BIOINFORMATICS MSc

PROBABILITY AND STATISTICS FORMULA SHEET

SET THEORY DEFINITIONS AND RESULTS

Events E and F are **mutually exclusive** if $E \cap F = \emptyset$ (the **empty set**).

Events $E_1, ..., E_k$ form a **partition** of event $F \subseteq S$ if

$$(a)\ E_i \cap E_j = \emptyset \text{ for all } i \text{ and } j \qquad \text{(b)} \bigcup_{i=1}^{k} E_i = E_1 \cup E_2 \cup ... \cup E_k = F.$$

THE RULES OF PROBABILITY: For any events E and F in sample space S,

(1) $0 \leq P(E) \leq 1$
(2) $P(S) = 1$
(3) If $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$

Corollaries :
$P(E') = 1 - P(E)$, $P(\emptyset) = 0$

If $E_1, ..., E_k$ are events such that $E_i \cap E_j = \emptyset$ for all i, j, then

$$P\left(\bigcup_{i=1}^{k} E_i\right) = P(E_1) + P(E_2) + ... + P(E_k)$$

.
If $E \cap F \neq \emptyset$, then $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

CONDITIONAL PROBABILITY

$P(E|F)$ is the probability that the event E occurs, given that F **has** occurred, for an event F such that P$(F) > 0$, and

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

The probability of the **intersection** of events $E_1, ..., E_k$ is given by the **chain rule**

$$P(E_1 \cap ... \cap E_k) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)...P(E_k|E_1 \cap E_2 \cap ... \cap E_{k-1})$$

Events F and F are **independent** if

$$P(E|F) = P(E) \quad \text{so that} \quad P(E \cap F) = P(E)P(F).$$

THEOREM OF TOTAL PROBABILITY : If events $E_1, ..., E_k$ form a partition of event $E \subseteq S$

$$P(E) = \sum_{i=1}^{k} P(E|E_i)P(E_i)$$

BAYES THEOREM: If events $E_1, ..., E_k$ form a partition of event $E \subseteq S$,

$$P(E_i|E) = \frac{P(E|E_i)P(E_i)}{P(E)} = \frac{P(E|E_i)P(E_i)}{\sum_{j=1}^{k} P(E|E_j)P(E_j)}$$

DISCRETE PROBABILITY DISTRIBUTIONS

The probability distribution of a *discrete* random variable X is described by the **probability mass function** f_X, specified by

$$f_X(x) = P[X = x] \qquad x \in \mathbb{X} = \{x_1, x_2, ..., x_n, ...\}$$

- Properties of the mass function :

$$\text{(i) } f_X(x_i) \geq 0 \qquad \text{(ii)} \sum_i f_X(x_i) = 1$$

- The cumulative distribution function or c.d.f., F_X, is defined by

$$F_X(x) = P[X \leq x] \qquad x \in \mathbb{R}$$

- Fundamental relationship between f_X and F_X :

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) \qquad \begin{aligned} f_X(x_1) &= F_X(x_1) \\ f_X(x_i) &= F_X(x_i) - F_X(x_{i-1}) \quad \text{for } i \geq 2 \end{aligned}$$

CONTINUOUS PROBABILITY DISTRIBUTIONS:

The probability distribution of a *continuous* random variable X is defined by the continuous **cumulative distribution function** or **c.d.f.**, F_X, specified by

$$F_X(x) = P[X \leq x] \qquad \text{for } x \in \mathbb{X}$$

- The **probability density function**, or **p.d.f.**, f_X, is defined by

$$f_X(x) = \frac{d}{dx}\{F_X(x)\} \quad \text{so that} \quad F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

- **Properties of the density function**

$$\text{(i) } f_X(x) \geq 0 \quad x \in \mathbb{X} \qquad \text{(ii)} \int_{\mathbb{X}} f_X(x) dx = 1.$$

EXPECTATION AND VARIANCE

For a **discrete** random variable X taking values in set $\mathbb{X}$ with mass function f_X, the **expectation** of X is defined by

$$E_{f_X}[X] = \sum_{x \in \mathbb{X}} x f_X(x)$$

For a **continuous** random variable X taking values in interval $\mathbb{X}$ with pdf f_X, the expectation of X is defined by

$$E_{f_X}[X] = \int_{\mathbb{X}} x f_X(x)\, dx.$$

The **variance** of X is defined by

$$Var_{f_X}[(X - E_{f_X}[X])^2] = E_{f_X}[X^2] - \{E_{f_X}[X]\}^2.$$

DISCRETE PROBABILITY DISTRIBUTIONS

The Bernoulli Distribution $X \sim Bernoulli(\theta)$
Range : $\mathbb{X} = \{0, 1\}$
Parameter : $\theta \in [0, 1]$
Mass function :

$$f_X(x) = \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1\}$$

The Binomial Distribution $X \sim Binomial(n, \theta)$
Range : $\mathbb{X} = \{0, 1, ..., n\}$
Parameters : $n \in Z^+$, $\theta \in [0, 1]$
Mass function :

$$f_X(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \qquad x \in \{0, 1, ..., n\}$$

The Geometric Distribution $X \sim Geometric(\theta)$
Range : $\mathbb{X} = \{1, 2, ...\}$
Parameter : $\theta \in (0, 1]$
Mass function :

$$f_X(x) = (1-\theta)^{x-1} \theta \qquad x \in \{1, 2, ...\}$$

Distribution function

$$F_X(x) = 1 - (1-\theta)^x \qquad x \in \{1, 2, ...\}$$

The Negative Binomial Distribution $X \sim NegBin(n, \theta)$
Range : $\mathbb{X} = \{n, n+1, n+2, ...\}$
Parameter : $n \in Z^+$, $\theta \in (0, 1]$
Mass function :

$$f_X(x) = \binom{x-1}{n-1} \theta^n (1-\theta)^{x-n} \qquad x \in \{n, n+1, n+2, ...\}\,.$$

The Poisson Distribution $X \sim Poisson(\lambda)$
Range : $\mathbb{X} = \{0, 1, 2, ...\}$
Parameter : $\lambda \in \mathbb{R}^+$
Mass function :

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x \in \{0, 1, 2, ...\}$$

CONTINUOUS PROBABILITY DISTRIBUTIONS

The Exponential Distribution $X \sim Exponential(\lambda)$
Range : $\mathbb{X} = \mathbb{R}^+$
Parameter : $\lambda > 0$
Density function :

$$f_X(x) = \lambda e^{-\lambda x} \qquad x \geq 0$$

Distribution function:

$$f_X(x) = 1 - e^{-\lambda x} \qquad x \geq 0$$

The Gamma Distribution $X \sim Gamma(\alpha, \beta)$
Range : $\mathbb{X} = \mathbb{R}^+$
Parameters : $\alpha, \beta > 0$
Density function :

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \geq 0 \qquad \text{where} \qquad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt \qquad \alpha > 0.$$

If $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, so if $\alpha = 1, 2, ...,$ $\Gamma(\alpha) = (\alpha - 1)!$.
If $\alpha = 1, 2, ...,$ then the $Gamma(\alpha/2, 1/2)$ distribution is known as the **Chi-squared distribution** with α **degrees of freedom**, denoted χ^2_α.
If $X_1, X_2 \sim Exponential(\lambda)$ are independent, then $Y = X_1 + X_2 \sim Gamma(2, \lambda)$.

The Normal Distribution $X \sim N(\mu, \sigma^2)$
Range : $\mathbb{X} = \mathbb{R}$
Parameters : $-\infty < \mu < \infty, \sigma > 0$
Density function :

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad -\infty < x < \infty$$

If $\mu = 0, \sigma = 1$, then $Y \sim N(0, 1)$ has a **standard** normal distribution
If $X \sim N(0, 1)$, and $Y = \sigma X + \mu$, then $Y \sim N(\mu, \sigma^2)$
If $X \sim N(0, 1)$, and $Y = X^2$, then $Y \sim Gamma(1/2, 1/2) = \chi^2_1$.
If $X \sim N(0, 1)$ and $Y \sim \chi^2_\alpha$ are independent random variables, then random variable $T = X/\sqrt{Y/\alpha}$ has a **t distribution** with α **degrees of freedom**.

THE POISSON PROCESS

In the Poisson process model for events that occur at random in continuous time with constant rate λ, there are three related probability distribution results

- the numbers of events occurring in disjoint intervals of lengths $t_1, t_2, t_3, ...$ are independent random variables $X_1, X_2, X_3, ...$ with $X_i \sim Poisson\,(\lambda t_i)$
- the times between the occurrences of events are independent continuous random variables $T_1, T_2, T_3, ...$ with $T_i \sim Exponential\,(\lambda)$
- the time of the nth event is a continuous random variable Y_n with $Y_n \sim Gamma\,(n, \lambda)$

THE CENTRAL LIMIT THEOREM

THEOREM: Suppose $X_1, ..., X_n$ are i.i.d. random variables with $E_{f_X}[X_i] = \mu$, $Var_{f_X}[X_i] = \sigma^2$. If Z_n is defined by

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$$

Then, as $n \longrightarrow \infty$, $Z_n \longrightarrow Z \sim N(0,1)$ **irrespective** of the distribution of $X_1, ..., X_n$.

MAXIMUM LIKELIHOOD INFERENCE

Suppose a sample $x_1, ..., x_n$ has been obtained from a probability model specified by mass or density function $f(x;\theta)$ depending on parameter(s) θ lying in parameter space Θ. The **maximum likelihood estimate** or **m.l.e.** is produced as follows;

STEP 1 Write down the **likelihood function**

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

STEP 2 Take the natural log of the likelihood, and collect terms involving θ.

STEP 3 Find the value of θ, $\hat{\theta}$, for which $log L(\theta)$ is maximized in Θ.

STEP 4 Verify that $\hat{\theta}$ maximizes $log L(\theta)$.

SAMPLING DISTRIBUTIONS

THEOREM If $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then if

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are the **mean**, **variance**, and **adjusted variance**, then it can be shown that

$$(1) \quad : \quad \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$(2) \quad : \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

(3) : $\bar{X}$ and s^2 are statistically independent.

HYPOTHESIS TESTING FOR NORMAL DATA

ONE-SAMPLE TESTS

Suppose $x_1, ..., x_n \sim N\left(\mu, \sigma^2\right)$, with observed sample mean and adjusted variance $\bar{x}, s^2$. To test the **hypothesis**

$$H_0 : \mu = c$$
$$H_1 : \mu \neq c$$

if σ is known, use the **Z-test**

$$z = \frac{\bar{x} - c}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{if } H_0 \text{ is TRUE.}$$

If σ is unknown, use the **T-test**

$$t = \frac{(\bar{x} - c)}{s/\sqrt{n}} \sim Student(n-1) \qquad \text{if } H_0 \text{ is TRUE}$$

where t_{n-1} is the $Student\,(n-1)$ distribution.

To test $H_0 : \sigma^2 = c$, calculate test statistic q

$$q = \frac{(n-1)s^2}{c} \sim \chi^2_{n-1} \qquad \text{if } H_0 \text{ is TRUE}$$

TWO-SAMPLE TESTS

For two data samples of size n_1 and n_2, where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, and s_1^2 and s_2^2 are the adjusted sample variances; to test the hypothesis

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

if $\sigma_1 = \sigma_2 = \sigma$ is **known** use the statistic z, defined by

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1) \qquad \text{if } H_0 \text{ is TRUE}$$

If $\sigma_1 = \sigma_2 = \sigma$ is **unknown**, use the statistic t, defined by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_P\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad \text{if } H_0 \text{ is TRUE}$$

where $s_P^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)$ is the **pooled** estimate of σ^2.

To test the hypothesis $H_0 : \sigma_1 = \sigma_2$, use the F statistic

$$F = \frac{s_1^2}{s_2^2} \sim Fisher\,(n_1 - 1, n_2 - 1) \qquad \text{if } H_0 \text{ is TRUE}$$

95 % CONFIDENCE INTERVALS FOR PARAMETERS

Let $t_k(p)$ be the pth percentile of a Student t distribution with k degrees of freedom.

ONE-SAMPLE: 95 % Confidence interval for μ is

$$\begin{array}{lll} \bar{x} & \pm & 1.96\sigma/\sqrt{n} \quad \text{if } \sigma \text{ is known} \\ \bar{x} & \pm & t_{n-1}(0.975)s/\sqrt{n} \quad \text{if } \sigma \text{ is unknown} \end{array}$$

95 % Confidence interval for σ^2 is

$$\left[(n-1)s^2/c_2 : (n-1)s^2/c_1\right]$$

where c_1 and c_2 are the 0.025 and 0.975 points of the χ^2_{n-1} distribution.

TWO-SAMPLE: 95 % Confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x_1} - \bar{x_2} \pm 1.96\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{if } \sigma \text{ is known}$$

$$\bar{x_1} - \bar{x_2} \pm t_{n_1+n_2-2}(0.975)s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{if } \sigma \text{ is unknown}$$

95 % Confidence interval for σ_1^2/σ_2^2 is

$$\left[\frac{s_1^2}{(c_2 s_2^2)} : \frac{s_1^2}{(c_1 s_2^2)}\right]$$

where c_1 and c_2 are the 0.025 and 0.975 points of the $Fisher\,(n_1-1, n_2-1)$ distribution.

THE CHI-SQUARED AND LIKELIHOOD RATIO TEST

To test the goodness-of-fit of a probability model to a sample of size n, use the **chi-squared statistic**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

If H_0 is true, then χ^2 approximately has a with $k-d-1$ degrees of freedom, where d is the number of estimated parameters.

For a contingency table with r rows and c columns, the χ^2 statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

for a test of independence has a null distribution that is chi-squared with $(r-1)\times(c-1)$ degrees of freedom, where

$$\hat{n}_{ij} = n_{i.}\hat{p}_j = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, ..., r,\ j = 1, ..., c$$

and $n_{i.}$ is the total of the ith row, $n_{.j}$ is the total of the jth column, and n is the total number of observations.

The Likelihood Ratio statistic LR has the same approximate null distribution, and is defined by

$$LR = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\frac{n_{ij}}{\hat{n}_{ij}}$$

CLASSIFICATION FOR TWO CLASSES ($K = 2$)

Let $f_1(x)$ and $f_2(x)$ be the probability functions associated with a (vector) random variable X for two populations 1 and 2. An object with measurements x must be assigned to either class 1 or class 2. Let $\mathbb{X}$ denote the sample space. Let $\mathcal{R}_1$ be that set of x values for which we classify objects into class 1 and $\mathcal{R}_2 \equiv \mathbb{X}\backslash\mathcal{R}_1$ be the remaining x values, for which we classify objects into class 2.

The **conditional probability**, $P(2|1)$, of classifying an object into class 2 when, in fact, it is from class1 is:

$$P(2|1) = \int_{\mathcal{R}_2} f_1(x)\ dx.$$

Similarly, the conditional probability, $P(1|2)$, of classifying an object into class 1 when, in fact, it is from class 2 is:

$$P(1|2) = \int_{\mathcal{R}_1} f_2(x)\ dx$$

Let p_1 be the *prior* probability of being in class 1 and p_2 be the *prior* probability of 2, where $p_1 + p_2 = 1$. Then,

$$\begin{aligned} P\left(\text{Object correctly classified as class 1}\right) &= P(1|1)p_1 \\ P\left(\text{Object misclassified as class 1}\right) &= P(1|2)p_2 \\ P\left(\text{Object correctly classified as class 2}\right) &= P(2|2)p_2 \\ P\left(\text{Object misclassified as class 2}\right) &= P(2|1)p_1 \end{aligned}$$

Now suppose that the *costs* of misclassification of a class 2 object as a class 1 object, and vice versa are, respectively.$c(1|2)$ and $c(2|1)$. Then the expected cost of misclassification is therefore

$$c(2|1)P(2|1)p_1 + c(1|2)\,P(1|2)p_2.$$

The idea is to choose the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ so that this expected cost is minimized. This can be achieved by comparing the predictive probability density functions at each point x

$$\mathcal{R}_1 \equiv \left\{x : \frac{f_1(x)}{f_2(x)}\frac{p_1}{p_2} \geq \frac{c(1|2)}{c(2|1)}\right\} \qquad \mathcal{R}_2 \equiv \left\{x : \frac{f_1(x)}{f_2(x)}\frac{p_1}{p_2} < \frac{c(1|2)}{c(2|1)}\right\}$$

If $p_1 = p_2$, then

$$\mathcal{R}_1 \equiv \left\{x : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}\right\}$$

and if $c(1|2) = c(2|1)$, equivalently

$$\mathcal{R}_1 \equiv \left\{x : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}\right\}$$

and finally if $p_1 = p_2$ and $c(1|2) = c(2|1)$ then

$$\mathcal{R}_1 \equiv \left\{x : \frac{f_1(x)}{f_2(x)} \geq 1\right\} \equiv \{x : f_1(x) \geq f_2(x)\}$$