MATH 3339
Statistics for the Sciences
Live Lecture Help

James West
jdwest@uh.edu

University of Houston

Session 7

Office Hours: see schedule in the "Office Hours" channel on Teams
Course webpage: www.casa.uh.edu

Email policy

When you email me you **MUST** include the following

- MATH 3339 Section 20024 and a description of your issue in the **Subject Line**
- Your name and ID# in the **Body**
- Complete sentences, punctuation, and paragraph breaks
- Email messages to the class will be sent to your Exchange account (user@cougarnet.uh.edu)

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/

Outline

1. Updates and Announcements
2. Recap
3. Student submitted questions

Updates and Announcements

- Test 1 begins tomorrow!
- Test 1 Structure:
    - 7 MC problems worth 7 pts each
    - 3 FR problems worth 17 pts each

PMF, Mean, and Variance for Bernoulli Random Variable

If X has the Bernoulli distribution with probability of success p, the pmf for X is:

$$f_X(x) = P(X = x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{if } x \neq 0, 1 \end{cases}$$

The **mean** and **variance** of X are:

$$\begin{aligned} \mu_X &= E[X] = p \\ \sigma_X^2 &= Var[X] = p(1-p) \end{aligned}$$

PMF, Mean, and Variance for Binomial Random Variable

If $X \sim Binom$[redacted](n, p), the pmf is:

$$f_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0, 1, 2, ..., n \\ 0 & \text{if } x \neq 0, 1, 2, ..., n \end{cases}$$

The **mean** and **variance** of X are:

$$\begin{aligned} \mu_X &= E[X] = np \\ \sigma_X^2 &= Var[X] = np(1-p) \end{aligned}$$

CDF of a Binomial R.V.

If $X \sim Binomi$[redacted](n, p)

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \displaystyle\sum_{k=0}^{x^*} \binom{n}{k} p^k (1-p)^{n-k}, & \text{if } 0 \leq x \leq n \\ 1, & \text{if } n \leq x \end{cases}$$

where $x^* = \lfloor x \rfloor$, the first integer less than or equal to x.

Cumulative Distribution Function

Recall that a quantitative random variable X has a **cumulative distribution function** given by

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

for all $x \in \mathbb{R}$.

When we have a discrete random variable X, the cdf is related to the pmf in the following way:

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i)$$

where $x_1, x_2, \ldots$ are the values of X.

Cumulative Distribution Function Properties

Any cdf F has the following properties:

1. F is a non-decreasing function defined on $\mathbb{R}$
2. F is right-continuous, meaning for each a, $F(a) = F(a+) = \lim_{x \to a^+} F(x)$
3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
4. $P(a < X \leq b) = F(b) - F(a)$ for all real a and b, where $a < b$.
5. $P(X > a) = 1 - F(a)$.
6. $P(X < b) = F(b-) = \lim_{x \to b^-} F(x)$.
7. $P(a < X < b) = F(b-) - F(a)$.
8. $P(X = b) = F(b) - F(b-)$.

Hypergeometric Distribution

The **parameters** of a hypergeometric distribution are m, n, k. We write $X \sim Hyper(m, n, k)$. The probability mass function for a hypergeometric is:

$$f_X(x) = P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

Mean and Variance of a Hypergeometric Distribution

Let Y have a hypergeometric distribution with parameter, m, n, and k.

- The mean of Y is:

$$\mu_Y = E(Y) = k\left(\frac{m}{m+n}\right) = kp.$$

- The variance of Y is:

$$\sigma_Y^2 = var(Y) = kp(1-p)\left(1 - \frac{k-1}{m+n-1}\right).$$

- $1 - \frac{k-1}{m+n-1}$ is called the **finite population correction factor**. As, the population increases, this factor will get closer to 1.

The Probability Function of the Poisson Distribution

A random variable X with nonnegative integer values has a Poisson distribution if its frequency function is:

$$f_X(x) = P(X = x) = e^{-\mu}\frac{\mu^x}{x!}$$

for $x = 0, 1, 2, \ldots$, where $\mu > 0$ is a constant. If X has a Poisson distribution with parameter μ, we can write $X \sim Pois(\mu)$.

The Mean and Variance of the Poisson Distribution

Let $X \sim Pois(\mu)$

- The mean of X is μ per unit of measure. By the conditions of the Poisson distribution.
- The variance of X is also μ per unit of measure.
- The standard deviation of X is $\sqrt{\mu}$.

Jointly Distributed Variables

The probabilities in the middle of the table are called the joint probabilities. The joint probability mass function is given by $f(x, y) = P(X = x, Y = y)$.

Properties of the joint probability mass function:

1. $0 \leq f(x, y) \leq 1$
2. $\sum\sum_{(x,y)\in\mathbb{R}^2} f(x, y) = 1$
3. $P((X, Y) \in A) = \sum\sum_{(x,y)\in A} f(x, y)$

Theorem 4.7

Let X and Y have the joint frequency function $f(x, y)$. Then

1. $f_Y(y) = \sum_x f(x, y)$ for all y.
2. $f_X(x) = \sum_y f(x, y)$ for all x.
3. $f_{Y|X}(y|x) = \dfrac{f(x, y)}{f_X(x)}$ if $f_X(x) > 0$. This is the **conditional** frequency function.
4. X and Y are independent if and only if $f(x, y) = f_X(x) f_Y(y)$ for all x, y.
5. $\sum_X \sum_Y f(x, y) = 1$.

Jointly Distributed Variables

Conditional Probabilities:

$$f_{Y|X}(y|x) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f(x,y)}{f_X(x)}$$

$$f_{X|Y}(x|y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x,y)}{f_Y(y)}$$

Expected Values of Jointly Distributed Variables

For joint random variables X and Y, the expected values are:

$$\mu_X = E[X] = \sum_{(x,y)\in\mathbb{R}^2} x \cdot f(x,y) = \sum_x \sum_y x \cdot f(x,y) = \sum_x x \cdot f_X(x)$$

$$\mu_Y = E[Y] = \sum_{(x,y)\in\mathbb{R}^2} y \cdot f(x,y) = \sum_y \sum_x y \cdot f(x,y) = \sum_y y \cdot f_Y(y)$$

$$E[g(X,Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x,y) \cdot f(x,y)$$

Means of Sums and Differences

- If X and Y are two different random variables, then the mean of the sums of the pairs of the random variable is the same as the sum of their means:

  $$E[X+Y] = E[X] + E[Y].$$

  This is called the addition rule for means.

- The mean of the difference of the pairs of the random variable is the same as the difference of their means:

  $$E[X-Y] = E[X] - E[Y].$$

Variances of Sums and Differences

If X and Y are independent random variables

$$\sigma^2_{X+Y} = Var[X + Y] = Var[X] + Var[Y].$$

and

$$\sigma^2_{X-Y} = Var[X - Y] = Var[X] + Var[Y].$$

Jointly Distributed Variables

Given two random variables X and Y , the covariance of X and Y is given by:

$$cov(X,Y) = E[(x-\mu_x)(Y-\mu_y)]$$

There is an easier version of this calculation:

$$cov(X,Y) = E[XY] - E[X]E[Y]$$

and the correlation coefficient of X and Y is given by:

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Properties of Covariance

1. $cov(X, Y) = cov(Y, X)$

2. $cov(X, X) = var(X)$

3. If X, Y, and Z are jointly distributed and a and b are constants

$$cov(X, aY + bZ) = a[cov(X, Y)] + b[cov(X, Z)].$$

4. If X and Y are jointly distributed,

$$var(X + Y) = var(X) + var(Y) + 2cor(X, Y)sd(X)sd(Y)$$
$$var(X - Y) = var(X) + var(Y) - 2cor(X, Y)sd(X)sd(Y)$$

5. If X and Y are independent, $cov(X, Y) = 0$.

6. If jointly distributed random variables $X_1, X_2, \cdots, X_n$ are pairwise uncorrelated, then

$$var(X_1 + X_2 + \cdots + X_n) = var(X_1) + var(X_2) + \cdots + var(X_n)$$

Definition of a Density Function

- A **density function** is a nonnegative function f defined of the set of real numbers such that:
$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

- If f is a density function, then its integral $F(x) = \displaystyle\int_{-\infty}^{x} f(u)du$ is a continuous **cumulative distribution function** (cdf), that is $P(X \leq x) = F(x)$.

- If X is a random variable with this density function, then for any two real numbers, a and b
$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx.$$

Cumulative Distribution Function Properties

Any cdf F has the following properties:

1. F is a non-decreasing function defined on $\mathbb{R}$
2. F is right-continuous, meaning for each a, $F(a) = F(a+) = \lim_{x \to a^+} F(x)$
3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
4. $P(a < X \leq b) = F(b) - F(a)$ for all real a and b, where $a < b$.
5. $P(X > a) = 1 - F(a)$.
6. $P(X < b) = F(b-) = \lim_{x \to b^-} F(x)$. $F(b-) = F(b)$
7. $P(a < X < b) = F(b-) - F(a)$.
8. $P(X = b) = F(b) - F(b-)$.

Quantiles

Let F be a given cumulative distribution and let p be any real number between 0 and 1. The **(100p)th percentile** of the distribution of a continuous random variable X is defined as

$$F^{-1}(p) = min\{x|F(x) \geq p\}.$$

For continuous distributions, $F^{-1}(p)$ is the smallest number x such that $F(x) = p$.

Expected Values for Continuous Random Variables

The **expected** or **mean value** of a continuous random variable X with pdf $f(x)$ is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

More generally, if h is a function defined on the range of X,

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx.$$

The Uniform Distribution

Let $X \sim \text{Unif}(a,b)$

- The pdf of X is:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

- The cdf of X is:

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b < x \end{cases}$$

- $\mu = E(X) = \dfrac{a+b}{2}$
- $\sigma^2 = \text{Var}(X) = \dfrac{(b-a)^2}{12}$

Henry and Jim are waiting for a raft. The number of rafts floating by over intervals of time is a Poisson process with a rate of λ=0.6 rafts per day. They agree in advance to let the first raft go and take the second one that comes along. What is the probability that they will have to wait more than a week? Hint: If they have to wait more than a week, the number of rafts in a period of 7 days would be λ*7.

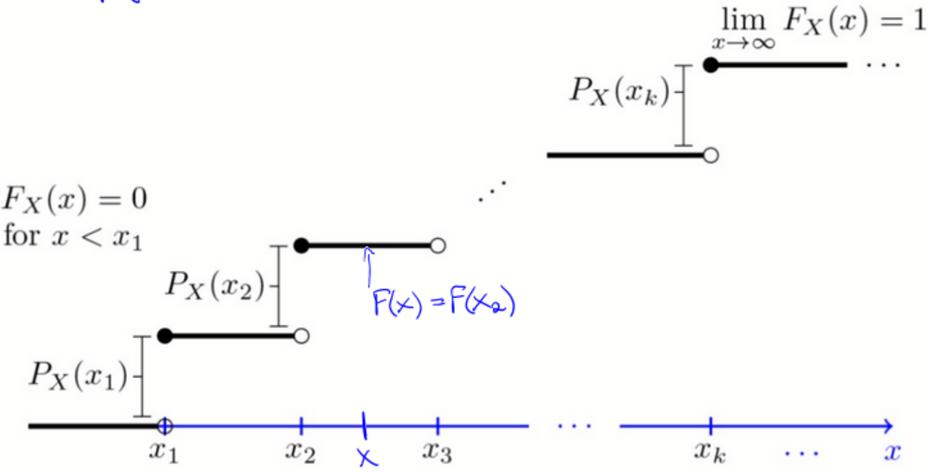
For a 7 day period $(\Delta t = 7)$, $\mu = \lambda \cdot \Delta t = (0.6)(7) = 4.2$

Let X be number of rafts in 7 days

want $P(X < 2)$, $X \sim \text{Pois}(4.2)$

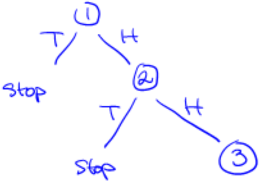
Graph of a CDF

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



A fair coin is tossed until either a tail occurs or 5 heads in a row have occurred. Let X denote the number of tosses. Find the frequency function, mean and variance of X.

What are the possible "words"?



Words

T

H T

H H T

H H H T

H H H H T

H H H H H

A biased coin has probability 0.8 of turning up heads. You win \$x if a head comes up and you lose \$y if a tail comes up. If your expected winnings is \$0, what is the relationship between x and y?

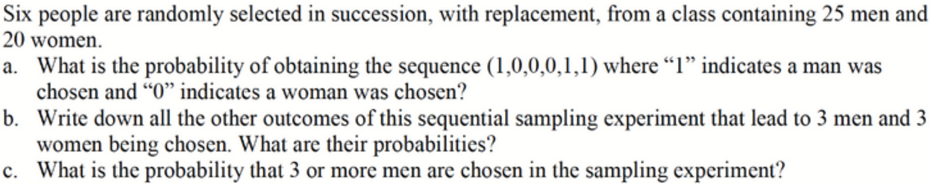
Let W denote your winnings

What are observable values for W?

There are only 2 possibilities

$W = x$ or $W = -y$

You define $P(W=w) = f_W(w)$



Six people are randomly selected in succession, with replacement, from a class containing 25 men and 20 women.

a. What is the probability of obtaining the sequence (1,0,0,0,1,1) where “1” indicates a man was chosen and “0” indicates a woman was chosen?

b. Write down all the other outcomes of this sequential sampling experiment that lead to 3 men and 3 women being chosen. What are their probabilities?

c. What is the probability that 3 or more men are chosen in the sampling experiment?

6. If you are using Rstudio click on the ”Packages” tab, then the checkbox next to the library MASS. Click on the word MASS and then the data set ”mammals” and read about it. If you are using R alone, in the Console window at the prompt > type

```
> data(mammals,package=”MASS”).
```

View the data with

```
> mammals
```

Make a scatterplot with the following commands and comment on the result.

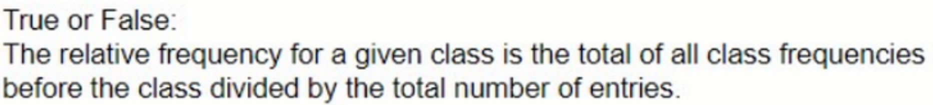
```
> attach(mammals)
```

```
> plot(body,brain)
```

Also make a scatterplot of the log transformed body and brain weights.

```
> plot(log(body),log(brain))
```

A recently discovered hominid species *homo floresiensis* had an estimated average body weight of 25 kg. Based on the scatterplots, what would you guess its brain weight to be?



True or False:

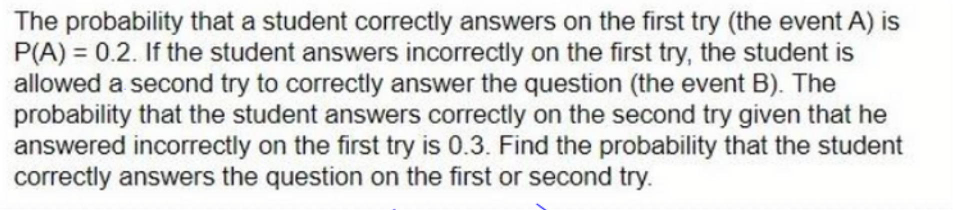
The relative frequency for a given class is the total of all class frequencies before the class divided by the total number of entries.

$[0,1), [1,2), [2,3), [3,4), [4,5)$

Of all customers purchasing automatic garage door openers, 75% purchase chain-driven model. Let X = the number among the next 15 purchasers who select the chain-driven model.

a. What is the frequency function (pmf) of X?

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



The probability that a student correctly answers on the first try (the event A) is P(A) = 0.2. If the student answers incorrectly on the first try, the student is allowed a second try to correctly answer the question (the event B). The probability that the student answers correctly on the second try given that he answered incorrectly on the first try is 0.3. Find the probability that the student correctly answers the question on the first or second try.

C is $A \cup (A^c \cap B)$

$$P(C) = P(A) + P(A^c \cap B)$$

What are we given? $P(A)=0.2$, $P(B|A^c)=0.3$

What $P(E \cap F) = P(E|F) \cdot P(F)$

or $P(E \cap F) = P(F|E) \cdot P(E)$

Use R’s “pbinom()” function to verify Chebyshev’s inequality for k=2 and k=3 when X follows *Binomial*(30, 0.2).

$$P\left(|X-\mu| > k\sigma\right) \leq \frac{1}{k^2}$$

$$X \sim \text{Binom}(30, 0.2)$$

$$\mu = 6, \quad \sigma = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{6 \cdot \tfrac{4}{5}} = \sqrt{\tfrac{24}{5}} \approx 2.19$$

$$|X-\mu| > k\sigma$$

$$\Rightarrow \quad X-\mu > k\sigma \quad \text{or} \quad -(X-\mu) > k\sigma$$

$$X > \mu + k\sigma \qquad\qquad X-\mu < -k\sigma$$

$$X < \mu - k\sigma$$

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/