ENV 710: Lecture 6

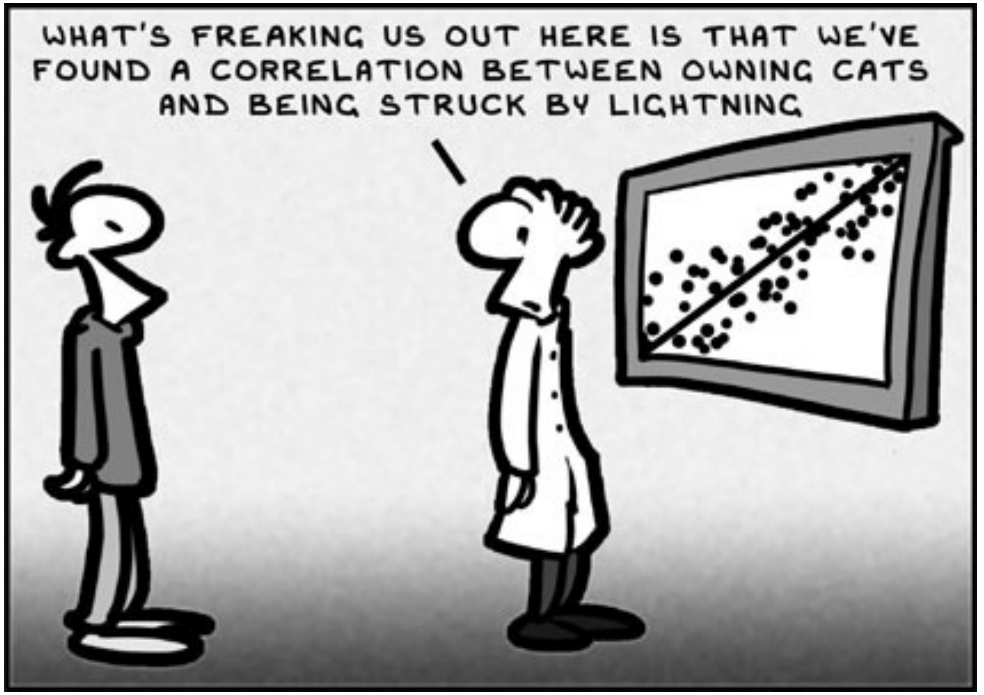
one- and two-sample tests

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT


COMMON SENSE

data dredging

data dredging is the failure to acknowledge that a correlation was in fact the result of chance

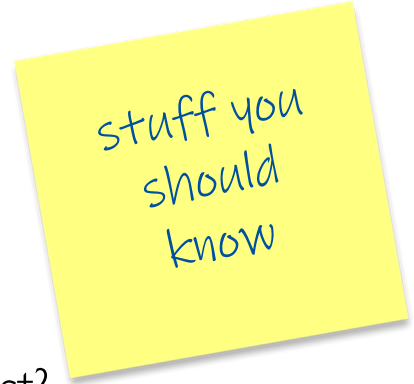


statistical inference

differences in means

learning goals

- what is the appropriate method for hypothesis testing?
- when should you use the z-test or the t-test to test differences in independent means?
- what is the difference between a 1-sample and 2-sample test?
- what is the difference between a 1-sided and 2-sided test?
- when should you use a paired t-test?
- what are the assumptions of different types of t-tests and how to conduct the tests in R?
- understand the results output from R output for t-tests

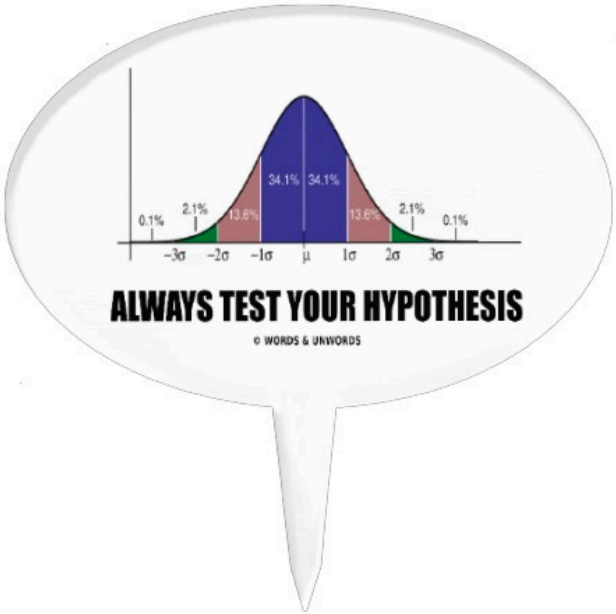


classical hypothesis testing

- establish *a priori* an α level (significance or error level) – historically, $\alpha = 0.05$

- state null hypothesis, H_0 (=)
- state alternative hypothesis, H_a (>,<, ≠)

- conduct statistical test
- reject H_0 if the p-value of the test statistic ≤ α



use (abuse) of hypothesis testing

- choice of level of significance, α, often arbitrary
- don't ignore lack of significance → p-value of 0.051?
- statistical significance is not the same as practical significance, especially with large sample sizes
- statistical inference is not valid for all sets of data
  - e.g., poorly designed study → randomize!!
- data dredging: do not run 100's of tests on 100's of variables looking for "significance"
  - a few will be significant by chance alone.

hypotheses make claims about parameters

How much more (or less) do college graduates work, on average, than non-graduates in the US?

parameter of interest

average difference between the number of hours worked per week by all Americans with a college degree compared to those without a college degree

$\mu_{grad} - \mu_{no}$

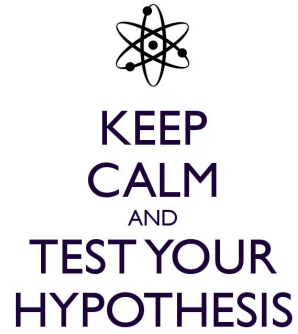
point of interest

average difference between the number of hours worked per week by sampled Americans with a college degree compared to those without a college degree

$\bar{x}_{grad} - \bar{x}_{no}$

testing for a difference between independent means

- null hypothesis: no difference $H_0: \mu_1 - \mu_2 = 0$
- alternative hypothesis: some difference $H_a: \mu_1 - \mu_2 \neq 0$



differences between independent means

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sigma_{\bar{x}_1 - \bar{x}_2}$$

- when the population σ is known

this equation gives the confidence interval for the difference between 2 means

standard error of difference between two independent means:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

differences between independent means

standard error of a statistic (e.g. mean) is the standard deviation of its sampling distribution

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t_{[df]\alpha/2} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

this equation gives the confidence interval for the difference between 2 means

- when the population σ is unknown

standard error of difference between two independent means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

conditions for comparing 2 independent means

1. independence
   - within groups: sampled observations must be independent
     - random sample/assignment
     - if sampling without replacement, n < 10% of population
   - between groups: the two groups must be independent of each other (non-paired)

2. sample size/skew: each sample must be at least 30 ($n_1 \geq 30$ and $n_2 \geq 30$), or larger if the population distributions are very skewed

How much more (or less) do college graduates work, on average, than non-graduates in the US?

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

- note use of *s* rather than σ
- when *N* is really big, can use *z* or *t*

	$\bar{x}$	s	n
grad	41.8	15.14	505
no grad	39.4	15.12	667

the table provides summary statistics, but these could also be calculated from two samples of data

How much more (or less) do college graduates work, on average, than non-graduates in the US?

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot SE_{\bar{x}_1 - \bar{x}_2}$$

	$\bar{x}$	s	n
grad	41.8	15.14	505
no grad	39.4	15.12	667

$$(41.8 - 39.4) \pm 1.96\sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}}$$

$$= 2.4 \pm 1.96(0.89) = 2.4 \pm 1.74 = (0.66, 4.14)$$

what does the confidence interval mean here?

Is the average number of hours worked by college graduates *significantly* different than the number of hours worked by non-graduates in the US?

$$\bar{x}_{grad} - \bar{x}_{no} = 2.4$$

$$SE = 0.89$$

$H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 \neq 0$

what is being asked here?

hint: the term 'significantly' is asking that we calculate the p-value (probability) of our statistic given the null hypothesis

sketch the null and alternative hypotheses on a probability distribution

Is the average number of hours worked by college graduates *significantly* different than the number of hours worked by non-graduates in the US?

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$

method 1

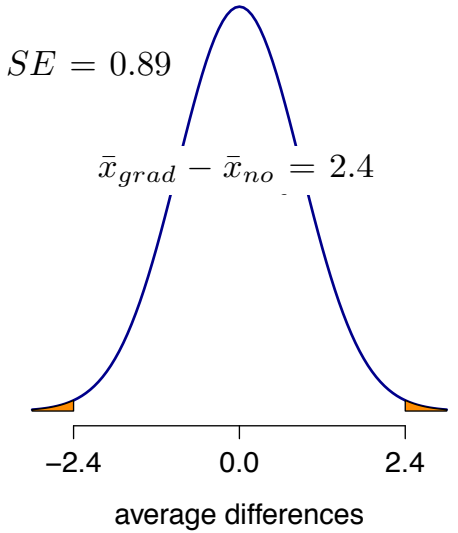
$$Z = \frac{2.4 - 0}{0.89} = 2.70$$

$$P(Z = 2.70) = 0.003(2) = 0.007$$

```
pnorm(q=-2.7, mean=0, sd=1)+
1-pnorm(q=2.7, mean=0, sd=1)
```

method 2

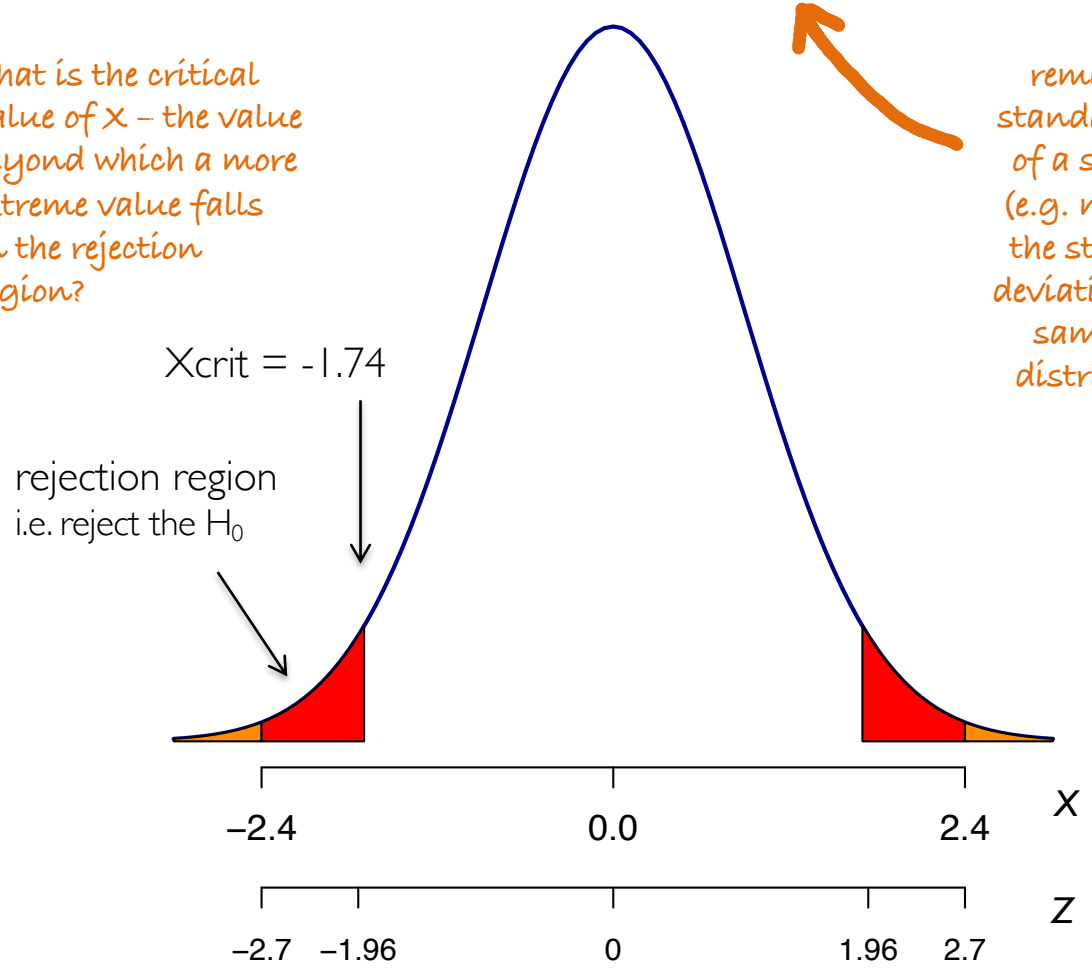
```
1- pnorm(q = 2.4, mean = 0, sd = 0.89) * 2
```



```
x_crit <- qnorm(0.025, mean = 0, sd = 0.89)
```

what is the critical value of X – the value beyond which a more extreme value falls in the rejection region?

remember: standard error of a statistic (e.g. mean) is the standard deviation of its sampling distribution



interpret the p-value in the context of the data & hypotheses

p-value = P(observed or more extreme statistic|H_0 true)

difference of 2.4 hours per week

no difference in average hours worked

there is a 0.7% chance of obtaining random samples of 505 graduates and 667 non-college graduates where the average difference in their weekly work hours is at least 2.4 hours under the null hypothesis

statistical inference

t-tests

definition

- degrees-of-freedom: number of values in the calculation of a statistic that are free to vary
  - sample variance has *n-1* df's because one parameter (mean) is computed as an intermediate step

$$s^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$$

assumptions

1-sample test

a one sample test compares the mean of a sample to a known value (often 0, but not always)

- test assumptions about a population with a sample
- a random sample of independent values has been obtained from a normal parent distribution
- if assumptions are not met, our results may not be valid…

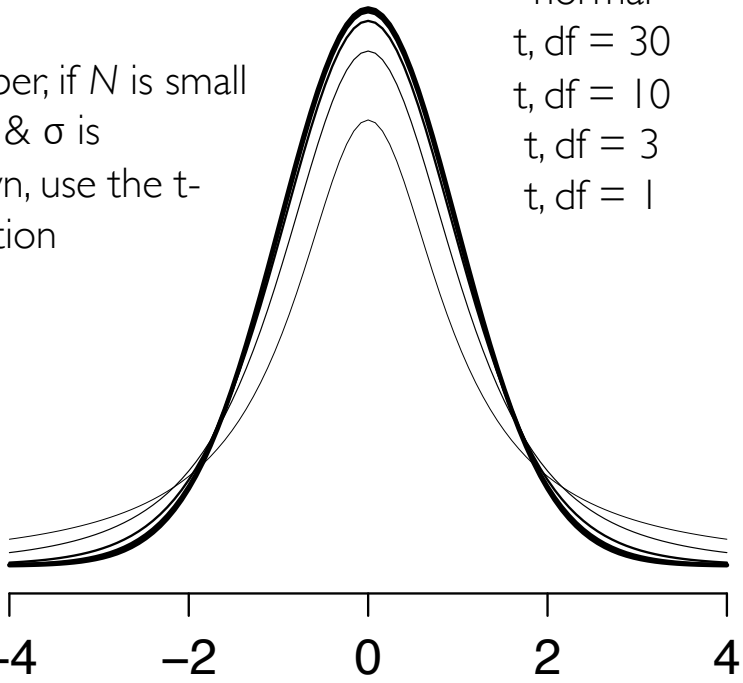
2-sample test

two-sample tests compare the means of two different samples

- two random samples of independent values from normal parent distributions with equal variances
  - if sample sizes are equal between samples, t-test is robust to unequal variances
  - Welch's t-test corrects for unequal variances regardless of whether sample sizes are similar
- long-tailedness/outliers and differences in skewness between samples can cause the test to be invalid

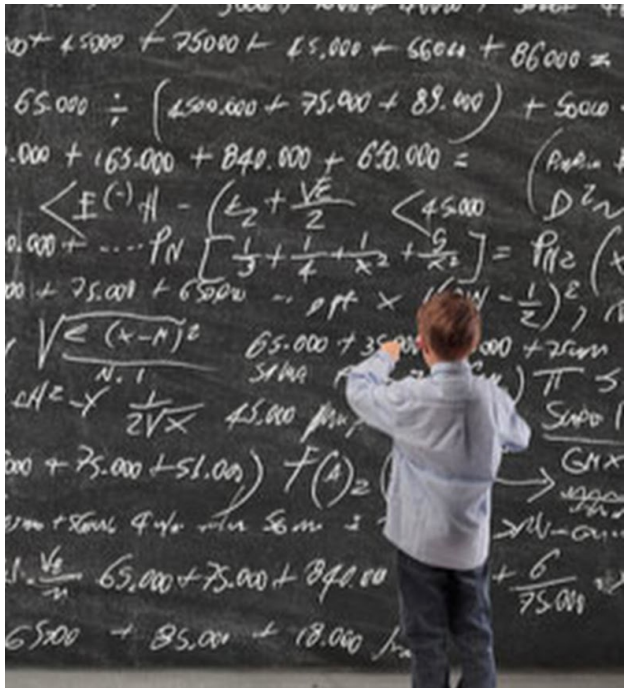
t-distribution

remember, if *N* is small (< 30) & σ is unknown, use the t-distribution



our process

1. calculate a statistic from the data
2. find the probability of getting that statistic, assuming the H_0 is true



t-test and t-statistic

- use to test sample means when:
  - σ is unknown
  - $n < 30$
- calculated the same as the z-statistic
- p-value (same definition)
  - one or two-tailed, depending on H_a

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$df = n - 1$$

- 1 sample test
- 2 sample test
- paired test

1-sample t-test

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$df = n - 1$$

for 1-sample test, we are testing our data to a hypothetical mean, μ

μ is a single number

calculate t and find its probability

or use R to do the same...

```
t.test(x, mu, alternative)
```

new light bulbs better?

A standard manufacturing process has produced millions of light bulbs, with a mean life of 1200 hours. A new process, recommended by the EPA, produces a sample of 25 bulbs with an average of 1265 hours (s = 300 hours). Are the new bulbs better, or is this just a sampling fluke?

null hypothesis?
alternative hypothesis?

calculations?

example

new light bulbs better?

A standard manufacturing process has produced millions of light bulbs, with a mean life of 1200 hours. A new process, recommended by the EPA, produces a sample of 25 bulbs with an average of 1265 hours (s = 300 hours). Are the new bulbs better, or is this just a sampling fluke?

H_0: $\mu = 1200$
H_a: $\mu > 1200$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad df = n - 1$$

$$t = \frac{1265 - 1200}{300/\sqrt{25}} = \frac{65}{60} = 1.08$$

```
P(t ≥ 1.08, df=24)

1- pt(q = 1.08, df = 24)
[1] 0.1454382
```

what does this p-value mean? how to cite it?

example

new light bulbs better?

A standard manufacturing process has produced millions of light bulbs, with a mean life of 1200 hours. A new process, recommended by the EPA, produces a sample of 25 bulbs with an average of 1265 hours (s = 300 hours). Are the new bulbs better, or is this just a sampling fluke?

```
bulbs <- c(1838, 818, 1069, 1218, 985, 1262, 1273, 1569,
1552, 1366, 1232, 1217, 1736, 1296, 1137, 990, 1290, 978,
1849, 769, 1541, 1458, 1388, 1010, 792)
```

```
t.test(x = bulbs, mu = 1200, alternative = "greater")
```

options: alternative = "less", "greater", or "two.sided"

example

new light bulbs better?

A standard manufacturing process has produced millions of light bulbs, with a mean life of 1200 hours. A new process, recommended by the EPA, produces a sample of 25 bulbs with an average of 1265 hours (s = 300 hours). Are the new bulbs better, or is this just a sampling fluke?

```
One Sample t-test

data:  bulbs
t = 1.0696, df = 24, p-value = 0.1477
alternative hypothesis: true mean is greater than 1200
95 percent confidence interval:
 1160.835      Inf
sample estimates:
mean of x
  1265.32
```

2-sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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

$$df = n_1 + n_2 - 2$$

- variances assumed to be equal
- 2 independent samples are normally distributed

```
t.test(x, y, var.equal = T)
```

- t calculates the difference between two samples
- the standard error is a pooled standard error for both samples

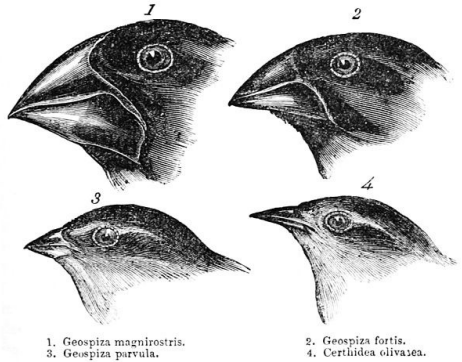
example

Grant's finch beaks

Grant collected measurements (mm) of the beaks of Darwin's finches in the Galapagos in both 1976 and 1978. Is there a difference in the size of the beaks between the two years?

1976	1978
9.1	7.1
10.0	8.0
8.8	7.9
…	…
9.5	11.7

1976	1978
$n_{76} = 89$	$n_{78} = 89$
$\bar{x}_{76} = 9.629$	$\bar{x}_{78} = 10.138$
$s^2_{76} = 0.897$	$s^2_{78} = 0.822$



example

Grant's finch beaks

Grant collected measurements (mm) of the beaks of Darwin's finches in the Galapagos in both 1976 and 1978. Is there a difference in the size of the beaks between the two years?

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}} = \sqrt{\frac{88 \cdot 0.897 + 88 \cdot 0.822}{(89+89-2)}} = 0.927$$

$$df = n_1 + n_2 - 2 = (89 + 89 - 2) = 176$$

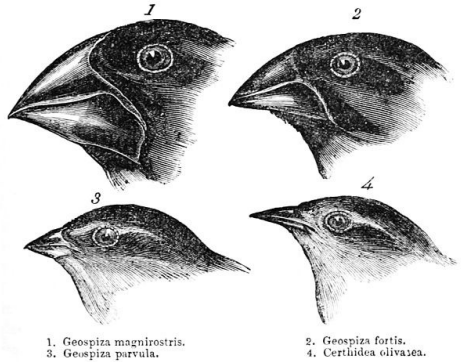
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{9.629 - 10.138}{0.927\sqrt{1/89 + 1/89}} = -3.662$$

what is the probability (p-value)?

$$P(-3.662 \leq t \geq 3.662)|\mu = 0)$$

1976	1978
9.1	7.1
10.0	8.0
8.8	7.9
…	…
9.5	11.7

1976	1978
$n_{76} = 89$	$n_{78} = 89$
$\bar{x}_{76} = 9.629$	$\bar{x}_{78} = 10.138$
$s^2_{76} = 0.897$	$s^2_{78} = 0.822$



example

Grant's finch beaks

Grant collected measurements (mm) of the beaks of Darwin's finches in the Galapagos in both 1976 and 1978. Is there a difference in the size of the beaks between the two years?

```
require(Sleuth3)
dat <- case0201
d1976 <- dat$Depth[dat$Year == "1976"]
d1978 <- dat$Depth[dat$Year == "1978"]
```

4 different ways to test equality of variances

car package

test assumption of equal variances, e.g. H_o: var(sample1) = var(sample2)

```
sd(d1976)/sd(d1978)
leveneTest(dat$Depth ~ factor(dat$Year))
bartlett.test(Depth ~ Year, data = dat)
var.test(x = d1976, y = d1978)
```

example

Grant's finch beaks

Grant collected measurements (mm) of the beaks of Darwin's finches in the Galapagos in both 1976 and 1978. Is there a difference in the size of the beaks between the two years?

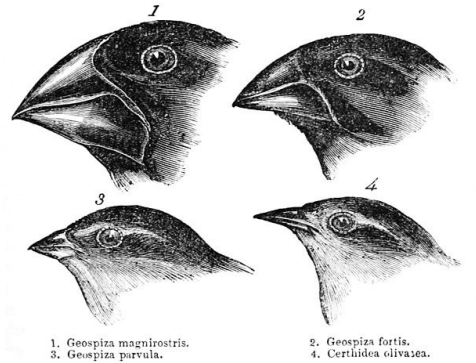
```
var.test(d1976, d1978)

        F test to compare two variances

data:  d1976 and d1978
F = 1.3045, num df = 88, denom df = 88, p-value = 0.2144
alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:
 0.8566123 1.9865508
sample estimates:
ratio of variances
          1.304494
```

example

Grant's finch beaks

Grant collected measurements (mm) of the beaks of Darwin's finches in the Galapagos in both 1976 and 1978. Is there a difference in the size of the beaks between the two years?

test difference in means between 2 samples




```
t.test(x=d1976, y=d1978, var.equal = T)
```

example

Grant's finch beaks

Grant collected measurements (mm) of the beaks of Darwin's finches in the Galapagos in both 1976 and 1978. Is there a difference in the size of the beaks between the two years?

```
t.test(x=d1976, y=d1978, var.equal = T)
```

```
            Two Sample t-test

data:  d1976 and d1978
t = -4.5833, df = 176, p-value = 8.65e-06
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
 -0.9564088 -0.3806698

sample estimates:
mean of x mean of y
 9.469663 10.138202
```

two sample t-test, unequal variances

- Satterthwaite's correction used to calculate *df*
- Welch's t-test will have a lower number of df's than (n_1 + n_2 - 2), which was used for the case of equal variances between samples

```
t.test(x, y, var.equal = F)
```

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}]}$$

paired t-test

- paired data are not independent!
- every data point in one sample is uniquely paired to a data point in a second sample
- data might come from the same observational unit (individual) or the same location

paired samples

$$d_i = x_{2_i} - x_{1_i}$$

$$t = \frac{\bar{d}_i}{s_{d_i}/\sqrt{n_{d_i}}}$$

$$df = n_{d_i} - 1$$

```
t.test(x, y, paired = T)
```

example

A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the same cars using the other kind of gasoline. Do cars get better mileage with premium?

```
reg = c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
prem = c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
```

```
t.test(reg, prem, paired = T)
```

```
        Paired t-test

data:  reg and prem
t = -4.4721, df = 9, p-value = 0.00155
alternative hypothesis: true difference
 in means is not equal to 0
95 percent confidence interval:
 -3.0116674 -0.9883326
sample estimates:
mean of the differences
                     -2
```

recap

1. if sample size < 30, use the *t*-test
2. evaluate whether data are normally distributed
   - separately for each group in a 2-sample test
3. calculate *s* (sample standard deviation) of each group to test equality of variances
   - if ratio of largest *s* to smallest *s* < 2, use t-test with equal variances, otherwise t-test without equal variances
   - or test with the *F*-test for variance, `var.test()`
4. calculate *t*-statistic and *p*-value

Post your questions to be answered during lecture