More on discrete random variables

CE 311S

CUMULATIVE DISTRIBUTION FUNCTION

The *cumulative distribution function* $F_X(x)$ of a random variable is the probability that X is less than or equal to x,

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

Remember that X is a labeling of outcomes; so, for example, $F_X(5)$ is the probability that the outcome which actually occurs is no more than 5.

Example

For a family with two children, the PMF for the number of girls was given by

x	$P_X(x)$
0	1/4
1	1/2
2	1/4

Example

To find the CDF, "add up" the values in the PMF column:

x	$P_X(x)$	$F_X(x)$
0	1/4	1/4
1	1/2	3/4
2	1/4	1

Example

I flip a coin three times. Let Y equal 1 if at least two of these flips are heads, and 0 otherwise.

y	$P_Y(y)$	$F_Y(y)$
0	$1/2$	$1/2$
1	$1/2$	1

Notice that the values in the CDF column are never decreasing, and that for the greatest value the random variable the CDF equals one.

Example

I flip a coin until it comes up heads. Let Z equal the number of flips before I stop (including the last one).

z	$P_Z(z)$	$F_Z(z)$
1	$1/2$	$1/2$
2	$1/4$	$3/4$
3	$1/8$	$7/8$
4	$1/16$	$15/16$
5	$1/32$	$31/32$
$\vdots$	$\vdots$	

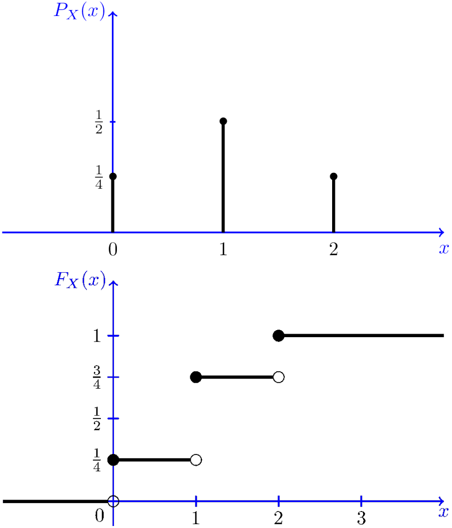
The values in the CDF column are still increasing; in this case there are an infinite number of values, so it never equals one; but as $z \to \infty$, $F_Z(z) \to 1$.

We can define the CDF for values in between those the random variable can take.

x	$P_X(x)$	$F_X(x)$
0	1/4	1/4
1	1/2	3/4
2	1/4	1

- $F_X(1.5) = P(X \leq 1.5) = 3/4$
- $F_X(0.5) = P(X \leq 0.5) = 1/4$
- $F_X(-5) = P(X \leq -5) = 0$
- $F_X(10) = P(X \leq 10) = 1$

The PMF can be plotted by showing the probability of each possible value for the random variable. The CDF can be plotted as horizontal lines which "jump" at each possible value for the random variable.



(Figures from the Pishro-Nik text.)

If we have the CDF, we can compute probabilities without having to sum up each of the possible outcomes.

In the infinite coin-flipping example, $F_Z(z) = 1 - 1/2^z$ whenever z is a positive integer. What is the probability that I flip the coin at least 5 times, but no more than 15 times?

"At least 5, but no more than 15" means $5, 6, 7, \ldots, 13, 14, 15$. (Pay attention to whether endpoints are included or not: "more than five" vs. "at least 5," "no more than 15" vs. "less than 15.")

I can write this statement in multiple ways:
$P(5 \leq Z \leq 15) = P(4 < Z \leq 15) = P(4 < Z < 16)$, etc.

If I write it in the form $P(4 < Z \leq 15)$, then this is just $F_Z(15) - F_Z(4)$. Why?

$F_Z(15)$ gives me the probability of seeing $1, 2, \ldots, 14, 15$.

$F_Z(4)$ gives the probability of seeing 1, 2, 3, or 4.

Subtracting these, what is left is the probability of $5, 6, 7, \ldots, 14, 15$ which is the goal.

$$P(5 \leq Z \leq 15) = F_Z(15) - F_Z(4) = (1 - 1/2^{15}) - (1 - 1/2^4) \approx 0.0625.$$

EXPECTED VALUE

In the first week of class, we developed descriptive statistics for data sets. (Why?)

```
000000000003 000000000066 01 0000 040508 2400 040508 0060 01 C 100    30   100
0000 0000 0000 0000 0000 0000 0000 0000 0000 333333777770000
'                '                            'US 290 4.5 miles W of FM 1960
01 12        0001 0002 0003 0004 0005 0006 0007 0008 0009 0010 0011 0012
00 00
00 00
00 00 3 0100 0054 0047 0039 0170 0192 0063 0083 0216 0227 0057 0051 0018
00 00 3 0200 0020 0015 0012 0108 0124 0038 0046 0150 0141 0039 0025 0009
00 00 3 0300 0011 0015 0008 0068 0100 0026 0038 0139 0134 0029 0030 0005
00 00 3 0400 0018 0008 0007 0079 0104 0015 0037 0116 0096 0030 0026 0005
00 00 3 0500 0009 0014 0013 0112 0157 0039 0035 0129 0101 0018 0027 0005
00 00 3 0600 0023 0022 0042 0214 0296 0139 0103 0242 0129 0073 0034 0013
00 00 3 0700 0062 0043 0085 0275 0384 0172 0305 0562 0380 0148 0078 0022
00 00 3 0800 0127 0093 0161 0398 0497 0262 0546 0768 0519 0270 0132 0085
00 00 3 0900 0178 0126 0284 0528 0640 0413 0653 0859 0645 0366 0190 0134
00 00 3 1000 0231 0170 0371 0663 0809 0534 0926 1009 0788 0526 0260 0212
00 00 3 1100 0288 0186 0396 0772 0896 0625 1086 1151 0935 0610 0322 0268
00 00 3 1200 0367 0237 0513 0845 1039 0731 1054 1160 1003 0657 0424 0262
00 00 3 1300 0344 0258 0460 0846 1086 0903 1085 1214 1095 0745 0460 0317
00 00 3 1400 0397 0351 0463 0956 1175 0993 1113 1217 1080 0713 0436 0317
00 00 3 1500 0407 0316 0556 0950 1208 1063 1144 1232 1116 0689 0461 0309
00 00 3 1600 0433 0318 0490 0971 1294 1089 1136 1203 1083 0665 0465 0298
00 00 3 1700 0440 0323 0502 1073 1304 1194 0876 1097 0996 0695 0455 0288
00 00 3 1800 0418 0314 0488 1043 1354 1230 0846 1090 0986 0631 0407 0290
00 00 3 1900 0399 0319 0441 1030 1286 1105 0707 0939 0896 0550 0390 0287
00 00 3 2000 0381 0258 0403 0933 1154 1006 0516 0777 0741 0460 0332 0245
00 00 3 2100 0337 0243 0214 0813 0976 0789 0360 0586 0632 0319 0266 0134
00 00 3 2200 0286 0193 0178 0669 0885 0607 0336 0560 0544 0247 0210 0132
00 00 3 2300 0153 0126 0137 0467 0547 0307 0277 0475 0424 0212 0152 0075
00 00 3 2400 0093 0081 0081 0387 0455 0214 0148 0304 0300 0146 0120 0060
```

1845

We want to do the same for random variables.

Consider families with two children. Let X be the number of girls. X has a value for each of the four outcomes.

Outcome	X
BB	0
BG	1
GB	1
GG	2

What would a **measure of location** or **measure of variability** mean for the random variable X?

Imagine that we repeat this experiment many, many times, and record the value of X for each.

$$\begin{array}{ccccccccccc} BB & \rightarrow & BG & \rightarrow & GG & \rightarrow & BB & \rightarrow & GB & \rightarrow & \cdots \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & \rightarrow & 1 & \rightarrow & 2 & \rightarrow & 0 & \rightarrow & 1 & \rightarrow & \cdots \end{array}$$

We can calculate the sample mean (and other descriptive statistics) from these values of X.

In the long run, we expect to see $X = 0$ for 25% of the sample values, $X = 1$ for 50% of them, and $X = 2$ for 25% of them. So the sample mean becomes a **weighted** average of the possible values of X, with weights according to the probability mass function.

Outcome	X
BB	0
BG	1
GB	1
GG	2

Mean of X is $\dfrac{0.25 \times 0 + 0.50 \times 1 + 0.25 \times 2}{0.25 + 0.50 + 0.25} = 1$. This is also called the **expected value**.

Because the denominator is the sum of the probabilities (which will always equal 1), the expected value can be written as

$$E[X] = \mu_X = \sum_{x \in R} x \cdot P_X(x)$$

where R is the set of all of the possible values X can take.

For example, we could have calculated the expected number of girls in a two-child family as follows:

x	$P_X(x)$	$x \cdot P_X(x)$
0	0.25	0
1	0.50	0.50
2	0.25	0.50
	1	**1**

Example: Here are the current odds for the Lotto Texas game.

Winnings	Odds
\$3	1 in 75
\$50	1 in 1526
\$2000	1 in 89,678
\$21,000,000	1 in 25,827,165

What is the expected winnings from a single ticket? If each ticket costs \$1, is it a good idea to play this game?

Write each possible outcome, the probability of occurrence, multiply and add.

3	1/75	0.040
50	1/1526	0.033
2000	1/89678	0.022
21×10^6	1/25827165	0.813
0	0.986	0
	1	**0.908**

For a \$1 ticket, on average you will only get 90.8¢ back.

Furthermore, excluding the very rare possibility of a jackpot, you will get less than 10¢ back for your dollar.

We can also calculate expected values of functions as well.

Assume that the temperature in Austin is either 95°F (with 10% probability), 90°F (with 20% probability), 80°F (with 30% probability), and 70 F (with 40% probability). What is the expected temperature in °F and °C?

We can also calculate expected values of functions as well.

Assume that the temperature in Austin is either 95°F (with 10% probability), 90°F (with 20% probability), 80°F (with 30% probability), and 70 F (with 40% probability). What is the expected temperature in °F and °C? If C is the temperature in Celsius and F is the temperature in Fahrenheit, $C = 5/9(F - 32)$.

f	Probability	$f \cdot P_F(f)$	$c(f) \equiv 5/9(f-32)$	$c(f) \cdot P_C(c)$
95	0.1	9.50	35	3.50
90	0.2	18.0	32.2	6.44
80	0.3	24.0	26.7	8.00
70	0.4	28.0	21.1	8.44
	1	**79.5**		**26.4**

(In general $E[g(X)] = \sum_{x_k \in R_X} g(x_k) P_X(x_k)$)

f	Probability	$f \cdot P_F(f)$	$c(f) \equiv 5/9(f-32)$	$c(f) \cdot P_C(c)$
95	0.1	9.50	35	3.50
90	0.2	18.0	32.2	6.44
80	0.3	24.0	26.7	8.00
70	0.4	28.0	21.1	8.44
	1	**79.5**		**26.4**

In this example, we could have just taken the expected temperature in Fahrenheit and converted to Celsius. **This won't always work.**

When I go to Double Daves, half of the time I eat a 12-inch medium pizza, and the other half of the time I eat the 18-inch Big Dave pizza. What is the expected diameter of my pizza?

If the number of calories C is related to the diameter of the pizza D by $C = 19D^2$, what is the expected number of calories I consume each time I go to Double Dave's?

$E[D] = 15$ and $E[C] = 4446 \neq 19 \times 15^2 = 4275$.

The only time that $E[g(X)] = g(E[X])$ reliably is when g is a linear function of X. In particular,

$E[aX + b] = a \cdot E[X] + b$ for any values of a and b

Why?

For a nonlinear function, we have to *first* compute $g(x)$ for every possible value, then compute $\sum g(x)P(x)$. This is called the **law of the unconscious statistician** (LOTUS):

$$E[g(X)] = \sum_{x_k \in R_X} g(x_k) P_X(x_k)$$

(The same formula works for linear functions too; but you can use the distributive property to manipulate the sum to get $aE[X] + b = g(E[X])$)

Consider this carnival game: a fair coin is tossed repeatedly until tails appears. The pot starts at one dollar and is doubled each time heads appears. Whenever tails appears, you win the entire pot.

How much would you pay to play this game?

What are your expected winnings?

Expected value does not always exist!

There must be more to the story than expected value alone...

VARIANCE

What can we do about an equivalent for variance and standard deviation?

$$
\begin{array}{ccccccccccc}
BB & \rightarrow & BG & \rightarrow & GG & \rightarrow & BB & \rightarrow & GB & \rightarrow & \cdots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \rightarrow & 1 & \rightarrow & 2 & \rightarrow & 0 & \rightarrow & 1 & \rightarrow & \cdots
\end{array}
$$

Do the same as with the mean (expected value). If we were to run the experiment many times, what would be the sample variance and sample standard deviation?

Remember, the formula for sample variance was $s^2 = \dfrac{\sum(x_i - \overline{x})^2}{n-1}$. Why did we have $n-1$ in the denominator?

1. We only had access to a sample, not the true (population) distribution, $n-1$ corrected for this. Here we have the true distribution.
2. In any case, when n is large, the difference between dividing by n or $n-1$ is small.

With n in the denominator, s^2 was simply the average value of $(x - \overline{x})^2$.

So, for random variables, we define the variance and standard deviation as follows:

Variance: $V[X] = \sigma_X^2 = E[(X - \mu_X)^2] = \sum_{x_k \in R_X} (x_k - \mu_X)^2 P_X(x_k)$
Standard deviation: $\sigma_X = \sqrt{V[X]}$

Example

What is the variance and standard deviation of the number of boys in a two-child family (remember $\mu_X = 1$)?

x	$P_X(x)$	$(x-\mu_X)^2$	$(x-\mu_X)^2 P_X(x)$
0	0.25	1	0.25
1	0.50	0	0
2	0.25	1	0.25
	1		**0.5**

So $V[X] = 0.5$ and $\sigma_X = \sqrt{0.5} = 0.71$.

VARIANCE SHORTCUTS AND FORMULAE

We can compute $V[X]$ more simply using

$$V[X] = E\left[X^2\right] - (E\left[X\right])^2$$

Why?

Example

What is the variance and standard deviation of the number of boys in a two-child family?

x	$P_X(x)$	x^2	$x^2 P_X(x)$
0	0.25	0	0
1	0.50	1	0.50
2	0.25	4	1.00
	1		**1.5**

So $E[X]^2 = 1.5$ and $V[X] = E[X]^2 - (E[X])^2 = 1.5 - 1^2 = 0.5$

The variance of a linear function can be calculated as follows:

$V[aX + b] = a^2 \cdot V[X]$ for any values of a and b

Why?

EXPECTED VALUE AND VARIANCE FOR SPECIAL DISCRETE RANDOM VARIABLES

For any binomial random variable

$E[X] = np$ and $V[X] = np(1-p)$

So, if you know X is a binomial random variable, you don't have to calculate a complicated sum; just use these formulas.

Example

- I flip a coin 50 times; the expected number of heads is $50(1/2) = 25$ and the variance is $50(1/2)(1/2) = 12.5$
- The probability of winning is 0.4; if I play 10 times, I expect to win $10(0.4) = 4$ times and the variance is $10(0.4)(0.6) = 2.4$.
- A family has 4 children; the expected number of boys is $4(1/2) = 2$ and the variance is $4(1/2)(1/2) = 1$.

For any hypergeometric random variable X, the expected value and variance are:

$$E[X] = \frac{kb}{b+r}$$

$$V[X] = k\frac{b}{b+r}\frac{r}{b+r}\frac{b+r-k}{b+r-1}$$

(Is this similar to the formulas for the binomial distribution when the population is large?)

For any negative binomial random variable, we have

$$E[X] = \frac{m}{p}$$

$$V[X] = \frac{m(1-p)}{p^2}$$

For any Poisson random variable X with an average rate of occurrence λ we have

$$E[X] = \lambda$$
$$V[X] = \lambda$$

SOME EXAMPLES

Example

Pennies minted before 1982 are mostly made of copper (after 1982, they are almost entirely zinc). Copper prices have risen to the point that a pre-1982 penny is actually worth 2.5 cents if melted down. Some people actually spend their time sifting pennies to find pre-1982 ones (roughly 25% of pennies in circulation).

On average, how many pennies will I look at before I earn a \$15 profit from this activity?

I earn 1.5 cents profit on each pre-1982 penny I see.

To make \$15 profit, I need to find 1000 pre-1982 pennies.

Let X be the number of **post**-1982 pennies I see first. X is negative binomial with $m = 1000$, $p = 1/4$

So $E[X] = 1000/(1/4) = 4000$

Assume that the number of students who stop by my office hours is a Poisson random variable with $\lambda = 2$. What is the probability that no students stop by my office hours? The probability that between 3 and 5 students stop by? What is the standard deviation of the number of students who stop by?

$$P(X = 0) = \frac{e^{-2}2^0}{0!} = 0.135$$

$$P(X = 3) + P(X = 4) + P(X = 5) = \frac{e^{-2}2^3}{3!} + \frac{e^{-2}2^4}{4!} + \frac{e^{-2}5^0}{5!} = 0.307$$

$$\sqrt{V[X]} = \sqrt{2} = 1.414$$

The number of callers to a technical support line is a Poisson random variable. On average, there are 2 calls per hour. What is the standard deviation of the number of calls in an eight-hour shift? What is the probability that there will be at least 2 calls during an eight-hour shift?

For the eight-hour shift, the average number of occurrences is 16, so $\lambda = 16$.

$$\sqrt{V[X]} = \sqrt{16} = 4$$

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{e^{-16}16^0}{0!} + \frac{e^{-16}16^1}{1!} = 0.999998$$

For next time...

I enter a casino playing a game with even odds (50% chance of winning). I adopt the following strategy: start by betting \$1. If I win, stop playing. If I lose, double my bet to \$2 and play again. Repeat until I win, and walk away \$1 richer.

Does this work?

First try to think about the problem intuitively. Then define random variables, calculate expected values and variances, and either confirm or revise your intuition.