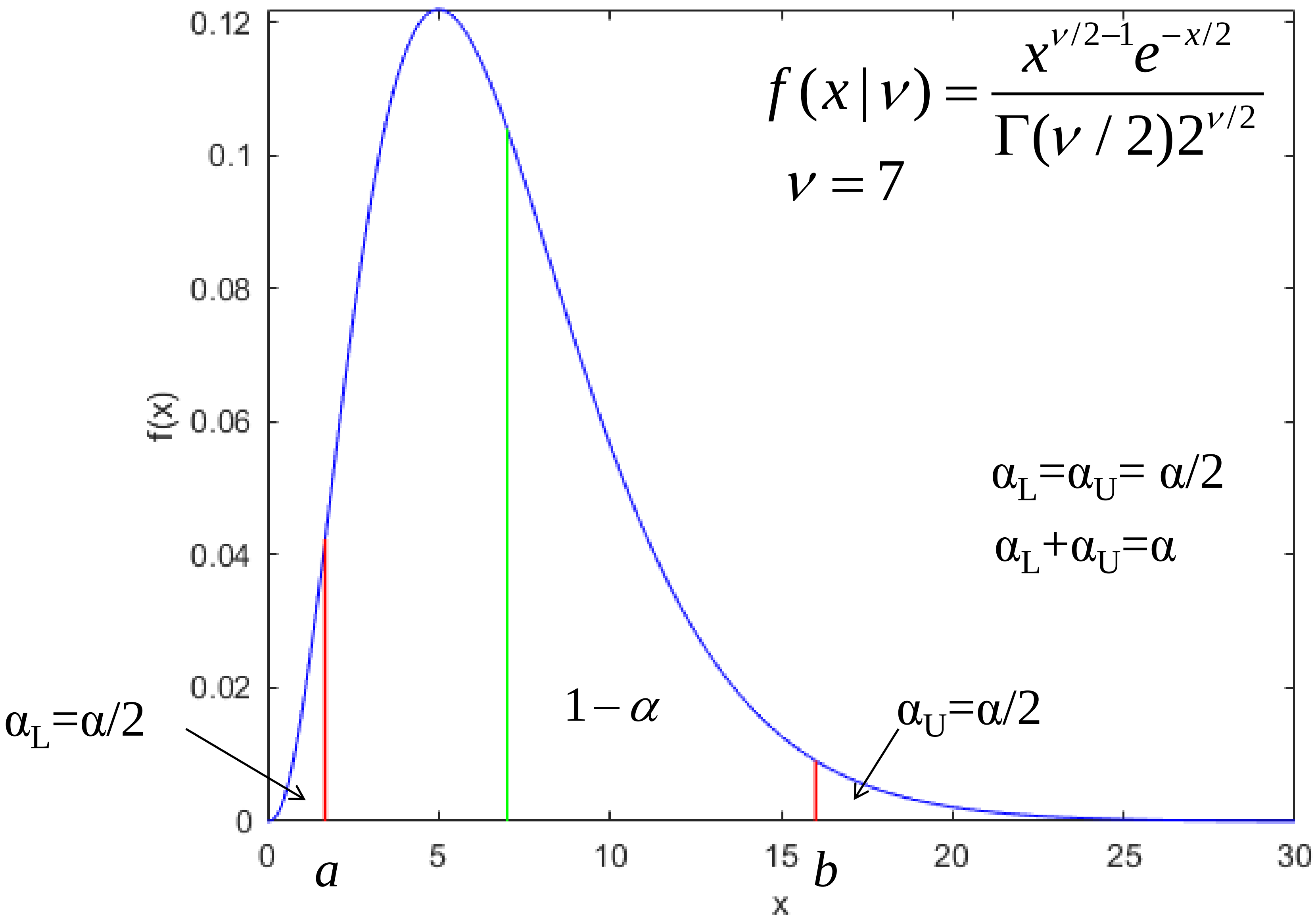
8.2½ Confidence Intervals for the Variance

So we should be able to find a and b such that

$$P\left\{a<\frac{(n-1)s^2}{\sigma^2}<b\right\}=1-\alpha$$

$$\int_0^a f(x)dx=\frac{\alpha}{2}$$

$$\int_0^b f(x)dx=1-\frac{\alpha}{2}$$



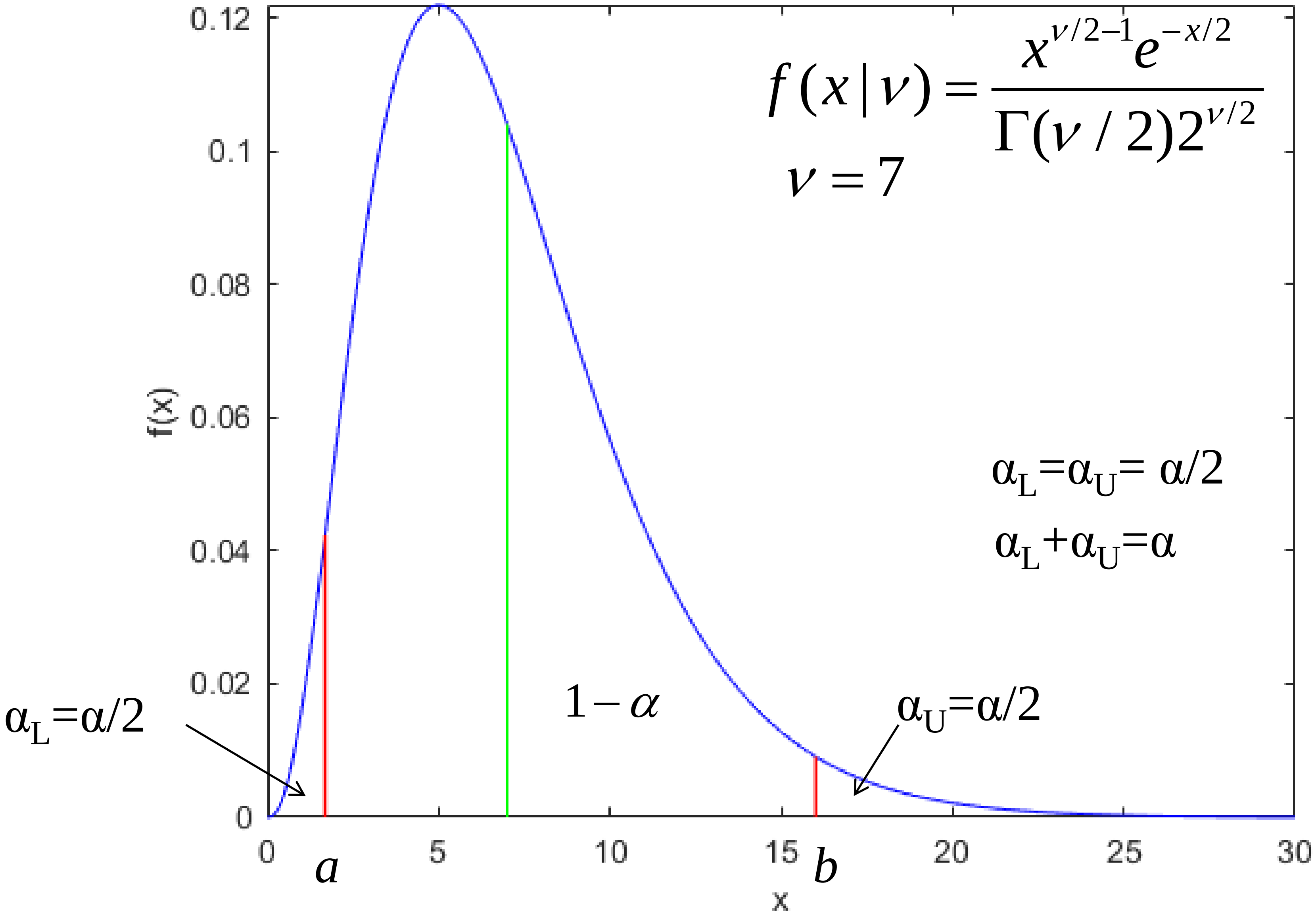
8.2½ Confidence Intervals for the Variance

Once we have a and b, we can look at

$$P\left\{a < \frac{(n-1)s^2}{\sigma^2} < b\right\} = 1-\alpha$$

then do a little algebra to get

$$P\left\{\frac{(n-1)s^2}{b} < \sigma^2 < \frac{(n-1)s^2}{a}\right\} = 1-\alpha \quad .$$



8.2½ Confidence Intervals for the Variance

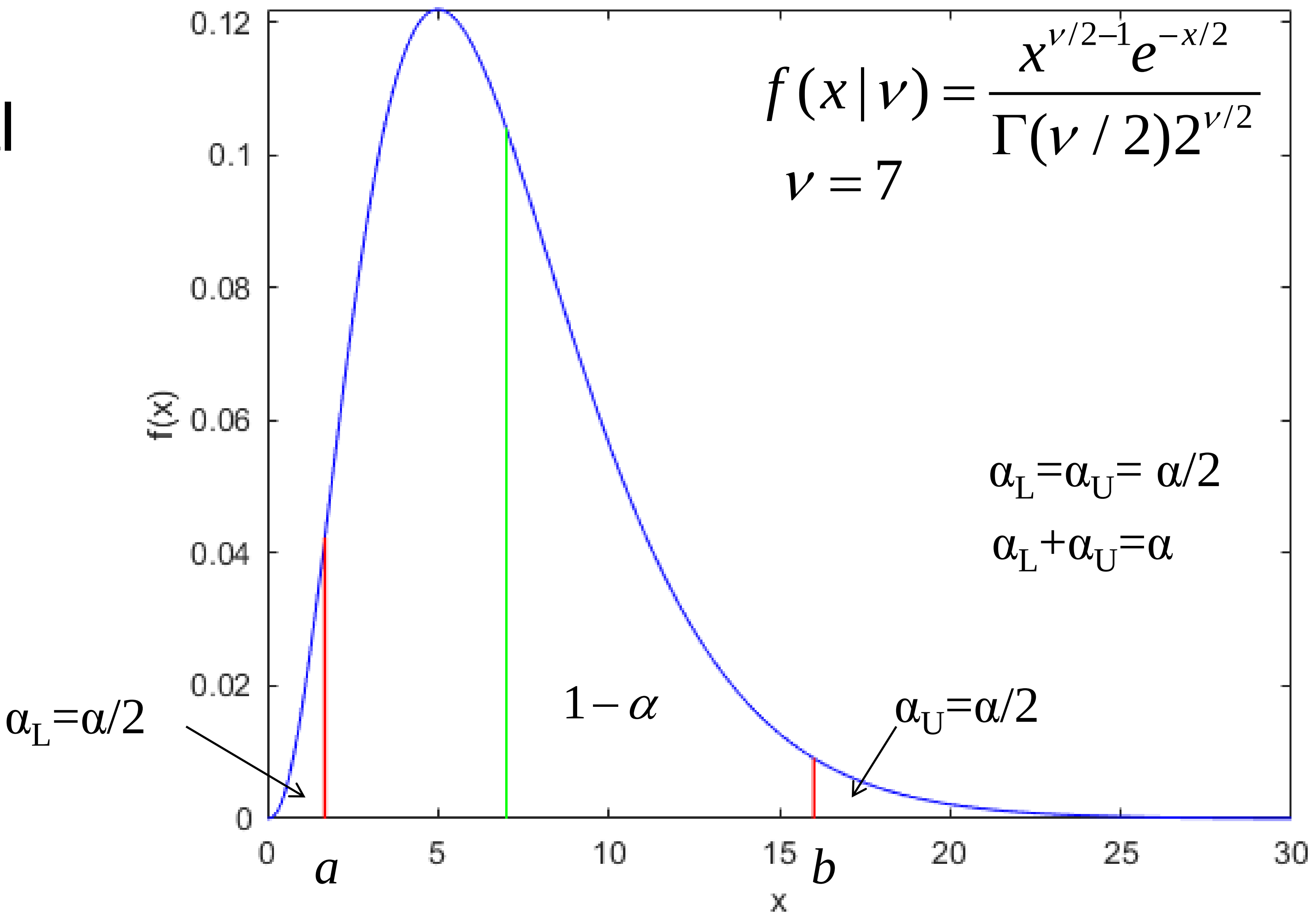
So $\frac{(n-1)s^2}{b} < \sigma^2 < \frac{(n-1)s^2}{a}$

is a $100(1\text{-}\alpha)\%$ confidence interval for σ^2.

$a = 1.6899$, $b = 16.0128$

$L = 0.0625s^2$, $U = 0.5918s^2$

$U\text{-}L = 0.5293s^2$



8.2½ Confidence Intervals for the Variance

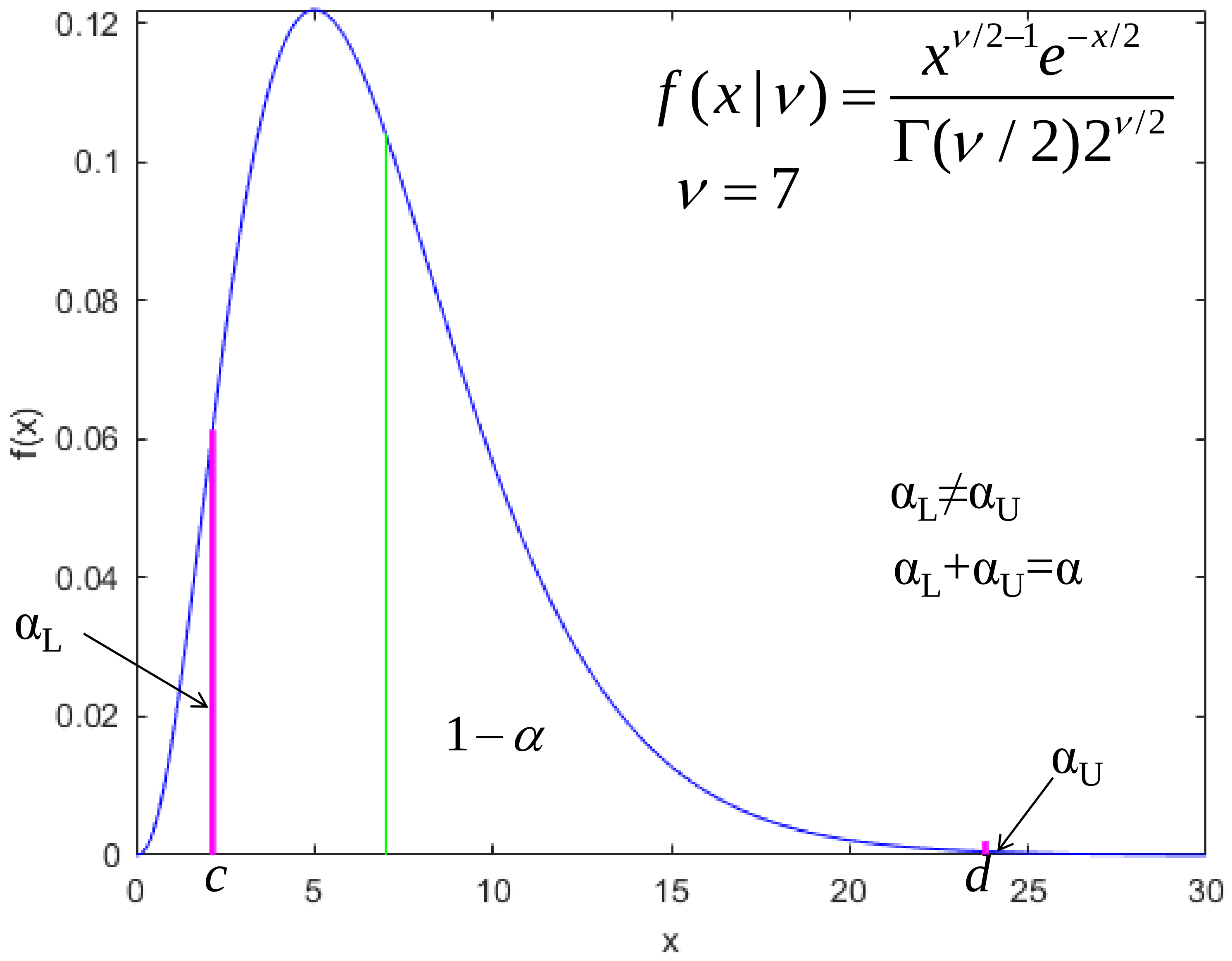
But this confidence interval

$$\frac{(n-1)s^2}{b} < \sigma^2 < \frac{(n-1)s^2}{a}$$

is not best!

We can find a minimum length confidence interval for σ^2 where the probability in each tail is not equal.

$$\frac{(n-1)s^2}{d} < \sigma^2 < \frac{(n-1)s^2}{c}$$



Tate and Klett, JASA 1959.

8.2½ Confidence Intervals for the Variance

So the goal is to minimize

$$\frac{(n-1)s^2}{d} < \sigma^2 < \frac{(n-1)s^2}{c}$$

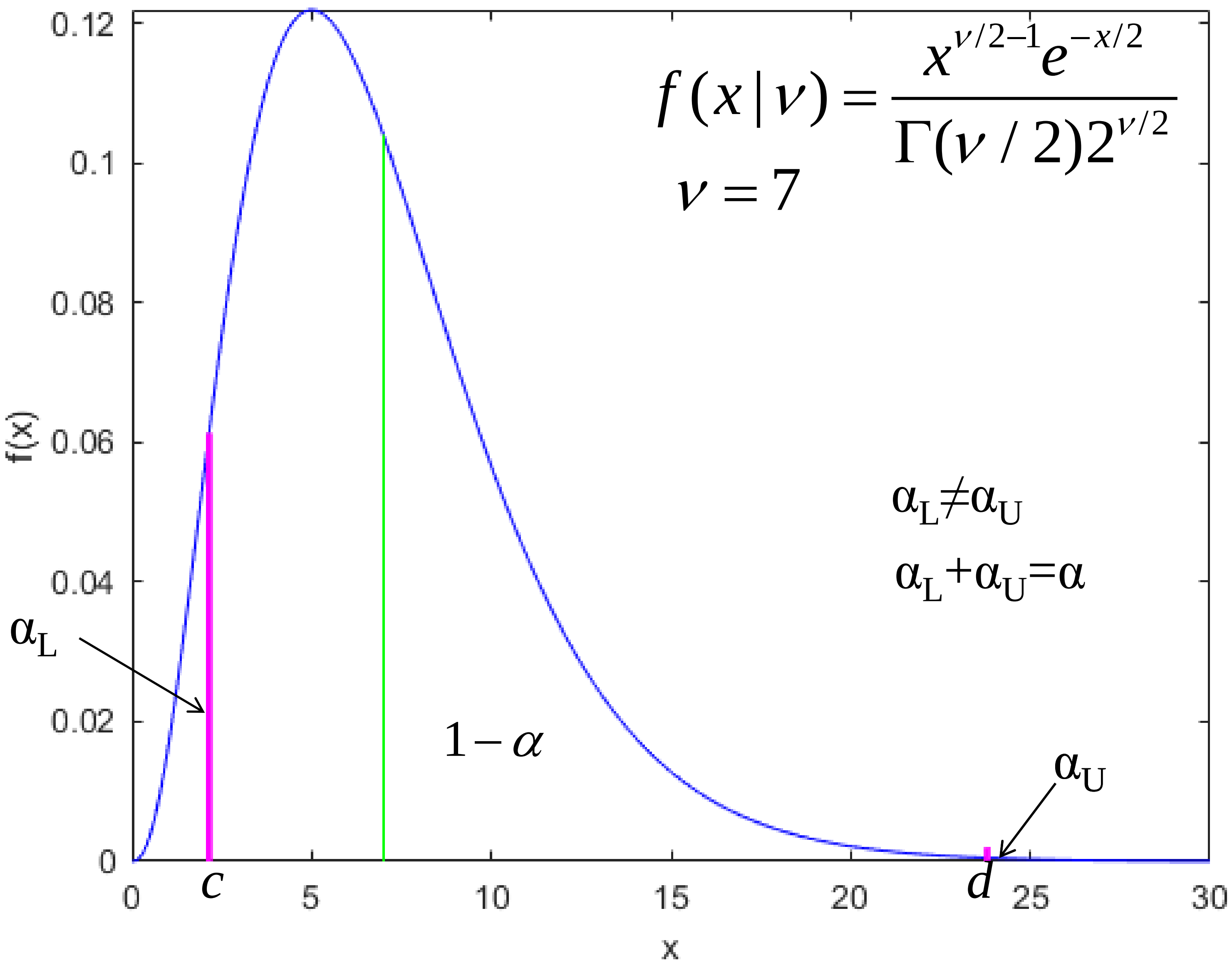
subject to the constraint

that $\int_c^d f(x)\,dx = 1-\alpha$.

Some amount α_L in lower tail

and some amount α_U in upper tail.

$\alpha_L + \alpha_U = \alpha$



Tate and Klett, JASA 1959.

8.2½ Confidence Intervals for the Variance

$$\frac{(n-1)s^2}{d} < \sigma^2 < \frac{(n-1)s^2}{c}$$

In terms of a cost/score function,

$$\phi = \left(\frac{1}{c} - \frac{1}{d}\right)(n-1)s^2 + \lambda\left(\int_c^d f(x)\ dx - 1 + \alpha\right)$$

where λ is the Lagrange multiplier.

a =1.6899, b =16.0128

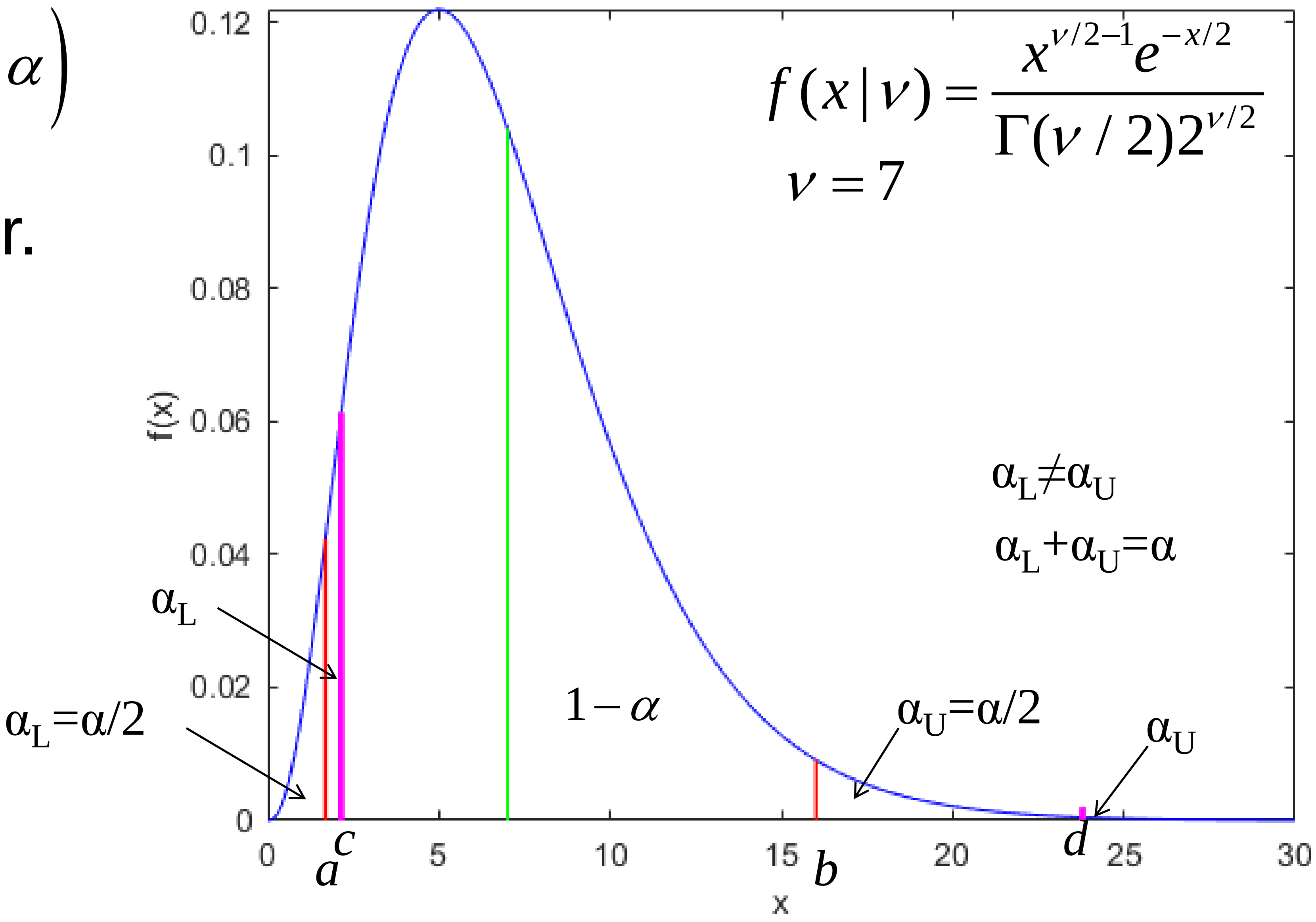
L=0.0625(n-1)s^2, U=0.5918(n-1)s^2

U-L=**0.5293**(n-1)s^2

c =2.1473, d =23.7944

L=0.0420(n-1)s^2, U=0.4657(n-1)s^2

U-L=**0.4237**(n-1)s^2.



Tate and Klett, JASA 1959.

8.3 The Bootstrapping Technique for Estimating Mean Squares

Assume that $X_1,\ldots,X_n$ are independent and identically distributed from cumulative distribution function F.

If θ is a parameter of interested and $g(X_1,\ldots,X_n)$ an estimator, we would like to estimate the value of

$$MSE(F) = E_F[(g(X_1,\ldots,X_n) - \theta(F))^2]$$

we can usually estimate it analytically if F is known

8.3 The Bootstrapping Technique for Estimating Mean Squares

But when F is not known, all we have is $X_1,\ldots,X_n$.

As we know we can estimate F by the empirical CDF

$$F_e(x) = \frac{\text{number of } i : X_i \leq x}{n}$$

F_e should be "close" to F especially if n is large and

F_e converges to F as $n \to \infty$.

8.3 The Bootstrapping Technique for Estimating Mean Squares

Let's examine the bootstrap approximation to the MSE.

when we don't need it. Assume $\theta=\mu$ and $g(X_1,\ldots,X_n) = \bar{X}$.

Then we know that $MSE = E[(\bar{X}-\mu)^2] = \sigma^2/n$,

which we would estimate by S^2/n.

To estimate the MSE via bootstrap, we have to calculate

$$MSE(F_e) = E_{F_e}[(g(X_1,\ldots,X_n) - \theta(F_e))^2]$$

8.3 The Bootstrapping Technique for Estimating Mean Squares

If we think of $X_1,\ldots,X_n$ as a population of values, then the vector $(x_1,\ldots,x_n)$, where each element is drawn from $X_1,\ldots,X_n$ with replacement can take on n^n possible values.

The MSE is then approximately

$$MSE(F_e) = \sum_{i_n} \cdots \sum_{i_1} \frac{[(g(X_{i_1},\ldots,X_{i_n}) - \theta(F_e))^2]}{n^n} \qquad i_j \in \{1,\ldots,n\},\ j = 1,\ldots,n$$

8.3 The Bootstrapping Technique for Estimating Mean Squares

The MSE is approximately

$$MSE(F_e) = \sum_{i_n} \cdots \sum_{i_1} \frac{[(g(X_{i_1},\ldots,X_{i_n}) - \theta(F_e))^2]}{n^n} \qquad i_j \in \{1,\ldots,n\},\ j = 1,\ldots,n$$

But this requires summing n^n terms, a daunting task.

If $n=20$, then there are 1.0486×10^{26} terms!

To get around this, we use simulation and approximate the empirical MSE.

8.3 The Bootstrapping Technique for Estimating Mean Squares

From $X_1,\ldots,X_n$, generate r samples of size n with replacement

$$
\begin{array}{ccccc}
X_1^{(1)},\ldots,X_n^{(1)} & & Y_1 & = & [(g(X_1^{(1)},\ldots,X_n^{(1)})-\theta(F_e)]^2 \\
X_1^{(2)},\ldots,X_n^{(2)} & \longrightarrow & Y_2 & = & [(g(X_1^{(2)},\ldots,X_n^{(2)})-\theta(F_e)]^2 \\
\vdots & & \vdots & & \\
X_1^{(r)},\ldots,X_n^{(r)} & & Y_r & = & [(g(X_1^{(r)},\ldots,X_n^{(r)})-\theta(F_e)]^2
\end{array}
$$

$$
Y_1,Y_2,\ldots,Y_r \longrightarrow MSE(F_e) \approx \frac{1}{r}\sum_{i=1}^{r} Y_i
$$

8.3 The Bootstrapping Technique for Estimating Mean Squares

From $X_1,\ldots,X_n$, generate r samples of size n with replacement

$$
\begin{array}{ccccc}
X_1^{(1)},\ldots,X_n^{(1)} & & Y_1 & = & [s_{(1)}^2 - s^2(X_1,\ldots,X_n)]^2 \\
X_1^{(2)},\ldots,X_n^{(2)} & \longrightarrow & Y_2 & = & [s_{(2)}^2 - s^2(X_1,\ldots,X_n)]^2 \\
\vdots & & \vdots & & \\
X_1^{(r)},\ldots,X_n^{(r)} & & Y_r & = & [s_{(r)}^2 - s^2(X_1,\ldots,X_n)]^2
\end{array}
$$

$$
Y_1,Y_2,\ldots,Y_r \longrightarrow MSE(s_{Fe}^2) \approx \frac{1}{r}\sum_{i=1}^{r} Y_i
$$

Discussion

Questions?

Homework 10

1. Generate 10^6 sets of 8 random data values from a normal $\mu=100$, $\sigma=3$.
   Calculate s^2 for each.
   Make a histogram and form eCDF.
   Compare the eCDF percentiles to the theoretical percentiles.

2. Find the 4% minimum length CI for σ^2 when we have $\nu=7$.
   Compare the min length Confidence Interval values to the usual
   2% in each tail. Generically assume $s^2=1$. Comment.

Homework 10

$$E[\bar{x}] = \mu \qquad E[s^2] = \sigma^2$$
$$\text{var}[\bar{x}] = \frac{\sigma^2}{n} \qquad \text{var}[s^2] = \frac{2\sigma^4}{n-1}$$

3. Generate n=25 random numbers from a normal distribution with μ=100 and σ=5. Compute $\bar{x}$ and s^2.
   Generate m=10^5 bootstrap samples of size n=25 from your sample.
   a) Compute the mean and variance of each sample.
   b) Make a histograms of means and variances in a).
   c) Compute mean and variance of means and variances in a).
   d) Compute bootstrap estimate of $var(s^2)$.
   e) Compare theoretical values to bootstrap values.
   g) Repeat with larger/smaller n.
   h) Comment.