Hypothesis testing

Directional and two-sided inference for a population mean

Today's agenda

1. Reading quiz [2pm section] [4pm section]
2. Hypothesis tests for a population mean
3. Lab: t-tests in R
4. (If time) Exploring decision errors

DDT data

The following are 15 measurements of the pesticide DDT in kale in parts per million (ppm). Each measurement was taken by a different laboratory.

2.79, *2.93*, *3.22*, *3.78*, *3.22*, *3.38*, *3.18*, *3.33*, *3.34*, *3.06*, *3.07*, *3.56*, *3.08*, *4.64* and *3.34*

> C. E. Finsterwalder (1976) Collaborative study of an extension of the Mills et al method for the determination of pesticide residues in food. J. Off. Anal. Chem. 59, 169–171.

Imagine the target level for safety considerations is 3ppm or less, and you want to use this data to determine whether the mean DDT level is within safe limits.

Hypothesis testing

This is an example of a *hypothesis testing* problem: we want to test the hypothesis that mean DDT in kale is within safe limits. Hypothesis testing is another form of statistical inference.

The general pattern for performing a hypothesis test is:

1. Formulate the hypothesis to test in terms of the values of a population parameter.
2. Assess the likelihood of the data under the hypothesis through use of a “test statistic”.
3. Conclude whether the data provide evidence favoring an alternative.

Today we’ll cover each step in turn in the context of tests for a population mean.

1. Formulating hypotheses

Hypotheses cannot be tested in isolation, but must be considered relative to a specified alternative.

To articulate the hypotheses for a test, we need:

- population parameter of interest
- **null hypothesis** H_0: possible value(s) under the claim to be tested
- **alternative hypothesis** H_A: possible value(s) if the claim is found to be false

In the context of the DDT example…

$$H_0 :$$

$$H_A :$$

2. Test statistic

Test statistics are data summaries that:

- depend on the null value of the population parameter
- have a known sampling distribution

For a population mean, we use:

$$T = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

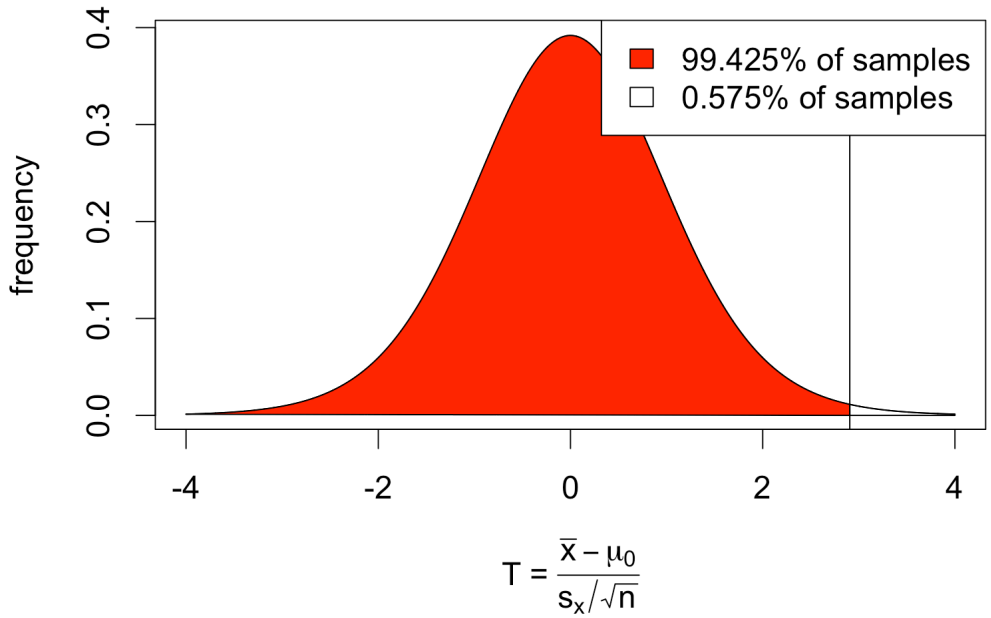
This is well-described by a t_{n-1} model when $\mu = \mu_0$. It is useful for the test because:

- large (absolute) values of T are unlikely if $\mu = \mu_0$
- small (absolute) values of T are expected if $\mu = \mu_0$

3. Drawing a conclusion

In the DDT example, $T = 2.906$. This favors H_A, but by how much?

According to the t model, less than 1% of samples would produce a result *more* favorable to H_A.



Point estimate:

mean	se
3.328	0.1129

Test statistic:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})} =$$

> This is strong evidence *against* the claim that the DDT level is 3ppm or less and *in favor of* the claim that the DDT level exceeds 3ppm.

Recap of DDT example

> Is the mean DDT level in kale 3ppm or less?

Data are measurements of DDT levels in ppm from 15 labs.

Population parameter:

H_0 :

H_A :

$\mu_0 =$

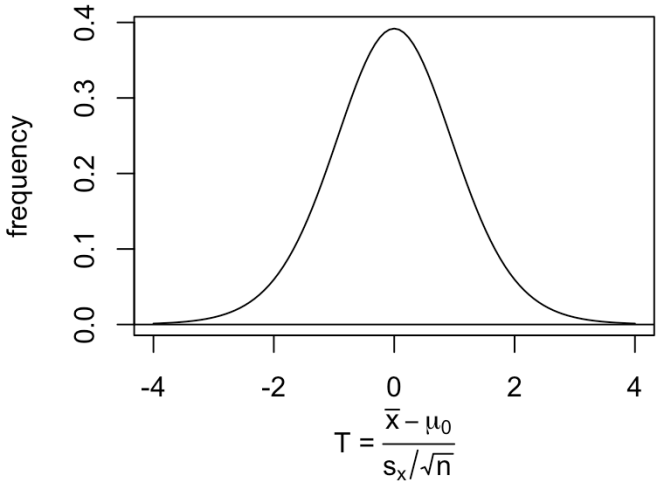
$T =$

% of samples more favorable to H_A $\approx$

Summary statistics:

mean	sd	se
3.328	0.4372	0.1129

t_{14} model:



Another example: sleep

> Does the average U.S. adult sleep at least 7 hours per night?

Data are reported average hours of sleep per night from 135 NHANES respondents.

Population parameter:

H_0 :

H_A :

$\mu_0 =$

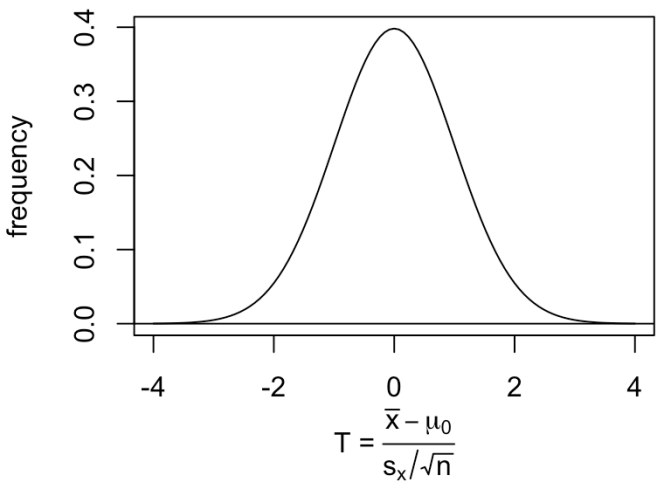
$T =$

% of samples more favorable to H_A $\approx$

Summary statistics:

mean	sd	se
6.896	1.394	0.12

t_{134} model:



Your turn: body temperatures

> Is mean body temperature actually 98.6 °F, or is it lower?

Data are 130 observations of body temperature (°F) derived from a JAMA study.

Population parameter:

H_0 :

H_A :

$\mu_0 =$

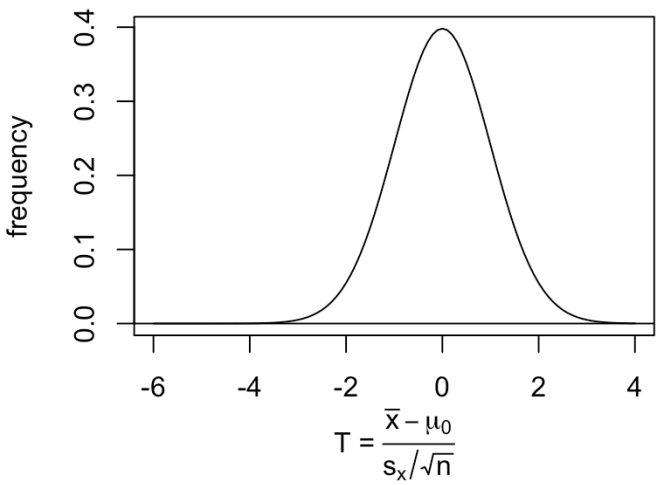
$T =$

% of samples more favorable to H_A $\approx$

Summary statistics:

mean	sd	se
98.25	0.7332	0.0643

t_{129} model:



Strength of evidence

The result that 0.575% of samples would produce a test statistic more strongly favoring the alternative hypothesis is an example of a ***p*-value**:

> the probability under H_0 of obtaining a sample for which the test statistic is at least as favorable to H_A as the value actually observed

In other words, p-values assume the null hypothesis is true, and then ask, "what is the chance I'd obtain data at least as suggestive as what I have that the alternative is more likely than the null?"

- smaller p-values: if H_0 is true, equally or more favorable results are not expected often by chance
- larger p-values: if H_0 is true, equally or more favorable results are expected often by chance

Evidence thresholds

It remains to define an *evidence threshold* above which we decide to reject H_0.

A heuristic is to fix a **significance level** α and reject H_0 whenever $p < \alpha$.

- represents an evidence threshold
- conventionally, $\alpha = 0.05$
- controls error rates

Imagine that indeed mean DDT in kale is 3ppm. Then 0.575% of samples produce test statistics at least as favorable to the alternative as what we saw in the study.

- so if we set the evidence threshold for rejecting H_0 exactly here ($\alpha = 0.00575$) we'll be wrong 0.575% of the time
- if we set the evidence threshold lower (say $\alpha = 0.01$) we'll be wrong more than 0.575% of the time (in fact 1% of the time)

Interpreting results

Because of the anatomy of hypothesis tests, there are two possible findings:

- [above evidence threshold] reject H_0 in favor of H_A
- [below evidence threshold] fail to reject H_0

> Since the null is assumed to be true to perform the test, the test can only result in (a) evidence against this assumption or (b) no evidence against this assumption. But because it's an assumption, we don't affirm it if the test fails.

In the DDT example: *the data provide sufficiently strong evidence (p = 0.00575) to reject the hypothesis that mean DDT in kale is at most 3ppm in favor of the hypothesis that mean DDT in kale exceeds 3ppm*.

Composite hypotheses

The null hypothesis is a *composite* of values, so why did we choose just one ($\mu_0 = 3$) to perform the test?

Any null value farther from the alternative will produce stronger results.

Null value	**Test statistic numerator**	**Proportion of samples more favorable to H_A**
$\mu_0 = 3$	0.328	0.00575
$\mu_0 = 2.99$	0.338	0.00483
$\mu_0 = 2.95$	0.378	0.00239

So by using $\mu_0 = 3$, we are choosing the most conservative null value for the test.

Components of a test

Component	Explanation	DDT example
Population parameter	The quantity of interest	Mean DDT μ
Null hypothesis	The claim to be tested	$\mu \leq 3$
Alternative hypothesis	The alternative claim	$\mu > 3$
Test statistic	A function of the sample data and the null value of the population parameter	$T = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} = 2.91$
Model	Sampling distribution of the test statistic under H_0	$t_{df=14}$ model
p-value	Probability under H_0 of obtaining a result at least as favorable to H_A	0.575% of samples are more favorable to H_A
Decision	Reject or fail to reject H_0	Reject at $\alpha = 0.05$

Performing tests in R

Inputs:

1. data vector
2. null value of parameter
3. alternative hypothesis

Outputs:

4. test statistic
5. degrees of freedom for t model
6. p-value
7. confidence interval
8. point estimate

`t.test` performs all calculations. Locate each input (1-3) and output (4-7) below:

```
1  t.test(ddt, mu = 3, alternative = 'greater')
```

```
    One Sample t-test

data:  ddt
t = 2.9059, df = 14, p-value = 0.005753
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 3.129197      Inf
sample estimates:
mean of x
    3.328
```

Directional hypotheses

Tests for the mean can involve directional or non-directional alternatives. We refer to these as one-sided and two-sided tests, respectively.

Test type	Null	Alternative	Favors alternative
Upper-sided	$\mu \leq \mu_0$	$\mu > \mu_0$	positive T
Lower-sided	$\mu \geq \mu_0$	$\mu < \mu_0$	negative T
Two-sided	$\mu = \mu_0$	$\mu \neq \mu_0$	large $\lvert T \rvert$

The direction of the test affects the p-value calculation (and thus decision), but *won't* change the test statistic.

```
# upper-sided
t.test(ddt, mu = mu_0, alternative = 'greater')

# lower-sided
t.test(ddt, mu = mu_0, alternative = 'less')

# two-sided (default)
t.test(ddt, mu = mu_0, alternative = 'two.sided')
```

Lab: t-tests in R

Open up `lab6-hypotesting` in the class workspace. The goals for this lab are:

1. Learn how to implement t tests in R and interpret output
2. Practice formulating and testing hypotheses from simple research questions
3. (If time) Explore decision errors