## **Exam 2 review** (draft: 2019/04/07-22:31:25)

Exam 2 covers all material since the first exam. You will still need to remember much of the material since the first day of class though, including, but not limited to, many of the rules of probability.

My recommendation is to study the two quizzes, make sure you understand everything on those perfectly. Make sure you have looked over any example problems included in the course notes that I have posted online (for chapters 3 and 4). Make sure you understand the webwork homework problems. (also note that the chapter numbers for the webwork assignments do not necessarily align with the chapter numbers for my notes). Of course, you should also understand everything and all examples covered in class. Sometimes these vary somewhat form what is seen on homework or in my online notes. You should also pay attention to what you did in the R Activity assignments.

Here is a summary list of topics (note that this list is not completely exhaustive):

- Random variables (discrete and continuous)
- Probability distributions (pmf, pdf, cdf)
- Expected value and variance of RVs
- Binomial, geometric, Poisson, exponential, normal
- Central limit theorem and law of large numbers

Example/practice problems:

1. Flip a biased coin with probability of heads being 0.2. Refer to getting heads as a success.

   (a) If we flip the coin 20 times, what is the probability of getting 13 heads?
   *Solution:*
   Let $X = \{\# \text{ of heads (successes)}\}$ so we have that $X \sim Bin(p = 0.2)$. The pmf for binomial is:

   $$P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

   Thus

   $$P(X = 13) = f(13) = \binom{20}{13} (0.2)^{13} (0.8)^7 \approx 0.00001331783.$$

   (b) Now, let's flip the coin until the first head occurs. Write the probability mass function for $X = \{\text{the total number of flips up to and including the first head}\}$. What is the probability that there are 5 tails before the first head?
   *Solution:*
   Now, this random variable is geometric, $X \sim Geom(p = 0.2)$. The pmf is

   $$P(X = x) = p(1-p)^{x-1}.$$

   You don't necessarily have to have this memorized if you understand that it is a sequence of $x - 1$ failures and then a single success, and that the trials are independent. Getting 5 tails before the first head would be 6 total trials, thus $X = 6$:

   $$P(X = 6) = 0.2(0.8)^5 \approx 0.065536.$$

2. Consider a population of animals where adult weight is normally distributed with mean 7.3 kg and standard deviation 1.8 kg.

   (a) What is the probability that a randomly selected individual weighs less than 5 kg?

   *Solution:*

   $P(X < 5) =$ `pnorm(5,7.3,1.8)` $\approx 0.1006639$.

   (b) What is the probability that a randomly selected individual's weight will be between 8.2 and 9.5 kg?

   *Solution:*

   $P(8 < X < 9) =$ `pnorm(9.5,7.3,1.8)-pnorm(8.2,7.3,1.8)` $\approx 0.1977257$.

   (c) What is the probability that a randomly selected individual's weight is between 5.5 and 9.1 kg?

   *Solution:*

   $7.3 - 1.8 = 5.5$ and $7.3 + 1.8 = 9.1$ so this is plus and minus 1 standard deviation. The probability of that range is approximately 68% by the 68-95-99.7 rule.

   (d) What is the probability that a randomly selected individual's weight is above 10.6 kg?

   *Solution:*

   $7.3 + 2 \cdot 1.8 = 10.6$ so this is 2 standard deviations above the mean. The probability of that range is approximately $\frac{1}{2}(1 - 95\%) = 2.5\%$ by the 68-95-99.7 rule.

   Generally, when you have an exact number of standard deviations, it is ok to use this rule for approximation of probabilities.

   (e) If a sample of 12 individuals is taken, what is the probability the mean weight will be less than 5 kg?

   *Solution:*

   $P(\overline{X}_{12} < 5) =$ `pnorm(5,7.3,1.8/sqrt(12))` $\approx 4.792003(10)^{-6}$.

   (f) What is the probability that the mean weight is exactly 5 kg?

   *Solution:*

   $P(\overline{X} = 5) = 0$. It is a continuous random variable, so the probability on each individual value is zero.

   (g) If an iid sample of 25 individuals is selected, what is the probability that every individual's weight is in the range of 5 to 10 kg?

   *Solution:*

   $P(5 < X < 10) =$ `pnorm(10,7.3,1.8)-pnorm(5,7.3,1.8)` $\approx 0.83$ for each individual. "iid" means they are independent so we just multiply all the probabilities: $P(\text{all between 5 and 10}) \approx (0.83)^{25} \approx 0.01$.

   (h) What is the probability that $\frac{X_3 - 9}{1.8} < -1$?

   *Solution:*

   $\frac{X-\mu}{\sigma} = Z \sim N(0,1)$ thus $P(\frac{X_3-9}{1.8} < -1) \approx \frac{1}{2}(1 - 0.68)$ by the 68-95-99.7 rule.

3. Consider a manufacturer producing a thin laminate material for protective coatings of surfaces. The laminate material is produced continually with a width of 1 m and rolled onto spools that hold a total length of 200 m. Assume it is known that on average there are 3 imperfections per 40 m$^2$.

   (a) If a particular customer is to buy a complete spool of this material, what is the probability that the total number of imperfections is greater than 25?

   *Solution:*

The number of imperfections on the 200 m spool is Poisson with rate $\lambda = 200 \cdot 3/40 = 15$. $P(\text{more than 25 imperfections}) = P(X > 25) =$ `1-ppois(25,lambda=15)` $\approx 0.006184904$.

(b) If that customer needs at least four 1 m by 10 m sections with no imperfections (a total of 1 m by 40 m), what is the probability that the first 40 m on the spool have no imperfections?

*Solution:*

The number of imperfections in the first 40 m is Poisson with rate $\lambda = 40 \cdot 3/40 = 3$. $P(\text{no imperfections}) = P(X = 0) =$ `dpois(0,lambda=3)` $= e^{-3} \approx 0.04978707$.

(c) Imagine that the production machinery is turned on and begins churning out a sheet of laminate and that it will be stopped once the first imperfection occurs. What is the probability that the machine will produce at least 20 m of laminate before being stopped?

*Solution:*

Here $X$ is the length of material produced between imperfections. The length of material produced before the next imperfection is exponentially distributed with rate $\lambda = 3/40$, $X \sim Exp(3/40)$ so $P(X > 20) = \int_{20}^{\infty} \lambda e^{-\lambda x} dx = e^{-\lambda \cdot 20} = e^{-3/2} \approx 0.2231302$.

4. Consider the pdf $f(x) = a(16 - x^2)$ for $0 \le x \le 4$.

(a) Find $a$.

*Solution:*

$\int_0^4 a(16 - x^2)dx = a(16x - \frac{1}{3}x^3)\Big|_0^4 = a(64 - 64/3)$ thus $a = \frac{3}{128}$.

(b) Find the cdf.

*Solution:*

$F(x) = \int_0^x a(16 - x^2)dx = a(16x - \frac{1}{3}x^3)\Big|_0^x = a(16x - x^3/3)$. So plugging in what we got for $a$ and simplifying a bit gives $F(x) = \frac{1}{128}x(48 - x^2)$.

Note that $F(0) = 0$ and $F(4) = \frac{1}{128} \cdot 4 \cdot (48 - 4^2) = \frac{1}{32} \cdot 32 = 1$ as required for a cdf.

(c) Find $E(X)$.

*Solution:*

$E(X) = \int_0^4 ax(16 - x^2)dx = a(8x^2 - \frac{1}{4}x^4)\Big|_0^4 = a(128 - 4^3) = a \cdot 64 = \frac{3}{2}$

5. Consider the pdf $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$. Note that this is the exponential pdf.

(a) Find the median.

*Solution:*

The cdf is $F(x) = 1 - e^{-\lambda x}$. The median $\tilde{\mu}$ is such that $F(\tilde{\mu}) = 0.5$. So we set $F(x) = 0.5$ and solve for $x$, that will be the median.

$0.5 = 1 - e^{-\lambda x}$ gives $x = -\frac{\ln(0.5)}{\lambda} = \frac{\ln 2}{\lambda}$. Thus $\tilde{\mu} = \frac{\ln 2}{\lambda}$. Note that we have used a few logarithm properties here, notably $r \log M = \log M^r$. You should make sure you remember all properties of logs and exponentials.

(b) Find the first quartile.

*Solution:*

The first quartile $Q_1$ is the $25^{th}$ percentile so we set $F(x) = 0.25$ and solve for $x$.

$0.25 = 1 - e^{-\lambda x}$ gives $x = -\frac{\ln(0.75)}{\lambda} = \frac{\ln \frac{4}{3}}{\lambda} = \frac{\ln 4 - \ln 3}{\lambda}$.

*Some further info:*

Thus to find the $p$-percentile, we set $p = F(x)$ and solve for $x$:

$$\tilde{x}_p = -\frac{\ln(1-p)}{\lambda}$$

6. Consider the pmf

| $x$ | 0 | 10 | 20 |
|---|---|---|---|
| $f(x)$ | 0.10 | 0.60 | 0.30 |

(a) Calculate $E(X)$ and $Var(X)$.

*Solution:*

$E(X) = \sum x \cdot f(x) = 0(0.1) + 10(0.6) + 20(0.3) = 0 + 6 + 6 = 12.$

We know that: $Var(X) = E(X^2) - E(X)^2$. SO we calculate $E(X^2)$.

$E(X^2) = \sum x \cdot f(x) = 0^2(0.1) + 10^2(0.6) + 20^2(0.3) = 0 + 60 + 120 = 180.$

Thus $Var(X) = 180 - 12^2 = 36.$

(b) Calculate $E(100 + 2X)$ and $Var(100 + 2X)$.

*Solution:*

$E(aX + b) = aE(X) + b$ and $Var(aX + b) = a^2 Var(X)$.

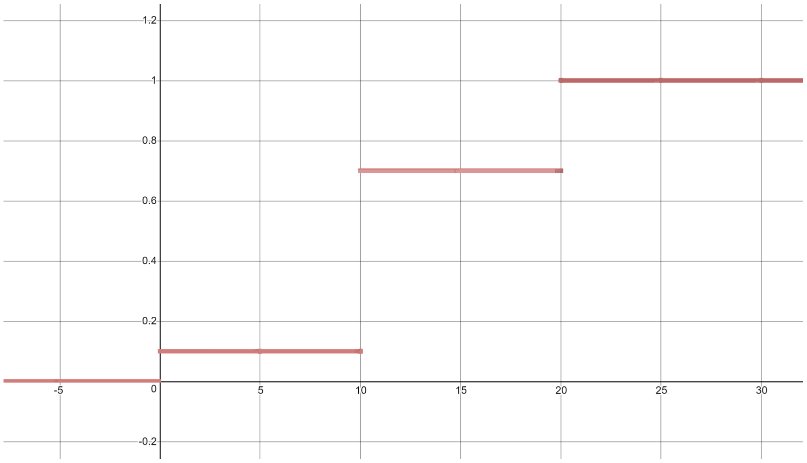
So $E(100 + 2X) = 100 + 2 \cdot 180 = 460$ and $Var(100 + 2X) = 4 \cdot 36 = 128$.

(c) Write the formula for and sketch a graph of the cdf.

*Solution:*

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.1 & \text{for } 0 \leq x < 10 \\ 0.7 & \text{for } 10 \leq x < 20 \\ 1 & \text{for } 20 \leq x \end{cases}$$

Here is the graph:



*Some further info:*

The cdf always starts of at 0 as $x$ approaches $-\infty$ and goes up to 1 as $x$ goes to $\infty$. For a discrete random variable, the cdf will be a step function where the steps occur at the possible values of $x$ and the size of the jump is the probability for that $x$-value.

(d) If two independent $X$ values are sampled (a sample of size 2) are selected, with replacement, what is the probability that the sample mean is 0? $P(\overline{X}_2 = 0)$.

*Solution:*

The only way the sample mean can be 0 is if both $X_1$ and $X_2$ are zero, and they are independent, so this occurs with probability $(0.1)^2 = \frac{1}{100}$.

(e) For the sample of size 2, what is the probability the sample variance is zero?

*Solution:*

The variance will only be zero when both $X$'s are the same.

To understand this, recall that the sample variance is calculated as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

For $s^2 = 0$, we need $(x_i - \overline{x})^2 = 0$ for all $i$, which means that all $x_i = \overline{x}$. Which mean that all sampled data values are the same and are equal to the sample mean.

Again, they are independent, thus $P(\text{zero sample variance}) = 0.10^2 + 0.60^2 + 0.30^2$.

(f) What is the probability that the sample mean for a sample of size two that is 10 or less?

*Solution:*

Let's generate all possible samples of size 2 that will give a sample mean less than or equal to 10. These samples written as $(X_1, X_2)$ are:

$$(0,0), (0,10), (10,0), (10,10), (0,20), (20,0).$$

Note that there are only 3 possible ordered pairs excluded. We just calculate the probability of each of these and add them up.

$P(\overline{X}_2 \le 10) = (0.1)^2 + 2(0.1)(0.6) + (0.6)^2 + 2(0.1)(0.3)$.

7. Consider the uniform random variable $X \sim U(0, 100)$.

(a) Write the pdf.

*Solution:*

$f(x) = \frac{1}{100}$ for $0 \le x \le 100$.

(b) Write the cdf.

*Solution:*

$F(x) = \int_0^x \frac{1}{100} du = \frac{x}{100}$

(c) Find $E(X)$.

*Solution:*

$E(x) = \int_0^{100} x \cdot \frac{1}{100} du = \frac{x^2}{200}|_0^{100} = \frac{100^2}{200} = 50.$

*Some further info:*

Note that for a uniform random variable, the expected value is always in the middle. $X \sim U(a, b)$ then $E(X) = \frac{a+b}{2}$.

(d) Find $P(50 < X < 70)$.

*Solution:*

This one we can calculate geometrically since the pdf is a constant function. This probability is the area of a rectangle with width $70 - 50 = 20$ and height $\frac{1}{100}$ thus the probability is $P(50 < X < 70) = 20 \cdot \frac{1}{100} = 20\%$.

(e) Find $b$ so that $P(30 < X < b) = 0.6$.

*Solution:*

$0.6 = P(30 < X < b) = (b - 30) \cdot \frac{1}{100}$ thus $60 = b - 30$ so $b = 90$.

(f) Find the first quartile.

*Solution:*

$0.25 = F(x) = \frac{x}{100}$ thus $x = 25$. So we can see that $Q_1 = 25$. In general for this pdf, the $p(100)^{th}$ percentile will be $100p$.

*Some further info:*

In general for a uniform distribution $X \sim U(a,b)$ the cdf is $F(x) = \frac{x-a}{b-a}$ and so the $p(100)^{th}$ percentile will be given by $p = \frac{\tilde{x}_p - a}{b-a}$ thus $\tilde{x}_p = a + p(b-a)$.

Example: for $X \sim U(1,5)$ then the $20^{th}$ percentile is $\tilde{x}_{0.2} = 1 + 0.2(5-1) = 1.8$.

8. Explain the following R code.

```
> x=c(0,2,4,6,8,10)
  p=c(0.69,0.01,0.05,0.12,0.03,0.10)
  m=sum(x*p)
  v=sum(x^2*p)-m^2
```

*Solution:*

This code calculates the expected value and variance for random variable $X$ with given pmf:
`x` is a list of possible values for discrete random variable $X$,
`p` give the probabilities for each value,
`m` is the expected value of $X$, and
`v` is the variance of $X$.

9. Explain the following R code.

```
> x=rnorm(100,mean=25,sd=5)
  hist(x,breaks=seq(from=0,to=50,by=2))
```

*Solution:*

This code generates 100 random samples from a normal distribution with mean 25 and standard deviation 5, and plots a histogram.

10. Consider the experiment where we are tracking the time between arrivals of data packets at a server. It is known that the mean time between data packet arrivals is 479 $\mu$s (microseconds). A data scientist records the arrival times for data packets, and has collected a list of times for 1 million packets.

(a) Calculate the approximate probability that the average time between packets is greater than 479.5 $\mu$s.

*Solution:*

Let $X_1$ be the wait time until the first packet, and $X_2$ be the wait time between the first and second packets, ... and generally, let $X_i$ be the time elapsed between packets $i-1$ and $i$. So we have a sample of size $n = 1$ million: $\{X_1, X_2, X_3, \ldots, X_{999{,}999}, X_{1{,}000{,}000}\}$.

The wait time between packets is modeled by an exponential random variable with rate $\lambda = \frac{1}{479}$ with time measured in microseconds.

$$X_i \sim Exp(\lambda = \frac{1}{479}) \quad \text{thus} \quad E(X_i) = \frac{1}{\lambda} = 479 \quad \text{and} \quad Var(X_i) = \frac{1}{\lambda^2} = 479^2$$

By the central limit theorem, the average time between packets is approximately normally distributed with mean $E(X)$ and variance $\frac{Var(X)}{n}$.

$$\overline{X}_n \sim N(\mu = 479, \sigma = \frac{479}{\sqrt{n}})$$

so

$$\overline{X}_{1000000} \sim N(\mu = 479, \sigma = 0.479)$$

Thus the probability that the mean time between packet arrivals is greater than 479.5 $\mu$s denoted by $P(\overline{X}_{1,000,000} > 479.5)$ is given as

$$\boxed{\texttt{1-pnorm(479.5,mean=479,sd=479/sqrt(10^6))} \approx 0.1482794}$$

(b) Calculate the approximate probability that the total time taken to collect this dataset is less than 8 minutes.

*Solution:*

By the central limit theorem, the total time for $n$ packets is approximately normally distributed with mean $n \cdot E(X)$ and variance $n \cdot Var(X)$.

$$\text{total time } = S_n \sim N(\mu = 479 \cdot n, \sigma = 479 \cdot \sqrt{n})$$

so

$$\overline{X}_{1000000} \sim N(\mu = 479(10)^6, \sigma = 479000)$$

Thus the probability that the mean time between packet arrivals is less than 8 minutes = 480 seconds = $480(10)^6$ $\mu$s is given as

$$\boxed{\texttt{pnorm(480*10^6,mean=479*10^6,sd=479*sqrt(10^6))} \approx 0.9815868}$$

*Additional info:*

Since $n > 30$ is the general rule for the CLT to apply when data is not normally distributed, $n = 1$ million should make this CLT approximation really accurate.

(c) If the scientist continues to collect data on the arrival times of packets, what will happen to the average between packet time in the sample?

*Solution:*

The set of wait times between packets is $\{X_1, X_2, \ldots, X_n\}$. The sample mean is $\overline{X}_n$. As $n$ gets really large (the size of the sample grows), the sample mean $\overline{X}_n$ is increasingly unlikely to vary to far from the mean time of 479 $\mu$s. This is due to the **law of large numbers**.

(d) For $n = 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^9$ calculate the probability that the sample mean is less than 478.95 $\mu$s.

*Solution:*

Thus the probability that the mean time between packet arrivals is less than 478.95 $\mu$s denoted by $P(\overline{X}_{1,000,000} < 478.95)$ is given as

`pnorm(478.95,mean=479,sd=479/sqrt(10^6))` $\approx 0.4584323$

`pnorm(478.95,mean=479,sd=479/sqrt(10^7))` $\approx 0.3706654$

`pnorm(478.95,mean=479,sd=479/sqrt(10^8))` $\approx 0.1482794$

`pnorm(478.95,mean=479,sd=479/sqrt(10^9))` $\approx 0.0004818484$

`pnorm(478.95,mean=479,sd=479/sqrt(10^10))` $\approx 8.277728(10)^{-26}$ (VERY tiny, basically zero) `pnorm(478.95,mean=479,sd=479/sqrt(10^11))` $\approx 3.000654(10)^{-239}$ (this is VERY near “machine zero”!)

`pnorm(478.95,mean=479,sd=479/sqrt(10^12))` $\approx 0$ (it’s not exactly zero, but we are limited by the computer’s ability to calculate)