Probability & Statistics Final Exam Note Sheet

Chapter 1: Probability

Replacement	Order	Number of Outcomes
With	Matters	n^r
Without	Matters	${}_nP_r = \frac{n!}{(n-r)!}$
With	Does Not Matter	${}_nC_r = \frac{n!}{(n-r)!r!} = \binom{n}{r}$

- *Conditional Probability:* $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. If i.i.d.: $P(A \mid B) = P(A)$.

Examples:

1. Four people are choosing an integer from $\{1, 2, \ldots, 10\}$ at random—assume that each choice is equally likely. What is the probability that all four choose different numbers? ***Sol.*** For all 4 people, each integer has a $\frac{1}{10}$ chance of being picked. That means there are $10^4 = 10000$ possible combinations. We calculate ${}_{10}P_4 = 5040$. We find $\frac{5040}{10000} = 50.4\%$.
2. A pair of fair, 6-sided dice are thrown. Find the probability that the sum has a total above 9. ***Sol.*** Each two pair outcome has a $\frac{1}{36}$ chance of being picked. There are one, two, and three ways to get a 12, 11, and 10, respectively. Thus, there is a $\frac{1+2+3}{36} = \frac{1}{6}$ chance of getting a sum above 9.
3. Three cards are drawn from a standard 52-card deck. Find the probability that at least one card is an Ace. ***Sol.*** The probability of getting 3 cards that have no Aces in them is $P(OOO) = \frac{48}{52} \cdot \frac{47}{51} \cdot \frac{46}{50} = \frac{4324}{5525}$. We can find the probability of at least one Ace by finding the complement of this probability: $P(\text{at least 1 Ace}) = P'(OOO) = 1 - \frac{4324}{5525} = \frac{1201}{5525}$. Therefore, we have approximately a $\boxed{21.73\%}$ chance of getting at least one Ace.
4. In a certain state, car license plates have the format of three letters followed by three numbers. How many possible license plates are there? ***Sol.*** Since we can use duplicate letters and numbers in the license plates, we find the total possible license plates with $26^3 \cdot 10^3 = 17576000$.
5. A fair coin is flipped three times. Given that you have at least one Head, find the probability that you have at least two Heads. ***Sol.*** Let event A be the probability of getting at least 2 Heads, and event B being at least 1 Head. The sample space consists of 8 outcomes, only one of which does not contain at least one head: $\{TTT\}$. Hence, $P(B) = \frac{7}{8}$. Then, there are 4 outcomes that contain at least one head, so $P(A) = \frac{4}{8} = \frac{1}{2}$. It follows that $P(A \cap B) = P(A)$ given that having at least two heads implies having at least one. Thus, $P(\text{At least } 2H \mid \text{At least } 1H) = \frac{P(A \cap B)}{P(B)} = \frac{1/2}{7/8} = \frac{4}{7}$.
6. Suppose a moment generating function for X is $M(t) = \frac{4}{9} + \frac{1}{3}e^t + \frac{2}{9}e^{3t}$. Find $P(X = 0)$, $P(X = 2)$, $E(X)$, $Var(X)$. ***Sol.*** From $M(t)$, we get $f(x) = \frac{4}{9}, \frac{1}{3}, \frac{2}{9}$. Thus, $P(X = 0) = \frac{4}{9}$, $P(X = 2) = 0$ (no prob for 0). To find $E(X)$, derive the mgf and set $t = 0$. For the variance, you get $E(X^2)$ from $M''(t = 0)$. Use that in the formula $E(X^2) - (\mu)^2$.

Chapter 1: (cont)

Examples:

1. An urn contains 5 Red, 7 Blue, and 1 White marbles. Two balls will be selected, without replacement.
   (a) Assuming that order matters, write the sample space for this experiment, and each outcome's probability. [Hint: If you draw a Red, followed by a White, you might write $(R, W), 0.12345$. A probability tree might be useful here.]

   Sol.

2. The flu comes in two forms: Mild and severe. 4% of students have mild, and 2% the severe. A test is 97% reliable, but cannot tell between strains.

   (a) Suppose you notice you have the Σs (i.e., you have either strain). What is the probability that you have the severe strain? ***Sol.***
   (b) Suppose you take the test and it comes back positive. What is the probability that you are infected with either strain? ***Sol.***
   (c) If you test is negative, what is the probability that you are infected with the severe strain? ***Sol.***

Chapter 1: (cont)

Examples:

1. An urn contains 10 marbles: 4 Red and 6 Blue. A second urn contains 16 Red marbles and an unknown number of Blue marbles. A single marble is drawn from each urn. The probability that both marbles are the same color is 0.44. How many blue marbles are there in the second urn? ***Sol.*** We have two urns: Urn_1: 4 Red, 6 Blue: 10 total, and Urn_2: 16 Red, b Blue: $16+b$ total. Let event A be getting both red. Thus, we find the probability to be: $P(A) = \frac{4}{10} \cdot \frac{16}{16+b}$. Similarly, let event B be getting both blue. Hence: $P(B) = \frac{6}{10} \cdot \frac{b}{16+b}$. Thus, the total probability is $P(A \cup B) = \frac{4}{10} \cdot \frac{16}{16+b} + \frac{6}{10} \cdot \frac{b}{16+b} = \frac{(64+6b)}{10(16+b)}$. Then, set the equation equal to 0.44 and solve for b.

Chapter 2: Discrete Distributions

Examples:

1. Consider a basketball player that makes only 60% of their shots. Assuming that each throw is independent of each other:
   (a) Find the probability that in 14 shots, they make exactly 11. ***Sol.*** Let X count the number of shots. $X \sim \mathit{bimnom}(14, .6, 11)$.
   (b) " 14 shots, make at least 9. ***Sol.*** $P(X \geq 9) = 1 - \text{binomcdf}(14, .6, 8)$.
   (c) They take 14 shots. Given that they make at least 9, find the probability that they make exactly 11. ***Sol.*** $P(X = 11 \mid X \geq 9) = \text{(a)}/\text{(b)}$. Since 11 is contained in $X \geq 9$, $P(A \cap B) = P(A)$.
   (d) Find the probability that they make their first shot in their 3rd attempt. ***Sol.*** Y counts num of attempts: $Y \sim \mathit{geometric}$. $P(Y = 3) = 0.096$.
   (e) Find probability that their 4th make occurs on 7th shot. ***Sol.*** $Z \sim NegBinom$. $\binom{7-1}{4-1} p^4 (1-p)^{7-4}$.
2. This course contains 20 total students, 8 of whom are declared math majors. If 5 are selected *without replacement* (hypergeometric), at random, find the probability that exactly 2 are math majors. ***Sol.*** 12 are non-math majors. $N_1 = 8$, $N_2 = 12$, $x = 2$, $n = 5$. Thus, $f(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$

Chapter 6: Point Estimation

Definitions:

- **pdf of X:** $f : S \to \mathbb{R}$ has the following properties:
  - $f(x) \geq 0$ for all $x \in S$; $\int_S f(x)\,dx = 1$; if $(a, b) \subseteq S$, then $P(a \leq X \leq b) = \int_a^b f(x)\,dx$
- **CDF of X:** $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$
- **Expected Value of X:** $E(X) = \mu = \int_S x f(x)\,dx$
- **Variance of X:** $\text{Var}(X) = \sigma^2 = \int_S (x-\mu)^2 f(x)\,dx = E(X^2) - \mu^2$
- **mfg of X:** $M_X(t) = E(e^{tX}) = \int_S e^{tx} f(x)\,dx$
- **Percentiles:** The pth percentile π_p is the value such that $P(X \leq \pi_p) = p$. Thus, $\int_{-\infty}^{\pi_p} f(x)\,dx = p$.

Chapter 3: Continuous Distributions

- **Integration by Parts:** $\int u\,dv = uv - \int v\,du$.
- **Gamma Function:** $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\,dt$, for $\alpha > 0$.
  - $\Gamma(x) = (x-1)\Gamma(x-1)$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(x) = \frac{\Gamma(x+1)}{x}$

Examples:

1. $\int_0^1 2xe^{-x^2}\,dx \implies u = x^2,\, du = 2x\,dx \implies \int_0^1 e^{-u}\,du = 1 - e^{-1}$
2. Consider the function $f(x) = cx^5$ on $[0, 1]$.
   (a) Find c. (Solve $\int_0^1 cx^5\,dx = 1$.)
   (b) Determine $E(X)$. (Solve $\int_0^1 x \cdot cx^5\,dx$.)
   (c) Determine $\text{Var}(X)$. (Solve $\int_0^1 (x-\mu)^2 \cdot cx^5\,dx$.)
   (d) Determine the 25th percentile. (Solve $\int_0^{\pi_{25}} cx^5\,dx = 0.25$.)
3. A pdf is given by $f(x) = \frac{1}{2}$. for $x \in [0, 1] \cup [2, 3]$, and 0 otherwise.
   (a) Determine $E(X)$. (Solve $\int_0^1 x \cdot \frac{1}{2}\,dx + \int_2^3 x \cdot \frac{1}{2}\,dx$. To do the variance, just replace x with x^2 and then subtract μ^2.)
4. The mean airspeed of an unladen swallow is 30 ft/s, with a std. dev. of 4.3 ft/s. Assuming the airspeed is normally distributed:
   (a) What is the probability that a randomly selected swallow has an airspeed between 28 and 38 ft/s? ($P(28 \leq X \leq 38) =$ `normalcdf(28, 38, 30, 4.3)`)
   (b) Suppose that 12 swallows are selected at random. What is the probability that exactly 8 have their speed between 28 and 38 ft/s? (Use binomial formula with $n = 12$, $k = 8$, and p from previous problem.)

Chapter 3 (cont.)

5. An auto insurance company insures an automobile worth \$15,000 for one year under a policy with a \$1,000 deductible. During the policy year there is a 0.04 chance of partial damage to the car and a 0.02 chance of a total loss of the car. If there is partial damage to the car, the amount X of damage (in thousands) follows a distribution with density function $f(x) = 0.5003e^{-x/2}$, for $0 \leq x \leq 15$. Find the expected value of the payment the insurance company makes in a year.
   Solution: Let E = expected yearly payment. There are 3 cases: 1. No damage: $P = 0.94$, payment $= 0$. 2. Total loss: $P = 0.02$, payment $= 14{,}000$. 3. Partial damage: $P = 0.04$, $X \sim f(x) = 0.5003e^{-x/2}$, $0 \leq x \leq 15$ (in thousands). The payment for $X > 1$ is $1000(X-1)$ and 0 otherwise. Expected payout for partial damage is $\int_1^{15} 1000(X-1)0.5003e^{-x/2}\,dx = 1211.96$. Thus, $E = 0.94 \cdot 0 + 0.02 \cdot 14000 + 0.04 \cdot 1211.96 = 328.48$.

Chapter 4: Bivariate Distributions

Definitions:

- **Joint pmf:** $f(x,y) = P(X = x, Y = y)$
- **Marginal pmf of X:** $f_X(x) = \sum_y f(x,y) = P(X = x)$ for $x \in S_X$
- **Mean of X_i:** $\mu_{X_i} = E(X_i) = \sum_{x_i} x_i f_{X_i}(x_i)$
- **Variance of X_i:** $\mathrm{Var}(X_i) = \sigma^2_{X_i} = \sum_{x_i}(x_i - \mu_{X_i})^2 f_{X_i}(x_i) = E(X_i^2) - \mu^2_{X_i}$
- **Covariance of X, Y:** $\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$. If $\mathrm{Cov}(X,Y) = 0$, then X and Y are uncorrelated; if pos, $X+$ and $Y+$, if neg, $X+$ and $Y-$. If independent, then $\mathrm{Cov}(X,Y) = 0$.
- **Correlation of X and Y:** $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$, or $\rho = 0$ if independent.
- **Conditional pmf of X given Y:** $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$ for $y \in S_Y$
- **Conditional Expectation of X given Y:** $E(X|Y = y) = \sum_x x f_{X|Y}(x|y)$
- **Conditional Variance of X given Y:** $\mathrm{Var}(X|Y = y) = E(X^2|Y = y) - (E(X|Y = y))^2$
- **Least Squares Regression Line:** $y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_x)$
- **Independence Criteria:** If $f(x,y) = f_X(x)f_Y(y)$ for all $x \in S_X$, $y \in S_Y$.

Chapter 4 (cont.)

Examples:

1. An actuary determines that the annual number of tornadoes in counties P and Q are jointly distributed as follows:

County P	County Q: 0	1	2	3
0	0.12	0.06	0.05	0.02
1	0.13	0.15	0.12	0.03
2	0.05	0.15	0.10	0.02

   (a) Determine the conditional expected number of tornados in county Q, given that there are no tornados in county P.
   $(Q|P = 0) = 0 \cdot \frac{0.12}{0.25} + 1 \cdot \frac{0.06}{0.25} + 2 \cdot \frac{0.05}{0.25} + 3 \cdot \frac{0.02}{0.25} = 0.88.$

   (b) Calculate the conditional variance of the annual number of tornadoes in county Q, given that there are no tornadoes in county P.
   $\mathrm{Var}(Q|P = 0) = E(Q^2|P = 0) - (E(Q|P = 0))^2 = 1.76 - 0.88^2 = 0.9856.$

   (c) Are counties P and Q independent? Justify your answer.
   No, because $f(0,0) = 0.12 \neq f_P(0)f_Q(0) = 0.25 \cdot 0.30 = 0.075$.

2. Let X be the weight of robin eggs, in grams, and Y be the daily high temperature, in degrees Celsius. Assume that X and Y have a bivariate normal distribution with $\mu_X = 145.2$, $\sigma^2_X = 109.2$, $\mu_Y = 23.5$, $\sigma^2_Y = 21.8$ and $\rho = -0.34$.

   (a) Find the least squares regression line for predicting Y from X.
   $y = 23.5 + (-0.34)\frac{\sqrt{21.8}}{\sqrt{109.2}}(X - 145.2) = 23.5 - 0.152(X - 145.2)$

   (b) Using the line from part (a), predict the daily high temperature when the weight of a robin egg is 150 grams.
   $y = 23.5 - 0.152(150 - 145.2) = 22.76$

   (c) Find the probability that a robin egg weights between 142 and 152 grams.
   $P(142 \leq X \leq 152) =$ `normalcdf(142, 152, 145.2, 10.45)`

   (d) Given $Y = 25.1$, find the probability that a robin egg weighs between 142 and 152 grams. (Use CDF formula w updated μ and σ:)
   $\mu_{X|Y=25.1} = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(25.1 - \mu_Y) = 143.9824$
   $\sigma^2_{X|Y=25.1} = (1-\rho^2)\sigma^2_X = 96.5765 \implies \sigma_{X|Y=25.1} = \sqrt{96.5765} = 9.82$

Chapter 5: Distributions of Functions of Random Variables

The formula for a z-score is $z = \frac{x-\mu}{\sigma}$. The formula for converting a z-score to an x-value is $x = \mu + z\sigma$.

Chapter 5 (cont.)

Examples:

1. Let X be a random variable with pdf $f(x) = cx^3$ for $0 < x < 2$. Let $Y = X^4$. Determine the pdf of Y. (cdf $= F(y) = P(Y \le y) = P(X^4 \le y) = P(X \le y^{1/4}) = F_X(y^{1/4})$, then differentiate to get pdf.)
2. Let X and Y be independent random variables with $f(x) = 2x, 0 < x < 1$ and $g(y) = 4y^3, 0 < y < 1$. Find $(0.5 < X < 1$ and $0.4 < Y < 0.8)$. (Since they are independent, just multiply their individual probabilities.)
3. Three basketball players each make 70% of their free throws. They each take 10 free throws. Assuming independence among the players, what is the probability that at least one player makes more than 8? (This is the complement of all 3 making 8 or fewer. Hence, $1-($`binomcdf(10, 0.7, 8)`$)^3 = 0.3844$.)
4. Suppose that each of $Z_1, Z_2, \ldots, Z_9$ are distributed as $N(0,1)$. Let $W = Z_1^2 + Z_2^2 + \ldots + Z_2^9$. Find $P(7.2 < W < 16)$. (Since W is chi-square with 9 degrees of freedom, use `x2cdf(7.2, 16, 9)`.)
5. 30 students roll a single 4-sided die until they get exactly 5 ones. Let $\overline{X}$ be the average number of rolls needed. Use the CLT to approximate $P(17 < \overline{X} < 19.5)$. (The number of rolls needed follows a negative binomial distribution with $p = \frac{1}{4}$ and $r = 5$. Thus, $\mu_X = \frac{r}{p} = 20$ and $\sigma_X^2 = \frac{r(1-p)}{p^2} = 60 \implies \sigma_X = \sqrt{60} = 7.746$. By the CLT, $\overline{X} \sim N(20, \frac{60}{30}) = N(20, 2)$. Hence, $P(17 < \overline{X} < 19.5) =$ `normalcdf(17, 19.5, 20, sqrt(2))` $= 0.3449$.)
6. A certain exam has a mean score of 73 and std. dev. of 6.8. Use CLT to approximate the probability that $n = 19$ students who took the exam have a mean larger than 75. (By CLT, $\overline{X} \sim N(73, \frac{6.8^2}{19}) = N(73, 2.434)$. Thus, $P(\overline{X} > 75) =$ `normalcdf(75, 1E99, 73, sqrt(2.434))` $= 0.0999$.)
7. A certain brand of LED lightbulbs advertise that they have a mean lifetime of 12.4 years, and we will assume their lifetime is exponentially distributed. You purchase a random sample of size $n = 34$, and find that the mean lifetime of your sample is 9.5 years. Use the Central Limit Theorem to (quantitatively) comment on how likely or unlikely this is? [Hint: Since the normal distribution is continuous, it does not make sense to talk about the probability that a sample has a particular mean. It might make sense to talk about the proportion of samples whose mean is at most some particular number, however.]
   We know that an individual bulb has a lifetime with $\mu = 12.4$ and $\sigma^2 = 12.4^2 = 153.76$, assuming that the company's claim is correct. The mean lifetime of our sample would then follow $N(12.4, 153.76/34)$. Thus, the probability of getting a sample with your mean lifetime or even shorter is given by `normalcdf(-1e99, 9.5, 12.4, 12.4/sqrt(34))` $= 0.08633$. This is reasonably unlikely, either you got a fairly unusual sample or the company's claim about the mean lifetime is incorrect.

Chapter 6: Point Estimation

Empirical Rule. Approximately 68%, 95%, and 99.7% live within 1, 2, and 3 standard deviations of the mean if the data is roughly bell shaped.
Order Statistics: Suppose X is a continuous random variable with CDF $F(x)$ and pdf $f(x)$ on the interval $x \in (a,b)$ and we select a random sample of size n. Then, the random variables $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics; that is, Y_1 is the smallest of X_i, Y_2 is the second smallest, Y_n is the largest.

The CDF of Y_r is: $G_r(y) = P(Y_r \le y) = \sum_{k=r}^{n} \binom{n}{k} [F(y)^k][1-F(y)]^{n-k}$.
The pdf of Y_r is: $g_r(y) = P(Y_r = y) = \frac{n!}{(r-1)!(n-r)!}[F(y)^{r-1}][1-F(y)]^{n-r} f(y)$.
Maximum Likelihood: Find $L(\theta) = \sum_{i=1}^{n} f(x_i, \theta)$ where $f(x_i, \theta)$ is from the distribution that you are looking for. Then, find $\frac{d}{d\theta} L'(\theta) = 0$.
Sample Mean: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. **S. Variance:** $s^2 = \frac{(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)}{n-1}$.

Examples:

1. Suppose that X has CDF $F(x) = x^2$ for $0 < x < 1$. Let $n = 6$ and Y_r be the 6 order statistics. Find $P(Y_4 < 0.7)$. (We can either use CDF as is, or use the pdf and integrate from 0 to 1. If asked for $E(Y)$, use pdf and add a y.)
   (a) *CDF Version:* $\displaystyle\sum_{k=4}^{6} {}_6C_k (.7^2)^k (1-.7^2)^{6-k}$.
   (b) *pdf Version:* $\displaystyle\int_0^{.7} \frac{6!}{3!2!} (y^2)^3 (1-y^2)^2 (2y)\, dy$.
2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta) = (1/\theta^2) x e^{-x\theta}$, for $0 < x < \infty$ and $0 < \theta < \infty$. Find $\hat{\theta}$.
   Solution. $L(\theta) = \displaystyle\prod_{i=1}^{n} (1/\theta^2) x e^{-x/\theta} = \left(\prod_{i=1}^{n} 1/\theta^2\right)\left(\prod_{i=1}^{n} x_i\right)\left(\prod_{i=1}^{n} e^{-x_i/\theta}\right) \Rightarrow$
   $\ln(L(\theta)) = -2n\ln(\theta) + \sum \ln(x_i) - \dfrac{1}{\theta}\sum x_i \Rightarrow \ln(L'(\theta)) = \dfrac{-2n}{\theta} + \dfrac{1}{\theta^2}\sum x_i$
   $\Rightarrow 0 = \ln(L'(\theta)) \Rightarrow \dfrac{2n}{\theta} = \dfrac{1}{\theta^2}\sum x_i \Rightarrow \theta = \left(\sum x_i\right)/2n \Rightarrow \theta = \overline{x}/2$.
3. Suppose $X \sim N(\mu, 1)$. Find $\hat{\mu}$.
   Solution. $L(\mu) = \displaystyle\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$
   $\Rightarrow \ln(L(\mu)) = \sum \ln\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)\right) = \sum\left(\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2}\right)$
   $\Rightarrow \dfrac{d}{d\mu} \ln(L(\mu)) = \sum (x_i - \mu) \Rightarrow 0 = n\overline{x} - n\mu \Rightarrow \mu = \overline{x}$.
4. Consider the data given by 6, 10, 14, 16, 18, 18, 18, 20. (Store in L_1 and run `1-Var Stats`)
   (a) Find the sample mean, $\overline{x}$: *Solution.* $1/8(6 + 10 + \cdots + 20) = 15$.
   (b) Find the sample standard deviation, s:
   Solution: $s^2 = \frac{1}{(8-1)}(6^2 + \cdots + 20^2 - 8(15)^2) = \frac{170}{7} \implies s = 4.78$