More one-sample tests

Reporting test results, understanding error types, and paired tests for differences

Today’s agenda

1. Reading quiz [2pm section] [4pm section]
2. More one-sample inference: reporting test results; decision errors
3. Comparing two population means via paired differences
4. Lab: more t-tests in R

From last time

1. What hypotheses were tested?
2. What was the test statistic and p-value?
3. What was the sample size?
4. What is the conclusion of the test?
5. Interpret the confidence interval.

Inference for the body temperature data.

```
data(thermometry)
body_temps <- thermometry$body.temp
t.test(body_temps,
       mu = 98.6,
       alternative = 'two.sided')
```

```

	One Sample t-test

data:  body_temps
t = -5.4548, df = 129, p-value = 2.411e-07
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
 98.12200 98.37646
sample estimates:
mean of x 
 98.24923 
```

Significance conventions

Convention 1: statistical significance

- $p < 0.05$: reject H_0
- $p \geq 0.05$: fail to reject H_0

> "The data provide **significant evidence at level α = 0.05** against the hypothesis that mean body temperature is 98.6 °F in favor of the alternative that mean body temperature differs from 98.6 °F (*T* = -5.4548 on 129 degrees of freedom, *p* = .0000002411)."

Convention 2: weight of evidence against H_0

- $p < 0.01$: strong evidence
- $0.01 \leq p < 0.05$: moderate evidence
- $0.05 \leq p < 0.1$: weak evidence
- $0.1 \leq p$: no evidence

> "The data **provide strong evidence** against the hypothesis that mean body temperature is 98.6 °F in favor of the alternative that mean body temperature differs from 98.6 °F (*T* = -5.4548 on 129 degrees of freedom, *p* = .0000002411)."

You may use either convention to interpret test results.

Choosing alternatives

> Is the mean body temp less than 98.6?

Which test should you use? Consider the interpretations:

a. [lower] *evidence favoring* lower temp
b. [upper] *no evidence against* lower temp

The conclusions are consistent but not equivalent – (a) is a better answer.

Your alternative should be the claim you hope to support with evidence, and your null the claim you hope to refute.

```
t.test(body_temps,
       mu = 98.6,
       alternative = 'less')
```

```

	One Sample t-test

data:  body_temps
t = -5.4548, df = 129, p-value = 1.205e-07
alternative hypothesis: true mean is less than 98.6
95 percent confidence interval:
     -Inf 98.35577
sample estimates:
mean of x 
 98.24923 
```

```
t.test(body_temps,
       mu = 98.6,
       alternative = 'greater')
```

```

	One Sample t-test

data:  body_temps
t = -5.4548, df = 129, p-value = 1
alternative hypothesis: true mean is greater than 98.6
95 percent confidence interval:
 98.14269      Inf
sample estimates:
mean of x 
 98.24923 
```

Reporting test results

To report the result of a hypothesis test, you should:

1. Lead with your conclusion
2. State the hypotheses tested
3. Interpret the conclusion of the test in context
4. Provide the T statistic, degrees of freedom, and p-value
5. State and interpret the point estimate and interval for the mean

Our results suggest mean body temperature is less than 98.6 °F. We tested the null hypothesis that mean body temperature is 98.6 °F or greater against the alternative that mean body temperature is less than 98.6 °F. The data provide sufficiently strong evidence against the hypothesis that mean body temperature is 98.6 °F or greater in favor of the alternative that mean body temperature is less than 98.6 °F (T = -5.4548 on 129 degrees of freedom, p-value = .0000001205). With 95% confidence, the mean nightly hours of sleep is estimated to be at most 98.36 °F, with a point estimate of 98.23 (SE = 0.0643).

Your turn: sleep data

Open `lab7-moretests` in the class workspace.

Work with your group on just *one* of the questions below.

1. Do US adults sleep 7.5 hours per night on average?
2. Do US adults sleep less than 7.5 hours per night on average?
3. Do US adults sleep more than 7.5 hours per night on average?
4. Do US adults sleep more than 6.5 hours per night on average?

Your task is to determine and carry out an appropriate test, and then write a complete report of the test outcome:

- answer the question
- hypotheses tested
- test conclusion, interpreted in context
- test statistic, degrees of freedom, p-value
- confidence interval and point estimate, interpreted in context

These elements should be summarized together in complete sentences.

Tests and intervals

Notice that `t.test` produces a confidence interval.

- for the two-sided test, the interval is *exactly* the one we calculated before
- for the one-sided tests, the interval is a lower/upper confidence bound:

$$\begin{array}{llcl} \bar{x} - c \times SE(\bar{x}) & \text{lower confidence bound} & \Longleftrightarrow & \text{lower-sided test} \\ \bar{x} + c \times SE(\bar{x}) & \text{upper confidence bound} & \Longleftrightarrow & \text{upper-sided test} \end{array}$$

  where c is chosen to ensure a specified coverage level

The tests and intervals correspond in the following sense:

> The level α test rejects just in case the $(1-\alpha)$ confidence interval excludes μ_0

```
# changing confidence level for interval
t.test(body_temps, mu = 98.6, alternative = 'less', conf.level = 0.99)
```

Decision errors

There are two ways to make a mistake in a hypothesis test:

1. reject a true H_0
2. fail to reject a false H_0

These are known as **type I** and **type II** errors.

Because rejecting H_0 is a stronger conclusion, type I errors are considered more severe.

		Test conclusion	
		Fail to reject H_0	Reject H_0 in favor of H_A
Reality	H_0 True	Correct Decision	Type 1 Error
	H_A True	Type 2 Error	Correct Decision

> The significance level of a test is a cap on the type I error rate.

Exploring type I error rates

In `lab7-moretests`, there are codes to draw a sample from a mock population and **test a true null hypothesis**. Run these and take note of your p-value.

In this situation, a type I error is rejecting H_0.

- rule: reject if $p < \alpha$.
- we will tally rejections for various α values

Significance level	Error frequency
$\alpha = 0.2$	
$\alpha = 0.1$	
$\alpha = 0.05$	
$\alpha = 0.02$	

> Takeaway: using larger significance thresholds leads to [more/less] type 1 errors

Exploring type II errors

Now let's **test a false null hypothesis**.

A type II error is failing to reject H_0.

- rule: $p < 0.05$
- use example commands to test each hypothesis at right
- use a two-sided test

Null value μ_0	Error frequency
4.2	
4.6	
4.9	
5.1	
5.4	
5.7	

> Notice that the type II error is quite high for null values near the true mean; this indicates the test has little power to detect such alternatives.

Swimsuit data

Let’s consider our first *two-sample* problem.

> Are swimmers faster in bodysuits than in regular swimsuits?

Below are the first few observations of the average velocity of competitive swimmers in a 1500m; one measurement was taken in a swimsuit, the other in a bodysuit.

swimmer.number	wet.suit.velocity	swim.suit.velocity
1	1.57	1.49
2	1.47	1.37
3	1.42	1.35

Can you formulate a pair of hypotheses to test to answer this question using methods from last time?

Comparing two means

We can formulate the question as a *comparison of two means*:

$$H_0 : \mu_{\text{bodysuit}} \leq \mu_{\text{swimsuit}}$$

$$H_A : \mu_{\text{bodysuit}} > \mu_{\text{swimsuit}}$$

It is common to express hypotheses of this form in terms of a difference in means:

$$H_0 : \underbrace{\mu_{\text{bodysuit}} - \mu_{\text{swimsuit}}}_{\delta} \leq 0$$

$$H_A : \underbrace{\mu_{\text{bodysuit}} - \mu_{\text{swimsuit}}}_{\delta} > 0$$

Pairing

The velocities in the swimsuit dataset are paired because every swimmer is measured in both suits.

Data are **paired** just in case the measurements in each group are taken on exactly the same study units.

> if you can calculate a difference for each study unit, your data are paired; otherwise, your data are not paired

This is the easiest situation to handle, because it reduces to a one-sample problem with the observed differences.

Inference for paired data

1. Calculate paired differences.

swimmer.number	wet.suit.velocity	swim.suit.velocity	velocity.diff
1	1.57	1.49	0.08
2	1.47	1.37	0.1
3	1.42	1.35	0.07

2. Perform test as before.

```r
# test for paired difference
t.test(diffs, mu = 0, alternative = 'greater')
```

```
    One Sample t-test

data:  diffs
t = 12.318, df = 11, p-value = 4.443e-08
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.06620114        Inf
sample estimates:
mean of x
   0.0775
```

3. Report the result of the test. *You try.*