Probability and Conditional Probability

Week 9

Yunkyu Sohn
School of Political Science and Economics
Waseda University

2019 Spring Statistics II

Why Learn Statistics

- Statistics is the language of science
  - Soon, Statistics will be the language of common sense
- You do not need to be a STEM major to utilize statistics/R
  - Amazing opportunities in between Human/Social + Data
- Social science majors with substantive + statistics knowledge
  - Business / Consulting / Finance
  - Policy making / Program evaluation
  - IT companies: survey design, field experiment designers
- Statistical literacy is becoming arithmetics(+−x÷) of 21C
- Learning statistics is the first step for your success in the long run

This Class

- The first step
  - Essentials for learning statistics
    - Syllabus: https://github.com/ysohn/stats

Lecture / Lab

1. Overview / Introduction to R and R Studio 1 (week starting 0408)
2. Experiments (2.1-2.4) / Introduction to R and R Studio 2 (week starting 0415)
3. Observational Studies (2.5-2.7) / Data Wrangling in R (week starting 0422)
4. Survey sampling (2.1-2.4) / Base Graphics in R (week starting 0506)
5. Correlation and Regression (3.6, 4.2) / Programming Loops in R (week starting 0513)
6. Regression and Prediction (4.2, 4.1) / Regression in R (week starting 0520)
7. Regression and Causation (4.3) / R Quiz (0528 5th period at 3-801) (week starting 0527)
8. Final Exam for Statistics I (no lab; exam: 0604)

This Class

- The first step
  - Essentials for learning statistics
    - Syllabus: https://github.com/ysohn/stats

Lecture / Lab

1. Probability and Conditional Probability / Advanced Graphics in R (week starting 0610)
2. Random Variables and Their Distributions / Probability and Simulations in R (week starting 0617)
3. Estimation / Monte Carlo Simulations in R 1 (week starting 0624)
4. Hypothesis Testing / Monte Carlo Simulations in R 2 (week starting 0701)
5. Regression with Uncertainty 1 / Hypothesis Testing in R (week starting 0708)
6. Regression with Uncertainty 2 / Regression in R 2 (week starting 0715)
7. Review / R Quiz (week starting 0722)
8. Final Exam for Statistics II (no lab; exam: 0730)

Textbook

- Imai (2017)
  - An accessible book
    - with a number of R examples in social science
    - Please visit https://github.com/kosukeimai/qss
    - written in up-to-date causal inference context
    - practical guide to data analysis
    - Math involved! ➠ *Some* efforts expected
  - In this class
    - We need to cut down several subsections
    - Refer to the book for details

Course Basics

- Website: https://github.com/ysohn/stats
- Instructor: Yunkyu SOHN (SPSE Political Science)
  - Fields: Statistical Methods, Computational Social Sci.
  - I am in my first year too
  - email: ysohn.teaching@gmail.com
- For R-related concerns:
  - TA: Masanori KIKUCHI
  - email: waseda.statistics@gmail.com
  - Regular office hours (2hr/week):
    - Thursday 11AM - noon + Wed 4:30 - 5:30 PM

RStudio Trouble-shooting guide

- RStudio, R, R packages, OS update regularly
  - They do not know each others' changes ➠ error/crash
- Google the message you see and your circumstance
  - General guide:
    - https://support.rstudio.com/hc/en-us/articles/200488498-Troubleshooting-Guide-Using-RStudio
    - e.g. install package access denied
    - mostly likely to find a working solution from
      - github or stack overflow
- Ask TA or the instructor when Googling did not help

Evaluation

- Participation (checking attendance from week 3)
- Final Exam for Statistics II (0730)
  - In-class exam
  - Mostly theoretical questions
- R quiz for Statistics II
  - In-class exam
  - R coding questions

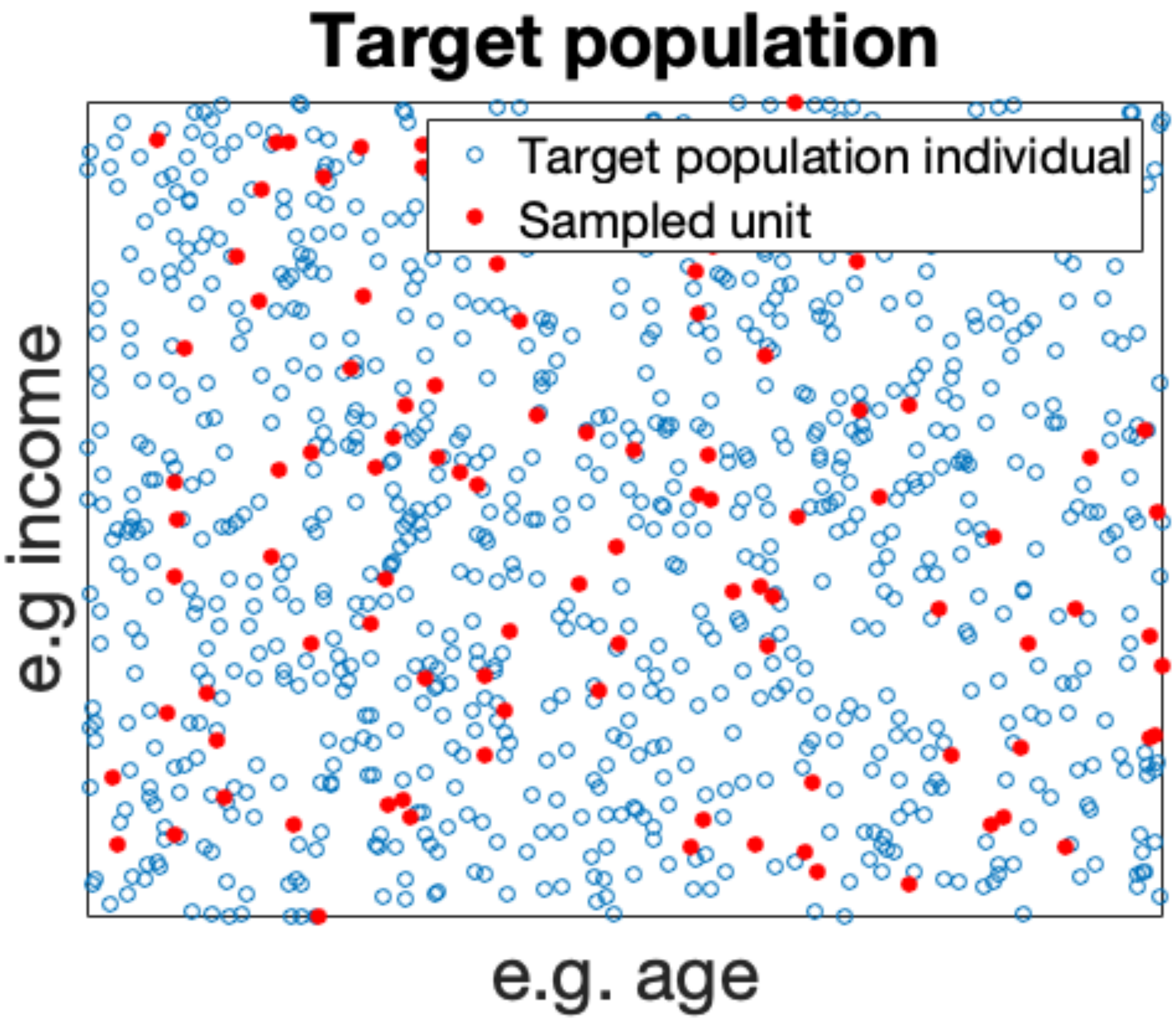
Overview of Statistics I

- All under the sprit of treatment effect identification
  - Randomized Control Trials
    - e.g. call-back rate given race
  - Observational studies (panel data)
    - e.g. minimum wage policy on full-time employment rate
  - List experiment: truthful opinions
  - Causal interpretation of regression coefficients
  - Quantities of interest ➡ single numbers = point estimates
    - Sample treatment effect

Overview of Statistics I

- Limitations of point estimates
  - = Limitations of using single numbers for causal estimates
  - We only can talk about samples
    - Never beyond the samples
      - Our ultimate subject of interest: Population
  - e.g. The case of list experiment
    - Conducting surveys in war zones
      - What percentage of total population can you approach?
        - Given such small number what can we tell about the rest?

Overview of Statistics II



- Statistics I: Using sample (•) to say only about the sample itself
- Statistics II: Using sample (•) to say about the population (o)
  - and all the equipments one needs to have

Overview of Statistics II

- We are going to revisit things covered in Statistics I
  - From a slightly different perspective
    - Probabilistic approach
    - How likely is it that we would observe a pattern in our sample, given what we know about the underlying distribution in the population?
    - e.g. average treatment effect as an interval estimate
  - It will be a long and winding road
    - The basic language of statistics must be mastered beforehand
      - This week: probability and random variables

Contents

Introduction to Probability

- Experiment: an action or a set of actions that produce stochastic events of interest
  - Rolling a dice
  - Coming to the class
  - Voting in an election
- Sample space: a set of all possible outcomes of the experiment, typically denoted by Ω
  - $\{1,2,3,4,5,6\}$
  - {abstain, attend}
  - {abstain, LDP, CDP, DPFP, JCP, ...}
- Event: any subset of the sample space Ω; simple event compound event
  - $\{1\}$, $\{1,5\}$, $\{1,2,3,4,5,6\}$
  - {abstain}, {attend}, {abstain, attend}
  - {abstain}, {LDP}, {CDP, DPFP, JCP}

Introduction to Probability

- Experiment: an action or a set of actions that produce stochastic events of interest
- Sample space: a set of all possible outcomes of the experiment, typically denoted by Ω
- Event: any subset of the sample space Ω
- A^c: complement of a set A
- $P(A)$: probability that event A occurs
  - e.g. tossing coin 3 times: if all outcomes are equally like to occur
    - Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
    - A = {HHH, HHT, THT}
    - $P(A) =$
    - $P(A^c) = 1 - P(A)$

Introduction to Probability

- Probability axioms
  - Probability of any event is non-negative

$$P(A) \geq 0$$

  - Prob. that one of the outcomes in the sample space occurs is 1

$$P(\Omega) = 1$$

  - Addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

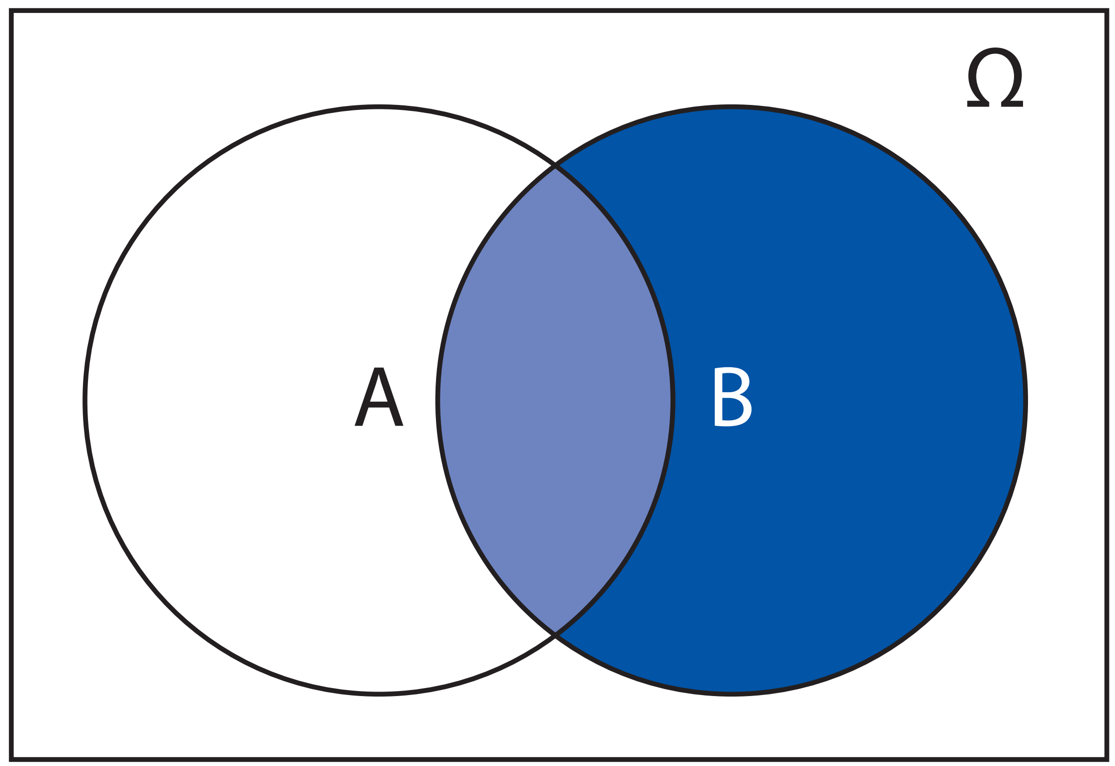
  - Law of total probability

$$P(A) = P(A \text{ and } B) + P(A \text{ and not } B)$$

Introduction to Probability

- Addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



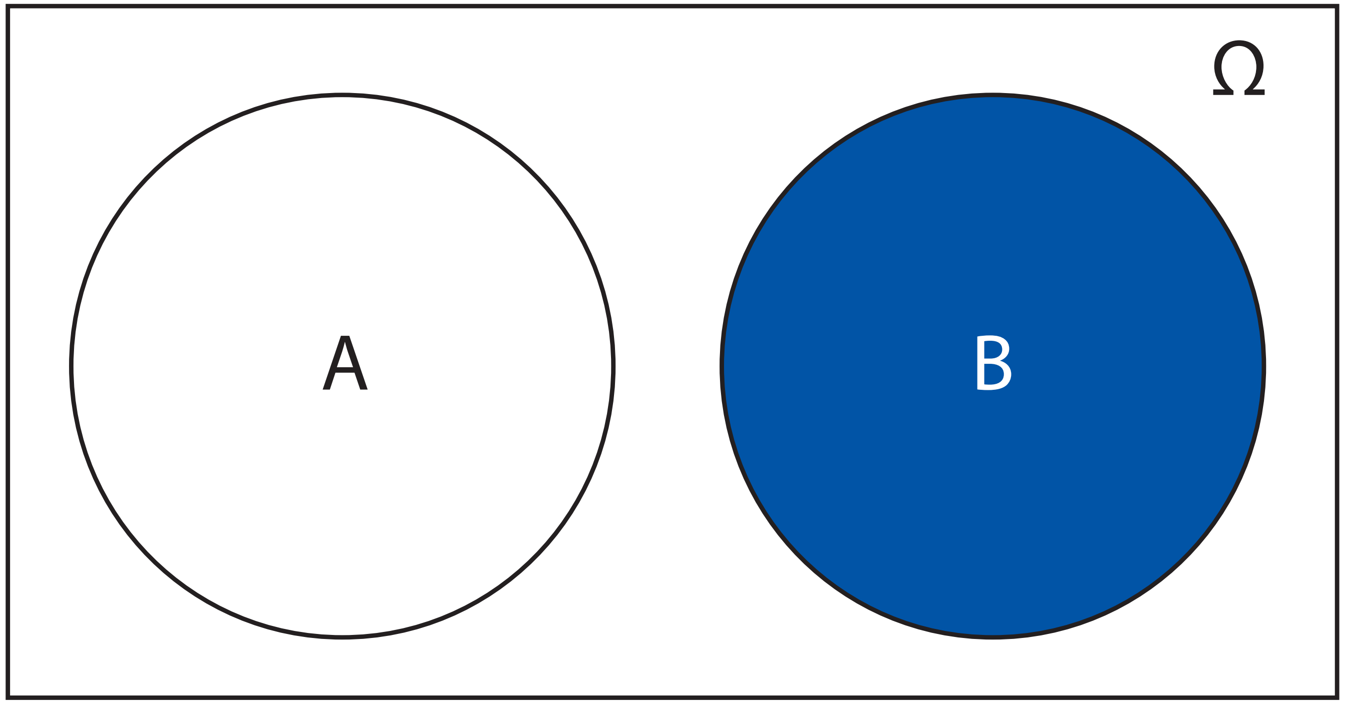
Venn diagram

  - ex1) A: Rain falls; B: Watching Aladin
  - ex2) 3 coin flips: A: H on the first flip; B: T on the second flip
    - Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Introduction to Probability

- Addition rule

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

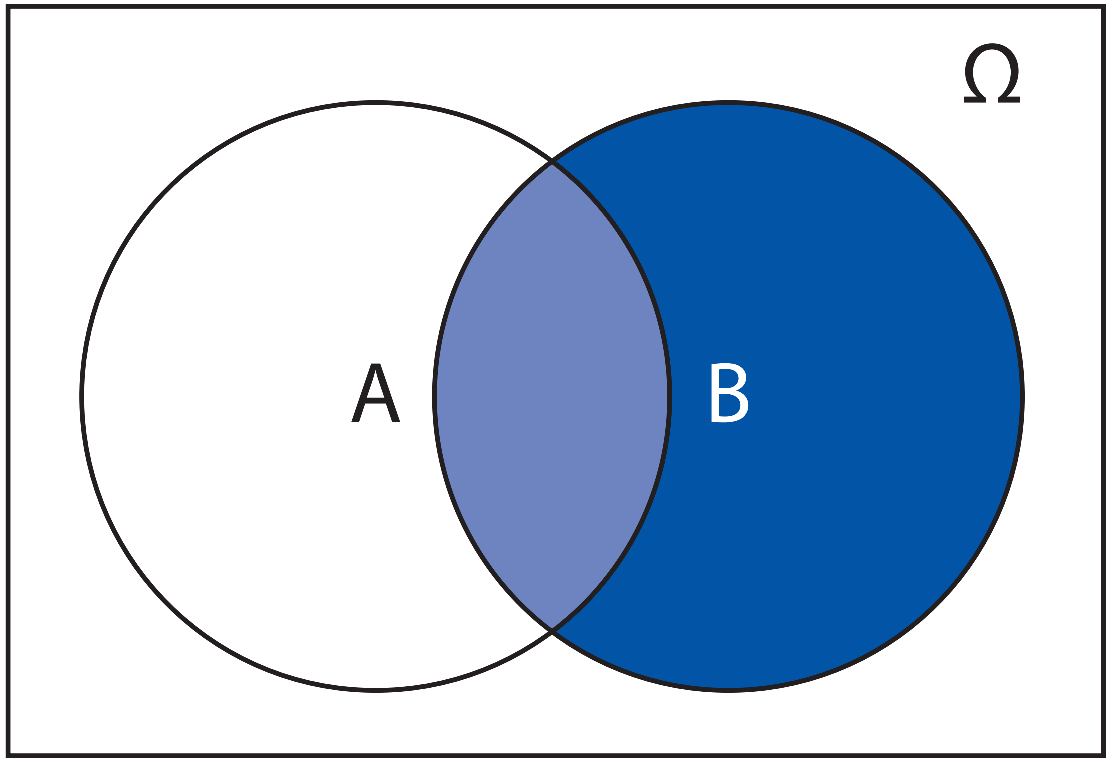


- Mutually exclusive events
  - e.g. A: age = 20 ; B: age = 40
  - e.g. A: abstain; B: vote in an election

Introduction to Probability

- Law of total probability

$$P(A) = P(A \text{ and } B) + P(A \text{ and not } B)$$

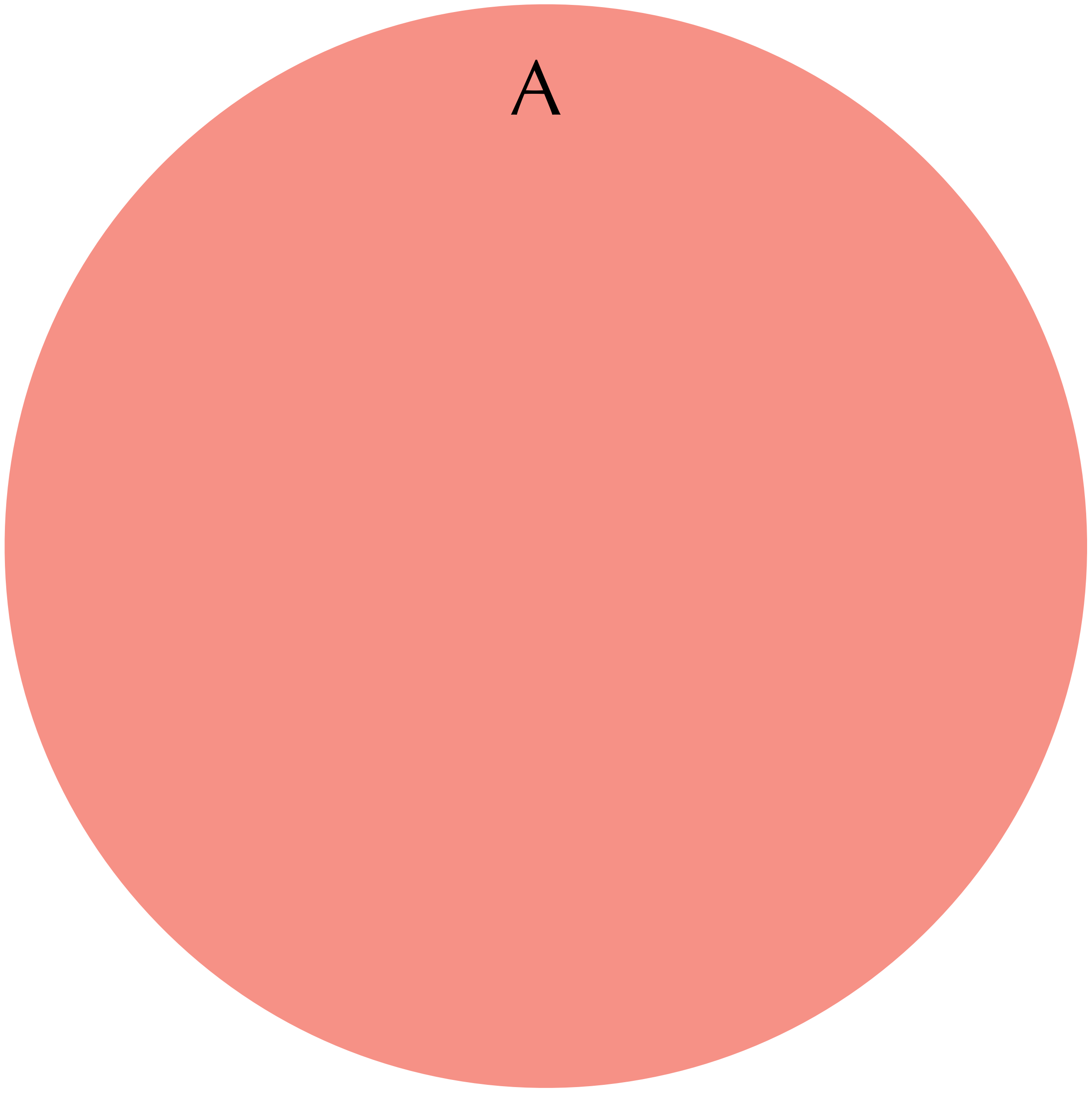


  - ex1) A: Rain falls; B: Watching Aladin
  - ex2) 3 coin flips: A: H on the first flip; B: T on the second flip

Introduction to Probability

- Law of total probability, more generally

$$P(A) = \sum_{i=1}^{N} P(A \text{ and } B_i)$$



Introduction to Probability

- P(A|B) is conditional probability of event A occurring given that event B occurs.

$$P(A \mid B) = \frac{\text{joint probability}}{\text{marginal probability}} = \frac{P(A \text{ and } B)}{P(B)}$$

  - e.g. P(Trump wins | I vote for Trump)

- Independence: Events A and B are independent if and only if

$$P(A \text{ and } B) = P(A)P(B)$$

  - e.g. *A*: Rain falls in Hawaii; B: Watching Aladin

- If A and B are independent:

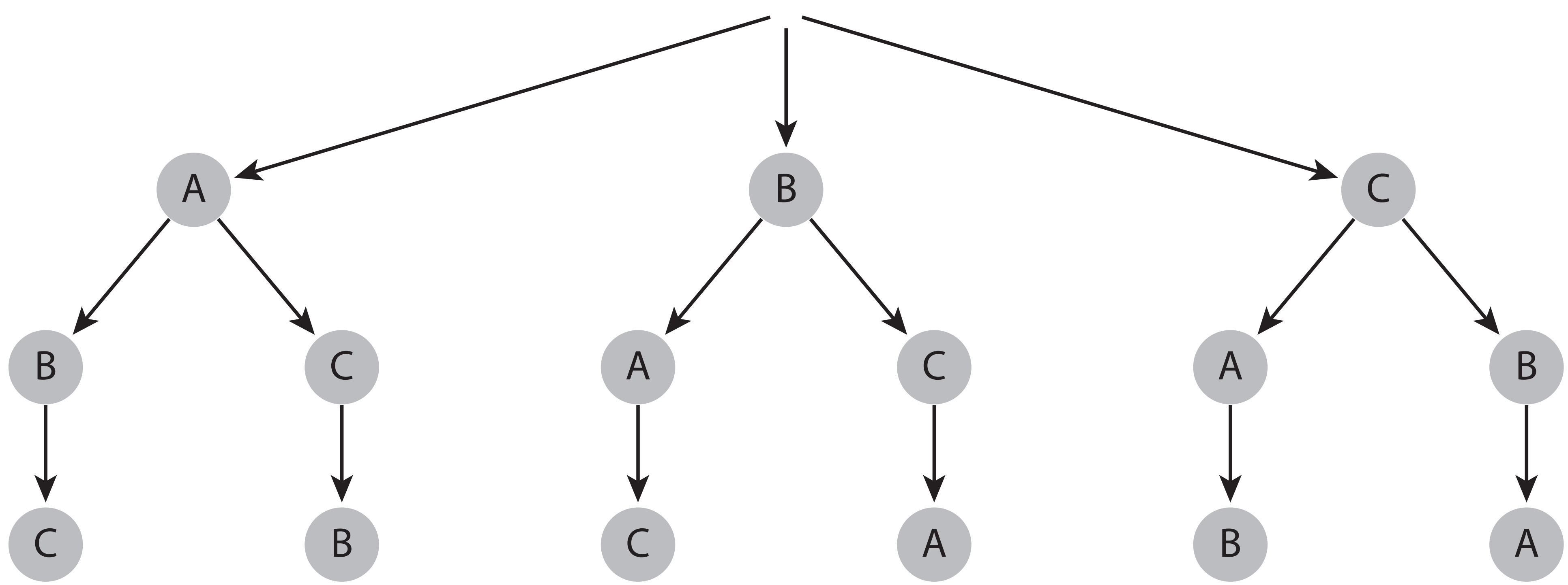
$$P(A \mid B) = P(A) \qquad P(B \mid A) = P(B)$$

Introduction to Probability

- Sampling without replacement 1: #orderings
  - k!(k factorial): k(k-1)(k-2)⋯1   e.g. 5!: 5 x 4 x 3 x 2 x 1 =
  - Permutations: ordering unique k elements out of n

$$_nP_k = n \times (n-1) \times \cdots \times (n-k+2) \times (n-k+1) = \frac{n!}{(n-k)!}$$

  - Ordering 3 unique objects (A, B, C): n=3, k=3

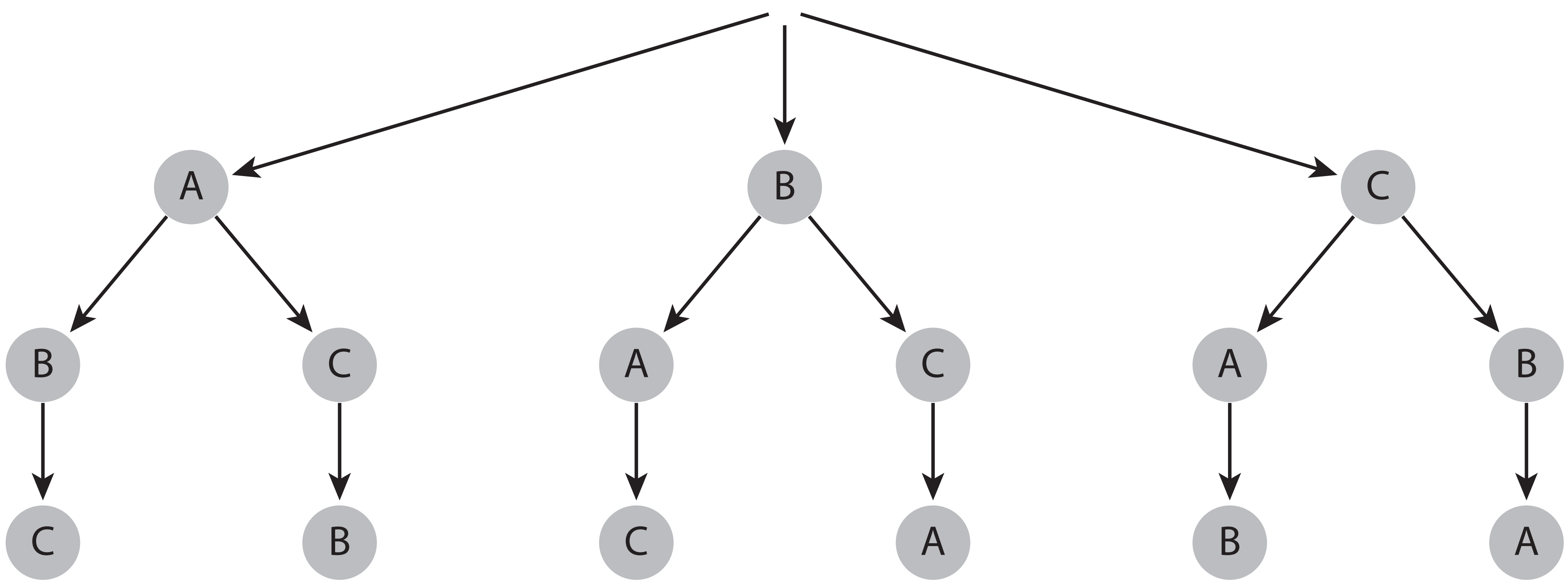


Introduction to Probability

- Sampling without replacement 2: Combinations
- #combinations when choosing k distinct elements from n elements

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

  - How does this differ from permutations? ignore ordering!



Introduction to Probability

- e.g. committee formation
  - selecting 5 among 20 (10 men & 10 women)
    - P (at least 2 women on the committee)

  - Permutations?: ordering matters
  - Combinations?: ordering does not matter

Introduction to Probability

- Schwarzenegger's veto message in 2009.
  - The case of random line breaks..
  - 85 words and 7 lines (6 line breaks)

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many
unnecessary bills come to me for consideration. Water reform, prison reform, and health
care are major issues my Administration has brought to the table, but the Legislature just
kicks the can down the alley.

Yet another legislative year has come and gone with out the major reforms Californians
overwhelmingly deserve. In light of this, and after careful consideration, I believe it is
unnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

Introduction to Probability

- Example: The birthday problem
  - Probability that at least 2 in this classroom having same birthday
  - Analytical derivation
  - Sampling without replacement VS sampling with replacement
  - person $\#1$, $\#2$, $\#3$, ... $\#k$

Introduction to Probability

- Example: The birthday problem
  - Computer calculation

```
birthday <- function(k) {
    logdenom <- k * log(365) + lfactorial(365 - k) # log denominator
    lognumer <- lfactorial(365) # log numerator
    ## P(at least two have the same bday) = 1 - P(nobody has the same bday)
    pr <- 1 - exp(lognumer - logdenom) # transform back
    return(pr)
}
```

  - log-transformation & transform back for large numbers (p.249)

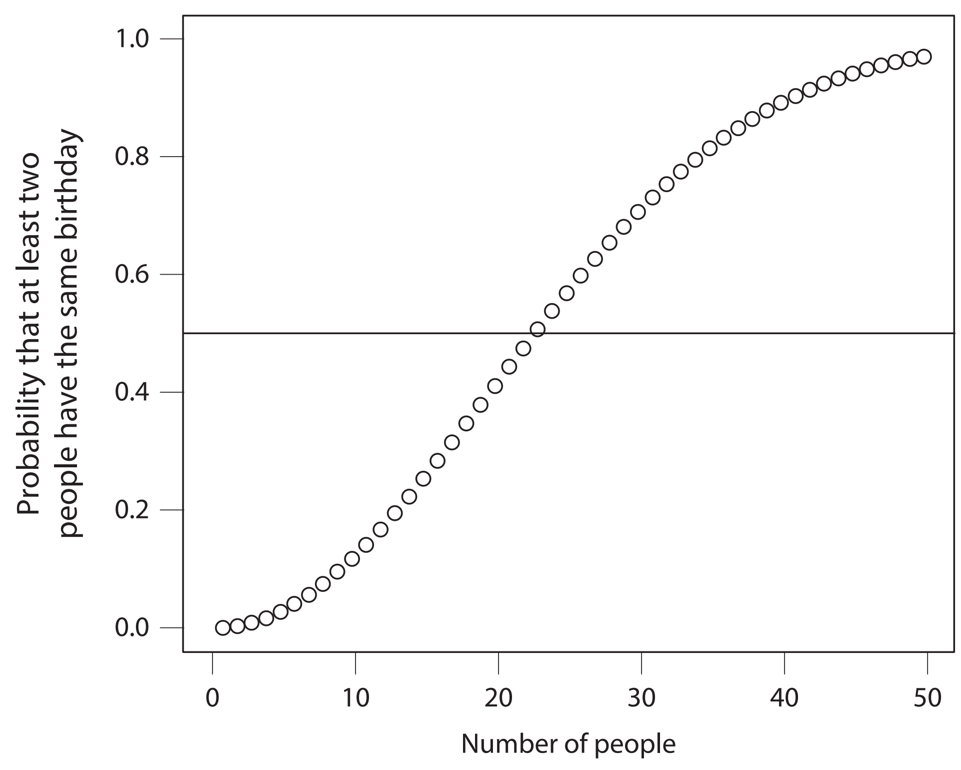
$$\log AB = \log A + \log B, \quad \log \frac{A}{B} = \log A - \log B, \quad \text{and} \quad \log A^B = B \log A$$

Introduction to Probability

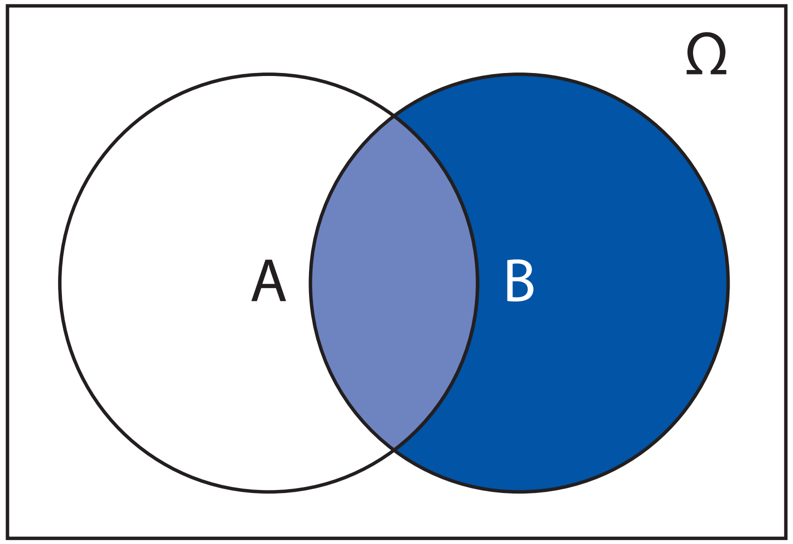
- Example: The birthday problem
  - Computer calculation

```r
k <- 1:50
bday <- birthday(k)  # call the function
names(bday) <- k  # add labels
plot(k, bday, xlab = "Number of people", xlim = c(0, 50), ylim = c(0, 1),
     ylab = "Probability that at least two\n people have the same birthday")
abline(h = 0.5) # horizontal 0.5 line
bday[20:25]

##        20        21        22        23        24        25
## 0.4114384 0.4436883 0.4756953 0.5072972 0.5383443 0.5686997
```



Conditional Probability



- Conditional probability of event A occurring given that event B occurred

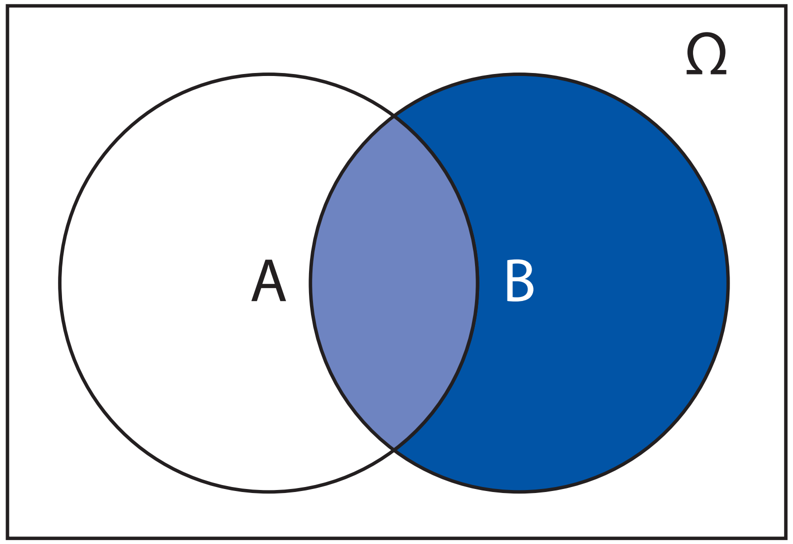
$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- Joint probability of both event A and event B occurring

$$P(A \text{ and } B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Marginal probability of event B: $P(B)$

Conditional Probability



- Conditional probability of event A occurring given that event B occurred

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

- Joint probability of both event A and event B occurring

$$P(A \text{ and } B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Marginal probability of event B: $P(B)$
- Law of total probability (marginal prob. <- conditional/joint prob.)

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$$

Conditional Probability

- e.g. The case of 2 set of twins for 2 couples
  - Couple 1 got informed 1 of 2 is a boy before the delivery
    - For couple 1, what is the probability of both being boys?

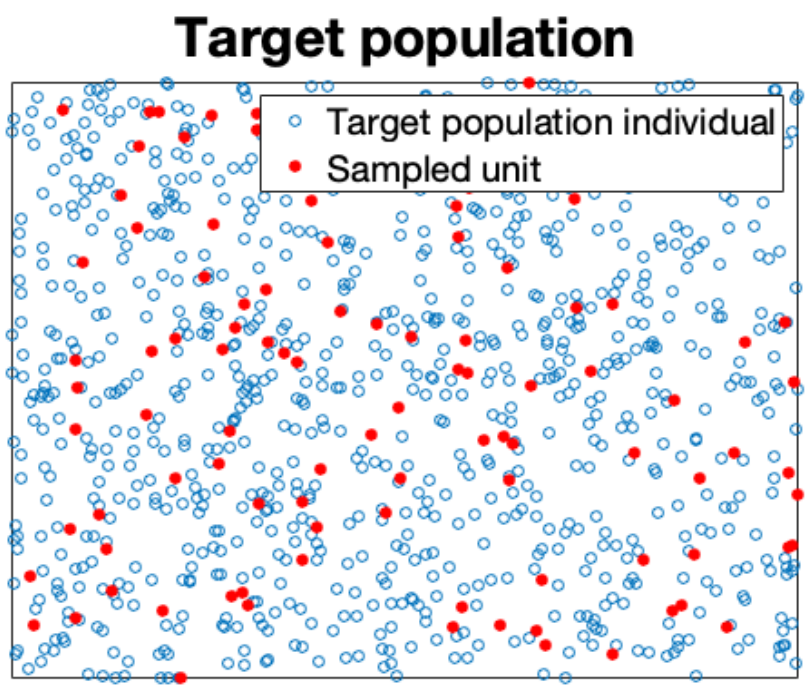
    $P(BB \mid \text{at least one is a boy})$

  - Couple 2 learned the first baby is a boy at the delivery
    - For couple 2, what is the probability of both being boys?

    $P(BB \mid \text{elder twin is a boy})$

Florida Registered Voter List Sample Example

- Inferring population characteristics from a sample



Variable	*Description*
`surname`	surname
`county`	county ID of the voter's residence
`VTD`	voting district ID of the voter's residence
`age`	age
`gender`	gender: `m` = male and `f` = female
`race`	self-reported race

Florida Registered Voter List Sample Example

- Marginal probability

```
margin.race <- prop.table(table(FLVoters$race))
margin.race
```

```
##
##       asian       black    hispanic      native       other
## 0.019203336 0.131021617 0.130802151 0.003182267 0.034017338
##       white
## 0.681773291
```

```
margin.gender <- prop.table(table(FLVoters$gender))
margin.gender
```

```
##
##         f         m
## 0.5358279 0.4641721
```

Florida Registered Voter List Sample Example

- Conditional probability

```
prop.table(table(FLVoters$race[FLVoters$gender == "f"]))
```

```
##
##       asian       black    hispanic      native       other
## 0.016997747 0.138849068 0.136391563 0.003481466 0.032357157
##       white
## 0.671922998
```

- Joint probability

```
joint.p <- prop.table(table(race = FLVoters$race, gender = FLVoters$gender))
joint.p
```

```
##           gender
## race                 f           m
##   asian    0.009107868 0.010095468
##   black    0.074399210 0.056622408
##   hispanic 0.073082410 0.057719741
##   native   0.001865467 0.001316800
##   other    0.017337869 0.016679469
##   white    0.360035115 0.321738176
```

```
rowSums(joint.p)
```

```
##       asian       black    hispanic      native       other       white
## 0.019203336 0.131021617 0.130802151 0.003182267 0.034017338 0.681773291
```

Florida Registered Voter List Sample Example

- Law of total probability

$$P(A) = \sum_{i=1}^{N} P(A \text{ and } B_i)$$

- Marginalizing joint probability

$$P(\text{black}) = P(\text{black and female}) + P(\text{black and male})$$

```
rowSums(joint.p)

##       asian       black    hispanic      native       other
## 0.019203336 0.131021617 0.130802151 0.003182267 0.034017338
##       white
## 0.681773291
```

$$\begin{aligned} P(\text{female}) &= P(\text{female and asian}) + P(\text{female and black}) \\ &\quad + P(\text{female and hispanic}) + P(\text{female and native}) \\ &\quad + P(\text{female and other}) + P(\text{female and white}). \end{aligned}$$

```
colSums(joint.p)

##         f         m
## 0.5358279 0.4641721
```

Florida Registered Voter List Sample Example

- Calculating conditional probability from marginal/joint probabilities

```
colSums(joint.p)

##         f         m
## 0.5358279 0.4641721
```

```
joint.p <- prop.table(table(race = FLVoters$race, gender = FLVoters$gender))
joint.p

##           gender
## race                 f           m
##   asian    0.009107868 0.010095468
##   black    0.074399210 0.056622408
##   hispanic 0.073082410 0.057719741
##   native   0.001865467 0.001316800
##   other    0.017337869 0.016679469
##   white    0.360035115 0.321738176
```

$$P(\text{black} \mid \text{female}) = \frac{P(\text{black and female})}{P(\text{female})} \approx \frac{0.074}{0.536} \approx 0.139$$

Florida Registered Voter List Sample Example

- Joint probability table

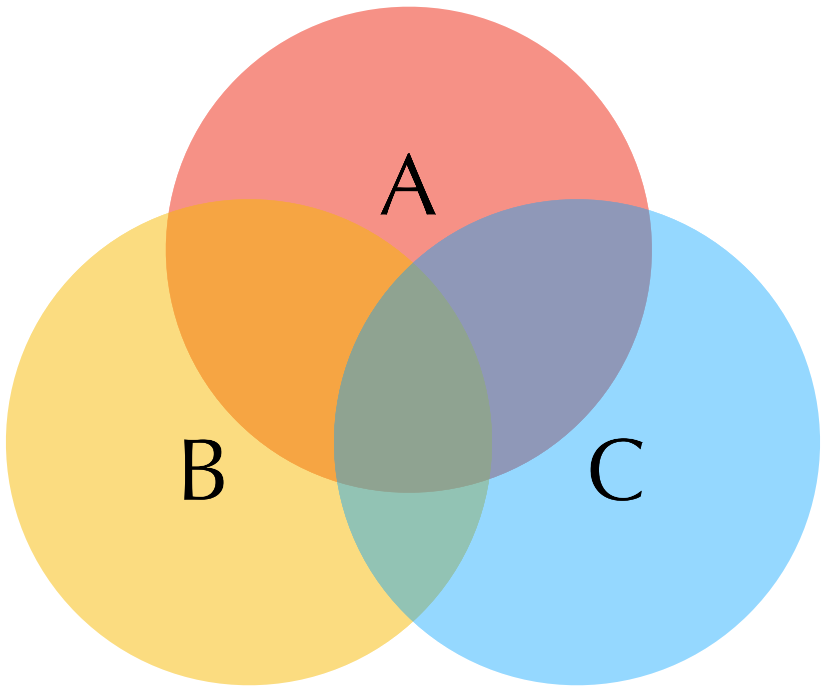
	Gender		
Racial groups	*Female*	*Male*	*Marginal prob.*
Asian	0.009	0.010	0.019
Black	0.074	0.057	0.131
Hispanic	0.073	0.058	0.131
Native	0.002	0.001	0.003
White	0.360	0.322	0.682
Marginal prob.	0.536	0.464	1

Florida Registered Voter List Sample Example

- Generalizing to 3 categories (check p.259 - 261)

$$P(A \text{ and } B \mid C) = \frac{P(A \text{ and } B \text{ and } C)}{P(C)},$$

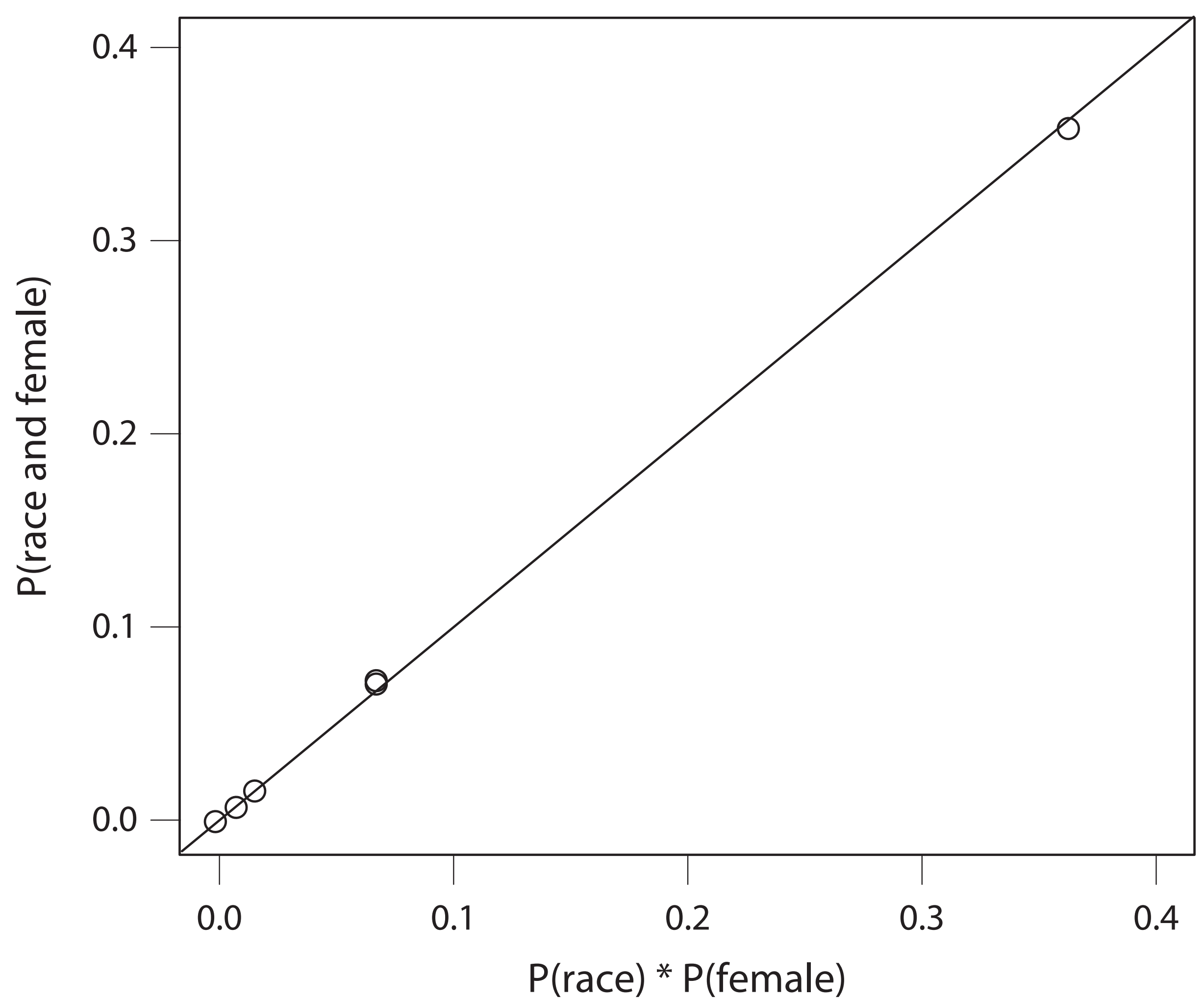
$$P(A \mid B \text{ and } C) = \frac{P(A \text{ and } B \text{ and } C)}{P(B \text{ and } C)} = \frac{P(A \text{ and } B \mid C)}{P(B \mid C)}$$



Florida Registered Voter List Sample Example

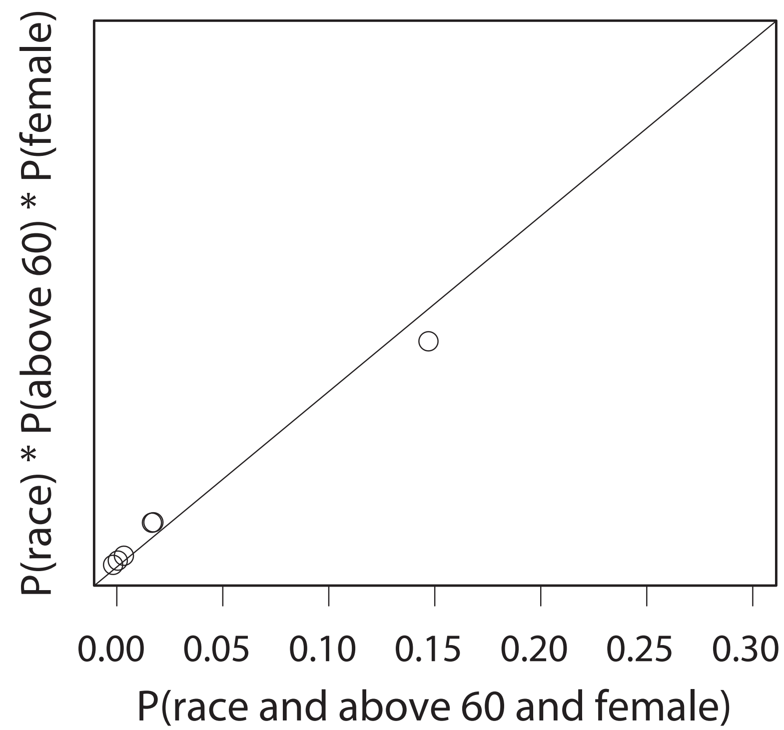
- Check independence: $P(A \text{ and } B) = P(A)P(B)$

```
plot(c(margin.race * margin.gender["f"]), # product of marginal probs.
     c(joint.p[, "f"]), # joint probabilities
     xlim = c(0, 0.4), ylim = c(0, 0.4),
     xlab = "P(race) * P(female)", ylab = "P(race and female)")
abline(0, 1) # 45-degree line
```

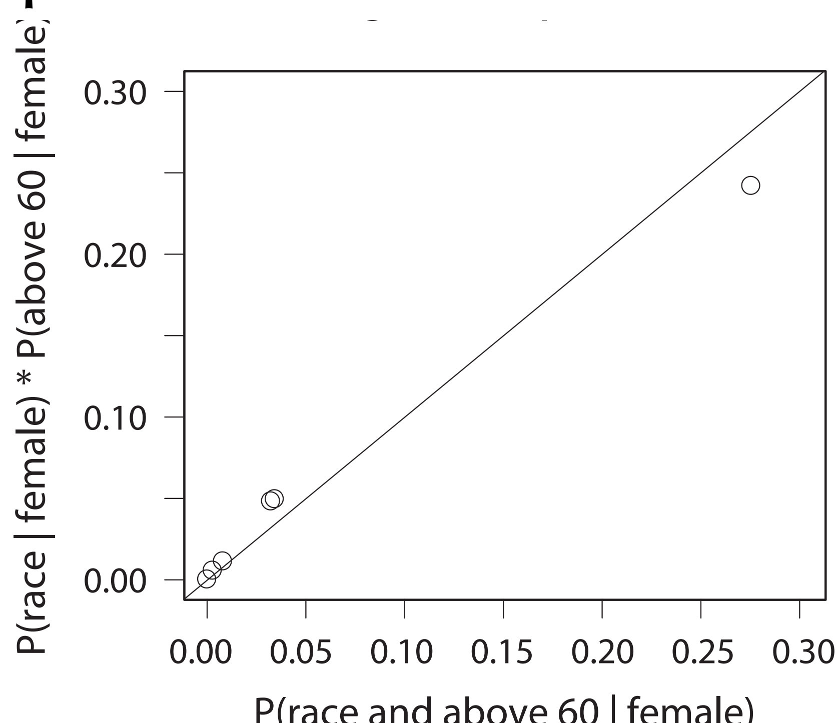


Florida Registered Voter List Sample Example

- Joint independence: $P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C)$



- Conditional independence: $P(A \text{ and } B \mid C) = P(A \mid C)P(B \mid C)$



Conditional Probability

- Bayes' rule (obvious if you know definitions / axioms)

$$\underbrace{\Pr(A \mid B)}_{\textit{conditional probability}} = \frac{\overbrace{\Pr(A \text{ and } B)}^{\textit{joint probability}}}{\underbrace{\Pr(B)}_{\textit{marginal probability}}}$$

$$= \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid \text{not } A)\Pr(\text{not } A)}$$

  - Bayesian updating: prior belief Pr(A) -> posterior belief Pr(A|B)
    - Only using conditional probability of B given A or not A
    - check textbook 6.2.4

Example: Monty Hall Problem

- The most famous probability problem
  - http://www.youtube.com/watch?v=mhlc7peGlGg

  - Even a great mathematician failed!

Example: Monty Hall Problem

- You pick door A. Monty opens door C that has a goat.
  - Should you switch to door B?
- Prior belief: $P(A) = P(B) = P(C) = 1/3$
- Data: Monty reveals C (i.e. MC)
- Posterior belief (inferential goals): $P(A \mid \text{MC})$ and $P(B \mid \text{MC})$
- Question: $P(A \mid \text{MC}) < P(B \mid \text{MC})$
- What do we need? Bayes' rule
  - Key: Monty's behavior is constrained to open a goat door

Summary

- Probability (Chapter 6.1.)
  - Definitions
  - Axioms
  - Permutations
  - Combinations
- Conditional Probability (Chapter 6.2.)
  - Conditional probability
  - Joint probability
  - Independence
- Please read the textbook chapters 6.1 and 6.2 to get familiar

See you next week.