Module 6: Supplementary Slides

Additional R Coding

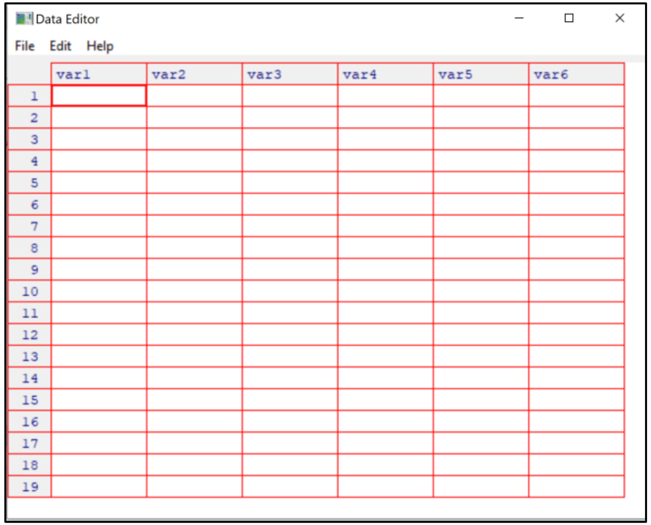
Input and Output in R

Entering Data from the Keyboard

For very small datasets, use the c() for vectors.

You can also create an empty dataset (data frame) and then use "edit" to fill it

```
> scores <- c(61, 66, 90, 88, 100)

> scores <- data.frame() # Create empty data frame
> scores <- edit(scores) # edit the data frame
```



Input and Output in R

Creating a Dataset on Keyboard

```
nameofdataframe = data.frame(
  var1=c(…),
  var2=c(…),
  var3=c(…)
  )
```

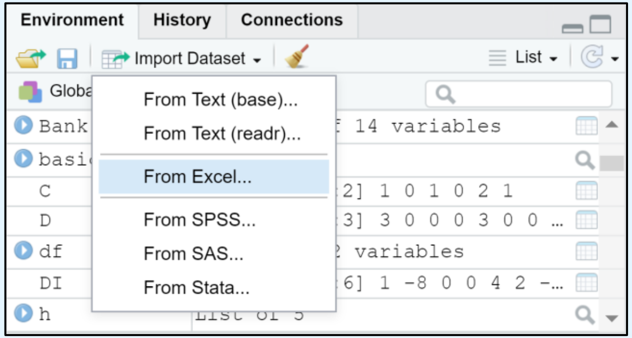
```
scores = data.frame(
  label=c("Low", "Mid", "High"),
  lbound=c( 0, 0.67, 1.64),
  ubound=c(0.674, 1.64, 2.33)
  )

> scores
  label   lbound      ubound
1 Low      0.00       0.674
2 Mid      0.67       1.640
3 High     1.64       2.330
```

Input and Output in R

Importing a Dataset From Excel

```
library(readxl)
Data_name = read_excel(“enter the address of the excel file", + sheet = “enter the sheet name")
```

An easier way: Import Dataset → From Excel




```
> library(readxl)
> Bank <- read_excel(“address…", + sheet = “mysheet")
> View(Bank)
```

(Use the posted UniversalBank.xlsx to practice these codes)

Input and Output in R

Exporting a Dataset (Data Frame) in Excel

```
library(writexl)

write_xlsx(list(Nameofsheet1 = dataforsheet1,
                Nameofsheet2 = dataforsheet2,
                Nameofsheet3 = dataforsheet3),
            "filename.xlsx")
```

```
> library(writexl)
> write_xlsx(list(mysheet=df, mysheet2=df),"C:/Users/Vortex/Dropbox/Math course/Bank5.xlsx")
```

Some Other Statistical Tools

Choose a Random Sample

```
sample(x, size, replace = FALSE)

x: the vector
size: size of the sample
relace: False/True whether the sample is with replacement or not
```

In the UniversalBank.xlsx case, after importing the dataset as a new data frame Bank, we can code like this:

```
> sample(Bank$ID,10)
[1] 1981 3717 2055 1622 868 2893 4634 1231 1202 385
```

Some Other Statistical Tools

Generating a Random Sequence

```
sample(x, size, replace = TRUE,Prob)

x: the vector
size: size of the sample
relace: False/True whether the sample is with replacement or not
Prob: a vector of probability weights for obtaining the elements of the vector being sampled.
```

```
> sample(c("H","T"),10,replace=TRUE,c(0.5,0.5))
[1] "T" "T" "T" "H" "H" "T" "H" "T" "T" "H"
```

Some Other Statistical Tools

Tabulating and Creating Contingency Tables

table(x) create a summary table based on different categories in x

Table(x,y) creates a contingency table

x, y: categorical variables

```
> table(Bank$Family)
   1       2           3          4
1472    1296        1010       1222


> table(Bank$Family,Bank$Education)
     1   2   3
1  678 326 468
2  657 265 374
3  349 383 278
4  412 429 381
```

Some Other Statistical Tools

Converting Data to Z-Scores

```
scale(x)

x: a vector
```

```
> v=scale(Bank$Income)
> v[1:5]
[1] -0.5381750 -0.8640230 -1.3636566 0.5697084 -0.6250678
```

Random Variables

Two Important Properties

Cauchy-Schwartz Inequality: If X and Y are two random variables, then

$$\left(E(XY)\right)^2 \leq E(X^2)E(Y^2)$$

Proof: For any $z \in \mathbb{R}$, set $Z = zX - Y$. Then, we have

$$0 \leq E\left(Z^2\right) = E\left(z^2X^2 - 2zXY + Y^2\right) = z^2E\left(X^2\right) - 2zE(XY) + E(Y^2), \text{ for all } z \in \mathbb{R}$$

The quadratic term $az^2 + bz + c$ is non-negative if and only if $b^2 - 4ac \leq 0$ and $a > 0$. Therefore,

$$4\left(E(XY)\right)^2 - 4E\left(X^2\right)E\left(Y^2\right) \leq 0 \Rightarrow \left(E(XY)\right)^2 \leq E(X^2)E(Y^2)$$

Jensen's Inequality: Let X be a random variable and $f: \mathbb{R} \to \mathbb{R}$ be a convex function. Then

$$E\left(f(X)\right) \geq f(E(X))$$

Proof: Since f is convex, the tangent line at point $(\mu, f(\mu))$ lies below the graph, i.e., $f(x) \geq f(\mu) + f'(\mu)(x - \mu)$ for x. Therefore, when x is a random variable it also holds, $f(X) \geq f(\mu) + f'(\mu)(X - \mu)$. Ta Taking the expectation from both sides gives the result.

Example

Expectation of a Random Variable

In an on-campus housing lottery with the participation of N students, a bowl contains the names of the N students on sealed envelops. Students, take turns in random order and then each randomly picks one of the envelopes and reads the name written inside (the opened envelops are not put back in the bowl). If the name written inside the envelope matches the student's name s/he will win an on-campus housing. On average how many students will win an on-campus housing?

Solution

Let $X = X_1 + X_2 + \cdots + X_N$ be the number of people who pick their hats correctly, where $X_i, i = 1, \dots, N$ is

$$P(X_i = x_i) = \begin{cases} \dfrac{1}{N} \quad if\ x_i = 1 & \text{(if the ith person picks his/her hat correctly)} \\ 1 - \dfrac{1}{N}\ \ if\ x_i = 0 & \text{(if the ith person does not pick his/her hat correctly)} \end{cases}$$

Then:

$$E(X_i) = \frac{1}{N}(1) + \left(1 - \frac{1}{N}\right)(0) = \frac{1}{N}$$

$$E(X) = E(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_\mathrm{N}) = \sum_{i=1}^{N} E(X_i) = N\left(\frac{1}{N}\right) = 1$$

No matter the size of the party, we expect only one person to take his/her hat correctly.

Example

Mean and Variance of Independent Random Variables

Let $X_i, i = 1,2, \dots, n$ be *identically independently distributed* (i.i.d) random variables from a distribution with mean μ and standard deviation σ. Find the mean and variance of the random variable $\bar{X}$ defined as below ($\bar{X}$ is said to be the sample mean.)

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

Solution:

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

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

Discrete Distributions: Bernoulli Distribution

Bernoulli Distribution

Suppose that a trial, or experiment, which results in a “success” with probability p and in a failure with probability $1 - p$, is performed. If $X = 1$ when the outcome is a success and $X = 1$ if it is a failure then X is said to be a *Bernoulli* random variable with the probability mass function given by:

$$p(x) = P\{X = x\} = p^x(1 - p)^{1-x}, \qquad x = 0,1$$

Example

Expectation of a Binomial Random Variable Using Bernoulli

A trial is run for n times independently. Each trial will have two outcomes: success with probability p and failure with probability $1-p$. Suppose $X_i = 1$ if the i^{th} trial is a success, and it is zero otherwise (X_i is called a Bernoulli random variable). Also suppose $X = X_1 + X_2 + \cdots + X_n$ is a random variable representing the number of successes in n trials (X_i is a Binomial random variable). Find $\mathrm{E}(X)$.

Solution

It is not difficult to see that for each random variable X_i defined as:

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

the expected value of X_i is:

$$\mathrm{E}(X_i) = p,$$

Hence,

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) = np$$

Example

Variance of a Binomial Random Variable Using Bernoulli

A trial is run for n times independently. Each trial will have two outcomes: success with probability p and failure with probability $1-p$. Suppose $X_i = 1$ if the i^{th} trial is a success, and it is zero otherwise (X_i is called a Bernoulli random variable). Also suppose $X = X_1 + X_2 + \cdots + X_n$ is a random variable representing the number of successes in n trials (X_i is a Binomial random variable). Find $\mathrm{var}(X)$.

Solution

It is not difficult to see that for each random variable X_i defined as:

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

the expectation and variance are:

$$E(X_i) = p, \qquad E\left(X_i^2\right) = p, \qquad var(X_i) = E\left(X_i^2\right) - E(X_i)^2 = p - p^2 = p(1-p)$$

Hence,

$$var(X) = var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i) = np(1-p)$$

Discrete Distributions:
Binomial: Additional Details

Binomial Distribution

Suppose that n independent trials, each of which results in a "success" with probability p and in a failure with probability $1-p$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a *binomial* random variable with parameters (n, p).

The probability mass function of a binomial random variables with parameters (n, p) is given by:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0,1,2, \dots, n$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Note that the probabilities sum to one, that is,

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \big(p + (1-p)\big)^n = 1$$

Binomial Mean and Variance

If X is a binomial random variable with parameters (n, p) then

$$\mu = \mathrm{E}(X) = np$$

$$\sigma^2 = var(X) = np(1-p)$$

Proof:

$$\mathrm{E}(X) = \sum_{x=1}^{n} xp(x) = \sum_{x=1}^{n} x\binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1}(1-p)^{n-x}$$

$$= np\sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k} = np\big(p + (1-p)\big)^{n-1} = np$$

Binomial Mean and Variance

$$\mathrm{E}\left(X^2\right) = \sum_{x=1}^{n} x^2 p(x) = \sum_{x=1}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = np \sum_{x=1}^{n} \frac{(x-1+1)(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$$

$$= np \sum_{x=2}^{n} \frac{(n-1)!}{(x-2)!(n-x)!} p^{x-1} (1-p)^{n-x} + np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$$

$$= n(n-1)p^2 \sum_{l=0}^{n-2} \frac{(n-2)!}{l!(n-2-l)!} p^l (1-p)^{n-2-l} + np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k}$$

$$= n(n-1)p^2 + np = n^2p^2 - np^2 + np$$

Hence,

$$var(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$$

Discrete Distributions:
Poisson Distribution: Additional Details

Poisson Distribution

A random variable X taking one of the values $x = 0,1,2, \dots$ is said to be a *Poisson* random variable with parameter $\mu = \lambda t$, if for some $\mu > 0$,

$$p(x) = \frac{e^{-\mu}\mu^x}{x!}, \qquad x = 0,1,2, \dots$$

μ is often the expected number of events as in the Bank's example. In that case, $\mu = \lambda t$ where λ is the rate of occurrence of the events and t is the length of time or area of a surface or volume, etc.

Note that the probabilities sum to one, that is,

$$\sum_{x=0}^{\infty} p(x) = \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^x}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!} = e^{-\mu}e^{\mu} = 1$$

Poisson Mean and Variance

If X is a Poisson random variable with parameter μ then

$$\mathrm{E}(X) = \mu$$

$$var(X) = \mu$$

Proof:

$$\mathrm{E}(X) = \sum_{x=0}^{\infty} xp(x) = \sum_{x=0}^{\infty} x\frac{e^{-\mu}\mu^x}{x!} = \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^x}{(x-1)!} = \mu e^{-\mu}\sum_{x=0}^{\infty}\frac{\mu^{x-1}}{(x-1)!} = \mu e^{-\mu}e^{\mu} = \mu$$

$$\mathrm{E}\left(X^2\right) = \sum_{x=0}^{\infty} x^2 p(x) = \sum_{x=0}^{\infty} x^2\frac{e^{-\mu}\mu^x}{x!} = \sum_{x=1}^{\infty}\frac{(x-1+1)e^{-\mu}\mu^x}{(x-1)!} = e^{-\mu}\mu^2\underbrace{\sum_{x=2}^{\infty}\frac{\mu^{x-2}}{(x-2)!}}_{e^{\mu}} + \mu e^{-\mu}\underbrace{\sum_{x=1}^{\infty}\frac{\mu^{x-1}}{(x-1)!}}_{e^{\mu}} = \mu^2 + \mu$$

Therefore:

$$\mathrm{var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 = \mu^2 + \mu - \mu^2 = \mu$$

Discrete Distributions: Geometric Distribution

Geometric Distribution

Suppose that independent trials, each having probability p of being a success, are performed until a success occurs. If we let X be the number of trials required until the first success, then X is said to be a *geometric* random variable with parameter p. Its probability mass function is given by:

$$p(x) = P\{X = x\} = (1-p)^{x-1}p, \qquad x = 1,2,\ldots$$

Note that the probabilities sum to one, that is,

$$\sum_{x=1}^{\infty} p(x) = \sum_{x=1}^{n} (1-p)^{x-1}p = \frac{p}{1-(1-p)} = 1$$

$$E(X) = \sum_{x=1}^{\infty} xp(x) = \sum_{x=1}^{n} x(1-p)^{x-1}p = p\frac{d}{dx}\left(-\sum_{x=1}^{n}(1-p)^x\right) = p\frac{d}{dx}\left(-\frac{1}{p}\right) = p\frac{1}{p^2} = \frac{1}{p}$$

$$E(X^2) = \sum_{x=1}^{\infty} x^2p(x) = \sum_{x=1}^{n} x^2(1-p)^{x-1}p = p\frac{d}{dx}\left(\sum_{x=1}^{n}\left(-x(1-p)^x + (1-p)^x\right)\right) = p\frac{d}{dx}\left(-\frac{1}{p}\right) = p\frac{1}{p^2} = \frac{1}{p}$$

Example

If a fair coin is successively flipped, what is the probability that a head first appears on the fifth trial?

Solution:

$$p(5) = P\{X = 5\} = (1-p)^{5-1}p = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

Geometric Distribution

If X is a geometric random variable with parameter p then

$$E(X) = \frac{1}{p}, \qquad var(X) = \frac{1-p}{p^2}$$

$$E(X) = \sum_{x=1}^{\infty} xp(x) = \sum_{x=1}^{n} x(1-p)^{x-1}p = p\frac{d}{dp}\left(-\sum_{x=1}^{n}(1-p)^x\right) = p\frac{d}{dp}\left(-\frac{1}{p}\right) = p\frac{1}{p^2} = \frac{1}{p}$$

$$E(X(X-1)) = p(1-p)\sum_{x=1}^{n} x(x-1)(1-p)^x = p(1-p)\frac{d^2}{dp^2}\left(\sum_{x=1}^{n}(1-p)^x\right) = p(1-p)\frac{d}{dp}\left(\frac{-1}{p^2}\right) = \frac{2(1-p)}{p^2}$$

$$\Rightarrow\ E\left(X^2\right) = E(X(X-1)) + E(X) = \frac{2(1-p)}{p^2} + \frac{1}{p} = \frac{2(1-p)+p}{p^2} = \frac{2-p}{p^2}$$

$$\Rightarrow var(X) = E\left(X^2\right) - E(X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

Example

Suppose that a powerplant electricity generation is handled by a number of independent control systems that operate in parallel to one another. Any of the control systems can fail, during the electricity generation process, with probability $1-p$ independently from other control systems. Suppose that the powerplant can generate electricity successfully during a day if at least 50 percent of its control systems remain operative. For what values of p having a powerplant with 4 parallel control systems is more preferable to a powerplant with two control system?

Solution

The probability that a four-engine plane will make a successful flight is

$$\binom{4}{2}p^2(1-p)^2 + \binom{4}{3}p^3(1-p)^1 + \binom{4}{4}p^4(1-p)^0 = 6p^2(1-p)^2 + 4p^3(1-p) + p^4$$

The probability that a two-engine plane will make a successful flight is

$$\binom{2}{1}p^1(1-p)^1 + \binom{2}{2}p^2(1-p)^0 = 2p(1-p) + p^2$$

The four-engine plane is safe if

$$6p^2(1-p)^2 + 4p^3(1-p) + p^4 \geq 2p(1-p) + p^2$$

$$\Rightarrow 6p(1-p)^2 + 4p^2(1-p) + p^3 \geq 2\,(1-p) + p \Rightarrow 6p + 6p^3 - 12p^2 + 4p^2 - 4p^3 + p^3 \geq 2\; - 2p + p$$

$$\Rightarrow 3p^3 - 8p^2 + 7p - 2 \geq 0$$

The left side has two roots $p = \frac{2}{3}, 1$. It is easy to verify that it is positive when $p \geq \frac{2}{3}$.

$f\ = 3p^3 - 8p^2 + 7p - 2$	
$0 \leq p \leq \frac{2}{3}$	$\frac{2}{3} \leq p \leq 1$
$f \leq 0$	$f \geq 0$

Correlation Property

$$-1 \leq \rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \leq 1$$

Proof.

Consider the random variable $Z = X - Y\frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}$. Then,

$$0 \leq \mathrm{var}(Z) = \mathrm{cov}(Z, Z) = \mathrm{cov}\left(X - Y\frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}, X - Y\frac{\mathrm{cov}(X,Y)}{\mathrm{var}(Y)}\right) = \mathrm{var}(x) - \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(Y)} \Rightarrow \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)} = \rho^2 \leq 1 \Rightarrow -1 \leq \rho \leq 1$$

Remark: $\mathrm{cov}(X, Y) = 0$ does *not* mean independence:

- To see why, suppose $\mathrm{P}(X = 1) = \mathrm{P}(X = -1) = 0.5$ and $Y = X^2$. Then $\mathrm{cov}(X, Y) = 0$ but X and Y are nonlinearly related.