Unit 2: Probability and distributions

3. Normal and binomial distributions

Sta 104 - Summer 2018, Term 2

Duke University, Department of Statistical Science

Prof. White

Slides posted at *https://www2.stat.duke.edu/courses/Summer18/sta104.001-2/*

Announcements

- PA 2 is due today (11:55 pm)
- PS 2 is due today (11:55 pm)
- Lab 2 is due Thursday (12:45 pm)
- RA 3 is Thursday in class.

1. Two types of probability distributions: discrete and continuous

- A *discrete probability distribution* lists all possible events and the probabilities with which they occur
  - The events listed must be disjoint
  - Each probability must be between 0 and 1
  - The probabilities must total 1

  Example: Binomial distribution
- A *continuous probability distribution* differs from a discrete probability distribution in several ways:
  - The probability that a continuous random variable will equal to any specific value is zero.
  - As such, they cannot be expressed in tabular form.
  - Instead, we use an equation or a formula to describe its distribution via a probability density function (pdf).
  - We can calculate the probability for ranges of values the random variable takes (area under the curve).

  Example: Normal distribution

Clicker question

Speeds of cars on a highway are normally distributed with mean 65 miles / hour. The minimum speed recorded is 48 miles / hour and the maximum speed recorded is 83 miles / hour. Which of the following is most likely to be the standard deviation of the distribution?

(a) -5
(b) 5
(c) 10
(d) 15
(e) 30

3. Z scores serve as a ruler for any distribution

A Z score creates a common scale so you can assess data without worrying about the specific units in which it was measured.

How can we determine if it would be unusual for an adult woman in North Carolina to be 96" (8 ft) tall?

How can we determine if it would be unusual for an adult alien woman(?) to be 103 metreloots tall, assuming the distribution of heights of adult alien women is approximately normal?

3. Z scores serve as a ruler for any distribution

$$Z = \frac{obs - mean}{SD}$$

- Z score: number of standard deviations the observation falls above or below the mean
- Z distribution (also called the *standardiZed normal* distribution, is a special case of the normal distribution where $\mu = 0$ and $\sigma = 1$

$$Z \sim N(\mu = 0, \sigma = 1)$$

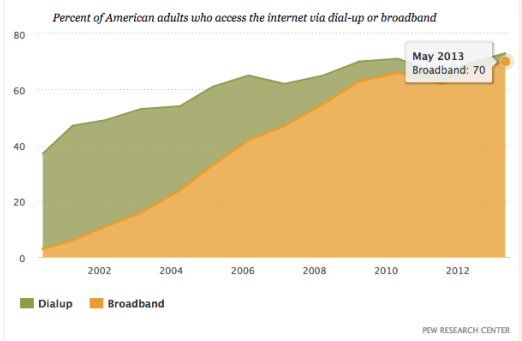
- Defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles
- Observations with $|Z| > 2$ are usually considered *unusual*

Clicker question

Scores on a standardized test are normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z scores, which of the following statements will be correct?

(a) The mean will equal 0, but the median cannot be determined.
(b) The mean of the standardized Z-scores will equal 100.
(c) The mean of the standardized Z-scores will equal 5.
(d) Both the mean and median score will equal 0.
(e) A score of 70 is considered unusually low on this test.

High-speed broadband connection at home in the US



- Each person in the poll be thought of as a *trial*
- A person is labeled a *success* if s/he has high-speed broadband connection at home, *failure* if not
- Since 70% have high-speed broadband connection at home, *probability of success* is *p = 0.70*

Considering many scenarios

Suppose we randomly select three individuals from the US, what is the probability that exactly 1 has high-speed broadband connection at home?

Let's call these people Anthony (A), Barry (B), Cam (C). Each one of the three scenarios below will satisfy the condition of "exactly 1 of them says Yes":

Scenario 1: $\underset{\text{(A) } yes}{0.70} \times \underset{\text{(B) no}}{0.30} \times \underset{\text{(C) no}}{0.30} \approx 0.063$

Scenario 2: $\underset{\text{(A) no}}{0.30} \times \underset{\text{(B) } yes}{0.70} \times \underset{\text{(C) no}}{0.30} \approx 0.063$

Scenario 3: $\underset{\text{(A) no}}{0.30} \times \underset{\text{(B) no}}{0.30} \times \underset{\text{(C) } yes}{0.70} \approx 0.063$

The probability of exactly one 1 of 3 people saying Yes is the sum of all of these probabilities.

$$0.063 + 0.063 + 0.063 = 3 \times 0.063 = 0.189$$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 3$ trials), and we calculated this probability as

$$\#\ of\ scenarios \times P(single\ scenario)$$

- $P(single\ scenario) = p^k\ (1-p)^{(n-k)}$

  probability of success to the power of number of successes, probability of failure to the power of number of failures
- number of scenarios: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

The *Binomial distribution* describes the probability of having exactly k successes in n independent trials with probability of success p.

Binomial distribution (cont.)

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k\ (1-p)^{(n-k)}$$

Note: You can also use R for the calculation of number of scenarios:

```
> choose(5,3)
```

```
[1] 10
```

Note: And to compute probabilities

```
> dbinom(1, size = 3, prob = 0.7)
```

```
[1] 0.189
```


Clicker question

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(a) the trials must be independent
(b) the number of trials, n, must be fixed
(c) each trial outcome must be classified as a *success* or a *failure*
(d) the number of desired successes, k, must be greater than the number of trials
(e) the probability of success, p, must be the same for each trial

Clicker question

According to the results of the Pew poll suggesting that 70% of Americans have high-speed broadband connection at home, is the probability of exactly 2 out of 15 randomly sampled Americans having such connection at home pretty high or pretty low?

(a) pretty high
(b) pretty low

Clicker question

According to the results of the Pew poll 70% of Americans have high-speed broadband connection at home, what is the probability that exactly 2 out of 15 randomly sampled Americans have such connection at home?

(a) $0.70^2 \times 0.30^{13}$
(b) $\binom{2}{15} \times 0.70^2 \times 0.30^{13}$
(c) $\binom{15}{2} \times 0.70^2 \times 0.30^{13}$
(d) $\binom{15}{2} \times 0.70^{13} \times 0.30^2$

Expected value and standard deviation of binomial

According to the results of the Pew poll suggestion that 70% of Americans have high-speed broadband connection at home, among a random sample of 100 Americans, how many would you expect to have such connection at home?

- $100 \times 0.70 = 70$
  - Or more formally, $\mu = np = 100 \times 0.7 = 7$
- But this doesn't mean in every random sample of 100 Americans exactly 70 will have high-speed broadband connection at home. In some samples there will be fewer of those, and in others more. How much would we expect this value to vary?
  - $\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.70 \times 0.30} \approx 4.58$

Note: Mean and standard deviation of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Shape of the binomial distribution

https://gallery.shinyapps.io/dist_calc/

You can use the normal distribution to approximate binomial probabilities when the sample size is large enough.

S-F rule: The sample size is considered large enough if the expected number of successes and failures are both at least 10

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

What is the probability that among a random sample of 1,000 Americans at least three-fourths have high-speed broadband connection at home?

$$\textit{Binom}(n = 1000, p = 0.7)$$

$$P(K \geq 750) = P(K = 750) + P(K = 751) + P(K = 752) + \cdots + P(K = 1000)$$

1. Using R:

```
> sum(dbinom(750:1000, size = 1000, prob = 0.7))
```

```
[1] 0.00026
```

2. Using the normal approximation to the binomial: Since we have at least expected successes $(1000 \times 0.7 = 700)$ and 10 expected failures $(1000 \times 0.3 = 300)$,

$$\textit{Binom}(n = 1000, p = 0.7) \sim$$
$$N(\mu = 1000 \times 0.7, \sigma = \sqrt{1000 \times 0.7 \times 0.3})$$

Summary of main ideas

1. Two types of probability distributions: discrete and continuous
2. Normal distribution is unimodal, symmetric, and follows the 68-95-99.7 rule
3. Z scores serve as a ruler for any distribution
4. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
5. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
6. Shape of the binomial distribution approaches normal when the S-F rule is met