Unit 3: Foundations for inference

4. Decision errors and significance levels

Sta 101 - Fall 2015

Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

Slides posted at *http://bit.ly/sta101_f15*

Outline

- MT Review session: Saturday 4-5pm at Old Chem 116

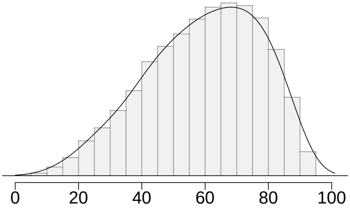
Outline

Refer to previous slide deck.

Outline

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

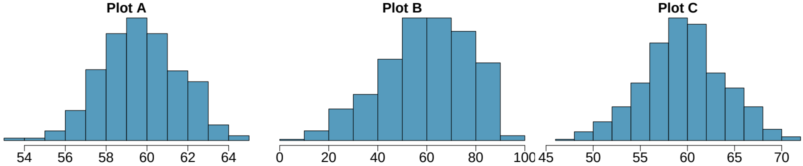


(a) (2) - B; (3) - A; (4) - C

(b) (2) - A; (3) - B; (4) - C

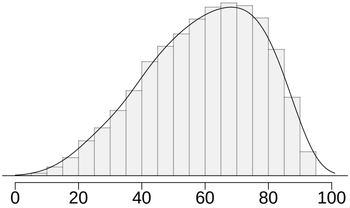
(c) (2) - C; (3) - A; (4) - D

(d) (2) - B; (3) - C; (4) - A



Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

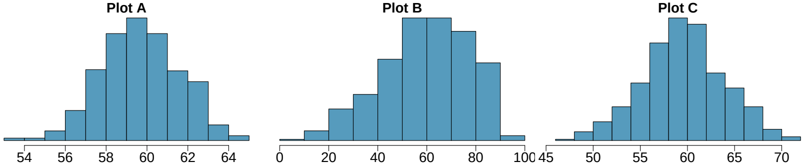


(a) (2) - B; (3) - A; (4) - C

(b) (2) - A; (3) - B; (4) - C

(c) (2) - C; (3) - A; (4) - D

(d) *(2) - B; (3) - C; (4) - A*



A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Would you expect most houses in Topanga to cost more or less than \$1.3 million? Hint: What is most likely the shape of this distribution?

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Would you expect most houses in Topanga to cost more or less than $1.3 million? Hint: What is most likely the shape of this distribution?

Since the distribution is probably right skewed, the median would be less than the mean, and a majority of observations would be lower than the mean.

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?

(a) yes

(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than \$1.4 million using the normal distribution?

(a) yes

(b) *no*

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

(a) yes
(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

(a) *yes*

(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X}$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\Bigg(mean = 1.3,$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

$$P(\bar{X} > 1.4) \quad = \quad P\left(Z > \frac{1.4 - 1.3}{0.0387}\right)$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

$$\begin{aligned} P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \\ &= P(Z > 2.58) \end{aligned}$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

$$\begin{aligned} P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \\ &= P(Z > 2.58) \\ &= 1 - 0.9951 \end{aligned}$$

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

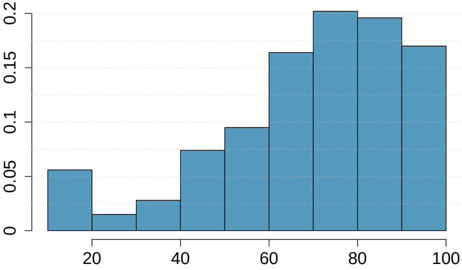
In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

$$\begin{aligned} P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \\ &= P(Z > 2.58) \\ &= 1 - 0.9951 = 0.0049 \end{aligned}$$

Clicker question

Which of the following is false?



(a) The box plot would have outliers only on the lower end.

(b) The median is between 70 and 80.

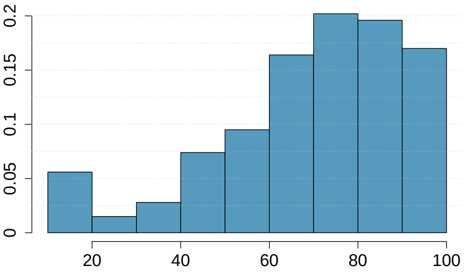
(c) More than 25% of the data is above 90.

(d) More than 50% of the data have positive Z scores.

(e) The mean is likely to be smaller than the median.

Clicker question

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



(a) The box plot would have outliers only on the lower end.

(b) The median is between 70 and 80.

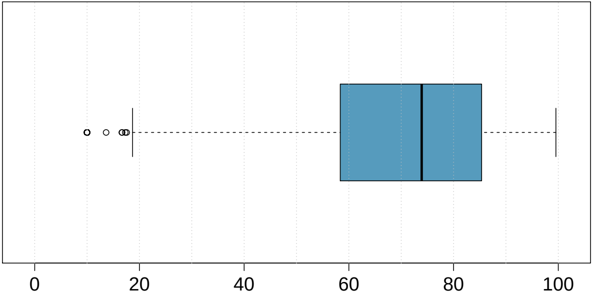
(c) *More than 25% of the data is above 90.*

(d) More than 50% of the data have positive Z scores.

(e) The mean is likely to be smaller than the median.

Clicker question

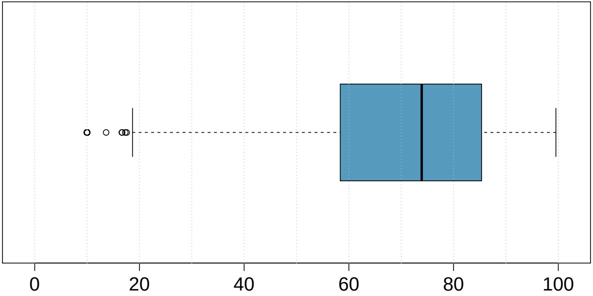
Which of the following is not necessarily true?



(a) Fewer observations are above 90 than below 90.
(b) Fewer observations are below 60 than above 60.
(c) Fewer observations are below 50 than above 50.
(d) The distribution is left skewed.

Clicker question

Which of the following is not necessarily true?



(a) Fewer observations are above 90 than below 90.

(b) *Fewer observations are below 60 than above 60.*

(c) Fewer observations are below 50 than above 50.

(d) The distribution is left skewed.

Outline

Clicker question

Which of the following is true?

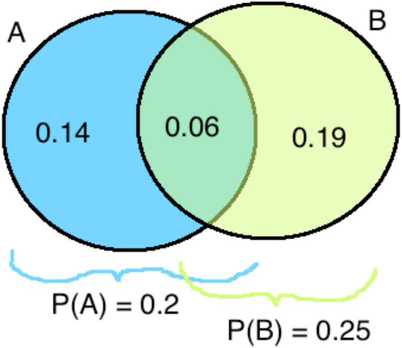
(a) A and B are independent.

(b) P(A but not B) = 0.2

(c) P(A | B) = 0.06 / 0.14

(d) P(A or B) = 0.14 + 0.06 + 0.19

(e) P(neither A nor B) = 1 - 0.06



Clicker question

Which of the following is true?

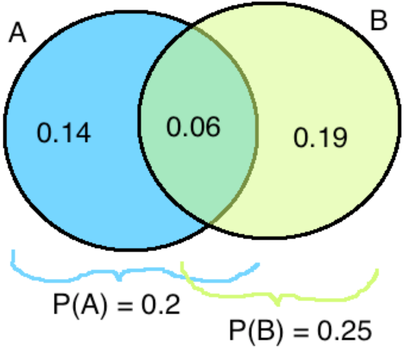
(a) A and B are independent.

(b) $P(A \text{ but not } B) = 0.2$

(c) $P(A \mid B) = 0.06 / 0.14$

(d) *$P(A \text{ or } B) = 0.14 + 0.06 + 0.19$*

(e) $P(\text{neither } A \text{ nor } B) = 1 - 0.06$

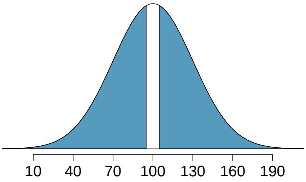


Clicker question

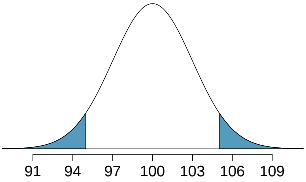
We want to conduct the following hypothesis test. Which is the correct distribution/p-value sketch associated with it?

$$H_0 : \mu = 100; H_A : \mu \neq 100 \qquad \bar{x} = 95, s = 30, n = 100$$

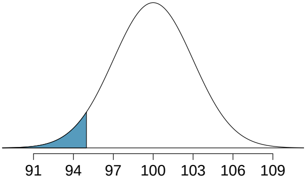
(a)



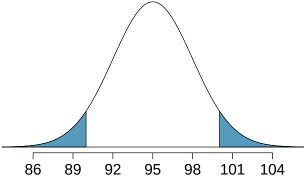
(c)



(b)

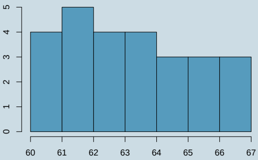


(d)



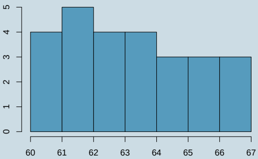
A random sample of 36 female college-aged dancers was obtained and their heights (in inches) were measured. Provided below are some summary statistics and a histogram of the distribution of these dancers' heights. The average height of all college-aged females is 64.5 inches. Do these data provide convincing evidence that the average height of female college-aged dancers is lower from this value?

n	36
$mean$	63.6 inches
sd	2.13 inches



A random sample of 36 female college-aged dancers was obtained and their heights (in inches) were measured. Provided below are some summary statistics and a histogram of the distribution of these dancers' heights. The average height of all college-aged females is 64.5 inches. Do these data provide convincing evidence that the average height of female college-aged dancers is lower from this value?

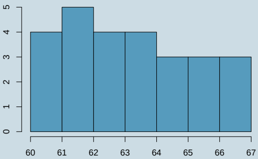
n	36
$mean$	63.6 inches
sd	2.13 inches



$$H_0 : \mu = 64.5$$
$$H_A : \mu < 64.5$$

A random sample of 36 female college-aged dancers was obtained and their heights (in inches) were measured. Provided below are some summary statistics and a histogram of the distribution of these dancers' heights. The average height of all college-aged females is 64.5 inches. Do these data provide convincing evidence that the average height of female college-aged dancers is lower from this value?

n	36
$mean$	63.6 inches
sd	2.13 inches



$$H_0 : \mu = 64.5$$
$$H_A : \mu < 64.5$$

$$\bar{x} = 63.6, s = 2.13, n = 36, \alpha = 0.05$$

$$\bar{x} \sim N\left(mean = 64.5, SE = \frac{2.13}{\sqrt{36}} = 0.355\right)$$

$$\bar{x} \sim N\left(mean = 64.5, SE = \frac{2.13}{\sqrt{36}} = 0.355\right)$$

$$Z = \frac{63.6 - 64.5}{0.355} = -2.54$$

$$\bar{x} \sim N\left(mean = 64.5, SE = \frac{2.13}{\sqrt{36}} = 0.355\right)$$

$$Z = \frac{63.6 - 64.5}{0.355} = -2.54$$

$$p - value = P(Z < -2.54) = 0.0055$$

$$\bar{x} \sim N\left(mean = 64.5, SE = \frac{2.13}{\sqrt{36}} = 0.355\right)$$

$$Z = \frac{63.6 - 64.5}{0.355} = -2.54$$

$$p - value = P(Z < -2.54) = 0.0055$$

Since p-value $<$ 0.05, reject H_0. The data provide convincing evidence that the average height of female college-aged dancers is lower than 64.5 inches.

Clicker question

Which of the following is the correct interpretation of the p-value?

(a) If in fact the average height of college aged dancers is less than 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches or lower is 0.0055.

(b) If in fact the average height of college aged dancers is 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches or lower is 0.0055.

(c) If in fact the average height of college aged dancers is 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches is 0.0055.

(d) The probability that the average height of college aged dancers is 64.5 inches is 0.0055.

Clicker question

Which of the following is the correct interpretation of the p-value?

(a) If in fact the average height of college aged dancers is less than 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches or lower is 0.0055.

(b) *If in fact the average height of college aged dancers is 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches or lower is 0.0055.*

(c) If in fact the average height of college aged dancers is 64.5 inches, the probability of observing a random sample of 36 where the average height is 63.6 inches is 0.0055.

(d) The probability that the average height of college aged dancers is 64.5 inches is 0.0055.

Clicker question

What is the equivalent confidence level for this one-sided hypothesis test with $\alpha = 0.05$?

(a) 80%
(b) 90%
(c) 95%
(d) 99.7%
(e) 97.5%

Clicker question

What is the equivalent confidence level for this one-sided hypothesis test with $\alpha = 0.05$?

(a) 80%
(b) *90%*
(c) 95%
(d) 99.7%
(e) 97.5%

Clicker question

If we were to calculate a 95% confidence interval for the average height of college-aged dancers, would this interval include the null value (64.5 inches)?

(a) Yes
(b) No
(c) Cannot tell without calculating the interval

Clicker question

If we were to calculate a 95% confidence interval for the average height of college-aged dancers, would this interval include the null value (64.5 inches)?

(a) Yes

(b) No

(c) *Cannot tell without calculating the interval*

CPR is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. The chest compressions involved with this procedure can also cause internal injuries. Blood thinners that are often given to help release a clot that is causing the heart attack may also negatively affect such internal injuries. An experiment was designed to evaluate if blood thinners have an impact on survival after a heart attack. Patients were randomly divided into a treatment group (received a blood thinner) or the control group (no blood thinner). The outcome variable of interest was whether the patients survived for at least 24 hours.

Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival proportion in the control group and p_t represent the survival proportion for the treatment group.

H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group. $p_t - p_c = 0$.

H_A: Blood thinners do have an impact on survival. $p_t - p_c \neq 0$.

Clicker question

Given these hypotheses, what is the sample statistic?

$$H_0 : p_t - p_c = 0 \qquad H_A : p_t - p_c \neq 0$$

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

(a) (11 / 25) - (39 / 65) = -0.16

(b) (14 / 40) - (11 / 50) = 0.13

(c) (14 / 90) - (11 / 90) = 0.033

(d) (40 / 90) - (50 / 90) = -0.111

Clicker question

Given these hypotheses, what is the sample statistic?

$$H_0 : p_t - p_c = 0 \qquad H_A : p_t - p_c \neq 0$$

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

(a) (11 / 25) - (39 / 65) = -0.16

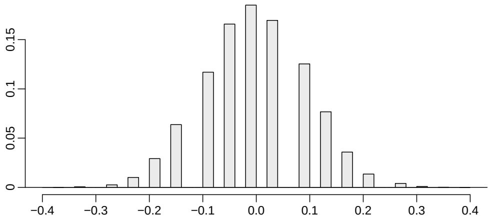
(b) *(14 / 40) - (11 / 50) = 0.13*

(c) (14 / 90) - (11 / 90) = 0.033

(d) (40 / 90) - (50 / 90) = -0.111

Clicker question

A randomization test was conducted to evaluate these hypotheses. Based on the randomization distribution below, what is the conclusion?



These data

(a) provide convincing evidence that blood thinners

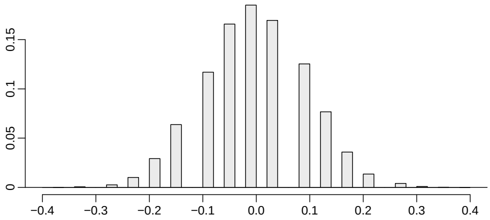
(b) provide convincing evidence that blood thinners do not

(c) do not provide convincing evidence that blood thinners

(d) do not provide convincing evidence that blood thinners do not

Clicker question

A randomization test was conducted to evaluate these hypotheses. Based on the randomization distribution below, what is the conclusion?



These data

(a) provide convincing evidence that blood thinners

(b) provide convincing evidence that blood thinners do not

(c) *do not provide convincing evidence that blood thinners*

(d) do not provide convincing evidence that blood thinners do not

Clicker question

Which of the following probabilities should be calculated using the Binomial distribution?

Probability that

(a) a basketball player misses 3 times in 5 shots

(b) train arrives on the time on the third day for the first time

(c) height of a randomly chosen 5 year old is greater than 4 feet

(d) a randomly chosen individual likes chocolate ice cream best

Clicker question

Which of the following probabilities should be calculated using the Binomial distribution?

Probability that

(a) *a basketball player misses 3 times in 5 shots* $\rightarrow$ *k successes in n trials*

(b) train arrives on the time on the third day for the first time

(c) height of a randomly chosen 5 year old is greater than 4 feet

(d) a randomly chosen individual likes chocolate ice cream best

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

- One possible scenario is that she misses the first three shots, and makes the last two. The probability of this scenario is:

$$0.4^3 \times 0.6^2 \approx 0.023$$

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

- One possible scenario is that she misses the first three shots, and makes the last two. The probability of this scenario is:

$$0.4^3 \times 0.6^2 \approx 0.023$$

- But this isn't the only possible scenario:

1. *MMM*HH
2. *MM*H*M*H
3. *M*H*MM*H
4. H*MMM*H
5. H*MM*H*M*
6. H*M*H*MM*
7. HH*MMM*
8. *M*H*M*H*M*
9. *M*HH*MM*
10. *MM*HH*M*

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

- One possible scenario is that she misses the first three shots, and makes the last two. The probability of this scenario is:

$$0.4^3 \times 0.6^2 \approx 0.023$$

- But this isn't the only possible scenario:

1. *MMM*HH
2. *MM*H*M*H
3. *M*H*MM*H
4. H*MMM*H
5. H*MM*H*M*
6. H*M*H*MM*
7. HH*MMM*
8. *M*H*M*H*M*
9. *M*HH*MM*
10. *MM*HH*M*

- Each one of these scenarios has 3 *M*s and 2 Hs, therefore the probability of each scenario is 0.023.

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

- One possible scenario is that she misses the first three shots, and makes the last two. The probability of this scenario is:

$$0.4^3 \times 0.6^2 \approx 0.023$$

- But this isn't the only possible scenario:

1. *MMM*HH
2. *MM*H*M*H
3. *M*H*MM*H
4. H*MMM*H
5. H*MM*H*M*
6. H*M*H*MM*
7. HH*MMM*
8. *M*H*M*H*M*
9. *M*HH*MM*
10. *MM*HH*M*

- Each one of these scenarios has 3 *M*s and 2 Hs, therefore the probability of each scenario is 0.023.
- Then, the total probability is $10 \times 0.023 = 0.23$.

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

$$\binom{5}{3} \times 0.4^3 \times 0.6^2 \quad = \quad \frac{5!}{3! \times 2!} \times 0.4^3 \times 0.6^2$$

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

$$
\begin{aligned}
\binom{5}{3} \times 0.4^3 \times 0.6^2 &= \frac{5!}{3! \times 2!} \times 0.4^3 \times 0.6^2 \\
&= 10 \times 0.023
\end{aligned}
$$

Suppose the probability of a miss for this basketball player is 0.40. What is the probability that she misses 3 times in 5 shots?

$$\begin{aligned}\binom{5}{3} \times 0.4^3 \times 0.6^2 &= \frac{5!}{3! \times 2!} \times 0.4^3 \times 0.6^2 \\ &= 10 \times 0.023 \\ &= 0.23\end{aligned}$$

Clicker question

Which of the following highlights the correct outcomes for "at most 3 misses in 5 shots"?

(a) {*0, 1, 2*, 3, 4, 5}

(b) {*0, 1, 2, 3*, 4, 5}

(c) {0, 1, 2, *3, 4, 5*}

(d) {0, 1, 2, *3*, 4, 5}

(e) {0, 1, 2, *3, 4*, 5}

Clicker question

Which of the following highlights the correct outcomes for "at most 3 misses in 5 shots"?

(a) {*0, 1, 2*, 3, 4, 5}

(b) *{0, 1, 2, 3, 4, 5}*

(c) {0, 1, 2, *3, 4, 5*}

(d) {0, 1, 2, *3*, 4, 5}

(e) {0, 1, 2, *3, 4*, 5}

Clicker question

Which of the following is the correct calculation for “P(at most 3 misses in 5 shots)”?

Note: P(k) means P(k misses in 5 shots), calculated using the binomial formula.

(a) P(0) + P(1) + P(2)

(b) P(3) + P(4) + P(5)

(c) 1 - P(0)

(d) 1 - [P(0) + P(1) + P(2)]

(e) 1 - [P(4) + P(5)]

Clicker question

Which of the following is the correct calculation for “P(at most 3 misses in 5 shots)”?

Note: P(k) means P(k misses in 5 shots), calculated using the binomial formula.

(a) P(0) + P(1) + P(2)

(b) P(3) + P(4) + P(5)

(c) 1 - P(0)

(d) 1 - [P(0) + P(1) + P(2)]

(e) *1 - [P(4) + P(5)]*

Outline

Suppose that the proportion of people infected with AIDS in a large population is 0.01. If AIDS is present, a certain medical test is positive with probability 0.997 (called the sensitivity of the test). If AIDS is not present, the test is negative with probability 0.985 (called the specificity of the test). If a person tests positive, what is the probability that they have AIDS?

Suppose that the proportion of people infected with AIDS in a large population is 0.01. If AIDS is present, a certain medical test is positive with probability 0.997 (called the sensitivity of the test). If AIDS is not present, the test is negative with probability 0.985 (called the specificity of the test). If a person tests positive, what is the probability that they have AIDS?

- Let's assume there are 1 million individuals in this population.

Suppose that the proportion of people infected with AIDS in a large population is 0.01. If AIDS is present, a certain medical test is positive with probability 0.997 (called the sensitivity of the test). If AIDS is not present, the test is negative with probability 0.985 (called the specificity of the test). If a person tests positive, what is the probability that they have AIDS?

- Let's assume there are 1 million individuals in this population.
- How many are expected to have AIDS, and how many are not expected to have AIDS?

Suppose that the proportion of people infected with AIDS in a large population is 0.01. If AIDS is present, a certain medical test is positive with probability 0.997 (called the sensitivity of the test). If AIDS is not present, the test is negative with probability 0.985 (called the specificity of the test). If a person tests positive, what is the probability that they have AIDS?

- Let's assume there are 1 million individuals in this population.
- How many are expected to have AIDS, and how many are not expected to have AIDS?
  - Have AIDS: $1,000,000 \times 0.01 = 10,000$

Suppose that the proportion of people infected with AIDS in a large population is 0.01. If AIDS is present, a certain medical test is positive with probability 0.997 (called the sensitivity of the test). If AIDS is not present, the test is negative with probability 0.985 (called the specificity of the test). If a person tests positive, what is the probability that they have AIDS?

- Let's assume there are 1 million individuals in this population.
- How many are expected to have AIDS, and how many are not expected to have AIDS?
  - Have AIDS: $1,000,000 \times 0.01 = 10,000$
  - Don't have AIDS: $1,000,000 \times 0.99 = 990,000$

From *http://www.pitt.edu/~nancyp/stat-1000-s07/week6.pdf*.

Clicker question

How many of the people with AIDS would we expect to test positive?

(a) 30
(b) 9,850
(c) 9,970
(d) 987,030
(e) 997,000

Clicker question

How many of the people with AIDS would we expect to test positive?

(a) 30

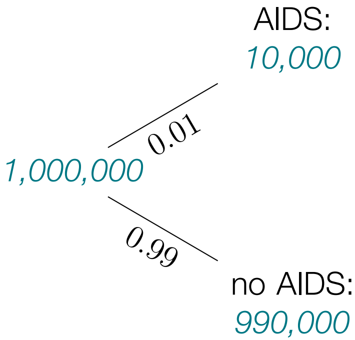
(b) 9,850

(c) *9,970* $\rightarrow 10,000 \times 0.997 = 9970$

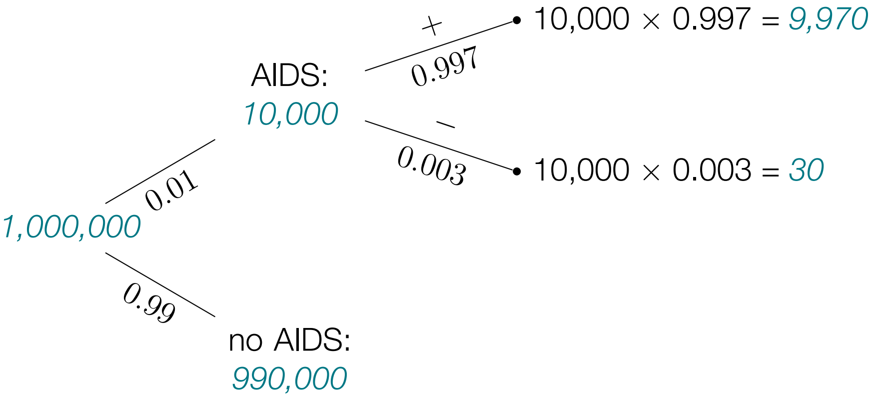
(d) 987,030

(e) 997,000

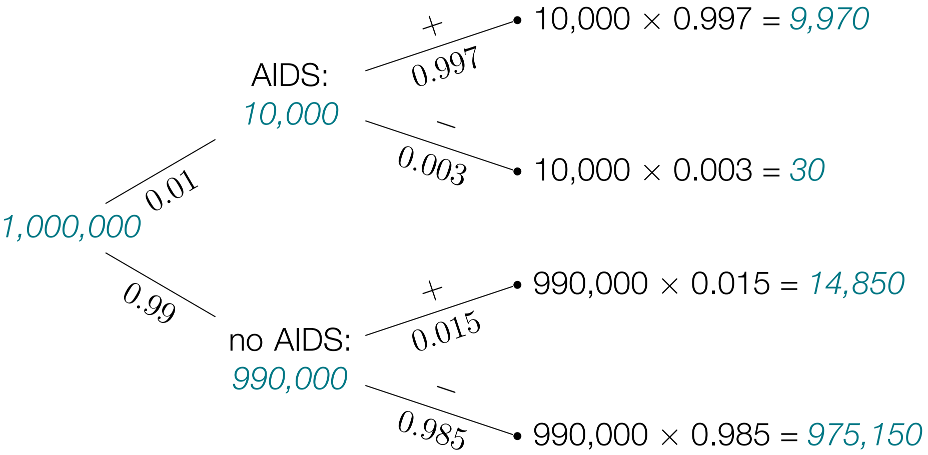
Testing for AIDS -- with counts (cont.)

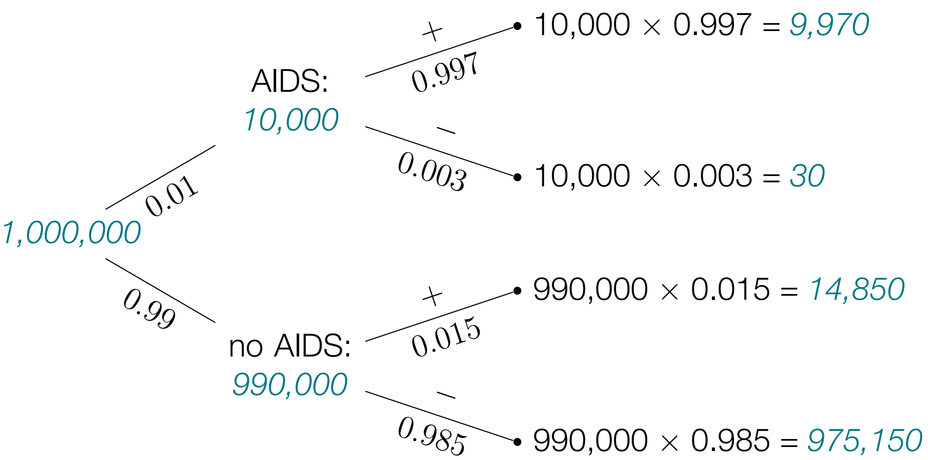


Testing for AIDS -- with counts (cont.)



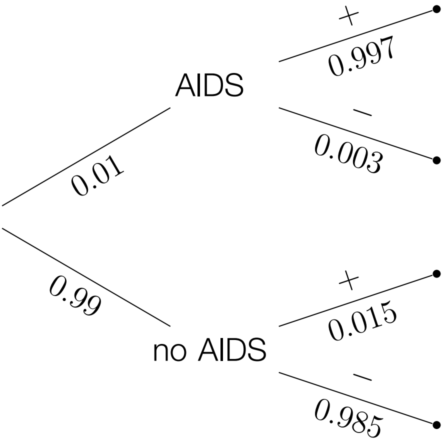
Testing for AIDS -- with counts (cont.)





$$P(AIDS|+) = \frac{9,970}{9,970 + 14,850} \approx 0.40$$

Testing for AIDS -- with probabilities

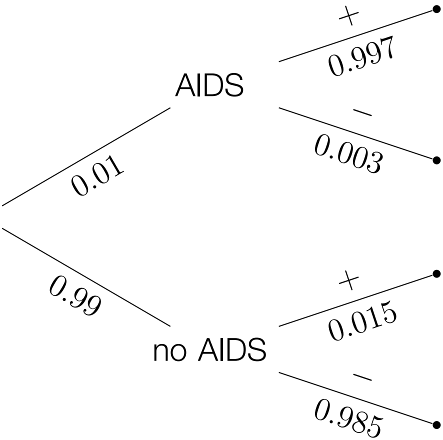


$$P(AIDS\,\&\,+) = 0.01 \times 0.997$$

$$P(AIDS\,\&\,-) = 0.01 \times 0.003$$

$$P(no\ AIDS\,\&\,+) = 0.99 \times 0.015$$

$$P(no\ AIDS\,\&\,-) = 0.99 \times 0.985$$



$$P(AIDS \,\&\, +) = 0.01 \times 0.997$$

$$P(AIDS \,\&\, -) = 0.01 \times 0.003$$

$$P(no\ AIDS \,\&\, +) = 0.99 \times 0.015$$

$$P(no\ AIDS \,\&\, -) = 0.99 \times 0.985$$

$$P(AIDS|+) = \frac{0.01 \times 0.997}{0.01 \times 0.997 + 0.99 \times 0.015} \approx 0.40$$

- In the first stage of testing:
  - Prior: P(AIDS)
    = P(person has AIDS before we collect any data on them) = 0.01
  - Posterior: P(AIDS | test +)
    = P(person has AIDS *given* that they tested positive) = 0.40
- In the second stage of testing:
  - Prior = Posterior from the previous test = 0.40

If the person tests positive for AIDS in the first test, will the prior probability be higher or lower than 1% (prior in the first test)? Why?

- In the first stage of testing:
  - Prior: P(AIDS)
    = P(person has AIDS before we collect any data on them) = 0.01
  - Posterior: P(AIDS | test +)
    = P(person has AIDS *given* that they tested positive) = 0.40
- In the second stage of testing:
  - Prior = Posterior from the previous test = 0.40

If the person tests positive for AIDS in the first test, will the prior probability be higher or lower than 1% (prior in the first test)? Why?

Higher, we're more likely to think that they have AIDS, compared to an average person from this population, since they tested positive once.