Week 1  QLB 408 / 508  Spring 2020

www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

SNP data

Example: Genotype  CC  CT  TT
           Xij  0  1  2

Individuals

SNPs

0 2 2 1 1 0 1
0 2 1 0 1
2 . . . .
Genome-wide association studies

©Pasieka, Science Photo Library

Allele frequencies in human populations

Median differentiated SNP in HGDP data
RNA-Seq - Gene Expression Quantification

Robinson et al. (2015) NAR

![Graph showing gene expression quantification](image-url)
Foundations of Applied Statistics -- Storey -- jdstorey.org/fas
Probability

Probability space \((\Omega, \mathcal{F}, \Pr)\)

\(\Omega\) = set of outcomes, sample space
\(\Pr\) = probability measure

Events \(A \subseteq \Omega\), calculate \(\Pr(A)\)

\(\mathcal{F}\) = \(\sigma\)-algebra, all events \(A\) where \(\Pr(A)\) is meaningful

Examples of \(\Omega\)

\(\Omega = \{TT, HT, TH, HH\}\) coin flips

\(\Omega = \{CC, CT, TT\}\) diploid genotypes

\(\Omega = \{C, T\}\) haploid genotypes

\(\Omega = \mathbb{R}\) stock returns

\(\Omega = [0,\infty)\) height

Mathematical Probability

1. The probability of any event \(A\) is such that \(0 \leq \Pr(A) \leq 1\).
2. \( P(\mathcal{S}) = 1 \)

3. Let \( A^c \) be the complement of \( A \), then \( P(A^c) + P(A) = 1 \).

4. For any \( n \) events such that \( A_i \) \( A_i \cap A_j = \emptyset \quad \forall \ i \neq j \), then
\[
P(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j)
\]
\[
P(A \cup B) = P(A) + P(B) - P(A \cap B)
\]
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}
\]

**Independence**

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

- \( P(A \cap B) = P(A) \cdot P(B) \)
- \( A \) and \( B \) are independent

Events \( A \) and \( B \) are independent if (all equivalent):
- \( \Pr(A \cap B) = \Pr(A) \)
- \( \Pr(A \cap B) = \Pr(B) \)
- \( \Pr(A \cap B) = \Pr(A) \cdot \Pr(B) \)

**Bayes' Theorem**

\[
\Pr(B \mid A) = \frac{\Pr(A \mid B) \cdot \Pr(B)}{\Pr(A)}
\]

\[
\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B) = \Pr(B \mid A) \Pr(A)
\]

**Law of Total Probability**

Events \( A_1, A_2, \ldots, A_n \) such that \( A_i \cap A_j = \emptyset \) for \( i \neq j \) and \( \bigcup A_i = \Omega \), then for any event \( B \),

\[
\Pr(B) = \sum_{i=1}^{n} \Pr(B \mid A_i) \cdot \Pr(A_i)
\]

\[\bigcup_{i=1}^{n} (A_i \cap B) = B \text{ and disjoint} \]
\[ P(B) = \sum_{i=1}^{n} P(A \cap B_i) \]

\[ \gamma = P(B | A_i) P(A_i) \]

Random Variable

A random variable (RV) \( X \) is a function:

\[ X : \Omega \rightarrow \mathbb{R} \]

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

The "range" of \( X \) is:

\[ \mathbb{R} = \{ X(w) : w \in \Omega \} \]

where \( \mathbb{R} = \mathbb{R} \).

Example

\( \Omega = \{ CC, CT, TT \} \) SNP genotypes

\( X(CC) = 0 \)

\( P(X=0) = P(CC \cap \Omega) \)

\( X(CT) = 1 \)

\( X(TT) = 2 \)
Distribution of RV's

cumulative distribution function (cdf):
\[ F(y) = \Pr (X \leq y) \]

Example: \[ F(1) = \Pr(X \leq 1) = \Pr(3CC, Ct3) \]
\[ F(1.1) = F(1) \]

Discrete RV's have a discrete \( \mathbb{R} \)

E.g. \( \mathcal{X} = 0, 1, 2, \ldots, 103 \)
\( \mathcal{X} = 0, 1, 3, 4, \ldots, 3 \)

Continuous RV's have a continuous \( \mathbb{R} \)

E.g. \( \mathcal{X} = [0, 1] \)
\( \mathcal{X} = \mathbb{R} \)
Probability mass or density functions

**Discrete**
Probability mass function (pmf) is
\[ f(x) = P(X = x) \text{ for all } x \in \mathbb{R} \]
\[ f(x) = F(x) - F(b) \text{ as } b \to x \]

**Continuous**
Probability density function (pdf)
\[ f(x) = \frac{d}{dx} F(x) \]
\[
\text{Discrete} \quad F(y) = \sum_{x \leq y} f(x) = \Pr(X \leq y)
\]

\[
\text{Continuous} \quad F(y) = \int_{-\infty}^{y} f(x) \, dx = \Pr(X \leq y)
\]

Note that \( \Pr(X = x) = 0 \)

\underline{Median of a distribution (aka RV)}:

A value \( y \) s.t. \( F(y) = 0.5 \)

\underline{Expected value or "population mean"}:

\[
E(X) = \sum_{x \in \mathbb{R}} x f(x) \quad \text{discrete}
\]

\[
= \int x \, dF(x) \quad \text{for continuous}
\]

\[
= \int x \, dF(x) \, dx \quad \text{for measure theory}
\]
Population Variance

\[ \text{Var}(X) = E \left[ (X - E[X])^2 \right] \]

\[ = \sum (x - E[X])^2 f(x) \quad \text{discrete number} \]

\[ = \int (x - E[X])^2 f(x) \, dx \quad \text{continuous} \]

\[ \text{SD}(X) = \sqrt{\text{Var}(X)} \]

Covariance of rv's X and Y

\[ \text{Cov}(X, Y) = E \left[ (X - E(X))(Y - E(Y)) \right] \]

\[ \text{Var}(X) = \text{Cov}(X, X) \]

\[ \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{SD(X) \cdot SD(Y)} \]

\[ -1 \leq \text{Cor}(X, Y) \leq 1 \]
Discrete rv's

**Uniform**

\[ X \sim \text{Uniform}(1, 2, \ldots, n) \]
\[ R = 1, 2, \ldots, n \]

\[ f(x; n) = \frac{1}{n} \quad \text{for } x \in R \]

"sample" in \( R \)

**Bernoulli**

\[ X \sim \text{Bernoulli}(p) \]
\[ R = 0, 1 \]

\[ f(x; p) = (1-p)^{1-x} \cdot p^x \]

\[ f(0) = (1-p), \quad f(1) = p \]

\[ E[X] = p = 0 \cdot f(0) + 1 \cdot f(1) \]

\[ \text{Var}(X) = p(1-p) \]

**Binomial**

\[ X \sim \text{Binomial}(n, p) \]

sum of \( n \) independent Bernoulli(\( p \))
$R = \{0, 1, \ldots, n\}$

$$f(x; \rho) = \binom{n}{x} \rho^x (1-\rho)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

number ways to choose $x$ from $n$ without order

$E[X] = \nu \rho$, $\text{Var}(X) = \nu \rho (1-\rho)$

Example: Under Hardy-Weinberg equilibrium, $X = \#$ of $T$ alleles,

$X \sim \text{Binomial}(2, \rho)$

where $\rho$ is the allele frequency of $T$.

CC:

$P(X=0) = (1-\rho)^2$

$P(X=1) = 2 \rho (1-\rho)$

$P(X=2) = \rho^2$

Poisson

$X \sim \text{Poisson}(\lambda)$
\( R = \{0, 1, 2, \ldots \} \)

\[ f(x; \lambda) = \frac{x^x e^{-\lambda}}{x!} \]

\[ E[X] = \lambda, \quad \text{Var}(X) = \lambda \]

\[ \pm \begin{array}{c}
\text{d pois} & \rightarrow \text{pmf} \\
p \text{pois} & \rightarrow \text{cdf} \\
q \text{pois} & \rightarrow \text{quantile} \\
r \text{pois} & \rightarrow \text{random draws}
\end{array} \]

? Distributions

Continuous

\begin{align*}
\text{Uniform (0, 1)} & , \quad \text{Normal (\mu, \sigma)} & , \quad \text{Beta (a, b)}
\end{align*}

\[ X \sim \text{Uniform (0, 1)} \]

\[ R = [0, 1] \]

\[ f(x) = 1, \quad x \in [0, 1] \]

\[ F(y) = y, \quad y \in [0, 1] \]
\[ E[X] = \frac{1}{2} \quad V_{01}(X) = \frac{1}{12} \]

Uniform \((0, \theta)\)

\[ f(x; \theta) = \frac{1}{\theta} \quad F(y; \theta) = \frac{y}{\theta} \]

\( \theta = [0, \theta] \)

Beta

\( X \sim \text{Beta}(\alpha, \beta) \quad \alpha, \beta > 0 \)

\( \theta = (0, 1) \)

\[ f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad x \in (0, 1) \]

\[ \int_0^1 f(x; \alpha, \beta) \, dx = 1 \]
$$\Pi(z) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx$$

$$E[X] = \frac{\alpha}{\alpha + \beta} \quad Var(X) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

Beta PDF's