Unit 3: Foundations for inference

2. Confidence intervals and hypothesis tests

Sta 104 - Summer 2015

Duke University, Department of Statistical Science

May 27, 2015

Dr. Çetinkaya-Rundel

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

1. Housekeeping

2. Main ideas
    1. Use hypothesis tests to make decisions about population parameters

3. Exercises
    1. Sample vs. sampling distributions
    2. Working with the CLT
    3. Inference for a mean - mechanics
    4. Inference for a mean - interpretations

4. Summary

- Peer evals - please complete asap after class today
- PS2 feedback:
  - 2.8 (b) - Venn diagrams: intersection is P(A and B), and this value should be subtracted from the values show outside the intersection so that the total probability in the circle for an event adds up to that event's marginal probability.
  - 3.16 -

$$
\begin{aligned}
P(X > 2100 | X > 1900) &= \frac{P(X > 2100 \ and \ X > 1900)}{P(X > 1900)} \\
&= \frac{P(X > 2100)}{P(X > 1900)}
\end{aligned}
$$

  - Pay attention to which problems are assigned

1. Housekeeping

2. Main ideas
    1. Use hypothesis tests to make decisions about population parameters

3. Exercises
    1. Sample vs. sampling distributions
    2. Working with the CLT
    3. Inference for a mean - mechanics
    4. Inference for a mean - interpretations

4. Summary

1. Housekeeping

2. Main ideas
    1. Use hypothesis tests to make decisions about population parameters

3. Exercises
    1. Sample vs. sampling distributions
    2. Working with the CLT
    3. Inference for a mean - mechanics
    4. Inference for a mean - interpretations

4. Summary

Hypothesis testing framework:

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

1. Set the hypotheses
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq null\ value$

1. Set the hypotheses
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq null\ value$
2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Sample size / skew: $n \geq 30$ (or larger if sample is skewed), no extreme skew

1. Set the hypotheses
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq null\ value$
2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Sample size / skew: $n \geq 30$ (or larger if sample is skewed), no extreme skew
3. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE},\ \ where\ SE = \frac{s}{\sqrt{n}}$$

1. Set the hypotheses
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq null\ value$
2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Sample size / skew: $n \geq 30$ (or larger if sample is skewed), no extreme skew
3. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE},\ \ where\ SE = \frac{s}{\sqrt{n}}$$

4. Make a decision, and interpret it in context of the research question
   - If p-value $< \alpha$, reject H_0, data provide evidence for H_A
   - If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A

Application exercise: 3.2 Hypothesis testing for a single mean

See course website for details.

Clicker question

Which of the following is the correct interpretation of the p-value from App Ex 3.2?

(a) The probability that average GPA of Duke students has changed since 2001.
(b) The probability that average GPA of Duke students has not changed since 2001.
(c) The probability that average GPA of Duke students has not changed since 2001, if in fact a random sample of 63 Duke students this year have an average GPA of 3.58 or higher.
(d) The probability that a random sample of 63 Duke students have an average GPA of 3.58 or higher, if in fact the average GPA has not changed since 2001.
(e) The probability that a random sample of 63 Duke students have an average GPA of 3.58 or higher or 3.16 or lower, if in fact the average GPA has not changed since 2001.

Clicker question

Which of the following is the correct interpretation of the p-value from App Ex 3.2?

(a) The probability that average GPA of Duke students has changed since 2001.

(b) The probability that average GPA of Duke students has not changed since 2001.

(c) The probability that average GPA of Duke students has not changed since 2001, if in fact a random sample of 63 Duke students this year have an average GPA of 3.58 or higher.

(d) The probability that a random sample of 63 Duke students have an average GPA of 3.58 or higher, if in fact the average GPA has not changed since 2001.

(e) *The probability that a random sample of 63 Duke students have an average GPA of 3.58 or higher or 3.16 or lower, if in fact the average GPA has not changed since 2001.*

1. P-value is the probability that the null hypothesis is true
   A p-value is the probability of getting a sample that results in a test statistic as or more extreme than what you actually observed (in the direction of H_A*, if in fact* H_0 *is correct. It is a conditional probability, conditioned on* H_0 *being correct.*

1. P-value is the probability that the null hypothesis is true
   A p-value is the probability of getting a sample that results in a test statistic as or more extreme than what you actually observed (in the direction of H_A, if in fact H_0 is correct. It is a conditional probability, conditioned on H_0 being correct.

2. A high p-value confirms the null hypothesis.
   A high p-value means the data do not provide convincing evidence for H_A and hence that H_0 can't be rejected.

1. P-value is the probability that the null hypothesis is true
   A p-value is the probability of getting a sample that results in a test statistic as or more extreme than what you actually observed (in the direction of H_A, if in fact H_0 is correct. It is a conditional probability, conditioned on H_0 being correct.

2. A high p-value confirms the null hypothesis.
   A high p-value means the data do not provide convincing evidence for H_A and hence that H_0 can't be rejected.

3. A low p-value confirms the alternative hypothesis.
   A low p-value means the data provide convincing evidence for H_A, but not necessarily that it is confirmed.

1. Housekeeping

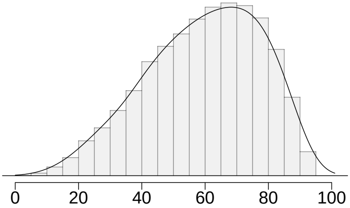
2. Main ideas
    1. Use hypothesis tests to make decisions about population parameters

3. Exercises
    1. Sample vs. sampling distributions
    2. Working with the CLT
    3. Inference for a mean - mechanics
    4. Inference for a mean - interpretations

4. Summary

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.

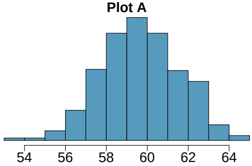


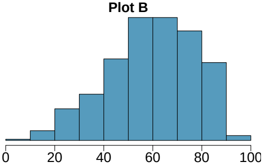
(a) (2) - B; (3) - A; (4) - C

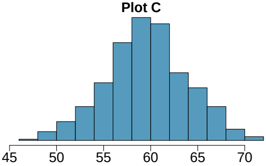
(b) (2) - A; (3) - B; (4) - C

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

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

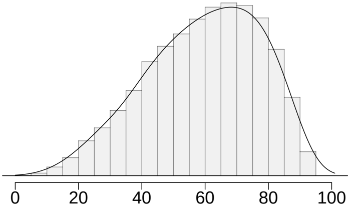






Clicker question

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.

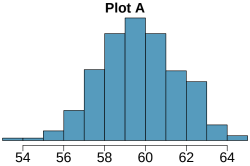


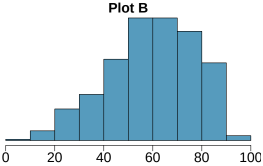
(a) (2) - B; (3) - A; (4) - C

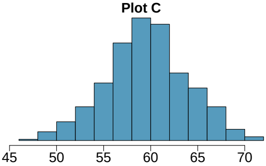
(b) (2) - A; (3) - B; (4) - C

(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,

$$\bar{X} \sim N\left(mean = 1.3,\right.$$

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,

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

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,

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

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

Application exercise: 3.3 Inference for a mean - mechanics

See course website for details.

1. Housekeeping

2. Main ideas
    1. Use hypothesis tests to make decisions about population parameters

3. Exercises
    1. Sample vs. sampling distributions
    2. Working with the CLT
    3. Inference for a mean - mechanics
    4. Inference for a mean - interpretations

4. Summary

Application exercise: 3.4 Inference for a mean - interpretations

See course website for details.

1. Housekeeping

2. Main ideas
    1. Use hypothesis tests to make decisions about population parameters

3. Exercises
    1. Sample vs. sampling distributions
    2. Working with the CLT
    3. Inference for a mean - mechanics
    4. Inference for a mean - interpretations

4. Summary

1. Use hypothesis tests to make decisions about population parameters