CHEM. ENG. II - PROBABILITY AND STATISTICS

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**).

For events E, F, and G, the following results hold;

ASSOCIATIVITY	$(E \cup F) \cup G = E \cup (F \cup G)$
	$(E \cap F) \cap G = E \cap (F \cap G)$
DISTRIBUTIVITY	$E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$
	$E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$
also	$(E \cup F)' = E' \cap F'$, $(E \cap F)' = E' \cup F'$

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

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

The rules of probability : For any events E and F in sample space S,

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

Corollaries :
$\mathrm{P}(E') = 1 - \mathrm{P}(E)$, $\mathrm{P}(\emptyset) = 0$
If $E_1, ..., E_k$ are events such that $E_i \cap E_j = \emptyset$ for all i, j, then

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

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

Conditional probability :

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

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

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

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

Events F and F are **independent** if

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

Theorem of Total Probability :

If events $E_1, ..., E_k$ form a partition of event $E \subseteq S$, then $\mathrm{P}(E) = \sum_{i=1}^{k} \mathrm{P}(E|E_i)\mathrm{P}(E_i)$

Bayes Theorem :

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

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

Discrete probability distributions:

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

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

Properties of the mass function :

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

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

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

Fundamental relationship between f_X and F_X :

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) \ ,$$

and

$$f_X(x_1) = F_X(x_1)$$
$$f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \quad \text{for } i \geq 2$$

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

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

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 :

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

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

$$\mathrm{E}_{f_X}[X] = \sum_{x \in \mathbb{X}} x f_X(x)$$

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

$$\mathrm{E}_{f_X}[X] = \int_{\mathbb{X}} x f_X(x)\ dx.$$

The **variance** of X is defined by

$$\mathrm{E}_{f_X}[(X - \mathrm{E}_{f_X}[X])^2] = \mathrm{E}_{f_X}[X^2] - \{\mathrm{E}_{f_X}[X]\}^2.$$

Special 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 :

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

The Geometric Distribution $X \sim Geometric(\theta)$

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

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

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 :

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

The Poisson Distribution $X \sim Poisson(\lambda)$

Range : $\mathbb{X} = \{0, 1, 2, ...\}$
Parameter : $\lambda \in \mathbb{R}^+$
Mass function :

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

Special Continuous Probability Distributions

The Exponential Distribution $X \sim Exponential(\theta)$

Range : $\mathbb{X} = \mathbb{R}^+$
Parameter : $\lambda \in \mathbb{R}^+$
Density function :

$$f_X(x) = \lambda e^{-\lambda x} \qquad x \in \mathbb{R}^+$$

The Gamma Distribution $X \sim Gamma(\alpha, \beta)$

Range : $\mathbb{X} = \mathbb{R}^+$
Parameters : $\alpha, \beta \in \mathbb{R}^+$
Density function :

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in \mathbb{R}^+$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt \qquad \alpha > 0.$$

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 Beta Distribution $X \sim Beta(\alpha, \beta)$

Range : $\mathbb{X} = (0, 1)$
Parameters : $\alpha, \beta \in \mathbb{R}^+$
Density function :

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad x \in (0, 1).$$

The Normal Distribution $X \sim N(\mu, \sigma^2)$

Range : $\mathbb{X} = \mathbb{R}$
Parameters : $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$
Density function :

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad x \in \mathrm{R}.$$

Notes :
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 Convolution Theorem

If X_1 and X_2 are discrete independent random variables with probability mass/density functions f_{X_1} and f_{X_2} respectively, then random variable Y, defined by $Y = X_1 + X_2$, has probability mass/density function given by

$$f_Y(y) = \begin{cases} \displaystyle\sum_{x_1=-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1) & \text{DISCRETE} \\[2ex] \displaystyle\int_{-\infty}^{\infty} f_{X_1}(x_1) f_{X_2}(y - x_1)\, dx_1 & \text{CONTINUOUS} \end{cases}$$

Note : Terms in the sum/integral may be zero on intervals of $\mathbb{R}$.

The Central Limit Theorem

THEOREM

Suppose $X_1, ..., X_n$ are i.i.d. random variables with $\mathrm{E}_{f_X}[X_i] = \mu$, $\mathrm{Var}_{f_X}[X_i] = \sigma^2$. If Z_n is defined by

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

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

Maximum Likelihood Estimation

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 $logL(\theta)$ is maximized in Θ.

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

Sampling Distributions

THEOREM

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

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

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

$$(1)\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad (2)\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}, \quad (3)\bar{X} \text{ and } s^2 \text{ are statistically independent.}$$

Hypothesis Testing for Normal data

One-sample tests

Suppose $x_1, ..., x_n \sim N(\mu, \sigma^2)$, 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**

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

If σ is unknown, use the **T-test**

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \sim t_{n-1} \quad \text{if } H_0 \text{ is TRUE}$$

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

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

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

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

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

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

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

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

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.

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

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} \quad \text{if } H_0 \text{ is TRUE}$$

95 % Confidence Intervals for Parameters

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

One-sample: 95 % Confidence interval for μ is

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

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

$$\begin{array}{ll} \bar{x_1} - \bar{x_2} \pm 1.96\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} & \text{if } \sigma \text{ is known} \\ \bar{x_1} - \bar{x_2} \pm t_{n_1+n_2-2}(0.975)\; s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} & \text{if } \sigma \text{ is unknown} \end{array}$$

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

$$\left[s_1^2/(c_2 s_2^2) : s_1^2/(c_1 s_2^2)\right]$$

where c_1 and c_2 are the 0.025 and 0.975 points of the F_{n_1-1,n_2-1} distribution.

The Chi-squared 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 chi-squared distribution with $k-d-1$ degrees of freedom, where d is the number of estimated parameters.

Linear Regression Analysis

Suppose that we have n measurements of two variables X and Y, denoted $\{(x_i, y_i) : i = 1, ..., n\}$, and there is a **linear regression** relationship between X and Y,

$$\mathrm{E}[Y|X = x] = \alpha + \beta x.$$

Then the least-squares estimates of α and β are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2}$$

where

$$S_x = \sum_{i=1}^{n} x_i \quad S_y = \sum_{i=1}^{n} y_i \quad S_{xx} = \sum_{i=1}^{n} x_i^2 \quad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Note : the correlation coefficient r is given by

$$r = \frac{nS_{xy} - S_x S_y}{\sqrt{(nS_{xx} - S_x^2)(nS_{yy} - S_y^2)}}$$

Estimates of Error Variance and Residuals

The maximum likelihood estimate of σ^2 is,

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = S^2$$

The **corrected** estimate, s^2, of the error variance is defined by

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ is the **fitted value** of Y at $X = x_i$. The ith **residual**, e_i is given by $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$.

Standard Errors of Estimators

$$s.e.(\hat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} \qquad s.e.(\hat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}$$

Confidence Intervals for Parameters

$$\alpha \;:\; \hat{\alpha} \pm t_{n-2}(0.975)\, s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}}$$

$$\beta \;:\; \hat{\beta} \pm t_{n-2}(0.975)\, s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}$$

where $t_{n-2}(0.975)$ is the 97.5th percentile of a t distribution with $n - 2$ degrees of freedom.