UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE
Multiomics Data Science

BIOSTATISTICS

Lecture 7

Hypothesis about Means and Proportions of Two Populations

Petr Nazarov

Email: petr.nazarov@lih.lu
Skype: pvn.public

26-03-2021

UNIVERSITÉ DU LUXEMBOURG

OUTLINE

Lecture 7

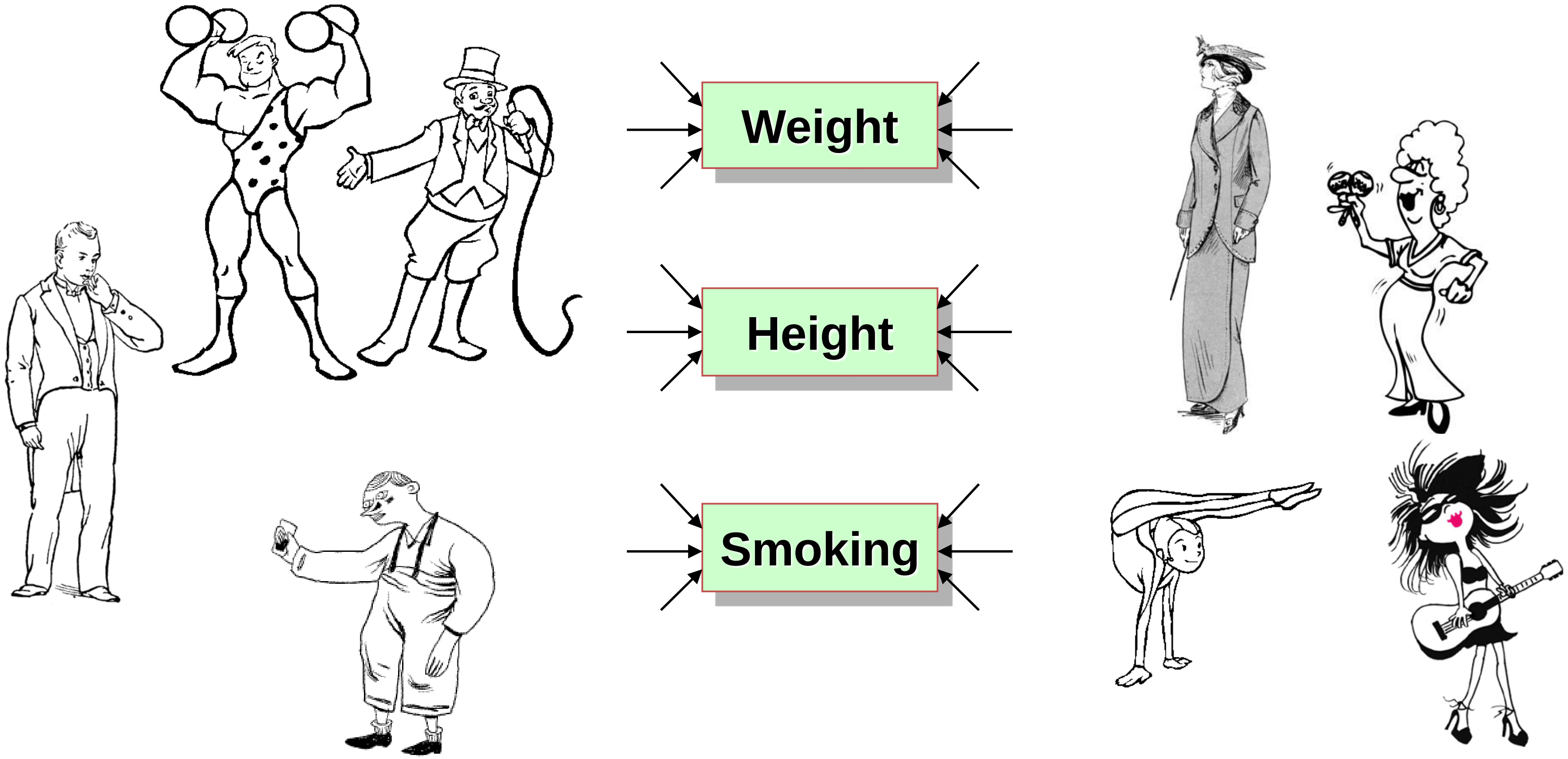
- Independent and dependent samples
- Comparison of means: t-test
- Paired t-test
- Comparison of two proportions

TWO POPULATIONS

Independent Samples

Independent samples

Samples selected from two populations in such a way that the elements making up one sample are chosen independently of the elements making up the other sample.



TWO POPULATIONS

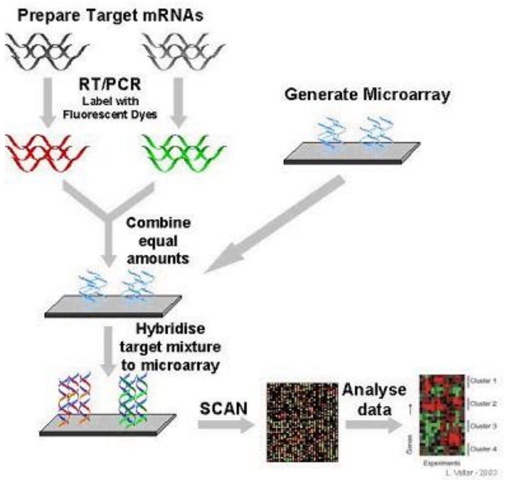
Dependent Samples

Matched samples

Samples in which each data value of one sample is matched with a corresponding data value of the other sample.

Before treatment | **After treatment**

Analysis

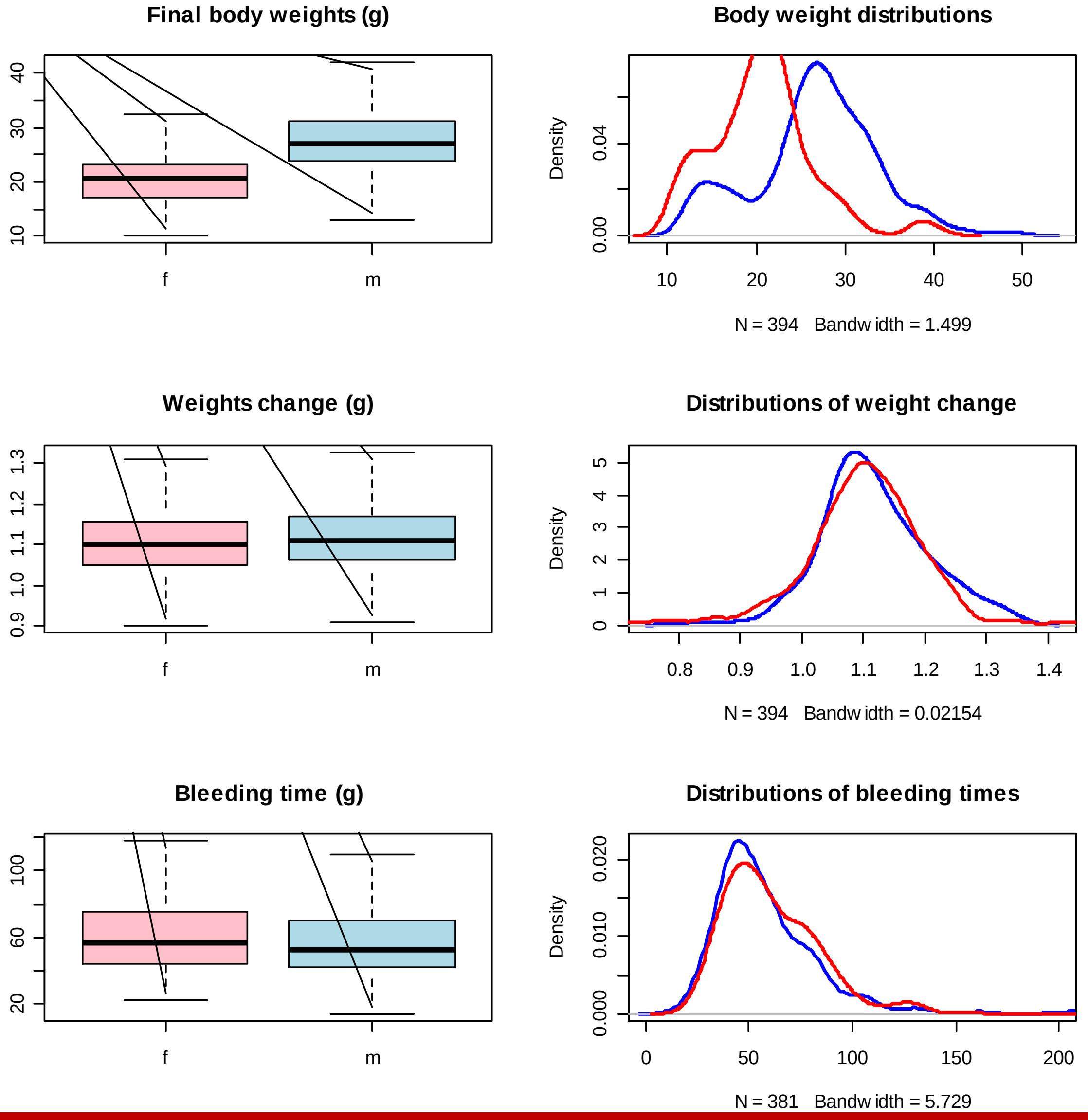


MEANS OF TWO POPULATIONS

Independent Samples: Example

mice.xls

outliers are removed from boxplots



Q1: Is body weight for male and female significantly different?

Q2: Is weight change for male and female significantly different?

Q3: Is bleeding time for male and female significantly different?

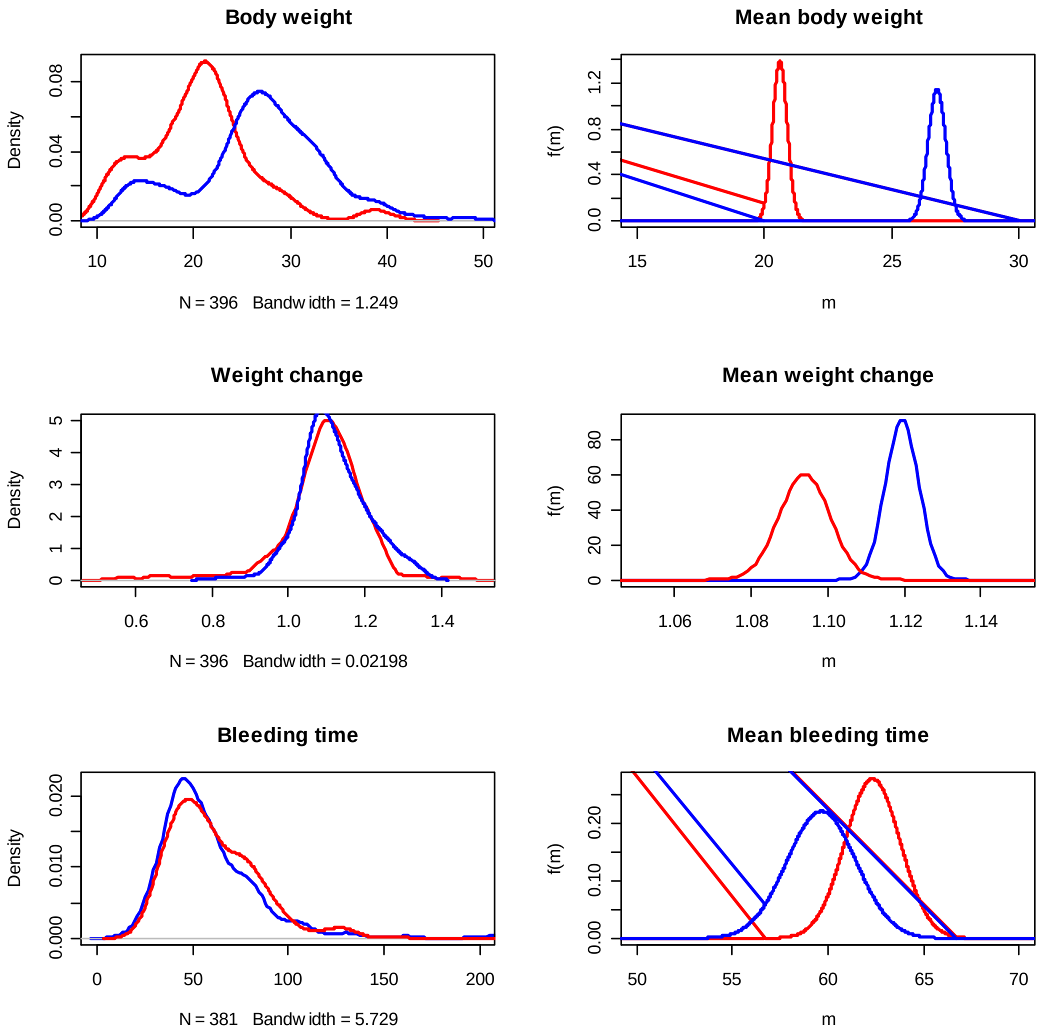
MEANS OF TWO POPULATIONS

Example

Q1: Is body weight for male and female significantly different?

Q2: Is weight change for male and female significantly different?

Q3: Is bleeding time for male and female significantly different?



HYPOTHESES

Theory

Two tail hypothesis

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

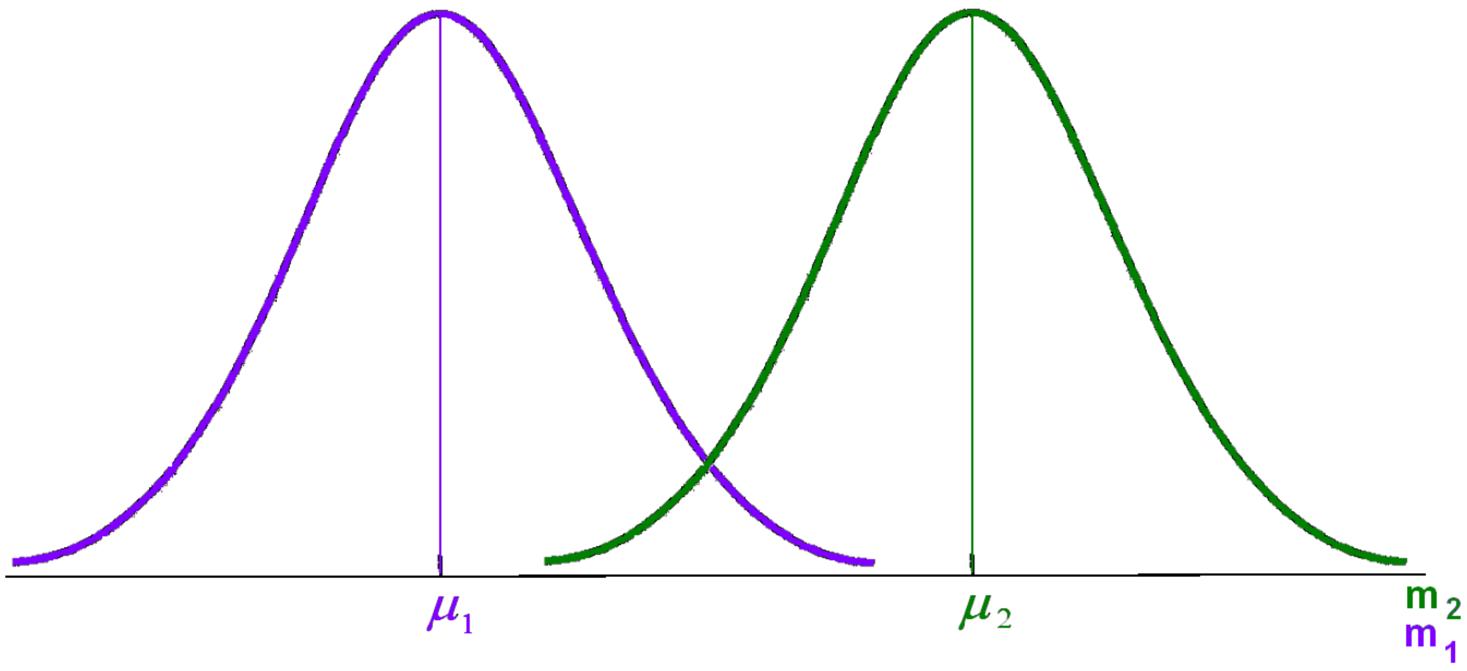
One tail hypothesis

H_0: $\mu_1 \geq \mu_2$

H_a: $\mu_1 < \mu_2$

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$



COMPARING MEANS

Theory

As we know how to work with standard hypotheses (comparison with constant μ_0), let us transform our hypothesis:

$H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

⇒

$H_0: \mu_2 - \mu_1 = 0$

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

To use it, we need to know what is the distribution of $D = m_2 - m_1$

Distribution of sum or difference of 2 normal random variables

The sum/difference of 2 (or more) normal random variables is a normal random variable with **mean equal to sum/difference** of the means and **variance equal to SUM** of the variances of the compounds.

Variables	m_1	m_2	$m_2 - m_1$
Means	μ_1	μ_2	$\mu_2 - \mu_1$
Variances	σ_1^2	σ_2^2	$\sigma_1^2 + \sigma_2^2$

COMPARING MEANS

Theory

H_0: $\mu_2 - \mu_1 = D_0$

H_a: $\mu_2 - \mu_1 \neq D_0$

$$D_0 = \mu_2 - \mu_1$$

$$\sigma_{m_2 - m_1} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

⇒

$$D_0 = m_2 - m_1$$

$$s_{m_2 - m_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Statistics to be used for hypothesis testing:

if σ is known: z-statistics

$$z = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

if σ is unknown: t-statistics

$$t = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This is what we call a t-test !!!

COMPARING MEANS

Unpaired t-test: Algorithm

H_0: $\mu_2 - \mu_1 = D_0$

H_a: $\mu_2 - \mu_1 \neq D_0$

$$D_0 = m_2 - m_1$$

Usually $D_0 = 0$

$$s_{m_2 - m_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

1. Build the statistics to be used for hypothesis testing:

$$t = \frac{m_2 - m_1 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-distribution has following degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$df = (n-1)\frac{\left(s_1^2 + s_2^2\right)^2}{\left(s_1^4 + s_2^4\right)}$$

$(n_1+n_2)/2 < df < n_1+n_2$

2. Calculate the p-value

◆ `= TDIST(ABS(t),df,2)`

☺. Or simply do:

!

```
= T.TEST (array1, array2, 2, 3)
```

```
t.test(x, y)
```

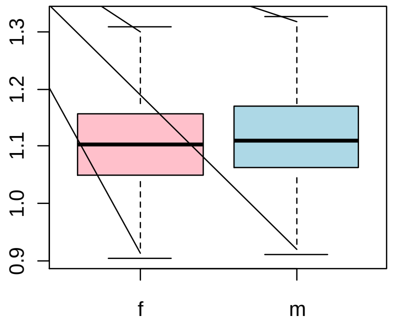
UNPAIRED T-TEST

Example

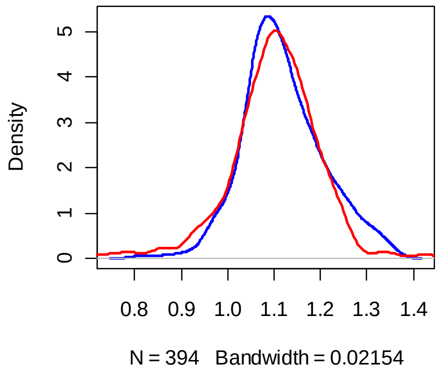
mice.xls

Q2: Is the mean of weight change for male and female significantly different?

Weights change (g)



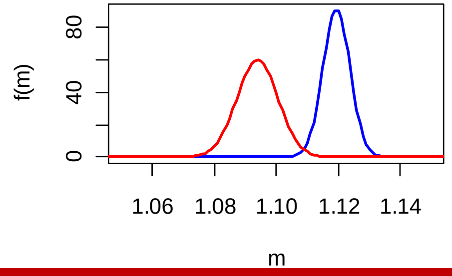
Distributions of weight change



Parameter	Value	Command
Female		
m1	1.0939	=AVERAGE($B$2:$B$397)
s1	0.1320	=STDEV($B$2:$B$397)
n1	396	=COUNTIF($B$2:$B$397,">=0")
Male		
m2	1.1194	=AVERAGE($B$398:$B$791)
s2	0.0867	=STDEV($B$398:$B$791)
n2	394	=COUNTIF($B$398:$B$791,">=0")
m2-m1	0.0255	=F8-F3
s(m2-m1)	0.0079	=SQRT(F4^2/F5+F9^2/F10)
t	3.2067	=F12/F13
df ~	700	
p-value (1)	0.001403856	=TDIST(F14,700,2)
p-value (2)	0.00140541	=TTEST($B$2:$B$397,$B$398:$B$791,2,3)

- **= T.TEST(data1, data2, 2, 3)**

p-value = 0.0014



PAIRED T-TEST

Theory and Example

Paired t-test

In a paired t-test, instead of testing H_0: $\mu_2 - \mu_1 = 0$, use following steps:

1. **Build a new random value $y = x_1 - x_2$ (subtract matched values).**
2. **Test whether one-sample mean $\mu_y = 0$ (see Lecture 6)**

```
bloodpressure
```

The systolic blood pressures of n=12 women between the ages of 20 and 35 were measured before and after usage of a newly developed oral contraceptive.

Systolic blood pressure (mmHg)

Subject	BP before	BP after
1	122	127
2	126	128
3	132	140
4	120	119
5	142	145
6	130	130
7	142	148
8	137	135
9	128	129
10	132	137
11	128	128
12	129	133

Q: Does the treatment affect the systolic blood pressure?

Test	p-value
unpaired	0.414662
paired	0.014506

Unpaired test

```
= T.TEST (array1, array2, 2, 3)
```

Paired test

```
= T.TEST (array1, array2, 2, 1)
```

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

COMPARING PROPORTIONS

Theory

$H_0: \pi_1 = \pi_2$

$H_a: \pi_1 \neq \pi_2$

$H_0: \pi_1 - \pi_2 = 0$

$H_a: \pi_1 - \pi_2 \neq 0$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Pooled estimator of π

An estimator of a population proportion obtained by computing a weighted average of the point estimators obtained from two independent samples.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\sigma_{p_1 - p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

```
= 2*NORM.S.DIST(-ABS(z),TRUE)
```

Exact test:

```
prop.test(x = c(x1,x2),
          n = c(n1,n2))
```

COMPARING PROPORTIONS

Example

SWR/J	MA/MyJ
f	f
f	f
f	f
f	f
f	f
f	f
f	f
f	f
f	m
f	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
m	m
	m
	m
	m
	m

mice.xls

Q: Is the male proportion significantly different in these mouse strains (0.47 and 0.65)?

Pooled proportion

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

if we apply directly:

Gives less precise estimation

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

```
= 2*NORM.S.DIST(-ABS(z),TRUE)
```

	SWR/J	MA/MyJ	*pooled*
count male	9	15	*24*
n	19	23	*42*
p	0.474	0.652	*0.571*
z	-1.16		
p-val	**0.244658997**		

p-value = 0.24

Exact test:

```
prop.test(x = c(x1,x2),
          n = c(n1,n2))
```

Thank you for your attention