Unit 2: Probability and distributions

3. Normal and binomial distributions

Sta 104 - Summer 2015

Duke University, Department of Statistical Science

May 21, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bit.ly/sta104su15

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

- Lab 2 + PA 2 due Sunday night
- PS 2 due Monday night
- Lab 2 tomorrow, no class on Monday
- RA 3 on Tuesday, covers Unit 3 - Parts 1 through 4 (not Part 5)
- Midterm next Friday, covers everything up to Unit 3 - Part 4 (Unit 3 - Part 5 not included)
- Any questions on the project?

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

- 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

- 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
- 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).

Discrete:
In a card game if you draw an ace from a well-shuffled full deck you win \$10. If you draw a red card, you lose \$2.

Discrete:
In a card game if you draw an ace from a well-shuffled full deck you win \$10. If you draw a red card, you lose \$2.

Outcome (\$)	X	P(X)
Win \$10 (black aces)	10	$\frac{2}{52}$
Win \$8 (red aces: 10 - 2)	8	$\frac{2}{52}$
Lose \$2 (non-ace reds)	-2	$\frac{24}{52}$
No win / loss	0	$\frac{24}{52}$
		$\frac{52}{52} = 1$

Discrete:
In a card game if you draw an ace from a well-shuffled full deck you win \$10. If you draw a red card, you lose \$2.

Outcome (\$)	X	P(X)
Win \$10 (black aces)	10	$\frac{2}{52}$
Win \$8 (red aces: 10 - 2)	8	$\frac{2}{52}$
Lose \$2 (non-ace reds)	-2	$\frac{24}{52}$
No win / loss	0	$\frac{24}{52}$
		$\frac{52}{52} = 1$

Continuous:
Distribution of weekly expenditures of entertainment for a family is right skewed with median of \$70.

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

Unimodal, symmetric, follows 68-95-99.7 rule

$$N(\mu, \sigma)$$

$$N(\mu, \sigma)$$

- Unimodal and symmetric (bell shaped) that follows very strict guidelines about how variably the data are distributed around the mean

$$N(\mu, \sigma)$$

- Unimodal and symmetric (bell shaped) that follows very strict guidelines about how variably the data are distributed around the mean
- *68-95-99.7 Rule:*
  - about 68% of the distribution falls within 1 SD of the mean
  - about 95% falls within 2 SD of the mean
  - about 99.7% falls within 3 SD of the mean
  - it is possible for observations to fall 4, 5, or more standard deviations away from the mean, but this is very rare if the data are nearly normal

$$N(\mu, \sigma)$$

- Unimodal and symmetric (bell shaped) that follows very strict guidelines about how variably the data are distributed around the mean
- *68-95-99.7 Rule:*
  - about 68% of the distribution falls within 1 SD of the mean
  - about 95% falls within 2 SD of the mean
  - about 99.7% falls within 3 SD of the mean
  - it is possible for observations to fall 4, 5, or more standard deviations away from the mean, but this is very rare if the data are nearly normal
- Lots of variables are nearly normal, but few are actually normal.

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

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 $\rightarrow$ *SD cannot be negative*

(b) *5* $\rightarrow 65 \pm (3 \times 5) = (50, 80)$

(c) 10 $\rightarrow 65 \pm (3 \times 10) = (35, 95)$

(d) 15 $\rightarrow 65 \pm (3 \times 15) = (20, 110)$

(e) 30 $\rightarrow 65 \pm (3 \times 30) = (-25, 155)$

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

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

- Z score: number of standard deviations it falls above or below the mean

3. Z scores serve as a ruler for any distribution

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

- Z score: number of standard deviations it falls above or below the mean
- Defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles

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

- Z score: number of standard deviations it falls above or below the mean
- Defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles
- A Z score creates a common scale so you can assess data without worrying about the specific units in which it was measured.

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

- Z score: number of standard deviations it falls above or below the mean
- Defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles
- A Z score creates a common scale so you can assess data without worrying about the specific units in which it was measured.
- Observations with $|Z| > 2$ are usually considered *unusual*.

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

- Linear transformations of a normally distributed random variable will also be normally distributed.
  If
  $$X \sim N(\mu, \sigma)$$
  and
  $$Y = a + bX,$$
  then
  $$Y \sim N(a + b\mu, b\sigma).$$

- Linear transformations of a normally distributed random variable will also be normally distributed.
  If
  $$X \sim N(\mu, \sigma)$$
  and
  $$Y = a + bX,$$
  then
  $$Y \sim N(a + b\mu, b\sigma).$$
- Hence, if
  $$Z = \frac{X - \mu}{\sigma}, \text{ where } X \sim N(\mu, \sigma),$$
  then
  $$Z \sim N(0, 1)$$

- Linear transformations of a normally distributed random variable will also be normally distributed.
  If
  $$X \sim N(\mu, \sigma)$$
  and
  $$Y = a + bX,$$
  then
  $$Y \sim N(a + b\mu, b\sigma).$$
- Hence, if
  $$Z = \frac{X - \mu}{\sigma}, \text{ where } X \sim N(\mu, \sigma),$$
  then
  $$Z \sim N(0, 1) \rightarrow \textit{standard normal}$$

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.

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.

Application exercise: 2.3 Normal distribution

See the course website for instructions.

Clicker question

Which of the following is <u>false</u>?

(a) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

(b) Majority of Z scores in a right skewed distribution are negative.

(c) In a normal distribution, Q1 and Q3 are more than one SD away from the mean.

(d) Regardless of the shape of the distribution (symmetric vs. skewed) the Z score of the mean is always 0.

Clicker question

Which of the following is false?

(a) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

(b) Majority of Z scores in a right skewed distribution are negative.

(c) *In a normal distribution, Q1 and Q3 are more than one SD away from the mean.*

(d) Regardless of the shape of the distribution (symmetric vs. skewed) the Z score of the mean is always 0.

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

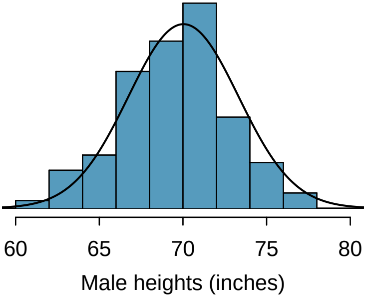
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis

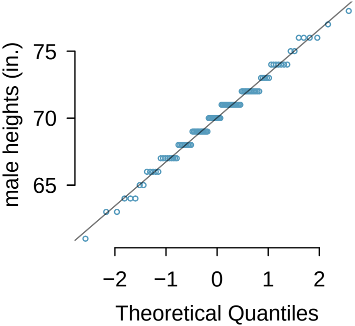
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis
- If there is a linear relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis
- If there is a linear relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution
- Since a linear relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis
- If there is a linear relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution
- Since a linear relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model
- Constructing a normal probability plot requires calculating percentiles and corresponding Z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots

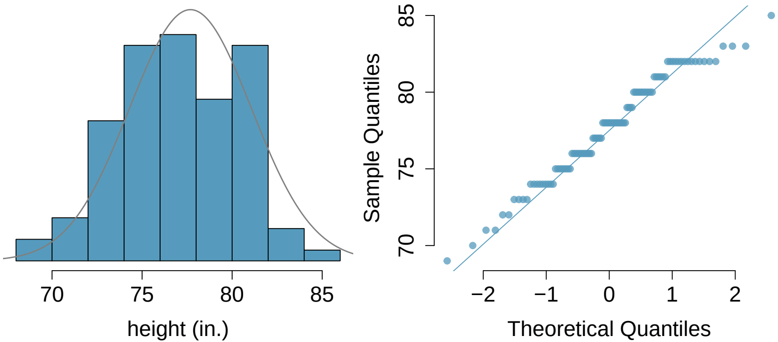
A histogram and *normal probability plot* of a sample of 100 male heights.



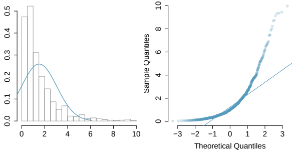


Why do the points on the normal probability have jumps?

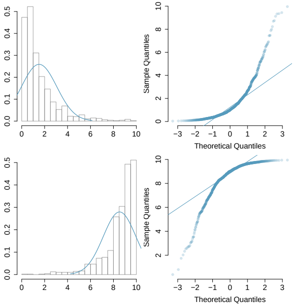
Below is a histogram and normal probability plot for the heights of Duke men's basketball players (from 1990s and 2000s). Do these data appear to follow a normal distribution?



Source: GoDuke.com

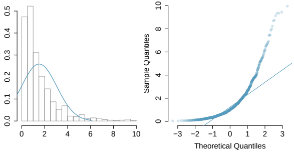


Right Skew - Points bend up and to the left

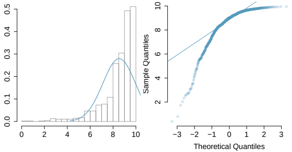


Right Skew - Points bend up and to the left

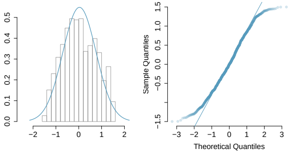
Left Skew - Points bend down and to the right



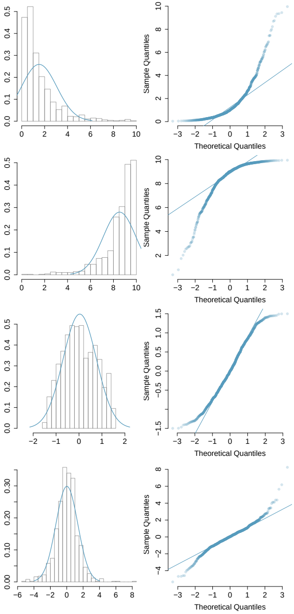
Right Skew - Points bend up and to the left



Left Skew - Points bend down and to the right



Skinny Tails - S shaped-curve indicating shorter than normal tails (narrower, less variable, than expected)



Right Skew - Points bend up and to the left

Left Skew - Points bend down and to the right

Skinny Tails - S shaped-curve indicating shorter than normal tails (narrower, less variable, than expected)

Fat Tails - Curve starting below the normal line, bends to follow it, and ends above it (wider, more variable, than expected)

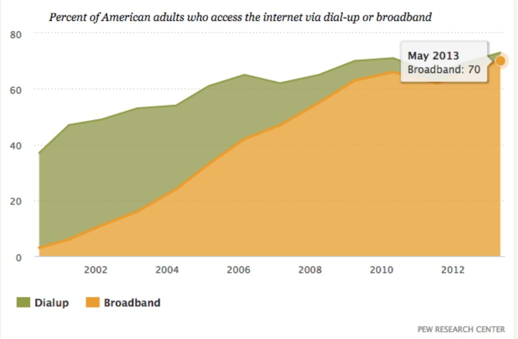
1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

1. Discrete & continuous probability distributions
2. Unimodal, symmetric, follows 68-95-99.7 rule
3. Z scores serve as a ruler for any distribution
4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
5. Normally distributed data plot as a straight line on the normal probability plot

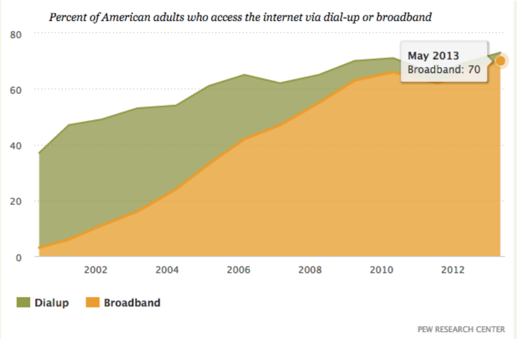
1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

High-speed broadband connection at home in the US

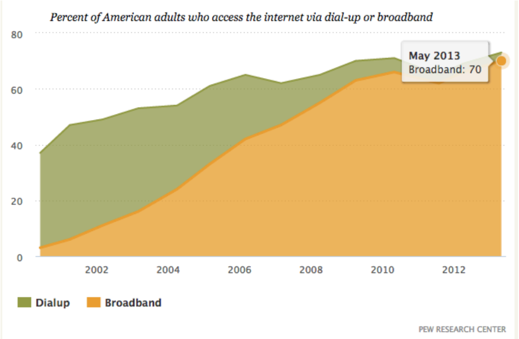


High-speed broadband connection at home in the US



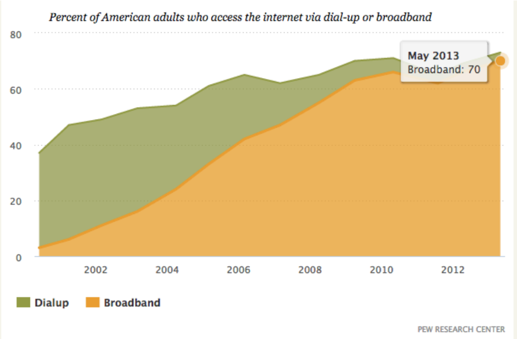
- Each person in the poll be thought of as a *trial*

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

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 $p = 0.70$

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":

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\# \text{ } of \text{ } scenarios \times P(single \text{ } scenario)$$

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

  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 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.

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

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

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

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

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

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

```
[1] 10
```

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

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

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

```
[1] 10
```

- And to compute probabilities

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

```
[1] 0.189
```

Clicker question

Which of the following is false? *Hint:* If you're not sure, pick any number for n (choose a low number to make your life easier) and calculate.

(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, $\binom{n}{n-1} = n-1$.

Clicker question

Which of the following is false? *Hint:* If you're not sure, pick any number for n (choose a low number to make your life easier) and calculate.

(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, $\binom{n}{n-1} = n-1$.*

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 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) $\binom{15}{2} \times 0.70^{13} \times 0.30^2$

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) $\binom{15}{2} \times 0.70^2 \times 0.30^{13}$
$= \frac{15!}{13! \times 2!} \times 0.70^2 \times 0.30^{13} = 105 \times 0.70^2 \times 0.30^{13} = 8.2e-06$

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

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

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?

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?

- $100 \times 0.70 = 70$

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, $\mu = np = 100 \times 0.7 = 7$

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?

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.

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

http://bitly.com/dist_calc

http://bitly.com/dist_calc

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

http://bitly.com/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

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

Clicker question

Below are four pairs of Binomial distribution parameters. Which distribution's shape can be approximated by the normal distribution?

(a) $n = 25, p = 0.45$

(b) $n = 100, p = 0.95$

(c) $n = 150, p = 0.05$

(d) $n = 500, p = 0.015$

Clicker question

Below are four pairs of Binomial distribution parameters. Which distribution's shape can be approximated by the normal distribution?

(a) $n = 25, p = 0.45 \rightarrow 25 \times 0.45 = 11.25$ *and* $25 \times 0.55 = 13.75$

(b) $n = 100, p = 0.95 \rightarrow 100 \times 0.95 = 95$ *but* $100 \times 0.05 = 5$

(c) $n = 150, p = 0.05 \rightarrow 150 \times 0.05 = 7.5$

(d) $n = 500, p = 0.015 \rightarrow 500 \times 0.015 = 7.5$

Application exercise: 2.4 Binomial distribution

See course website for details.

Why do we care?

1. Housekeeping
2. Main ideas - Normal distribution
    1. Discrete & continuous probability distributions
    2. Unimodal, symmetric, follows 68-95-99.7 rule
    3. Z scores serve as a ruler for any distribution
    4. Z distribution is normal with $\mu = 0$ and $\sigma = 1$
    5. Normally distributed data plot as a straight line on the normal probability plot
3. Summary
4. Main ideas - Binomial distribution
    1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
    2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
    3. Shape of the binomial distribution approaches normal when the S-F rule is met
5. Summary

1. Binomial distribution is used for calculating the probability of exact number of successes for a given number of trials
2. Expected value and standard deviation of the binomial can be calculated using its parameters n and p
3. Shape of the binomial distribution approaches normal when the S-F rule is met