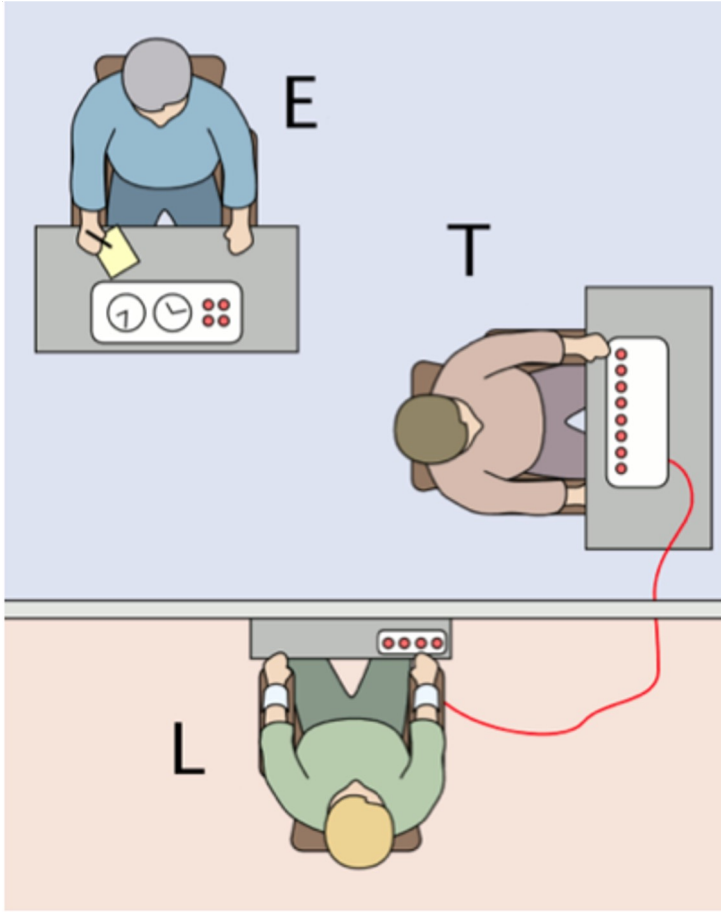
Useful Distributions

Bernouli Distribution

Milgram experiment

Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1961.

- 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 (cont.)

- 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*.

Bernoulli Distribution

- An RV X has the Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The probability function (pf) of X is

- $$f(x \mid p) = \begin{cases} p^x(1-p)^{1-x} & x = 0,1 \\ 0 & \text{otherwise} \end{cases}$$

- An experiment with two outcomes: "success", "failure", $X =$ number of successes.

  Parameter space: $p \in [0,1]$.

Bernoulli Distribution

- An RV X has the Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The probability function (pf) of X is

- $$f(x \mid p) = \begin{cases} p^x(1-p)^{1-x} & x = 0,1 \\ 0 & \text{otherwise} \end{cases}$$

- An experiment with two outcomes: "success", "failure", $X =$ number of successes.

  Parameter space: $p \in [0,1]$.

$$E(X) = p, \mathrm{Var}(X) = p(1-p)$$

Geometric Distribution

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?

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

... the tenth person?

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

Geometric distribution (cont.)

The *geometric distribution* describes the waiting time until a success for *independent and identically distributed (iid)* 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

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

Practice

Find the probability of rolling a 6 for the first time on the 6^{th} roll of a die using the geometric distribution

Practice

Find the probability of rolling a 6 for the first time on the 6^{th} roll of a die using the geometric distribution

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

Geometric distribution

An RV X has the Geometric distribution with parameters n and p a if the probability function (pf) of X is

$$f(x \mid p) = \begin{cases} f(x \mid p) = p(1-p)^{x-1} & x = 1, \dots n \\ 0 & \text{otherwise} \end{cases}$$

- An experiment with two outcomes: "success", "failure", $X =$ number of trials until the first success.
- Sometimes: Number of failures before the first success.

Parameter space $p \in [0,1]$.

- $E(X) = \frac{1-p}{p}, \mathrm{Var}(X) = \frac{1-p}{p^2}$.

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 1/p

$$E(X) = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

She is expected to test 2.86 people before finding the first one that refuses to administer the shock.

Expected value and Variance

- Mean and standard deviation of geometric distribution

$$E(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

- Going back to Dr. Smith's experiment:

$$\sigma_x = \sqrt{(1-p)/p^2} = \sqrt{0.65/0.35^2} = 2.3$$

- 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.

Memorylessness

Assume that Dr. Smith has tested 3 people and all of them administered the shock.

What is the probability she will need to test two more people before she finds one that refuses?

$$P(X = 5|X \geq 3) = P(X = 2)$$

$$P(X = k + t|X \geq k) = P(X = t)$$

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":

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

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

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

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

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 # of scenarios

Binomial coefficient.

The *binomial coefficient (n choose k)*is useful for calculating the number of ways to choose *k* successes in *n* trials.

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

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

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

Note: You can also use R for these calculations:

```
> choose(9,2)
[1] 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, $\binom{n}{n-1} = n - 1$.

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 (cont.)

An RV X has the Binomial distribution with parameters n and p a if the probability function (pf) of X is

$$f(x \mid p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,\dots,n \\ 0 & \text{otherwise} \end{cases}$$

n repetitions of an experiment with two outcomes: "success", "failure", X = number of successes.

- Parameter space: n positive integer, $p \in [0,1]$.

- $E(X) =$ $\quad\quad$ $\mathrm{Var}(X) =$.

Binomial distribution (cont.)

An RV X has the Binomial distribution with parameters n and p a if the probability function (pf) of X is

$$f(x \mid p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,\dots,n \\ 0 & \text{otherwise} \end{cases}$$

n repetitions of an experiment with two outcomes: "success", "failure", X = number of successes.

- Parameter space: n positive integer, $p \in [0,1]$.

- $E(X) = \mathrm{np}$ $\quad\quad$ $\mathrm{Var}(X) = \mathrm{np}(1-\mathrm{p})$.

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?

Gallup: http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx, January 23, 2013.

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

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) $\binom{10}{8} \times 0.262^8 \times 0.738^2 = 45 \times 0.262^8 \times 0.738^2 = 0.0005$

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

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, $E(X) = np$ = 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, variance

Mean and standard deviation of binomial distribution

$$E(X) = \frac{1}{p}, \qquad \mathrm{Var}(X) = np(1-p)$$

Going back to the obesity rate:

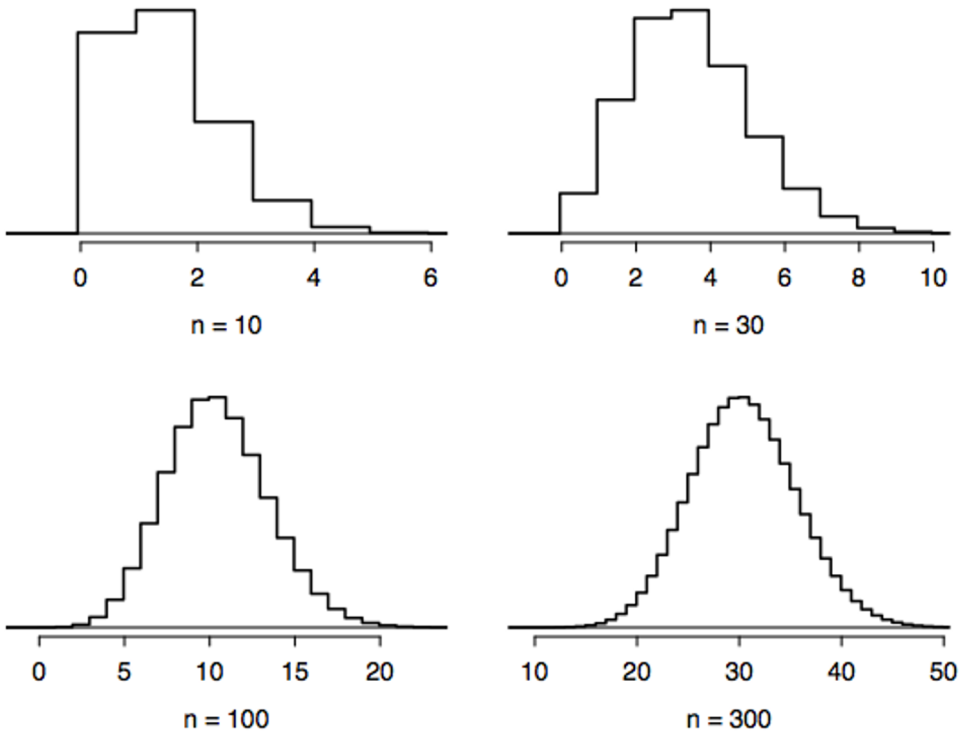
$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

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.

Distributions 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?



An analysis of Facebook users

A recent study found that “Facebook users get more than they give”. For example:

1. 40% of Facebook users in our sample made a friend request, but 63% received at least one request
2. Users in our sample pressed the like button next to friends' content an average of 14 times, but had their content ``liked'' an average of 20 times
3. Users sent 9 personal messages, but received 12
4. 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

Power users contribute much more content than the typical user.

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Practice

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

Practice

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$P(X \geq 70) = P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \ldots \text{ or } K = 245)$$
$$= P(K = 70) + P(K = 71) + P(K = 72) + \ldots + P(K = 245)$$

This seems like an awful lot of work...

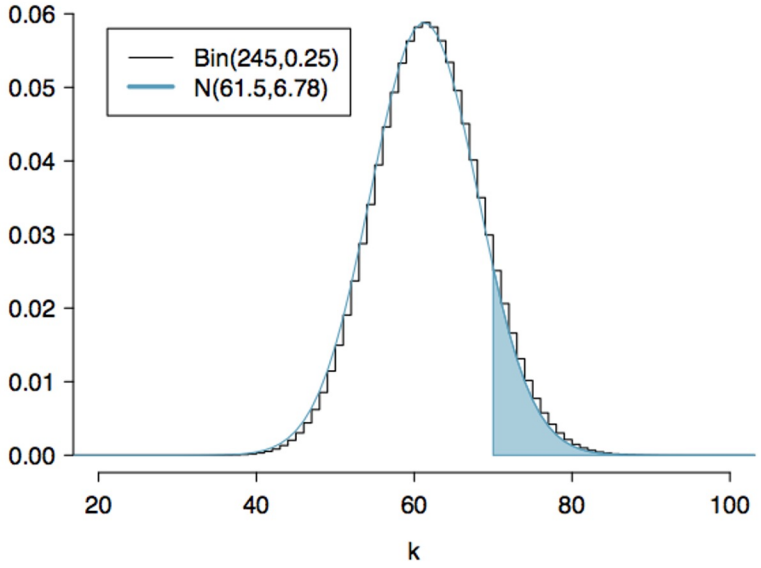
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

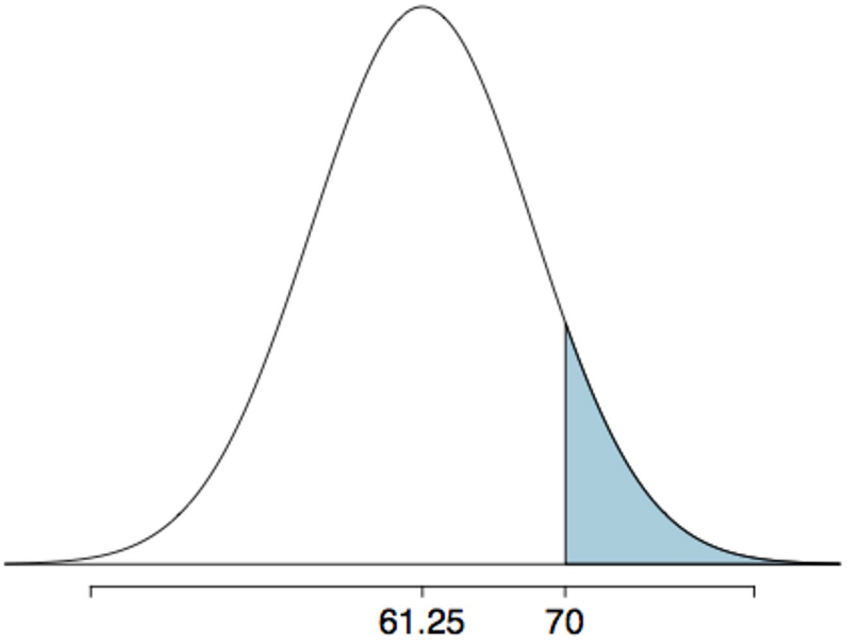
$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



Practice

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$