Teaching Statistics with Real World Data from IPUMS

Answer key to basic statistical concept exercises in R, using real-world census microdata from the IPUMS International database.

Exercise: Probability Answer Key

Topics covered:

- Probability of an event
- Probability of union and intersection of two events
- Disjoint and independent events
- Conditional probability

Required dataset: IPUMS-International

Required variables:

1. COUNTRY
2. YEAR
3. RELATE (relationship to household head)
4. OWNERSHIP (ownership of housing unit)
5. INTERNET (internet access)

[The only preselected variables that are needed in this exercise are COUNTRY and YEAR. Make sure to remove all of the other preselected variables by unchecking the blue boxes next to them. This will reduce the size of your data file and also make it easier to view the data in R.]

Recommended sample:

Fiji [2014]

Sample selection instructions:

[Note that both the variables – OWNERSHIP and INTERNET are household variables. We assume that head of the household represents a household. Hence, we record the OWNERSHIP and INTERNET values for the head of a household.]

1. Limit sample to 10,000 households.
2. In the "extract request" page, which comes after you, click "create data extract", first click on "select cases", choose "RELATE" and submit. Then, select "include only those persons meeting case selection criteria" and "1 Head".

3. Next select "customize sample size" and type 10 in the box under households and Fiji 2014. *[Note that the sample size is in 1000s]*

Notes:

- Round off all probabilities to the nearest two decimal places.
- Let us define event A as the event of having ownership of housing unit and event B as the event of having internet access.

• SECTION I

1. Describe the sampling method used to collect the customized samples. [*Hint: Check help in "customize sample sizes"*]

Systematic sampling method used to collect the customized samples. Let the total number of households in the population be N and the required sample size (10,000 in this case) be n. Then, the interval in systematic sampling, say I, is defined as I = N/n. The interval used for uniform spread over the population. A number is then selected randomly between 1 and I which may be denoted by R. Then the list of sample households are R, R+I, R+2I,……, R+(n-1)I, i.e., every I^{th} household is selected starting from the R^{th} household.

2. What is the universe for the two variables – OWNERSHIP and INETRNET?

The universe for both the variables is all households.

3. What does the values INTERNET = 0 and OWNERSHIP = 0 signify? Should these observations be included or excluded in calculation of probability?

The values INTERNET = 0 and OWNERSHIP = 0 signifies observations that were not in the universe. Hence, these observations should be excluded in calculation of probability.

4. What does the values INTERNET = 9 and OWNERSHIP = 9 signify? Should these observations be included or excluded in calculation of probability?

The values INTERNET = 9 and OWNERSHIP = 9 signifies observations with unknown responses. These observations should be excluded in calculation of probability as otherwise they will distort our estimate.

SECTION II

```r
library(ipumsr)

# Load the data
ddi <- read_ipums_ddi("ipumsi_00023.xml")
data_prob1 <- read_ipums_micro(ddi)
```

```r
ipums_val_labels(data_prob1$OWNERSHIP)
```

```
## # A tibble: 4 x 2
##     val lbl
##   <dbl> <chr>
## 1     0 NIU (not in universe)
## 2     1 Owned
## 3     2 Not owned
## 4     9 Unknown
```

```r
ipums_val_labels(data_prob1$INTERNET)
```

```
## # A tibble: 4 x 2
##     val lbl
##   <dbl> <chr>
## 1     0 NIU (not in universe)
## 2     1 No
## 3     2 Yes
## 4     9 Unknown
```

1. What is the probability that a household has ownership of housing unit?

$$\begin{aligned}
&P(\mathit{household\ owns\ the\ housing\ unit}\ )\\
&= P(A)\\
&= \frac{\mathit{Number\ of\ households\ that\ owns\ housing\ unit}}{\mathit{Total\ number\ of\ households}}\\
&= \frac{7686}{9851}\\
&\approx 0.78
\end{aligned}$$

```r
# Note that we need here only the variable OWNERSHIP is in consideration. We
need to remove unknown responses and NIU observations for only the variable
OWNERSHIP.

# Convert the variable - OWNERSHIP to factor
data_prob1<-within(data_prob1, OWNERSHIP<- as.factor(OWNERSHIP))
```

```
summary(data_prob1$OWNERSHIP) # Notice that there are no observations which
are not in the universe (NIU)

##    1    2    9
## 7686 2165   45

# Remove observations for which are the value of OWNERSHIP 9 (Unknown Respons
es)
data_prob1$OWNERSHIP[data_prob1$OWNERSHIP==9]<-NA
newdata_prob1<- na.omit(data_prob1)
summary(newdata_prob1$OWNERSHIP)

##    1    2    9
## 7686 2165    0

# 1: Owned, 2: Not Owned, 3: Unknown
# Probability that a household has ownership of housing unit
round(7686/(7686+2165),2)

## [1] 0.78

# Instead of using summary and manually typing the required numbers, we can
do the following:
# Number of households who has ownership of housing unit
a<- nrow(newdata_prob1[newdata_prob1$OWNERSHIP==1,])
a

## [1] 7686

# Total number of households
n1<- nrow(newdata_prob1)
n1

## [1] 9851

# Probability that a household has ownership of housing unit
round(a/n1, 2)

## [1] 0.78
```

2. What is the probability that a household does not have ownership of housing unit?

Method 1:

$$P(\text{household does not own the housing unit})$$

$$= P(A^c)$$

$$= \frac{\text{Number of households that does not owns housing unit}}{\text{Total number of households}}$$

$$= \frac{2165}{9851}$$

$$\approx 0.22$$

```
# Method 1: P(households that do not have ownership) = number of households
who do not have ownership / total number of households
# Number of households who do not have ownership of housing unit
b<- nrow(newdata_prob1[newdata_prob1$OWNERSHIP==2,])
b

## [1] 2165

# Probability that a household does not ownership of housing unit
round(b/n1, 2)

## [1] 0.22
```

Method 2:

$$
\begin{aligned}
&P(\textit{household does not own the housing unit })\\
&= 1 - P(\textit{household owns the housing unit })\\
&= 1 - 0.78\\
&= 0.22
\end{aligned}
$$

```
# Method 2: Use the fact that P(Ownership - Yes) + P(Ownership - No) = 1
# Probability that a household does not ownership of housing unit
round(1 - (a/n1), 2)

## [1] 0.22
```

3. What is the probability that a household has an internet connection?

$$
\begin{aligned}
&P(\textit{household has internet connection })\\
&= P(B)\\
&= \frac{\textit{Number of households that has internet connection}}{\textit{Total number of households}}\\
&= \frac{2169}{9851}\\
&\approx 0.22
\end{aligned}
$$

```
# Note that we need to load the dataset again as here the variable INTERNET
is in consideration. We need to remove unknown responses and NIU observations
for only the variable INTERNET.

# Load the original dataset
data_prob2 <- read_ipums_micro(ddi)
```

```r
# Convert the variable - INTERNET to factor
data_prob2<-within(data_prob2, INTERNET<- as.factor(INTERNET))
summary(data_prob2$INTERNET) # Notice that there are no observations which are not in the universe (NIU)
##    1    2    9
## 7682 2169   45

# Remove observations for which are the value of 9 (unknown responses)
data_prob2$INTERNET[data_prob2$INTERNET==9]<-NA
newdata_prob2<- na.omit(data_prob2)
summary(newdata_prob2$INTERNET)

##    1    2    9
## 7682 2169    0

# Number of households who have internet connection
c<- nrow(newdata_prob2[newdata_prob2$INTERNET==2,])
c

## [1] 2169

# Total number of households
n2<- nrow(newdata_prob2)
n2

## [1] 9851

# Probability that a household has internet connection
round(c/n2, 2)

## [1] 0.22
```

4. What is the probability that a household does not have internet connection?

$$\begin{aligned} &P(household\ does\ not\ own\ have\ internet\ connection\ ) \\ &\quad = P(B^c) \\ &\quad = 1 - \ P(household\ \ has\ internet\ connection) \\ &\quad = \ 1 - 0.22 \\ &\quad = 0.78 \end{aligned}$$

```r
# Use the fact that P(Internet - Yes) + P(Internet - No) = 1
# Probability that a household does not have internet connection
round(1-(c/n2), 2)

## [1] 0.78
```

- **SECTION III**

1. Are the variables OWNERSHIP and INTERNET disjoint (mutually exclusive)? Why/ why not?

The variables OWNERSHIP and INTERNET are not disjoint because both events can happen at the same time. It is possible for a household to own housing unit and have internet connection.

2. Are the variables OWNERSHIP and INTERNET independent? Why/ why not?

We know that if the two variables OWNERSHIP and INTERNET are independent then, $P(household\ owns\ housing\ unit\ and\ has\ interent\ access) = P(A\ \cap B) = P(A)P(B)$. Hence, we need to check if this condition holds.

From the R code below, we observe that:

$$P(A\ \cap B) = \frac{1443}{9851} \approx 0.15$$

$$P(A) = \frac{7686}{9851} \approx 0.78$$

$$P(B) = \frac{2169}{9851} \approx 0.22$$

Since, $P(A\ \cap B) = 0.15 \neq 0.78 * 0.22\ \approx 0.17 = P(A)P(B)$ we conclude that OWNERSHIP and INTERNET are not two independent variables.

```r
# Note that we need to load the dataset again as here both the variable
INTERNET and OWNERSHIP are in consideration. We need to remove unknown
responses and NIU observations for both the INTERNET and OWNERSHIP.

# Load the original dataset
data_prob3 <- read_ipums_micro(ddi)
# Convert the variable - INTERNET and OWNERSHIP into factors
data_prob3<-within(data_prob3, INTERNET<- as.factor(INTERNET))
data_prob3<-within(data_prob3, OWNERSHIP<- as.factor(OWNERSHIP))
summary(data_prob3$OWNERSHIP)

##    1    2    9
## 7686 2165   45

summary(data_prob3$INTERNET)

##    1    2    9
## 7682 2169   45
```

```
# Notice that there are no observations which are not in the universe (NIU)
# We need to remove observations for which both the INTERNET status and
OWNERSHIP status is 9 (Unknown).
data_prob3$OWNERSHIP[data_prob3$OWNERSHIP==9]<-NA
data_prob3$INTERNET[data_prob3$INTERNET==9]<-NA
newdata_prob3<- na.omit(data_prob3)

# Total number of households
n3<- nrow(newdata_prob3)
n3

## [1] 9851

# Number of households that have ownership of housing unit and internet  conn
ection
d<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==1 & newdata_prob3$INTERNET==2
,])
d

## [1] 1433

# Probability that a household has ownership of housing unit and internet con
nection - P(Ownership "Intersect" Internet)
e<- round(d/n3, 2)
e

## [1] 0.15

# Number of households that have ownership of housing unit
f<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==1,])
f

## [1] 7686

# Probability that a household has ownership of housing unit – P(A)
h<- round(f/n3, 2)
h

## [1] 0.78

# Number of households that have internet access
i<- nrow(newdata_prob3[newdata_prob3$INTERNET==2,])
i

## [1] 2169

# Probability that a household has internet access – P(B)
j<- round(i/n3, 2)
j

## [1] 0.22

round(h*j,2)
```

```
## [1] 0.17
```

3. What is the probability that a random household from Brazil has internet access given that household owns housing unit?

$$\begin{aligned} &P(household\ has\ interent\ access\ |household\ owns\ housing\ unit\ ) \\ &= P(B|A) \\ &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{0.15}{0.78} \\ &\approx 0.19 \end{aligned}$$

- **SECTION IV**

1. What is the probability that a random household from Brazil does not own housing unit and has internet access?

$$\begin{aligned} &P(household\ does\ not\ own\ housing\ unit\ and\ has\ interent\ access) \\ &= P(A^c \cap B) \\ &= \frac{736}{9851} \\ &\approx 0.07 \end{aligned}$$

```
# Number of households that do not own housing unit and have internet    conn
ection
k<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==2 & newdata_prob3$INTERNET==2
,])
k

## [1] 736

# Probability that a random household from Brazil does not own housing   unit
and have internet connection
round(k/n3, 2)

## [1] 0.07
```

Note: We can also use conditional probability to calculate the required probability as $P(A^c \cap B) = P(A^c|B)P(B)$

2. What is the probability that a random household from Brazil owns housing unit and does not have internet access?

$$
\begin{aligned}
&P(\textit{household owns the housing unit and does not have interent access}) \\
&\quad = P(A \cap B^c) \\
&\quad = \frac{6253}{9851} \\
&\quad \approx 0.63
\end{aligned}
$$

```
# Number of households that own housing unit and do not have internet
connection
l<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==1 & newdata_prob3$INTERNET==1
,])
l

## [1] 6253

# Probability that a random household from Brazil owns housing unit and do no
t have internet connection
round(l/n3, 2)

## [1] 0.63
```

3. What is the probability that a random household from Brazil does not own housing unit and does not have internet access?

$$
\begin{aligned}
&P(\textit{household does not own housing unit and does not have interent access}) \\
&\quad = P(A^c \cap B^c) \\
&\quad = \frac{1429}{9851} \\
&\quad \approx 0.15
\end{aligned}
$$

```
# Number of households that do not own housing unit and do not have      inte
rnet connection
m<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==2 & newdata_prob3$INTERNET==1
,])
m

## [1] 1429

# Probability that a random household from Brazil does not own housing unit
and does not have internet connection
round(m/n3, 2)

## [1] 0.15
```

• SECTION V

1. What is the probability that a randomly selected household from Brazil owns housing unit or has internet access?

$$\begin{aligned}
&P(household\ owns\ housing\ unit\ or\ has\ interent\ access)\\
&= P(A \cup B)\\
&= P(A) + P(B) - \ P(A \cap B)\\
&= 0.78 + 0.22 - 0.15\\
&= 0.85
\end{aligned}$$

The probability that a randomly selected household from Brazil owns housing unit or has internet access is 0.85.

```
# Number of households that own housing unit or they have internet access
o<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==1 | newdata_prob3$INTERNET==2
,])
o
## [1] 8422

# Probability that a random household from Brazil owns housing unit or   has
internet access
round(o/n3, 2)

## [1] 0.85
```

2. What is the probability that a randomly selected household from Brazil does not own housing unit or have internet access?

$$\begin{aligned}
&P(household\ does\ not\ own\ housing\ unit\ or\ have\ interent\ access)\\
&= P(A^c \cup B)\\
&= P(A^c) + P(B) - \ P(A^c \cap B)\\
&= 0.22 + 0.22 - 0.07\\
&= 0.37
\end{aligned}$$

The probability that a randomly selected household from Brazil does not own housing unit or have internet access is 0.37.

```
# Number of households that do not own housing unit or they have internet acc
ess
p<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==2 | newdata_prob3$INTERNET==2
,])
p

## [1] 3598
```

```
# Probability that a randomly selected household from Brazil does not own
housing unit or internet access
round(p/n3, 2)

## [1] 0.37
```

3. What is the probability that a randomly selected household from Brazil owns housing unit or does not have internet access?

$$
\begin{aligned}
&P(\text{household owns housing unit or does not have interent access})\\
&\quad = P(A \cup B^c)\\
&\quad = P(A) + P(B^c) - \ P(A \cap B^c)\\
&\quad = 0.78 + 0.78 - 0.63\\
&\quad = 0.93
\end{aligned}
$$

The probability that a randomly selected household from Brazil owns housing unit or does not have internet access is 0.93.

```
# Number of households that own housing unit or does not have internet   acce
ss
q<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==1 | newdata_prob3$INTERNET==1
,])
q

## [1] 9115

# Probability that a randomly selected household from Brazil owns housing uni
t or does not have internet access
round(q/n3, 2)

## [1] 0.93
```

4. What is the probability that a randomly selected household from Brazil does not own housing unit or does not have internet access?

$$
\begin{aligned}
&P(\text{household does not own housing unit or does not have interent access})\\
&\quad = P(A^c \cup B^c)\\
&\quad = P(A^c) + P(B^c) - \ P(A^c \cap B^c)\\
&\quad = 0.22 + 0.78 - 0.15\\
&\quad = 0.85
\end{aligned}
$$

The probability that a randomly selected household from Brazil does not own housing unit or does not have internet access is 0.85.

```
# Number of households that do not own housing unit or do not have internet
access
r<- nrow(newdata_prob3[newdata_prob3$OWNERSHIP==2 | newdata_prob3$INTERNET==1
,])
r

## [1] 8418

# Probability that a randomly selected household from Brazil does not own
housing unit or does not have internet access
round(r/n3, 2)

## [1] 0.85
```