Chapter 5
- Normal distribution (5.1)
- Central limit theorem (5.2)
- Normal distribution for p-values (5.2)
- Normal distribution for confidence intervals (5.2)
- Standard normal (5.2)

---

Malaria Parasites and Mosquitoes
- Mice were randomized to either eat from a malaria infected mouse or a healthy mouse.
- After infection, the parasites go through two stages:
  1) Oocyst (not yet infectious), Days 1-8
  2) Sporozoite (infectious), Days 9 – 28
- Response variable: whether the mosquito approached a human (in a cage with them)
- Does this behavior differ by infected vs control? Does it differ by infection stage?
- Dr. Andrew Read, Professor of Biology and Entomology and Penn State, is a co-author

---

Malaria Parasites and Mosquitoes
- Malaria parasites would benefit if:
  - Mosquitos sought fewer blood meals after getting infected, but before becoming infectious (oocyst stage), because blood meals are risky.
  - Mosquitoes sought more blood meals after becoming infectious (sporozoite stage), to pass on the infection.
- Does infecting mosquitoes with Malaria actually impact their behavior in this way?

---

Oocyst Stage
We’ll first look at the Oocyst stage, after the infected group has been infected, but before they are infectious.

\( p_I \): proportion of infecteds to approach human

\( p_C \): proportion of controls to approach human

What are the relevant hypotheses?

a) \( H_0: p_I = p_C, H_a: p_I < p_C \)

b) \( H_0: p_I = p_C, H_a: p_I > p_C \)

c) \( H_0: p_I < p_C, H_a: p_I = p_C \)

d) \( H_0: p_I > p_C, H_a: p_I = p_C \)

---

Data: Oocyst Stage

\[ \hat{p}_I - \hat{p}_C = \frac{20}{113} - \frac{36}{117} = 0.177 - 0.308 = -0.131 \]
Randomization Test

Randomization Distributions

Normal Distribution

- The symmetric bell-shaped curve we have seen for almost all of our distribution of statistics is called a *normal distribution*

- The normal distribution is fully described by its mean and standard deviation:
  \[ N(\text{mean}, \text{standard deviation}) \]

Randomization and Bootstrap Distributions

Malaria and Mosquitoes

- Which normal distribution should we use to approximate this?
  a) N(0, -0.131)
  b) N(0, 0.056)
  c) N(-0.131, 0.056)
  d) N(0.056, 0)

Normal Distribution

- We can compare the original statistic to this Normal distribution to find the p-value!
p-value from N(null, SE)

Exact same idea as randomization test, just using a smooth curve!

Statistics: Unlocking the Power of Data

Standardized Data

- Often, we standardize the statistic to have mean 0 and standard deviation 1
- How? z-scores!
  \[ z = \frac{x - \text{null value}}{\text{SE}} \]
- What is the equivalent for the null distribution?

Standardized Statistic

The standardized test statistic (also known as a z-statistic) is

\[ z = \frac{\text{statistic} - \text{null}}{\text{SE}} \]

- Calculating the number of standard errors a statistic is from the null lets us assess extremity on a common scale

Malaria and Mosquitoes:

- From original data: statistic = -0.131
- From null hypothesis: null value = 0
- From randomization distribution: SE = 0.056

\[ z = \frac{\text{statistic} - \text{null}}{\text{SE}} = \frac{-0.131 - 0}{0.056} = -2.34 \]

Compare to N(0,1) to find p-value...

Standard Normal

- The standard normal distribution is the normal distribution with mean 0 and standard deviation 1

\[ N(0,1) \]

- Standardized statistics are compared to the standard normal distribution

p-value from N(0,1)

If a statistic is normally distributed under H_0, the p-value can be calculated as the proportion of a N(0,1) beyond

\[ z = \frac{\text{statistic} - \text{null}}{\text{SE}} \]
Statistics: Unlocking the Power of Data

**Sporozoite Stage**

For the data from the Sporozoite stage, after infectious, what are the relevant hypotheses?

- $p_C$: proportion of controls to approach human
- $p_I$: proportion of infecteds to approach human

What are the relevant hypotheses?

a) $H_0: p_I = p_C$, $H_a: p_I < p_C$

b) $H_0: p_I = p_C$, $H_a: p_I > p_C$

c) $H_0: p_I < p_C$, $H_a: p_I = p_C$

d) $H_0: p_I > p_C$, $H_a: p_I = p_C$

Statistics: Unlocking the Power of Data

**Proportion of Infected**

- All mosquitoes in the infected group were exposed to the malaria parasites, but not all mosquitoes were actually infected
- Of the 201 mosquitoes in the infected group that we actually have data on, only 90 were actually infected ($90/201 = 0.448$)
- What proportion of mosquitoes eating from a malaria infected mouse become infected?
- We want a confidence interval!

**Sporozoite Stage**

The difference in proportions is 0.15 and the standard error is 0.05. Is this significant?

a) Yes

b) No

Statistics: Unlocking the Power of Data

**Data**

- **Oocyst Stage**
- **Sporozoite Stage**

The $p$-value is always the proportion in the tail(s) beyond the relevant statistic!

We have evidence that mosquitoes exposed to malaria parasites are less likely to approach a human before they become infectious than mosquitoes not exposed to malaria parasites.

The $p$-value from $N(0,1)$

Exact same idea as before, just standardized!
Bootstrap Interval

If a bootstrap distribution is normally distributed, we can write it as:

a) \( N(\text{parameter}, \text{sd}) \)
b) \( N(\text{statistic}, \text{sd}) \)
c) \( N(\text{parameter}, \text{se}) \)
d) \( N(\text{statistic}, \text{se}) \)

\( \text{sd} = \text{standard deviation of data values} \)
\( \text{se} = \text{standard error} = \text{standard deviation of statistic} \)

Normal Distribution

We can find the middle \( P\% \) of this Normal distribution to get the confidence interval!

CI from \( N(\text{statistic}, \text{SE}) \)

Same idea as the bootstrap, just using a smooth curve!

(Un)-standardization

- Standardized scale:
  \[
  z = \frac{x - \text{mean}}{\text{sd}}
  \]

- To un-standardize:
  \[
  z \cdot \text{sd} = x - \text{mean} \\
  x = \text{mean} + z \cdot \text{sd}
  \]
(Un)-standardization
- In testing, we go to a standardized statistic
- In intervals, we find \((-z^*, z^*)\) for a standardized distribution, and return to the original scale
- Un-standardization (reverse of z-scores):
  \[ x = mean + z \cdot sd \]
- What’s the equivalent for the distribution of the statistic? (bootstrap distribution)

P% Confidence Interval
1. Find values \((-z^*, z^*)\) that capture the middle P% of N(0,1)
2. Return to original scale with statistic \(\pm z^* \times SE\)

Confidence Interval using N(0,1)
If a statistic is normally distributed, we find a confidence interval for the parameter using
\[ statistic \pm z^* \times SE \]
where the proportion between \(-z^*\) and \(+z^*\) in the standard normal distribution is the desired level of confidence.

Confidence Intervals
Find \(z^*\) for a 99% confidence interval.
www.lock5stat.com/statkey
\[ z^* = 2.575 \]

Proportion of Infected
- Proportion of infected mosquitoes:
  - Sample statistic (from data): 90/201 = 0.448
  - \(z^*\) (from standard normal): 2.575
  - SE (from bootstrap distribution): 0.037
- Give a 99% confidence interval for the proportion of mosquitoes who get infected.

\[ z^* \]
- Why use the standard normal?
- \(z^*\) is always the same, regardless of the data!
- Common confidence levels:
  - 95%: \(z^* = 1.96\) (but 2 is close enough)
  - 90%: \(z^* = 1.645\)
  - 99%: \(z^* = 2.576\)
### Confidence Interval Formula

**IF SAMPLE SIZES ARE LARGE...**

- From $N(0,1)$
- From original data
- From bootstrap distribution

\[
\text{sample statistic} \pm z^* \times SE
\]

### Formula for p-values

**IF SAMPLE SIZES ARE LARGE...**

\[
z = \frac{\text{sample statistic} - \text{null value}}{SE}
\]

Compare $z$ to $N(0,1)$ for p-value

### Standard Error

- Wouldn’t it be nice if we could compute the standard error *without* doing thousands of simulations?
  - We can!!!
  - Or at least we'll be able to next class...

### Malaria and Mosquitoes

- **Should we limit our analysis to only those mosquitoes that actually got infected? Why or why not?**
- **In favor of yes:**
  - We care about whether mosquitoes behave differently after being infected, not just after being exposed to an infection
  - Including mosquitoes that didn’t actually get infected may weaken results
- **In favor of no:**
  - Mosquitoes were not randomized to be infected or not, they were randomized to the possibility of becoming infected.
  - We could have confounding variables and could no longer make conclusions about causality
- **Methods for this, but beyond the scope of this course**

### To Do

- Read Chapter 5
- Do HW 5.2 (due Friday, 10/30)