STAT512 and STAT513

Fall/Winter, 1995-6, 1996-7, and 1997-8

E. A. Thompson,

Department of Statistics,
University of Washington,
Box 354322, Seattle, WA 98195.

Chapter 2 and Chapter 3 only.

Chapter 2: More about Random Variables

2.1 Covariance and conditional expectation

2.1.1 Covariance and correlation (C&B 4.5)

Variance: $\text{var}(X) = \text{E}((X - \text{E}(X))^2) \;=\; \text{E}(X^2) - (\text{E}(X))^2$

Covariance: $\text{cov}(X,Y) \;=\; \text{E}((X - \text{E}(X))(Y - \text{E}(Y))) \;=\; \text{E}(XY) \;-\; \text{E}(X)\text{E}(Y)$

Properties:
(i) X, Y independent $\Longrightarrow$ $\text{cov}(X,Y) \;=\; 0$, but in general not conversely.
(ii) If (X,Y) is bivariate Normal, $\text{cov}(X,Y) \;=\; 0 \;\Longrightarrow\; X,\; Y$ independent (see 1.3.2, 3.1.2).
(iii) $\text{cov}(g_1(X), g_2(Y)) \;=\; 0$ for all functions $g_1()$ and $g_2()$, then $X,\; Y$ are independent.
(iv) $\text{var}(aX + bY) \;=\; a^2\ \text{var}(X) + 2a\ b\ \text{cov}(X,Y) + b^2\ \text{var}(Y)$.

Correlation: $\rho(X,Y) \;=\; \text{cov}(X,Y)/\sqrt{\text{var}(X)\text{var}(Y)}$

Note: For all real t,

$$\begin{aligned} 0 \le h(t) &= \text{E}(((X - \text{E}(X))\ t + (Y - \text{E}(Y)))^2) \\ &= t^2\ \text{var}(X) \;+\; 2t\ \text{cov}(X,Y) \;+\; \text{var}(Y) \end{aligned}$$

so $[2\text{cov}(X,Y)]^2 \le 4\text{var}(X)\text{var}(Y)$ or $-1 \le \rho(X,Y) \le 1$, with equality *iff*
$(X - \text{E}(X))t + (Y - \text{E}(Y)) \equiv 0$ for some t. That is, $Y = aX + b$ for some $a,\ b$.

2.1.2 Conditional expectation (C&B 4.4)

$\text{E}(X)$ is the best predictor of the value of X in the sense of minimising $\text{E}((X - m)^2)$ (see 1.2.2).
Similarly, given $Y = y$, $m(y) = \text{E}(X \mid Y = y)$ minimises $\text{E}((X - m(Y))^2 \mid Y = y)$
or $m(Y) = \text{E}(X \mid Y)$ minimises $\text{E}((X - m(Y))^2 \mid Y)$.
That is $\text{E}(X \mid Y)$ is the best predictor of X based on Y.

Properties:
(i) $\text{E}(X) \;=\; \text{E}(\text{E}(X \mid Y))$ since

$$\int_x \int_y x f_{X,Y}(x,y)dxdy \;\;=\;\; \int_y \left[\int_x x f_{X|Y=y}(x)dx\right] f_Y(y)dy$$

(ii) $\text{var}(X) \;=\; \text{E}(\text{var}(X \mid Y)) \;+\; \text{var}(\text{E}(X \mid Y))$ since

$$\begin{aligned} \text{E}(X^2) &= \text{E}(\text{E}(X^2 \mid Y)) \;=\; \text{E}(\text{var}(X \mid Y) \;+\; (\text{E}(X \mid Y))^2) \\ &= \text{E}(\text{var}(X \mid Y)) \;+\; \text{var}(\text{E}(X \mid Y)) \;+\; (\text{E}(\text{E}(X \mid Y)))^2 \\ &= \text{E}(\text{var}(X \mid Y)) \;+\; \text{var}(\text{E}(X \mid Y)) \;+\; (\text{E}(X))^2 \end{aligned}$$

2.1.3 Examples

(i) Given $N = n$, $T \equiv$ *number of recombinants* $\sim B(n,\theta)$. Suppose $N \sim \mathcal{P}(\lambda)$. Then

$$\text{E}(T \mid N) = N\theta, \quad \text{var}(T \mid N) \;=\; N\theta(1-\theta). \;\text{So}\; \text{E}(T) \;=\; \text{E}(N\theta) \;=\; \theta\ \text{E}(N) \;=\; \lambda\theta$$

$$\text{and}\;\; \text{var}(T) \;=\; \text{E}(N\theta(1-\theta))\ \text{var}(N\theta) \;=\; \lambda\theta(1-\theta) \;+\; \theta^2\lambda \;=\; \lambda\theta$$

In fact, $T \sim \mathcal{P}(\lambda\theta)$ (see 2.2.3, 2.3.9).

(ii) Given $N = n$, $T = \sum_{i=1}^{n} X_i$ where X_i are independent, each with mean μ and variance σ^2. Then

$$\mathrm{E}(T) = \mathrm{E}(N\mu) = \mu\mathrm{E}(N), \text{ and } \mathrm{var}(T) = \mathrm{E}(N\sigma^2) + \mathrm{var}(N\mu) = \sigma^2\,\mathrm{E}(N) + \mu^2\,\mathrm{var}(N)$$

(The result for the mean does not require that the X_i are independent.)

2.2 Moment Generating functions (C&B 2.3, 4.6)

2.2.1 Definitions and Theorems

Univariate case: The m.g.f of a real-values r.v X is $m_X(t) = \mathrm{E}(e^{tX})$, $-\infty < t < +\infty$. Note $m_X(0) = 1$. For $t \neq 0$, $m_X(t)$ may not exist (i.e. be $+\infty$.).

Multivariate case: The m.g.f. of vector random variable $\mathbf{X} = (X_1, ..., X_k)$ is

$$m_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}(e^{\mathbf{t}'\mathbf{X}}) = \mathrm{E}(\exp(\sum_{1}^{k} t_i X_i))$$

Two important facts: (actually theorems, but not proved here)
(i) **Uniqueness:** If there exists $\epsilon > 0$ s.t. $m_{\mathbf{X}}(\mathbf{t})$ exists for $|\mathbf{t}| < \epsilon$, then

$$m_{\mathbf{X}}(\mathbf{t}) = m_{\mathbf{Y}}(\mathbf{t}) \text{ for all } \mathbf{t} \text{ with } |\mathbf{t}| < \epsilon, \text{ iff } F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{Y}}(\mathbf{x}) \text{ for all } \mathbf{x}$$

(ii) **Convergence:** If there exists $\epsilon > 0$ s.t. $m_{\mathbf{X}_n}(\mathbf{t})$ exists for $|\mathbf{t}| < \epsilon$, for all n, then

$$m_{\mathbf{X}_n}(\mathbf{t}) \longrightarrow m_{\mathbf{X}}(\mathbf{t}) \text{ for all } \mathbf{t} \text{ with } |\mathbf{t}| < \epsilon, \text{ iff } F_{\mathbf{X}_n}(\mathbf{x}) \longrightarrow F_{\mathbf{X}}(\mathbf{x})$$

for all $\mathbf{x}$ at which $F_{\mathbf{X}}(\mathbf{x})$ is continuous.

2.2.2 Properties

(i) Assuming the expectations exist:

$$m_X^{(r)}(0) = \frac{\partial^r}{\partial t^r} m_X(t) \Big|_{t=0} = \mathrm{E}(X^r)$$

(ii) $\mathbf{X} = (X_1, ..., X_k)'$, $\mathbf{A}$ is $l \times k$, $\mathbf{t} = (t_1, ..., t_l)'$, $\mathbf{b} = (b_1, ..., b_l)'$.

$$m_{aX+b}(t) = \mathrm{E}(\exp((aX+b)t)) = \exp(bt)\mathrm{E}(\exp(atX)) = \exp(bt)\; m_X(at)$$

$$m_{\mathbf{AX}+\mathbf{b}}(\mathbf{t}) = \mathrm{E}(\exp(\mathbf{t}'(\mathbf{AX}+\mathbf{b}))) = \exp(\mathbf{t}'\mathbf{b})\mathrm{E}(\mathbf{t}'\mathbf{AX}) = \exp(\mathbf{t}'\mathbf{b})m_{\mathbf{X}}(\mathbf{A}'\mathbf{t})$$

(iii) If $\mathbf{X}_i$ are independent, and $\mathbf{Y} = \sum_{i=1}^{n} \mathbf{X}_i$ (vector or scalar)

$$m_{\mathbf{Y}}(\mathbf{t}) = \mathrm{E}(\exp(\mathbf{t}'\mathbf{Y})) = \mathrm{E}(\exp(\sum_{1}^{n} \mathbf{t}'\mathbf{X}_i)) = \mathrm{E}(\prod_{1}^{n} \exp(\mathbf{t}'\mathbf{X}_i)) = \prod_{1}^{n} \mathrm{E}(\exp(\mathbf{t}'\mathbf{X}_i)) = \prod_{1}^{n} m_{\mathbf{X}_i}(\mathbf{t})$$

(iv) From (ii) and (iii), if X_i are i.i.d, each with m.g.f $m_X(t)$,

$$\text{if } Y = \sum_{1}^{n} X_i, \quad m_Y(t) = (m_X(t))^n$$

$$\text{if } Y = \bar{X} = \frac{1}{n}\sum_{1}^{n} X_i, \quad m_Y(t) = \mathrm{E}(\exp(\sum(t\; X_i/n)))$$

$$= \mathrm{E}(\prod_{1}^{n} \exp(t\; X_i/n)) = \prod_{1}^{n} \mathrm{E}(\exp(t\; X_i/n)) = (m_X(t/n))^n$$

$$\text{if } Y = (\bar{X} - \mu)/\sqrt{n}\sigma, \quad m_Y(t) = exp(-\mu t/\sqrt{n}\sigma) m_{\bar{X}}(t/\sqrt{n}\sigma))$$

$$= exp(-\mu t/\sqrt{n}\sigma)(m_X(t/(n\sqrt{n}\sigma)))^n$$

(v) If $Y = \sum_1^N X_i$, where X_i are i.i.d. and N is a r.v.

$$\begin{aligned} m_Y(t) &= \mathrm{E}(\exp(t\sum_1^N X_i)) = \mathrm{E}(\mathrm{E}(\exp(t\sum_1^N X_i)) \mid N) \\ &= \mathrm{E}((m_X(t))^N) = \mathrm{E}(\exp(N \log(m_X(t)))) = m_N(\log(m_X(t))) \end{aligned}$$

(vi) In fact, $X_1, ..., X_k$ are jointly independent if and only if

$$m_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}(\exp(\mathbf{t}'\mathbf{X})) = \prod_1^k m_{X_i}(t_i)$$

The proof follows from the uniqueness of m.g.fs.

2.2.3 Examples

(i) Scale parameters in *Gammas* and *exponentials*

If $X \sim G(\alpha, \beta)$, $f_X(x) = \frac{1}{\Gamma(\alpha)}\beta^{-\alpha}x^{\alpha-1}e^{-x/\beta}$ on $0 < x < \infty$.
Direct integration gives $m_X(t) = (1-\beta t)^{-\alpha}$.
So $m_{X/\beta}(t) = m_X(t/\beta) = (1-t)^{-\alpha}$. That is, $X/\beta \sim G(\alpha, 1)$.
Similarly for exponentials: $\mathcal{E}(\beta) \equiv G(1, \beta)$. Here, β is a **scale parameter**.
Also, if X_i is $G(\alpha_i, \beta)$, independent. Then $\sum X_i \sim G(\sum \alpha_i, \beta)$.

(ii) The *Poisson-Bionomial* hierarchy

If $X \sim B(n, \theta)$, $m_X(t) = (1-\theta+\theta e^t)^n$.
If $Y \sim \mathcal{P}(\lambda)$. $m_Y(t) = \exp(\theta(e^t - 1))$.
If $(X|Y) \sim B(Y, \theta)$, and $Y \sim \mathcal{P}(\lambda)$,

$$\begin{aligned} m_X(t) &= \mathrm{E}(\mathrm{E}(e^{tX} \mid Y)) = \mathrm{E}((1-\theta+\theta e^t)^Y) = m_Y(\log(1-\theta+\theta e^t)) \\ &= \exp(\lambda((1-\theta+\theta e^t)-1)) = \exp(\lambda\theta(e^t-1)) \end{aligned}$$

That is $X \sim \mathcal{P}(\lambda\theta)$.

2.3 The Poisson Process

2.3.1 Definition

Events occur "independently and at random" (in time or space) at a given rate λ. That is, the probability that one event occurs in a small time (or space) interval δx is $\lambda\delta x + o(\delta x)$
and the probability of more than one event is $o(\delta x)$. Occurences of events in disjoint intervals are independent.

2.3.2 An example from genetics

A chromosome which ends up in a *gamete* (sperm or egg cell) is formed by copying the DNA from the pair of corresponding chromosomes that are in every (non-gamete) cell of the parent. One chromosome of the pair is the parent's maternal chromosome. The other is the parent's paternal chromosome. In forming the gamete chromosome, at random points the copying switches from one parental chromosome to the other; these points are known as *crossovers*. (Note this is a description of the outcome, not of the actual biological process of meiosis.)

Haldane (1919) proposed to model the occurrence of crossovers along the chromosome as a Poisson process. This is not exactly so, but it is quite an accurate model for most purposes. He defined the (additive) measure of genetic distance between two points on the chromosome as the expected

number of crossovers between them. He defined the Morgan as the distance over which 1 crossover is expected, so if we measure genetic distance in Morgans we have a Poisson process rate 1.

2.3.3 Waiting times are exponential $\mathcal{E}(\frac{1}{\lambda})$

Suppose we have a Poisson process rate λ, let X be the time (or distance) to the next event, and let $H(x) = P(X > x) = 1 - F_X(x)$. Then

$$H(x + \delta x) = H(x)(1 - \lambda \delta x) + o(\delta x)$$

$$\text{or } \frac{dH}{dx} = -\lambda H \quad \text{or } H(x) = \exp(-\lambda x) \quad (H(0) = 1)$$

so $F_X(x) = 1 - \exp(-\lambda x)$, $f_X(x) = \lambda \exp(-\lambda x)$, or $X \sim \mathcal{E}(1/\lambda)$

2.3.4 The forgetting property of exponentials

This follows directly from the cdf:
if $X \sim \mathcal{E}(1/\lambda)$, $H(x) = P(X > x) = \exp(-\lambda x)$, so $H(x + c) = H(x)H(c)$ or
$P(X > x + c \mid X > c) = P(X > x)$.
However, the result also makes sense from the definition of the Poisson process; the occurrences (or not) of events in disjoint time intervals are independent.

2.3.5 The time to n^{th} event is Gamma

The waiting time to the next event is *exponential*. The time to the next is independent of past occurrences. The sum of i.i.d exponentials is *gamma*. So the total time to the n^{th} event is $G(n, 1/\lambda)$.

2.3.6 The number of events in interval T is Poisson mean λT

Let $P_n(x) = P(Y = n)$ be the probability of n events in interval length x in a Poisson process rate λ. Then

$$P_n(x + \delta x) = (1 - \lambda \delta x) P_n(x) + \lambda \delta x P_{n-1}(x) + o(\delta x)$$

$$\text{or} \quad \frac{dP_n}{dx} = -\lambda P_n(x) + \lambda P_{n-1}(x)$$

Let $m_Y(x, t) = \sum_{n=0}^{\infty} e^{tn} P_n(x)$ be the m.g.f. of Y. Then

$$\frac{\partial}{\partial x} m_Y(x, t) = -\lambda m_Y(x, t) + e^t \lambda m_Y(x, t) = \lambda(e^t - 1) m_Y(x, t)$$

$$\text{Hence,} \quad m_Y(x, t) = \exp(\lambda x (e^t - 1)) \quad (m_Y(0, t) = 1, \ \ m_Y(x, 0) = 1)$$

But this is the m.g.f. of a $\mathcal{P}(\lambda x)$, so $Y \sim \mathcal{P}(\lambda x)$.

2.3.7 Given an event in interval T, its time $\sim U(0, T)$

Suppose it is given that one event occurred in $[0, T]$. Let X be the time at which it occurred.

$$P(X \le x \mid 1 \textit{ event in } [0, T]) = \frac{P(1 \textit{ event in } [0, x], 0 \textit{ in } (x, T])}{P(1 \textit{ in } [0, T])} = \frac{P_1(x) P_0(T - x)}{P_1(T)} = \frac{x}{T}$$

which is the required $U(0, T)$ cdf. Note since events occur independently, the same result holds for each given event.

2.3.8 Superposed Poisson processes make another Poisson process

Consider a set of Poisson processes rate λ_i, $i = 1, ..., k$. Combining all the events we still have a Poisson process, rate $\sum_{i=1}^{k} \lambda_i$. This follows from the original definition, or from the fact that sum of independent Poisson r.vs is Poisson.

Note this ties in with the minimum of independent exponentials being exponential. The time to the first event in process i is $\mathcal{E}(1/\lambda_i)$. The time to the first event in *any* of the processes is the minimum of the times for each process, and is $\mathcal{E}(1/\sum_i \lambda_i)$.

2.3.9 Given $W = w$ events in total process, the number Y_i from one process is Binomial

In the combined process, given an event occurs, the probability it is from i^{th} subprocess is $\theta_i = \lambda_i / \sum_{j=1}^{k} \lambda_j$, again from original definition. This is true independently for each of w events. So, given $W = w$, $Y_i \sim B(w, \theta_i)$. Note this is the same result from Hwk Exercise 4.15.

Conversely, if we start with $(Y_i \mid W) \sim B(W, \theta_i)$ and give W a Poisson distribution, we get back $Y_i \sim \mathcal{P}(\lambda_i)$, as shown previously, using m.g.fs. (see 2.2.3).

2.4 Three important inequalities (C&B 4.7)

2.4.1 Hölder's inequality

$$|\,\mathrm{E}(XY)| \;\leq\; \mathrm{E}(|XY|) \;\leq\; (\mathrm{E}(|X|^p))^{1/p}(\mathrm{E}(|Y|^q))^{1/q}$$

where $p, q > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$.

Proof: First take v, $w > 0$, and p, q as above (note $p, q > 1$). Let

$$g(v, w) \;=\; \frac{1}{p}v^p \;+\; \frac{1}{q}w^q \;-\; vw$$

Minimising $g(v, w)$ w.r.t. v, we have $\frac{\partial g}{\partial v} = v^{p-1} - w$ and $\frac{\partial^2 g}{\partial v^2} = (p-1)v^{p-2} > 0$ and substituting $v^{p-1} = w$ we obtain that the minimised value of $g(v, w)$ is 0.

Now we note first that $-|XY| \leq XY \leq |XY|$ to get the first inequality, and then substitute $v = |X|/(\mathrm{E}(|X|^p))^{1/p}$ and $w = |Y|/(\mathrm{E}(|Y|^q))^{1/q}$ in $g(v, w) \geq 0$ to get the final result.

Note: Cauchy-Schwarz inequality is a special case of Hölder's inequality applied to $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$, with $p = q = 2$.

2.4.2 Chebychev's inequality If $g()$ is a non-negative real-valued function

$$P(g(X) \geq b) \;\leq\; \mathrm{E}(g(X))/b$$

Proof:

$$\begin{aligned}\mathrm{E}(g(X)) \;&=\; \textstyle\int_{-\infty}^{\infty} g(x) f_X(x)dx \;=\; \int_{A^c} g(x) f_X(x)dx \;+\; \int_A g(x) f_X(x)dx \\ &\geq\; 0 \;+\; b\int_A f_X(x)dx \;=\; bP(g(X) \geq b)\end{aligned}$$

where $A = \{x : g(x) \geq b\}$.

Most often this is applied to $g(X) = (X - \mathrm{E}(X))^2$, to obtain

$$P(|X - \mathrm{E}(X)| > a) \;=\; P((X - \mathrm{E}(X))^2 > a^2) \leq \mathrm{E}((X - \mathrm{E}(X))^2)/a^2 \;=\; \mathrm{var}(X)/a^2$$

Note also this is a very weak inequality. It is rather useless for practical examples. However it is very useful to give rates of convergence in large samples (see later).

2.4.3 Jensen's inequality

For a convex function $g()$,

$$\mathrm{E}(g(X)) \geq g(E(X)),$$

with strict inequality if $g()$ is strictly convex, unless $\mathrm{var}(X) = 0$.

Proof: For a convex function $g(x) \geq ax + b$ where a and b are chosen s.t. the line $y = ax + b$ touches the curve $y = g(x)$, at any chosen point, say at $\mathrm{E}(X)$. That is, $g(\mathrm{E}(X)) = a\mathrm{E}(X) + b$. Then

$$\mathrm{E}(g(X)) \geq \mathrm{E}(aX + b) \;=\; a\mathrm{E}(X) + b \;=\; g(\mathrm{E}(X))$$

Examples: $\mathrm{E}(X^2) \geq (\mathrm{E}(X))^2$, $\mathrm{E}(1/X) \geq 1/\mathrm{E}(X)$, $\mathrm{E}(-\log(X)) \geq -\log(\mathrm{E}(X))$, the last two of these being for a positive r.v. X.

Chapter 3: Random samples and Convergence

3.1 Some notes about Normal distributions

3.1.1 The (univariate) Normal distribution

(i) Define $X \sim N(\mu, \sigma^2)$ if $f_X(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-(x-\mu)^2/2\sigma^2)$.

(ii) Let $Y = (X-\mu)/\sigma$ and transform the density to obtain $f_Y(y) = (2\pi)^{-\frac{1}{2}} \exp(-y^2/2)$ or $Y \sim N(0,1)$.

(iii) $(X-\mu)$ has a distribution not depending on μ; μ is a location parameter.
$(X-\mu)/\sigma$ has a distribution not depending on μ or σ; σ is a scale parameter.
The family of Normal distributions is a location-scale family.

(iv) Now if $Y \sim N(0,1)$, direct integration gives $m_Y(t) = \exp(t^2/2)$. Hence, differentiating the m.g.f. $\mathrm{E}(Y) = 0$, $\mathrm{var}(Y) = \mathrm{E}(Y^2) = 1$.

(v) Transforming back to $X = \mu + \sigma Y$, $\mathrm{E}(X) = \mu$, $\mathrm{var}(X) = \sigma^2$, and transforming the m.g.f

$$m_X(t) = \mathrm{E}(\exp(t(\mu+\sigma Y))) = \exp(\mu t) m_Y(\sigma t) = \exp(\mu t + \tfrac{1}{2}\sigma^2 t^2)$$

(vi) Now let $Z = aX + b$, then

$$m_Z(t) = \exp(bt) m_X(at) = \exp((a\mu + b)t + \tfrac{1}{2}(a^2\sigma^2)t^2)$$

or $Z \sim N(a\mu + b, a^2\sigma^2)$.
Any linear transformation of a Normal r.v. is Normal.

3.1.2 Multivariate Normal distributions

(i) Let $\mathbf{X} = (X_1, ..., X_k)$ be i.i.d $N(0,1)$, so $m_{\mathbf{X}}(\mathbf{t}) = \exp(\frac{1}{2}\mathbf{t}'\mathbf{t})$.

(i) Define $\mathbf{Y}$ to be multivariate Normal (MVN) is $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, where X_i are i.i.d. N(0,1), and A is (for convenience) a $k \times k$ non-singular matrix. Hence $\mathbf{X} = A^{-1}(Y - \mathbf{b})$.

(ii) Hence $\mathrm{E}(\mathbf{Y}) = \mathbf{b} = \mu$, $\mathrm{var}(\mathbf{Y}) = \Sigma = AA'$, and $m_{\mathbf{Y}}(\mathbf{t}) = \exp(\mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})$.
Hence, the dsn of $\mathbf{Y}$ is determined by its mean, μ, and variance-covariance matrix Σ.

(iii)

$$\begin{aligned} m_{(Y_i,Y_j)} &= m_{\mathbf{Y}}(0, ..., 0, t_i, 0, ..., 0, t_j, 0, ..., 0) \\ &= m_{Y_i}(t_i) m_{(Y_j)}(t_j) \exp(t_i t_j (\Sigma_{ij} + \Sigma_{ji})) \end{aligned}$$

Thus the joint m.g.f factorises **if and only if** $\mathrm{cov}(Y_i, Y_j) = \Sigma_{ij} = \Sigma_{ji} = 0$.
Jointly Normal random variables are independent is and only if they have zero correlation.

(iv) The marginal dsn of any subset of the Y_i is Normal; this follows directly from the mgf.
The conditional dsn of any subset of the Y_i, conditional on the other components, is Normal; this is harder to prove.

(v) Using $\mathbf{X} = A^{-1}(\mathbf{Y} - \mu)$, we can transform from the joint density of the i.i.d. $N(0,1)$ components of $\mathbf{X}$, to get the density of $\mathbf{Y}$:

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-k/2} (det(\Sigma))^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mathbf{y} - \mu)'\Sigma^{-1}(\mathbf{y} - \mu))$$

3.1.3 Independence of sample mean and sample variance

(i) Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$.
Define the sample mean $\overline{X}_n = n^{-1} \sum X_i$ so $\overline{X}_n \sim N(\mu, \sigma^2/n)$,
and the sample variance $S^2 = (n-1)^{-1} \sum (X_i - \overline{X}_n)^2$.

(ii) Let $Y_i = (X_i - \overline{X}_n)$ for $i = 1, ..., (n-1)$, and note then $(X_n - \overline{X}_n) = -\sum_1^{n-1} Y_i$.
Then $(Y_1, ..., Y_{n-1}, \overline{X}_n)$ is multivariate Normal, but $\text{cov}(Y_i, \overline{X}_n) = 0$ (see 4.2.4(c)), so $(Y_1, ..., Y_{n-1})$ is independent of $\overline{X}_n$.

(iii) Now $(n-1)S^2 = \sum_1^{n-1} Y_i^2 + (-\sum_1^{n-1} Y_i)^2$ a function of $(Y_1, ..., Y_{n-1})$. So S^2 and $\overline{X}_n$ are independent.

3.1.4 Some notes about χ^2 distributions

(i) Define a χ^2_ν dsn as the dsn of $\sum_1^\nu Z_i^2$, where Z_i are i.i.d. $N(0,1)$.

(ii) Consider $Y \sim \chi^2_1 \equiv Z^2$ where $Z \sim N(0,1)$.

Direct integration gives $m_Y(t) = (1-2t)^{-\frac{1}{2}}$. Thus $Y \sim G(\frac{1}{2}, 2)$, and $\chi^2_\nu \equiv G(\nu/2, 2)$.

(iii) Now suppose $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$. Then

$$\sum_1^n (X_i - \mu)^2 \equiv \sum_1^n (X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2$$
$$W = (n-1)S^2/\sigma^2 + V$$

where $W \sim \chi^2_n$ and $V \sim \chi^2_1$. Also S^2 and V are independent. Hence the m.g.f. of $(n-1)S^2/\sigma^2$ is

$$m_W(t)/m_V(t) = (1-2t)^{-n/2}/(1-2t)^{-1/2} = (1-2t)^{-(n-1)/2}$$

Thus

$$(n-1)S^2 \sim \sigma^2 \chi^2_{n-1} \equiv \sigma^2 G((n-1)/2, 2) \equiv G((n-1)/2, 2\sigma^2)$$

3.2 Convergence and limit theorems (C&B 5.3)

3.2.1 Convergence in probability

Let $X_1, ..., X_n$ be a sequence of r.v.s defined on the same probability space. X_n converges in probability to X (written $X_n \xrightarrow{\text{p}} X$) if, given any $\epsilon > 0$

$$P(|X_n - X| > \epsilon) \to 0 \quad \text{as} n \to \infty$$

Example 1: $P(X_n = 7 + \frac{1}{n}) = 1 - \frac{1}{n}$, $P(X_n = n^2) = \frac{1}{n}$,
Given ϵ, take $n > \frac{1}{\epsilon}$, then $P(|X_n - 7| > \epsilon) = P(X_n = n^2) = \frac{1}{n} \to 0$ as $n \to \infty$. That is $X_n \xrightarrow{\text{p}} 7$.
(Note the limiting r.v. X is degenerate here: $P(X = 7) = 1$.)

3.2.2 Theorem (Weak Law of Large Numbers) (WLLN)

Let $Y_1, Y_2, ...$ be a sequence of i.i.d. r.vs with $\text{E}(|Y_i|) < \infty$. Let $\overline{Y}_n = \frac{1}{n} \sum_1^n Y_i$. Then $\overline{Y}_n \xrightarrow{\text{p}} \text{E}(Y_1)$.
Proof: (for $\text{var}(Y_i) = \sigma^2 < \infty$)
Let $\text{E}(Y_i) = \mu$, $\text{var}(Y_i) = \sigma^2$. Then $\text{E}(\overline{Y}_n) = \mu$ and $\text{var}(\overline{Y}_n) = \sigma^2/n$.
Hence by Chebychev's inequality, $P(|\overline{Y}_n - \mu| > \epsilon) \leq var(\overline{Y}_n)/\epsilon^2 = \sigma^2/n\epsilon^2 \to 0$ as $n \to \infty$.

Example 2: Y_i i.i.d., $\text{var}(Y_i) = \sigma^2 < \infty$, $\text{E}(Y_i) = \mu$, $Z_n = \frac{1}{n} \sum_1^n Y_i^2$.
By W.L.L.N, $\overline{Y}_n \xrightarrow{\text{p}} \mu$, and $Z_n \xrightarrow{\text{p}} \text{E}(Y_1^2) = \mu^2 + \sigma^2$.

(a) $X_n \xrightarrow{\text{p}} k \implies X_n^2 \xrightarrow{\text{p}} k^2$, since, assuming $k > 0$,

$$P(|X_n^2 - k^2| > \epsilon) \leq P(X_n < \sqrt{k^2 - \epsilon}) + P(X_n > \sqrt{k^2 + \epsilon}) \to 0$$

(b) $X_n \xrightarrow{\text{p}} X$, $W_n \xrightarrow{\text{p}} W \implies (X_n - W_n) \xrightarrow{\text{p}} (X - W)$
since $|(X_n - W_n) - (X - W)| \leq |X_n - X| + |W_n - W|$.
(c) Hence, $Z_n - (\overline{Y}_n)^2 \xrightarrow{\text{p}} \sigma^2 + \mu^2 - (\mu)^2 = \sigma^2$.

3.2.3 Convergence almost surely

Let $X_1, X_2, ...$ be a sequence of random variables defined on the same probability space. X_n converges almost surely to X (written $X_n \xrightarrow{\text{a.s}} X$) if, for any $\epsilon > 0$, $P(\lim_{n\to\infty} |X_n - X| > \epsilon) = 0$. Note this is convergence of X_n as a function; it is stronger than convergence in probability.

3.2.4 Theorem (Strong Law of Large Numbers) (SLLN) without proof.

Let $Y_1, Y_2, ...$ be a sequence of i.i.d. r.vs with $\mathrm{E}(|Y_i|) < \infty$. Let $\overline{Y}_n = \frac{1}{n}\sum_1^n Y_i$. Then $\overline{Y}_n \xrightarrow{\text{a.s}} \mathrm{E}(Y_1)$.

(i) **Example 1 above**, assuming the X_i are independent, given any k,
$P(X_n = n^2, \text{ for some } n > k) = 1$, so X_n does not converge almost surely to $X \equiv 7$.

(ii) C& B, P.215: If $2^{r-1} \leq n < 2^r$, $X_n = I((n - 2^{r-1})/2^{r-1}, (n + 1 - 2^{r-1})/2^{r-1})$.

3.2.5 Convergence in distribution

Defn: A sequence of random variables $X_1, X_2, ...$ converges in distribution to a random variable X ($X_N \xrightarrow{\text{D}} X$) if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all points x at which $F_X(x)$ is continuous.

Example 1 above (again): (with $n \geq 3$ say)

$$\left.\begin{array}{cc} F_{X_n}(x) = 0 & x < 7 + n^{-1} \\ F_{X_n}(x) = 1 - n^{-1} & 7 + n^{-1} \leq x < n^2 \\ f_{X_n}(x) = 1 & n^2 \leq x \end{array}\right\} \longrightarrow \left\{\begin{array}{cc} 0 & x \leq 7 \\ 1 & x > 7 \end{array}\right\}$$

So the limit is equal to $F_X(x) = I_{[7,\infty)}(x)$ except at the discontinuity point $x = 7$.

Theorem $X_n \xrightarrow{\text{p}} X \implies X_n \xrightarrow{\text{D}} X$
Proof:

$$\begin{aligned} F_X(x - \epsilon) &= P(X \leq x - \epsilon) = P(X \leq x - \epsilon,\ X_n \leq x) + P(X \leq x - \epsilon,\ X_n > x) \\ &\leq P(X_n \leq x) + P(|X - X_n| > \epsilon) = F_{X_n}(x) + \delta_n \\ 1 - F_X(x + \epsilon) &= P(X > x + \epsilon) = P(X > x + \epsilon,\ X_n \leq x) + P(X > x + \epsilon,\ X_n > x) \\ &\leq P(|X - X_n| > \epsilon) + P(X_n > x) = \delta_n + (1 - F_{X_n}(x)) \end{aligned}$$

where $\delta_n = P(|X - X_n| > \epsilon) \to 0$ as $n \to \infty$. So

$$F_X(x - \epsilon) - \delta_n \leq F_{X_n}(x) \leq F_X(x + \epsilon) + \delta_n \text{ where } \delta_n \to 0$$

But $F_X(x \pm \epsilon) \to F_X(x)$ as $\epsilon \to 0$ is F_X is continous at x.

Lemma $X_n \xrightarrow{\text{D}} c \implies X_n \xrightarrow{\text{p}} X$
Proof Given $\epsilon > 0$,

$$\begin{aligned} X_n \xrightarrow{\text{D}} c \implies & F_{X_n}(x) \to 0 \text{ for } x < c, \quad F_{X_n}(x) \to 1 \text{ for } x > c \\ \implies & P(X_n < c - \epsilon) \to 0, \quad \text{and } P(X_n > c + \epsilon) \to 0 \\ \implies & P(|X_n - c| > \epsilon) \to 0 \quad \text{or} \quad X_n \xrightarrow{\text{p}} c \end{aligned}$$

3.2.6 Central limit theorem

Let $Y_1, Y_2, ...$ be a sequence of i.i.d. r.vs with $\mathrm{E}(Y_i) = \mu$, and $0 < \mathrm{var}(Y_i) = \sigma^2 < \infty$.
Let $\overline{Y}_n = \frac{1}{n}\sum_1^n Y_i$. Then $\sqrt{n}(\overline{Y}_n - \mu)/\sigma \xrightarrow{\mathrm{D}} Z$ where $Z \sim N(0,1)$.

Proof: (for the case when $m_{Y_i}(t)$ exists for $|t| < h$, some $h > 0$.)
Let $Z_i = (Y_i - \mu)/\sigma$, so $\mathrm{E}(Z_i) = 0$, $\mathrm{var}(Z_i) = 1$ and $\overline{Z}_n = (\overline{Y}_n - \mu)/\sigma$.
Let $m(t)$ be the m.g.f. of Z_i, and let $m_n(t)$ be the m.g.f. of $\sqrt{n}\overline{Z}_n = \sqrt{n}(\overline{Y}_n - \mu)/\sigma$. So

$$\begin{aligned} m_n(t) &= \mathrm{E}(\exp(\sqrt{n}\overline{Z}_n t)) = \mathrm{E}(\exp((t/\sqrt{n})\sum_1^n Z_i)) = (m(t/\sqrt{n}))^n \\ &= (1 + m'(0)t/\sqrt{n} + \tfrac{1}{2}m''(0)t^2/n + ...)^n = (1 + 0 \times t/\sqrt{n} + \tfrac{1}{2} \times 1 \times t^2/n + ...)^n \\ &= (1 + t^2/n + ...)^n \longrightarrow \exp(\tfrac{1}{2}t^2) \end{aligned}$$

which is the m.g.f. of a $N(0,1)$ r.v., so by the theorem about convergence of m.g.fs we have the result.

3.3 Slutsky's Theorem and the delta-method

3.3.1 Slutsky's Theorem (with "sort-of" proof)

If $X_n \xrightarrow{\mathrm{D}} X$ and $Y_n \xrightarrow{\mathrm{p}} c$ (a constant), and $g(x,y)$ is a continuous function, then $g(X_n, Y_n) \xrightarrow{\mathrm{D}} g(X, c)$.

Note: The result is also true for vectors $\mathbf{X}_n$, $\mathbf{X}$ in $\Re^k$ and $\mathbf{Y}_n$, $\mathbf{c}$ in $\Re^l$ and functions $\mathbf{g} : \Re^{k+l} \to \Re^d$. Then $(\mathbf{X}_n, \mathbf{Y}_n)$ is a r.v. of dimension $k+l$, and $\mathbf{g}(\mathbf{X}_n, \mathbf{Y}_n)$ is a d dimensional r.v.

Idea of proof: There are theorems due to Skorokhod (which are beyond this course) which say that if $\mathbf{X}_n \xrightarrow{\mathrm{D}} \mathbf{X}$ then there are r.vs $\mathbf{X}_n^*$, $\mathbf{X}^*$ with the same distributions as each $\mathbf{X}_n$ and $\mathbf{X}$, s.t. $\mathbf{X}_n^* \xrightarrow{\mathrm{a.s.}} \mathbf{X}^*$.

So we have $(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{\mathrm{D}} (\mathbf{X}, \mathbf{c})$, so we transfer to $(\mathbf{X}_n^*, \mathbf{Y}_n^*) \xrightarrow{\mathrm{a.s.}} (\mathbf{X}^*, \mathbf{c})$.

Once we are in the world of a.s. convergence, everything about functions that is true in analysis is also true here, so that $\mathbf{g}(\mathbf{X}_n^*, \mathbf{Y}_n^*) \xrightarrow{\mathrm{a.s.}} \mathbf{g}(\mathbf{X}^*, \mathbf{c})$.

But convergence a.s. implies convergence in dsn, so $\mathbf{g}(\mathbf{X}_n^*, \mathbf{Y}_n^*) \xrightarrow{\mathrm{D}} \mathbf{g}(\mathbf{X}^*, \mathbf{c})$.

But now we are only talking about the dsns of these r.v.s, so we can transfer back to the original (non-star) ones with the same dsns, and $\mathbf{g}(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{\mathrm{D}} \mathbf{g}(\mathbf{X}, \mathbf{c})$.

3.3.2 Corollary; the delta-method

Suppose $\sqrt{n}(X_n - \mu) \xrightarrow{\mathrm{D}} N(0, \sigma^2)$ and $g(x)$ is differentiable at $x = \mu$.
Then $\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{\mathrm{D}} N(0, (g'(\mu))^2\sigma^2)$.

Proof: $g(x) = g(\mu) + (x - \mu)\, g'(\mu) + O((x-\mu)^2)$

$$\begin{aligned} \sqrt{n}(g(X_n) - g(\mu)) &= \sqrt{n}(X_n - \mu)g'(\mu) + \sqrt{n}O((X_n - \mu)^2) \\ &= \sqrt{n}g'(\mu)(X_n - \mu) + (\sqrt{n})^{-1}O((\sqrt{n}(X_n - \mu))^2) \end{aligned}$$

and the first term converges in dsn to $N(0, (g'(\mu))^2\sigma^2)$ and the second term converges in probability to 0.

3.3.3 The multivariate delta-method

Suppose $\sqrt{n}(\mathbf{X}_n - \mu) \xrightarrow{\mathrm{D}} N_k(0, \Sigma)$ and $\mathbf{g}: \ \Re^k \ \rightarrow \ \Re^r$ is differentiable at $\mathbf{x} = \mu$.
Then $\sqrt{n}(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\mu)) \xrightarrow{\mathrm{D}} N_r(0, (\nabla \mathbf{g}(\mu))\Sigma(\nabla \mathbf{g}(\mu))^t)$, where $\nabla \mathbf{g}(\mu)$ is the $r \times k$ matrix of derivatives $\left(\frac{\partial g_i}{\partial x_j}\right)$ evaluated at $\mathbf{x} = \mu$.

3.3.4 Examples

(i) Propagation of variance.

Suppose Y_i are i.i.d with mean μ and variance σ^2, so that $\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{\mathrm{D}} N(0, \sigma^2)$.

Then, taking $g(x) = x^{-1}$, and assuming $\mu \neq 0$, $g'(x) = -x^{-2}$, so $\sqrt{n}({\overline{Y}_n}^{-1} - \mu^{-1}) \xrightarrow{\mathrm{D}} N(0, \sigma^2\mu^{-4})$.

Or, taking $g(x) = \log(x)$, and assuming $\mu > 0$, $g'(x) = x^{-1}$, so $\sqrt{n}(\log(\overline{Y}_n) - \log(\mu)) \xrightarrow{\mathrm{D}} N(0, \sigma^2\mu^{-2})$

(ii) Variance-stabilising transformations.

If we can choose g so that $(g'(\mu))^2\sigma^2 \ = \ 1$, the limit distribution will be $N(0, 1)$.

For example, if Y_i are i.i.d. $\mathcal{P}(\theta)$, so $\overline{Y}_n \sim n^{-1}\mathcal{P}(n\theta)$, then $\sqrt{n}(\overline{Y}_n - \theta) \xrightarrow{\mathrm{D}} N(0, \theta)$.
$(g'(\theta))^2\theta = 1$ gives $g'(\theta) = \theta^{-\frac{1}{2}}$ or $g(\theta) = 2\sqrt{\theta}$. So $2\sqrt{n}(\sqrt{(\overline{Y}_n)} - \sqrt{(\theta)}) \xrightarrow{\mathrm{D}} N(0, 1)$.

Or, if Y_i are i.i.d. $\mathcal{E}(\lambda)$, so $\overline{Y}_n \sim G(n, \lambda/n)$, then $\sqrt{n}(\overline{Y}_n - \lambda) \xrightarrow{\mathrm{D}} N(0, \lambda^2)$.
$(g'(\lambda))^2\lambda^2 = 1$ gives $g'(\lambda) = \lambda^{-1}$ or $g(\lambda) = \log \lambda$. So $\sqrt{n}(\log(\overline{Y}_n) - \log(\theta)) \xrightarrow{\mathrm{D}} N(0, 1)$.

3.4 Sample quantiles and order statistics

3.4.1 The empirical distribution function (edf).

3.4.2 The edf approach to convergence of sample quantiles

3.4.3 Relationship of sample of U[0,1] to i.i.d exponentials

3.4.4 I.i.d. exponentials approach to convergence of sample quantiles

3.5 Relationships among some standard distributions