Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

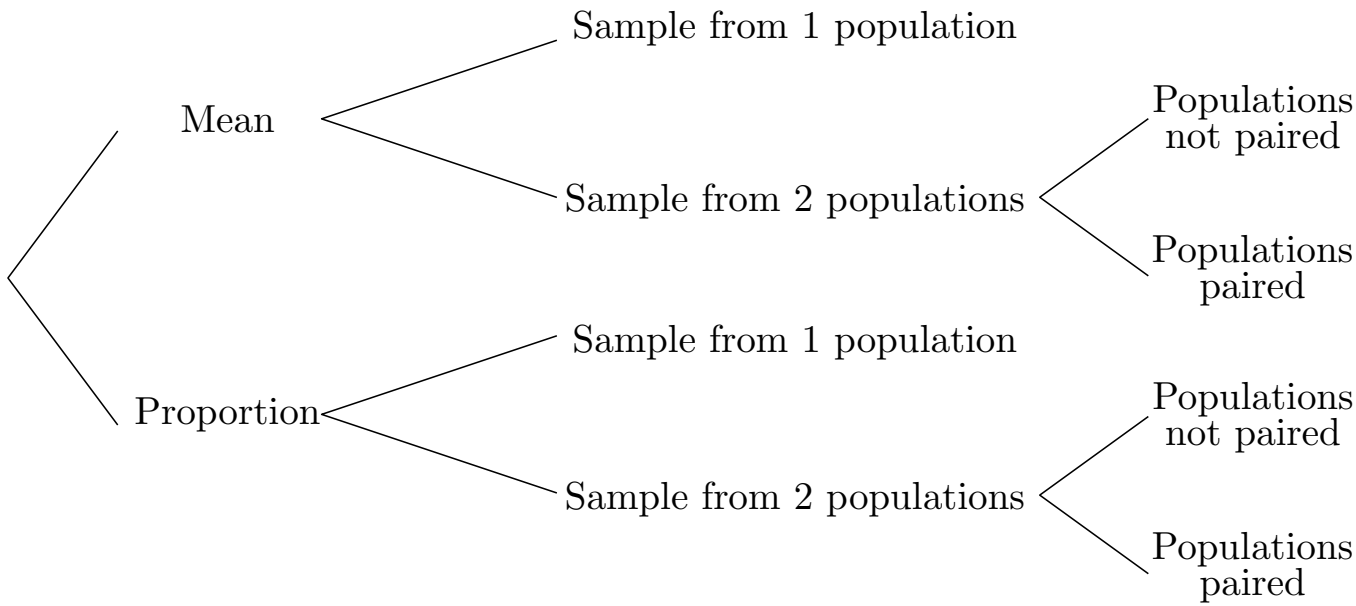
eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 5 (continued)

- Differentiate paired and unpaired data in two-sample setups
- Understand how paired data reduce to a one-sample analysis on differences
- Shape, center, and spread of sampling distribution for difference of means
- Check conditions for doing inference on two-sample problems

Statistics in the Large

Where we stand: We know how to build C.I.'s and run hypothesis tests for one sample means. So we can build a range of plausible values for a single parameter, or compare a single parameter to a known value



Today: Extend these ideas to two parameters (two populations)
Later: Extend these ideas to three or more parameters

Nice thing: The approach we use (largely) remains the same.

Examples of Two-Populations Problems

- Average SAT score in men VS women at UCSD
- Average height of aliens on planets X and Y
- Average age of husbands and wives
- Average income of children compared to their parents

Something should sound different about these examples...

Are Your Two Populations Really Independent?

Two extremes:

- Knowing info about members of one population gives no helpful info about members in the other population
  (Independent samples, 2-sample T-test)
- The members of the two populations have some direct link where each member of one population is paired with a member of the other
  (Paired samples, 1-sample T-test)

Pre-weight	Post-weight	Difference
171	168	-3
203	204	1
130	135	5
⋮	⋮	⋮

Husband Age	Wife Age	Difference
24	22	2
37	40	-3
81	72	8
⋮	⋮	⋮

To analyze paired data, just do analysis on the differences!

C.I. for the Mean Difference of Paired Samples

You decide to research global warming in the U.S. You choose 62 random cities and look up the high temperature on Jan 1st, 1970 and Jan 1st 2017.

Clearly, the data are paired: hot locations will have high readings in both 1970 and 2016. Cold locations will have low readings both times.

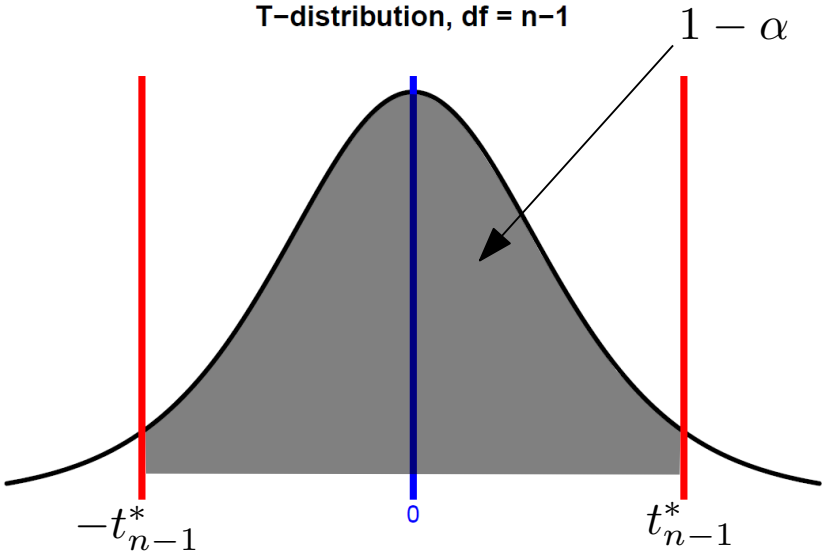
You calculate

$$d = temp_{2016} - temp_{1970}$$

for each location, and find the differences d have $\bar{d} = 1.1°F$ with $s_d = 4.9°F$.

Our differences will follow a T-distribution with $df = 62 - 1 = 61$.

We must calculate $\bar{d} \pm t^*_{61} \times SE_d$.



We get $t^*_{61} \simeq 2$.

```
> qt(0.975, df = 61)
[1] 1.999624
```

$$SE_{\bar{d}} = \frac{s_{\bar{d}}}{\sqrt{n}} = \frac{4.9}{\sqrt{62}} \simeq 0.622.$$

Thus, we have

$$1.1 \pm 2 \times 0.622 = (-0.144, 2.344).$$

We are 95% confident that temperature rose, on average (at the same location), between $-0.144°F$ and $2.344°F$ in the U.S. between Jan 1st, 1970 and Jan 1st, 2017.

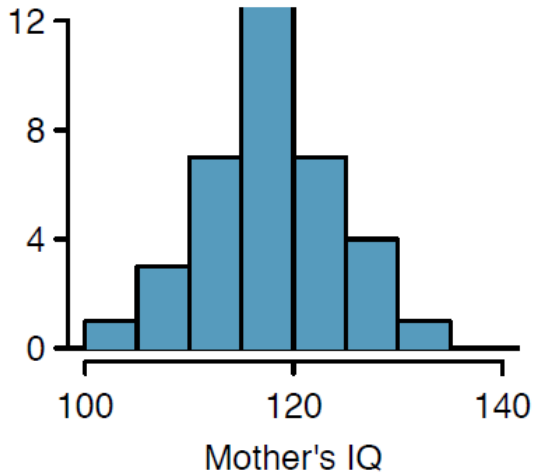
Wait! What About the Conditions We Must Check?

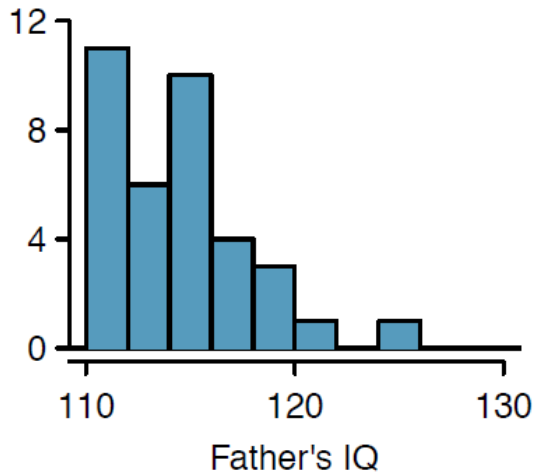
Since two-sample paired data reduce to a 1-sample T-interval (or T-test) on the **differences**, we must simply check our usual conditions on the **differences** (which are the sample undergoing T-testing).

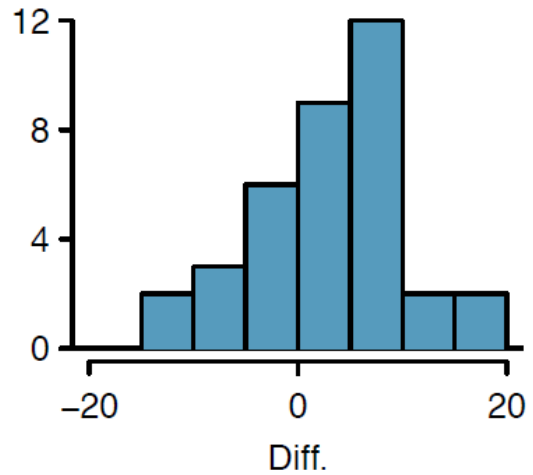
- Independence: The **differences** must be independent of one another. Since the differences are tied to the same location/person/couple, we just need those paired units to be independent of one another. This is usually checked via the Randomization Condition and the $< 10\%$ Condition.
- Nearly Normal Condition: The **differences** must look nearly normal. As n gets larger, you can weaken this condition.

IQ of Parents of Gifted Children

Researchers collected IQ data on parents of 36 children identified as "gifted". Below are the results and histogram of the IQ differences of the parnts.
Run a test to see if mothers and fathers of gifted children have different average IQ's.

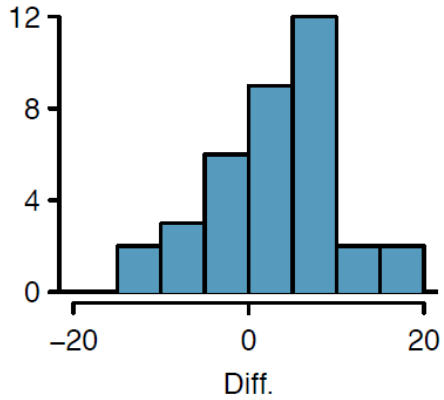






	Mother	Father	Diff.
Mean	118.2	114.8	3.4
SD	6.5	3.5	7.5
n	36	36	36

IQ of Parents of Gifted Children



	Mother	Father	Diff.
Mean	118.2	114.8	3.4
SD	6.5	3.5	7.5
n	36	36	36

The parents were chosen randomly, so the differences will be independent. The histogram appears nearly normal (slight left skew, but $n = 36 > 30$).

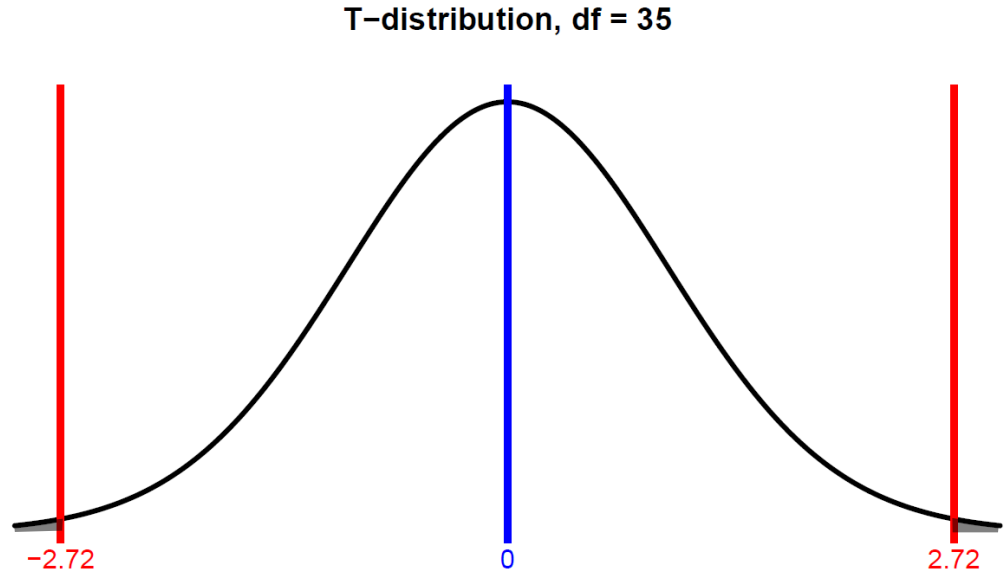
If we assume H_0: $\mu_d = 0$, then the average sample differences follow a $t_{36-1} = t_{35}$ distribution with center 0 and $SE = \dfrac{s_d}{\sqrt{n}} = \dfrac{7.5}{\sqrt{36}} \simeq 1.25$.

The t-score for our observed difference is

$$T = \frac{\bar{d} - 0}{SE} = \frac{3.4}{1.25} \simeq 2.72.$$

IQ of Parents of Gifted Children

If the alternative hypothesis is "H_A: $\mu_d \neq 0$", we find the following shaded area.

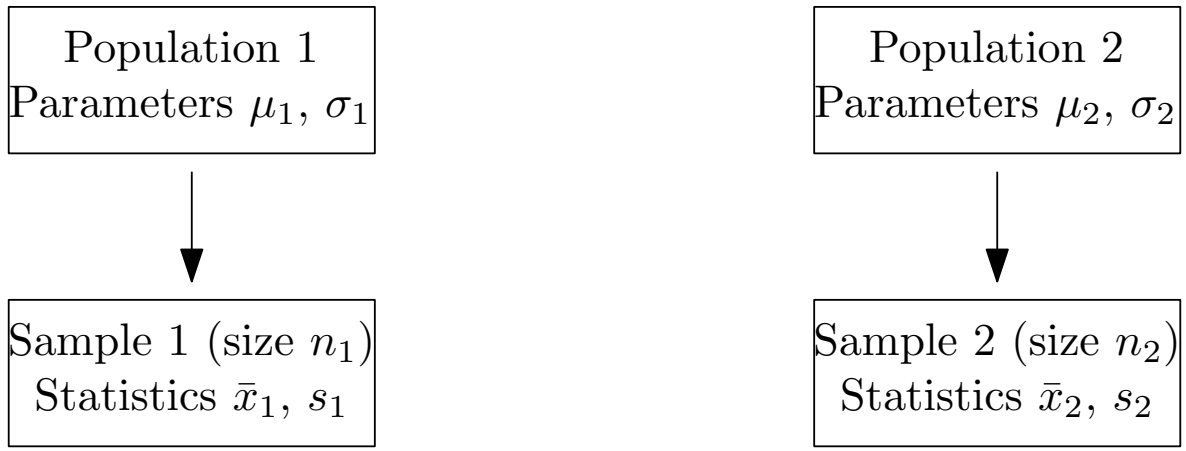



```
> 2*pt(-2.72,df=35,lower.tail=T)
[1] 0.01009512
```

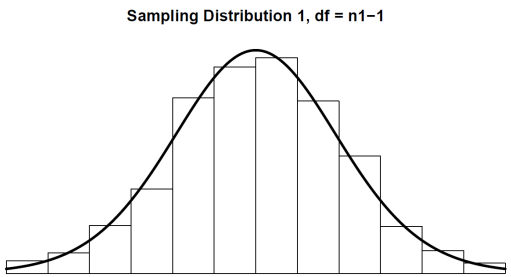
We get a p-value $p = 0.0101$.

We reject the null hypothesis. It does appear that there is a difference in the average IQ's of parents of gifted children.

Unpaired Independent Populations



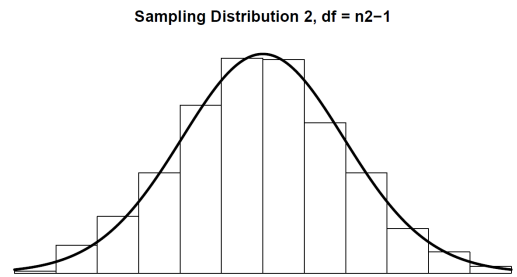
(**Note**: the samples may have different sizes)



T-distribution with
$df = n_1 - 1$
centered at μ_1
with $SE = \dfrac{s_1}{\sqrt{n_1}}$.

What does the sampling distribution of $\bar{x}_1 - \bar{x}_2$ look like?

Shape?
Center?
Spread?



T-distribution with
$df = n_2 - 1$
centered at μ_2
with $SE = \dfrac{s_2}{\sqrt{n_2}}$.

Unpaired Independent Populations

If $\bar{X} = t_{n_1-1}$ and $\bar{Y} = t_{n_2-1}$ are independent random variables both modelled by T-distributions, then $\bar{X} - \bar{Y}$ is also a T-distribution with

$$df = \min(n_1 - 1, n_2 - 1).$$

Furthermore, $\bar{X} - \bar{Y}$ is centered at

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2,$$

and has a SE which is found using the formula for the variance of a difference:

$$\begin{aligned} SE_{\bar{X}-\bar{Y}} = \sqrt{Var(\bar{X} - \bar{Y})} &= \sqrt{Var(\bar{X}) + Var(\bar{Y})} \\ &\simeq \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \end{aligned}$$

All C.I.'s are variations of one another

C.I. for	Formula	SE	df
1 sample	$\bar{x} \pm t^*_{df} SE_{\bar{x}}$	$\frac{s}{\sqrt{n}}$	$n-1$
2 paired samples	$\bar{d} \pm t^*_{df} SE_{\bar{d}}$	$\frac{s}{\sqrt{n}}$	$n-1$
2 independent samples	$\bar{x}_1 - \bar{x}_2 \pm t^*_{df} SE_{\bar{x}_1 - \bar{x}_2}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\min(n_1 - 1, n_2 - 1)$

Keep in mind: When you change the scenario being discussed, you change the sampling distribution, and hence, the critical value and standard error.
To make C.I.'s of new ideas, we just need to know what the sampling distribution is and we are all set!

Beetle Study (Again!)

We study beetle biodiversity in a pasture. For this, we collect a biodiversity index (Steinhaus index) in 2 different types of parcels:

1. in $n_1 = 12$ parcels where no animal grazes
2. in $n_2 = 13$ parcels with sheeps are grazing

```
> data1
 [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(data1)
[1] 0.2505833
> sd(data1)
[1] 0.09591138
> data2
 [1] 0.653 0.540 0.427 0.427 0.457 0.687 0.482 0.460 0.377
     0.507 0.622 0.323 0.463
> mean(data2)
[1] 0.4942308
> sd(data2)
[1] 0.1067459
```

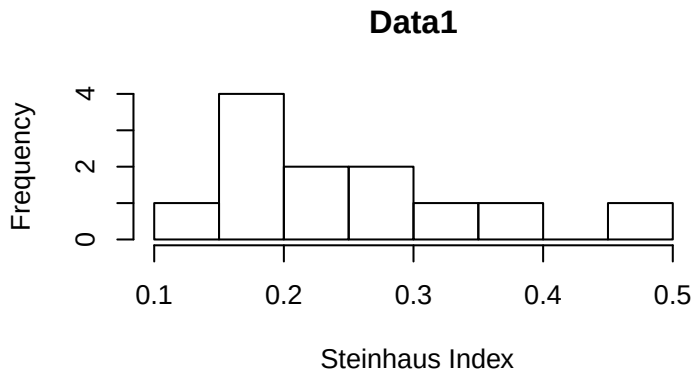
Build a test with level of confidence $\alpha = 5\%$ to determine if animal grazing influences the biodiversity of beetles.

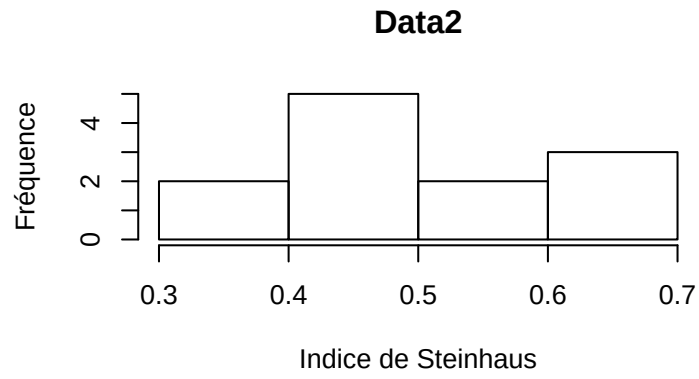
Beetle Study

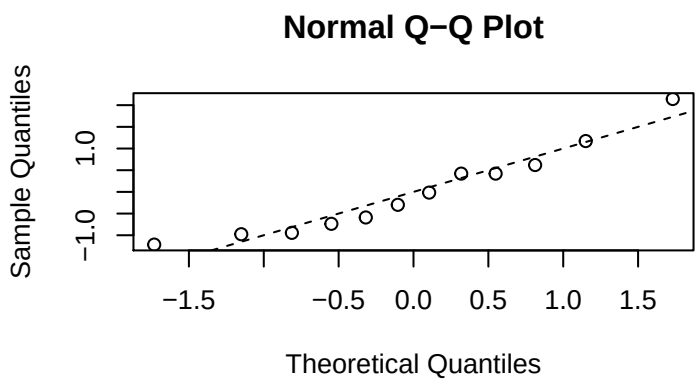
1) We build hypotheses for the situation

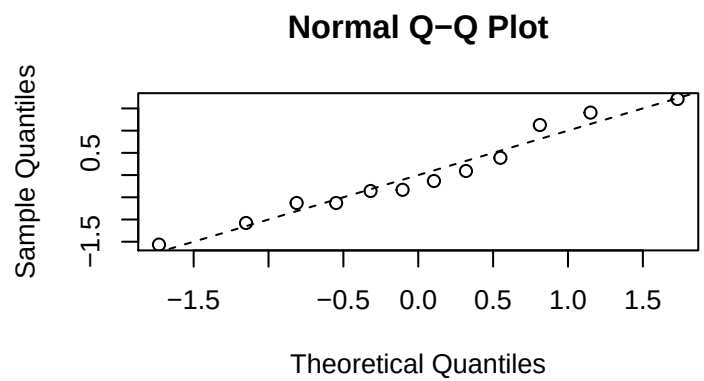
$$H_0\colon\ \mu_{grazed} = \mu_{not\ grazed},$$
$$H_A\colon\ \mu_{grazed} \neq \mu_{not\ grazed}.$$

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





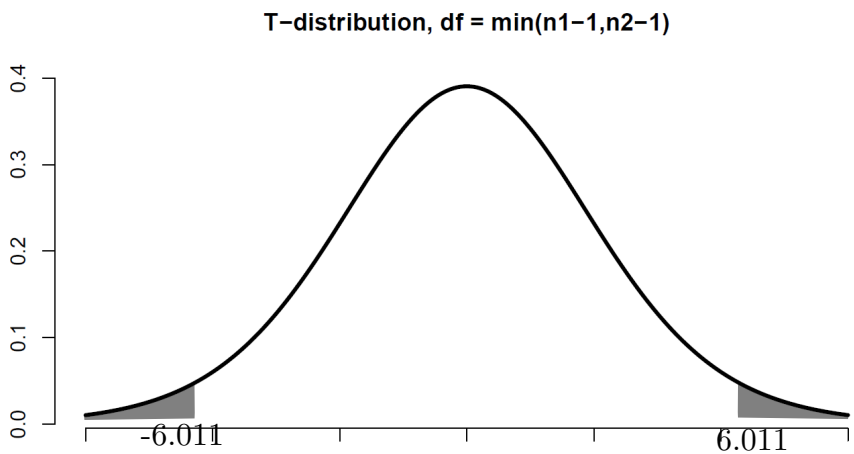




3) T-score your data

$$T = \frac{\text{point estimate} - \text{ null value}}{SE}$$
$$= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
$$= \frac{0.251 - 0.494}{\sqrt{\frac{0.095^2}{12} + \frac{0.107^2}{13}}} \simeq -6.011.$$

4) Compute the p-value. Here, $df = \min(12 - 1, 13 - 1) = 11$.




```
> 2*pt(-6.011, df= 11)
[1] 8.786302e-05
```

Since $p \simeq 9 \cdot 10^{-5} \ll 0.05$, we reject H_0 and favor H_A.

There is (a very) strong evidence that the animal grazing influences beetle biodiversity.

C.I.'s for Unpaired Data

Researchers were interested if smoking was linked with lower birth weights of babies. They sampled 150 random North Carolina mothers and found the below data.

	smoker	non-smoker
mean weight (lbs)	6.78	7.18
st. dev.	1.43	1.60
sample size	50	100

Find a 90% confidence interval for $\mu_{non-smoke} - \mu_{smoke}$.

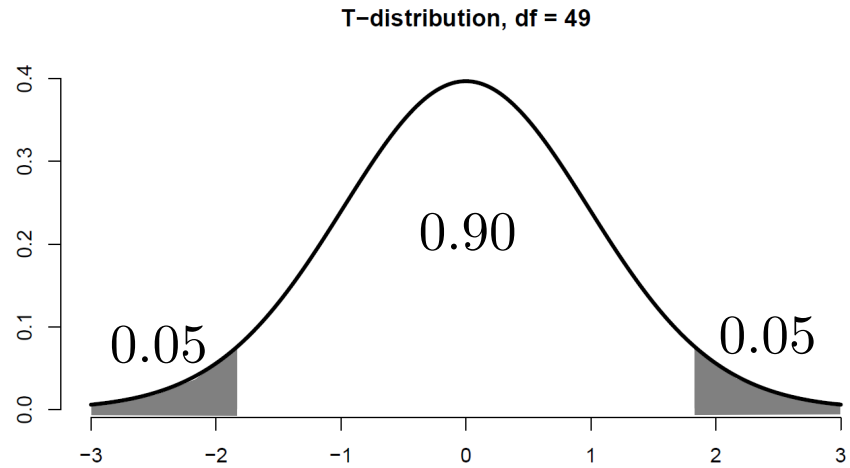
We must find $(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} SE_{\bar{x}_1 - \bar{x}_2}$.

Here, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1.6^2}{100} + \frac{1.43^2}{50}} \simeq 0.258.$

The sampling distribution for the difference in the sample means is a T-distribution with $df = \min(50 - 1, 100 - 1) = 49$.

Need to find the critical value t^*_{df}.

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	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72
	36	1.31	1.69	2.03	2.43	2.72
	37	1.30	1.69	2.03	2.43	2.72
	38	1.30	1.69	2.02	2.43	2.71
	39	1.30	1.68	2.02	2.43	2.71
	40	1.30	1.68	2.02	2.42	2.70
	41	1.30	1.68	2.02	2.42	2.70
	42	1.30	1.68	2.02	2.42	2.70
	43	1.30	1.68	2.02	2.42	2.70
	44	1.30	1.68	2.02	2.41	2.69
	45	1.30	1.68	2.01	2.41	2.69
	46	1.30	1.68	2.01	2.41	2.69
	47	1.30	1.68	2.01	2.41	2.68
	48	1.30	1.68	2.01	2.41	2.68
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68



We find $t^*_{49} = 1.68$.

Since $\bar{x}_1 - \bar{x}_2 = 7.18 - 6.78 = 0.4$, we have

$$CI = 0.4 \pm 1.68 \times 0.258 = (-0.03, 0.83).$$

We are 90% confident that babies born to non-smoking NC women are about 0.83 to -0.03 lbs heavier than babies born to smoking NC women.

Don't Forget Sampling Conditions!

To get each of the individual sampling distribution to be a T-distribution, we need (in each sample):

- Independence of items in the sample (usually shown through randomization and <10% rules)
- Nearly normal distribution

To be able to substract the T approximations for each sampling distribution and use our variance formula to get the SE of the difference:

- Independence of the two samples (no datum in one sample should help you predict any datum in the other sample)

Your Turn!

Which of the following scenarios involve paired data?

1. Comparing students' self-reports of "love for statistics" before and after E. Aamari's class.
2. Assessing the gender-related salary gap by comparing salaries of men and women in the same randomly sampled positions at the same companies.
3. Comparing lung capacity changes in athletes before and after six weeks of training.
4. Assessing the claim that Uber is better than Lyft by dividing 70 random people intro two groups of 35 and asking for their feedback on the one service they were assigned.
5. Exploring the average attractiveness of husbands and wives in couples who own a yacht.

Answer:

1. Paired. The linkage in the student.
2. Paired. The linkage is the common job.
3. Paired. The linkage is the athlete.
4. No paired. Paired data would be people trying both.
5. Paired. The linkage is marriage.