Chapter 7: Inference for Numerical Data

Math 140
Based on content in OpenIntro Stats, 4th Ed

Hitchman

April 7, 2022

Section 7.1
Inference on one mean

Run, KBear, Run!

Example: How long does it take my dog to race around the trampoline?

Start Video

- Parameter: μ
  - the true average time it takes Kaizo to race around the trampoline.
- Point estimate: $\overline{x}$
  - a sample mean. For example, I might time him on 8 occasions, and compute the average time for these 8 trials.
- Question: How close is $\overline{x}$ likely to be to μ? In other words, What is the sampling distribution for $\overline{x}$?

Kaizo Data

Here are the times (in seconds) recorded for Kaizo racing around the trampoline.

$$\texttt{times} = c(5.1, 5.4, 5.1, 5.3, 5.2, 5.2, 5.8, 5.1)$$

We note:
sample size: $n = \texttt{length(times)} = 8$.
sample mean: $\overline{x} = \texttt{mean(times)}$ =5.275 seconds.
sample standard deviation $s = \texttt{sd(times)} = 0.2375$ seconds.

Sampling Distribution for a sample mean

Let's simulate:
web app

Central Limit Theorem for the sample mean

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$\text{Mean} = \mu \qquad \text{Standard Error } (SE) = \frac{\sigma}{\sqrt{n}}$$

For instance, suppose for Kaizo, $\mu = 5.0$ seconds and $\sigma = 0.3$ seconds.

1. Assuming the population (of times) is normal, what is the probability that Kaizo will take more than 5.2 seconds to complete a **single lap**?

For instance, suppose for Kaizo, $\mu = 5.0$ seconds and $\sigma = 0.3$ seconds.

1. Assuming the population (of times) is normal, what is the probability that Kaizo will take more than 5.2 seconds to complete a **single lap**?

   Goal: Find $P(X > 5.2)$.

   Strategy: Convert to z-scores and use `pnorm()`.

   $$z = \frac{(5.2 - 5.0)}{0.3} = 2/3,$$

   so

   $$P(X > 5.2) = P(Z > 2/3) = 1 - \texttt{pnorm}(2/3) \approx 0.2525.$$

For instance, suppose for Kaizo, $\mu = 5.0$ seconds and $\sigma = 0.3$ seconds.

1. Assuming the population (of times) is normal, what is the probability that Kaizo will take more than 5.2 seconds to complete a **single lap**?

   Goal: Find $P(X > 5.2)$.

   Strategy: Convert to z-scores and use `pnorm()`.

   $$z = \frac{(5.2 - 5.0)}{0.3} = 2/3,$$

   so

   $$P(X > 5.2) = P(Z > 2/3) = 1 - \texttt{pnorm}(2/3) \approx 0.2525.$$

2. What is the probability that an independent sample of $n = 8$ measurements has a **sample mean** greater than 5.2 seconds?

Suppose for Kaizo, $\mu = 5.0$ seconds and $\sigma = 0.3$ seconds.

2. What is the probability that an independent sample of $n = 8$ measurements has a **sample mean** greater than 5.2 seconds?

Goal: Find $P(\overline{x} > 5.2)$.

Strategy: Convert to z-scores and use `pnorm()`.

Suppose for Kaizo, $\mu = 5.0$ seconds and $\sigma = 0.3$ seconds.

2. What is the probability that an independent sample of $n = 8$ measurements has a **sample mean** greater than 5.2 seconds?

Goal: Find $P(\overline{x} > 5.2)$.

Strategy: Convert to z-scores and use `pnorm()`.

By the CLT, $\overline{x} \sim N(5.0, 0.3/\sqrt{8})$, so

$$z = \frac{(5.2 - 5.0)}{(0.3/\sqrt{8})} \approx 1.886,$$

so $P(\overline{x} > 5.2) = P(z > 1.886) = 1 - pnorm(1.886) = 0.0296$.
So, while there's about a 25% chance that a single lap takes longer than 5.2 seconds, there's only about a 3% chance that the mean time for 8 laps is greater than 5.2 seconds.

Eggs

A certain hen lays eggs with weights that are normally distributed, with $\mu = 50$ grams and standard deviation $\sigma = 2$ grams.

a What is the probability that a single egg weighs more than 51 grams?

b What is the probability that the average weight of four eggs is greater than 51 grams?

c What is the probability that the average weight of a dozen eggs is greater than 51 grams?

d What is the probability that a dozen eggs weigh more than 620 grams?

Eggs

A certain hen lays eggs with weights that are normally distributed, with $\mu = 50$ grams and standard deviation $\sigma = 2$ grams.

a. What is the probability that a single egg weighs more than 51 grams?
b. What is the probability that the average weight of four eggs is greater than 51 grams?
c. What is the probability that the average weight of a dozen eggs is greater than 51 grams?
d. What is the probability that a dozen eggs weigh more than 620 grams?

Answers: (a) 0.3085; (b) 0.1587; (c) 0.0416; (d) 0.0019 (Hint: first convert 620 grams to an average of $\overline{x} = 620/12 = 51.667$.

Before diving into inference on a population mean, we first need to cover two topics related to the CLT:

- certain conditions must be satisfied.
- we rarely know σ. What can we use instead?

Meeting CLT conditions

Independence. The sample observations must be independent, The most common way to satisfy this condition is when the sample is a simple random sample from the population.

Normality. When a sample is small, we also require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger and larger sample sizes.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size if the population distribution you are drawing from is nearly normal.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size if the population distribution you are drawing from is nearly normal.
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets.

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds true for any sample size if the population distribution you are drawing from is nearly normal.
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets.
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.
    - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

Rules of thumb

n $<$ 30: If the sample size n is less than 30 and there are no clear outliers in the data, and the underlying distribution of individual observations is nearly normal, then it is reasonable to assume the sampling distribution of $\overline{x}$ is nearly normal too.

n $\geq$ 30: If the sample size n is at least 30 and there are no *particularly extreme* outliers, then we typically assume the sampling distribution of $\overline{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

What if we don't know σ?

The CLT for means says

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n}).$$

If we don't know σ, is it reasonable to use the sample standard deviation s as an estimate, and say

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})?$$

Not necessarily...

The trouble with using s

The strategy of using s as an estimate for σ tends to work well when we have a lot of data and can estimate σ using s accurately.

However, the estimate is less precise with smaller samples, and this leads to problems when using the normal distribution to model $\bar{x}$.

For small samples, we'll find it useful to use a new distribution for inference calculations called the t-distribution.

Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

The t distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t-distribution*.

The t distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t-distribution*.
- t-distribution page.

Recap: The t-distribution

- Always centered at zero, like the standard normal (z) distribution.
- This distribution also has a bell shape, but its tails are *thicker* than the standard normal model.
- Extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)
- Has a single parameter: *degrees of freedom* (*df*).

Q: *What happens to shape of the t-distribution as df increases?*

Recap: The t-distribution

- Always centered at zero, like the standard normal (z) distribution.
- This distribution also has a bell shape, but its tails are *thicker* than the standard normal model.
- Extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)
- Has a single parameter: *degrees of freedom* (*df*).

Q: *What happens to shape of the t-distribution as df increases?*
Approaches normal.

Kaizo's Lap time around the trampoline

- From the sample, give a 95% confidence interval for μ the average time it takes Kaizo to complete a lap around the trampoline.
- Recall the data: $n = 8$, $\overline{x} = 5.275$, and $s = 0.2375$.

Confidence interval for a mean when n is small.

- Confidence intervals are always of the form

$$\text{point estimate} \pm MOE$$

Confidence interval for a mean when n is small.

- Confidence intervals are always of the form

$$\text{point estimate} \pm MOE$$

- MOE is always calculated as the product of a critical value and SE.

Confidence interval for a mean when n is small.

- Confidence intervals are always of the form

$$\text{point estimate} \pm MOE$$

- MOE is always calculated as the product of a critical value and SE.
- Since sample means from small samples follow a t-distribution (and not a z-distribution), the critical value is a $t^\star$ (as opposed to a $z^\star$).

$$\text{point estimate} \pm t^\star \times SE$$

- More precisely, a confidence interval in this setting looks like:

$$\overline{x} \pm t^* \times \frac{s}{\sqrt{n}}$$

Finding the critical t ($t^{\star}$)

First note that the sample size is $n = 8$, so df $= n - 1 = 7$.
What is the critical t score for 95% confidence in this case?
Using R, we use `qt` (rather than `qnorm`) since we're finding a t^* instead of a z^*:

```
> qt(p = 0.975, df = 7)

[1] 2.364624
```

95% CI for Kaizo mean lap time

Recall the data: $n = 8$, $\bar{x} = 5.275$, and $s = 0.2375$.

$$5.275 \pm 2.3646 \times \frac{0.2375}{\sqrt{8}}$$

$$5.275 \pm 0.199$$

5.076 to 5.474 seconds

Interpreting the CI

Which of the following is the best interpretation for the confidence interval we just calculated?

$$\mu = (5.08, 5.47)$$

(a) 95% of all of Kaizo's lap times will be between 5.08 and 5.47 seconds.

(b) We are 95% confident that on any given lap, Kaizo's time will be between 5.08 and 5.47 seconds.

(c) We are 95% confident that Kaizo's average lap time is between 5.08 and 5.47 seconds.

Interpreting the CI

Which of the following is the best interpretation for the confidence interval we just calculated?

$$\mu = (5.08, 5.47)$$

(a) 95% of all of Kaizo's lap times will be between 5.08 and 5.47 seconds.

(b) We are 95% confident that on any given lap, Kaizo's time will be between 5.08 and 5.47 seconds.

(c) *We are 95% confident that Kaizo's average lap time is between 5.08 and 5.47 seconds.*

Hypotheses

The standard of excellence for trampoline laps is 5 seconds. Do we have reason to believe that Kaizo's true average lap time μ exceeds 5 seconds? Which hypotheses do we want to test this question?

(a) $H_0 : \mu = 5$
$H_A : \mu \neq 5$

(b) $H_0 : \bar{x} = 5$
$H_A : \bar{x} > 5$

(c) $H_0 : \mu = 5$
$H_A : \mu > 5$

(d) $H_0 : p = 5$
$H_A : p > 5$

Hypotheses

The standard of excellence for trampoline laps is 5 seconds. Do we have reason to believe that Kaizo's true average lap time μ exceeds 5 seconds? Which hypotheses do we want to test this question?

(a) $H_0 : \mu = 5$
$H_A : \mu \neq 5$

(b) $H_0 : \bar{x} = 5$
$H_A : \bar{x} > 5$

(c) $H_0 : \mu = 5$
$H_A : \mu > 5$

(d) $H_0 : p = 5$
$H_A : p > 5$

Conditions

- *Independence:* We are told to assume the sample is independent.

Conditions

- *Independence:* We are told to assume the sample is independent.
- *Sample size / skew:*
  - The sample distribution does not appear to be extremely skewed (plotting the 8 times in a histogram in RStudio), but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not – perhaps slightly skewed right, with the occasional slow lap time, but it seems reasonable that the data is nearly normal, and the sample has no extreme outliers.
  - We do not know σ and n is too small to assume s is a reliable estimate for σ.

Conditions

- *Independence:* We are told to assume the sample is independent.
- *Sample size / skew:*
  - The sample distribution does not appear to be extremely skewed (plotting the 8 times in a histogram in RStudio), but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not – perhaps slightly skewed right, with the occasional slow lap time, but it seems reasonable that the data is nearly normal, and the sample has no extreme outliers.
  - We do not know σ and n is too small to assume s is a reliable estimate for σ.

Q: *So what do we do when the sample size is small?*

Finding the test statistic

Test statistic for inference on a mean for a small sample

The test statistic for inference on a small sample mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

Finding the test statistic

Test statistic for inference on a mean for a small sample

The test statistic for inference on a small sample mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

$$point\ estimate \quad = \quad \bar{x} = 5.275$$

Finding the test statistic

Test statistic for inference on a mean for a small sample

The test statistic for inference on a small sample mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

$$\begin{aligned} point\ estimate &= \bar{x} = 5.275 \\ SE &= \frac{s}{\sqrt{n}} = \frac{.2375}{\sqrt{8}} = 0.084 \end{aligned}$$

Finding the test statistic

Test statistic for inference on a mean for a small sample

The test statistic for inference on a small sample mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

$$\begin{aligned} point\ estimate &= \bar{x} = 5.275 \\ SE &= \frac{s}{\sqrt{n}} = \frac{.2375}{\sqrt{8}} = 0.084 \\ T &= \frac{5.275 - 5}{0.084} = 3.274 \end{aligned}$$

Finding the test statistic

Test statistic for inference on a mean for a small sample

The test statistic for inference on a small sample mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

$$\begin{aligned}
point\ estimate &= \bar{x} = 5.275 \\
SE &= \frac{s}{\sqrt{n}} = \frac{.2375}{\sqrt{8}} = 0.084 \\
T &= \frac{5.275 - 5}{0.084} = 3.274 \\
df &= 8 - 1 = 7
\end{aligned}$$

Note: Null value is 5 because in the null hypothesis we set $\mu = 5$.

Finding the p-value

- The p-value is, once again, calculated as a tail area, this time under the t_7 distribution greater than our test statistic (because $H_a : \mu > 5$

Finding the p-value

- The p-value is, once again, calculated as a tail area, this time under the t_7 distribution greater than our test statistic (because $H_a : \mu > 5$
- Using R, we use pt() instead of pnorm()

```
> 1-pt(3.274, df = 7)

[1] 0.006799545
```

Finding the p-value

- The p-value is, once again, calculated as a tail area, this time under the t_7 distribution greater than our test statistic (because $H_a : \mu > 5$
- Using R, we use pt() instead of pnorm()

```
> 1-pt(3.274, df = 7)

[1] 0.006799545
```

- Using a web app:
  https://gallery.shinyapps.io/dist_calc/

Finding the p-value

- The p-value is, once again, calculated as a tail area, this time under the t_7 distribution greater than our test statistic (because $H_a : \mu > 5$
- Using R, we use pt() instead of pnorm()

```
> 1-pt(3.274, df = 7)

[1] 0.006799545
```

- Using a web app:
  https://gallery.shinyapps.io/dist_calc/
- Or when these aren't available, we can use a t-table (see course resource page)

Conclusion of the test

Q: *What is the conclusion of this hypothesis test?*

Conclusion of the test

Q: *What is the conclusion of this hypothesis test?*
Since the p-value is quite low, we conclude that the data provide strong evidence against the null in favor of the alternative. That is, we have strong evidence that Kaizo's average lap time exceeds 5 seconds.

Conclusion of the test

Q: *What is the conclusion of this hypothesis test?*
Since the p-value is quite low, we conclude that the data provide strong evidence against the null in favor of the alternative. That is, we have strong evidence that Kaizo's average lap time exceeds 5 seconds.
But don't fret! He's getting stronger every day!!

Synthesis

Q: *Does the conclusion from the hypothesis test agree with the findings of the confidence interval?*

Synthesis

Q: *Does the conclusion from the hypothesis test agree with the findings of the confidence interval?*

Yes, the hypothesis test found significant evidence that Kaizo's lap time exceeds 5 seconds, and the CI doesn't contain 5, but gives an interval of numbers greater than 5.

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$.

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by random sample, and if sampling w/o replacement, $n < 10\%$ of population)
  - no extreme skew

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by random sample, and if sampling w/o replacement, $n < 10\%$ of population)
  - no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by random sample, and if sampling w/o replacement, $n < 10\%$ of population)
  - no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

- Confidence interval: point estimate $\pm\, t^{\star}_{df} \times SE$