Ma 3 Book Notes

Lim Soon Wei Daniel

1. **DeMorgan's Laws** $(A \cap B)^c = A^c \cup B^c, (A \cup B)^c = A^c \cap B^c$.

2. **Probability function axioms**:

   (a) Axiom 1: $P(A) \geq 0$

   (b) Axiom 2: $P(S) = 1$, S =sample space.

   (c) Axiom 3: For two mutually exclusive events A and B, $P(A \cup B) = P(A) + P(B)$.

   (d) Axiom 4: When S has an infinite number of members, let $A_1, A_2, \ldots$ be events defined over S. If $A_i \cap A_j = 0$ for each $i \neq j$, then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

3. **Generalized inclusion-exclusion principle**:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

   Generally,

$$P\left(\cup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \leq i_1 \leq \cdots \leq i_k \leq n} P(A_{i_1} \cap \ldots \cap A_{i_k}) \right)$$

4. **Higher-order intersections**

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_n | A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \cdot P(A_{n-1} | A_1 \cap \cdots \cap A_{n-2}) \cdots P(A_2 | A_1) \cdot P(A_1)$$

   Prove by repeated use of $P(A \cap B) = P(A|B)P(B)$.

5. **Derangement:** The probability of a random shuffle of n elements producing a result where no element is in the correct position is given by:

$$Q = 1 - \frac{1}{n!} \sum_{p=1}^{n} (-1)^{p-1} \frac{n!}{p!} = \sum_{p=0}^{n} \frac{(-1)^p}{p!}$$

   which is the nth order truncation of the Taylor expansion of e^{-1}.

6. **Conditional Probability**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

7. **Rule of average conditional probabilities** For a partition $B_1, \ldots, B_n$ of sample space Ω, we have:

$$P(A) = P(A|B_1)P(B_1) + \ldots P(A|B_n)P(B_n)$$

   $P(A)$ is the weighted average of the conditional probabilities with weights $P(B_i)$.

8. **Bayes' Rule** For a partition $B_1, \ldots, B_n$ of all possible outcomes, we have:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \ldots + P(A|B_n)P(B_n)}, i = 1, 2, \ldots, n = \frac{P(A|B_i)P(B_i)}{P(A)}$$

9. **Multiplication rule for multiple events**

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

10. **Binomial Distribution**

- Mean $\mu = np$
- Mode $= \begin{cases} \lfloor (n+1)p \rfloor, & (n+1)p = 0 \text{ or } (n+1)p \notin \mathbb{Z} \\ (n+1)p \text{ and } (n+1)p - 1, & (n+1)p \in \mathbb{Z} \\ n, & p = 1 \end{cases}$.
- PDF: $f(k; n, p) = P(X = k, X \sim Binom(n, p)) = \binom{n}{k} p^k (1-p)^{n-k}$,

- Normal Approximation: given $\sqrt{npq}$ large, $\mu = np, \sigma = \sqrt{npq}$.

11. **Normal Distribution**
    - PDF: $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$.
    - Converting to standard normal $Z = \frac{x-\mu}{\sigma}$.
    - With Continuity correction: $P(a \leq X \leq b, X \sim Binom(n,p)) = N\left(\frac{b+\frac{1}{2}-\mu}{\sigma}\right) - N\left(\frac{a-\frac{1}{2}-\mu}{\sigma}\right)$.
    - Relation to error function: $N(x) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{x}{\sqrt{2}}\right)$.
12. **Skewness of the Binomial function** $Skewness(n,p) = (1-2p)/\sigma = (1-2p)/\sqrt{npq}$. Skewness is positive for $p < 1/2$ and called skewed to the right. It is negative for $p > 1/2$ and called skewed to the left.
13. **Skew-normal approximation**: For n independent trials with success probability p, $P(0$ to b successes$) \approx \Phi(z) - \frac{1}{6} Skewness(n,p)(z^2-1)\phi(z)$ where $z = \frac{b+\frac{1}{2}-\mu}{\sigma}$, Φ is the standard normal CDF and $\phi(z)$ is the standard normal curve.
14. **Poisson Approximation to Binomial** Condition: n large and p small. $P($k successes$) = e^{-\mu}\frac{\mu^k}{k!}$.
15. **Approximate 95% confidence interval, sampling with replacement** $p \pm \frac{1}{\sqrt{n}}$.
16. **Sampling**: Consider a population of size N with G good and B bad elements, with $N = G + B$. For a sample of size $n = g + b, 0 \leq g \leq n$, the probability of getting g good elements and b bad elements is:
    - For sampling with replacement: P(g good and b bad)=$\binom{n}{g}\frac{G^g B^b}{N^n}$.
    - For sampling without replacement: P(g good and b bad) = $\frac{\binom{G}{g}\binom{B}{b}}{\binom{N}{n}}$.
17. **Hypergeometric Distribution:** Number of "good" elements from a sample of size n without replacement from a population of N elements, G of which are "good". Three parameters: n, N, G. Gives the probability P(g good, n-g bad) for $g = 0, 1, 2, \ldots, n$.
18. **Random Walk Stuff**
    - Ballot theorem (ties allowed) - alternatively, non-negative random walk. So the walker can return to zero, but cannot go under. Then the probability that this occurs is $\frac{p+1-q}{p+1}$ where p is the number of positive steps, and q is the number of negative steps. Alternatively, write $k = p-q, t = p+q$ so that $p = \frac{t+k}{2}, q = \frac{t-k}{2}$ and the probability is $\frac{(t+k)/2+1-(t-k)/2}{(t+k)/2+1} = \frac{t+k+2-t+k}{t+k+2} = \frac{2k-2}{t+k+2}$. Compare this to the no-tie version probability: $\frac{p-q}{p+q} = \frac{k}{t}$.
    - Deriving the tie case from the no tie case. Note the number of non-tie sequences with $p+1$ positive movements is equal to the number of tie sequences with p positive movements. See this by noting that the first step has to be $(0,0) \to (1,1)$. Then use $(1,1)$ as the origin for the tie sequences. Appending the first movement to all these tie sequences, we obtain the non-tie sequences with $p+1$ positive movements. Number of non-tie sequences with $p+1$ is then $\frac{p+1-q}{p+1+q}\binom{p+1+q}{q} = \frac{p+1-q}{p+1}\binom{p+q}{q}$. Hence the probability of a tie sequence with p positive movements is $\frac{p+1-q}{p+1}$.

Progress: Larsen (5th ed), stopped at page 38. Pitman Page 98

Chapter 1

Border notes

2.2.2 Definition: Odds The odds against an event is $\frac{P(E^c)}{P(E)}$.

2.3.1 Definition: Probability space A probability space is a triple $(S, \mathcal{E}, P)$ where S is a nonempty sample space, $\mathcal{E}$ is the set of events, P is a countably additive probability measure on $\mathcal{E}$.

2.4.1 Definition: Random Variable A random variable on a probability space $(S, \mathcal{E}, P)$ is a real-valued function on S such that for every interval $I \subset \mathbb{R}$, the inverse image of I is an event.

2.5.1 Boole's Inequality $P((\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$.

2.7.1 Independence of complements If A and B are independent, A and B^c, A^c and B^c, A^c and B are all independent.

2.8.2 Inclusion-exclusion principle $P\left(\cup_{i=1}^n A_i\right) = \sum_i p(A_i) - \sum_{i<j} p(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - \ldots + (-1)^{n+1} P(A_1 A_2 \cdots A_n)$

2.10.5 Binomial identity $\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}$. $\sum_{i=0}^k \binom{n}{i}(-1)^i = (-1)^k \binom{n-1}{k}$.

3.2.2 Binomial Theorem $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$.

4.2.2 Stirling's Formula $n! \approx e^{-n} n^n \sqrt{2\pi n}(1+\epsilon_n)$.

4.3 Multinomial distribution Let there be m possible outcomes, ith outcome has probability p_i, and we take n independent trials, k_i of which result in outcome i, then:

$$P(k_i \text{ outcomes of type } i, i = 1, \ldots, m) = \frac{n!}{k_1! k_2! \cdots k_m!} p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$$

4.5.1 Bayes' Rule $P(B|A) = P(A|B)\frac{P(B)}{P(A)}$. Let $B_1, \ldots B_n$ be a partition of S. Then:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}$$

5.4.1 Definition: Distribution of random variable The distribution of the random variable $X : S \to \mathbb{R}$ on the probability space $(S, \mathcal{E}, P)$ is the probability measure P_X defined on $\mathbb{R}$ by $P_X(B) = P(X \in B)$. The probability mass function is $p_X(x) = P(X = x)$ for the discrete random variable. The cumulative distribution function is $F_X(t) = P(X \leq t) = P_X(-\infty, t]$.

5.9 Definition: Stochastic dominance X stochastically dominates Y if for all $t \in \mathbb{R}, P(X \geq t) \geq P(Y \geq t)$ and for some t this is a strict inequality. Write this also as $F_X(t) \leq F_Y(t)$.

5.10.1 Defintion: Expectation $EX = \sum_{s \in S} X(s)P(s) = \int xp(x)dx$.

5.11 Expectation of a composition Let X be a discrete RV on $(S, \mathcal{E}, P)$ and $g : \mathbb{R} \to \mathbb{R}$. Then $Eg \circ X = \sum_{s \in S} g\left(X(s)\right) P(s) = \sum_{x \in \text{range X}} g(x)p(x)$.

6.2 Definition: Stochastic Independence A set of random variables is stochastically independent if for every finite subset of random variables $\{X_1, \ldots, X_n\}$ and every collection of subsets of $\mathbb{R}$ $\{B_1, \ldots, B_n\}$ we have $P(X_1 \in B_1, \ldots, X_n \in$

$B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n)$.

6.5 Uniform Distribution CDF: $F(t) = \begin{cases} 0, t < a \\ 1, t > b, \frac{t-a}{b-a}, a \le t \le b \end{cases}$.

6.9.2 Definition: Convex Function f is convex on (a, b) iff $f''(x) \ge 0, \forall x \in (a, b)$.

6.9.3 Jensen's Inequality Let X be a random variable with finite expectation and let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then $E(f(X)) \ge f(EX)$. If X is non degenerate and f is strictly convex, the inequality is strict.

6.11.1 Holder's inequality If $1 \le p < q$, if $E|X^q|$ is finite, then $E|X^p|$ is also finite.

7.11.1 Multinomial identity

$$(1 + x_1)(1 + x_2) \cdots (1 + x_n) = \sum_{k=0}^{n} \sum_{i_1 < \cdots < i_k} x_{i_1} \cdots x_{i_k}$$
$$(1 - x_1)(1 - x_2) \cdots (1 - x_n) = \sum_{k=0}^{n} \sum_{i_1 < \cdots < i_k} (-1)^k x_{i_1} \cdots x_{i_k}$$

7.1.3 Indicator function algebra $1_{AB} = 1_A \cdot 1_B = \min(1_A, 1_B)$. $1_{A^c} = 1 - 1_A$. $1_{A \cup B} = \max(1_A, 1_B) = 1_A + 1_B - 1_A \cdot 1_B$.

7.5 Definition: Standard Normal Distribution $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

7.5.1 Gaussian integral $\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi}$

7.8 Normal Approximation to Binomial (deMoivre-Laplace Limit theorem Let $X \sim B(n, p)$, then $EX = np, Var(X) = np(1 - p)$ so $\forall a, b \in \mathbb{R}$:

$$\lim_{n \to \infty} P\left(a \le \frac{X - np}{\sqrt{np(1-p)}} \le b \right) = \frac{1}{2\pi} \int_a^b e^{-z^2/2} dz$$

8.2.1 Markov's inequality Let X be a non-negative RV with mean $\mu < \infty$. For every $a > 0$, $P(X \ge a) \le \frac{\mu}{a}$.

8.2.2 Chebychev's inequality Let X be a random variable with finite mean μ and variance σ^2. For every $a > 0$: $P(|X - \mu| \ge a) \le \frac{\sigma^2}{a^2}$.

9.2 Joint Distributions The Joint Distribution of $X_1, \ldots, X_n$ is $P_{\mathbf{X}}(B) = P(\mathbf{X} \in B)$ for $B \in \mathbb{R}^n$. The Joint CDF is $F_{\mathbf{X}}(t_1, \ldots, t_n) = P(X_i \le t_i, i = 1, 2, \ldots, n)$. The Joint PMF is $p_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$.

9.3 Expectation of Function on Joint Distribution $Eg(X, Y) = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$.

9.5 Marginal Distribution $p_X = P(X = x) = \sum_y p_{X,Y}(x, y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$.

9.7 Distribution of a Sum Let $Z = X + Y$. Then $P(Z = z) = \sum_x p_{X,Y}(x, z - x) = \int_{-\infty}^{\infty} f_{X,Y}(t - y, y) dy$. If X and Y are independent, $f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t - y) f_Y(y) dy$.

9.8 Covariance $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) = Var(X) + Var(Y) + 2E(X - EX)(Y - EY)$. Also, $Cov(X, Y) = E(XY) - E(X)E(Y)$.

9.10.1 Variance of a linear combination Let $Z = a_1 X_1 + \ldots + a_n X_n$. Then let $\mathbf{X} = (X_1, \ldots, X_n), \mathbf{a} = (a_1, \ldots, a_n)$ so $Z = \mathbf{a} \cdot X$. Define the covariance matrix $\Sigma = [Cov(X_i, X_j)]_{i,j=1}^n$. Then $Var(Z) = \mathbf{a}\Sigma\mathbf{a}^T = \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j) a_i a_j$. Alternatively, $\Sigma = \mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$ so that $\sigma_{i,j} = \Sigma_{i,j} = E[(X_i - EX_i)(X_j - EX_j)]$.

9.11.2 Cauchy-Schwartz Inequality $E(XY)^2 \le E(X^2)E(Y^2)$ with equality only if X and Y are linearly independent. Implies $|Cov(X, Y)|^2 \le Var(X)Var(Y)$.

9.13.1 Sum of independent normals is normal If $X \sim N(\mu, \sigma^2), Y \sim N(\lambda, \tau^2)$ then $X + Y \sim N(\mu + \lambda, \sigma^2 + \tau^2)$.

10.1.1 Quantile function For any $p \in (0,1)$ and $x \in \mathbb{R}$. $Q(p) \leq x \iff p \leq F(x)$

10.2 Stochastic dominance on expectation Let X stochastically dominate Y and let g be nondecreasing. Then $Eg(X) \geq Eg(Y)$ with equality when $F_X = F_Y$.

10.4.1 Central Limit Theorem v1 Let $X_1, X_2 \ldots$ be a sequence of iid random variables. Let $\mu = EX_i$ and $\sigma^2 = Var(X_i)$. Let $S_n = \sum_{i=1}^n X_i$. Then: $\frac{S_n - n\mu}{\sqrt{n}\sigma} \to N(0,1)$ converges in distribution.

11.2 Definition: Gamma Function $\Gamma(s) = \int_0^s t^{s-1}e^{-t}dt$. $\Gamma(m) = (m-1)!, m \in \mathbb{Z}^+$.

11.3 Defintion: Gamma Distribution For a Gamma(r, λ) distribution, $f(t) = \frac{\lambda^r}{\Gamma(r)}t^{r-1}e^{-\lambda t}, t > 0$.

11.4 Random Lifetime Let T be chosen with CDF $F(t)$. The survival function is $G(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(s)ds$. The hazard rate is $\lambda(t) = \lim_{h \to 0} \frac{P(T \in (t, t+h) | T > t)}{h} = \frac{f(t)}{G(t)}$, the instantaneous probability of failure.

11.5 Definition: Exponential distribution Exponential(λ) has density $f(t) = \lambda e^{-\lambda t}$, hazard rate λ constant, mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$. Exponential is memoryless: $P(T > t + s | T > t) = P(T > s)$.

11.8 Sum of independent exponentials The sum of n iiid Exponential(λ) RV has a Gamma(n, λ) distribution. $f(t) = \lambda^n e^{-\lambda t}\frac{t^{n-1}}{(n-1)!}$.

11.10 Definition: Poisson Distribution If $X \sim Poisson(\lambda)$, then the PMF is $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$. Mean and variance is λ.

11.12 Poisson Approximation to Binomial $Poisson(\lambda) \approx B(n, \lambda/n)$ when n is large and p is small.

11.15 Sum of Poissons is also Poisson $X \sim Poisson(\mu), Y \sim Poisson(\lambda) \implies X + Y \sim Poisson(\mu + \lambda)$.

12.1 Defintion: Order Statistics 1st order statistic is the minimum. nth order statistic is the maximum.

12.2 CDF of order statistics The CDF of the kth order statistic from a sample of n of iid RVs with individual CDF $F(x)$ is $F_{k,n}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^n \binom{n}{j}(1 - F(x))^{n-j}F(x)^j$.

12.3 PDF of order statistics $f_{k,n}(x) = n\binom{n-1}{k-1}(1 - F(x))^{n-k}F(x)^{k-1}f(x)$.

12.5.1 Definition: Beta Function $B(r,s) = \int_0^1 t^{r-1}(1-t)^{s-1}dt = \frac{\Gamma(s)\Gamma(r)}{\Gamma(r+s)}$.

12.5.2 Definition: Beta Distribution The Beta(r,s) distribution has density $f(x) = \frac{1}{B(r,s)}x^{r-1}(1-x)^{s-1}, x \in [0,1]$ and zero elsewhere. Mean is $\frac{r}{r+s}$. The (k,n) order statistic of a Uniform[0,1] distribution has a beta(k,n-k+1) distribution with mean $\frac{k}{n+1}$.

12.6 Conditioning on Random Variable Let X and Y be discrete RV with joint PMF $p(x,y)$. Then $P(Y = y | X = x) = \frac{P(Y=y, X=s)}{P(X=x)} = \frac{p(x,y)}{p_X(x)}$.

12.7 Conditional Expectation $E(Y|X = x) = \sum_y y\frac{p(x,y)}{p_X(x)}$.

12.9 Conditional Expectation is linear $E(aY + bZ|X) = aE(Y|X) + bE(Z|X)$. If iterated, $E(E(Y|X)) = EY$ (Statement 12.10).

13.1 Definition: Markov Chain A discrete time stochastic process where the conditional distribution of $X_{t_{n+1}}$ given $X_{t_n}, \ldots, X_{t_1}$ is the same as that given just X_{t_n} alone. A Markov chain is time-invariant if the distribution of $X_{t+s}|X_t$ does not depend on t.

13.2 Definition: Transition Matrix A time-invariant Markov chain can be represented by a transition matrix $P = [p(i,j)] = [P(X_{t+1} = j | X_t = i)]$. Each row probabilities must sum to one.

13.2 Definition: Reachable and Communicate State j is reachable from i if $p^n(i,j) > 0$ for some n. If states i and j are mutually reachable, they communicate. When every state communicates with every other state, the chain is irreducible.

13.4 Definition: Invariant Distribution A probability distribution x on states is invariant if $xP = x$, which is the eigenvector of P corresponding to the eigenvalue 1.

13.5 Definition: Martingales A martingale is a stochastic process $\{X_t : t \in T\}$ such that $E|X_t| < \infty, \forall t \in T$ and $E(X_{t_{n+1}}|X_{t_n}, \ldots, X_{t_1}) = X_{t_n}$. The expectation conditioned on past values is the present value. A submartingale has $E(X_{t_{n+1}}|X_{t_n}, \ldots, X_{t_1}) \geq X_{t_n}$, and a supermartingale has $E(X_{t_{n+1}}|X_{t_n}, \ldots, X_{t_1}) \leq X_{t_n}$.

13.5.2 Martingale Convergence Theorem Let $\{X_n\}$ be a martingale. If $\lim_{n\to\infty} E|X_n| = M < \infty$, then there is a random variable X_∞ with $E|X_\infty|$ such that $X_n \to X_\infty$ as $n \to \infty$ almost surely. If $X_n \geq 0$ for all n or $X_n \leq 0$ for all n, then $M < \infty$ is satisfied and the martingale converges.

13.8 Definition: Stopping Times Given a discrete-time stochastic process $X_1, \ldots, X_n \ldots$ a stopping time is an integer-valued random variable N such that $P(N < \infty) = 1$. and the event $(N = k)$ belongs to $\sigma(X_1, \ldots, X_k)$,. which is the σ-algebra of events generated by $X_1, \ldots, X_n$.

13.9 Definition: Stopped martingale $\bar{Z}_n = Z_{\min(N,n)}$. Identity: $\bar{Z}_n = \bar{Z}_{n-1} + \mathbf{1}_{N\geq n}(Z_n - Z_{n-1})$.

13.9.2 Martingale Stopping Theorem $EZ_n = EZ_1$ if any of the conditions hold:

- The stopped martingales $\bar{Z}_n$ are uniformly bounded.
- N is bounded.
- $EN < \infty$ and there is some $M < \infty$ such that for all n, $E(Z_{n+1} - Z_n|Z_n) < M$.

14.1 Defintion: Rademacher Random Variable $X_t = \begin{cases} 1, p = 1/2 \\ -1, p = -1/2 \end{cases}$. $EX_t = 0, Var(X_t) = 1$.

14.3 Criterion for reachability in random walk For (t, k) to be reachable, there must be non-negative integers p, m where p is the number of plus ones and m is the number of minus ones such that:

$$p + m = t, \quad p - m = k$$
$$p = \frac{t+k}{2}, \quad m = \frac{t-k}{2}$$

so both $t + k$ and $t - k$ must be even $\implies$ t, k have the same parity.

14.3.3 Number of paths $N_{t,k} = \binom{t}{\frac{t+k}{2}} = \binom{t}{\frac{t-k}{2}} = \binom{p+m}{p} = \binom{p+m}{m}$. Divide by 2^t to get the probability. If starting from a point (t_0, k_0) that is not the origin, the number of paths to (t_1, k_1) is $N_{t_1-t_0,k_1-k_0}$.

14.3.6 Reflection Principle Let (t_1, k_1) be reachable from (t_0, k_0) and on the same side of the time axis. Then there is a one-to-one correspondence between the set of paths from (t_0, k_0) to (t_1, k_1) that meet the time axis and the set of paths from $(t_0, -k_0)$ to (t_1, k_1).

14.3.7 Ballot Theorem If $k > 0$ there are exactly $\frac{k}{n}N_{n,k}$ paths from the origin to (n, k) satisfying $s_t > 0, t = 1, \ldots, n$, that is, paths that never return to zero. Hence the probability that the path does not return to zero is $\frac{k}{n} = \frac{p-m}{p+n}$.

14.4.1 Probability of equalizing $u_{2m} = P(S_{2m} = 0) = \frac{N_{2m,0}}{2^{2m}} = \binom{2m}{m}\frac{1}{2^{2m}}$.

14.5 Main Lemma The following probabilities are equal:

$$P(S_2m = 0) = u_{2m}$$
$$P(S_1 \neq 0, \ldots, S_{2m} \neq 0)$$
$$P(S_1 \geq 0, \ldots, S_{2m} \geq 0)$$
$$P(S_1 \leq 0, \ldots, S_{2m} \leq 0)$$
$$2P(S_1 > 0, \ldots, S_{2m} > 0)$$
$$2P(S_1 < 0, \ldots, S_{2m} < 0)$$

14.5.1 Definition: Types of paths Let $\mathcal{Z}_t$ be the set of paths satisfying $s_t = 0$. Let $\mathcal{P}_t$ be the set of paths satisfying $s_1 > 0, \ldots, s_t > 0$. Let $\mathcal{N}_t$ be the set of paths satisfying $s_1 \geq 0, \ldots, s_t \geq 0$.

14.5.2/3 One-one correspondences There is a one-to-one correspondence between $\mathcal{P}_{2m}$ and $\mathcal{N}_{2m-1}$. Nelson's Lemma: There is a one-to-one correspondence between $\mathcal{Z}_{2m}$ and $\mathcal{N}_{2m}$. A path in $\mathcal{Z}_{2m}$ with minimum value $-k$ corresponds to a path in $\mathcal{N}_{2m}$ with terminal value $2k$.

14.6.2 First Return to zero Let $f_t = f_{2m}$ denote the probability of the first return to zero occurring at t. $f_0 = 0$ by definition. Then $f_{2m} = u_{2m-2} - u_{2m} = \frac{1}{2m-1}\binom{2m}{m}\frac{1}{2^{2m}}$.

14.8 Definition: Last Return Let L_{2m} be the epoch of the last visit to zero, up t o and including $2m$. Let $\alpha_{2k,2m} = P(L_{2m} = 2k)$, the probability that the last return occurred at $2k$ after $2m$ epochs.

14.8.1 Arc-Sine Law for last returns $\alpha_{2k,2m} = u_{2k}u_{2(m-k)} \approx \frac{1}{m}\frac{1}{\pi\sqrt{\frac{k}{m}\left(1-\frac{k}{m}\right)}}$. Hence for $0 < \rho < 1$, $P(L_{2m} \leq \rho 2m) \approx \frac{2}{\pi}\sin^{-1}\sqrt{\rho}$. Note that $\alpha_{m,2m} = \frac{1}{2}$.

14.9 Definition: Dual Walk Define the dual walk: $S_t^* = S_n - S_{n-t}, t = 1, \ldots, n$ for a fixed n. Every event in S has a dual event in S^* that has the same probability. The dual walk can be seen as a rotation 180° around the origin, then sliding the left corner to the origin.

14.10 Probability of First Visit P(first visit to k occurs at epoch n) $= \frac{k}{n}\binom{n}{\frac{n-k}{2}}\frac{1}{2^n}$ provided $n-k$ is not a negative integer.

14.11 Expected number of visits before equalization Let M_k be the number of epochs for nonzero k which $S_n = k$ before the first return to zero. For all k, $EM_k = 1$.

14.12 Sign changes A sign change occurs when S_{t-1} and S_{t+1} have opposite signs. Note that $S_t = 0$ and t must be even. Theorem: P(there are exactly c sign changes before epoch t) $= 2P(S_t = 2c + 1)$.

14.14 Definition: Rademacher(p) random variable $X_t = \begin{cases} 1, \text{ probability p} \\ -1, \text{ probability 1-p} \end{cases}$ so $EX_t = 2p - 1, Var(X_t) = 4p(1-p)$.

14.15 Probability of reaching zero for asymmetric walk Let p be the probability of an uptick. Start at $S_0 = m$. Then the probability of reaching zero from m is $z_m = z_1^m$, where $z_1 = \begin{cases} 1, p \leq 1/2 \\ \frac{1-p}{p}, p \geq 1/2 \end{cases}$.

15.3.1 Definition: Estimator An estimator is the function $T : S \to \Theta$, a map from the sample space to the set of possible parameter values.

16.1.1 Definition: Unbiasedness An estimator $T : \chi \to \Theta$ is unbiased if for every $\theta \in \Theta$, $E_\theta T(X) = \int T(x)f(x,\theta)dx = \theta$.

16.1.2. Definition: Consistent Let T_n be the estimator of θ based on n replications. An estimator is consistent if $\text{plim}_{n\to\infty} T_n = \theta$. That is, $\forall\theta \in \Theta, \epsilon > 0, P_\theta(|T_n - \theta| > \epsilon) \to 0, n \to \infty$. Strongly consistent if $P_\theta(T_n \to \theta) = 1$.

16.1.3. Definition: Efficient T is efficient if for $\theta \in \Theta$, T has the minimum variance of any unbiased estimator.

16.3 Multivariable maximization $f : \mathbb{R}^n \to \mathbb{R}$ is maximized at $\mathbf{x}$ interior to its domain if all the first order partial derivatives vanish and the Hessian matrix (matrix of second order partials) is negative definite. That is, for all columns x of H, $x^*Hx < 0$.

16.4.1. Maximum likelihood estimator of Bernoulli trial $L(p;x) = px+(1-p)(1-x), x = 0, 1$. $\hat{p}(x) = \begin{cases} 1, x = 1 \\ 0, x = 0 \end{cases}$.

16.4.2. Maximum likelihood estimator of Binomial $L(p;k) = \binom{n}{k}p^k(1-p)^{n-k}$. $\hat{p}(x) = \frac{x}{n}$.

16.4.3 Maximum likelihood estimators of i.i.d. normals $L(\mu, \sigma^2; x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$. $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^n x_i}{n}$, $\hat{\sigma^2}_{MLE} = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n}$. But $E\hat{\sigma^2}_{MLE} = \frac{n-1}{n}\sigma^2$. Hence define $s^2 = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-1} = \frac{n}{n-1}\hat{\sigma^2}_{MLE}$.

16.5 MLE of composition The MLE of $g(\theta)$ is $g(\hat{\theta}_{MLE})$. But it may not be unbiased (Jensen's inequality).

16.6 Sufficient statistic Let $T = \psi(X_1, \ldots, X_n)$ be a statistic with density $f_T(t, \theta)$. If the likelihood function factors as: $L(\theta; x_1, \ldots, x_n) = f_T(\psi(x_1, \ldots, x_n); \theta) b(x_1, \ldots, x_n)$ so θ enters the likelihood function only through the distribution of T, then T is a sufficient statistic for θ. It suffices to maximise $f_T(\psi(x_1, \ldots, x_n); \theta)$. For example, for iid normals, $L(\mu, \sigma^2; \bar{x}, s^2) = \frac{-n}{2}\left[\log(2\pi) + \log(\sigma^2) + \frac{1}{\sigma^2}\left(\frac{n-1}{n}s^2 - 2\mu\bar{x} + \bar{x}^2 + \mu^2\right)\right]$ so $(\bar{x}, s^2)$ are sufficient for (μ, σ^2). In other words, no other statistic that can be calculated from the same sample provides any additional information about the value of the parameter.

17.1 Definition: Mean Square Error Let T be an estimator of $g(\theta)$. $MSE_T(\theta) = E_\theta\left[(T - g(\theta)^2\right] = \int \left(T(x) - g(\theta)\right)^2 f(x; \theta) dx$. If T is unbiased, $MSE_T(\theta) = Var(T)$.

17.1 Definition: Bias $b_T = E_\theta(T) - g(\theta)$. Then $MSE_T(\theta) = Var_\theta(T) + (b_T(\theta))^2$.

17.2 Expectation of log-likelihood $E_\theta \frac{\partial \mathcal{L}}{\partial \theta} = 0, \mathcal{L}(\theta; x) = \ln L(\theta; x)$.

17.3 Cramer-Rao Lower Bound Let f be continuously differentiable and let the support of x not depend on θ. Let T be an estimator of θ with differentiable bias function $b(\theta)$. Then $Var_\theta(T)$ is bounded below by: $Var_\theta(T) \geq \frac{[1 + b'(\theta)]^2}{nE_\theta\left[\left(\frac{\partial}{\partial \theta} \log f(X; \theta)\right)^2\right]}$.

17.3.2 Theorem: Unbiased sufficient statistic achieves lower bound of variance Let the likelihood factor as $L(\theta; x) = b(x) f_T(T(x); \theta)$ so T is a sufficient statistic. If $f_T(t; \theta)$ has the form $e^{a(\theta)t + b(\theta)}$ and T is unbiased, then its variance achieves the Cramer-Rao lower bound so T is the minimum variance unbiased estimator of θ.

17.4.1 When MLEs are consistent Multiple conditions; see notes.

17.6 Method of moments The kth sample moment is $\frac{\sum_{i=1}^n x_i^k}{n}$ and the kth distribution moment is $\int x^k f(x; \theta_1, \ldots, \theta_m) dx$. Solve for the set of parameters $(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ that satisfy the first m moments: $\int x^k f(x; \hat{\theta}_1, \ldots, \hat{\theta}_m) = \frac{\sum_{i=1}^n x_i^k}{n}, k = 1, \ldots, m$.

17.8 Confidence interval bounds (Normal, known SD) Define $z_\alpha = \Phi^{-1}(1 - \alpha)$. Then $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. So the $1 - \alpha$ confidence interval when σ is known is: $I = \left[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right], P(\mu \in I) = 1 - \alpha$. By the symmetry of the normal distribution, the symmetric interval is the shortest such interval that satisfies $P(\mu \in I) = 1 - \alpha$.

17.10 Construction confidence intervals

1. Choose the shortest interval $[a, b]$ containing MLE $\hat{\theta}$ that has $P_{\hat{\theta}}([a, b]) = 1 - \alpha$. This is when the density is largest.
2. Choose an interval such that $P(\theta < a) = P(\theta > b) = \frac{\alpha}{2}$.

18.2 Posterior density The posterior density $\phi(\theta_0|x)$ given the prior density $\phi(\theta)$ with likelihood $f(x; \theta)$ with observation x is given by Bayes law as: $\phi(\theta_0|x) = \frac{f(x; \theta_0)\phi(\theta_0)}{\int_\Theta f(x; \theta)\phi(\theta) d\theta} \propto f(x; \theta_0)\phi(\theta_0)$.

18.3 Conjugate Prior A parametric family of distributions is conjugate to a likelihood function if the posterior belongs to the family whenever the prior does.

18.3 Example: Beta Conjugate Prior Given the Binomial(n,p) likelihood function: $L(p; k, n) = \binom{n}{k} p^k (1 - p)^{n-k}$, if the prior density is $\phi(p) = Beta(s, f) = p^{s-1}(1 - p)^{f-1}$, then the posterior density is $\phi(p|k) \propto Beta(s + k, f + n - k)$.

18.3 Example: Exponential Conjugate Prior Given the exponential(λ) likelihood function $L(\lambda; T, n) \propto \lambda^n e^{-\lambda T}$, the conjugate prior $\phi(\lambda) = \lambda^{n_0 - 1} e^{-\lambda T_0}, n_0, T_0 > 0$ is a $Gamma(n_0, T_0)$ and has posterior density $\phi(\lambda|k, n) = \lambda^{n_0 - 1 + n} e^{-\lambda(T_0 + T)} = Gamma(n_0 + n, T_0 + T)$.

18.3 Example: Poisson Conjugate Prior Given the likelihood function $L(\mu; k, n) \propto \mu^k e^{-n\mu}$, the conjugate prior is $\phi(\mu) = \mu^{k_0 - 1} e^{-n_0 \mu} = Gamma(k_0, n_0)$ and the posterior is $\phi(\mu|k, n) \propto \mu^{k_0 + k - 1} e^{-\mu(n_0 + n)} = Gamma(k_0 + k, n_0 + n)$.

18.4 Definition: Loss function Function of the parameter and the estimate satisfying $L(\hat{\theta}, \theta) \geq 0$ and $L(\theta, \theta) = 0$.

18.4 Definition: Risk Function $\int_\Theta L(\hat{\theta}, \theta)\phi(\theta|x) d\theta$. A Bayesian estimate chooses $\hat{\theta}$ to minimize the risk.

18.4 Examples: Risk-minimizing function When $L(\hat{\theta},\theta) = |\hat{\theta}-\theta|$, the risk-minimising $\hat{\theta}$ is the median of $f(\theta|x)$. When $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$, the risk-minimizing $\hat{\theta}$ is the mean of $f(\theta|x) : \int \theta f(\theta,x)d\theta$.

19.4 Definition: Significant value for composite null hypothesis The size α is $\sup_{\theta\in\Theta_0}\{P_\theta(T\in C)\}$, where C is the critical region (in which we reject the null hypothesis in favour of the alternative hypothesis).

19.5 Definition: Type I error H_0 is rejected when it is true. Always α.

19.5 Definition: Type II error H_0 fails to be rejected when it is false. Power is $1-P(\text{Type II error}|\theta)$.

19.5 Definition: Uniformly Most Powerful Test Let a test be characterized by (T,C), the test statistic and the critical region. Let $\beta_{T,C}(\theta)$ be the probability of obtaining a type II error. If a test (T^*,C^*) has the maximum value of $1-\beta T,C(\theta)$ for all tests with the same significance level α, then it is called the uniformly most powerful test (UMP).

19.6 Definition: Likelihood ratio test Define f_0 to be the area of the PDF under the null hypothesis, and f_1 to be the area of the PDF under the alternative hypothesis. Then define $\lambda = \frac{f_1(x)}{f_0(x)}$. Choose a cutoff $k>0$, and reject the null hypothesis if $\lambda \geq k$.

19.7 Definition: Monotone Likelihood ratio property (MLRP) When $\Theta\in\mathbb{R}$, the probability model satisfies MLRP if there exists real-valued statistic $T(x)$ such that for all $\theta<\theta'$, $\frac{f(x;\theta')}{f(x;\theta)}$ is non-decreasing in $T(x)$. Alternatively, write $(\theta<\theta', T(x)<T(x')) \implies \frac{f(x;\theta')}{f(x;\theta)} \leq \frac{f(x';\theta')}{f(x';\theta)}$.

19.9 Likelihood ratio test for composite hypothesis without MLRP Let $\hat{\theta}_0$ be the MLE of θ over Θ_0 and $\hat{\theta}_1$ be the MLE for θ over Θ_1. Then: $\lambda(x) = \frac{L(\hat{\theta}_0(x);x)}{L(\hat{\theta}_1(x);x)}$ is a test. This is the ratio of the maximum likelihoods at the MLE. Choose critical value λ^* and reject the null hypothesis if $0\leq\lambda(x)\leq\lambda^*$. The size of the test is $P(\Lambda\leq\lambda^*|H_0 \text{ is true}) = \alpha$.

20.3 Definition: Chi-square distribution If $X_1,\ldots,X_n$ are i.i.d $N(\mu,\sigma^2)$ RVs, then (1). $\bar{X}$ and S^2 are independent, (2). $\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$, (3). $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\bar{X})^2 \sim \chi^2(n-1)$.

20.4 Definition: F-distribution Let $U\sim\chi^2(n), V\sim\chi^2(m)$ be independent. Then $\frac{V/m}{U/n}\sim F_{m,n}$.

20.5 Definition: Student t-distribution Let $Z\sim N(0,1), U\sim\chi^2(n)$ be independent. Then $T_n = \frac{Z}{\sqrt{U/n}}\sim t(n)$.

20.6 Mean with estimated SD For a sample of n $N(\mu,\sigma^2)$ random variables, $\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim T_{n-1}$

20.11 Difference of means with same variance Given $X_1,\ldots,X_n$, and $Y_1,\ldots,Y_m$ normal with same variance, we test for the null hypothesis $\mu_X=\mu_Y$ using the test statistic: $t = \frac{\bar{x}-\bar{y}}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(n+m-2)$ with $s_p = \frac{\sum_{i=1}^n (x_i-\bar{x})^2+\sum_{j=1}^m (y_j-\bar{y})^2}{n+m-2}$.

20.12 Difference of means, unknown variance Define the test statistic $W = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}}\sim t(v)$ approximately. Use $TTest[data1, data2]$.

20.13 Paired Data Define $X_i-Y_i = (\mu_X-\mu_Y)+(\epsilon_i-\epsilon_i')$ as a new set of data points and test it using a t-test against a mean of zero.

20.14 Confidence interval for standard deviation Define $\chi^2_{\alpha,n}$ to be the quantile function for the chi-square distribution with n degrees of freedom. The $1-\alpha$ confidence interval for σ^2 is: $\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$.

20.15 Testing difference of variances Given n samples from $N(\mu_X,\sigma_X^2)$ and m samples from $N(\mu_y,\sigma_Y^2)$, construct $\frac{\frac{(m-1)S_Y^2}{\sigma_Y^2}}{\frac{(n-1)S_X^2}{\sigma_X^2}}\sim F_{m-1,n-1}$.

20.16.1 Testing difference of variances alternative Test $H_0=\sigma_X^2=\sigma_Y^2$: two tailed: reject H_0 if $\frac{s_Y^2}{s_X^2}\leq F_{\alpha/2,m-1,n-1}, \frac{s_Y^2}{s_X^2}\geq F_{1-\alpha/2,m-1,n-1}$. One tailed: replace $\alpha/2$ with α.

21.2 Empirical CDF $F_n(x) = \frac{|\{i:i\leq n \& X_i\leq x\}|}{n}$ for i.i.d. random variables $X_1,\ldots X_n,\ldots$. In terms of indicator functions, $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)$.

21.2 Kolmogorov-Smirnov Test Null hypothesis: The CDF of the data $X_1, \ldots, X_n$ is F_0. Transform each X_i using $Y_i = F_0(X_i)$ so that the range is now $[0, 1]$. If X_i indeed follows F_0, then the transformed variables are drawn from Uniform$[0, 1]$. Calculate the test statistic: $K = \sup_{0 \leq y \leq 1} |G_n(y) - y| = \sup_x |F_n(x) - F_0(x)|$ where G_n is the empirical CDF of the Y_i transformed random variables. Reject the null hypothesis if K is larger than the cut-off value.

21.5.1 Chi-Squared Test Consider the multinomial distribution with t possible results, each of which having probability p_i. Let X_i be the number of occurrences of the ith result in n independent trials. Let $\mathbf{p}_0$ be a t-dimensional column vector of observed probabilities that sum to one. Let the null hypothesis be that the this is the true parameter set$H_0 : \mathbf{p} = \mathbf{p}_0$. Estimate s parameters that determine $\mathbf{p}$ using maximum likelihood. Then calculate the test statistic: $D = \sum_{n=1}^{t} \frac{(X_i - np_i)^2}{np_i}$ which has a chi-squared distribution with $t - 1 - s$ degrees of freedom. Ensure each $np_i \geq 5$.

21.8 Testing independence Given pairs of observations $(X_i, Y_i), i = 1, \ldots, n$. Bin the data into columns containing the X values and rows containing the Y values. Calculate the relative frequency of each column and row (sum of counts in that column or row divided by the total count). If X and Y are independent, the relative frequency of each cell should be the row frequency multiplied by the column frequency. The number of degrees of freedom is the (number of rows -1) times (number of columns -1).

21.10 Minimum chi-squared estimators If $\mathbf{p}$ depends on parameter vector $\boldsymbol{\theta}$, we can choose $\hat{\boldsymbol{\theta}}$ to minimise the test statistic, which is equivalent to minimizing $\sum_{i=1}^{t} \frac{X_i^2}{p_i(\boldsymbol{\theta})}$.

22.3.1 Definition: Orthogonal Matrix An orthogonal matrix has its transpose as its inverse: $A^T = A^{-1}$. Orthogonal matrices preserve norms $||Ax|| = ||x||$ and inner products: $(Ax) \cdot (Ay) = x \cdot y$. These are equivalent definitions of orthogonal matrices.

22.3.4 Definition: Quadratic forms $x^T Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$. Identities:

$$(x + y)^T A(x + y) = x^T Ax + 2x^T Ay + y^T Ay$$
$$\nabla(a^T Ax) = A^T a$$
$$\nabla(x^T Ax) = 2Ax$$

A matrix is positive definite if $x^T Ax > 0$ whenever $x \neq 0$ and is positive semidefinite if $x^T Ax \geq 0$ whenever $x \neq 0$.

The rows and columns of an orthogonal matrix A are orthonormal.

22.4.2 Principal Axis Theorem Consider an $n \times n$ symmetric matrix A. Let C be the matrix of orthonormal eigenvectors of A. Then C is orthogonal and $\Lambda = C^{-1}AC$ where Λ is the diagonal matrix with eigenvalues on the diagonal.

22.5.1 Orthogonal Complement Theorem Consider a linear subspace M. Any vector x can be written as a unique sum of $x_M + x_\perp$ where $x_M \in M$ and $x_\perp \in M_\perp$. x_M is the orthogonal projection of x onto M, and is the point on M closest to x.

22.5.3 Projection is linear $(x + z)_M = x_M + z_M, (\alpha x)_M = \alpha x_M$.

22.5.3 Projection Operator Let $x_1, \ldots, x_k$ be a basis for M who lives in n dimensional space, and let X be a $n \times k$ matrix whose column are x_k. Then the projection operator that maps y to y_M is:

$$y_M = Py = X(X^T X)^{-1} X^T y$$

22.6 Normal Parameters of Linear Combination Let $\mathbf{X}$ be a column vector of n random variables with column vector of means $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{Y} = \mathbf{AX}$ where A is an $m \times n$ matrix of constants. Then $\mathbf{EY} = \mathbf{A}\boldsymbol{\mu}$ and $Var(\mathbf{Y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$.

22.8.1 Defintion: Multivariate Normal A random vector $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ has a multivariate normal distribution if for every constant vector $\mathbf{T} \in \mathbb{R}^n$, the linear combination $\mathbf{T}^T\mathbf{X} = \sum_{i=1}^{n} T_i X_i$ has a normal $N(\mu_T, \sigma_T^2)$ distribution. By Corollary 22.8.3, if $X_1, \ldots, X_n$ are independent normals, then the vector $\mathbf{X}$ has a multivariate normal distribution. This holds in the other direction: By Proposition 22.8.5, every component of a multivariate normal distribution is a normal distribution. By Proposition 22.8.4, $\mathbf{AX}$ where $\mathbf{A}$ is a constant $m \times n$ matrix is an m-dimensional Normal random vector.

22.8.9 Multivariate Normal Density Let $\mathbf{X} = (X_1, \ldots, X_n)$ with a non-singular variance-covariance matrix Σ. The density is:

$$f(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

22.9.1 Multivariate Normal and Chi-Square Let $\mathbf{X} \sim N(0, \mathbf{I}_n)$. Then $\mathbf{X}^T \mathbf{A} \mathbf{X} \sim \chi^2(k)$ iff $\mathbf{A}$ is symmetric, idempotent and has rank k. If $\mathbf{X} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, then $\left(\frac{\mathbf{X}-\boldsymbol{\mu}}{\sigma}\right)^T \mathbf{A} \left(\frac{\mathbf{X}-\boldsymbol{\mu}}{\sigma}\right) \sim \chi^2(k)$ iff $\mathbf{A}$ is symmetric, idempotent and has rank k.

22.9.3 Testing independence with idempotent matrices Let $\mathbf{X} \sim N(0, \sigma^2 \mathbf{I})$ and let $\mathbf{A}_1$ and $\mathbf{A}_2$ be symmetric idempotent matrices that satisfy $A_1 A_2 = A_2 A_1 = 0$. Then $\mathbf{X}^T \mathbf{A}_1 \mathbf{X}$ and $\mathbf{X}^T \mathbf{A}_2 \mathbf{X}$ are independent.

22.10.1 Theorem (A Covariance Menagerie) *We have:*

1. $\boldsymbol{E}(X_i X_j) = (\boldsymbol{E}\, X_i)(\boldsymbol{E}\, X_j) = \mu^2$, *for* $i \neq j$ *(by independence).*
2. $\boldsymbol{E}(X_i^2) = \sigma^2 + \mu^2$.
3. $\boldsymbol{E}(X_i T) = \sum_{\substack{j=1 \\ j \neq i}}^n \boldsymbol{E}(X_i X_j) + \boldsymbol{E}(X_i^2) = \boldsymbol{E}(X_i^2) = \sigma^2 + n\mu^2$.
4. $\boldsymbol{E}(X_i \bar{X}) = \boldsymbol{E}(X_i T/n) = (\sigma^2/n) + \mu^2$.
5. $\boldsymbol{E}(T) = n\mu$.
6. $\boldsymbol{Var}(T) = n\sigma^2$.
7. $\boldsymbol{E}(T^2) = n\sigma^2 + n^2\mu^2$.
8. $\boldsymbol{E}(\bar{X}) = \mu$.
9. $\boldsymbol{Var}(\bar{X}) = \sigma^2/n$.
10. $\boldsymbol{E}(\bar{X}^2) = (\sigma^2/n) + \mu^2$.
11. $\boldsymbol{E}(D_i) = 0$, $i = 1, \ldots, n$.
12. $\boldsymbol{Var}(D_i) = \boldsymbol{E}(D_i^2) = (n-1)\sigma^2/n$

$$\boldsymbol{Var}(D_i) = \boldsymbol{E}(X_i - \bar{X})^2 = \boldsymbol{E}(X_i^2) - 2\,\boldsymbol{E}(X_i \bar{X}) + \boldsymbol{E}(\bar{X}^2) = (\sigma^2 + \mu^2) - 2\Big((\sigma^2/n) + \mu^2\Big) + \Big((\sigma^2/n) + \mu^2\Big) = \left(1 - \frac{1}{n}\right)\sigma^2.$$

13. $\boldsymbol{Cov}(D_i, D_j) = \boldsymbol{E}(D_i D_j) = -\sigma^2/n$, *for* $i \neq j$.

$$\boldsymbol{E}(D_i D_j) = \boldsymbol{E}\Big((X_i - \bar{X})(X_j - \bar{X})\Big) = \boldsymbol{E}(X_i X_j) - \boldsymbol{E}(X_i \bar{X}) - \boldsymbol{E}(X_j \bar{X}) + \boldsymbol{E}(\bar{X}^2)$$
$$\mu^2 - \Big[(\sigma^2/n) + \mu^2\Big] - \Big[(\sigma^2/n) + \mu^2\Big] + \big[(\sigma^2/n) + \mu^2\big] = -\sigma^2/n.$$

14. $\boldsymbol{Cov}(D_i, T) = \boldsymbol{E}(D_i T) = 0$.

$$\boldsymbol{E}(D_i T) = \boldsymbol{E}\Big(\Big(X_i - (T/n)\Big)T\Big) = \boldsymbol{E}(X_i T) - \boldsymbol{E}(T^2/n) = (\sigma^2 + n\mu^2) - (n\sigma^2 + n^2\mu^2)/n = 0.$$

15. $\boldsymbol{Cov}(D_i, \bar{X}) = \boldsymbol{E}(D_i \bar{X}) = \boldsymbol{E}(D_i T)/n = 0$.

23.1 Assumptions on Error $\mathbf{E}(\epsilon|X) = 0, \mathbf{Var}(\epsilon|X) = \mathbf{Var}(\epsilon\epsilon^T|X) = \sigma^2 I_{n\times n}$. Latter is homoskedacity.

23.2 Sum of Squared Residuals (SSR) $SSR(b) = (y - Xb)^T(y - Xb) = y^T y - 2y^T Xb + b^T X^T Xb$.

23.2 Normal Equation for OLS parameters $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$.

24.2 Statistics of the OLS estimator $\mathbf{E}\hat{\beta}_{OLS} = \beta$ (unbiased), $\mathbf{Var}\hat{\beta}_{OLS} = \sigma^2(X^TX)^{-1}$.

24.3.1 Gauss-Markov Theorem In the standard linear model, if $\mathbf{X}$ has rank K, then the OLS estimator $\hat{\beta}_{OLS}$ is the Best Linear Unbiased Estimate in that in among all the estimators b of β which are linear in y and which satisfy $\mathbf{E}(b) = \beta$ for any possible value of β, $\mathbf{Var}(b) = \mathbf{Var}(\hat{\beta}_{OLS}) + \mathbf{P}$ where $\mathbf{P}$ is positive semi-definite.

24.3.2 Unbiased estimator of error variance Assume $(\epsilon) \sim N(0, \sigma^2\mathbf{I})$. Then $\hat{\beta}_{OLS} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. The unbiased estimator of σ^2 is $s^2 = \frac{e^Te}{N-K}$. Also, $\frac{(N-K)s^2}{\sigma^2} \sim \chi^2(N-K)$. Also, for any K-vector w of weights, note that $w^T(\hat{\beta}_{OLS} - \beta) \sim N(0, \sigma^2 w^T(X^TX)^{-1}w)$ and $\frac{w^T(\hat{\beta}-\beta)}{s\sqrt{w^T(X^TX)^{-1}w}} \sim t(N-K)$. s is the residual standard error.

24.3.4 T-distribution of OLS estimators $\frac{\hat{\beta}_k - \beta_k}{s\sqrt{(X^TX)^{-1}_{kk}}} \sim t(N-K)$. The standard error of $\hat{\beta}_{OLS} = s\sqrt{(X^TX)^{-1}_{kk}}$. Note that the kth OLS parameter is associated with the kth column of X.

24.6 Testing complicated Hypotheses Consider the null hypothesis $H_0 : a = A\beta$ where A is a $q \times K$ constant matrix of constraints. The test statistic $F = \frac{1}{qs^2}(a - A\hat{\beta}_{OLS})^T[A(X^TX)^{-1}A^T](a - A\hat{\beta}_{OLS}) \sim F(q, N-K)$.

24.10 Coefficient of multiple correlation $1 - R^2 = \frac{e^Te}{y^Ty}$.

24.10 Adjusted R-squared $(1 - \bar{R}^2) = \frac{N-1}{N-K}(1-R^2)$, which penalises for too many regressors.

24.11 Confidence intervals for regression predictions Consider the prediction $y_* = x_*\hat{\beta}_{OLS}$ where x_* is a chosen value and the value of y_* is desired. Then: $\frac{x_*\hat{\beta}_{OLS} - y_*}{s\sqrt{x_*(X^TX)^{-1}x_* + 1}} \sim t(N-K)$. Hence the confidence interval is: $[y_* - t_{\alpha/2,N-K}s\sqrt{x_*(X^TX)^{-1}x_* + 1}, y_* + t_{\alpha/2,N-K}s\sqrt{x_*(X^TX)^{-1}x_* + 1}]$.

25.3 ANOVA model Consider $Y_{ij} = \mu_j + \epsilon_{ij}, i = 1, \ldots, n_j, i = 1, \ldots, k$ where there are k different factor levels and n_j different observations for the jth factor level. Let $n = n_1 + \ldots + n_k$ be the total number of observations. Nomenclature:

- y_{ij} is the response of the i^{th} observation at level j.
- $T_{\bullet j} = \sum_{i=1}^{n_j} y_{ij}$ is the **response total** at level j.
- $\bar{Y}_{\bullet j} = \dfrac{T_{\bullet j}}{n_j}$ is the sample mean at level j.
- $T_{\bullet\bullet} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \sum_{j=1}^{n} T_{\bullet j}$ is the sample overall total response.
- $\bar{Y}_{\bullet\bullet} = \dfrac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij} = \dfrac{1}{n}\sum_{j=1}^{k} T_{\bullet j}$ is the sample overall average response.
- The **treatment sum of squares** SSTR is defined to be

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$$

- $\mu - \sum_{j=1}^{k} \frac{n_j}{n}\mu_j$ is the overall average of the (unobserved) μ_js.

$$E(SSTR) = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j(\mu_j - \mu)^2$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{\bullet j})^2}{n_j - 1}$$

$$SSE = \sum_{j=1}^{k}(n_j - 1)s_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{\bullet j})^2$$

where SSE is the error sum of squares.

25.4.1 SSE and chi-squared $\frac{SSE}{\sigma^2} \sim \chi^2(n-k)$ and SSE and SSTR are stochastically independent.

25.4.1 Total sum of squares SSTOT=SSTR+SSE=$\sum_{j=1}^{k}(n_j-1)s_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{\bullet\bullet})^2$.

25.5 ANOVA testing equality of means Define the test statistic: $F = \frac{SSTR/(k-1)}{SSE/(n-k)}$ and reject the null hypothesis if $F \geq F_{1-\alpha,k-1,n-k}$.

Anova table format:

Source	df	SS	MS	F	P
Treatment	k-1	SSTR	$\frac{\text{SSTR}}{k-1}$	$\frac{\text{SSTR}/(k-1)}{\text{SSE}/(n-k)}$	$F \leqslant F_{k-1,n-k}$
Error	n-k	SSE	$\frac{\text{SSE}}{n-k}$		
Total	n-1	SSTOT			

25.7 Testing contrast hypotheses Define the weighted linear combination $C = w^T\boldsymbol{\mu}$. To test the hypothesis that $C = 0$, weight the same means $\hat{C} = \sum_{j=1}^{k} w_j\bar{y}_{\bullet j}$ and define $SS_C = \frac{\hat{C}^2}{\sum_{j=1}^{k}\frac{w_j^2}{n_j}}$. Then $F = \frac{SS_C}{SSE/(n-k)} \sim F(1, n-k)$. Reject the null hypothesis if $F \geq F_{1-\alpha,1,n-k}$.

Final Review

Likelihoods Consider a distribution $Exp(\lambda)$ with n draws from it. Then the likelihood of obtaining some sample is (up to some scaling factor):

$$L(\lambda; x_i) = \prod_{i=1}^{n} f(x_i; \lambda) = \prod_{i=1}^{n} \mathbf{1}_{x_i > 0} e^{-\lambda x_i} = \begin{cases} 0, \exists i : x_i \leq 0 \\ e^{-\lambda \sum_{i=1}^{n} x_i}, \text{else} \end{cases}$$

The first product follows from independence. Note that if any of the x_is are negative, then it is not possible that it came from the exponential distribution and hence the likelihood is zero.

MLE By definition $\hat{\theta}_{MLE}$ is such that for all x:

$$\max_{\theta \in \Theta} L(\theta; x) = L(\hat{\theta}_{MLE}; x)$$

Checking unbiasedness $E_{\mu_0}[\bar{x}] = \mu_0$. Use linearity of expectation.

Likelihood ratio tests Let $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$. Then define:

$$\lambda(x) \equiv \frac{\max_{\theta \in \theta_0} L(\theta, x)}{\max_{\theta \in \theta_1} L(\theta, x)}$$

1. α = Type I error = P(reject $H_0 | H_0$ is true)
2. Power = 1-P(type II error) = P(reject $H_0 | H_0$ is false). Note that for a simple hypothesis, where the null and alternate hypotheses are single points, the power is just a number, not a function.
3. Type II error = β = P(accept $H_0 | H_0$ false)

UMP test A test is uniformly most powerful if for a given $\alpha \in (0, 1)$ it yields the best possible power $1 - \beta$.

Neyman-Pearson Lemma For simple hypotheses, the likelihood ratio test is UMP.

Theorem 19.8.1 Given composite hypotheses $H_0 \leq \theta_0, H_1 : \theta > \theta_0$, if the likelihood ratio $\lambda(x)$ satisfies the maximum likelihood ratio property (MLRP), then there exists a test that is UMP and this test is constructed explicitly using LR tests.

Standard linear model Assumptions: the data comes from sampling:

$$y_i = \beta_1 x_i + \beta_0 + \epsilon_i$$

with unknown $\beta_0, \beta_1 \in \mathbb{R}$ and $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$ and independent with σ^2 unknown.

Testing assumptions Ensure that the data points are distributed linearly by eye power. Use Kolmogorov-Smirnov test to check the normality of the errors (do not include the distribution type!).

Chi-squared test Choose your bins such that the *expected* number of counts in each bin is ≥ 5.