Revision Notes for CO245 Probability and Statistics

Spring 2018

Results in grey are included in the formula sheet.

1 Probability

Sample Spaces and Events

- **Sample space** S: Range of possible outcomes of a random experiment.
- **Event**: Subset of sample space.
  - **Null event**: $\emptyset$.
- Events are **mutually exclusive** if $\forall i, j. E_i \cap E_j = \emptyset$.

The σ-algebra A collection $\mathfrak{S}$ of subsets of S is a σ-field or σ-algebra if it satisfies:

1. Nonempty: $S \in \mathfrak{S}$.
2. Closed under complements: if $E \in \mathfrak{S}$ then $\overline{E} \in \mathfrak{S}$.
3. Closed under countable union: if $E_1, E_2, \dots \in \mathfrak{S}$ then $\cup_{i=1}^{\infty} E_i \in \mathfrak{S}$.

Axioms:

1. For any E in $\mathfrak{S}$, $0 \leq P(E) \leq 1$.
2. $P(S) = 1$.
3. Countably additive: $P(\cup_i E_i) = \sum_i P(E_i)$.

Properties:

1. $P\left(\overline{E}\right) = 1 - P(E)$.
2. $P(\emptyset) = 0$.
3. For any events E and F, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Independence

- Events E and F are independent iff $P(E \cap F) = P(E) P(F)$.
- If E and F are independent, $\overline{E}$ and F are also independent.

Conditional Probability

1. $P(E \mid F)$ is called a **conditional** probability.
2. $P(E \cap F)$ is called a **joint** probability.
3. $P(E)$ is called a **marginal** probability.

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

- Events E_1 and E_2 are **conditionally independent** given F iff $P(E_1 \cap E_2 \mid F) = P(E_1 \mid F) P(E_2 \mid F)$.
- **Bayes theorem** (easily derived from definition above) states:

$$P(E \mid F) = \frac{P(E) P(F \mid E)}{P(F)}$$

  Read the question and **check carefully**. Make sure you know the difference between $P(A \cap B)$ and $P(A \mid B)$!

- For a set of events $\{F_1, F_2, \dots\}$ which form a partition of S, the **partition rule** (derived from $E = E \cap S$) states:

$$P(E) = \sum_i P(E \mid F_i) P(F_i)$$

Likelihood and Posterior Probability For parameters θ and evidence X:

1. **Likelihood function** is $P(X \mid \theta)$.
2. **Posterior probability** is $P(\theta \mid X)$.
3. **Prior probability** is $P(\theta)$.

By Bayes theorem:

$$P(\theta \mid X) = \frac{P(X \mid \theta) P(\theta)}{P(X)}$$

2 Random Variables

Mapping from sample space to $\mathbb{R}$ (e.g. $X : S \to \mathbb{R}$).

- **Probability distribution function** $P_X(x) = P\left(X^{-1}(x)\right)$.
  - Probabilities are between 0 and 1.
  - Sum to 1.
- **Cumulative distribution function** $F_X(x) = P_X(X \leq x)$.
  - For every real number x, $0 \leq F_X(x) \leq 1$.
  - F_X is monotonic.
  - $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
- A random variable is **simple** iff it can only take a finite number of possible values.

2.1 Discrete Random Variables

2.1.1 Definition

X is discrete iff range (X) is countable.

$$p(x_i) = F(x_i) - F(x_{i-1})$$

$$F(x_i) = \sum_{j=1}^{i} p(x_j)$$

- p_X is the **probability mass function**.
- F_x is the **cumulative distribution function**.

2.1.2 Expectation and Probability Generating Function

Mean $E(X)$

$$E_X(X) = \sum_x x p_X(x)$$

$$E_X(g(X)) = \sum_x g(x) p_X(x)$$

$$E_X(aX + b) = aE_X(X) + b$$

Variance Var (X)

$$\mathrm{Var}_X(X) = E_X\left[(X - E_X(X))^2\right] = E\left(X^2\right) - (E(X))^2$$

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

Standard Deviation sd$_X(X)$

$$\mathrm{sd}_X(X) = \sqrt{\mathrm{Var}_X(X)}$$

Don't forget the square root!

Skewness γ_1

$$\gamma_1 = \frac{E_X\left[(X - E_X(X))^3\right]}{\mathrm{sd}_X(X)^3}$$

Probability Generating Function $G_X(z)$

$$G_X(z) = E_X\left(z^X\right) = \sum_x p_X(x) z^x$$

Look out for well-known series results (e.g. Maclaurin series / geometric series).

Moments M_n

- The nth **moment** of a random variable X is $M_n = E(X^n)$.
- The nth **factorial moment** is $M_n^f = E(X(X-1)\ldots(X-n+1)) = G^{(n)}(1)$.

$$\begin{aligned}
M_0 &= & M_0^f &= & G(1) &= 1 \\
M_1 &= & M_1^f &= & G'(1) & \\
M_2 &= & M_2^f + M_1^f &= & G''(1) + G'(1) &
\end{aligned}$$

Sums of Random Variables Where $S_n = \sum_{i=1}^{n} X_i$ is a sum of random variables, and S_n/n is their average:

$$E(S_n) = \sum_{i=1}^{n} E(X_i) \qquad E\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n}$$

$$\mathsf{Var}(S_n) = \sum_{i=1}^{n} \mathsf{Var}(X_i) \quad \mathsf{Var}\left(\frac{S_n}{n}\right) = \frac{\sum_{i=1}^{n} \mathsf{Var}(X_i)}{n^2} \qquad (X_i \text{ are independent})$$

$$G_{S_n}(z) = \prod_{i=1}^{n} G_{X_i}(z) \qquad (X_i \text{ are independent})$$

2.1.3 Discrete Distributions

Bernoulli Bernoulli (p)

- $p(x) = p^x (1-p)^{1-x}$ for $x = 0, 1$.
- $\mu = p$.
- $\sigma^2 = p(1-p)$.

Binomial Binomial (n, p)

- n identical **independent** Bernoulli (p) trials.
- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 1, 2, \ldots, n$. (Remember $\binom{n}{x} = \frac{n!}{x!(n-x)!}$).
- $\mu = np$.
- $\sigma^2 = np(1-p)$.
- $\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$.

Geometric Geometric (p)

- Potentially infinite sequence of **independent** Bernoulli (p) trials.
- $p(x) = p(1-p)^{x-1}$ for $x = 1, 2, \ldots$.
- $\mu = \frac{1}{p}$.
- $\sigma^2 = \frac{1-p}{p^2}$.
- $\gamma_1 = \frac{2-p}{\sqrt{1-p}}$.

Poisson Poi (λ)

- $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, 2, \ldots$.
- $\mu = \sigma^2 = \lambda$.
- $\gamma_1 = \frac{1}{\sqrt{\lambda}}$.
- $G(z) = e^{-\lambda(1-z)}$.
- When p is small and n is large, Binomial (n, p) is approximated by Poi (n, p).

Uniform U $(\{1, 2, \ldots, n\})$

- $p(x) = \frac{1}{n}$ for $x = 1, 2, \ldots, n$.
- $\mu = \frac{n+1}{2}$.
- $\sigma^2 = \frac{n^2-1}{12}$.
- $\gamma_1 = 0$.

2.2 Continuous Random Variables

2.2.1 Definition

X is a **continuous random variable** if $\exists f_X : \mathbb{R} \to \mathbb{R}$ s.t.

$$P_X(B) = \int_{x \in B} f_X(x)\,\mathrm{d}x$$

- f_X is the **probability density function**.
- The **cumulative distribution function** is

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$$

- Note that $f_X(x) = F_X'(x)$.

Properties of a pdf

1. For all $x \in \mathbb{R}$, $f_X(x) \ge 0$.
2. $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

Transformed Random Variables E.g. $Y = g(X)$ for some $g : \mathbb{R} \to \mathbb{R}$ where g is **continuous** and **strictly monotonic** (so it has an inverse).

- $F_Y(y) = P(Y \le y) = P(g(X) \le Y) = P\left(X \le g^{-1}(Y)\right) = F_X\left(g^{-1}(y)\right)$.
- By the chain rule, we get $f_Y(y) = F_Y'(y) = f_X\left(g^{-1}(y)\right)\left|g^{-1'}(y)\right|$.

2.2.2 Mean, Variance and Quantiles

Mean $E(X)$

$$E_X(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$$

$$E_X(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$$

Properties:

1. **Linearity**: $E(aX+b) = aE(X) + b$.
2. **Additivity**: $E(g(X) + h(X)) = E(g(X)) + E(h(X))$.

Variance Var (X)

$$\begin{aligned}\mathrm{Var}_X(X) &= E\left((X-\mu_X)^2\right) \\ &= \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x - \mu_X^2 \\ &= E\left(X^2\right) - (E(X))^2\end{aligned}$$

- $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$.

Moment Generating Function $M_X(t)$

$$M_X(t) = E\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{d}x$$

- Might not exist (for some t).
- The nth moment is $M_n = \left.\frac{\mathrm{d}^n M_X(t)}{\mathrm{d}t^n}\right|_{t=0}$.

Characteristic Function $\phi_X(t)$

$$\phi_X(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\,\mathrm{d}x$$

- Always exists (Fourier transform of pdf).
- The nth moment is $M_n = (-i)^n \left.\frac{\mathrm{d}^n \phi_X(t)}{\mathrm{d}t^n}\right|_{t=0}$.

Probability Generating Functions

$$M(t) = G\left(e^t\right) \text{ and } \phi(t) = G\left(e^{it}\right)$$

Sum of Random Variables For independent random variables $X_1, X_2, \ldots, X_n$, and $S_n = \sum_{j=1}^{n} X_j$:

$$\phi_{S_n}(t) = \prod_{j=1}^{n} \phi_{X_j}(t) \text{ and } M_{S_n}(t) = \prod_{j=1}^{n} M_{X_j}(t)$$

Quantiles

$$Q_X(\alpha) = F_X^{-1}(\alpha)$$

E.g. **median** is $F_X^{-1}\left(\frac{1}{2}\right)$. I.e. the solution to $F_X(x) = \frac{1}{2}$.

2.2.3 Continuous Distributions

Uniform U (a,b)

- $f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$.
- $F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x < b \\ 0 & x \ge b \end{cases}$.
- $\mu = \frac{a+b}{2}$.
- $\sigma^2 = \frac{(b-a)^2}{12}$.

Exponential Exp (λ)

- $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$.
- $F(x) = 1 - e^{-\lambda x}$ for $x \ge 0$.
- $\mu = \frac{1}{\lambda}$.
- $\sigma^2 = \frac{1}{\lambda^2}$.
- **Memoryless**: $P(X > x) = e^{-\lambda x}$ and $P(X > x + s \mid X > s) = e^{-\lambda x}$.
- If the number of events is distributed by $N \sim \mathrm{Poi}(\lambda)$ then the time between consecutive events is distributed by $T \sim \mathrm{Exp}(\lambda)$.

Normal $\mathsf{N}\left(\mu, \sigma^2\right)$

- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$.

Standard Normal:

- $Z \sim \mathsf{N}(0,1)$.
- $f(z)$ is written as $\phi(z)$ and $F(z)$ as $\Phi(z)$.
- $\phi(-z) = \phi(z)$ and $\Phi(z) = 1 - \Phi(-z)$.

Standardising Normal RVs:

- $X \sim \mathrm{N}\left(\mu, \sigma^2\right) \implies \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$.

Central Limit Theorem:

- For $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, $\lim_{n\to\infty} \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathsf{N}(0,1)$.
- I.e. for large n, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- For large n, Binomial $(n,p) \approx \mathsf{N}(np, np(1-p))$

Log-Normal Distribution:

- Y is said to follow a log-normal distribution if $Y = e^X$ and $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$.

2.3 Joint Random Variables

Definitions Has joint cdf:

$$F_{XY}(x,y) = P_{XY}\left((-\infty, x], (-\infty, y]\right)$$

We can recover marginal cdfs:

$$F_X(x) = F_{XY}(x, \infty)$$
$$F_Y(y) = F_{XY}(\infty, y)$$

Has joint pmf:

$$p_{XY}(x,y) = P_{XY}(X = x, Y = y)$$

We can recover marginal pmfs:

$$p_X(x) = \sum_y p_{XY}(x,y)$$

$$p_Y(y) = \sum_x p_{XY}(x,y)$$

Definitions for Jointly Continuous Variables Has joint cdf:

$$F_{XY}(x,y) = \int_{t=-\infty}^{y} \int_{s=-\infty}^{x} f_{XY}(s,t)\,\mathrm{d}s\mathrm{d}t$$

Has joint pdf:

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$$

We can recover marginal pdfs:

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{XY}(x,y)\,\mathrm{d}y$$

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{XY}(x,y)\,\mathrm{d}x$$

Take care with integration. E.g. don't ignore constants!

Conditional Distributions

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

Finding Probability of $X < Y$

$$P(X < Y) = \int_{y=-\infty}^{\infty} F_{X|Y}(y \mid y)\, f_Y(y)\,\mathsf{d}y$$
$$= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{y} f_{XY}(x,y)\,\mathsf{d}x\mathsf{d}y$$

Expectation

$$E_{XY}(g(x,y)) = \sum_y \sum_x g(x,y)\, p_{XY}(x,y)$$
$$E_{XY}(g(x,y)) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} g(x,y)\, p_{XY}(x,y)\,\mathsf{d}x\mathsf{d}y$$

Conditional Expectation

$$E_{Y|X}(Y|X=x) = \sum_y y\, p_{Y|X}(y|x)$$

$$E_{Y|X}(Y|X=x) = \int_{y=-\infty}^{\infty} y\, f_{Y|X}(y|x)\,\mathsf{d}y$$

Tower Rule

$$E_Y(Y) = E_X\left(E_{Y|X}(Y \mid X)\right)$$

Covariance Measures how two RVs change in tandem with one another.

$$\sigma_{XY} = E_{XY}\left((X - \mu_X)(Y - \mu_Y)\right)$$

Correlation Invariant to scale of the RVs X and Y.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

3 Estimation

For $\underline{X} = (X_1, \ldots, X_n)$ representing n iid data samples from a population with distribution P_X, we observe $\underline{x} = (x_1, \ldots, x_n)$.

Statistic A random variable:

$$T = T(X_1, \ldots, X_n) = T(\underline{X})$$

Observed statistic is $t = t(\underline{x})$.

Estimator A statistic $T(\underline{X})$ when used to approximate parameters of the distribution $P_{X|\theta}(x \mid \theta)$. An **estimate** is the realised value for the estimator for a particular sample $t(\underline{x})$.

Bias Of an estimator T for a parameter θ:

$$\text{bias}(T) = E(T \mid \theta) - \theta$$

- **Unbiased** if bias is 0.
- Sample mean $\overline{x}$ is an unbiased estimate for population mean μ.
- Biased-corrected sample variance $S^2_{n-1} = \frac{n}{n-1} S^2$ is an unbiased estimate for σ^2.

Efficiency For two unbiased estimators, $\hat{T}$ and $\tilde{T}$, $\hat{T}$ is **more efficient** than $\tilde{T}$ if:

1. For all θ, $\text{Var}_{\hat{T}|\theta}\left(\hat{T} \mid \theta\right) \leq \text{Var}_{\tilde{T}|\theta}\left(\tilde{T} \mid \theta\right)$, and
2. There is some θ with $\text{Var}_{\hat{T}|\theta}\left(\hat{T} \mid \theta\right) < \text{Var}_{\tilde{T}|\theta}\left(\tilde{T} \mid \theta\right)$.

$\hat{T}$ is **efficient** if it is more efficient that any other possible estimator.

Consistency An estimator T is **consistent** if

$$\forall \epsilon > 0 \qquad P(|T - \theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

If T is unbiased and $\lim_{n\to\infty} \text{Var}(T) = 0$ then T is consistent.

3.1 Maximum Likelihood Estimation

1. Find the likelihood function $L(\theta)$ where:
$$L(\theta \mid \underline{x}) = \prod_{i=1}^{n} p_{X|\theta}(x_i) \text{ or } \prod_{i=1}^{n} f_{X|\theta}(x_i)$$
2. Take the natural log of the likelihood $l(\theta \mid \underline{x})$, and collect terms involving θ:
$$l(\theta \mid \underline{x}) = \sum_{i=1}^{n} \log\left(p_{X|\theta}(x_i)\right) \text{ or } \sum_{i=1}^{n} \log\left(f_{X|\theta}(x_i)\right)$$
3. Find the value of θ for which log-likelihood is maximised: usually find the $\hat{\theta}$ that solves
$$\frac{\delta}{\delta\theta} l\left(\hat{\theta}\right) = \frac{\delta}{\delta\theta} \log\left(L\left(\hat{\theta}\right)\right) = 0$$
4. Ensure that the estimate $\hat{\theta}$ corresponds to a maximum by checking that the second derivative satisfies
$$\frac{\partial^2}{\partial\theta^2} l\left(\hat{\theta}\right) < 0$$

The MLE is not necessarily unbiased, but it is consistent and efficient, if an efficient estimator exists.

Confidence Intervals The $100(1-\alpha)\%$ confidence interval for μ is given by:

$$\left[\overline{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

- **Normal distribution with known variance**: confidence interval above is exact.
- **Other distributions**: confidence interval above is an approximate, by CLT.
- **Normal distribution with unknown variance**: need to use bias-corrected variance: exact CI is given by $\left[\overline{x} - t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}, \overline{x} + t_{n-1,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}\right]$.

3.2 Bayesian Estimation

Maximum Likelihood Estimator

- Doesn't take into account any prior information about the MLE.
- Only returns a single and specific value of θ.

Using Prior Information Use Bayes theorem

$$\underbrace{P(\theta \mid X)}_{\text{posterior}} = \underbrace{P(X \mid \theta)}_{\text{likelihood}} \times \underbrace{P(\theta)}_{\text{prior}} \times \underbrace{\frac{1}{P(X)}}_{\text{evidence}}$$

Maximum a Posteriori Estimator Instead of maximising $\prod_{i=1}^{n} P(X = x_i|\theta)$, maximise $\prod_{i=1}^{n} P(\theta \mid X = x_i) = \prod_{i=1}^{n} P(X = x_i \mid \theta) \times P(\theta)$.

Prior Distributions Often we use the **beta distribution**: Beta $(\theta; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$ with $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\mathrm{d}x$

4 Hypothesis Testing

4.1 Testing Population Mean

Hypotheses

- $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ is a **two-sided** test.
- $H_0 : \theta > \theta_0$ vs. $H_1 : \theta < \theta_0$ is a **one-sided** test.

Tests Statistics and Rejection Regions

- Choose a test statistic $T(\underline{X})$ for which we can find the distribution under H_0.
- Calculate the rejection region R such that $P(T \in R \mid H_0) = \alpha$ for some small probability α.
- Compute the realised value of the test statistic and conclude appropriately.

Errors and Power

1. **Type I Error**: Reject H_0 when it was true.
2. **Type II Error**: Not rejecting H_0 when H_1 is true.
3. **Power**: Probability of rejecting H_0 when H_1 is true.

Samples from Two Populations Use **bias-corrected pooled sample variance**:

$$S^2_{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{n_2}\left(Y_i - \overline{Y}\right)^2}{n_1+n_2-2} = \frac{n_1-1}{n_1+n_2-2}S^2_{n_1} + \frac{n_2-1}{n_1+n_2-2}S^2_{n_2}$$

4.2 Goodness of Fit

Chi-Square Statistic

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

- Approximation is valid only if $\forall j E_j \geq 5$.
- The **rejection region** at the $100\alpha\%$ level is given by

$$R = \left\{x^2 \mid x^2 > \chi^2_{k-p-1,1-\alpha}\right\}$$

where k is the number of terms summed and p is the number of parameters being estimated.

4.3 Independence Testing

1. Write up a contingency table.

	y_1	$\dots$	y_l	
x_1	n_{11}		n_{1l}	$n_{1\bullet}$
$\vdots$				
x_k	n_{k1}		n_{kl}	$n_{k\bullet}$
	$n_{\bullet 1}$		$n_{\bullet l}$	n

2. The expected value in each cell is $\hat{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$.
3. Compute the x^2 statistic.
4. Compare against χ^2 dist. with $kl - (k-1) - (l-1) - 1 = (k-1)(l-1)$ degrees of freedom.