ECE 307 – Techniques for Engineering Decisions

13. Data Uses

George Gross

Department of Electrical and Computer Engineering

University of Illinois at Urbana-Champaign

FOCUS OF DATA USAGE TOPIC

- Use of historical data for the construction of probability distributions
- The interpretation of probability information
- Use of estimators
- Application example

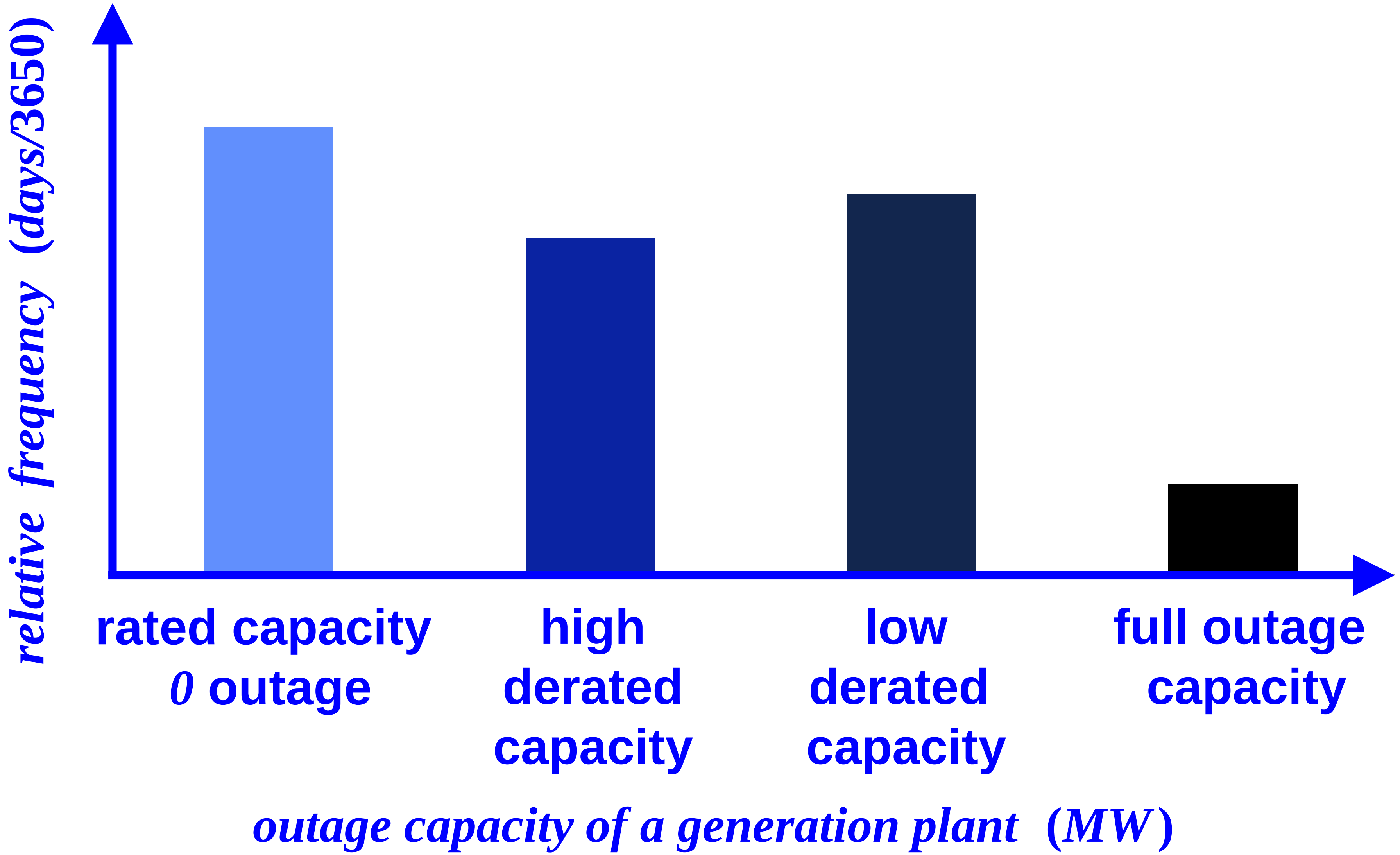
EXAMPLE

- Consider the interpretation of the statement

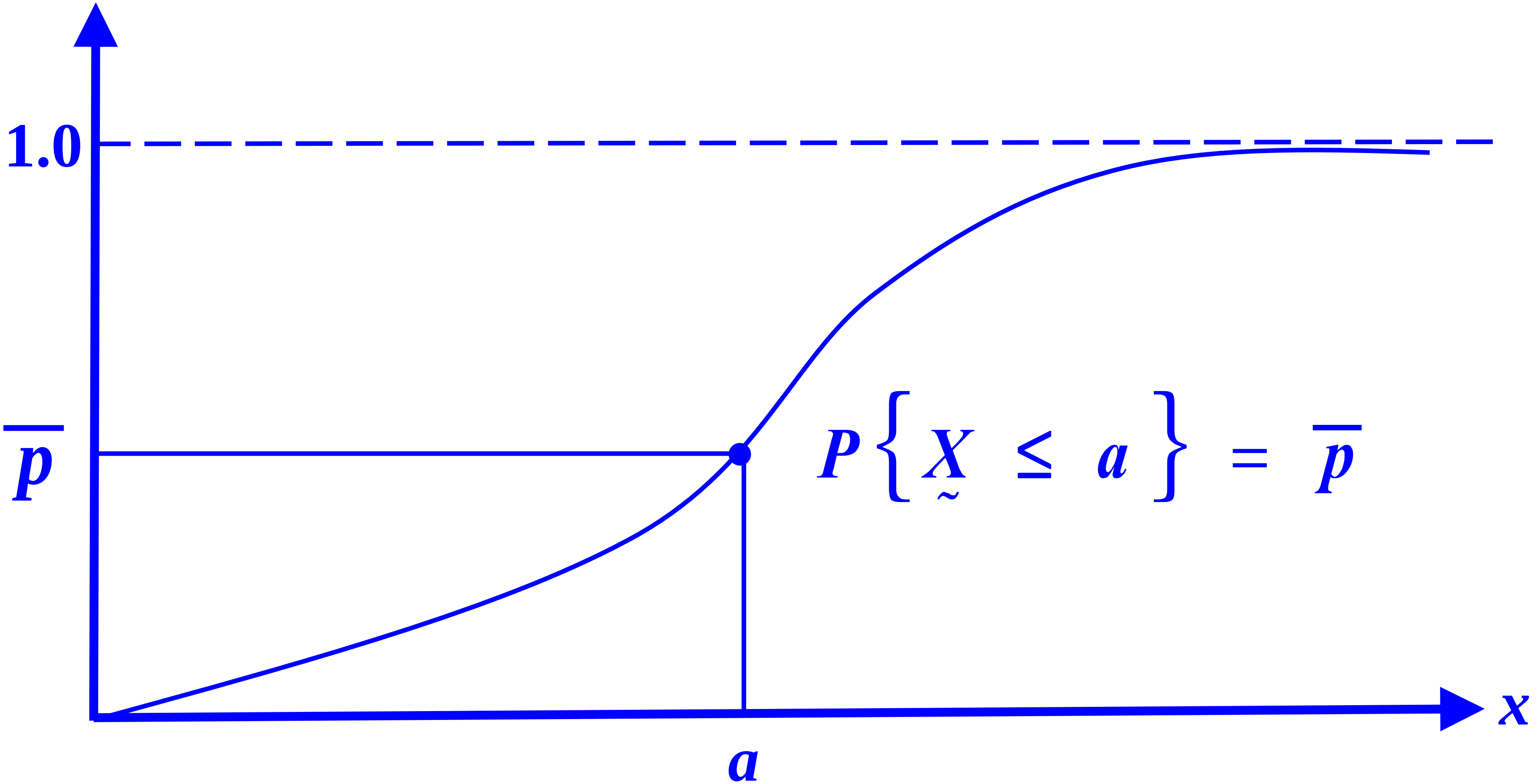
$$P\left\{ \textit{sunny day in June in Champaign} \right\} = 0.53$$

- We obtain this probability from 20 years of June weather data in Champaign, with each day classified as either *sunny* or *not sunny*
- The 600 June days of data indicate that 318 or 53 % of these days are classified as sunny
- Given the long – term historical behavior in the data, the probability of 0.53 makes sense

USE OF HISTOGRAMS



CONSTRUCTION OF THE *c.d.f.*



STATISTICAL PARAMETER ESTIMATORS

- An estimator is a *r.v.* that can be used to estimate the value of a parameter of interest
- Consider a *r.v.* $\underset{\sim}{X}$ whose statistical parameters we wish to estimate
- We consider a set of *r.v.*s $\left\{ \underset{\sim}{X}_i , i = 1, 2, \ldots, n \right\}$ where each $\underset{\sim}{X}_i$ is independent of $\underset{\sim}{X}_j$, $i \neq j$, and each $\underset{\sim}{X}_i$ has the same distribution as $\underset{\sim}{X}$; we refer to this set as a set of n *independent, identically distributed* or *i.i.d. r.v.*s

STATISTICAL PARAMETER ESTIMATORS

- We use the set of n *i.i.d. r.v.*s $\left\{ \underset{\sim}{X}_i ,\ i = 1, 2, \ldots, n \right\}$ to construct estimators for the moments of $\underset{\sim}{X}$
- We focus on the estimators for two key parameters of $\underset{\sim}{X}$:
  - mean of $\underset{\sim}{X}$
  - variance of $\underset{\sim}{X}$

STATISTICAL PARAMETER ESTIMATORS

- The *sample mean estimator* is the *r.v.*

$$\underset{\sim}{\bar{X}} = \frac{\sum_{i=1}^{n} \underset{\sim}{X}_i}{n}$$

- In practice, we obtain an estimate of the mean by using the (observed) realizations of the n *r.v.*s $\underset{\sim}{X}_i$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

STATISTICAL PARAMETER ESTIMATORS

- The estimator of the sample variance is given by the *r.v.*

$$\underset{\sim}{S}^{2} = \frac{\sum_{i=1}^{n}\left(\underset{\sim}{X}_{i} - E\{\underset{\sim}{X}\}\right)^{2}}{n-1}$$

- We obtain an estimate of the variance by using the observed realizations of the *n r.v.*s $\underset{\sim}{X}_{i}$

$$s^{2} = \frac{\sum_{i=1}^{n}\left(x_{i} - \bar{x}\right)^{2}}{n-1}$$

STATISTICAL PARAMETER ESTIMATORS

- An equivalent way to think about the computation of the estimate is to draw n random samples from the sample space $\underset{\sim}{X}$
- We collect the set of n random samples $\left\{ x_1, x_2, \ldots, x_n \right\}$ of the *r.v.* $\underset{\sim}{X}$: these are n *randomly* drawn values from the sample space of $\underset{\sim}{X}$

STATISTICAL PARAMETER ESTIMATORS

- The value $\bar{x}$ computed with the set of random samples provides an estimate of

$$\mu = E\{\underset{\sim}{X}\}$$

- The value s^2 computed with the set of random samples provides an estimate of

$$\sigma^2 = var\{\underset{\sim}{X}\}$$

EXAMPLE: TACO SHELLS

- This application example focuses on taco shells and is concerned with the high breakage rate in the shipment of most taco shells: typical rate is 10 – 15 %
- A company with a new shipping container claims to have a lower – approximately 5 % breakage rate
- This company's price is $ 25 for a 500 – taco shell box vs. $ 23.75 for a 500 – taco shell box of the current supplier

EXAMPLE: TACO SHELLS

- A test run using 12 boxes from the new company and 18 boxes from the current company is performed and used for comparison purposes: we randomly pick the elements to construct the set

$$\left\{ x_1, x_2, \ldots, x_{12} \right\}$$

from the sample space of the *r.v.* $\underset{\sim}{X}$ to represent

EXAMPLE: TACO SHELLS

the number of unbroken shells from the new company and the elements to construct the set

$$\left\{ y_1, y_2, \ldots, y_{18} \right\}$$

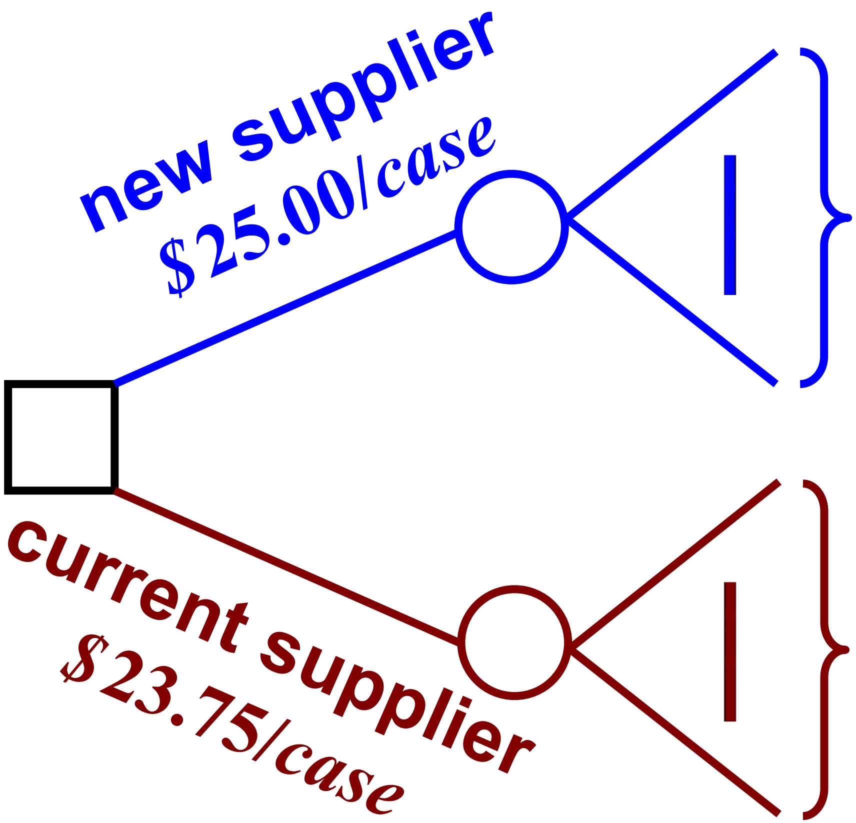
from the sample space of the *r.v.* $\underset{\sim}{Y}$ to represent those of the current company

- We tabulate the data of the useable shells from the two suppliers

UNBROKEN TACO SHELLS

new supplier		*current supplier*		
468	467	444	441	450
474	469	449	434	444
474	484	443	427	433
479	470	440	446	441
482	463	439	452	436
478	468	448	442	429

EXAMPLE: TACO SHELLS



12 samples of the number of unbroken shells (x)

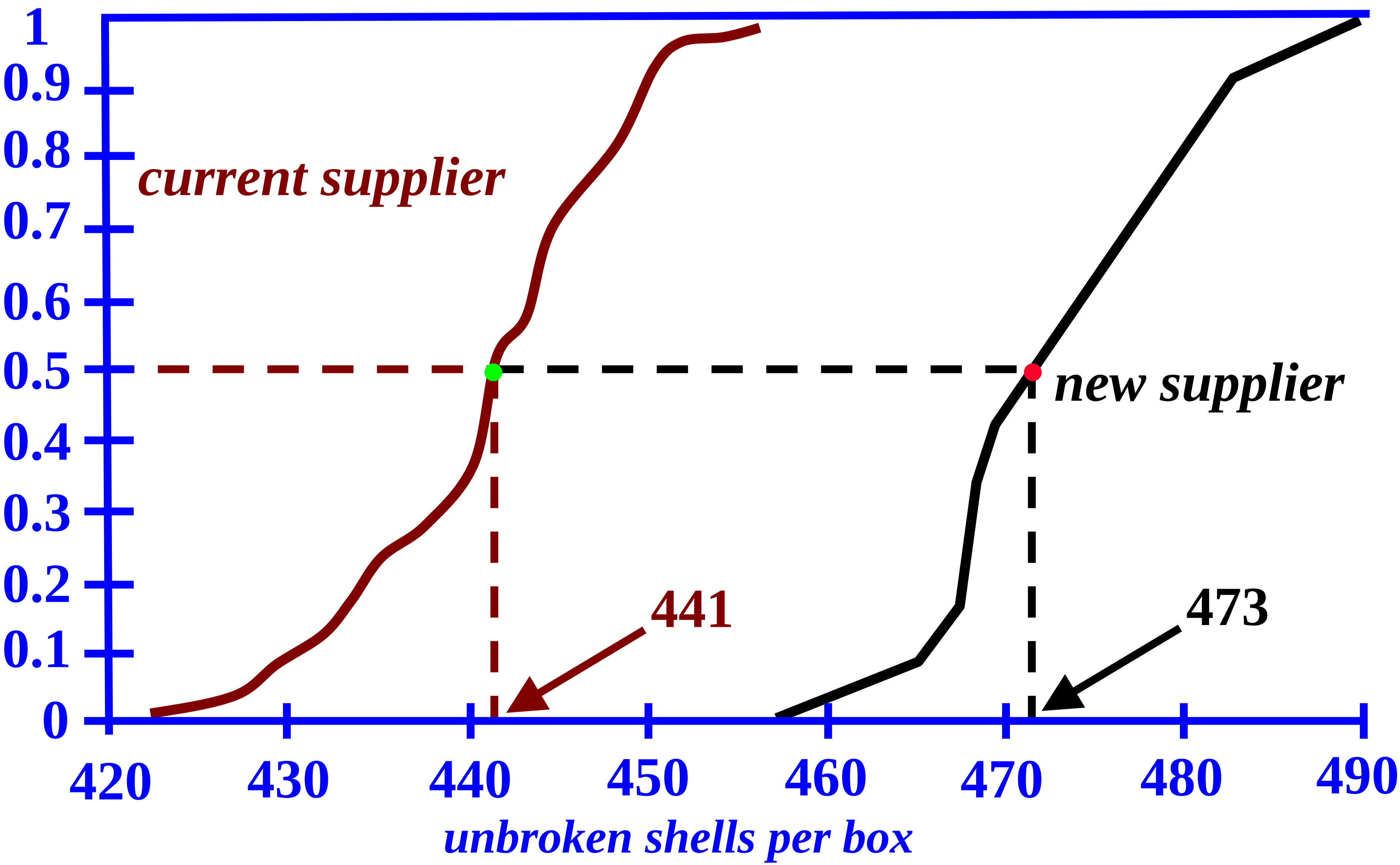
costs per unbroken shell

$$\frac{25}{x}$$

$$\frac{23.75}{y}$$

18 samples of the number of unbroken shells (y)

c.d.f.s CONSTRUCTED FOR THE TWO SUPPLIERS



*c.d.f.*s OF THE TWO SUPPLIERS

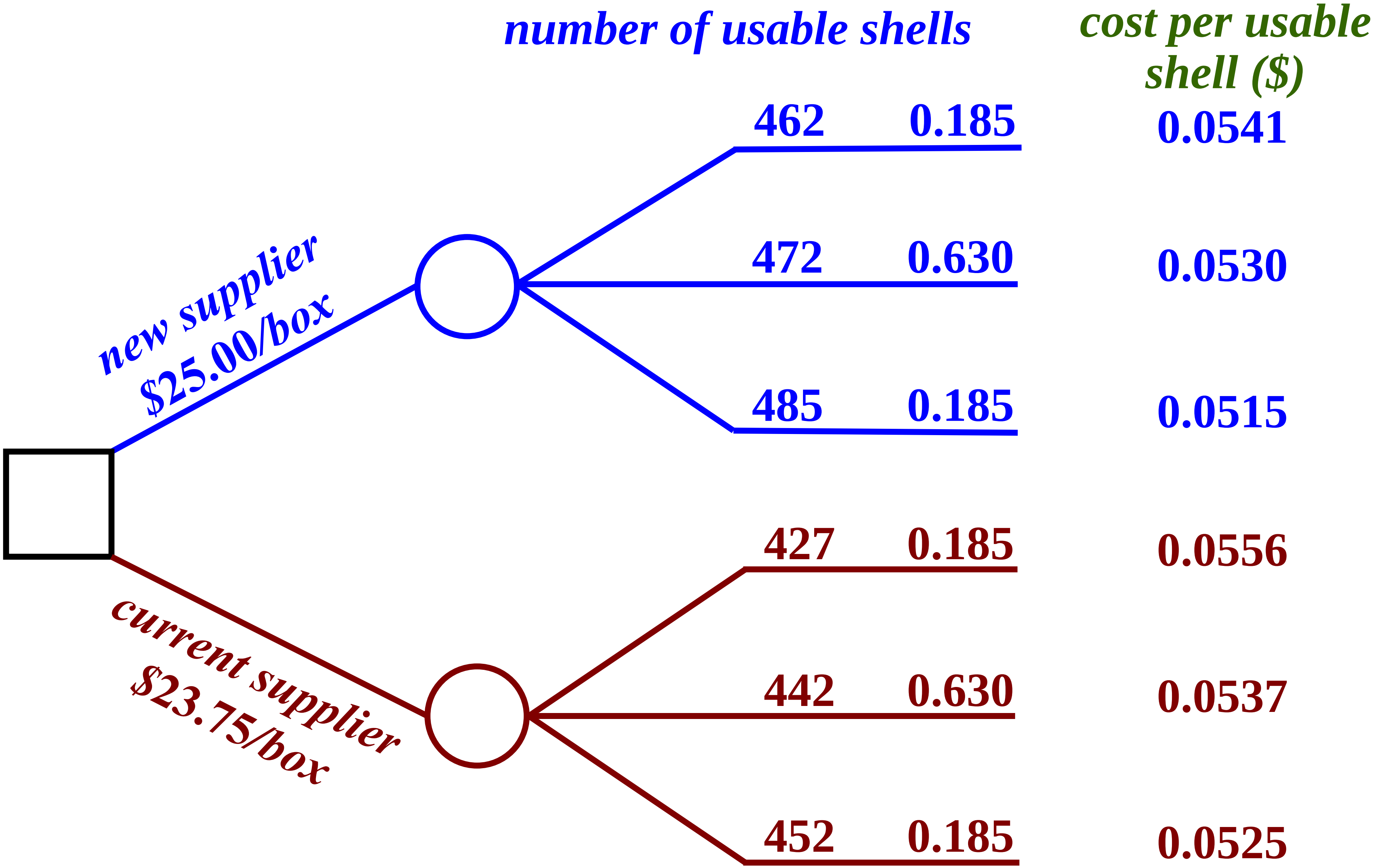
- **Clearly, the new supplier has the higher expected number of useable shells per box; the two distributions, however, are highly similar**
- **The mean number of useable shells for the new supplier is 473 and so the expected costs per**

*c.d.f.*s OF THE TWO SUPPLIERS

useable shell is $0.0529; the minimum (maximum) number of useable shells is 463(482)

- The mean number of useable shells for the current supplier is 441 and so the expected costs per useable shell is $0.0539; the minimum (maximum) number of useable shells is 429(452)

REDUCED ORDER REPRESENTATION OF THE TEST RUN DATA



COMMENTS

- We use the *c.d.f.*s to estimate the means of the two populations of suppliers

- In general for an arbitrary *r.v.* $\underset{\sim}{X}$

$$E\left\{\frac{1}{\underset{\sim}{X}}\right\} \neq \left[E\left\{\underset{\sim}{X}\right\}\right]^{-1}$$

COMMENTS

and so we cannot use the approximation

$$E\left\{\frac{25}{\underset{\sim}{X}}\right\} \approx \frac{25}{E\left\{\underset{\sim}{X}\right\}}$$

- **This example demonstrates the usefulness of the** ***c.d.f.*****s in applications even when they can only be approximated for the limited data available**