QBS 120 - Lecture 1
Class overview and probability review
(Rice Chapter 1)

Rob Frost

COVID procedures

- TSA survey (link on QBS and Dartmouth homepages): must complete before class
- Class attendance survey (https://dartgo.org/qbsa): complete for each class you attend
- Wear mask
- Sanitize hands and clean desk before and after class

Outline

- Class overview
- Sample spaces
- Probability measures
- Computing probabilities
- Conditional probability
- Independence

People

- Instructor: Rob Frost
- TAs: Courtney Schiebout, Mina Stevanovic

Personal background

- Assistant Professor of Biomedical Data Science
- Graduate students:
  →Xingyu Zheng
  →Quang Nguyen
  →Courtney Schiebout
- Teaching:
  → QBS 120 (Biostats I)
  → Single Cell Interest Group

Path to academic research

- Initial training in mechanical engineering
- ∼10 years in commercial software
- First QBS PhD cohort, defended in 2014

  → Research Associate

  → Instructor (K01)

  → Assistant Professor (research line)

  → Assistant Professor (tenure-track)

Research interests

Theme: methods for the analysis of high-dimensional genomic data

Applied research:

- Gene set testing
- Single cell transcriptomics
- Cancer/immunology genomics
- Gene-gene, gene-environment interactions

Statistical areas of interest:

- Hypothesis aggregation/weighting
- Penalized regression
- Dimensionality reduction (PCA)
- Random matrix theory

Overview

- First of 3 course statistics series (QBS 120, 121, 122)
- Pre-reqs: calculus, linear algebra
- Provide a sufficient foundation for both applied and methodological research
- Provide some exposure to relevant theory
- Combine computational and analytical techniques (R used for computational problems)
- **Calculus-based graduate statistics course; will be challenging and time intensive**

Logistics

- Lectures: Tues/Thurs from 8:30-10:00am in Kellog from 9/15-11/13.
- Problem sets: 1 per week (corrections required)
- Exams: take home final
- Text book: Rice, Mathematical Statistics and Data Analysis, 3rd Edition
- Problem session: once per week, lead by TAs via Zoom (see Canvas for details)
- Office hours: see syllabus/Canvas
- Canvas: all material will be posted on Canvas and PS and exams must be submitted via Canvas.

Expectations

- Attend lecture (in person or via Zoom)
- Finish assigned reading prior to lecture
- Complete all assignments and correct based on solutions.
- Attend problem sessions or office hours as needed
- Final exam is take home and open book but must be completed independently

Grading

- HP/P/LP scale
- Grade weight: PS 50%, final 50%
- Everyone should be able to get a P (do all of the expected work and seek help when needed)
- HP requires exceptional work

Problem Sets

- 1 per week, distributed on Wednesday, due following Monday
- Mix of analytical and computational work
- No handwritten work - recommend Sweave/R markdown type of system
- Free to work in groups but important that you understand how to solve problems
- Solutions may exist out on the web. Aside from answers in the back of Rice, please do not look at solutions until they are posted.
- Random problem(s) will be graded
- Must correct and resubmit based on distributed solutions

5 exposures to content each week

1. **Reading before lecture**
2. Lecture
3. Problem set
4. Problem session
5. Problem set corrections

Exams

- Final (worth 50% of grade)
- Take home and open book but **must be done independently**
- If you complete and understand all of the problem sets, you'll be fine on the exam

- Class overview
- **Sample spaces**
- Probability measures
- Computing probabilities
- Conditional probability
- Independence

Sample spaces

- Ω: Set of all possible random outcomes in a given experiment.
- Size of Ω can be finite or infinite.
- Subsets of Ω are termed events.

Sample space examples

Finite example:

- Nucleotide at specific position in DNA, $\Omega = \{adenine(A), cytosine(C), guanine(G), thymine(T)\}$
- Example event A that base is cytosine, $A = \{C\}$

Infinite example:

- BMI of a human subject, $\Omega = \{b|b \geq 0\}$
- Example event A that subject is obese, $A = \{b|b \geq 30\}$

Set theory basics

- Complement: A^C, event that A does not occur
- Empty set: $\emptyset$, event with no outcomes
- Intersection: $A \cap B$: event that both A and B occur
- Disjoint: $A \cap B = \emptyset$
- Union: $A \cup B$, event that A and/or B occur
- Subset: $A \subseteq B$, all events in A are also in B $(A \cup B = B, A \cap B = A)$
- Proper subset: $A \subset B$, B has some elements not in A $(A^C \cap B \neq \emptyset)$

Set theory basics, continued

- Commutative laws: $A \cap B = B \cap A$, $A \cup B = B \cup A$,
- Associative laws: $(A \cap B) \cap C = A \cap (B \cap C)$, $(A \cup B) \cup C = A \cup (B \cup C)$
- Distributive laws: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, $(A \cap B) \cup C = (A \cup C) \cap (A \cup C)$

- Class overview
- Sample spaces
- **Probability measures**
- Computing probabilities
- Conditional probability
- Independence

Probability measures

A probability measure on Ω is a function $P()$ that maps subsets of Ω to real numbers and satisfies three axioms:

1. $P(\Omega) = 1$
2. $A \subset \Omega \rightarrow P(A) \geq 0$
3. $A \cap B = \emptyset \rightarrow P(A \cup B) = P(A) + P(B)$

   $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Probability measure properties

- Property A: $P(A^C) = 1 - P(A)$
- Property B: $P(\emptyset) = 0$
- Property C: $A \subset B \rightarrow P(A) \leq P(B)$
- Property D (Addition Law):
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof of property A

$$
\begin{align}
A \cap A^C &= \emptyset && \text{by defn} \tag{1}\\
A \cup A^C &= \Omega && \text{by defn} \tag{2}\\
P(A \cup A^C) &= P(A) + P(A^C) && \text{by axiom 3 \& 1} \tag{3}\\
P(\Omega) &= P(A) + P(A^C) && \text{by 2} \tag{4}\\
1 &= P(A) + P(A^C) && \text{by axiom 1} \tag{5}\\
P(A^C) &= 1 - P(A) && \square \tag{6}
\end{align}
$$

Coin toss example

Experiment:

- Coin is flipped twice
- $\Omega = \{hh, ht, th, tt\}$
- $P(hh) = P(ht) = P(th) = P(tt) = 0.25$

What is the probability that the coin is heads on the first or second toss?

Coin toss example, continued

- Event A: heads on first toss, $A = \{hh, ht\}$, $P(A) = P(hh) + P(ht) = 0.5$
- Event B: heads on second toss, $B = \{hh, th\}$, $P(B) = P(hh) + P(th) = 0.5$
- Event C: heads on either first or second toss, $C = A \cup B$

$$P(C) = P(A) + P(B) - P(A \cap B) \quad \text{by property D}$$
$$P(C) = 0.5 + 0.5 - P(hh) = 0.75 \quad \text{by defn}$$

- Class overview
- Sample spaces
- Probability measures
- **Computing probabilities**
- Conditional probability
- Independence

Equally likely outcomes

- $\Omega = \{\omega_1, ..., \omega_N\}$
- If all ω_i are equally likely, $P(\omega_i) = 1/N$
- More generally, if event A can occur in n mutually exclusive ways of equal probability:
  $P(A) = n/N$

In many scenarios, computing n is simple but many realistic cases require non-trivial counting procedures.

Multiplication principle

- Assume there are two independent experiments A and B:

$$\Omega_A = \{a_1, ..., a_n\}$$
$$\Omega_B = \{b_1, ..., b_m\}$$

- The number of outcomes for both experiments is nm.
- Proof: possible outcomes can be held in an $n \times m$ matrix:

$$\begin{array}{ccc} a_1, b_1 & \cdots & a_1, b_m \\ \vdots & \ddots & \vdots \\ a_n, b_1 & \cdots & a_n, b_m \end{array}$$

Extended multiplication principle

- Assume there are p independent experiments each with $n_i, i = 1, ..., p$ outcomes.
- The number of outcomes for all p experiments is $n_i \times ... \times n_p$.
- Proof: by induction

Coin flipping

- Coin is flipped 8 times.
- Consider each flip a separate experiment with two outcomes:

$$\Omega_i = \{h, t\}, i = 1, ..., 8$$

- Number of possible outcomes of all 8 coin flips $= 2^8 = 256$

Permutations

- A permutation is an ordered arrangement of objects.
- How many permutations are possible for r objects taken from set $S = \{s_1, ..., s_n\}$?
- Answer depends on whether duplicates are allowed:
  - Duplicates allowed: **sampling with replacement**
  - Duplicates not allowed: **sampling without replacement**

Permutation examples

How many 10 base DNA sequences are possible?

- Each base is determined by drawing **with replacement** from $\{A, C, T, G\}$

How many ways are there to order 10 people in line?

- Each position is determined by drawing **without replacement** from $\{p_1, ..., p_{10}\}$

Counting permutations: sampling with replacement

Number of ordered samples of size r drawn with replacement from a set of size n can be computed using the extended multiplication principle:

- Treat each draw as an independent experiment (r total experiments)
- With replacement, each experiment has n possible outcomes
- From extended multiplication principle, total number of outcomes is therefore n^r.

Counting permutations: sampling without replacement

Number of ordered samples of size r drawn without replacement from a set of size n can also be computed using the extended multiplication principle:

- Treat each draw as an independent experiment
- Without replacement, the number of possible outcomes decreases by one for each subsequent experiment.
- From extended multiplication principle, total number of outcomes is therefore $n(n-1)(n-2)...(n-r+1)$

If $r = n$, the number of ordered samples without replacement, or orderings of n elements, is $n(n-1)(n-2)...1 = n!$.

Permutation examples

How many 10 base DNA sequences are possible?

- $n^r = 4^{10} = 1048576$

How many ways are there to order 10 people in line?

- $n = r = 10, 10! = 3628800$

Combinations

- A combination is an unordered arrangement of objects.
- How many combinations are possible for r objects taken from set $S = \{s_1, ..., s_n\}$?
- Answer again depends on whether duplicates are allowed but usually interested in sampling without replacement.

Counting combinations

Number of unordered samples of size r drawn from a set of size n can be computed using the multiplication principle:

$$N_p = N_c * N_{oc}$$

where:

- N_p: number of permutations $= n(n-1)...(n-r+1)$
- N_c: number of combinations
- N_{oc}: number orderings of each combination $= r!$

$$N_c = \frac{N_p}{N_{oc}} = \frac{n(n-1)...(n-r+1)}{r!} = \frac{n!}{(n-r)!r!} = \binom{n}{r}$$

Binomial coefficients

- $\binom{n}{r}$ is referred to as ”n choose r”
- Also called ”binomial coefficients” as they represent the coefficients in the polynomial expansion:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

- For $a + b = 2$, simplifies to:

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

- 2^n is therefore the number of subsets of a set of n elements.

Capture/recapture example (Example 1.4 I in Rice)

Technique used to estimate size of a wildlife population:

- Capture, tag and release 10 animals.
- Subsequently capture 20 animals and find 4 that were previously tagged.

What is the best estimate of the total size of the population?

Capture/recapture example, continued

- Call the scenario where 4 of the 10 originally tagged animals are among the 20 captured event A.
- $P(A) =$ (Number of ways A can occur)/(Total number of possible outcomes)
- Assume total population size is n and that all animals are equally likely to be captured.
- Number of possible outcomes is a combination: $\binom{n}{20}$
- Determine numerator from multiplication principle:
  # ways A can occur = (# of ways to select 4 tagged) * (# of ways to select 16 untagged)
- # of ways to select 4 tagged $= \binom{10}{4}$
- # of ways to select 16 untagged $= \binom{n-10}{16}$

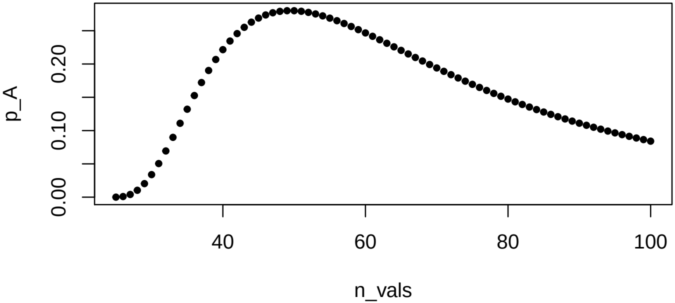
Capture/recapture example, continued

$$P(A) = \frac{\binom{10}{4}\binom{n-10}{16}}{\binom{n}{20}}$$

How to estimate n? Use method of maximum likelihood, i.e., pick n that maximized $P(A)$.

Capture/recapture example, continued

```
n_vals = 25:100
p_A = sapply(n_vals, function(x) {
return ( (choose(10,4)*choose(x-10,16))/choose(x,20))
})
plot(n_vals, p_A, pch=20)
```



Likelihood is maximized at $n \sim 50$.

Grouping into multiple classes (multinomial coefficients)

How many ways can n objects be grouped into r classes of sizes $n_i, i = 1, ..., r, \sum_{i=1}^{r} n_i = n$?

$$\binom{n}{n_1 n_2 ... n_r} = \frac{n!}{n_1! n_2! ... n_r!}$$

Can derive via multiplication principle:

- $\binom{n}{n_1}$ ways to choose elements of first group
- $\binom{n-n_1}{n_2}$ ways to choose elements of second group, etc.

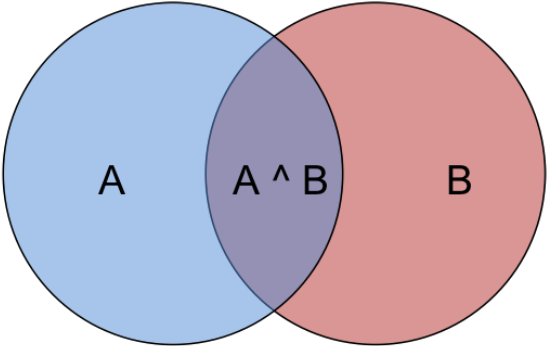
$$\binom{n}{n_1}\binom{n-n_1}{n_2}...\binom{n-n_1-n_2...-n_{r-1}}{n_{r-1}} = \binom{n}{n_1 n_2 ... n_r}$$

Note that normal combination rule is special case for $r = 2$.

- Class overview
- Sample spaces
- Probability measures
- Computing probabilities
- **Conditional probability**
- Independence

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$



$P(A|B)P(B) = P(A \cap B)$ Multiplication Law

Example: 2×2 contingency table

	No disease	Disease	
Negative test	A	B	r_1
Positive test	C	D	r_2
	c_1	c_2	n

- $c_1 = A + C$, $c_2 = B + D$, $r_1 = A + B$, $r_2 = C + D$
- $n = A + B + C + D = c_1 + c_2 = r_1 + r_2$
- Unconditional probability of disease:
  $P(disease) = c_2/n$
- Conditional probability of disease given positive test:
  $P(disease|positive) =$
  $P(disease \cap positive)/P(positive) = \frac{D/n}{r_2/n} = D/r_2$

Example: urn (Rice Example 1.5 A)

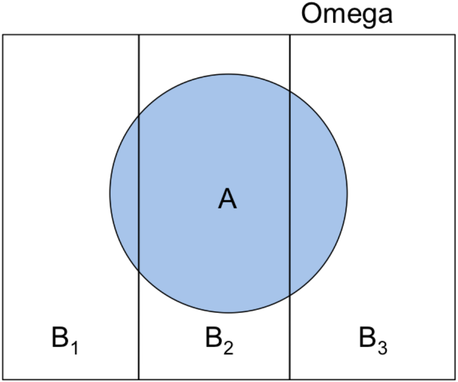
- Urn contains 3 red balls and 1 blue ball.
- 2 balls are selected without replacement.
- What is the probability that both balls are red?
- Let R_1 and R_2 represent events that red is drawn first and second.
- Prob that both are red:

$$\begin{aligned} P(R_1 \cap R_2) &= P(R_1)P(R_2|R_1) \\ &= 3/4 * P(R_2|R_1) \\ &= 3/4 * 2/3 \\ &= 1/2 \end{aligned}$$

Law of total probability

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

where $\cup_{i=1}^{n} B_i = \Omega$, $B_i \cap B_j = \emptyset$ for $i \neq j$, $P(B_1) > 0$



Example: urn (Rice Example 1.5 C)

- What is probability that red ball is drawn second?
- Use law of total probability with R_1 and B_1 as disjoint events that cover Ω:

$$\begin{aligned}
P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|B_1)P(B_1) \\
&= P(R_2|R_1) * 3/4 + P(R_2|B_1) * 1/4 \\
&= 2/3 * 3/4 + 1 * 1/4 \\
&= 1/2 + 1/4 = 3/4
\end{aligned}$$

Bayes' rule

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$

where $\cup_{i=1}^n B_i = \Omega$, $B_i \cap B_j = \emptyset$ for $i \neq j$, $P(B_1) > 0$

Derive as follows:

$$\begin{aligned} P(B_j|A) &= \frac{P(B_j \cap A)}{P(A)} && \text{defn of cond prob} \\ &= \frac{P(A|B_j)P(B_j)}{P(A)} && \text{symm of mult law} \\ &= \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i))} && \text{law of total prob} \end{aligned}$$

When to use Bayes' rule? When $P(B_j|A)$ is desired but know $P(A|B_j)$ and $P(B_j)$.

Bayes' rule example (Rice Example 1.5 F)

What to find the probability that a person who fails a lie-detector test is telling the truth.

- T is event they are telling the truth; L is event they are lying.
- $+$ is event the test indicates a lie; $-$ the event the test does not indicate a lie
- From lie-detector studies know that $P(+|L) = 0.88$, $P(-|T) = 0.86$ Also know complements: $P(-|L) = 0.12$, $P(+|T) = 0.14$
- Subjects are unlikely to lie: $P(T) = 0.99, P(L) = 0.01$

Use Bayes' rule to find $P(T|+)$:

$$\begin{aligned} P(T|+) &= P(+|T)P(T)/(P(+|T)P(T) + P(+|L)P(L)) \\ &= (0.14)(0.99)/((0.14)(0.99) + (0.88)(0.01)) \\ &= 0.94 \end{aligned}$$

A 94% false positive rate!

- Class overview
- Sample spaces
- Probability measures
- Computing probabilities
- Conditional probability
- **Independence**

Independence

- Events are independent if knowledge of one event doesn't change probability of other event.
- For events A and B, this implies that $P(A|B) = P(A)$ and $P(B|A) = P(B)$
- Multiplication rule becomes:

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

Q: are disjoint events independent? No!

Example: die rolling

- Role a six-sided die twice.
- What is the probability of rolling two sixes?
- Let A be the event that first roll is a six and B the event that the second role is a six.
- For a fair die, $P(A) = P(B) = 1/6$.
- Since the two rolls are independent: $P(A \cap B) = P(A)P(B) = 1/36$.

Example: uninformative test

	Disease	Not disease	
Negative test	10	40	50
Positive test	20	80	100
	30	120	150

- $P(D) = 30/150 = 0.2$, $P(D|+) = 20/100 = 0.2$, $P(D|-) = 10/50 = 0.2$
- $P(D \cap +) = 20/150 = 2/15 = P(D)P(+) = 30/150 * 100/150 = 1/5 * 2/3 = 2/15$

Pairwise and mutual independence

- Pairwise independence: Set of events $A_1, ..., A_n$ are pairwise independent if any two $A_i, A_j, i \neq j$ are independent.
- Mutual independence: Events are mutually independent if for any subcollection :

$$P(A_{i_1} \cap ... \cap A_{i_m}) = P(A_{i_1})...P(A_{i_m})$$

Example: coin flipping (Rice Example 1.6 C)

- Coin is flipped twice. Let A be event that first flip is heads, B be event that second flip is heads and C event that there is only one heads.
- A, B and C are pairwise independent:
    - A and B are clearly independent.
    - A and C are independent: $P(C) = P(C|A) = 0.5$
    - A and B are independent: $P(C) = P(C|B) = 0.5$
- But A, B and C are not mutually independent (C is disjoint from $A \cap B$):

$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$$