INTRODUCTION TO PROBABILITY MODELS

Lecture 35

Qi Wang, Department of Statistics

Nov 14, 2018

EXAMPLE 1

Heights of Pokémon are Normally Distributed with a mean of 59 inches and a standard deviation of 17 inches.

1. What is the standardized height (z-score) for Blastoise, who is 63 inches tall?
2. Using the Normal Table, find the probability that any Pokémon is taller than Blastoise
3. Knowing that a Pokémon is taller than Blastoise, what is the probability that it is taller than 70 inches?
4. What height corresponds to the top 10% of Pokémon heights?

NORMAL APPROXIMATION TO THE BINOMIAL

If a Binomial distribution has a large enough combination of n and p, it behaves much like a Normal distribution, which means we can use the Normal distribution to approximate the original Binomial distribution

- If $X \sim Bin(n, p)$, and $np > 5, n(1-p) > 5$
- Then we can use $X^\star \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$, to approximate X

You may notice that Binomial is Discrete, and Normal is Continuous. This means the approximation comes at a cost of accuracy that we must try to correct. When we use the approximation, we need to perform a continuity correction:

- If you're looking for: $P(a \leq X \leq b)$
- Use $P(a - 0.5 < X^\star < b + 0.5)$

TIME FOR QUIZ

CROSSTABS TABLE/CONTINGENCY TABLE

CROSSTABS TABLE/CONTINGENCY TABLE

- Describes the relationship between two categorical variables.
- Represents a table of counts (can include percentages).

Examples

- Gender versus major
- Political party versus voting status

Sometimes one or both variables are quantitative, but we classify them into categories for data collection and/or analysis. For example, suppose our variables are years of college education and income. We decide to group years of education into four classes: none, some college, Bachelor's degree, and post-graduate. We also decide to classify annual income in dollars into four classes: $< 10,000, 10,000 - 30,000, 30,001 - 50,000,$ and $> 50,000$.

EXAMPLE 2

An instructor taught four sections of a large statistics course and had the following distribution of grades when the semester was finished.

Grade	One	Two	Three	Four	To
A	12	18	10	12	
B	26	26	16	16	
C	28	20	24	18	
D	6	8	20	18	
F	4	4	8	12	
Total					

JOINT DISTRIBUTION

The **joint distribution** of the 2 categorical variables is the proportion of total cases in a cell

$$JointProbability = \frac{TotalInCell}{OverallTotal}$$

All the joint distributions should add to 1 (or 100%). For example: 18/306 = 0.0588 or 5.88% is the joint distribution for people with grade of A AND Class time One. Joint distributions use or imply "AND". (i.e intersection)

Fill in the table of Joint distributions:

Grade	One	Two	Three	Four	Tota
A		5.88%		3.92%	16.9
B	8.50%		5.23%	5.23%	27.4
C	9.15%	6.54%	7.84%		29.4
D		2.61%	6.54%	5.88%	16.9
F	1.31%	1.31%	2.61%		9.15
Total	24.84%	24.84%	25.49%	24.84%	100

MARGINAL DISTRIBUTION

The **marginal distribution** allows us to study 1 variable at a time. The marginal distributions of each categorical variable are obtained from row and column totals. Basically we are examining the distributions of a single variable in the two-way table. Marginal distributions allow us to compare the relative frequencies among the levels of a single categorical variable

1. The marginals for the row variable should add to 1 (or 100%).
2. The marginals for the column variable should add to 1 (or 100%)

Find the marginal distribution of Class Time for Example 2

	One	**Two**	**Three**	**Four**
Counts				
Percents				

Find the marginal distribution of Letter Grade for Example 2

	A	B	C	D	F
Counts					
Percents					

CONDITIONAL DISTRIBUTION

In **conditional distributions**, we find the distribution of one categorical variable given a common level of another categorical variable. Look for key words to indicate a conditional—“given”, “knowing”, etc.

Find the conditional distribution of Letter Grade for Class Time One

	A	B	C	D	F
Counts					
Percents					

Find the conditional distribution of Class Time for Letter Grade C.

	One	Two	Three	Four
Counts				
Percents				

ADDITIONAL QUESTIONS FOR EXAMPLE 2

1. What percent of students in Class time Four earned a B? Is this joint, conditional or marginal?
2. Of all students earning a B, what proportion were in Class time 4? Is this joint, conditional or marginal
3. What percent of students were enrolled in Class Time 3? Is this joint, conditional or marginal?
4. What proportion of students earned B's and were in Class time 2? Is this joint, conditional or marginal?

STEMPLOT

Stemplot or Stem-and-leaf plot is a technique that orders quantitative data points and provides insight about the shape of the distribution. To make a stem-and-leaf plot, the last digit of the number is the leaf and the rest of the number is the stem. Leaves are arrange in ascending order on the stem. Additionally, any stem that is not used, but is within the range of the data, is kept in the plot.

EXAMPLE 3

DATA SET is:
1, 3, 5, 7, 12, 15, 17, 19, 21, 21, 21, 30, 33, 39, and 56.
Create a stem-and-leaf plot of the data.