Some Common Commands

Command	Explanation	Abbreviation
use	opens a *.dta* file	
import delimited	imports a *.csv* file	
import excel	imports a *.xls* file	
mean x	gives mean, *se*, confidence interval for x	
ttest	executes a $T(n-1)$ hypothesis test	
display	displays scalars and scalar functions	di

Examples

```
use swedishfish.dta
import delimited mikeandike.csv
import excel sourpatchkids.xls
ttest x = 3
display ttail(552, 2)
```

Midterm 1, Winter 2014, Problem 2 (Long-Winded)

Confidence Intervals Explained

A 95% confidence interval is an interval constructed from a random sample in such a way that 95% of such intervals will contain the population mean, μ. (It is *not* correct to say that there is a 95% chance that the population mean lies within the interval. Subtle difference.)

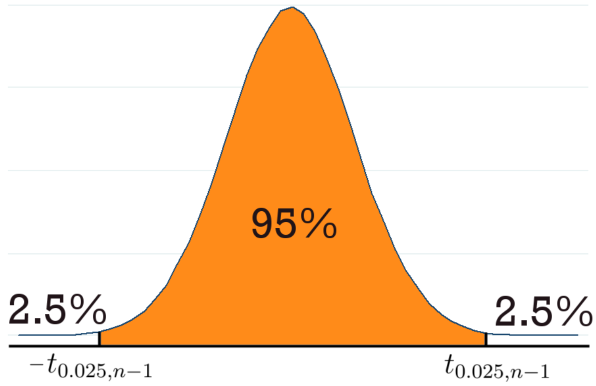
Ultimately we are trying to construct some values A and B, which depend on our data, such that

$$Pr\left(A \leq \mu \leq B\right) = 0.95.$$

One way to approach this is to standardize. We know it is approximately true (and sometimes exactly true – know when!) that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n-1).$$

Thus, we know there is a 95% probability that anything drawn from this distribution lies within the interval $[-t_{n-1,0.025}, t_{n-1,0.025}]$.



Hence we can write

$$\begin{aligned}
0.95 &= Pr\left(-t_{n-1,0.025} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1,0.025}\right)\\
&= Pr\left(-t_{n-1,0.025} \times \frac{S}{\sqrt{N}} \leq \bar{X}-\mu \leq t_{n-1.0.025} \times \frac{S}{\sqrt{N}}\right)\\
&= Pr\left(-\bar{X}-t_{n-1,0.025} \times \frac{S}{\sqrt{N}} \leq -\mu \leq -\bar{X}+t_{n-1,0.025} \times \frac{S}{\sqrt{N}}\right)\\
&= Pr\left(\bar{X}+t_{n-1,0.025} \times \frac{S}{\sqrt{N}} \geq \mu \geq \bar{X}-t_{n-1,0.025} \times \frac{S}{\sqrt{N}}\right).
\end{aligned}$$

The first step multiplied all sides by $S/\sqrt{n}$. The second step subtracted $\bar{X}$ from all sides. The third step multiplied all sides by -1 to turn the μ term positive.

So we have constructed the 95% confidence interval μ,

$$\left[\bar{X}-t_{n-1,0.025} \times \frac{S}{\sqrt{n}}, \bar{X}+t_{n-1,0.025} \times \frac{S}{\sqrt{n}}\right].$$

That's the formula to use. The Stata command for $t_{n-1,0.025}$ is `invttail(n-1, 0.025)`.

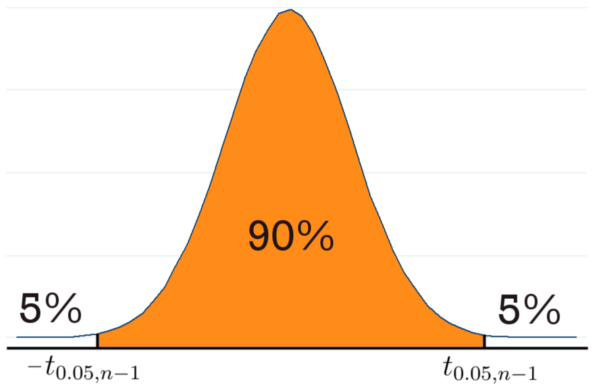
Part (a)

We are given a mean of $\bar{x} = 1.42$, a standard deviation of $s = 35.52$, and $N = 274$ observations. For a 95% confidence interval, we want to find `invttail(273, 0.025)`, which is given to us as $t_{273,.025} = 1.97$. Thus the 95% confidence interval is

$$\left[1.42 - 1.97 \times \frac{35.52}{\sqrt{274}}, 1.42 + 1.97 \times \frac{35.52}{\sqrt{274}}\right] \quad = \quad [-2.81, 5.65].$$

Part (b)

Now that we are doing a 90% confidence interval, it means we are instead considering a picture like



The only practical difference is that we're using different numbers for the tails, `invttail(273, 0.05)` which is given to us as $t_{273,.05} = 1.65$. Thus the 90% confidence interval is

$$\left[1.42 - 1.65 \times \frac{35.52}{\sqrt{274}}, 1.42 + 1.65 \times \frac{35.52}{\sqrt{274}}\right] \quad = \quad (-2.12, 4.96).$$

Notice that less confidence gives a smaller interval. Think back to the interpretation of a confidence interval: 90% means that a smaller percentage of our constructed intervals will actually contain μ, so it makes sense that the corresponding interval is a tighter one.

Hypothesis Testing Explained

Suppose we have some guess about what the population mean μ is. If it's a good guess, then intuitively it should be "close" to the sample mean $\bar{x}$. Hypothesis testing is a way of formalizing "closeness."

We start with a **null hypotheses**. This is our guess for what μ is. Let μ_0 be that guess. We express the null hypothesis as

$$H_0 : \mu = \mu_0.$$

In English: my null hypothesis H_0 is that the population mean μ equals my guess μ_0.

We need to test the null hypothesis against something – we call this the **alternative hypothesis**. The simplest case is that our guess is wrong, which we express as

$$H_1 : \mu \neq \mu_0.$$

Here is how the test proceeds in intuitive terms. We assume that our guess is true. Then we compute a difference between our guess and the sample mean. If we've made a good guess, then the difference should be nearly zero. If the difference is big, then our guess must have been bad, so we reject our guess.

Now let's carry the test out. The way to quantify closeness is with the expression

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

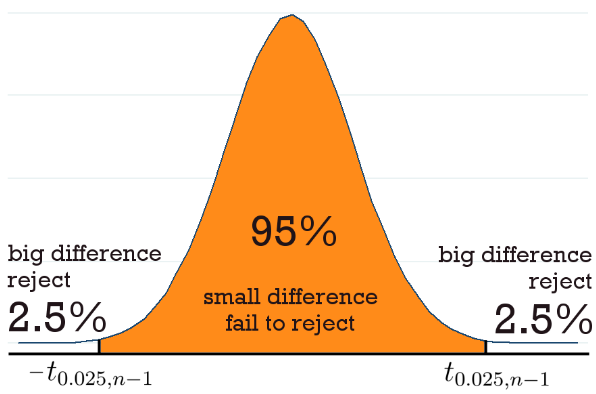
where the number T is referred to as a ***T* statistic**, or more generally a **test statistic**. If the null hypothesis is true, then the T statistic is $T(n-1)$ distributed. By definition, we know that 95% of the draws from a $T(n-1)$ distribution will fall within the interval

$$[-t_{n-1,0.025}, t_{n-1,0.025}].$$

Numbers $-t_{n-1,0.025}$ and $t_{n-1,0.025}$ are called **critical values**. If the test statistic falls beyond the critical values – in the **rejection region** – then we reject. Such is our **rejection rule**.

In English: If my guess is true, then there's a 95% chance that my guess should lie within this interval. But it doesn't lie within this interval. There's only a 5% chance of that actually happening, which is pretty unlikely. So my guess is probably bad.

If the guess does lie within the interval, then we *fail to reject the null hypothesis* at significance level 0.05. (We never say "we accept" or "we confirm" the null hypothesis – statistics is about falsification, not confirmation. See: Karl Popper.)



Here's another way to think about it. We're interested in the closeness of our guess to the sample mean. We can use absolute value as the "distance" between the two. If the distance is too big, then we reject the null. Then we can simplify and reject if $|T| > t_{n-1,0.025}$.

Part (c)

The null hypothesis is $H_0 : \mu = 0$. Thus we are working with the test statistic

$$\frac{1.42 - 0}{35.52/\sqrt{274}} = 0.66.$$

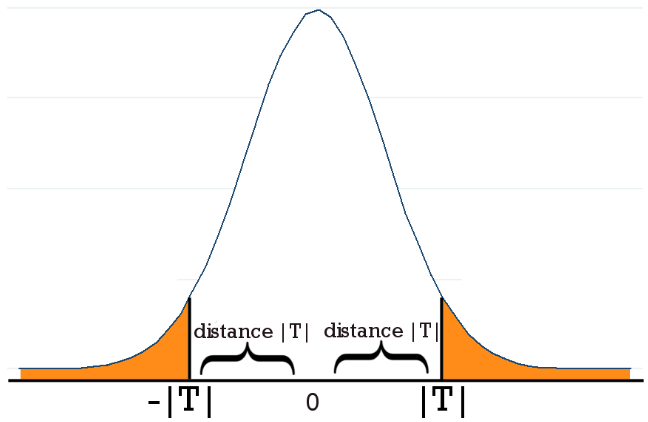
The critical values are given by $\pm$ `invttail(273,0.025)`, given as $t_{n-1,0.025} = 1.969$. Clearly 0.66 lies within the interval $(-1.969, 1.969)$, so we fail to reject the null that $\mu = 0$.

Alternatively, $|0.66| \leq 1.969$, so we fail to reject.

p-value Explained

The p-value tells you the probability of observing a number more extreme in magnitude than the T statistic, supposing that the null hypothesis is true.

Suppose you calculate your T statistic and find that $|T| = 1$. What is the probability of getting a random $T(n-1)$ draw that is greater than $|T| = 1$ in absolute value? It's the probability of drawing less than $-|1|$ plus the probability of drawing greater than $|1|$. In pictures, it's the probability of being in the orange region below:



Or to put it in the maths,

$$\begin{aligned} p &= P(T_{n-1} < -|T|) + P(T_{n-1} > |T|) \\ &= P(-T_{n-1} > |T|) + P(T_{n-1} > |T|) \\ &= P(|T_{n-1}| > |T|). \end{aligned}$$

In Stata, the p-value can be calculated with the command `2 * ttail(n-1, abs(T))`.

Part (d)

What is the probability of observing a $T(273)$ draw that is greater in magnitude than $|0.66|$? Formulate this mathematically as

$$\begin{aligned} p &= P(T_{n-1} < -|0.66|) + P(T_{n-1} > |0.66|) \\ &= P(-T_{n-1} > |0.66|) + P(T_{n-1} > |0.66|) \\ &= P(|T_{n-1}| > |0.66|). \end{aligned}$$

Turns out the answer is $p = 2$ * `ttail(273, abs(0.66))` $= .51$.

Note that a p-value less than 0.05 means there's a less than 5% chance of observing our statistic if the null is true – according to our rejection rule, this means we'd reject our null hypothesis.