Stat 411/511

Two-Sample T

Oct 12 2015

Charlotte Wickham

stat511.cwick.co.nz

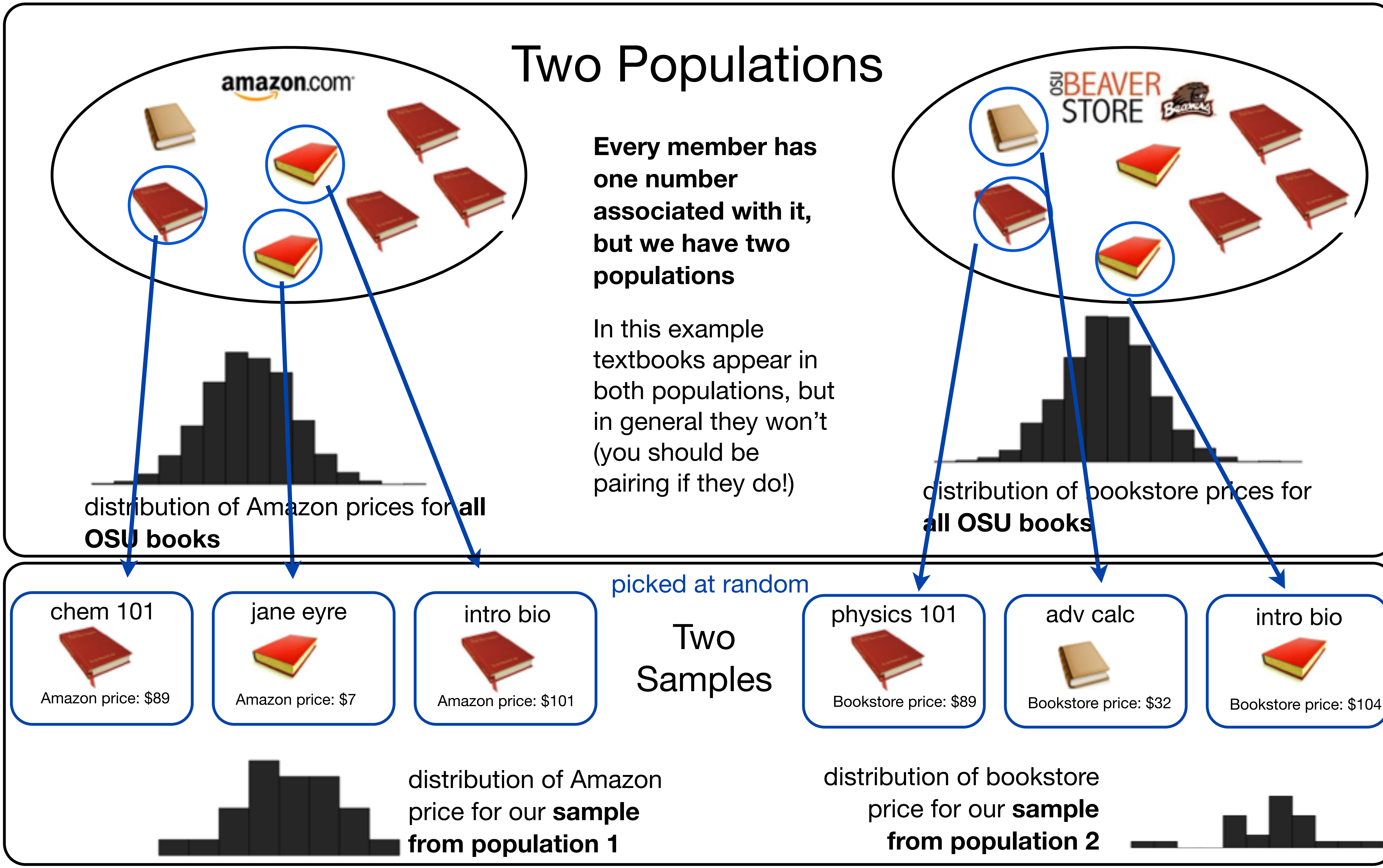
Today

The two sample model

The two sample t-test and CI

When sampling isn't random

Two sample sampling model



Two sample inference

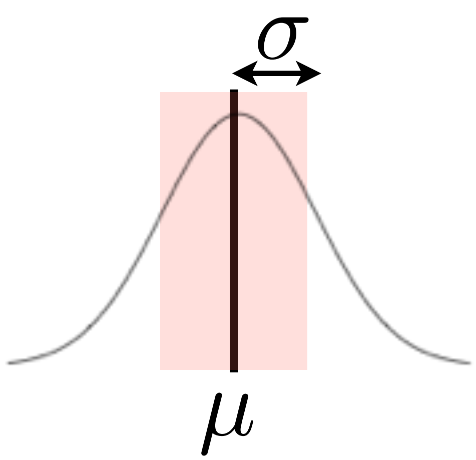
In the two sample model our questions are about the parameters of two populations.

We want to use our two samples to make inferences about the two populations, usually the difference in their means.

Paired case

one sample of differences

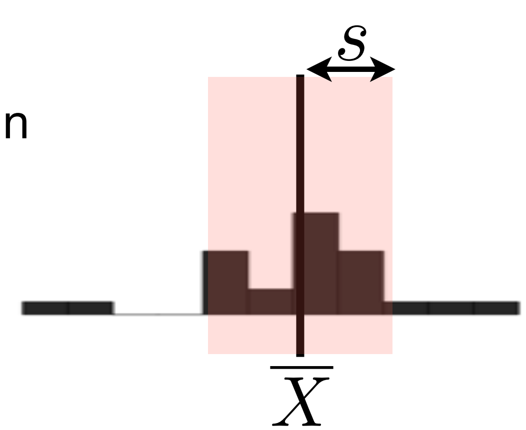
one population



the population distribution of differences
with unknown mean, μ
and standard deviation σ

one random sample

of size n

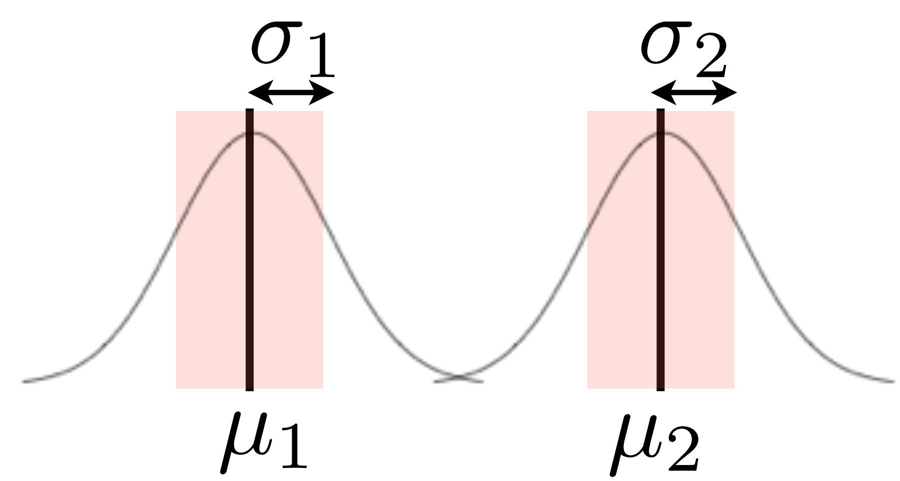


sample differences,
with sample average, $\overline{X}$
and sample standard deviation, s

use sample to make inferences about the **mean difference,** μ

Two sample case

two populations



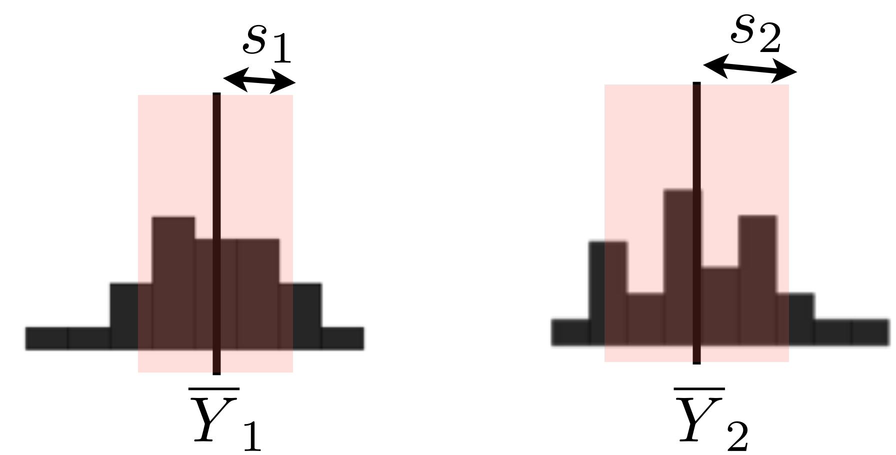
population 1
with unknown mean, μ_1
and standard deviation σ_1

population 2
with unknown mean, μ_2
and standard deviation σ_2

two random samples

of size n_1

of size n_2



sample 1,
with sample average, $\overline{Y}_1$
and sample standard
deviation, s_1

sample 2,
with sample average, $\overline{Y}_2$
and sample standard
deviation, s_2

use samples to make inferences about the **difference in means**, μ_2 - μ_1

Your turn

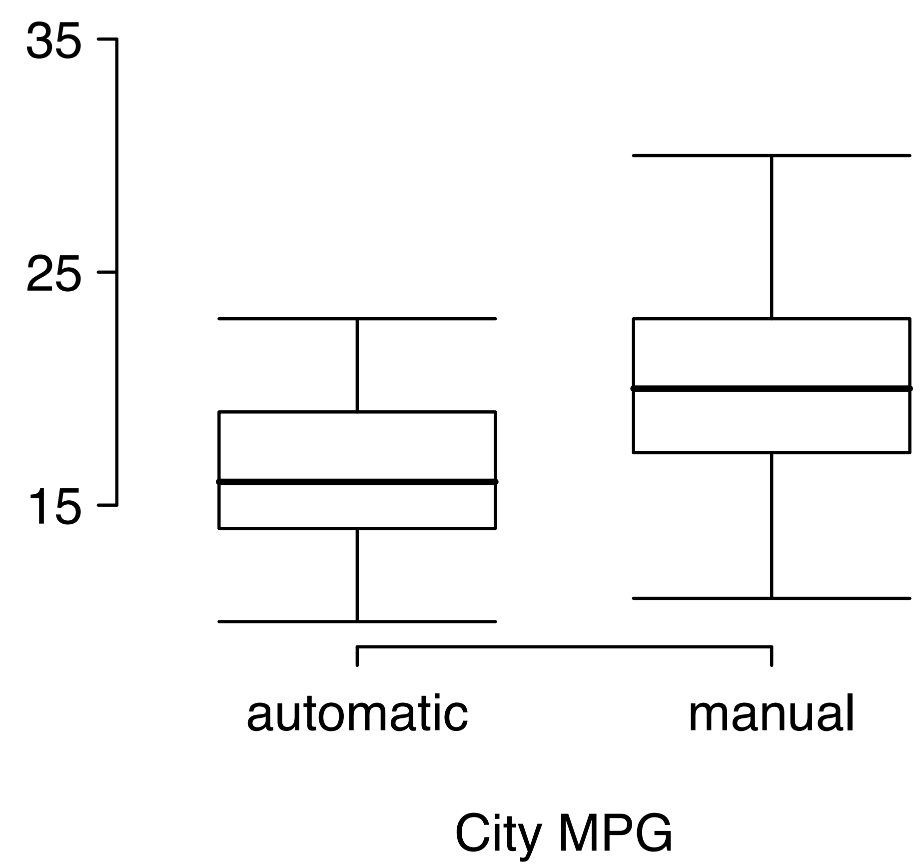
In a one-sample case, we started by looking at the sampling distribution of the sample average.

What would be a good **one number summary** for the two-sample problem?

from OpenIntro

5.30 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[45]

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



difference in sample averages
= 3.73

Facts about the sampling distribution for the

difference in two sample averages

assuming the samples are independent

The sampling distribution of $\overline{Y}_2 - \overline{Y}_1$:

1 will have the mean $\mu_2 - \mu_1$

2 have standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

3 and it's shape will be closer to a Normal distribution than the population distributions

(how close depends on the sample size and how close the population distributions were to Normal).

Use $\overline{Y_2} - \overline{Y_1}$ to make inferences about

$$\mu_2 - \mu_1$$

Assume that both groups have **Normal** population distributions with the **same standard deviation**.

Same idea as one-sample

If the **populations are Normal,**

the sampling distribution of the difference in sample averages is Normal,

but depends on the **unknown** population standard deviation.

If instead we look at the **two-sample t-ratio**,

then it’s sampling distribution doesn’t depend on the unknown population standard deviation.

The two sample t-ratio

If the populations are Normal,

and have the same standard deviation

Fact:

The two-sample *t*-ratio:

$$\frac{(\overline{Y_2} - \overline{Y_1}) - (\mu_2 - \mu_1)}{\mathrm{SE}_{\overline{Y_2} - \overline{Y_1}}}$$

can be described by a **Student's *t*-distribution** with $\mathbf{n_1 + n_2 - 2}$ degrees of freedom

The two sample t-ratio

If the populations are Normal,

and have the same standard deviation

Fact:

The two-sample *t*-ratio:

$$\frac{(\overline{Y_2} - \overline{Y_1}) - (\mu_2 - \mu_1)}{\mathrm{SE}_{\overline{Y_2} - \overline{Y_1}}}$$

can be described by a **Student's *t*-distribution** with $\mathbf{n_1 + n_2 - 2}$ degrees of freedom

The two-sample *t*-ratio:

$$\frac{(\overline{Y_2} - \overline{Y_1}) - (\mu_2 - \mu_1)}{\mathrm{SE}_{\overline{Y_2} - \overline{Y_1}}}$$

can be described by a **Student's *t*-distribution** with **n_1 + n_2 - 2** degrees of freedom

Leads to:

95% CIs

$$\overline{Y_1} - \overline{Y_2} \pm t_{n_1+n_2-2}(0.975) \times \mathrm{SE}_{\overline{Y_1} - \overline{Y_2}}$$

And tests

Null Hypothesis: The population means are equal $\mu_1 = \mu_2$
Alternative hypothesis: The population means are not equal $\mu_1 = \mu_2$

Compare the two sample $\text{t-statistic} = \dfrac{\overline{Y_1} - \overline{Y_2}}{\mathrm{SE}_{\overline{Y_1} - \overline{Y_2}}}$ to a t-distribution with $n_1 + n_2 - 2$ d.f.

What is $\mathrm{SE}_{\overline{Y_2}-\overline{Y_1}}$?

An **estimate** of the standard deviation of the sampling distribution of $\overline{Y_2} - \overline{Y_1}$

With our assumption that the populations have the same standard deviation,

$$\sigma_1 = \sigma_2 = \sigma$$

Then the standard deviation of the sampling distribution of $\overline{Y_2} - \overline{Y_1}$

is

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

We need to **estimate** σ

Pooled standard deviation

We have two samples each with their own standard deviation, s_1 and s_2.

Our assumption tells us these should each by estimating σ.

We need to combine them to get a **pooled** sample standard deviation.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Use to estimate σ

What is $\mathrm{SE}_{\overline{Y_2}-\overline{Y_1}}$?

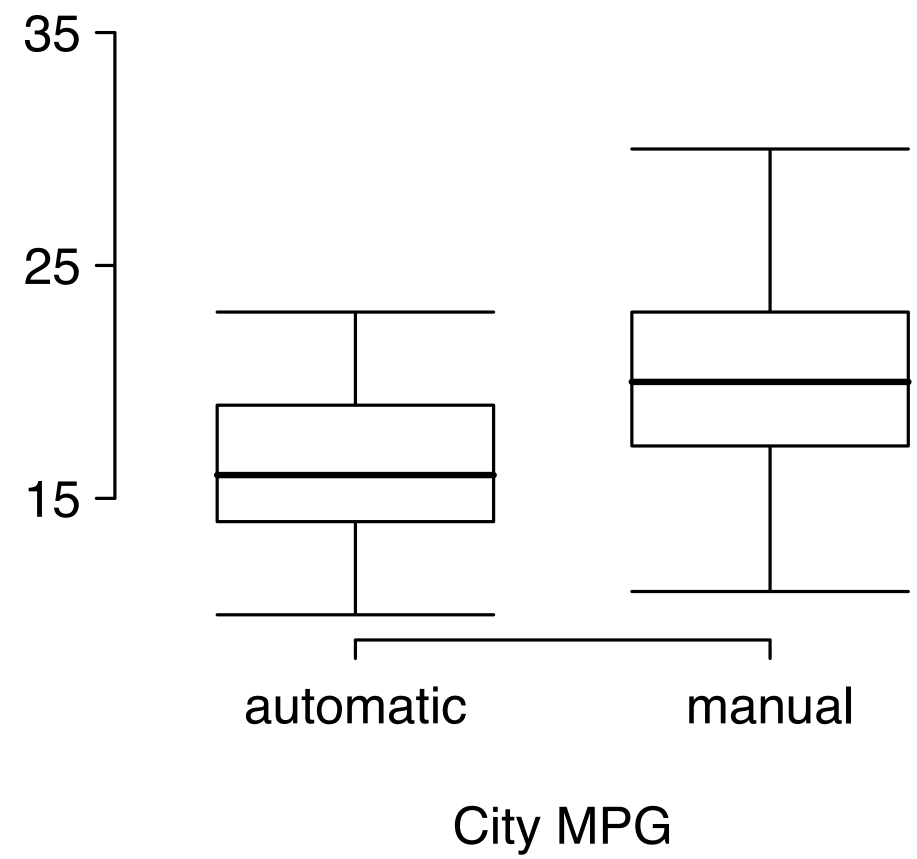
Our **estimate** of the standard deviation of the sampling distribution of $\overline{Y}_2 - \overline{Y}_1$

is

$$\mathrm{SE}_{\overline{Y_2}-\overline{Y_1}} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

5.30 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[45]

	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



Your turn:

Find the standard error on the difference in sample averages

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$\mathrm{SE}_{\overline{Y_2} - \overline{Y_1}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Two sample t-test in R

```
> t.test(city_mpg ~ trans, data = mpg_sample,
         var.equal = TRUE)

   Two Sample t-test
                                   2*(1 - pt(2.7278, 50))

data:  city_mpg by trans
t = -2.7278, df = 50, p-value = 0.008774
alternative hypothesis: true difference in means
   is not equal to 0
95 percent confidence interval:
 -5.2757919 -0.8011311
sample estimates:
  mean in group auto mean in group manual
            17.38462             20.42308
```

Statistical Summary

There is convincing evidence that the mean fuel efficiency of automatic cars manufactured in 2012 is not equal to the mean fuel efficiency of manual cars manufactured in 2012 (two sample t-test, two-sided p-value = 0.009).

The mean fuel efficiency of automatic cars manufactured in 2012 is estimated to be 3.0 mpg lower than the mean fuel efficiency of manual cars manufactured in 2012.

With 95% confidence the mean fuel efficiency of automatic cars is between 0.8 and 5.3 mpg lower than the population mean fuel efficiency of manual cars manufactured in 2012.

t-tools summary so far

The t-tools are motivated by the random sampling models (paired or two sample).

Which t-tool is appropriate (paired or two sample) depends on the design of the study.

The sampling distributions of the t-ratios are known exactly if you also assume Normal populations (and in the two sample case, equal population standard deviations).

Our conclusions are about the parameters of the populations (mean difference or difference in means).

What if you don't have random samples?

Often people proceed with the t-tools anyway.

The conclusions rely on an additional assumption,

"our data **is just like a random sample** from a population of interest"

This assumption is always suspect, and any deviations can lead to significant bias and misleading conclusions.

Arguments for why your **"not random"** sample is just like a **random** sample cannot be backed up statistically.

There is one situation where the t-tools can be used
without random sampling, but they become an approximation
this is where we are heading this week....

Some interesting reading about non-random samples:

http://www.stat.berkeley.edu/~census/berk2.pdf

Conventional statistical inferences (e.g., formulas for the standard error of the mean, t-tests, etc.) depend on the assumption of random sampling. This is not a matter of debate or opinion; it is a matter of mathematical necessity.[3] When applied to convenience samples, the random sampling assumption is not a mere technicality or a minor revision on the periphery; the assumption becomes an integral part of the theory.