Advanced Statistics
DS2003 (BDS-4A)
Lecture 02

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

17 February, 2022

Previous Lecture

- Random Variables
- Discrete Random Variables
  - How many heads in three coin tosses
  - Sum of the scores of two dice
- Continuous Random Variables
  - Uniform Distribution
- Cumulative Distribution Function (CDF)
- The Standard Normal Distribution
- IID
- Bernoulli Trials

IID

- In statistics, we usually say “random sample,” but in probability it’s more common to say “IID.”
- ***Identically Distributed*** means that there are no overall trends–the distribution doesn’t fluctuate and all items in the sample are taken from the same probability distribution.
- ***Independent*** means that the sample items are all independent events. In other words, they aren’t connected to each other in any way

Bernoulli Trials

- A single trial or experiment, for example, a *coin toss*
- Independent repeated trials of an experiment with exactly two possible outcomes are called Bernoulli trials.
- Call one of the outcomes *success* and the other outcome *failure*.
- Let *p* be the probability of *success* in a Bernoulli trial, and *q* be the probability of *failure*.
- Then the probability of *success* and the probability of *failure* sum to *one*, since these are complementary events:
  - "success" and "failure" are *mutually exclusive and exhaustive*.
- Thus one has the following relations:
  - $p + q = 1$

Permutations and Combinations

- "*My fruit salad is a combination of apples, grapes and bananas*" We don't care what order the fruits are in, they could also be "*bananas, grapes and apples*" or "*grapes, apples and bananas*", its the same fruit salad.
- "*The combination to the safe is 472*". Now we do care about the order. "*724*" won't work, nor will "*247*". It has to be exactly *4-7-2*.
- The language of Math needs more *precision*:
  - When the order doesn't matter, it is a ***Combination***.
  - When the order does matter it is a ***Permutation***.

Permutations

- Two basic types of permutations:
  - *Repetition is Allowed*: such as the lock above. It could be "333".
  - *No Repetition*: for example the first three people in a running race. You can't be first and second.
- Example (*Repetition Allowed*):
  - We have a lock, there are 10 numbers to choose from (0,1,2,3,4,5,6,7,8,9) and we choose 3 of them:
    - 10 × 10 × ... (3 times) = 10^3 = ***1,000 permutations***.
  - Formula = n^r
    - Where ***n*** is the number of things to choose from,
    - and we choose ***r*** of them,
    - repetition is allowed, and order matters.

Permutations

- Example (*No Repetition*):
    - What order could 16 pool balls be in?
        - 16 × 15 × 14 × 13 × ... = 20,922,789,888,000
    - But maybe we don't want to choose them all, just 3 of them, and that is then:
        - 16 × 15 × 14 = 3,360
    - Without repetition our *choices get reduced each time*.
    - Formula = $\frac{n!}{(n-r)!}$
        - Where ***n*** is the number of things to choose from,
        - and we choose ***r*** of them,
        - No repetitions, and order matters.

Permutations

- Notation:

$$P(n, r) = {}^{n}P_{r} = \frac{n!}{(n-r)!}$$

Combinations

- There are also two types of *combinations* (remember the *order does not matter* now):
  - *Repetition is Allowed*: such as coins in your pocket (5,5,5,10,10)
  - *No Repetition*: such as lottery numbers (2,14,15,27,30,33)

- Example (*Combinations Without Repetition*):
  - This is how lotteries work. The numbers are drawn one at a time, and if we have the lucky numbers (no matter what order) we win!
  - One easy way to explain it is to:
    - assume that the order does matter (i.e., permutations),
    - then alter it so the *order* does *not* matter.

Combinations

- Going back to our pool ball example, let's say we just want to know which 3 pool balls are chosen, not the order.
- We already know that 3 out of 16 gave us 3,360 permutations.
- But many of those are the same to us now, because we don't care what order!

Combinations

- For example, let us say balls *1, 2 and 3* are chosen. These are the possibilities:

Order does matter	Order doesn't matter
1 2 3	1 2 3
1 3 2	
2 1 3	
2 3 1	
3 1 2	
3 2 1	

- *So, the permutations have 6 times as many possibilities.*

Combinations

- In fact there is an easy way to work out how many ways "*1 2 3*" could be placed in order, and we have already talked about it. The answer is:
  - 3! = 3 × 2 × 1 = 6
- So we adjust our permutations formula to *reduce it* by how many ways the objects could be in order (because we aren't interested in their order any more):
  - $\frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r! \times (n-r)!}$
- That formula is so important it is often just written in big parentheses like this: $\frac{\boldsymbol{n}!}{\boldsymbol{r}! \times (\boldsymbol{n}-\boldsymbol{r})!} = \binom{\boldsymbol{n}}{\boldsymbol{r}}$

Geometric Distribution

When will the wait end, for Karachi Kings to win?

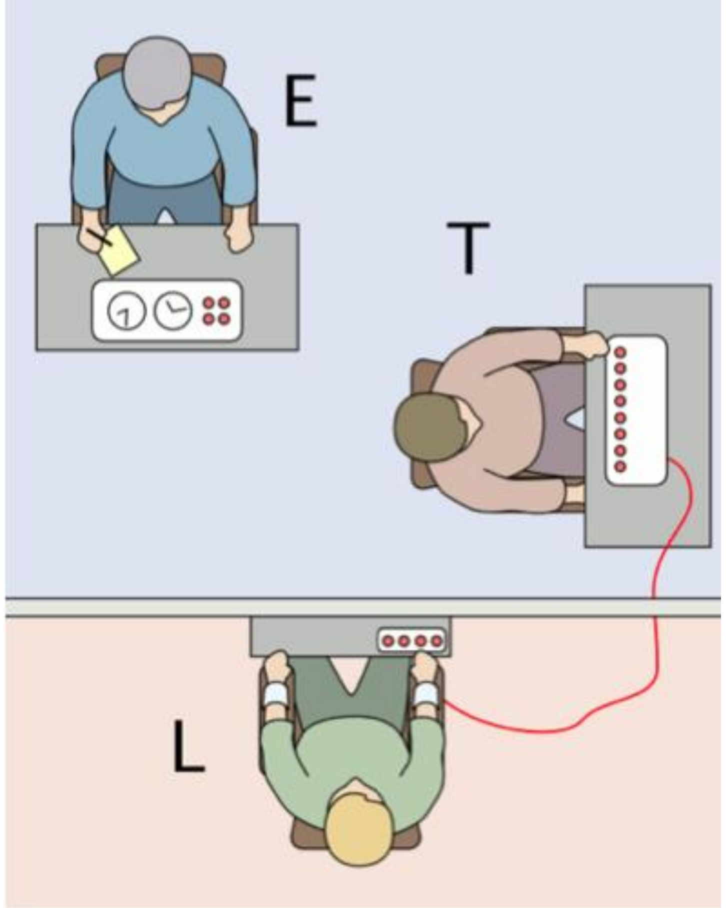
The following slides were developed by Mine Çetinkaya-Rundel of OpenIntro, and translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro

Milgram Experiment (continued)

Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.

- *Experimenter* (E) orders the *teacher* (T), the subject of the experiment, to give severe electric shocks to a *learner* (L) each time the learner answers a *question incorrectly*.

- The *learner* is actually an *actor*, and the electric shocks are not *real*, but a *pre-recorded sound* is played each time the teacher administers an *electric shock*.



http://en.wikipedia.org/wiki/File:Milgram_Experiment_v2.png

Milgram Experiment (continued)

- These experiments measured the willingness of study participants to *obey an authority figure* who instructed them to perform acts *that conflicted with their personal conscience*.

- Milgram found that about *65% of people would obey authority* and give such shocks.

- Over the years, additional research suggested this number is approximately *consistent across communities* and time.

Bernoulli Random Variables

- Each person in Milgram's experiment can be thought of as a *trial*.

- A person is labeled a *success* if she refuses to administer a severe shock, and *failure* if she administers such shock.

- Since only 35% of people refused to administer a shock, *probability of success* is $p = 0.35$

- When an individual trial has only two possible outcomes, it is called a *Bernoulli random variable*.

Geometric Distribution

- Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{st}\ person\ refuses) = 0.35$$

Geometric Distribution

- Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{st}\ person\ refuses) = 0.35$$

- ... the third person?

$$P(1^{st}\ and\ 2^{nd}\ shock,\ 3^{rd}\ refuses) = \underset{0.65}{\underline{S}} \times \underset{0.65}{\underline{S}} \times \underset{0.35}{\underline{R}} = 0.65^2 \times 0.35 \approx 0.15$$

- ... the tenth person?

$$P(9\ shock,\ 10^{th}\ refuses) = \underbrace{\underset{0.65}{\underline{S}} \times \cdots \times \underset{0.65}{\underline{S}}}_{9\ of\ these} \times \underset{0.35}{\underline{R}} = 0.65^9 \times 0.35 \approx 0.0072$$

Geometric Distribution

The *geometric distribution* describes the waiting time until a success for *independent and identically distributed (i.i.d.)* Bernoulli random variables.

- *Independence*: outcomes of trials don't affect each other
- *Identical*: the probability of success is the same for each trial

Geometric probabilities

If *p* represents probability of success, *(1 - p)* represents probability of failure, and *n* represents number of independent trials

$$P(\text{success on the } n^{th} \text{ trial}) = (1-p)^{n-1}p$$

Practice

- Can we calculate the probability of rolling a 6 for the first time on the 6th roll of a die using the geometric distribution?
  - Note that what was a success (rolling a 6) and what was a failure (not rolling a 6) are clearly defined and one or the other must happen for each trial.

- What is your opinion?
  1. no, on the roll of a die there are more than 2 possible outcomes
  2. yes, why not

$$P(6 \text{ } on \text{ } the \text{ } 6^{th} \text{ } roll) = \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right) \approx 0.067$$

Expected Value

- How many people is Dr. Smith expected to test before finding the first one that refuses to administer the shock?
- The expected value, or the mean, of a geometric distribution is defined as $\frac{1}{p}$

$$\mu = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

- She is expected to test 2.86 people before finding the first one that refuses to administer the shock.
- But how can she test a non-whole number of people?

Expected Value and it’s Variability

- Mean and standard deviation of geometric distribution:

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

- Going back to Dr. Smith's experiment:

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.35}{0.35^2}} = 2.3$$

- Dr. Smith is expected to test *2.86 people before finding the first one that refuses* to administer the shock, *give or take 2.3 people*.
- These values only make sense in the *context of repeating the experiment many many times*.

Binomial Distribution

- Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

- Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

$$\text{Scenario 1:}\quad \underset{\text{(A) }\textit{refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\text{Scenario 2:}\quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) }\textit{refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\text{Scenario 3:}\quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) }\textit{refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\text{Scenario 4:}\quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) }\textit{refuse}}{\underline{0.35}} = 0.0961$$

- The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

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* = 4 trials), and we calculated this probability as

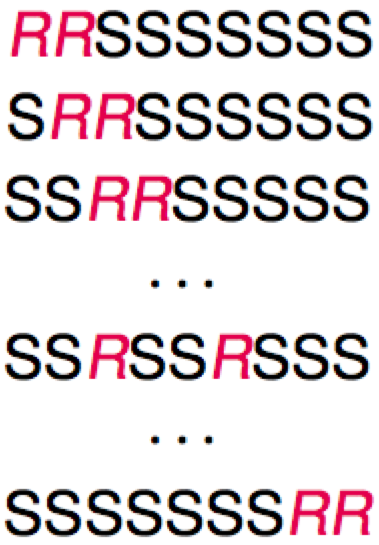
# of scenarios x P(single scenario)

- # of scenarios: there is a less tedious way to figure this out, we'll get to that shortly...
- $P(single\ scenario) = p^k(1-p)^{(n-k)}$
  - where *p* is the probability of success to the power of number of successes, probability of failure to the power of number of failures

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

Computing the # (number) of scenarios

- Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If *n* was larger and/or *k* was different than 1, for example, *n* = 9 and *k* = 2:



*RR*SSSSSSS

S*RR*SSSSSS

SS*RR*SSSSS

...

SS*R*SS*R*SSS

...

SSSSSSS*RR*

- Writing out all possible scenarios would be very tedious and prone to errors.

Computing the # of scenarios

Choose function

The *choose function* is useful for calculating the number of ways to choose *k* successes in *n* trials.

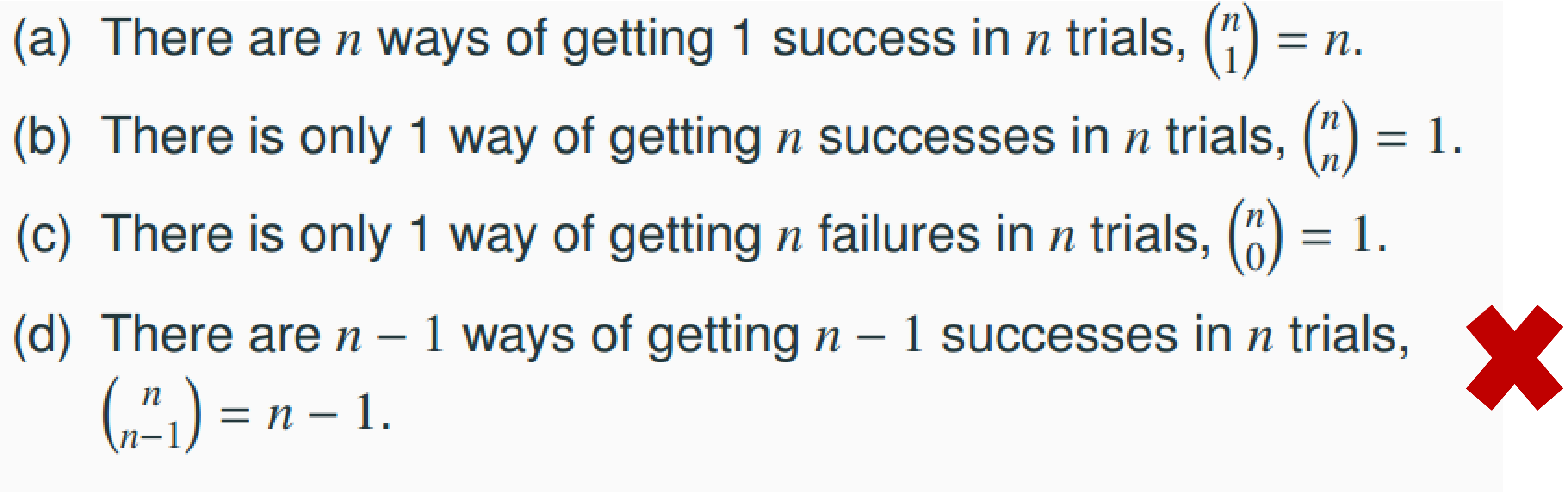
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$k = 1, n = 4: \binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$$

$$k = 2, n = 9: \binom{9}{2} = \frac{9!}{2!(9-2)!} = \frac{9\times8\times7!}{2\times1\times7!} = \frac{72}{2} = 36$$

Practice

- Which of the following is false?



(a) There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.

(b) There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.

(c) There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.

(d) There are $n - 1$ ways of getting $n - 1$ successes in n trials, $\binom{n}{n-1} = n - 1$. ✘

Binomial Distribution (continued)

Binomial probabilities

If *p* represents probability of *success*, *(1-p)* represents probability of *failure*, *n* represents number of *independent trials*, and *k* represents number of *successes*

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

Practice

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

Practice

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

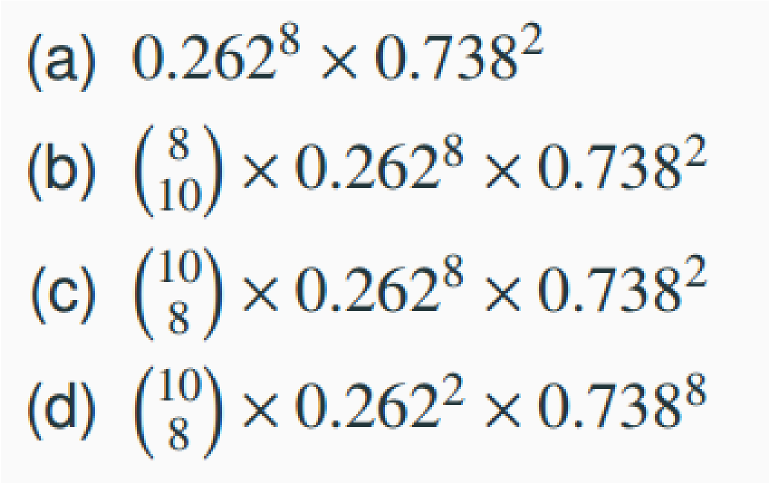
c) each trial outcome must be classified as a *success* or a *failure*

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

Practice

- A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?



(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$ ←

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Number of fraudulent transactions?

- Suppose it is known that 2% of all credit card transactions in a certain region are *fraudulent*
- If there are 50 transactions per day in a certain region
  - ***What is the probability that 3 or less than 3 transactions are fraudulent?***

Number of fraudulent transactions?

- Suppose it is known that 2% of all credit card transactions in a certain region are *fraudulent*
- If there are 50 transactions per day in a certain region
  - ***What is the probability that 3 or less than 3 transactions are fraudulent?***
    - P(X=3) = 0.06067
    - P(X=2) = 0.18580
    - P(X=1) = 0.37160
    - P(X=0) = 0.36417

Number of spam emails?

- Suppose it is known that 4% of all emails are spam
- If an account receives 20 emails in a given day:
  - ***What is the probability that we receive zero spam emails?***
  - ***What is the probability that we receive more than two spam emails?***

Number of spam emails?

- Suppose it is known that 4% of all emails are spam
- If an account receives 20 emails in a given day:
  - ***What is the probability that we receive zero spam emails?***
    - P(X = 0 spam emails) = **0.44200**
  - ***What is the probability that we receive more than two spam emails?***
    - P(X = 0 spam emails) = **0.44200**
    - P(X = 1 spam email) = **0.36834**
    - P(X = 2 spam emails) = **0.14580**
    - **P(X > 2) = 1 – (0.44200+ 0.36834+ 0.14580) = 0.04386**

Binomial Distribution – Expected Value

- A 2012 Gallup survey suggests that 26.2% of Americans are obese.
- Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, 100 x 0.262 = 26.2.
- Or more formally, $\mu = np$ = 100 x 0.262 = 26.2.
- But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Expected Value and it's Variability

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

- Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

- We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

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.

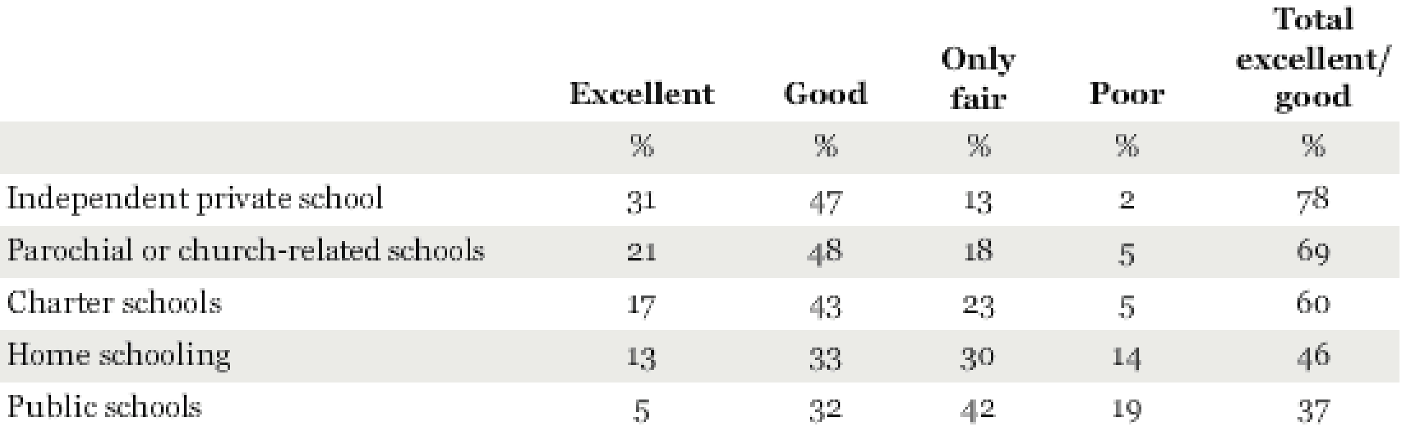
Unusual Observations

Using the notion that *observations that are more than 2 standard deviations away from the mean are considered unusual* and the mean and the standard deviation we just computed, we can calculate a range for the plausible number of obese Americans in random samples of 100.

26.2 ± (2 x 4.4) → (17.4, 35.0)

Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual? → Yes or No??



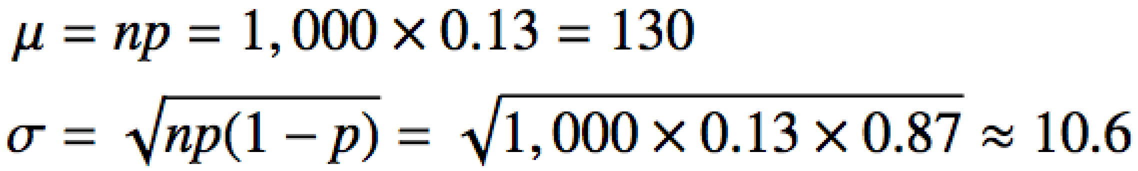
	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

Gallup, Aug. 9-12, 2012

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

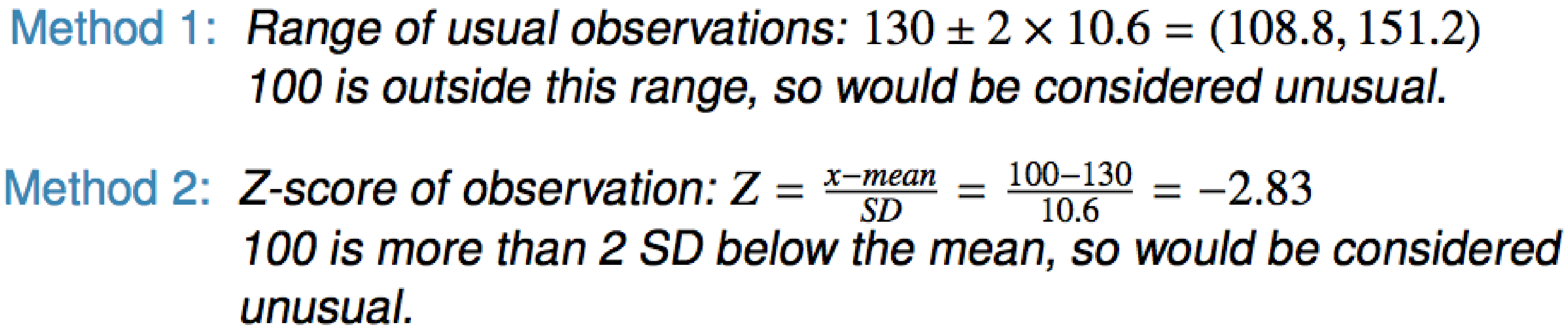
Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual? → Yes or No??



$$\mu = np = 1,000 \times 0.13 = 130$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$



Method 1: *Range of usual observations:* $130 \pm 2 \times 10.6 = (108.8, 151.2)$
100 is outside this range, so would be considered unusual.

Method 2: *Z-score of observation:* $Z = \frac{x-mean}{SD} = \frac{100-130}{10.6} = -2.83$
100 is more than 2 SD below the mean, so would be considered unusual.

Shapes of Binomial Distributions

For this activity you will use a web applet. Go to:

http://socr.stat.ucla.edu/htmls/SOCR_Experiments.html and choose Binomial coin experiment in the drop down menu on the left.

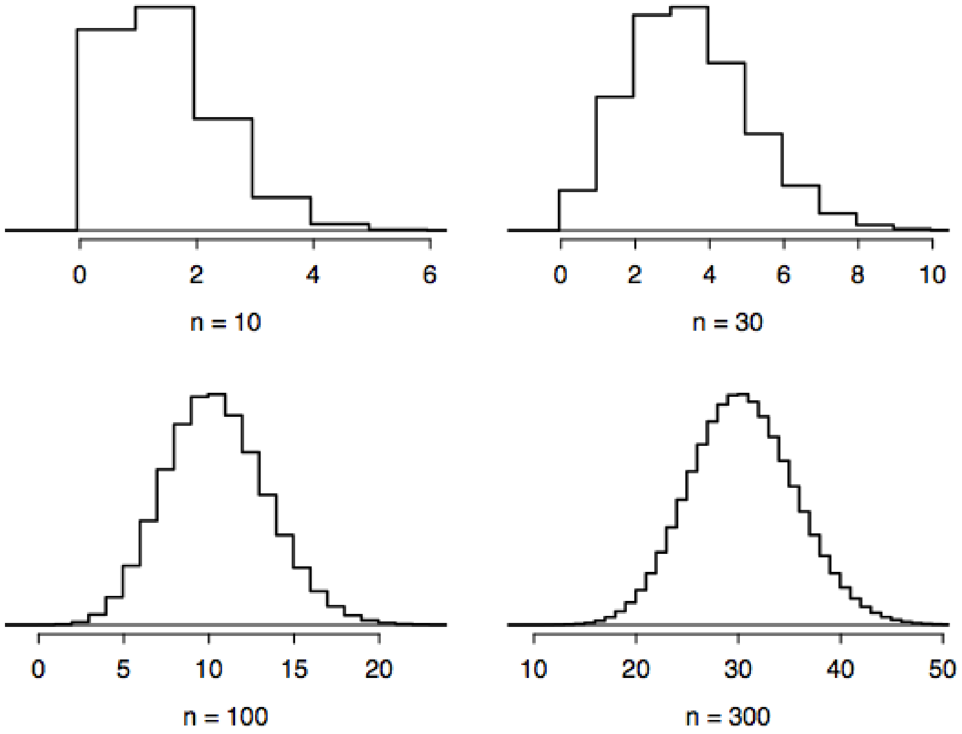
- Set the number of trials to 20 and the probability of success to 0.15. Describe the shape of the distribution of number of successes.

- Keeping *p* constant at 0.15, determine the minimum sample size required to obtain a unimodal and symmetric distribution of number of successes. Please submit only one response per team.

- Further considerations:
  - What happens to the shape of the distribution as *n* stays constant and *p* changes?
  - What happens to the shape of the distribution as *p* stays constant and *n* changes?

Distribution of Number of Successes

- Hollow histograms of samples from the binomial model where *p* = 0.10 and *n* = 10, 30, 100, and 300. What happens as *n* increases?



Sources

- https://en.wikipedia.org/wiki/Bernoulli_trial
- openintro.org/os
- https://byjus.com/maths/binomial-distribution/
- https://www.statology.org/binomial-distribution-real-life-examples/
- https://www.statology.org/binomial-distribution-calculator/
- Combinations and Permutations (mathsisfun.com)