Probability Background

Namrata Vaswani, Iowa State University

August 24, 2015

Probability recap 1: EE 322 notes

Quick test of concepts: Given random variables $X_1, X_2, \ldots X_n$. Compute the PDF of the second smallest random variable (2nd order statistic).

1 Some Topics

1. Chain Rule:
$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \ldots P(A_n|A_1, A_2, \ldots A_{n-1})$$

2. Total probability: if $B_1, B_2, \ldots B_n$ form a *partition* of the sample space, then
$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$
Partition: The events are mutually disjoint and their union is equal to the sample space.

3. Union bound: suppose $P(A_i) \geq 1 - p_i$ for small probabilities p_i, then
$$P(\cap_i A_i) = 1 - P(\cup_i A_i^c) \geq 1 - \sum_i P(A_i^c) \geq 1 - \sum_i p_i$$

4. Independence:
   - events A, B are independent iff
   $$P(A, B) = P(A)P(B)$$
   - events $A_1, A_2, \ldots A_n$ are mutually independent iff for any subset $S \subseteq \{1, 2, \ldots, n\}$,
   $$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

- analogous definition for random variables: for mutually independent r.v.'s the joint pdf of any subset of r.v.'s is equal to the product of the marginal pdf's.

5. Conditional Independence:
   - events A, B are conditionally independent given an event C iff

   $$P(A, B|C) = P(A|C)P(B|C)$$

   - extend to a set of events as above
   - extend to r.v.'s as above

6. Given X is independent of $\{Y, Z\}$. Then,
   - X is independent of Y; X is independent of Z
   - X is conditionally independent of Y given Z
   - $\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\mathbb{E}[Y|Z]$
   - $\mathbb{E}[XY|Z] = \mathbb{E}[X]\mathbb{E}[Y|Z]$

7. Law of Iterated Expectations:

   $$\mathbb{E}_{X,Y}[g(X, Y)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[g(X, Y)|Y]]$$

8. Conditional Variance Identity:

   $$Var_{X,Y}[g(X, Y)] = \mathbb{E}_Y[Var_{X|Y}[g(X, Y)|Y]] + Var_Y[\mathbb{E}_{X|Y}[g(X, Y)|Y]]$$

9. Cauchy-Schwartz Inequality:

   (a) For vectors v_1, v_2, $(v_1'v_2)^2 \leq \|v_1\|_2^2\|v_2\|_2^2$

   (b) For vectors:

   $$(\frac{1}{n}\sum_{i=1}^{n} x_i'y_i)^2 \leq (\frac{1}{n}\sum_{i=1}^{n} \|x_i\|_2^2)(\frac{1}{n}\sum_{i=1}^{n} \|y_i\|_2^2)$$

   (c) For matrices:

   $$\|\frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i\mathcal{Y}_i'\|^2 \leq \|\frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i\mathcal{X}_i'\|_2\|\frac{1}{n}\sum_{i=1}^{n} \mathcal{Y}\mathcal{Y}'\|_2$$

   (d) For scalar r.v.'s X, Y: $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$

   (e) For random vectors X, Y,

   $$(\mathbb{E}[X'Y])^2 \leq \mathbb{E}[\|X\|_2^2]\mathbb{E}[\|Y\|_2^2]$$

(f) Proof follows by using the fact that $\mathbb{E}[(X - \alpha Y)^2] \geq 0$. Get a quadratic equation in α and use the condition to ensure that this is non-negative

(g) For random matrices $\mathcal{X}, \mathcal{Y}$,

$$\|\mathbb{E}[\mathcal{X}\mathcal{Y}']\|_2^2 \leq \lambda_{\max}(\mathbb{E}[\mathcal{X}\mathcal{X}'])\lambda_{\max}(\mathbb{E}[\mathcal{Y}\mathcal{Y}']) = \|\mathbb{E}[\mathcal{X}\mathcal{X}']\|_2 \|\mathbb{E}[\mathcal{Y}\mathcal{Y}']\|_2$$

Recall that for a positive semi-definite matrix M, $\|M\|_2 = \lambda_{\max}(M)$.

(h) Proof: use the following definition of $\|M\|_2$: $\|M\|_2 = \max_{x,y:\|x\|_2=1,\|y\|_2=1} |x'My|$, and then apply C-S for random vectors.

10. Convergence in probability. A sequence of random variables, $X_1, X_2, \ldots X_n$ converges to a constant a in probability means that for every $\epsilon > 0$,

$$\lim_{n\to\infty} Pr(|X_n - a| > \epsilon) = 0$$

11. Convergence in distribution. A sequence of random variables, $X_1, X_2, \ldots X_n$ converges to random variable Z in distribution means that

$$\lim_{n\to\infty} F_{X_n}(x) = F_Z(x), \text{ for almost all points} x$$

12. Convergence in probability implies convergence in distribution

13. Consistent Estimator. An estimator for θ based on n random variables, $\hat{\theta}_n(\underline{X})$, is consistent if it converges to θ in probability for large n.

14. independent and identically distributed (iid) random variables: $X_1, X_2, \ldots X_n$ are iid iff they are mutually independent and have the same marginal distribution

   - For all subsets $S \subseteq \{1, 2, \ldots n\}$ of size s, the following two things hold

$$F_{X_i, i\in S}(x_1, x_2, \ldots x_s) = \prod_{i\in S} F_{X_i}(x_i) \quad \text{(independent) and}$$

$$F_{X_i}(x_i) = F_{X_1}(x_1) \quad \text{(iid)}$$

   - Clearly the above two imply that the joint distribution for any subset of variables is also equal

$$F_{X_i, i\in S}(x_1, x_2, \ldots x_s) = \prod_{i=1}^{s} F_{X_1}(x_i) = F_{X_1, X_2, \ldots X_s}(x_1, x_2, \ldots x_s)$$

15. Moment Generating Function (MGF) $M_X(u)$

$$M_X(u) := \mathbb{E}[e^{u^T X}]$$

- It is the two-sided Laplace transform of the pdf of X for continuous r.v.'s X
- For a scalar r.v. X, $M_X(t) := \mathbb{E}[e^{tX}]$, differentiating this i times with respect to t and setting $t = 0$ gives the i-th moment about the origin

16. Characteristic Function
$$C_X(u) := M_X(iu) = \mathbb{E}[e^{iu^T X}]$$
    - $C_X(-u)$ is the Fourier transform of the pdf or pmf of X
    - Can get back the pmf or pdf by inverse Fourier transform

17. Union bound: suppose $P(A_i) \geq 1 - p_i$ for small probabilities p_i, then
$$P(\cap_i A_i) = 1 - P(\cup_i A_i^c) \geq 1 - \sum_i P(A_i^c) \geq 1 - \sum_i p_i$$

18. Hoeffding's lemma: bounds the MGF of a *zero mean* and *bounded* r.v..
    - Suppose $\mathbb{E}[X] = 0$ and $P(X \in [a, b]) = 1$, then
$$M_X(s) := \mathbb{E}[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{8}} \text{ if } s > 0$$

    Proof: use Jensen's inequality followed by mean value theorem, see `http://www.cs.berkeley.edu/~jduchi/projects/probability_bounds.pdf`

19. Markov inequality and its implications
    (a) Markov inequality: for a non-negative r.v. i.e. for X for which $P(X < 0) = 0$
$$P(X > a) \leq \frac{\mathbb{E}[X]}{a}$$
    (b) Chebyshev inequality: apply Markov to $(Y - \mu_Y)^2$
$$P((Y - \mu_Y)^2 > a) \leq \frac{\sigma_Y^2}{a}$$
    if the variance is small, w.h.p. Y does not deviate too much from its mean
    (c) Chernoff bounds: apply Markov to e^{tY} for any $t > 0$.
$$P(X > a) \leq \min_{t>0} e^{-ta}\mathbb{E}[e^{tX}]$$
$$P(X < b) \leq \min_{t>0} e^{tb}\mathbb{E}[e^{-tX}]$$
    or sometimes one gets a simpler expression by using a specific value of $t > 0$

20. Using Chernoff bounding to bound $P(S_n \in [a,b])$, $S_n := \sum_{i=1}^n X_i$ when X_i's are iid

$$P(S_n \geq a) \leq \min_{t>0} e^{-ta} \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \min_{t>0} e^{-ta} (\mathbb{E}[e^{tX_1}])^n := p_1$$
$$P(S_n \leq b) \leq \min_{t>0} e^{tb} \prod_{i=1}^n \mathbb{E}[e^{-tX_i}] = \min_{t>0} e^{tb} (\mathbb{E}[e^{-tX_1}])^n := p_2$$

Thus, using the union bound with $A_1 = \{S_n < a\}$, $A_2 = \{S_n > b\}$

$$P(b < S_n < a) \geq 1 - p_1 - p_2$$

With $b = n(\mu - \epsilon)$ and $a = n(\mu + \epsilon)$, we can conclude that w.h.p. $\bar{X}_n := S_n/n$ lies b/w $\mu \pm \epsilon$

21. A similar thing can also be done when X_i's just independent and not iid. Sometimes have an upper bound for $\mathbb{E}[e^{tX_i}]$ and that can be used, for example Hoeffding lemma gives one such bound

22. Hoeffding inequality: Chernoff bound for sums of independent bounded random variables, followed by using Hoeffding's lemma

    - Given independent and *bounded* r.v.'s $X_1, \ldots X_n$: $P(X_i \in [a_i, b_i]) = 1$,

    $$P(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

    or let $\bar{X}_n := S_n/n$ and $\mu_n := \sum_i \mathbb{E}[X_i]/n$, then

    $$P(|\bar{X}_n - \mu_n| \geq \epsilon) \leq 2\exp(\frac{-2\epsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}) \leq 2\exp(\frac{-2\epsilon^2 n}{\max_i (b_i - a_i)^2})$$

    Proof: use Chernoff bounding followed by Hoeffding's lemma

23. Various other inequalities: Bernstein inequality, Azuma inequality

24. Weak Law of Large Numbers (WLLN) for i.i.d. scalar random variables, $X_1, X_2, \ldots X_n$, with finite mean μ. Define

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$$

For any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Proof: use Chebyshev if σ^2 is finite. Else use characteristic function

25. Central Limit Theorem for i.i.d. random variables. Given an iid sequence of random variables, $X_1, X_2, \ldots X_n$, with finite mean μ and finite variance σ^2 as the sample mean. Then $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution a Gaussian rv $Z \sim \mathcal{N}(0, \sigma^2)$

26. Many of the above results also exist for certain types of non-iid rv's. Proofs much more difficult.

27. Mean Value Theorem and Taylor Series Expansion

28. Delta method: if $\sqrt{N}(X_N - \theta)$ converges in distribution to Z then $\sqrt{N}(g(X_N) - g(\theta))$ converges in distribution to $g'(\theta)Z$ as long as $g'(\theta)$ is well defined and non-zero. Thus if $Z \sim \mathcal{N}(0, \sigma^2)$, then $g'(\theta)Z \sim \mathcal{N}(0, g'(\theta)^2\sigma^2)$.

29. If $g'(\theta) = 0$, then one can use what is called the second-order Delta method. This is derived by using a second order Taylor series expansion or second-order mean value theorem to expand out $g(X_N)$ around θ.

30. Second order Delta method: Given that $\sqrt{N}(X_N - \theta)$ converges in distribution to Z. Then, if $g'(\theta) = 0$, $N(g(X_N) - g(\theta))$ converges in distribution to $\frac{g''(\theta)}{2}Z^2$. If $Z \sim \mathcal{N}(0, \sigma^2)$, then $Z^2 = \sigma^2 \frac{g''(\theta)}{2}\chi_1^2$ where ch_1^2 is a r.v. that has a chi-square distribution with 1 degree of freedom.

31. Slutsky's theorem

2 Jointly Gaussian Random Variables

First note that a scalar Gaussian r.v. X with mean μ and variance σ^2 has the following pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Its characteristic function can be computed by computing the Fourier transform at $-t$ to get

$$C_X(t) = e^{j\mu t} e^{-\frac{\sigma^2 t^2}{2}}$$

jointly Gaussian r.v.'s. Any of the following can be used as a definition of j G.

1. The $n \times 1$ random vector X is jointly Gaussian if and only if the scalar

$$u^T X$$

   is Gaussian distributed for all $n \times 1$ vectors u

2. The random vector X is jointly Gaussian if and only if its characteristic function, $C_X(u) := \mathbb{E}[e^{iu^T X}]$ can be written as

$$C_X(u) = e^{iu^T \mu} e^{-u^T \Sigma u/2}$$

   where $\mu = \mathbb{E}[X]$ and $\Sigma = cov(X)$.

- Proof: X is j G implies that $V = u^T X$ is G with mean $u^T\mu$ and variance $u^T\Sigma u$. Thus its characteristic function, $C_V(t) = e^{itu^T\mu}e^{-t^2u^T\Sigma u/2}$. But $C_V(t) = \mathbb{E}[e^{itV}] = \mathbb{E}[e^{itu^TX}]$. If we set $t = 1$, then this is $\mathbb{E}[e^{iu^TX}]$ which is equal to $C_X(u)$. Thus, $C_X(u) = C_V(1) = e^{iu^T\mu}e^{-u^T\Sigma u/2}$.
- Proof (other side): we are given that the charac function of X, $C_X(u) = \mathbb{E}[e^{iu^TX}] = e^{iu^T\mu}e^{-u^T\Sigma u/2}$. Consider $V = u^TX$. Thus, $C_V(t) = \mathbb{E}[e^{itV}] = C_X(tu) = e^{iu^T\mu}e^{-t^2u^T\Sigma u/2}$. Also, $\mathbb{E}[V] = u^T\mu$, $var(V) = u^T\Sigma u$. Thus V is G.

3. The random vector X is jointly Gaussian if and only if its joint pdf can be written as

$$f_X(x) = \frac{1}{(\sqrt{2\pi})^n det(\Sigma)} e^{-(X-\mu)^T\Sigma^{-1}(X-\mu)/2} \tag{1}$$

   - Proof: follows by computing the characteristic function from the pdf and vice versa

4. The random vector X is j G if and only if it can be written as an affine function of i.i.d. standard Gaussian r.v's.

   - Proof uses 2.
   - Proof: suppose $X = AZ + a$ where $Z \sim \mathcal{N}(0, I)$; compute its c.f. and show that it is a c.f. of a j G
   - Proof (other side): suppose X is j G; let $Z := \Sigma^{-1/2}(X - \mu)$ and write out its c.f.; can show that it is the c.f. of iid standard G.

5. The random vector X is j G if and only if it can be written as an affine function of jointly Gaussian r.v's.

   - Proof: Suppose X is an affine function of a j G r.v. Y, i.e. $X = BY + b$. Since Y is j G, by 4, it can be written as $Y = AZ + a$ where $Z \sim \mathcal{N}(0, I)$ (i.i.d. standard Gaussian). Thus, $X = BAZ + (Ba + b)$, i.e. it is an affine function of Z, and thus, by 4, X is j G.
   - Proof (other side): X is j G. So by 4, it can be written as $X = BZ + b$. But $Z \sim \mathcal{N}(0, I)$ i.e. Z is a j G r.v.

Properties

1. If X_1, X_2 are j G, then the conditional distribution of X_1 given X_2 is also j G
2. If the elements of a j G r.v. X are pairwise uncorrelated (i.e. non-diagonal elements of their covariance matrix are zero), then they are also mutually independent.
3. Any subset of X is also j G.

3 Optimization: basic fact

Claim: $\min_{t_1,t_2} f(t_1,t_2) = \min_{t_1}(\min_{t_2} f(t_1,t_2))$

Proof: show that LHS $\geq$ RHS and LHS $\leq$ RHS

Let $[\hat{\hat{t}}_1, \hat{\hat{t}}_2] \in \arg\min_{t_1,t_2} f(t_1,t_2)$ (if the minimizer is not unique let $\hat{\hat{t}}_1, \hat{\hat{t}}_2$ be any one minimizer), i.e.

$$\min_{t_1,t_2} f(t_1,t_2) = f(\hat{\hat{t}}_1, \hat{\hat{t}}_2)$$

Let $\hat{t}_2(t_1) \in \arg\min_{t_2} f(t_1,t_2)$, i.e.

$$\min_{t_2} f(t_1,t_2) = f(t_1, \hat{t}_2(t_1))$$

Let $\hat{t}_1 \in \arg\min f(t_1, \hat{t}_2(t_1))$, i.e.

$$\min_{t_1} f(t_1, \hat{t}_2(t_1)) = f(\hat{t}_1, \hat{t}_2(\hat{t}_1))$$

Combining last two equations,

$$f(\hat{t}_1, \hat{t}_2(\hat{t}_1)) = \min_{t_1} f(t_1, \hat{t}_2(t_1)) = \min_{t_1}(\min_{t_2} f(t_1,t_2))$$

Notice that

$$\begin{aligned} f(t_1,t_2) &\geq \min_{t_2} f(t_1,t_2) \\ &= f(t_1, \hat{t}_2(t_1)) \\ &\geq \min_{t_1} f(t_1, \hat{t}_2(t_1)) \\ &= f(\hat{t}_1, \hat{t}_2(\hat{t}_1)) \end{aligned} \tag{2}$$

The above holds for all t_1, t_2. In particular use $t_1 \equiv \hat{\hat{t}}_1$, $t_2 \equiv \hat{\hat{t}}_2$. Thus,

$$\min_{t_1,t_2} f(t_1,t_2) = f(\hat{\hat{t}}_1, \hat{\hat{t}}_2) \geq \min_{t_1} f(t_1, \hat{t}_2(t_1)) = \min_{t_1}(\min_{t_2} f(t_1,t_2)) \tag{3}$$

Thus LHS $\geq$ RHS. Notice also that

$$\min_{t_1,t_2} f(t_1,t_2) \quad \leq \quad f(t_1,t_2) \tag{4}$$

and this holds for all t_1, t_2. In particular, use $t_1 \equiv \hat{t}_1$, $t_2 \equiv \hat{t}_2(\hat{t}_1)$. Then,

$$\min_{t_1,t_2} f(t_1,t_2) \quad \leq \quad f(\hat{t}_1, \hat{t}_2(\hat{t}_1)) = \min_{t_1}(\min_{t_2} f(t_1,t_2)) \tag{5}$$

Thus, LHS $\leq$ RHS and this finishes the proof.