Information about the Quiz

Also some practice problems

Autumn 2023

1. As we agreed, your individual assessment will be composed of a 50 minute quiz. **The quiz is scheduled to be held on Friday Dec 8 from 11:30 - 12:20 PM at ARC 160.**
2. You will need to bring your student ID, writing instruments and a scientific calculator. You will not be asked to do very elaborate calculations on the test. If a question requires a numeric answer that involves a laborious calculation, you can always leave the answer in its final form with all the numbers plugged in, but you don't need to actually perform the calculation (unless I specifically ask you to).
3. You will not be asked to write code on the test. I may however, have a few small parts where you might be required to fill in blanks. You should also be familiar with the syntax of R functions like "d","p", "q" for finding the density, CDF and the percentiles of the "celebrity" distributions.
4. The quiz will focus on discrete and continuous random variables. **This is covered in chapters 5 through 13 of our notes.** Please be aware that there is an inherently cumulative nature to mathematical topics, and so I am still expecting you to remember the content from the earlier chapters. I will not be directly testing you on them though.
5. You are not allowed to bring any notes/formula sheet. I have created a formula sheet which can also serve as a reading guide to facilitate your review of course content. This sheet along with some results on sums/series will be printed along with your test for you to refer to.[1] In addition, you will also get scratch paper.
6. Any questions on the format/content of the quiz must be posted publicly on Ed. We will not answer questions that are sent to our NETIDs or marked as private.

[1] I am not responsible for any typos. If you find any, please let me know by Wed Dec 6 (in class) at the latest so we can fix it before we print. You are responsible for checking that everything is accurately represented in this guide.

FORMULA SHEET (also use as reading guide)

The following is a summary of formulas we have seen this quarter. You will receive a copy of this summary with the final. You are not allowed to bring any notes/cheat sheets etc.

1. **Discrete Distributions (§ 5)**

a. A random variable is a *function* which maps each outcome in a sample space to a number. More informally, it is a variable whose value depends on the outcome of a random experiment.

Notation: uppercase X denotes the random variable as a function, lowercase x denotes a possible value or number.

b. PMF: probability of observing a specific value x

$$f(x) = P(X = x)$$

c. CDF: accumulated probability up till a specific value x

$$F(x) = P(X \leq x)$$

.....

d. Mean and variance (§ 7)

i. Mean: a number which represents the **average** value of random variable across separate replications of the experiment

Definition	$\mu = E[X] = \sum_{-\infty}^{\infty} x \cdot f(x)$
Linearity of Expectation	$E[aX + b] = aE[X] + b$
Law of the Unconscious Probabilist	$E[t(X)] = \sum_{-\infty}^{\infty} t(x) \cdot f(x).$

ii. Variance: a positive number which describes spread of the values of the random variable from the mean.

Definition	$\sigma^2 = Var[X] = \sum_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)$
Short cut for calculation	$\sigma^2 = E\left[X^2\right] - \mu^2.$
Variance of linear transformation	$Var[aX + b] = a^2 \cdot Var[X]$

iii. Standard deviation: positive square root of variance which is on the same units as data. It is interpretable as the *typical* deviation of the values from the mean.

iv. Chebychevs's inequality: a useful inequality which provides an upper bound for the probability that a random variable can be more than k standard deviations from the mean.

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

.....

d. Binomial Random Variable (§ 6)

i. A binomial random variable counts the number of successes in n independent trials where each trial results in a success with probability π or in a failure with probability $1 - \pi$. We write $X \sim Binom(n, \pi)$.

ii. Binomial PMF

$$f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, \ldots, n$$

iii. Mean of $X \sim Binom(n, \pi)$: $n\pi$

iv. Variance of $X \sim Binom(n, \pi)$: $n\pi(1-\pi)$

v. Relevant R functions:
- `dbinom(x, size, prob)` calculates $f(x) = P(X = x)$
- `pbinom(q, size, prob)` calculates $F(q) = P(X \leq q)$
- `pbinom(q, size, prob,lower.tail = F)` calculates $P(X > q)$.

e. Geometric random variable (§ 8.1)

i. A geometric random variable counts the number of failures *before* we see the first success when independent trials with probability π of observing a success are performed. We write $X \sim Geom(\pi)$.

ii. Geometric PMF

$$f(x) = \pi(1-\pi)^x, \ x = 0, 1, 2, \ldots$$

iii. For any non-negative integer k, we have the result

$$P(X \geq k) = (1-\pi)^k$$

iv. A geometric distribution is *memoryless*. This means for all non-negative integers x, k

$$P(X \geq x + k | X \geq k) = P(X \geq x)$$

v. Mean of $X \sim Geom(\pi)$: $\frac{1-\pi}{\pi}$.

vi. Relevant R functions:
- `dgeom(x, prob)` calculates $f(x) = P(X = x)$
- `pgeom(q, prob)` calculates $F(q) = P(X \leq q)$
- `pgeom(q, prob,lower.tail = F)` calculates $P(X > q)$.

f. Poisson (§ 8.2)

i. The Poisson random variable counts the number of occurrences of an event over a fixed time period or within a space. We write $X \sim Poisson(\lambda)$ where λ denotes the rate of occurrence.

ii. The PMF of a Poisson can be derived from a $Binom(n, \pi)$ by setting $\pi = \frac{\lambda}{n}$ in the binomial PMF and letting $n \to \infty$.

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2 \ldots$$

iii. Mean of $X \sim Pois(\lambda)$: λ

iv. Variance of $X \sim Pois(\lambda)$: λ

v. Relevant R functions:
- `dpois(x, lambda)` calculates $f(x) = P(X = x)$
- `ppois(q, lambda)` calculates $F(q) = P(X \leq q)$
- `ppois(q, lambda,lower.tail = F)` calculates $P(X > q)$.

---

2. **Continuous Distributions**

a. PDF and CDF (§ 9)

i. The PDF is any function which satisfies two properties:

$$f(x) \geq 0 \ \forall \ x, \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

ii. Probabilities are calculated as areas under the PDF:

$$P(a \leq X < b) = \int_a^b f(x)dx.$$

Since a single value has no area, $P(X = x) = 0$ for any x however.

iii. The CDF $F(x)$ is again the accumulated probability up til a value x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

The CDF has the following properties:

- it is non-decreasing
- it is right continuous
- $\lim_{x\to\infty} F(x) = 1$
- $\lim_{x\to-\infty} F(x) = 0$

iv. By the Fundamental Theorem of Calculus, we can write

$$f(x) = \frac{d}{dx}F(x).$$

. . . . . .

b. Mean and variance and higher moments (§ 12)

i. Mean: $\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$

ii. Variance: $\sigma^2 = Var[X] = E\left[X^2\right] - \mu^2$

iii. The results stated in 1d. for Discrete Distributions hold in the continuous case as well.

iv. In addition to the mean and variance, we can also calculate percentiles for a continuous distribution.

- The $100p$ percentile of a continuous distribution is the number q such that $P(X < q) = p$

. . . . . .

c. Uniform random variable (§ 10)

i. The uniform random variable is the continuous analog of the equally likely model in a discrete sample space. We write $X \sim Unif(a, b)$.

ii. PDF of a uniform

$$f(x) = \frac{1}{b-a}, \quad a \leq x < b$$

iii. Mean of $X \sim Unif(a, b)$: $(a + b)/2$

iv. Variance of $X \sim Unif(a, b)$: $(b - a)^2/12$

v. The $100p$th percentile of $X \sim Unif(a, b)$ is given by $a + (b - a) \times p$

vi. Relevant R functions: For $X \sim Unif(min, max)$

- `dunif(x, min, max)` calculates PDF $f(x)$
- `punif(q, min, max)` calculates $F(q) = P(X \leq q)$
- `punif(q, min, max, lower.tail = F)` calculates $P(X > q)$.
- `qunif(p, min, max)` calculates the $100p$th percentile

. . . . . .

d. Exponential random variable (§ 11)

i. The exponential distribution arises as the inter-event time in a Poisson model. However, it can be used as a model for any non-negative random variable! We write $X \sim Exp(\lambda)$ where $\lambda(> 0)$ is called the *rate* parameter.

ii. PDF of an exponential random variable:

$$f(x) = \lambda e^{-\lambda x}, \quad 0 \leq x < \infty$$

iii. CDF of an exponential random variable:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & 0 \leq x \end{cases} \end{aligned}$$

iv. The exponential distribution is *memoryless*: this means for $x, k > 0$ we have the result:

$$P(X \geq x + k | X \geq k) = P(X \geq x)$$

v. Mean of $X \sim Exp(\lambda)$: $\frac{1}{\lambda}$
vi. Variance of $X \sim Exp(\lambda)$: $\frac{1}{\lambda^2}$
vii. The $100pth$ percentile of $X \sim Exp(\lambda)$ is given by $-\frac{1}{\lambda}\ln(1-p)$.
viii. Relevant R functions: For $X \sim Exp(rate)$
- `dexp(x, rate)` calculates PDF $f(x)$
- `pexp(q, rate)` calculates $F(q) = P(X \leq q)$
- `pexp(q, rate, lower.tail = F)` calculates $P(X > q)$
- `qexp(p, rate)` calculates the $100p$th percentile

. . . . . .

e. Normal random variable (§ 13)
i. The normal distribution is often used as a model for biological measurements such as height, weight etc. It is also the limiting distribution for other models, such as the binomial, Poisson, etc. We write $X \sim Norm(\mu, \sigma)$.
ii. We can write $X = \mu + \sigma\, Z$ where $Z \sim Norm(0, 1)$.
iii. PDF of a normal:

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \qquad -\infty < x < \infty$$

iv. The mean of $X \sim Norm(\mu, \sigma)$: μ
v. The variance of $X \sim Norm(\mu, \sigma)$: σ^2.
vi. The 68-95-99.7 rule states that regardless of the value of μ and σ the area within 1/2/3 standard deviations of the mean is 68%/95%/99.7%.
vii. The $100p$th percentile of $X \sim Norm(\mu, \sigma)$ is $\mu + \sigma q$ where q is the corresponding percentile for the standard normal distribution.
viii. Relevant R functions: For $X \sim Norm(\mu, \sigma)$
- `dnorm(x, mean, sd)` calculates PDF $f(x)$
- `pnorm(q, mean, sd)` calculates $F(q) = P(X \leq q)$
- `pnorm(q, mean, sd, lower.tail = F)` calculates $P(X > q)$.
- `qnorm(p, mean, sd)` calculates the $100p$th percentile

Sums and Series

Binomial Theorem For any real numbers a and b and integer $n > 0$

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

Geometric Series For any real numbers a and r ($|r| < 1$)

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$$

Taylor series for e^x:

$$e^x = 1 + x + x^2 + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

The following questions are for practice. Not all the problems are a reflection of what is reasonable for me to ask on a timed test. They are just here to improve your problem solving skills. We will work on them in class starting Friday 12/1. Please also review as many chapter quizzes as you have time for.

1. Suppose $X \sim Binom(n, \pi)$.

a. Write the PMF of $Y = n - X$.
b. What is $E[Y]$? How about $SD[Y]$?
c. Suppose $\pi = \frac{3}{4}$. Use Chebychev's inequality to find the smallest n for which $P\left(0.9 \times \frac{n}{4} < X < 1.1 \times \frac{n}{4}\right)$ to be at least 90%.

2. Fred wants to know if his cat, Gus, prefers his right paw or if he uses both paws equally. So he dangles a ribbon in front of Gus and notes which paw Gus uses to bat at it.

   He does this 10 times and Gus bats at it with his right paw 8 times and his left paw 2 times. Then Gus gets bored and leaves.

   Let the random variable X denote the number of times that Gus batted with his right paw in 10 trials. (We observe $x = 8$). Suppose you decide to model X as a binomial random variable with $\pi = 0.5$ (meaning Gus is equally likely to use either paw).

a. What assumptions do you need to make in order for the binomial model to be a reasonable choice for this setting?

b. How unusual is the observed data under the presumed model? Give the code for calculating $P(X \geq 8)$.

c. Suppose the probability from part b is 0.0547. Do you believe the presumption that Gus is equally likely to use either paw? Why or why not?

3. A condition C among new-born babies occurs apparently independently in any baby with probability $p = 0.15$. Suppose in King County in 2014, there are 10,000 births.

a. Which probability distribution provides the most accurate model for computing the probability that more than 1600 babies with condition C are born in the county K in 2014? Explain your thinking. (You would not need to calculate the probability on the test, but go ahead and calculate it in R since this is just for practice)

b. What other probability distribution might be more computationally convenient, and would provide a good approximation for the probability in part a? Justify your thinking. (You would not need to calculate the probability on the test, but go ahead and calculate it since this is just for practice)

4. Karen is studying for a history exam, where the teacher is going to choose 5 essay questions randomly from the 10 he has given the class. Due to an upcoming probability exam, she only has limited time to prepare for the history exam. Suppose she decides to study 8 out of the 10 questions.

a. Let the random variable X denote the number of questions on the exam which Karen has studied. What is the range of X?

b. Calculate the probability that $X = 4$.

5. In a certain country commercial airplane crashes occur according to a Poisson process at the rate of 2.5 per year.

   For each part, clearly define the random variable and state its distribution and the probability you are trying to calculate BEFORE you start doing calculations.

a. Find the probability that the next two crashes will occur within three months of each other.

b. Given that the next crash occurs sometime in the next year (12 months), what is the probability that it occurs within the first three months?

6. The annual rainfall (in inches) in a certain region is normally distributed with $\mu = 40$ and $\sigma = 4$. What is the probability that, starting with this year, it will take over 10 years before a year occurs having a rainfall of over 50 inches? What assumption are you making? You are given that `pnorm(2.5) = 0.994`

   Hint: Let X denote the number of years before a year occurs having rainfall of over 50 inches. What is a reasonable distribution to assume for X?

7. Suppose $X \sim Exp(1)$, that is it has PDF

$$f_X(x) = e^{-x} \quad 0 \leq x < \infty$$

   What distribution does the random variable $Y = \frac{X}{\lambda}$ have? State the name of the distribution and also the value for any parameters. (Hint: Find the CDF of Y and then differentiate it to find a PDF)

8. Which distribution has the smaller 25th percentile? The $Unif(0, 1)$ or the $Exp(1)$?

9. The internal temperature in a gizmo is a random variable X with PDF (in appropriate units)

$$f(x) \quad = \quad \begin{cases} 11(1-x)^{10} & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

   The gizmo has a cutoff feature, so that whenever the temperature exceeds the cutoff (call it k), the gizmo turns off. It is observed that the gizmo shuts off with probability 10^{-22}. What is k?

10. Suppose $X \sim Pois(\lambda)$. That is, it has PMF

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad x = 0, 1, 2, \ldots$$

    Find $E\left[\frac{1}{X+1}\right]$.