# 3. Probability

**Readings:** **Rosner: 3.1 – 6**
**SAS:** **PROC FREQ**
**R:** **creating and manipulating data: matrix, data.frame, apply, sum, arithmetic operators**

**A) Definitions and Notation**
**B) Rules of Probability**
**C) Joint, Marginal and Conditional Probability**
**D) Calculating Probabilities**
**E) Independence**

**Examples:** How do we quantify the frequency of random events?

1) National Health Interview Survey 1980-81
Hearing impairments due to injury, age 17+

| Employment | Population | Impairments |
|---|---|---|
| Currently Employed | 98,917 | 552 |
| Currently Unemp | 7,462 | 27 |
| Not in labor force | 56,778 | 368 |
| Total | 163,157 | 947 |

What is the probability of impairment, in general? What is the probability of impairment for those who are unemployed? etc.

2) Diagnostic tests - Using the Pap smear to screen for cervical cancer, the probability of testing positive is 0.8375 for women with cancer and 0.1864 for women without cancer. If a woman tests positive, what is the probability

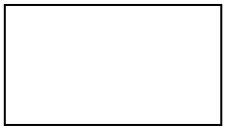
she has cancer? If she tests negative, what is the probability she is free of cancer?

## A) Definitions and Notation

**Random Trial:** Action where outcome is uncertain

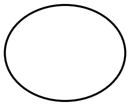
**Population:** elements of a population that have a characteristic in common

**Sample Space:** The set of all possible outcomes. Probability is based on the concept of sample space.

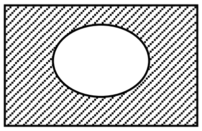
S = 

**Simple Event (or basic outcome):** One possible result of a random trial. Elements of sample space

**Event:** Any set of outcomes of interest. A subset of a sample space.

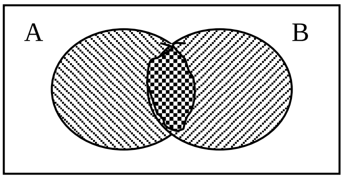


**Complement of Event:** All outcomes that are not in the Event set.

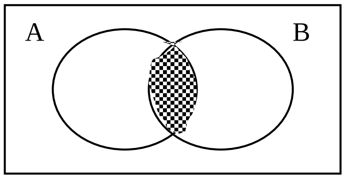


The complement of Event E is $E^c$ (all outcomes not in E).

**A ∪ B (A union B):** All outcomes in A or B, or both:

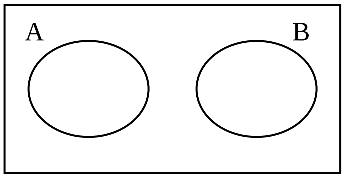


**$A \cap B$ (A intersect B):** All outcomes in A and B:



e.g. A = father smokes, B = mother smokes, $A \cap B$ = both smoke

**Mutually exclusive:** A and B are mutually exclusive if they have ***no*** common outcomes/elements



e.g. Sample space has 6 simple events

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |

Event $E_1 = \{1\ \ 2\ \ 3\ \ \}$ Event $E_2 = \{3\ \ 6\ \}$

$E_1^c = \{4\ \ 5\ \ 6\ \}$

$E_1 \cup E_2 = \{1\ \ 2\ \ 3\ \ 6\}$

$E_1 \cap E_2 = \{\ 3\ \}$

Events mutually exclusive to $E_1$ =

$\{4\ \ 5\ \ 6\}, \{\ 4\ \ 5\}, \{4\ \ 6\ \}, \{5\ \ 6\}, \{4\}, \{5\}, \{6\}$

e.g. Toss 3 coins, record H or T for each.

Sample Space: HHH, HHT, HTH, THH, HTT, THT, TTH, TTT

$E_1$: Two or more heads: {HHH, HHT, HTH, THH}

$E_2$: Three heads: {HHH}

$E_3$: Two or more tails: {TTT, TTH, THT, HTT}

Which are mutually exclusive?:

**Probability:** If a random experiment has *n* equally likely (mutually exclusive and exhaustive) outcomes, *m* of which satisfy an event A, then the probability of the occurrence of A is $P(A)=\frac{m}{n}$.

The probability of an event is the ***relative frequency*** of the event in a population or over an infinite number of repetitions of an experiment or a sample. With statistics, we collect data and try to infer something about the probability of the event from the data.

Objective or Frequentist definition: Repeat the experiment *n* times and record the occurrences of Event A (*m* times). Then $\frac{m}{n} \to P(A)\, as\, n \to \infty$

**Empirical Probabilities:** Probabilities calculated from a finite number of trials are estimates or empirical probabilities. We denote them by $\hat{P}$.

e.g. Recall the hearing impairment survey:

Hearing impairments due to injury, age 17+

| Employment | Population | Impairments |
|---|---|---|
| Currently Employed | 98,917 | 552 |
| Currently Unemp | 7,462 | 27 |
| Not in labor force | 56,778 | 368 |
| Total | 163,157 | 947 |

$\hat{P}$ (currently employed) =

$\hat{P}$ (impairment for currently employed) =

e.g. Results of Screening 135,000 Newborns for PKU.

| | **Screen +** | **Screen -** | **Total** |
|---|---|---|---|
| **PKU +** | 14 | 0 | 14 |
| **PKU -** | 67 | 134,919 | 134,986 |
| **Total** | 81 | 134,919 | 135,000 |

Consider results of screening 135,000 newborn infants for PKU in 115 maternity hospitals in Massachusetts during a 19-month period (Macready and Hussey, 1964, *J Public Health*). PKU is a hereditary metabolic disease that usually results in irreversible retardation if undetected. Since it is detectable and treatable by diet, PKU screening is required by law. No sampling involved, this was all of the newborns at the participating hospitals during the specified time period. Hence, the 135,000 might be considered the ***population*** of interest.

Children with phenylalanine levels in excess of 6 mg/dl were classified as positive. 14 of 81 positives were confirmed by a definitive blood test. There were no undetected cases.

Suppose we select at random one baby from the population screened: **random experiment**

135,000 possible choices: **simple outcomes, simple events**

In this case, selecting at random implies each is **equally likely:** P(John Smith) = 1/135,000

135,000 choices are **exhaustive**

Any two choices are **mutually exclusive** (both cannot occur at the same time)

## B) Probability – Rules and definitions

1) $0 \leq P(\theta_i) \leq 1$ for any simple event $\theta_i$
   0 = can't happen 1 = must happen
2) Probability of the union of *k* mutually exclusive events in the sample space is $\sum_{i=1}^{k} P\left(\vartheta_i\right)$
   e.g. P(head) + P(tail) = 1
3) P(S) = P (entire sample space) = 1

These first 3 are called the ***axioms*** of probability. (An axiom is a statement accepted as true as the basis for argument or inference.)

4) P (null event) = 0
5) $\sum P(\vartheta_i)=1$, summing over all simple events in space
6) $0 \leq P(E) \leq 1$ for any event
7) Probability of mutually exclusive events = P (A or B) = P (A) + P (B)
8) For mutually exclusive events: $P(\theta_1 \cap \theta_2) = 0$
9) Complement of Event A: $A^c$ : $P\left(A^c\right)$ = 1 – P(A)

## C) Joint, Marginal and Conditional Probability

When two or more variables are involved in defining events, several probabilities are of interest.

e.g. From a large population of men:

| | Non-Smoker | Smoker | Total |
|---|---|---|---|
| No respiratory problems | .50 | .30 | .80 |
| Respiratory problems | .05 | .15 | .20 |
| Total | .55 | .45 | 1.00 |

**Joint:** determined by two events.

P (A and B) = P (A $\cap$ B)

P (smoker and respiratory problems) = P (S $\cap$ RP) =

P (Nonsmoker and respiratory prob) = P (NS $\cap$ RP) =

**Marginal:** determined by one event. P (A)

P (smoker) = P(S and R) + P(S and $R^c$) =

P (respiratory problems) = P($S^c$ and R) + P(S and R) =

**Conditional:** determined by fixing one variable.

Probability of B, given A: P (B | A)

$P(B|A) = \dfrac{P(B \cap A)}{P(A)}$ limit sample space to A first; find the fraction of P(A) that is contained in P(A $\cap$ B)

P(respiratory problems | smoker) =

$$P(RP|S) = \frac{P(RP \cap S)}{P(S)} = \frac{.15}{.45} = .33$$

P(respiratory problems | nonsmoker) =

P(no respiratory problems | smoker) =

P(no respiratory prob | nonsmoker) =

*Note: All probabilities refer to this particular population. If empirical probabilities are computed from a non-random sample, much care is needed in interpreting these, especially the joint and marginal probabilities. They may not be meaningful.*

Let's return now to the PKU example to calculate some relevant joint, marginal and conditional probabilities.

e.g. Results of Screening 135,000 Newborns for PKU.

| | **Screen +** | **Screen -** | **Total** |
|---|---|---|---|
| **PKU +** | 14 | 0 | 14 |
| **PKU -** | 67 | 134,919 | 134,986 |
| **Total** | 81 | 134,919 | 135,000 |

**Marginal Probabilities:** numerator comes from margins of table

Probability that randomly selected baby has a positive test?
P(+) = 81/135,000 = 0.0006

Probability that baby will have PKU? P(PKU) = 14/135,000 = .0001

## Joint Probability:

Probability that selected baby does not have PKU and tests positive?

P (no PKU and +) =

Probability that selected baby has PKU and tests negative?

P(PKU and -) =

## Conditional Probabilities:

Probability of PKU for babies with positive test results? Restricting the group to the 81 babies with positive test results

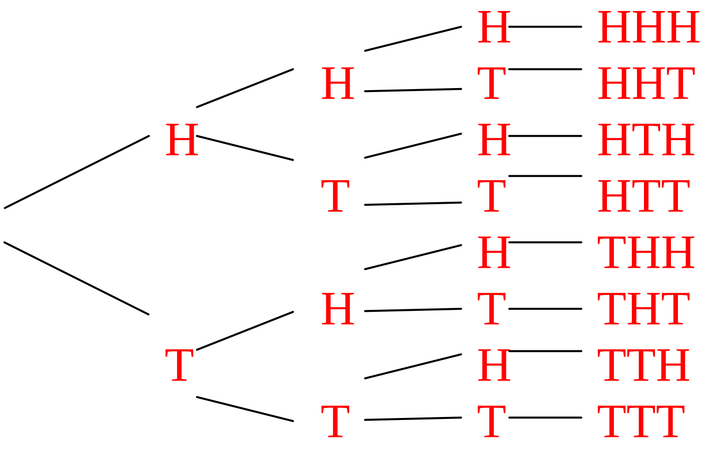
P(PKU | + ) =

## D) Calculating Probabilities

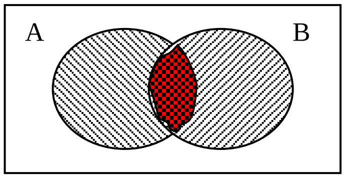
Trees can be useful

Toss a fair coin 3 times: 8 equally likely simple outcomes

Toss 1 Toss 2 Toss 3 Outcome



**Addition Theorem:** For any two events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



***Subtract out intersection/overlap so it's not counted twice***

This helps us combine events at the end of a tree.

e.g. P(2 or more heads ∪ first toss heads) =
4/8 + 4/8 -3/8 = 5/8, or directly

P(HHH or HHT or HTH or HTT or THH) = 5/8

**Complement Theorem:** $A^c$: $P\left(A^c\right)$ = 1 – P(A)

For any event E, $P(E^c) = 1 - P(E)$

e.g. P(no heads) = 1 – P(at least one head) = 1 – 7/8 = 1/8

**Multiplication Theorem:** For any two events A and B,

$$P(A \cap B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$$

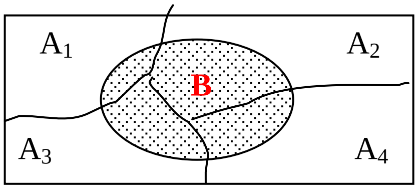
This follows from the definition of conditional probability.

e.g. P (H) = ½, P(T) = ½. Two tosses of a fair coin are assumed to be independent. What's the probability of a head on the first toss and a tail on the second toss?

P(H and T) = ½ x ½ = ¼

**Total Probability Rule:** a very useful result for computing marginal (overall or unconditional probabilities). The total (unconditional probability) of event B is:

$$P(B) = \sum_{i=1}^{k} P\left(B \cap A_i\right) = \sum_{i=1}^{k} P\left(B \middle| A_i\right) P\left(A_i\right)$$



e.g. Hearing impairments due to injury, age 17+

| Employment | Population | Impairments |
|---|---|---|
| Currently Employed | 98,917 | 552 |
| Currently Unemp | 7,462 | 27 |
| Not in labor force | 56,778 | 368 |
| Total | 163,157 | 947 |

P(B) = P(impairments) = P(I|CE)P(CE) + P(I|CU)P(CU) + P(I|NILF)P(NILF) =

## E) Independence

**Independent Events:** Intuitive definition - 2 events are independent if the occurrence of one event does not affect the probability of occurrence of the other.

e.g. Roulette wheel: outcome of one spin should not affect the outcome of subsequent spins.

e.g. In biostatistics, measurements on different people are assumed to be independent, if the people are unrelated.

Events A and B are independent, ***if and only if***:

Definition 1: $P(A \cap B) = P(A)\,P(B)$

or

Definition 2: $P(A | B) = P(A)$ or $P(B | A) = P(B)$

Knowing B (A) doesn't change the chance of A (B)

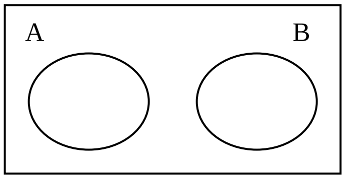
We can use 1) to show 2) : if A and B independent, then

$$P\left(B|A\right)=\frac{P\left(B\cap A\right)}{P\left(A\right)}=\frac{P\left(B\right)P(A)}{P\left(A\right)}=P(B)$$

**Non-Independent Events:** Intuitive ideas -

e.g. age and blood pressure measured on same person

e.g. body temperatures on one person measured 1 hour apart.



Be sure not to confuse independence with being mutually exclusive. Independent events are not mutually exclusive; there must be an intersection of the two independent events.

Mutually exclusive events are completely ***dependent***: if one event occurs we know that the probability of the other event occurring at the same time is 0.

e.g. P (H) = ½, P(T) = ½. What's the probability of heads on the first toss given tails on the first toss?

$P(H|\, T) = P(H \text{ and } T)/P(T) = P(\emptyset)/P(T) = 0/0.5 = 0$

```
pku <- matrix(c(14, 0, 67, 134919), nrow = 2, ncol = 2, byrow = TRUE)
pku

     [,1]   [,2]
[1,]   14      0
[2,]   67 134919

colnames(pku) <- c("screen.pos", "screen.neg")
rownames(pku) <- c("pku.pos", "pku.neg")
pku

        screen.pos screen.neg
pku.pos         14          0
pku.neg         67     134919

sum(pku[, 1])  #column 1 sum, or screen.pos sum

[1] 81

sum(pku[1, ])  #row 1 sum, or pku.pos sum

[1] 14

pku <- as.data.frame(pku)  #convert matrix to a data frame
sum(pku$screen.pos)  #you can refer to data frame columns by name.

[1] 81

# note: you cannot refer to rows by name
# the apply command allows you to apply a function by columns or rows
pku[3, ] <- apply(pku, 2, sum)  #apply the sum function to columns of pku
pku$total <- apply(pku, 1, sum)  #apply the sum function to the rows of pku
rownames(pku) <- c("pku.pos", "pku.neg", "col.total")  #have to reassign all 3 rownames
####
prob.postest <- pku[3, 1]/pku[3, 3]
prob.postest

[1] 6e-04

prob.pku <- pku[1, 3]/pku[3, 3]
prob.pku

[1] 0.0001037

round(prob.pku, 4)

[1] 1e-04
```

```
### another way to create data
population <- c(98917, 7462, 56778)  #create a vector for each column
impairments <- c(552, 27, 368)
hear.mat <- cbind(population, impairments)  #this creates a matrix
hearing <- data.frame(population, impairments)  #this creates instead a data frame
row.names(hearing) <- c("curr.employed", "curr.unemployed", "not.in.labor.force")
hearing

                   population impairments
curr.employed           98917         552
curr.unemployed          7462          27
not.in.labor.force      56778         368

sum(hearing$pop)  #you can abbreviate the variable name as long as it's unique

[1] 163157

sum(hearing$imp)

[1] 947

sum(hearing)  #sums the entire data frame

[1] 164104
```

## *Appendix: Code*

```r
pku<-matrix(c(14,0,67,134919),nrow=2,ncol=2,byrow=TRUE)
pku
colnames(pku)<-c("screen.pos","screen.neg")
rownames(pku)<-c("pku.pos","pku.neg")
pku
sum(pku[,1]) #column 1 sum, or screen.pos sum
sum(pku[1,]) #row 1 sum, or pku.pos sum
pku<-as.data.frame(pku) #convert matrix to a data frame
sum(pku$screen.pos) #you can refer to data frame columns by name.
#note: you cannot refer to rows by name

#the apply command allows you to apply a function by columns or rows
pku[3,]<-apply(pku,2,sum) #apply the sum function to columns of pku
pku$total<-apply(pku,1,sum) #apply the sum function to the rows of pku
rownames(pku)<-c("pku.pos","pku.neg","col.total") #have to reassign all 3 rownames

####
prob.postest<-pku[3,1]/pku[3,3]
prob.postest
prob.pku<-pku[1,3]/pku[3,3]
prob.pku
round(prob.pku,4)

###
#another way to create data
population<-c(98917,7462,56778) #create a vector for each column
impairments<-c(552,27,368)
hear.mat<-cbind(population,impairments) #this creates a matrix
hearing<-data.frame(population,impairments) #this creates instead a data frame
row.names(hearing)<-c("curr.employed","curr.unemployed","not.in.labor.force")
hearing
sum(hearing$pop) #you can abbreviate the variable name as long as it's unique
sum(hearing$imp)
sum(hearing) #sums the entire data frame
```