Lab 6: Hypothesis testing basics

STAT218

In class we discussed the $t$ test for the hypotheses:

$$\left\{\begin{matrix}H\_{0}:&μ=μ\_{0}\\H\_{A}:&μ\ne μ\_{0}\end{matrix}\right.$$

The objective of this lab is to learn to perform the basic calculations involved in this $t$ test “by hand” (without the use of the function that we’ll apply later):

* calculating the test statistic $T=\frac{‾−μ\_{0}}{SE\left(‾\right)}$
* calculating a critical value for a level $α$ test
* calculating a $p$-value

We’ll use the temps dataset to illustrate.

library(tidyverse)
load('data/temps.RData')
head(temps)

 body.temp sex heart.rate
1 98.8 female 69
2 98.6 female 85
3 98.4 male 68
4 97.2 female 66
5 99.5 male 75
6 97.1 male 82

### Checking test assumptions

The $t$ test only makes sense for unimodal populations; otherwise, the population mean isn’t an interpretable parameter.

Beyond that, the test is based on the assumption that either the underlying population distribution is symmetric or the sample size is not too small. “Too small” is relative to just how much funny business you see in the distribution of values: more pronounced skewness or outliers means that more data are required for the test to work well.[[1]](#footnote-20)

Focusing on the mean body temperature as the parameter of interest, we’ll start by checking data summaries of the body.temp variable to assess the properties mentioned above: unimodality and subsequently symmetry and presence of outliers. Here, we have 39 observations, which is neither small nor large but modest. So, we don’t need to be *too* sensitive unless strong skewness or extreme outliers are present.

# extract temperature variable
bodytemp <- temps$body.temp

# location measures
summary(bodytemp)

 Min. 1st Qu. Median Mean 3rd Qu. Max.
 96.40 97.85 98.40 98.41 99.00 100.80

# histogram
hist(bodytemp, breaks = 10)



The histogram suggests a unimodal population, so conceptually the test is sensible; moreover, there’s no indication of strong skewness or outliers, so the test should work just fine.

|  |
| --- |
|  Your turn 1 |
| Check the heart.rate variable for any skewness or outliers. Assess whether the $t$ test is appropriate, taking account of the sample size.# extract heartrate variable# location measures# histogram |

### Calculating the test statistic

To test the hypotheses:

$$\left\{\begin{matrix}H\_{0}:&μ=μ\_{0}\\H\_{A}:&μ\ne μ\_{0}\end{matrix}\right.$$

We use the test statistic:

$$T=\frac{‾−μ\_{0}}{SE\left(‾\right)}$$

In these expressions, $μ\_{0}$ is a placeholder for the hypothesized value of $μ$ in any particular problem. To test whether the mean body temperature is 98.6°F, set $μ\_{0}=98.6$ and follow the formula:

# store sample mean and standard error
bodytemp.mean <- mean(bodytemp)
bodytemp.mean.se <- sd(bodytemp)/sqrt(length(bodytemp))

# calculate t statistic
bodytemp.tstat <- (bodytemp.mean - 98.6)/bodytemp.mean.se
bodytemp.tstat

[1] -1.328265

Notice that this matches the value obtained in class. The interpretation is:

If mean body temperature were 98.6°F, the estimation error would be -1.328 standard errors.

|  |
| --- |
|  Your turn 2 |
| Calculate the test statistic to test the hypothesis that mean heart rate is 75 beats per minute (bpm).# store sample mean and standard error for heart rate# calculate t statistic to test whether the mean is 75bpm |

### Critical values

One way to make a test decision is to determine a threshold $q$ for the absolute value of $T$:

* for any larger $\left|T\right|>q$, you’ll reject the null hypothesis in favor of the alternative
* for any smaller value $\left|T\right|\leq q$, you’ll fail to reject the null hypothesis in favor of the alternative

The threshold is chosen to control the error rate $α$: how often you make the mistake of rejecting $H\_{0}$ when it’s true. To do this, we use the $1−\frac{α}{2}$ quantile of the $t\_{n−1}$ model, *i.e.*, the value such that:

$$P\left(T<q\right)=1−\frac{α}{2}$$

We use this specific quantile so that $P\left(\left|T\right|>q\right)=α$, meaning, if $H\_{0}$ is true, the proportion of samples for which $T$ exceeds the decision threshold – and thus an error is made – is $α$.

Several examples follow to help you get the hang of determining the correct inputs to obtain critical values based on controlling the error rate at $α$:

# to control error rate at 10%, use this critical value
qt(0.95, df = 38)

[1] 1.685954

# to control error rate at 5%, use this critical value
qt(0.975, df = 38)

[1] 2.024394

# to control error rate at 1%, use this critical value
qt(0.995, df = 38)

[1] 2.711558

|  |
| --- |
|  Your turn 3 |
| Calculate the critical value you’d use to control error rate at 20%.# to control error rate at 20%, use this critical value |

The error rate $α$ is also called a “significance level”, leading to the following interpretations:

* at the 20% significance level, we reject the null hypothesis $μ=98.6$
* at the 10% significance level, we fail to reject the null hypothesis $μ=98.6$
* at the 5% significance level, we fail to reject the null hypothesis $μ=98.6$
* at the 1% significance level, we fail to reject the null hypothesis $μ=98.6$

You’ll notice that the thresholds get more stringent as the significance level decreases; so we’ll reject for every significance level below 10%.

A 5% significance level is conventional. We’d report the test result as follows:

At the 5% significance level, the data do not provide evidence that mean body temperature differs from 98.6°F.

|  |
| --- |
|  Your turn 4 |
| Test the hypothesis that mean heart rate is 75 bpm at the 5% significance level and report the test result following the language above.# to control error rate at 5%, use this critical value# decision whether mean heart rate is 75bpm? |

### $p$-values

A $p$-value quantifies exactly how unusual a particular $T$ statistic is. Technically, it’s the proportion $p$ of samples for which $T$ exceeds the observed value; you can also think of it as the *minimum significance level at which the test rejects the null hypothesis*.

To calculate a $p$-value, we compute:

$$p=2×P\left(T>\left|T\_{observed}\right|\right)$$

In R, the $p$-value for the test of mean body temperature is:

2\*pt(abs(bodytemp.tstat), df = 38, lower.tail = F)

[1] 0.1920133

The interpretation of this quantity:

If mean body temperature is in fact 98.6°F then 19.2% of samples would produce a $T$ statistic as large or larger than the observed value.

|  |
| --- |
|  Your turn 5 |
| Compute the $p$-value for the test of whether mean heart rate is 75 bpm. Interpret the value in context.# p value for test of whether mean heart rate is 75bpm |

### Reporting test results

The conventional style for reporting a test is to include a statement of the outcome, interpreted in context, with supporting statistics provided parenthetically:

The data do not provide evidence that the mean body temperature differs from 98.6°F (*T* = -1.328 on 38 degrees of freedom, *p* = 0.192).

|  |
| --- |
|  Your turn 6 |
| Write a report of your test of whether mean heart rate is 75bpm following the convention above. |

### Practice problems

1. Use the nhanes dataset to test the hypothesis that the average U.S. adult gets 8 hours of sleep per night.
	1. Produce a histogram of the observations and assess, considering the sample size, whether the $t$ test is appropriate based on its shape and the presence or absence of outliers.
	2. Calculate a point estimate and standard error for the mean nightly hours of sleep among U.S. adults.
	3. Write the hypotheses to test in notation.
	4. Calculate the test statistic, critical value for a 5% significance level test, and $p$-value.
	5. Report the test result following the conventional style introduced in class.
	6. Calculate and interpret a 95% confidence interval for the mean nightly hours of sleep among U.S. adults.
1. I usually think in the following terms for the $t$-test:

	* small: $n\leq 20$
	* modest: $20<n\leq 50$
	* large: $50<n$Unless I’m in the ‘small’ regime, I’m not too worried about skew or outliers. In the ‘modest’ regime, I’m not concerned unless I spot very pronounced skew or outliers. In the ‘large’ regime, I’m really only concerned about (strong) multimodality. Interestingly, in the latter case, the $t$ test still works for multimodal populations, but the population mean isn’t meaningful. [↑](#footnote-ref-20)