Module 3

This is a single, concatenated file, suitable for printing or saving as a PDF for offline viewing. Please note that some animations or images may not work.

Module 3 Study Guide and Deliverables

Topics:	**Lecture 5:** Analyzing Risk: Modeling Input Data **Lecture 6:** Analyzing Risk: Dealing with Correlated Data
Readings:	Lectures 5 and 6 online content
Discussions:	No discussion.
Assignments:	**Tutorial:** 3 **Assignment 3**: Individual Assignment covering Lecture 1 and Lecture 6.
Assessments:	**Quiz 3**

Lecture 5

Learning Objectives

After you complete this lecture, you will be familiar with the following:

- The newsvendor problem
- Developing input models to be used in simulation
- Using resampling method to represent an uncertain input

- Fitting a probability distribution to the data on hand
- Chi-square test, Kolmogorov-Smirnov test
- Q-Q plot, P-P plot
- Using historical data itself in simulation
- Breakpoints method
- Mean and variability method

Newsvendor Problem[1]

"Banana Republic, a division of Gap, Inc., was trying to build a name for itself in fashion circles. In one recent holiday season, the company has forecasted that the blue would be the top-selling color in wool sweaters. They were wrong. The number one seller was green and they did not have enough of the green sweaters."[2]

This example describes a practical situation in which a one-time purchase decision has to be made in the face of uncertain demand.

This problem is commonly referred to as a *newsvendor problem* or *newsboy problem*: A newsboy (someone selling newspaper on the street) has to decide **how many newspapers to purchase** on a daily basis before observing the demand. Since the demand is uncertain, the newsboy runs the risk of over-purchasing or under-purchasing. More specifically, if he purchases too few newspapers, he will risk not satisfying the demand which will result in lost opportunity to increase profits as well as loss of goodwill. If he purchases too many newspapers, he will risk not being able to sell them all, which will result in loss because he has to discard the leftovers at the end of the day (today's newspaper is useless tomorrow).

In this module, we will be working with the following single-period purchasing decision problem:

> Suppose that a small retail store anticipates that on a Valentine's day, at least 40 wool socks (produced for Valentine's day) will be sold, but the actual amount is uncertain. A pair of socks costs $12 and sells for $18. After the Valentine's day, the unsold socks will be discounted 50% and are eventually sold.
>
> The question is how many socks should this store purchase in the face of uncertain demand?
>
> **Source:** *Adapted from Evans, J. R. (2016). Business analytics: Methods, models, and decisions (2nd edition). Pearson Education, Inc., p. 353*

In this module, our goal is *not* to answer this question (the question will be answered in Lecture 7). Rather, **our goal is to model the demand** in several ways. So far, in our examples, we assumed that the demand (more generally, the uncertain input variable) comes from a specific distribution such as the normal distribution. However, this need not be the case (and in some cases, it is not appropriate to use a probability distribution to model the uncertain variable(s)). In this module, we will look at the problem of modeling uncertain input variables in more detail.

Individual Exercise:

Before you read further, please try to build the spreadsheet model on your own. Assume a demand of 41 and purchase quantity of 44.

Figure 5.1 presents the R Code we use to model this problem, assuming a demand of 41 for now. We have made some assumptions just to be able to build our spreadsheet model for the newsboy problem. In particular, we have assumed a certain quantity of 41 for the demand. However, as we all know, in real life demand is uncertain. In the remainder of this module, we will be focusing on modeling the uncertain demand in several different ways.

Figure 5.1

```
> UnitPrice<-18
> UnitCost<-12
> UnitDisc<-9
> QtyPurchase<-44
> Demand<-41
> Profit<-UnitPrice*min(QtyPurchase,Demand)+
+   UnitDisc*(QtyPurchase-min(QtyPurchase,Demand))-
+   UnitCost*QtyPurchase
> Profit
[1] 237
```

Input Model Development

Input modeling is picking a model to represent the uncertainty or randomness in a stochastic simulation. Examples include

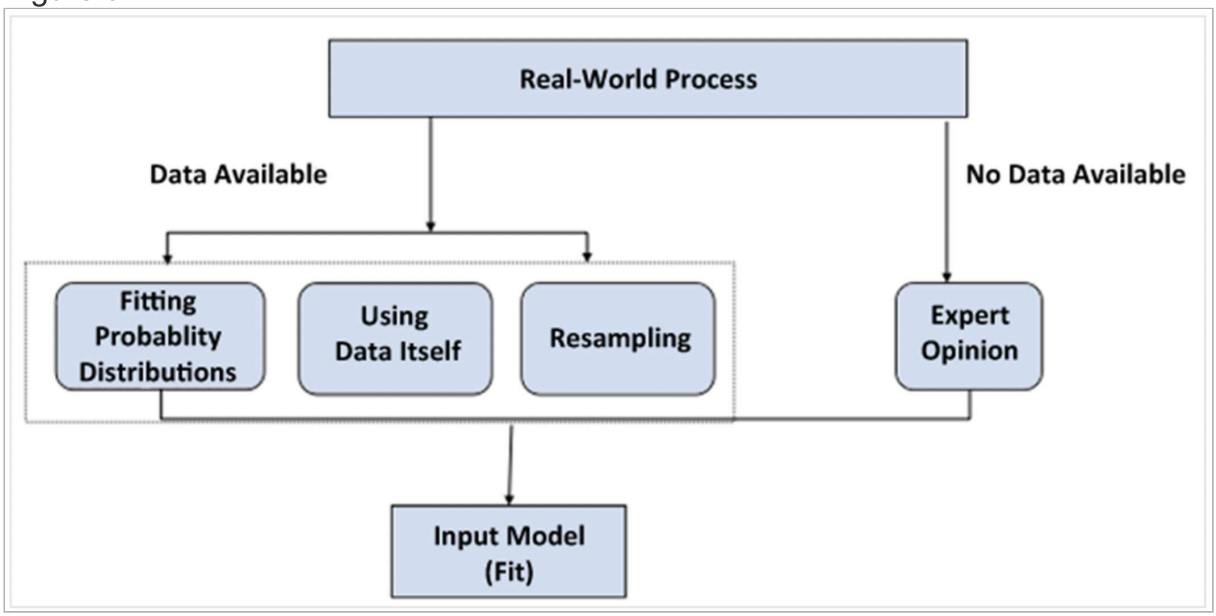
- demand per unit time for inventory of a product
- number of defective items in a shipment of goods
- times between arrivals of calls to a call center

So far, we assumed that we are given these probability distributions as models; e.g., we assumed demand is normally distributed with a given mean and variance. In this module, we will learn how we choose these models; i.e., we will cover the topic of input modeling.

It is important to understand that input modeling is an art and that the picked input model will be only an approximation of the true input model.

Figure 5.2 presents the process of developing an input model for an uncertain variable.

Figure 5.2



When data is available:

- If you think there is **a physical basis for the sample data** and you can use an analytical probability distribution to represent this physical phenomenon, then use the approach *"fitting probability distributions*". For instance, if the data represent the project completion times, then you can use the PERT distribution to model the data, as we covered in Lecture 3. Therefore, it is important we know the physical basis of the distributions, i.e., which distribution is appropriate to represent a particular uncertainty. We covered this also in Lecture 3.
- If you have **large number of observations** compared to the simulation replications you will perform, you can use the "*historical data itself"* approach.
- If you have **small number of observations** compared to the simulation replications you will perform, you can use the "*resampling"* approach.

When data is not available:

- If available, use expert opinion to model the data.

- If expert opinion is not available, use the literature for your industry and try to find example applications similar to yours.

Input Model Development when Data is Available

Now let us assume that our small retail store has kept records for the past 20 years on the number of wool socks sold at full price during the Valentine's Day. The data is shown in the spreadsheet in cells A2:A21 in Figure 5.3.

Figure 5.3

	A
1	Historical wool socks sales
2	42
3	45
4	40
5	46
6	43
7	43
8	46
9	42
10	44
11	43
12	47
13	41
14	41
15	45
16	51
17	43
18	45
19	42
20	44
21	48

Using Historical Data Itself to Model the Sales Data

This approach is useful especially when one has large amounts of data compared to the simulation trials to be performed. Although we have only 20 data points in our example, for the sake of completeness, we will still

illustrate this approach with our newsvendor problem.

We first read our data into a data frame in R. In this case, our data is contained in a .csv file, so we use the read.csv function. We then perform our profit calculation for each level of historical demand.

The R Code for this approach and the first 5 rows of data are reproduced in Figure 5.4.

Figure 5.4

```
library(tidyverse)

UnitPrice<-18
UnitCost<-12
UnitDisc<-9
QtyPurchase<-44

L5Data <- read.csv("~/L5Data.csv")
HSales<-L5Data[,1]
rm(L5Data)

trials<-data.frame(QtyDemand=HSales)
trials<-trials %>%
  mutate(QtySoldFull=pmin(QtyDemand,QtyPurchase),
         QtySoldDisc=QtyPurchase-QtySoldFull,
         Profit=UnitPrice*QtySoldFull+UnitDisc*QtySoldDisc-UnitCost*QtyPurchase)
```

```
> head(trials,5)
  QtyDemand QtySoldFull QtySoldDisc Profit
1        42          42           2    246
2        45          44           0    264
3        40          40           4    228
4        46          44           0    264
5        43          43           1    255
```

Using Resampling to Model the Sales Data

To model the demand, we can resample from the historical sales data. More specifically, we select a value randomly from the historical sales data to represent the demand in the model. In other words, instead of randomly sampling a value from a probability distribution for each trial of the simulation (as we have done so far in this course), we will sample a value for demand from the given historical sales data for each trial of the simulation. We can do this using the rdunif function from Lecture 3 to choose a *random* element of our list for each trial. The R code and resulting rows of data are reproduced in Figure 5.5.

Figure 5.5

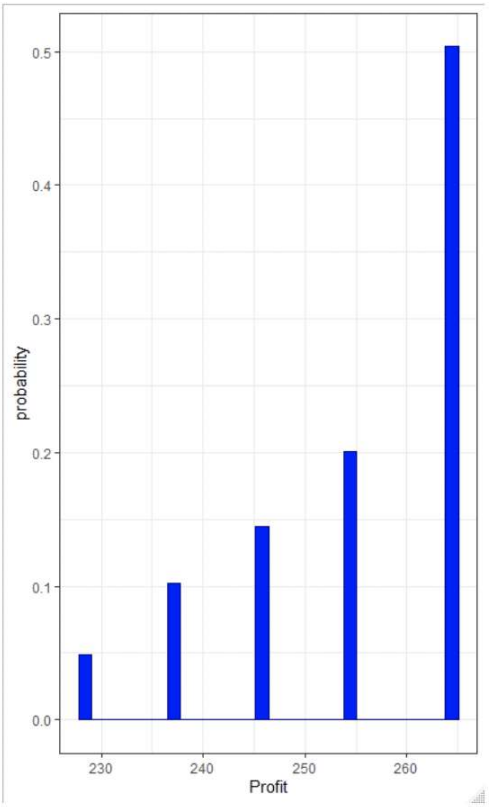
```
library(tidyverse)
n<-10000
UnitPrice<-18
UnitCost<-12
UnitDisc<-9
QtyPurchase<-44

L5Data <- read.csv("~/L5Data.csv")
HSales<-L5Data[,1]
rm(L5Data)

rdunif<- function(n,min=1,max=100){
  floor(runif(n,min,max+1))
}

trials<-data.frame(trial=1:n,
                   QtyDemand=HSales[rdunif(n,min=1,max=length(HSales))])

trials<-trials %>%
  mutate(QtySoldFull=pmin(QtyDemand,QtyPurchase),
         QtySoldDisc=QtyPurchase-QtySoldFull,
         Profit=UnitPrice*QtySoldFull+UnitDisc*QtySoldDisc-UnitCost*QtyPurchase)
```

```
> head(trials,5)
  trial QtyDemand QtySoldFull QtySoldDisc Profit
1     1        45          44           0    264
2     2        44          44           0    264
3     3        51          44           0    264
4     4        43          43           1    255
5     5        41          41           3    237
>
```

As before, we can produce a histogram and answer questions about the resulting profit distribution, seen in Figure 5.6.

Figure 5.6

```
> ggplot(trials)+geom_histogram(aes(y=..count../sum(..count..),x=Profit),color="dark blue",
fill="blue")+
+   ylab("probability")+theme_bw()
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
>
> cat("The mean is",mean(trials$Profit),"\nThe standard deviation is",sd(trials$Profit))
The mean is 255.0729
The standard deviation is 11.01309
```



Despite the ease in its use, resampling has some shortcomings:

- Useful when you have a small number of historical observations compared to the number of replications you want to run.
- You can never go outside your observed data. This may limit your ability to predict uncertain events that may occur because you are limited with the range of the observed data; you cannot include likely values outside the range of the data.
- It is difficult to reflect dependencies in the inputs.

Nonetheless, resampling is used frequently in practice.

Fitting a Probability Distribution to the Sales Data

The approach of fitting probability distributions to the demand data proceeds as follows:

Step 1: Find a probability distribution that will represent the data and fit this distribution to the data (i.e., determine values for its unknown parameters). One way of doing this is to use the physical characteristics of the process as illustrated in Lecture 3.

Step 2: Check the fit to the data via tests and graphical analysis.

Step 3: If the distribution does not fit, select another candidate and repeat the process, or use an empirical distribution (e.g., resampling or a step-wise function).

Now, let us elaborate on these steps (these steps are difficult to perform by hand but are automatable in R. Please review them and try to grasp the big picture. You do not need to understand the specifics for the purpose of this course, nor do you need to exactly follow the R code for this section):

Step 1: In this step, our goal is to find a probability distribution along with its parameters to represent the data on hand. A good way to start this step is to use the physical characteristics of the process. For instance, if you are modeling the project completion times, then you know that PERT is a good distribution to represent the project completion times.

After we choose a distribution to represent the data on hand, the next step is to estimate the parameters of the distribution. For instance, let us assume that we have chosen the normal distribution to represent the data. The next question is "what values should the parameters of the normal distribution take?"

The estimation of the parameters is usually done by a method called the "maximum likelihood estimation method." The resulting parameter estimates are called "maximum likelihood estimators." There are different maximum likelihood estimators for the parameters of different distributions.

Maximum likelihood estimation uses the data provided, but treats the parameters of the distribution as random variables, taking the mode of the resulting distribution.

For the normal distribution, for instance, the maximum likelihood estimator for the mean and variance are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

where x_i and n represent your data and the number of data points available. A Google search will bring out the maximum likelihood estimators of the parameters of several other distributions.

Step 2: After we find a probability distribution to represent the data, the next question we ask is "Is the distribution we identified really a good fit to the data?" We answer this question with the help of the goodness-of-fit tests and graphical comparisons:

- Goodness-of-fit tests
  - Chi-squared test
  - Kolmogorov-Smirnov (KS) test
  - Anderson-Darling (AD) test
  - AIC/BIC test
- Graphical comparisons
  - Histogram-based plots
  - Probability plots
  - P-P plot
  - Q-Q plot

Now, let us review the goodness-of-fit tests and the graphical methods briefly.

Goodness-of-fit tests: The Chi-squared test provides a formal comparison of a histogram (for continuous data) or line graph (for discrete data) with the fitted density (for continuous data) or mass function (for discrete data).

The KS and AD tests, on the other hand, compare an empirical distribution function with the distribution function of the hypothesized distribution. The AD test detects discrepancies in the tails and has higher power than the KS test.

AIC and BIC use a likelihood function to find the distribution that best fits the data on hand.

The KS, AD and AIC/BIC tests are only useful in the presence of data drawn from a continuous source. For discrete data we have to use the Chi-squared test.

> **Note:** In comparing several candidate distributions to represent the data, the smaller the KS test statistic/AD test statistic/Chi-square test statistics/AIC-BIC values, the better the fit.

Graphical comparisons: Histogram-based plots compare (graphically) a histogram of the data with the density function of the fitted distribution. They are sensitive to how we group the data when forming the histogram. Probability plots, on the other hand, graphically compare an estimate of the true distribution function of the data with the distribution function of the fit.

Q-Q (P-P) plot amplifies differences between the tails (middle) of the model and sample distribution functions. If the distribution chosen is a good fit, the Q-Q and the P-P plot will be approximately linear (i.e., will lie on the $y=x$ line).

Note:

Use every graphical tool to examine the fit.

If you're using a histogram-based tool, then play with the widths of the cells.

Checking the Q-Q plot is very highly recommended!

When different plots provide different insights, it is generally the Q-Q plot that we will rely on.

Source: Adapted from Biller B., & Nelson, B. L. (2002). *Answers to the top ten input modeling questions. Proceedings of the 2002 Winter Simulation Conference* (E. Yucesan, C. H. Chen, J. L. Snowdon, & J. M. Charnes, Eds.), p. 38.

Step 3: After careful analysis in Step 2, if you decide that the distribution that you have chosen in Step 1 is not a good fit to the data on hand, you need to go back to Step 1 and choose another distribution and perform Step 2 and Step 3 again. You continue this process until you are satisfied the distribution you chose in Step 1 represents the data on hand.

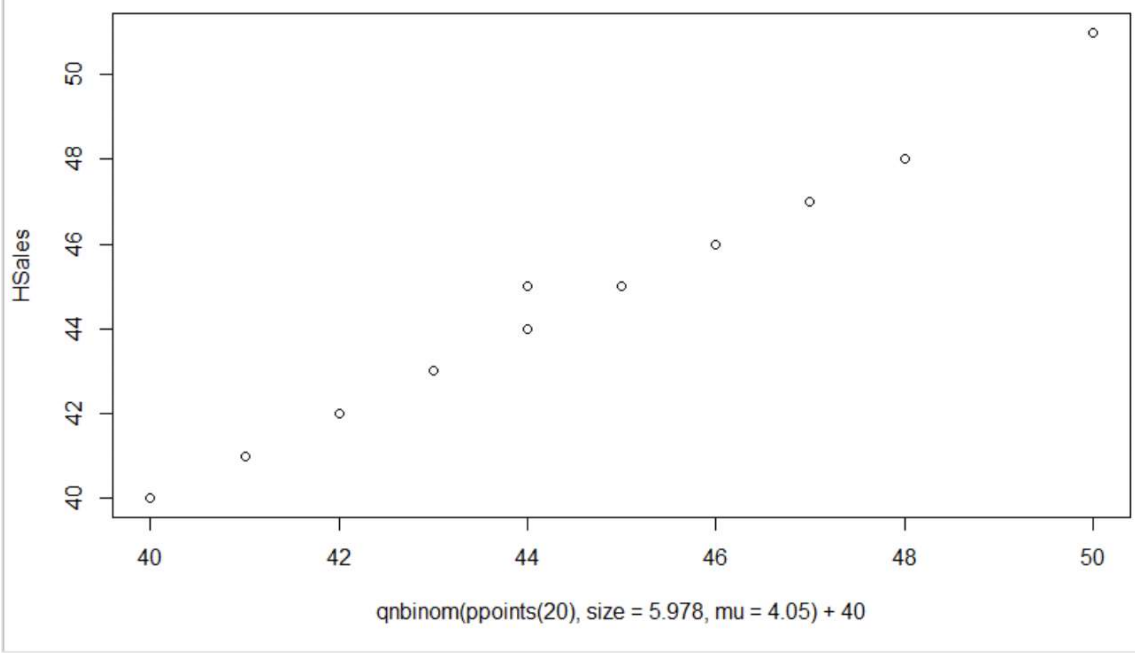
It is important to note that these three steps are very time-consuming and are not easy to perform without the help of a software program. Fortunately, R can automate this process with several distributions.

Fitting with R

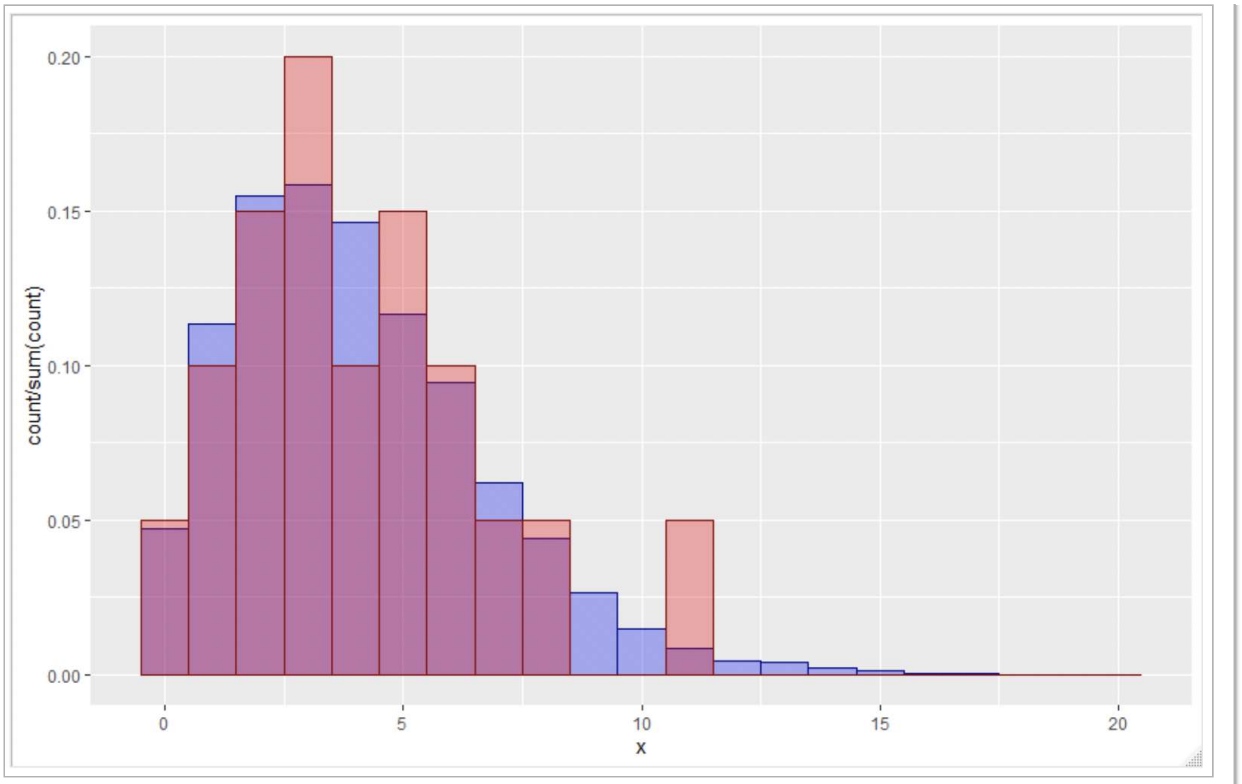
In order to fit a distribution with R, we combine the fitdistr function from the MASS library with the test statistic of our choice. Here I use the goodfit function from the vcd library to find the best fit for sales, comparing the three available distributions, shifted several times. In the next lecture, we will use the fitdistrplus library to perform both functions. The fitdistr function finds the MLE of parameters for the best fitting result. We find the best fit we can get is the negative binomial distribution, shifted to the right by 40.

The code, along with histograms comparing the two distributions, and a qqplot, are reproduced in Figure 5.7.

Figure 5.7



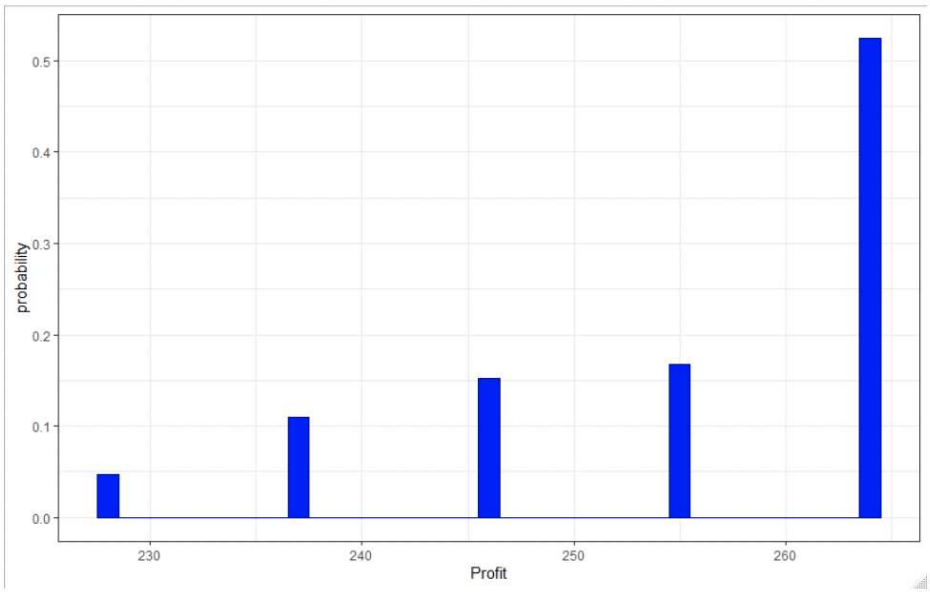

```
library(MASS)
library(vcd)

p<-lapply(31:40,function (X) goodfit(HSales-X,"poisson") %>%summary())
b<-lapply(31:40,function (X) goodfit(HSales-X,"binomial") %>%summary())
n<-lapply(31:40,function (X) goodfit(HSales-X,"nbinomial") %>%summary())

fitdistr(HSales-40,"negative binomial")

x<-rnbinom(n=10000,size=5.978,mu=4.05)

ggplot()+geom_histogram(aes(x=x,y=..count../sum(..count..)),color="dark blue",fill="blue",binwidth=1,alpha=.3)+
  geom_histogram(aes(x=HSales-40,y=..count../sum(..count..)),color="dark red",fill="red",binwidth=1,alpha=.3)

qqplot(qnbinom(ppoints(20),size=5.978,mu=4.05)+40,HSales)
```



If we're satisfied with the resulting fit, we can replace our historical data with the new distribution in a data frame. The code and resulting histogram are displayed in Figure 5.8.

Figure 5.8

```
n<-10000
UnitPrice<-18
UnitCost<-12
UnitDisc<-9
QtyPurchase<-44


trials<-data.frame(trial=1:n,
                   QtyDemand=rnbinom(n,size=5.978,mu=4.05)+40)

trials<-trials %>%
  mutate(QtySoldFull=pmin(QtyDemand,QtyPurchase),
         QtySoldDisc=QtyPurchase-QtySoldFull,
         Profit=UnitPrice*QtySoldFull+UnitDisc*QtySoldDisc-UnitCost*QtyPurchase)

ggplot(trials)+geom_histogram(aes(y=..count../sum(..count..),x=Profit),color="dark blue",fill="blue",binwidth=1)+
  ylab("probability")+theme_bw()

cat("The mean is",mean(trials$Profit),"\nThe standard deviation is",sd(trials$Profit))
```



Input Modeling when Data is Not Available

If we have no data to model an uncertain variable in our model, then we have to use anything we can find including:

- Expert opinion
- Physical or conventional limitations (can provide bounds)
- Physical basis of the process (can suggest appropriate distribution families)

Source: Adapted from Biller, B., & Gunes, C. (2010). *Introduction to simulation input modeling. Proceedings of the 2010 Winter Simulation Conference* (B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, & E. Yucesan, Eds.), section "When No Data are Available," p. 52.

In this section, we will learn about two general methods used to model the expert opinion:

- Breakpoints method
- Mean and variability method

Breakpoints Method

Assuming that we are trying to model the demand of a product and we are given by the expert that demand is likely to be no less than 100 units but no more than 500 units, we can use uniform distribution with parameters 100 and 500 to model this demand process.

Source: Biller, B., & Gunes, C. (2010). *Introduction to simulation input modeling. Proceedings of the 2010 Winter Simulation Conference* (B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, & E. Yucesan, Eds.), section "When No Data are Available," pp. 52-53.

Why the uniform distribution? Because we are only given bounds 100 and 500 for the demand and all values in between 100 and 500 are equally likely.

If we are additionally given the information that demand is most likely to be 350 units, then we can use the triangular distribution with parameters 100, 350, and 500 to model this demand process.

The ideal information that can be gained from the expert is the breakpoints of the process. Breakpoints are values and each value is associated with a percentage of being less than itself.

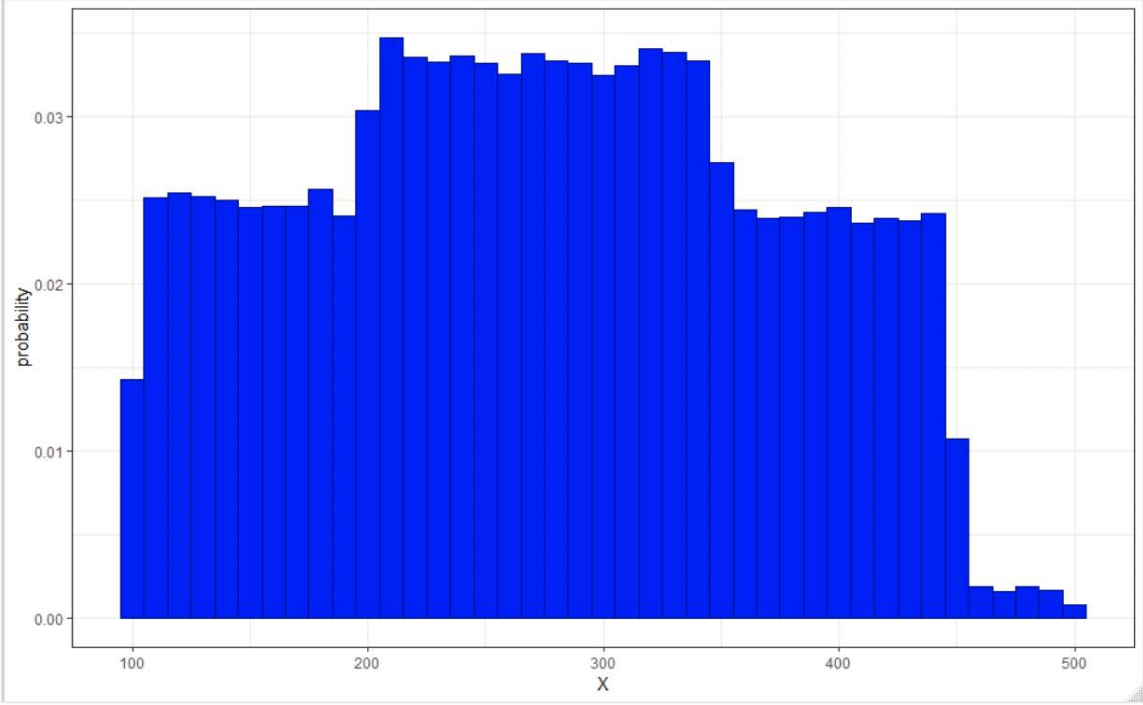
Source: Biller, B., & Gunes, C. (2010). *Introduction to simulation input modeling. Proceedings of the 2010 Winter Simulation Conference* (B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, & E. Yucesan, Eds.), section "When No Data are Available," pp. 52-53.

Now assume that we are given by the expert the following information: The demand of the product will be between 100 and 500 units with a 25% chance of being less than 200, a 75% chance of being less than 350, and a 99% chance of being less than 450.

For this we can combine a runif call with an rdunif call. The first will determine which of the three demand results occur. The code to do so is reproduced in Figure 5.9, along with a histogram giving an idea of the pmf.

Figure 5.9

```
y<-runif(100000)

rdunif<- function(n,min=1,max=100){
  floor(runif(n,min,max+1))
}

X<-sapply(X=y,FUN= function(t)
            ifelse(t<.25,rdunif(1,100,199),
                   ifelse(t<.75,rdunif(1,200,349),
                          ifelse(t<.99,rdunif(1,350,449),rdunif(1,450,500)
                                 )
                          )
                   )
          )
ggplot()+geom_histogram(aes(x=X,y=..count../sum(..count..)),
                        fill="blue",color="dark blue",binwidth=10)+
  ylab("probability")+theme_bw()
```



Lecture 5 Footnotes

[1] Adapted from Evans, J. R. (2016). *Business analytics: Methods, models, and decisions* (2nd edition). Pearson Education, Inc., p. 353.

[2] Lee, L. (2004, May 31). Yes, we have a new banana. *Business Week,* 70–72.

Lecture 5 References

Evans, J. R. (2016). *Business analytics: Methods, models, and decisions* (2nd edition). Pearson Education, Inc., p. 353.

Lee, L. (2004, May 31). Yes, we have a new banana. *Business Week,* 70–72.

Lecture 5 Summary Questions

1. What is "input modeling"?
2. Please comment on the following statement: "*There is no true input model for any uncertain input."*
3. What input modeling methods can we use in the presence of data?
4. What input modeling methods can we use in the absence of data?
5. What are the disadvantages of resampling?
6. Assume that we are given by the expert the following information: The demand of the product will be between 250 and 750 units with a 25% chance of being less than 300, a 75% chance of being less than 550, and a 99% chance of being less than 600. How would you model this situation?