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: ∅.

• Events are mutually exclusive if ∀i, j.Ei ∩ Ej = ∅.

The σ-algebra A collection S of subsets of S is a σ-field or σ-algebra if it satisfies:

1. Nonempty: S ∈ S.

2. Closed under complements: if E ∈ S then E ∈ S.

3. Closed under countable union: if E1,E2,···∈ S then ∪∞i=1Ei ∈ S.

Axioms:

1. For any E in S, 0 ≤ P (E) ≤ 1.

2. 3. P (S)=1.

Countably additive: P (∪iEi) = ∑i P (Ei).

Properties:

1. P (E) = 1 − P (E).

2. P (∅)=0.

3. For any events E and F, P (E ∪ F) = P (E) + P (F) − P (E ∩ F).

Independence

• Events E and F are independent iff P (E ∩ F) = P (E)P (F).

• If E and F are independent, E and F are also independent.

Conditional Probability

1. P (E | F) is called a conditional probability.

2. P (E ∩ F) is called a joint probability.

3. P (E) is called a marginal probability.

P (E | F) = P (E ∩ F)

P (F)

• Events E1 and E2 are conditionally independent given F iff P (E1 ∩ E2 | F) = P (E1 | F)P (E2 | F).

• Bayes theorem (easily derived from definition above) states:

P (E | F) = P (E)P(F | E)

P (F)

Read the question and check carefully. Make sure you know the difference between P (A ∩ B) and P (A | B)!

• For a set of events {F1,F2,...} which form a partition of S, the partition rule (derived from E = E ∩ S) states:

P (E) = ∑i

P (E | Fi)P (Fi)

Likelihood and Posterior Probability For parameters θ and evidence X:

1. Likelihood function is P (X | θ).

2. Posterior probability is P (θ | X).

3. Prior probability is P (θ).

By Bayes theorem:

P (θ | X) = P (X | θ)P (θ)

P (X)

1

2 Random Variables

Mapping from sample space to R (e.g. X : S → R).

• Probability distribution function PX (x) = P (X−1 (x)).

– Probabilities are between 0 and 1. – Sum to 1.

• Cumulative distribution function FX (x) = PX (X ≤ x).

– For every real number x, 0 ≤ FX (x) ≤ 1. – FX is monotonic. – FX (−∞)=0 and FX (∞)=1.

• A random variable is simple iff it can only take a finite number of possible values.

Variance Var(X)

VarX (X) = EX

[(X − EX (X))2]

[(X − EX (X))2]

[(X − EX (X))2]

= E (X2) − (E (X))2

= E (X2) − (E (X))2

= E (X2) − (E (X))2

Var(aX + b) = a2Var(X)

Standard Deviation sdX (X)sdX (X) = √VarX (X)

Don’t forget the square root!

2.1 Discrete Random Variables

2.1.1 Definition

X is discrete iff range(X) is countable.

Skewness γ1

γ1 = EX

[(X − EX (X))3]

[(X − EX (X))3]

sdX (X)3

sdX (X)3

p(xi) = F (xi) − F (xi−1)

Probability Generating Function GX (z)

F (xi) =

∑ij=1p(xj)

GX (z) = EX (zX) = ∑x

GX (z) = EX (zX) = ∑x

pX (x)zx

pX (x)zx

pX (x)zx

• pX is the probability mass function.

• Fx is the cumulative distribution function.

2.1.2 Expectation and Probability Generating Function

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

Moments Mn

• The nth moment of a random variable X is Mn = E (Xn).

Mean E (X)

EX (X) = ∑x

xpX (x)

EX (g (X)) = ∑x

g (x)pX (x)

• The nth factorial moment is Mn f= E (X (X − 1)...(X − n + 1)) = G(n) (1).

• The nth factorial moment is Mn f= E (X (X − 1)...(X − n + 1)) = G(n) (1).

• The nth factorial moment is Mn f= E (X (X − 1)...(X − n + 1)) = G(n) (1).

M0 = M0 f= G(1) = 1 M1 = M1 f= G (1) M2 = M2 f+ M1 f= G (1) + G (1)

M0 = M0 f= G(1) = 1 M1 = M1 f= G (1) M2 = M2 f+ M1 f= G (1) + G (1)

M0 = M0 f= G(1) = 1 M1 = M1 f= G (1) M2 = M2 f+ M1 f= G (1) + G (1)

EX (aX + b) = aEX (X) + b

Sums of Random Variables and Sn/n is their average:

Where Sn = ∑ni=1 Xi is a sum of random variables,

Where Sn = ∑ni=1 Xi is a sum of random variables,

2

∑E (Sn) =

∑nE (Xi) E

ni=1 E (Xi)

n

Var (Sn) =

(Snn

)

=

i=1 ∑ni=1

∑ni=1 Var(Xn2 i)

(Xi are independent)

GSn (z) =

(Snn

) Var (Xi) Var

=

∏ni=1GXi (z) (Xi are independent)

2.1.3 Discrete Distributions

Bernoulli Bernoulli(p)

• p(x) = px (1 − p)1−x for x = 0,1.

• μ = p.

• σ2 = p(1 − p).

Binomial Binomial(n, p)

• n identical independent Bernoulli(p) trials.

• p(x) =

( nx

)px (1 − p)n−x for x = 1,2,...,n. (Remember

( nx

)

=

n! x!(n−x)!).

• μ = np.

• σ2 = np(1 − p).

• γ1 = √np(1−p)1−2p .

Geometric Geometric(p)

• Potentially infinite sequence of independent Bernoulli(p) trials.

• p(x) = p(1 − p)x−1 for x = 1,2,....

• μ = p1.

• σ2 = 1−p

p2 .

• γ1 = √2−p 1−p.

Poisson Poi(λ)

• p(x) = e−λλx

x! for x = 0,1,2,....

• μ = σ2 = λ.

• γ1 = √1λ.

• G(z) = e−λ(1−z).

• When p is small and n is large, Binomial(n, p) is approximated by Poi(n, p).

Uniform U({1,2,...,n})

• p(x) = n 1for x = 1,2,...,n.

• μ = n+1 2 .

• σ2 = n212 −1 .

• γ1 = 0.

2.2 Continuous Random Variables

2.2.1 Definition

X is a continuous random variable if ∃fX : R → R s.t.

PX (B) =

ˆx∈B fX (x)dx

• fX is the probability density function.

• The cumulative distribution function is

FX (x) = P (X ≤ x) =

ˆ x−∞ fX (t)dt

• Note that fX (x) = F X (x).

Properties of a pdf

1. 2. For  ́ ∞−∞ all x ∈ R, fX (x) ≥ 0. fX (x)dx = 1.

Transformed Random Variables E.g. Y = g (X) for some g : R → R where g is continuous and strictly monotonic (so it has an inverse).

• FY (y) = P (Y ≤ y) = P (g (X) ≤ Y ) = P (X ≤ g−1 (Y )) = FX (g−1 (y)).

• By the chain rule, we get fY (y) = F Y (y) = fX (g−1 (y))∣∣∣g−1 (y)∣∣∣.

3

2.2.2 Mean, Variance and Quantiles

Mean E (X)

EX (X) =

ˆ ∞−∞ xfX (x)dx

EX (g (X)) =

ˆ ∞−∞ g (x)fX (x)dx 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)

VarX (X) = E

((X − μX)2)=

ˆ ∞−∞ (x − μX)2 fX (x)dx

=

ˆ ∞−∞ x2fX (x)dx − μ2X

= E (X2) − (E (X))2

• Var(aX + b) = a2Var(X).

Moment Generating Function MX (t)

MX (t) = E (etX) =

ˆ ∞−∞ etxfX (x) dx

• Might not exist (for some t).

• The nth moment is Mn = dnMdtXn

(t)

∣∣∣t=0.

Characteristic Function φX (t)

φX (t) = E (eitX) =

ˆ ∞−∞ eitxfX (x) dx

• Always exists (Fourier transform of pdf).

• The nth moment is Mn = (−i)n dnφdtX(t)

n

∣∣∣t=0.

Probability Generating Functions

M (t) = G(et) and φ(t) = G(eit)

Sum and Sn of = Random ∑nj=1 Xj:

Variables For independent random variables X1,X2,...,Xn,

φSn (t) =

∏nj=1φXj (t) and MSn (t) =

∏nj=1MXj (t)

Quantiles

QX (α) = F X −1 (α)

E.g. median is F X

−1 ( 12). I.e. the solution to FX (x) = 12.

2.2.3 Continuous Distributions

Uniform U(a, b)

• f (x) =

{ 1

b−a 0 a<x<b otherwise .

• F (x) =

0 x ≤ a

x−a 0 b−a a<x<b

x ≥ b

.

• μ = a+b

2 .

• σ2 = (b−a)12 2

.

Exponential Exp (λ)

• f (x) = λe−λx for x ≥ 0.

• F (x)=1 − e−λx for x ≥ 0.

• μ = λ1.

• σ2 = λ12 .

• Memoryless: P (X>x) = e−λx and P (X>x + s | X>s) = e−λx.

• If the number of events is distributed by N ∼ Poi(λ) then the time between consecutive events is distributed by T ∼ Exp (λ).

4

Normal N(μ, σ2)

• f (x) = σ√1

2πe− (x−μ)2σ2 2

.

• F (x) = σ√1

2π

́ x−∞ e− (t−μ)2σ2 2

dt.

Standard Normal:

• Z ∼ N(0,1).

• f (z) is written as φ(z) and F (z) as Φ(z).

• φ(−z) = φ(z) and Φ(z)=1 − Φ(−z). Standardising • X ∼ N(μ, Normal σ2) RVs:

=⇒ X−μ

σ ∼ N (0,1). Central Limit • For X =

Theorem: ∑ni=1 n Xi

where X1,X2,...,Xn are independent and identically dis- tributed random variables, limn→∞ σ/X−μ

√n ∼ N(0,1).

• I.e. for large n, X ∼ N

(μ, σn

2).

• For large n, Binomial(n, p) ≈ N(np, np(1 − p)) Log-Normal • Y Distribution:

is said to follow a log-normal distribution if Y = eX and X ∼ N (μ, σ2).

2.3 Joint Random Variables

Definitions Has joint cdf:

FXY (x, y) = PXY ((−∞,x],(−∞,y]) We can recover marginal cdfs:

FX (x) = FXY (x,∞) FY (y) = FXY (∞,y) Has joint pmf:

pXY (x, y) = PXY (X = x, Y = y) We can recover marginal pmfs:pX (x) = ∑y

pXY (x, y)

pY (y) = ∑x

pXY (x, y)

Definitions for Jointly Continuous Variables Has joint cdf:

FXY (x, y) =

ˆ y

t=−∞

ˆ xs=−∞ fXY (s, t)dsdt

Has joint pdf:

fXY (x, y) = ∂x∂y∂2

FXY (x, y)

We can recover marginal pdfs:

fX (x) =

ˆ ∞y=−∞ fXY (x, y)dy

fY (y) =

ˆ ∞x=−∞ fXY (x, y)dx

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

Conditional Distributions

pY |X (y | x) = pXY (x, y)

pX (x)

Finding Probability of X<Y

P (X<Y ) =

ˆ ∞y=−∞ FX|Y (y | y)fY (y)dy

=

ˆ ∞y=−∞

ˆ yx=−∞ fXY (x, y)dxdy

Expectation

EXY (g (x, y)) = ∑y

∑x

g (x, y)pXY (x, y)

EXY (g (x, y)) =

ˆ ∞y=−∞

ˆ ∞x=−∞ g (x, y) pXY (x, y)dxdy

Conditional ExpectationEY |X (Y |X = x) = ∑y

y pY |X (y|x)

EY |X (Y |X = x) =

ˆ ∞y=−∞ y fY |X (y|x)dy

5

Tower Rule

EY (Y ) = EX (EY |X (Y | X))

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

σXY = EXY ((X − μX)(Y − μY ))

Correlation Invariant to scale of the RVs X and Y .

ρXY = σXY

σXσY

3 Estimation

For X = (X1,...,Xn) representing n iid data samples from a population with dis- tribution PX, we observe x = (x1,...,xn).

Statistic A random variable:

T = T (X1,...,Xn) = T (X)

Observed statistic is t = t(x).

Estimator A statistic T (X) when used to approximate parameters of the distribu- tion PX|θ (x | θ). An estimate is the realised value for the estimator for a particular sample t(x).

Bias Of an estimator T for a parameter θ:

bias (T) = E (T | θ) − θ

• Unbiased if bias is 0.

• Sample mean x is an unbiased estimate for population mean μ.

• Biased-corrected sample variance S2n−1 = nn−1S2 is an unbiased estimate for σ2.

Efficiency For two unbiased estimators, T ˆand T,  ̃T ˆis more efficient than T  ̃if:

1. For all θ, Var T|θ

ˆ( T ˆ| θ)

≤ Var T|θ

̃( T ̃| θ), and

2. There is some θ with Var T|θ

ˆ( T ˆ| θ)

< Var T|θ

̃( T ̃| θ).

T ˆis efficient if it is more efficient that any other possible estimator.

Consistency An estimator T is consistent if

∀ε > 0 P (|T − θ| > ε) → 0 as n → ∞

If T is unbiased and limn→∞ Var(T)=0 then T is consistent.

3.1 Maximum Likelihood Estimation

1. Find the likelihood function L(θ) where:

L(θ | x) =

∏ni=1pX|θ (xi) or

∏ni=1fX|θ (xi)

2. Take the natural log of the likelihood l (θ | x), and collect terms involving θ:

l (θ | x) =

∑ni=1

log(pX|θ (xi)) or

∑nlog(fX|θ (xi)) i=1

3. Find the value of θ for which log-likelihood is maximised: usually find the θ ˆthat

solves δδθl

(θˆ)

= δθ δlog(L(θˆ))

= 0

4. Ensure that the estimate θ ˆcorresponds to a maximum by checking that the

second derivative satisfies ∂θ∂2

2l

(θˆ)

< 0

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

Confidence Intervals The 100 (1 − α)% confidence interval for μ is given by:

[x − z1− α2

√σn]

6

√σn,x + z1− α2

• 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 vari- ance: exact CI is given by

[]x − tn−1,1− α2 √n−1S .

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

P } (θ {{ | X) } posterior

√n−1S ,x + tn−1,1− α2

= P } (X {{ | θ) }

× likelihood

1 } P {{ (X) } evidence

Maximum maximise ∏ni=1 a Posteriori P (θ | X = Estimator xi) = ∏ni=1 P Instead (X of = xi maximising | θ) × P (θ).

∏ni=1 P (X = xi|θ),

Prior θα−1 (1 Distributions − θ)β−1 with B Often (α, β) we = use Γ(α)Γ(β)

Γ(α+β) the beta and distribution: Γ(z) = ́ 0 ∞× P } {{ (θ) }

prior

Beta(θ;α, β) = B(α,β) 1

× xz−1e−xdx

4 Hypothesis Testing

4.1 Testing Population Mean

Hypotheses

• H0 : θ = θ0 vs. H1 : θ = θ0 is a two-sided test.

• H0 : θ>θ0 vs. H1 : θ<θ0 is a one-sided test.

Tests Statistics and Rejection Regions

• Choose a test statistic T (X) for which we can find the distribution under H0.

• Calculate the rejection region R such that P (T ∈ R | H0) = α for some small probability α.

• Compute the realised value of the test statistic and conclude appropriately.

Errors and Power

1. Type I Error: Reject H0 when it was true. 2. Type II Error: Not rejecting H0 when H1 is true. 3. Power: Probability of rejecting H0 when H1 is true.

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

Sn21+n2−2 =

∑ni=1

1 (Xi − X)2 + ∑ni=1

2 (n1 + n2 − 2 Yi − Y )2

= n1 n+ 1 n2 − 1

− 2Sn21 + n1 n+ 2 n2 − 1

− 2Sn22

4.2 Goodness of Fit

Chi-Square Statistic

χ2 =

∑ki=1

(Oi − Ei)2 Ei

• Approximation is valid only if ∀jEj ≥ 5.

• The rejection region at the 100α% level is given by

R = {x2 | x2 > χ2k−p−1,1−α}

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.y1 ... yl

x...1 n11 n1l n1• xk nk1 nkl nk• n•1 n•l n

2. The expected value in each cell is ˆnij = ni•×nn •j

. 3. Compute the x2 statistic. 4. Compare against χ2 dist. with kl − (k − 1) − (l − 1) − 1 = (k − 1) (l − 1)

degrees of freedom.

7