QCB 408 / 508 – Notes on Week 1

Student

2020-03-01

Summary

Topics covered in Week 1 included:

- Components of Applied Statistics (Mon.)
- Central Dogma of Statistical Inference (Mon.)
- Features of SNP and RNA-seq data (Mon.)
- Probability, including axioms, conditional probability, independence, Bayes theorem, law of the total probability (Wed.)
- Random variables and their distributions, including cdf, pmf/pdf and measures of center (mean,median) and spread (variance, standard deviation, covariance and correlation) (Wed.)
- Special and common cases of random variables, including those that are discrete (uniform, Bernoulli, Binomial and Possion distrubutions) and continuous (uniform, Normal and Beta distributions) (Wed.)

Notes for Mon.

- Components of Applied Statistics include study design, data wrangling, data analysis, interpretation, decision and communication.
- Central Dogma of Statistical Inference
- SNP data can be presented as a matrix in which rows, columns and values are SNPs, individuals and corresponding genetypes, respectively.
- RNA-seq data are count-based, which means the raw values of its matrix are read counts. Batch effects are one of the major source of technical variations that effect expression levels (relative proportions) of genes and samples.

Notes for Wed.

1. Probability

1.1 Probability Space $(\mathbf{\Omega}, F, Pr)$

Definitions 1.1.1:

- $\mathbf{\Omega}$: set of all possible outcomes, sample space
- Pr: probability measure
- Events $A \subseteq \Omega$, calculate $Pr(A)$
- F: σ-algebra, all events A where $Pr(A)$ is meaningful

Examples ($\mathbf{\Omega}$):

- $\mathbf{\Omega} = \{TT, HT, TH, HH\}$, coin flip
- $\mathbf{\Omega} = \{1, 2, 3, 4, 5, 6\}$, roll a die
- $\mathbf{\Omega} = \{CC, CT, TT\}$, diploid genotypes
- $\mathbf{\Omega} = \{C, T\}$, haploid genotypes
- $\mathbf{\Omega} = R$, stock returns
- $\mathbf{\Omega} = [0, +\infty)$, height

1.2 Mathematical Probability

Definitions 1.2.1 (Set Operation): Let A and B be events in a sample space $\mathbf{\Omega}$.

- $A \cap B$ is the set of outcomes that are in both A and B.
- $A \cup B$ is the set of outcomes that are in either A or B (or both).
- A^c is the set of outcomes that are not in A (but are in $\mathbf{\Omega}$).
- $A|B$ is the set of outcomes that are in A and not in B.

Definitions 1.2.2 (Mathematical Probability):

- the probability of any event A such that $0 \leq Pr(A) \leq 1$.
- $Pr(\Omega) = 1$, the probability of the sample space is 1.
- Let A^c be the complement of A, then $Pr(A^c) + Pr(A) = 1$.
- Probabilities are countably additive. For any n events such that $A_i \cap A_j = \varnothing, \forall i \neq j$ (pairwise disjoint), then $Pr(\cup_{j=1}^{n} A_j) = \sum_{j=1}^{n} Pr(A_j)$.

Example: $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

1.3 Conditional Probability

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

Example (Ice Cream): 80% of your friends like Chocolate, and 40% like Chocolate and like Vanilla. What percent of those who like Chocolate also like Vanilla? Hint: $P(V|C) = P(V \cap C)/P(C) = 0.4/0.8 = 0.5$

1.4 Independence

Events A and B are independent if (all equivalent):

$$Pr(A|B) = Pr(A)$$
$$Pr(B|A) = Pr(B)$$
$$Pr(A \cap B) = Pr(A)Pr(B)$$

Example: Two dice are rolled. Let A, and B be the events "The first die is a 3" and "The sum of the dice is 8", respectively. A and B are independent. Hint: $Pr(B) = \frac{6}{36} = \frac{1}{6}$, $Pr(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$

1.5 Bayes Theorem

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$$
$$Pr(A \cap B) = Pr(B|A)Pr(A) = Pr(A|B)Pr(B)$$

1.6 Law of total probability

Events $A_1, A_2, ..., A_n$ such that $A_i \cap A_j = \varnothing, \forall i \neq j$ and $\cup A_i = \Omega$, then for any event B,

$$Pr(B) = \sum_{i=1}^{n} Pr(B|A_i)Pr(A_i)$$

$A_i \cap B, i = 1, 2, ..., n \cup_{i=1}^{n} A_i \cap B = B$ and disjoint

$$Pr(B) = \sum_{i=1}^{n} Pr(B \cap A_i) = \sum_{i=1}^{n} Pr(B|A_i)Pr(A_i)$$

2. Random Variable (rv)

2.1 Definition of rv's

A **random variable** (rv) X is a function:

$$X : \mathbf{\Omega} \to R$$

Take any outcome $w \in \mathbf{\Omega}$, the $X(w)$ produces a real value.

The "range" of X is:

$$\Re = \{X(w), w \in \Omega\}, \Re \subseteq R$$

Notes:

- Discrete rv's have a discrete $\Re$. E.g. $\Re = 0, 1, 2, ..., 10$, $\Re = 0, 1, 3, 4, ...$
- Continuous rv's have a continuous $\Re$. E.g. $\Re = [0, 1]$, $\Re = R$.

Examples: $\mathbf{\Omega} = \{CC, CT, TT\}$, SNP genotype. X refers to numbers of T alleles.
$X(CC) = 0$, $Pr(X = 0) = Pr(\{CC\})$
$X(CT) = 1$, $Pr(X = 1) = Pr(\{CT\})$
$X(TT) = 2$, $Pr(X = 2) = Pr(\{TT\})$

2.2 Distribution of rv's

2.2.1 cumulative distribution function (cdf)

Definitions 2.2.1:

For both discrete and continuous rv's, the **Cumulative Distribution Function** (cdf) is:

$$F(y) = Pr(X \leq y)$$

Example:

- $F(1) = Pr(X \leq 1) = Pr(\{CC, CT\})$
- $F(1.1) = F(1)$

2.2.2 Probability mass or density functions (pmf or pdf)

Definitions 2.2.2:

For Discrete rv's, the **Probability Mass Function** (pmf) is

$$f(x) = Pr(X = x), \forall x \in \Re$$

$$f(x) = F(x) - F(b), b \uparrow x$$

For continuous rv's, the **Probability Density Function** (pdf) is

$$f(x) = \frac{d}{dx} F(x)$$

Notes: cdf calculation differs with discrete and continuous rv's.

- Discrete rv's:
$$F(y) = \sum_{x \in y, x \in \Re} f(x) = Pr(x \leq y)$$
- Contiuous rv's (Note that $Pr(X = x) = 0$):
$$F(y) = \int_{-\infty}^{y} f(x)dx = Pr(X \leq y)$$

2.3 Median, mean, variance, covariance and correlation of rv's

2.3.1 Median

Median of a distribution (aka rv): a value y s.t. $F(y) = 0.5$

2.3.2 Expected value of "population mean"

$$\mathrm{E}[X] = \sum_{x \in \Re} x f(x), discrete$$

$$\mathrm{E}[X] = \int x f(x) dx, continuous$$

$$\mathrm{E}[X] = \int x F(x) dx, measure\ theory$$

2.3.3 Population variance

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sum_{x \in \Re} (x - \mathrm{E}[X])^2 f(x), discrete$$

$$\mathrm{Var}[X] = \int (x - \mathrm{E}[X])^2 f(x) dx, continuous$$

2.3.4 Standard deviation

$$SD[X] = \sqrt{\mathrm{Var}[X]}$$

Notes: Both $\mathrm{Var}[X]$ and $SD[X]$ show the typical size of the deviation of the rv X from $E[X]$.

2.3.5 Covariance of rv's X and Y

$$\mathrm{Cov}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$
$$\mathrm{Var}[X] = \mathrm{Cov}[X, X]$$

2.3.6 Correlation of rv's X and Y

$$Cor[X, Y] = \frac{\mathrm{Cov}[X, Y]}{SD[X] SD[Y]}$$

Notes:

- correlation is a scaled version of covariance.
- (*example*) X:heights; Y:weights; $-1 \leq \mathrm{Cov}[X, Y] \leq 1$

3. Special and common cases of discrete rv's

Table Summary

rv's	Uniform	Bernoulli	Binomial	Poisson
Notation	Uniform(1,2,...,n)	Bernoulli(p)	Binomial(n,p)	Poisson(λ)
$\Re$	{1,2,...,n}	{0,1}	{0,1,2,...,n}	{0,1,2,...}
pmf	$f(x;n) = \frac{1}{n}$	$f(x;p) = (1-p)^{1-x} p^x$	$f(x;p) = \binom{n}{x} p^x (1-p)^{n-x}$	$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$
E[X]	$\frac{n}{2}$	p	np	λ
Var[X]	$\frac{(n+1)^2 - 1}{12}$	$p(1-p)$	$np(1-p)$	λ

Notes:

- Uniform: `sample` in R
- Bernoulli:
  $f(0) = 1 - p, f(1) = p$
  $\mathrm{E}[X] = 0 \times f(0) + 1 \times f(1) = p$
  $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = (0-p)^2 \times f(0) + (1-p)^2 \times f(1) = p(1-p)$
- Binomial:
  sum of independent Bernoulli(p)
- Poisson:
  `dpois` -> pmf
  `ppos` -> cdf
  `qpos` -> quantile
  `rpois` -> random draws

Examples:

- Bernoulli:
  Toss a coin. Arbitrarily define heads to be success. Then $p = 0.5$;
  Roll a die. Arbitrarily define rolling a six to be success. Then $p = \frac{1}{6}$.
- Binomial: Under Hardy-Weinberg Equilibrium, X refers to numbers of T alleles, X~Binomial(2,p), where p is the allele frequency of T.
  $Pr(X = 0) = (1-p)^2$
  $Pr(X = 1) = 2p(1-p)$
  $Pr(X = 2) = p^2$

4. Special and common cases of continuous rv's

Table Summary:

rv's	Uniform(0,1)	Uniform(0,θ)	Normal(μ, σ^2)	Beta(α,β)
Notation	Uniform(0,1)	Uniform(0,θ)	Normal(μ, σ^2)	Beta(α,β), $\alpha, \beta > 0$
$\Re$	$[0, 1]$	$[0, \theta]$	$(-\infty, +\infty)$	$(0, 1)$
pdf	$f(x) = 1$	$f(x;\theta) = \frac{1}{\theta}$	$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$	$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
cdf	$F(y) = y$	$F(y) = \frac{y}{\theta}$	-	-
E[X]	$\frac{1}{2}$	$\frac{\theta}{2}$	μ	$\frac{\alpha}{\alpha+\beta}$
Var[X]	$\frac{1}{12}$	$\frac{\theta^2}{12}$	σ^2	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Session Information

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS  10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] forcats_0.4.0   stringr_1.4.0   dplyr_0.8.1     purrr_0.3.2
 [5] readr_1.3.1     tidyr_0.8.3     tibble_2.1.1    ggplot2_3.1.1
 [9] tidyverse_1.2.1 knitr_1.22

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.1       cellranger_1.1.0 pillar_1.4.0     compiler_3.6.0
 [5] plyr_1.8.4       tools_3.6.0      digest_0.6.18    lubridate_1.7.4
 [9] jsonlite_1.6     evaluate_0.13    nlme_3.1-140     gtable_0.3.0
[13] lattice_0.20-38  pkgconfig_2.0.2  rlang_0.3.4      cli_1.1.0
[17] rstudioapi_0.10  yaml_2.2.0       haven_2.1.0      xfun_0.7
[21] withr_2.1.2      xml2_1.2.0       httr_1.4.0       hms_0.4.2
[25] generics_0.0.2   grid_3.6.0       tidyselect_0.2.5 glue_1.3.1
[29] R6_2.4.0         readxl_1.3.1     rmarkdown_1.12   modelr_0.1.4
[33] magrittr_1.5     backports_1.1.4  scales_1.0.0     htmltools_0.3.6
[37] rvest_0.3.3      assertthat_0.2.1 colorspace_1.4-1 stringi_1.4.3
[41] lazyeval_0.2.2   munsell_0.5.0    broom_0.5.2      crayon_1.3.4
```