Math 183
Statistical Methods

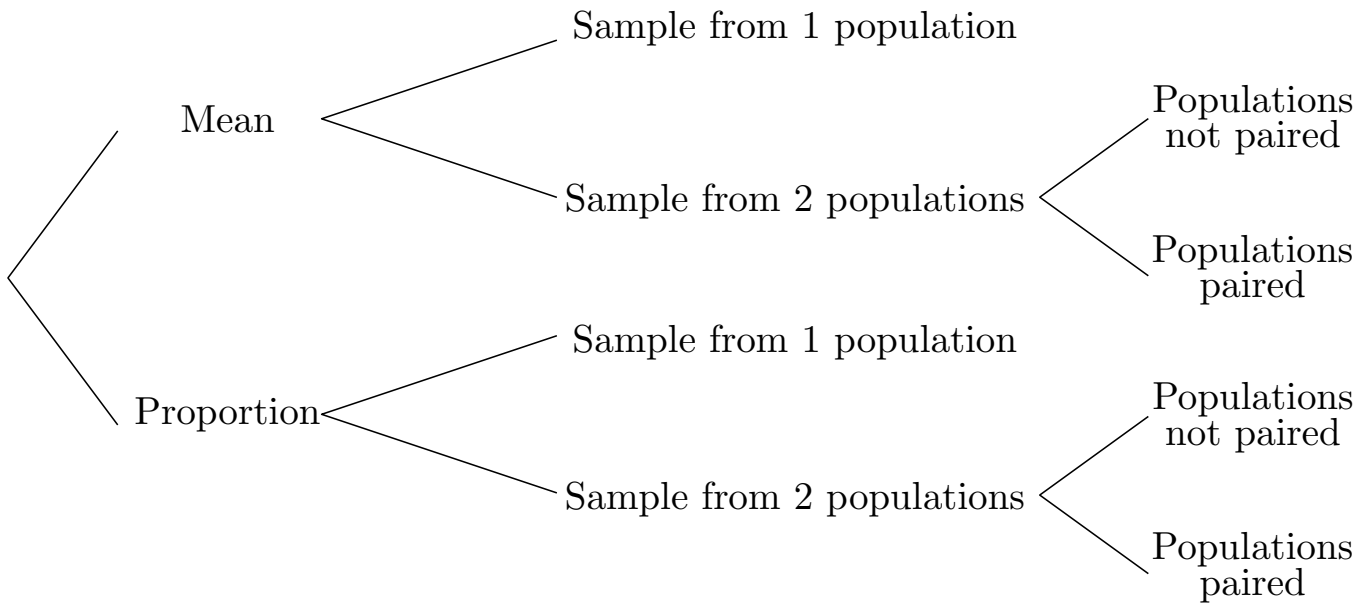
Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 5 (beginning)

- Using the T-distribution instead of the normal model in inference problems
- Realize what conditions are truly needed to use the normal model and accept that we never have them in practice
- Practice using a T-table

Statistics in the Large



We've seen that the sample distribution for the mean from a single population is well approximated by $N(\mu, \sigma/\sqrt{n})$.
Haven't we already done the top branch?

Z vs T

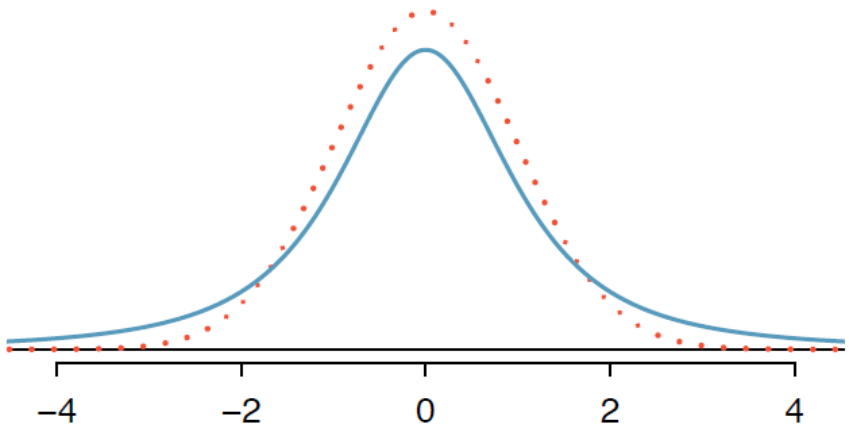
With

- Small sample size ($n < 30$),

or

- When you don't know σ (and must approximate it using s_x),

there is a better approximation of the sampling distribution than the Normal model



—: T-distribution

···: Normal (Z) distribution

The T-distribution is unimodal and symmetric.
The tails of the T-distribution are thicker, and this changes the areas under the curve
(and hence p-values).

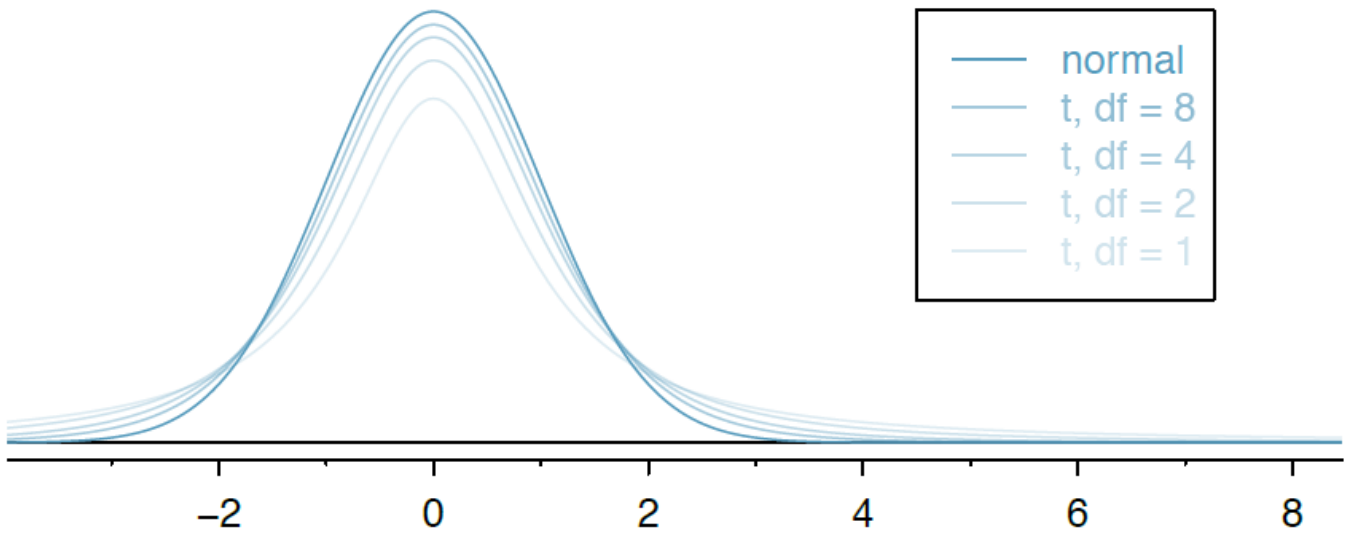
When should you use the Normal distribution from now on?

If you know σ or If n is huge

In other cases, use a T-distribution.

From T to Z

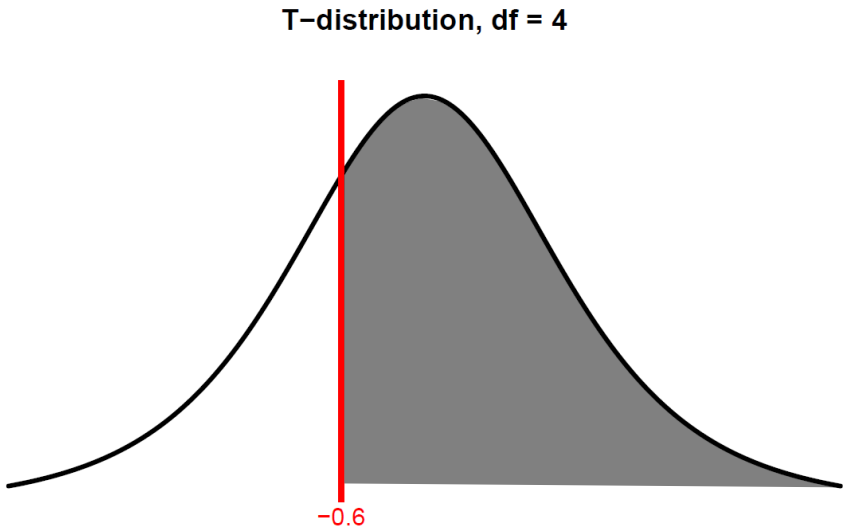
In practice, we tend not to use the Normal curve as the approximation of the sampling distribution because the T-distribution gives us more precise results.



There are many curves in the T-distribution family. You choose the appropriate one based on the size of your sample.
With n data points in the sample, you use the T-distribution with $df = n - 1$.

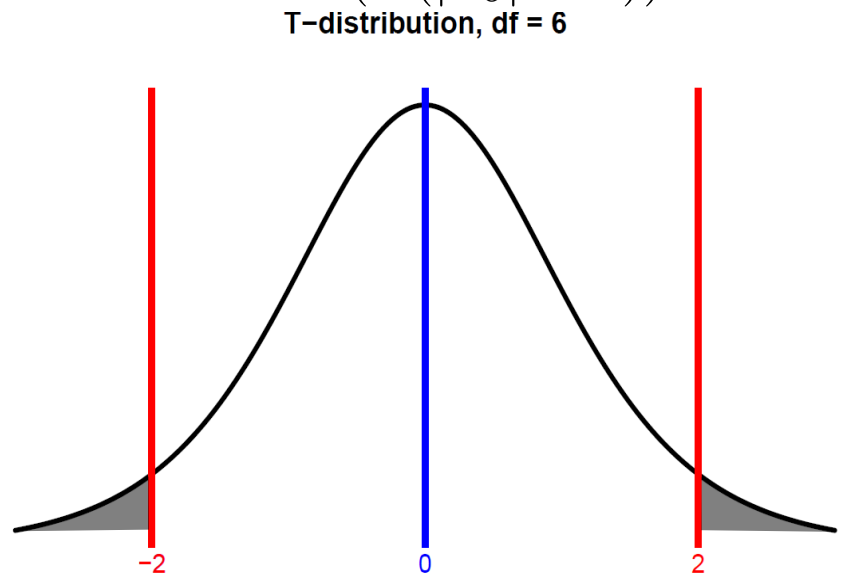
df = "degrees of freedom", and as it gets bigger, the T-distribution morphs into a standard Normal distribution $N(0, 1)$.

How much area is to the right of -0.6 on t_4?




```
> pt(-0.6, df=4, lower.tail=F)
[1] 0.7095794
```

What percent of the t_6-distribution is more than 2 units from 0? ($P(|T_6| > 2)$)




```
> 2*pt(-2, df = 6)
[1] 0.09242631
```

`pt` works like `pnorm`.

T-distributions are always centered at 0. Treat the T-distributions like the standard normal.

There is nothing like `pt(2, df = 6, mean = 4, sd = 3)`.

You'll have to translate your situation to the "standard" T-distribution.

T-Tables?

Since they don't want to print a new table for every possible *df* value, they print one table, but it is not as good as a z-table (standard normal) Also, a T-table is usually reversed (areas on outside, critical values in the table).

Let's find $P(|T_6| > 2)$ using tables only.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50

Here, we find

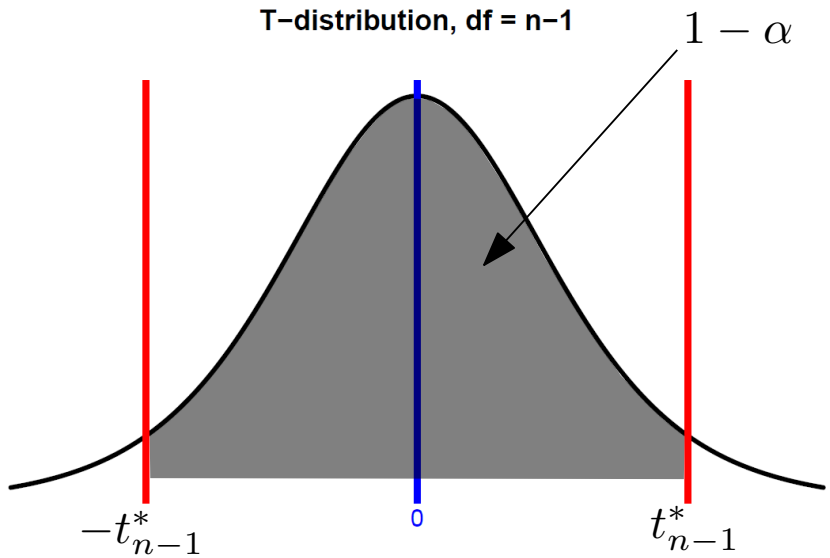
$$0.05 \leq P(|T_6| > 2) \leq 0.10,$$

(R gave us the exact value 0.092)

Improved C.I.'s for Means

For means, our C.I.'s will use the same setup as before, but we live on a T-distribution, not a Z-distribution!

$$\bar{x} \pm t^*_{n-1} \times SE_{\bar{x}}$$



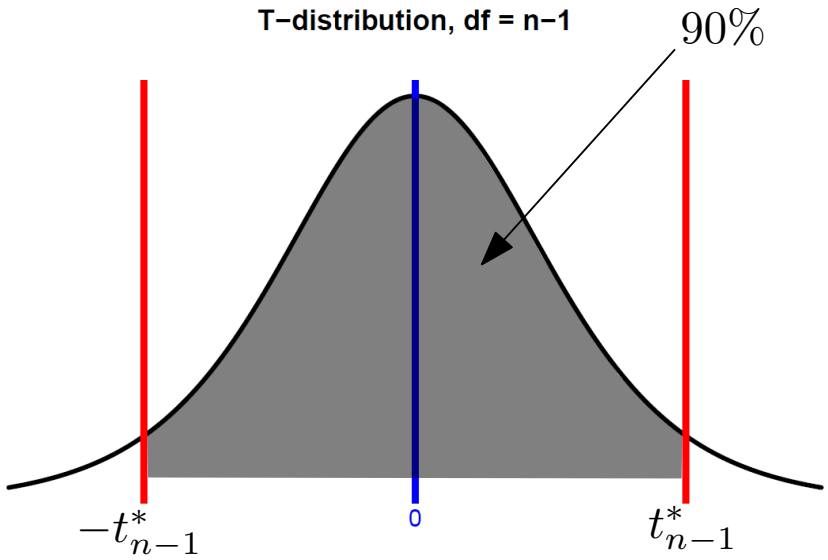
In a sample with n data points, the best model for the sampling distribution is t_{n-1}. Here, we have

$$SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}.$$

First C.I. With the T-Distribution

On average, how much do U.S. baby girls weigh? To find out, you sneak into random hospitals and collect illegally the weight of 12 random newborn babies. If $\bar{x} = 7.3$ lbs and $s_x = 2$, find a 90% C.I. for μ, the average weight of all U.S. female babies.

1) Decide on the model and find the critical value for your confidence level.
Since $n = 12$, our sampling distribution is modeled by $t_{n-1} = t_{11}$.



one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05

```
> qt(0.95, df=11)
[1] 1.795885
```

2) Break out the C.I. formula.

$$\bar{x} \pm t^*_{n-1} \times SE_{\bar{x}} \simeq 7.3 \pm 1.8 \times \frac{2}{\sqrt{12}}$$
$$= (6.26, 8.34).$$

3) Interpret in plain English.

We are 90% confident that μ is between 6.26 and 8.34 pounds.

Remark: The actual value for μ is 7.5 pounds. Our C.I. was one of the lucky 90%.

Your Turn!

You decide to build an 80% C.I. for some mean you are estimating. What is the critical value if your sample has size 15?

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81

- 1.34
- 1.35
- 1.75
- 1.76

Answer: $t^*_{n-1} = t^*_{14} = 1.35$.

Your Turn!

About how much area is to the left of -2 on t_{15}?

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11
	12	1.36	1.78	2.18	2.68	3.05
	13	1.35	1.77	2.16	2.65	3.01
	14	1.35	1.76	2.14	2.62	2.98
	15	1.34	1.75	2.13	2.60	2.95
	16	1.34	1.75	2.12	2.58	2.92
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81

- Between 0.1 and 0.2
- Between 0.05 and 0.1
- Between 0.025 and 0.05
- Between 0.001 and 0.025

Answer: Between 0.025 and 0.05

Your Turn!

Suppose we want to do inference on a mean using a T-distribution. When does a T-distribution actually model the sample distribution?

1. It always does. It doesn't matter what the sample size is or what the population distribution looks like.
2. We only need the data points in our sample to be independent
3. If the data in the sample were chosen at random and are <10% of the population
4. We need randomization, <10 % population, and the population to be reasonably normal looking (skew is ok with a larger n).

Answer: 4.

Hypothesis Testing with T-Distributions

15 years ago, the average finishing time of a Marathon was 3.683 hours. You are curious if runners are getting faster or slower in the current year.

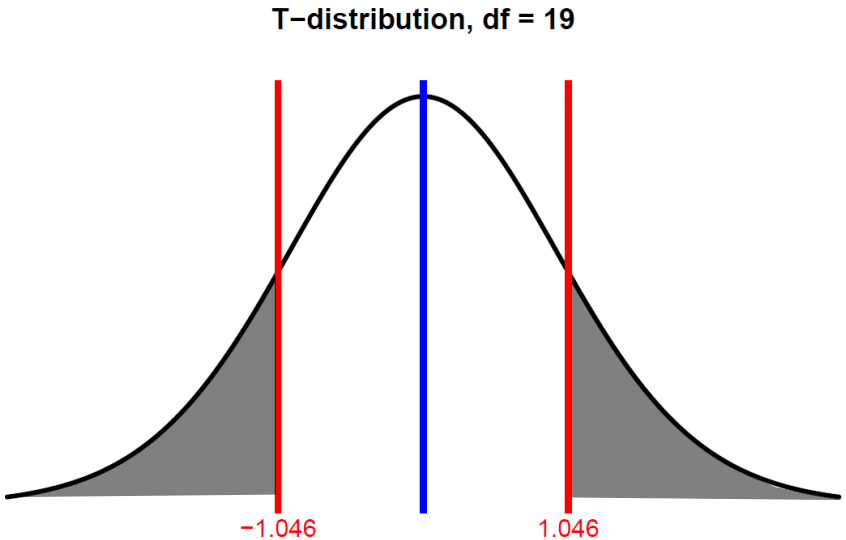
Let μ be the average time of finishers in the current year.

$$H_0 : \mu = 3.683 \; ; \; H_A : \mu \neq 3.683$$

You collect data for 20 random runners in the current year, and get $\bar{x} = 3.8$ with $s_x = 0.5$. Run a hypothesis test with $\alpha = 0.05$.

You should always convert to the standard T-distribution. To do so, you'll need the T-score:

$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{3.8 - 3.683}{0.5/\sqrt{20}} \simeq 1.046.$$




```
> 2*pt(-1.046, df= 19)
[1] 0.3086866
```

Since $p = 0.31 > 0.05$, we do not reject H_0. Our data are not strong enough to show a difference (if there is one).

Remark: With T-tables you only get $p > 0.2$, which is still sufficient to conclude.

Beetle Study

We study beetle biodiversity in a pasture. For this, we collect a biodiversity index (Steinhaus index) in 12 parcels and get the following data:

```
> data
 [1] 0.249 0.291 0.291 0.134 0.194 0.157 0.310 0.222 0.160
     0.363 0.180 0.456
> mean(data)
[1] 0.2505833
> sd(data)
[1] 0.09591138
```

An environmental engineer, specialist of beetle populations, tells you than an average biodiversity index lower than 0.49 in the pasture would be worrying.

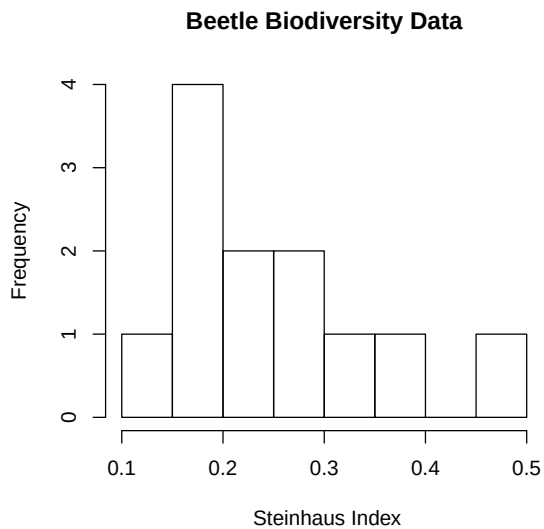
Build a test with level of confidence $\alpha = 5\%$ for determining if the state of the pasture is worrying.

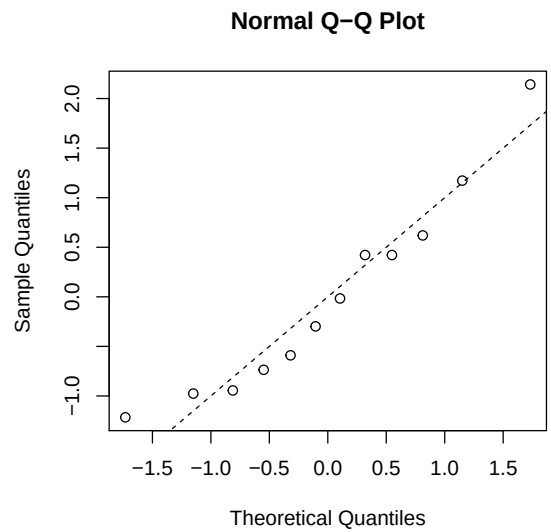
Beetle Study

1) We build hypotheses for the situation

$$H_0\colon\ \mu = 0.49,$$
$$H_A\colon\ \mu < 0.49.$$

2) We want to build a hypothesis test using the T-distribution, so we have to check the normality of our population distribution.





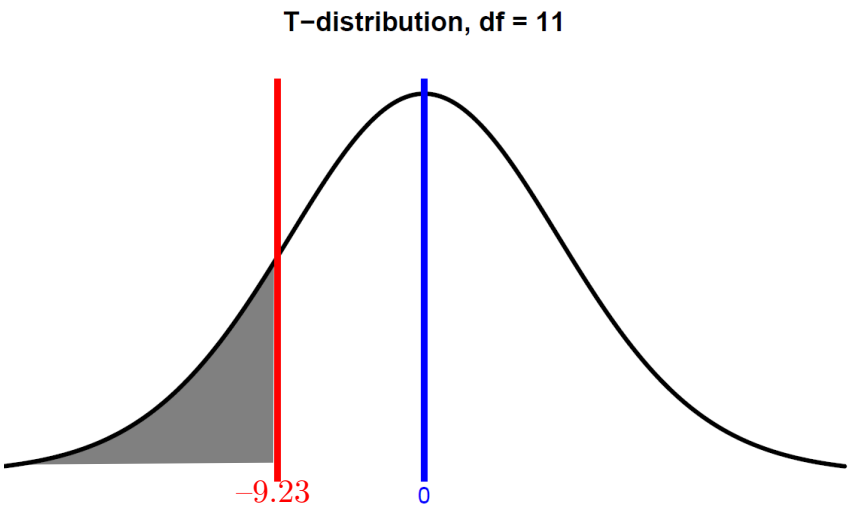
Remark: Normal quantile-quantile plot in R: `qqnorm`.

Beetle Study

3) T-score your data

$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{0.25 - 0.49}{0.09/\sqrt{12}} \simeq -9.23.$$

4) Compute the p-value




```
> pt(-9.23, df= 11)
 [1] 8.182258e-07
```

Since $p \simeq 8 \cdot 10^{-7} \ll 0.05$, we reject H_0 and favor H_A.

There is (a very) strong evidence that the true average biodiversity index is smaller than 0.49.