

Statistics 220A Topics

Autumn 2004

A list of the topics covered in Stat 120A appears below. Readings and homework problems are listed with each topic. While you should be able to do all the problems listed, you will only need to turn in a subset with your weekly homework. Details will be announced each week in class.

Basic Probability

1. Sample Spaces

- Reading: Sections 1.1 and 1.2.
- Homework
 - From the Text: Chapter 1: 5, 9.
 - Additional Problem A: A probability-minded despot offers her prisoner a final chance to gain his freedom. The prisoner is given 20 chips, 10 pink and 10 purple. All 20 are to be placed into two boxes, according to any allocation scheme the prisoner wishes, with the one proviso being that each box contains at least one chip. The executioner will then pick one of the two boxes at random and from that box, one chip at random. If the selected chip is pink, the prisoner will be freed, otherwise he will perish. Characterize the sample space by describing the possible allocation options. Intuitively, which allocation is most beneficial to the prisoner?

2. Probability Measures [Discrete and Continuous]

- Reading: Section 1.3.
- Homework
 - From the Text: Chapter 1: 7.
 - Additional Problem A: Consider the sample space $\Omega = \{2, 3, 4, \dots\}$, let $\Pr(\omega) = k \cdot (\frac{2}{3})^\omega$. For what value of k is $\Pr(\omega)$ a probability function?
 - Additional Problem B: Repeat Problem A with $\Omega = \{0, 1, 2, \dots\}$ and $\Pr(\omega) = k/(s^\omega \omega!)$.
 - Additional Problem C: Suppose X follows a standard normal distribution, compute (a) $\Pr(X > 0.5)$, (b) $\Pr(-1.5 < X < 1)$, and (c) the value of x such that $\Pr(X > x) = 0.05$.
 - Additional Problem D: Consider the model for describing human mortality,

$$f(t) = kt^2(100 - t)^2 \quad \text{for } 0 \leq t \leq 100,$$

where t denotes the age at which a person dies. Find k and $\Pr(\text{A Person lives past 25})$.

- Additional Problem E: Given that the

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right] \quad -\infty < x < \infty$$

is a probability function for any real μ and positive σ , evaluate

$$\int_0^\infty e^{-4x^2} dx.$$

[Hint: Use the basic properties of probability functions along with the fact that $f(x)$ is a probability function for any real μ and any positive σ .]

3. Conditional Probability

- Reading: Section 1.5.
- Homework: Chapter 1: 50, 51, 55, 56, 58, 60.

4. Independence

- Reading: Section 1.6.
- Homework
 - From the Text: Chapter 1: 65, 77, 79.
 - Additional Problem A: Prove that if $\Pr(A|B) = \Pr(A)$ then the events A and B are independent.
 - Additional Problem B: Jill, Donna, and Wilma have gotten in a disagreement over a male acquaintance, Charley, and decide to settle their dispute with a three-way pistol duel. Of the three Jill is the best shot, she never misses. Donna makes her shot half the time and Wilma only hits 30% of the time. The rules they agree to are simple: they are to fire at the target of their choice in succession, and cyclically in the order Wilma, Donna, Jill, Wilma, Donna, Jill, and so on until only one is left standing. Discuss Wilma's optimal strategy. Should she take aim at Jill, at Donna, or at the ground?

Discrete Probability Functions Based on Combinatorics

5. Combinatorics

- Reading: Sections 1.4, 1.4.1, 1.4.2.
- Homework: Chapter 1: 16, 31.

6. Combinatorial Probability

- Homework
 - From the Text: Chapter 1: 11, 14, 19, 20, 22, 24.
 - Additional Problem A: Five cards are dealt from a Poker deck. Compute the probabilities associated with each of the following hands: two pairs, three-of-a-kind, four-of-a-kind, flush (five cards of the same suit, but not having all consecutive denominations), and a royal flush (10, J, Q, K, and ace in the same suit).
 - Additional Problem B: A somewhat inebriated mathematician finds herself in the rather embarrassing situation of being unable to predict whether her next step will be forward or backward. What is the probability that after hazarding n such steps she will have stumbled forward a distance of r steps? [Hint: Let x denote the number of steps she will take forward and y the number backward. Then $x + y = n$ and $x - y = r$.]

7. The Hypergeometric Distribution

- Reading: Section 2.1.4.
- Homework
 - From the Text: Chapter 1: 30.
 - Additional Problem A: Describe how you might try to estimate the number of cars on the MBTA red line. Critique your plan in terms of its underlying model assumptions and practicality.

8. The Binomial Distribution

- Reading: Sections 2.1.1 and 2.1.2.
- Homework
 - From the Text: Chapter 2: 12, 13, 16.
 - Additional Problem A: One reason cited for the mental deterioration so often seen in the very elderly is the reduction in cerebral blood flow that accompanies the aging process. Addressing itself to this notion, a study was done (Ball and Taylor, 1969) to see whether cyclandelate, a vasodilator, might be able to stimulate the cerebral circulation and thereby slow the rate of mental deterioration. Blood circulation time can be measured using a radioactive tracer. Let X and Y be the mean blood circulation time *before treatment* and *after treatment* respectively, for a randomly selected elderly patient.

(a) Let

$$D = \begin{cases} 0 & \text{if } Y < X \\ 1 & \text{if } Y > X \end{cases}.$$

What is the distribution of D ?

- (b) Consider the skeptical hypothesis that cyclandelate has no effect on mean circulation time and any differences that are observed before and after treatment are due to chance. What is the distribution of D under this hypothesis?
- (c) Now suppose we select a random sample of n patients, and let $\Delta = \sum_{i=1}^n D_i$, what is the distribution of Δ under the skeptical hypothesis?
- (d) The drug was given to eleven subjects and blood flow was measured before and after treatment as described above. The data appear below. What is the observed value, δ , of Δ ? How likely is it that we would see a value as extreme or more extreme than this if the skeptical hypothesis were true? What do you conclude about the skeptical hypothesis?

Univariate Random Variables

9. Probability Density (Mass) Functions and Cumulative Distribution Functions

- Reading: Sections 2.1, 2.2, and 2.2.1.

- Homework

- From the Text: Chapter 2: 2, 34, 37, 40(Also: Suppose X_1, \dots, X_{10} are iid random variables with the given pdf. What is the probability that exactly four of these random variables are greater than 0.5?).
- Additional Problem A: Let Y be a random variable with pdf

$$f_Y(y) = \begin{cases} \frac{2}{3} & 0 \leq y \leq 1 \\ \frac{1}{3} & 3 \leq y \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Graph $f_Y(y)$ and $F_Y(y)$.

- Additional Problem B: Suppose $f_Y(y)$ is a symmetric and continuous pdf, that is $f_Y(y) = f_Y(-y)$, for $y > 0$. Show that $\Pr(-a < Y < a) = 2F_Y(a) - 1$ for $a > 0$.

10. The Poisson Distribution

- Reading: Section 2.1.5.

- Homework

- From the Text: Chapter 2: 26, 30, 42.
- Additional Problem A: The table below records the number of bombs that fell in each of 576 regions of the south of London during World War II. Equating the Poisson model average with the average of the data suggests setting $\lambda = 0.93$. Compute the expected frequencies under this model and comment on how well the Poisson model applies.

Number of hits	Frequency
0	229
1	211
2	93
3	35
4	7
5	1

- Additional Problem B: Suppose commercial airline crashes occur at a rate of about 2.5 per year in the US. Using the Poisson model, find the probability of five or more crashes next year. Also, find the probability that the next two crashes will occur within two months of each other.

11. The Gamma Distribution

- Reading: Section 2.2.2.
- Homework: Chapter 2: 49.

12. The Geometric and Negative Binomial Distributions

- Reading: Section 2.1.3.
- Homework: Chapter 2: 22, 24.

13. The Normal Distribution

- Reading: Section 2.2.3.
- Homework
 - From the Text: Chapter 2: 53, 54.
 - Additional Problem A: At a certain Ivy League University, the average score of freshmen students on the verbal part of the SAT is 570, with a standard deviation of 70. Using the normal model, what proportion of students have verbal SAT scores over 640? Under 450? Find the sixtieth percentile of verbal SAT scores.
 - Additional Problem B: About 75% of 20 year old women weigh between 103.5 and 148.5 lb. Using the normal model, and assuming that 103.5 and 148.5 are equal distant from the mean, μ , calculate σ .
 - Additional Problem C: Traffic fatality rates (deaths per 100 million motor-vehicle miles) for each of the fifty states are given in the table below. Make a histogram of these observations using classes 2.0-2.9, 3.0-3.9, etc. Using the normal model (with $\mu = 5.3$ and $\sigma = 1.3$) compute the expected frequency for each class and comment on the fit of the normal model for this data.

AL	6.4	LA	7.1	OH	4.5
AS	8.8	ME	4.6	OK	5.0
AR	6.2	MA	3.5	OR	5.3
AK	5.6	MD	3.9	PA	4.1
CA	4.4	MI	4.2	RI	3.0
CO	5.3	MN	4.6	SC	6.5
CN	2.8	MS	5.6	SD	5.4
DE	5.2	MO	5.6	TN	7.1
FL	5.5	MT	7.0	TX	5.2
GA	6.1	NC	6.2	UT	5.5
HI	4.7	ND	4.8	VA	4.5
ID	7.1	NE	4.4	VT	4.7
IN	4.3	NV	8.0	WV	6.2
IL	5.1	NH	4.6	WA	4.3
IA	5.9	NJ	3.2	WI	4.7
KA	5.0	NM	8.0	WY	6.5
KY	5.6	NY	4.7		

14. The Expected Value

- Reading: Section 4.1; pages 111-117.
- Homework
 - From the Text: Chapter 4: 3.8: 12, 20.
 - Additional Problem A: Show that

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad -\infty < x < \infty$$

is a pdf, but has no mean.

- Additional Problem B: There are six barstools at The Tasty. The fry cook predicts that if two strangers come into the bar they will sit in such a way as to leave at least two stools between them. Suppose two strangers come in and choose their seats at random. What is the probability that the fry cook’s prediction will come true? Also compute the expected number of seats between the two customers under this assumption.
- Additional Problem C: Suppose $X \sim \text{Gamma}(\alpha, \beta)$, compute $E(X^n)$ for $n = 1, 2, \dots$

15. The Variance

- Reading: Sections 4.2 and 4.2.1.
- Homework
 - From the Text: Chapter 4: 4, 16.
 - Additional Problem A: Referring to Additional Problem A under 13 above, if there are 1000 freshmen, what is the mean and standard deviation of the number of students scoring between 550 and 570, inclusive?
 - Additional Problem B: Compute the variance of a Gamma random variable with parameters α and β .

16. Univariate Transformations

- Reading: Sections 2.3 and 2.4.
- Homework
 - From the Text: Chapter 2: 60, 64, and 67 (hint: do additional problem C first).
 - Additional Problem A: Suppose that the random variable X takes on $-3, 1, 5$ each with probability one third. Find the pmf of $y = 3X - 4$.
 - Additional Problem B: Suppose X is a continuous rv, g is a continuous strictly increasing function or a continuous strictly decreasing function, and $Y = g(X)$. Show

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

Hint: Consider increasing and decreasing functions separately.

- Additional Problem C: Use Problem B to verify that $Y \sim \text{exp}(\lambda)$ if $Y = -\log(X)/\lambda$ with $X \sim \text{Unif}(0, 1)$.
- Additional Problem D: Suppose $X \sim N(\mu, \sigma^2)$ and $Y = a + bX$. Find the distribution of Y . Hint: Use the result in Problem B.

Multivariate Random Variables

17. Joint Distributions

- Reading: Sections 3.1, 3.2 and 3.3.
- Homework: Chapter 3: 3, 6, 8a, 8b, 19.

18. Independence

- Reading: Section 3.4.
- Homework: Chapter 3: 14a, 15a-d.

19. More on Expectations

- Reading: pages 117-118.
- Homework
 - From the Text: Chapter 4: 22, 30.
 - Additional Problem A: Two fair dice are tossed one time. Let X denote the number of 3’s that appear and Y , the number of 4’s. Let $Z = XY^2$. Find $E[g(X, Y)]$ two ways.

20. The Covariance

- Reading: Section 4.3.
- Homework: Chapter 4: 39, 42, 44, 53.

21. The Multivariate Normal Distribution

- Reading: Section 3.3: Example F and Section 4.3: Example E.
 - Additional Problem A: Suppose

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\mu = \begin{pmatrix} 2 \\ -5 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 4 \end{pmatrix} \right).$$

Compute $\Pr(X_1 > 0)$ and $\Pr(X_2 < -6)$.

- Additional Problem B: Suppose X_1 and X_2 follow the Bivariate Normal Distribution and that $\text{Cov}(X_1, X_2) = 0$. Show that X_1 and X_2 are independent.
- Additional Problem C: If the joint pdf of the random variables X and Y is

$$f_{X,Y}(x,y) = ke^{-(2/3)[(1/4)x^2 - (1/2)xy + y^2]}$$

find $E(X)$, $E(Y)$, $\text{Var}(X)$, $\text{Var}(Y)$, $\text{Cov}(X, Y)$ and k .

22. Multivariate Transformations:

- Reading: Sections 3.6.1 and 3.6.2.
- Homework
 - From the Text: Chapter 3: 34, 36, 44, 52, 54.
 - Additional Problem: Recall the light bulb example. Let X be the lifetime of a light bulb and suppose $f_X(x) = \frac{1}{\lambda}e^{-x/\lambda}$ (you may assume $\lambda = 500$). We have discovered that $Y = 1/X$, the cost per hour of the light bulb has infinite expectation. Suppose we buy two light bulbs with lifetimes X_1 and X_2 , which are independent and distributed as X . Find the distribution of $W = X_1 + X_2$, and the $E(2/W)$. What do you conclude?

23. Conditional Distributions and Regression

- Reading: Sections 3.5.1, 3.5.2, 4.4.1, and 4.4.2.
- Homework
 - From the Text: Chapter 3: 1b, 10, 14b, 21, 23, and Chapter 4: 62, 64, and 71.
 - Additional Problem A: Suppose $X \sim N(\mu, \sigma^2)$ and $Y|X \sim N(\alpha + \beta X, \tau^2)$. What is the joint distribution of X and Y ? What is the marginal distribution of Y ? What is the conditional distribution of X given Y .
 - Additional Problem B: Income (Y) and Years of Education (X) for individuals can be modeled as bivariate normal random variables. For white men age 35-54 in 1988, $\mu_X = 13.5$, $\sigma_X = 3$ (in years), $\mu_Y = 33,900$, $\sigma_Y = 22,000$ (in dollars), and $\rho = 0.39$. For black men age 35-54 in 1988, $\mu_X = 12.3$, $\sigma_X = 3$ (in years), $\mu_Y = 24,200$, $\sigma_Y = 16,000$ (in dollars), and $\rho = 0.42$. Compute the conditional distribution of income as a function of years of education for both groups. What are the values of α , β , and σ and what do they mean? What can you conclude based on the differences in the two conditional distributions?
 - Additional Problem C: The number of offspring of an organism is a discrete random variable with mean μ and variance σ^2 . Each of its offspring reproduce in the same manner. Find the expected number of offspring in the third generation and its variance.
 - Additional Problem D: Suppose $X \sim \text{gamma}(r, \lambda)$ and $Y|X \sim \text{Poisson}(X)$. Find the conditional distribution of X given Y (this is a standard distribution). What are the conditional mean and variance of X ? How do they compare with the unconditional mean and variance? Find is the marginal distribution of Y (this is also a standard distribution).

Linear Combinations of Random Variables

24. Means

- Reading: Section 4.1.2.

25. Variances

- Reading: pages 4.2.1.
- Homework
 - From the Text: Chapter 4: 45, 46, 49, 50.
 - Additional Problem A: Suppose

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\mu = \begin{pmatrix} 2 \\ -5 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 4 \end{pmatrix} \right).$$

Compute $\Pr(X_1 + X_2 > -2)$.

- Additional Problem B: Suppose (X_1, X_2, \dots, X_n) are an iid sample from $N(\mu, \sigma^2)$. Compute the expectation and variance of the estimate $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. How does this estimate compare with S^2 ? Which estimate do you prefer?

26. Distributions (The Central Limit Theorem)

- Reading: Sections 4.5, 5.3.
- Homework: Chapter 4: 75, 76, 77, 78; Chapter 5: 6, 12, 13, 14, 18.

27. Distributions derived from the Normal Distribution.

- Reading: Sections 6.1, 6.2, and 6.3.
- Homework
 - From the Text: Chapter 6: 4, 6, 10.
 - Additional Problem A: A manufacturer of small appliances employs a market research firm to estimate retail sales of products by gathering information from a sample of retail stores. This month a sample of 75 stores finds these stores sold an average of 158 of the manufacturer's power drills with standard deviation 25. Do you believe that sales have increased since the same month last year when the average number sold in all stores that carried the product was 150?

28. Some Computer Problems

- Homework
 - Problem A: You will find the file `football.dct` on the Stat 120 home page. The file contains several variables pertaining to the 1989-91 seasons of the National Football League. We are interested in the two variables `pts` and `diff`. The variable `pts` records the odds-makers "point spread" for each game, which is a value assigned before each game to serve as a handicap for whichever is perceived to be the better team. Thus, to win against the point spread, the "favorite" team must beat the "underdog" team by more points than the spread. The underdog "wins" against the spread if it wins the game outright or manages to lose by fewer points than the spread. In theory, the point spread should represent the "expert" prediction as to the game's outcome. In practice, it more usually denotes a point at which an equal amount of money will be wagered both for and against the favored team. The variable `diff` represents the difference between the score of the favored team and the underdog in the actual game. Use Stata to fit a bivariate normal model to these data. (You may assume the parameter values are equal to the sample values.) Find the conditional distribution of `diff` given `pts`. Does the model seem appropriate? What is your evidence? What do you conclude about the relationship between `diff` and `pts`?

To read the data into Stata, type

```
infile using football.dct
```

at the prompt. **If you are using a MAC**, the file name is more complicated, and you should select Filename from the File menu after typing `infile` using

- You can then find the file in the dialog box that appears. Additional Stata commands appear on the last two pages of this homework handout and in the STATA handouts that are available on the Stat 120 homepage.
- Problem B: Suppose X follows the gamma distribution with parameters r and λ . The pdf of X is

$$f_X(x) = \frac{\lambda^r}{r!} x^{r-1} e^{-\lambda x} \text{ for } x > 0.$$

For this distribution, $E(X) = r/\lambda$ and $Var(X) = r/\lambda^2$. Simulate 1000 iid observations from this distribution with $\lambda = 2$ and r equal to each of 1, 5, 25, and 100. What do you notice about the distribution of X as r increases? [Describe the shape, center and spread.] Why might you expect this behavior? [Hint: If Y_1, \dots, Y_r are iid exponential random variables with parameter λ , $\sum_{i=1}^r Y_i \sim \text{gamma}(r, \lambda)$.] In Stata, λ is the scale parameter, and r is the shape parameter.

Convergence Theorems and Inequalities

29. Inequalities

- Reading: pages 125-126.
- Homework
 - From the Text: 38.
 - Additional Problem A: A fair die is tossed 100 times. Let X_k denote the outcome on the k th roll. Find a lower bound on the probability that $X = \sum_{k=1}^{100} X_k$ is between 300 and 400. Also use the CLT to approximate this probability. How do the answers compare?
 - Additional Problem B: Suppose X and Y are two random variables each with finite mean and variance. Prove $-1 \leq \rho_{XY} \leq 1$ by using the fact that

$$Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \text{ and } Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$$

are both positive quantities.

30. Modes of Convergence and the Law of Large Numbers

- Reading: Sections 5.1 and 5.2.
- Homework: Chapter 5: 1, 17, 19, 20, 22.