CS 554m

controlled experiments II

Joanna McGrenere

today: part I

learning goals:

what is an analysis of variance (ANOVA)?

what is the important terminology in ANOVA?

what are the different types of ANOVA?

when would one choose to use an ANOVA?

what is the difference between statistical and practical significance?

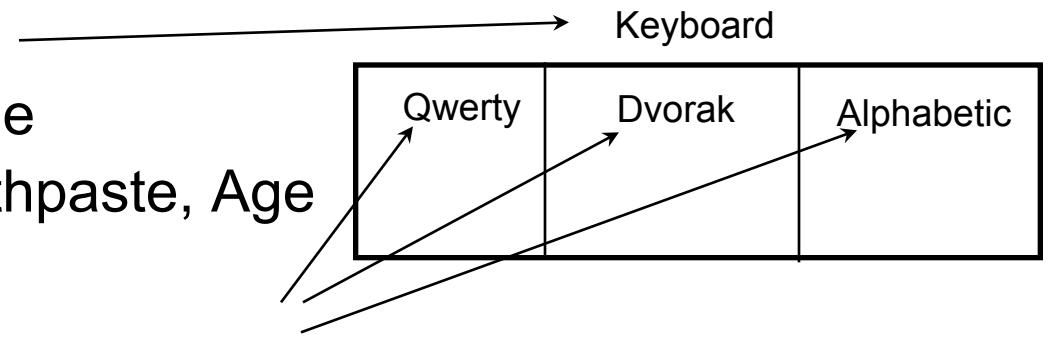
analysis of variance (ANOVA)

a workhorse

- allows moderately complex experimental designs (relative to t-test)

terminology

- factor
  - independent variable
  - i.e., Keyboard, Toothpaste, Age



- factor level
  - specific value of independent variable
  - i.e., Qwerty, Crest, 5-10 years old

ANOVA terminology

between subjects

- a subject is assigned to only one factor level of treatment
- problem: greater variability, requires more subjects

Keyboard

Qwerty	Dvorak	Alphabetic
S1-20	*S21-40*	*S41-60*

within subjects

- subjects assigned to all factor levels of a treatment
- requires fewer subjects
- less variability as subject measures are paired
- problem: order effects (e.g., learning)
- partially solved by counter-balanced ordering

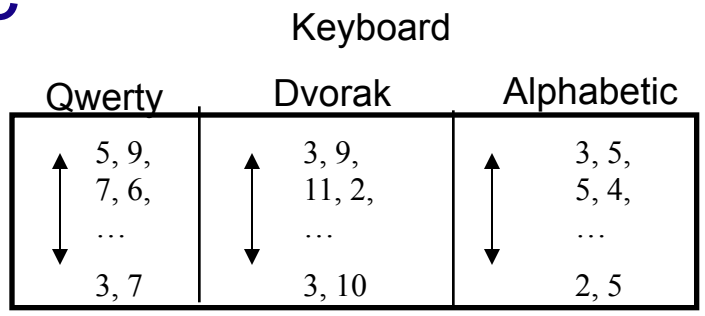
Keyboard

Qwerty	Dvorak	Alphabetic
S1-20	*S1-20*	*S1-20*

f statistic

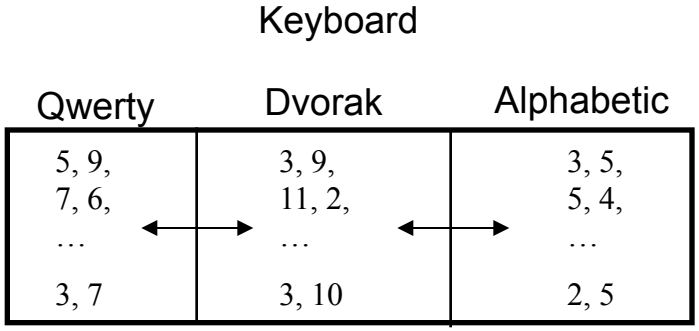
within group variability (WG)

1.
2.



between group variability (BG)

1.
2.
3.



these two variabilities combine to give total variability

we are mostly interested in _____ variability because we are trying to understand the effect of the treatment

f statistic

$$f = \frac{BG}{WG} = \frac{\text{treatment} + \text{id} + \text{m.error}}{\text{id} + \text{m.error}} = ?$$

= 1, if there are no treatment effects

> 1, if there are treatment effects

within-subjects design: the id component in numerator and denominator factored out, therefore a more powerful design

f statistic

similar to the t-test, we look up the f value in a table, for a given α and degrees of freedom to determine significance

thus, f statistic is sensitive to sample size

- Big N ⟶ Big Power ⟶ Easier to find significance
- Small N ⟶ Small Power ⟶ Difficult to find significance

what we (should) want to know is the effect size

- does the treatment make a big difference (i.e., large effect)?
- or does it only make a small difference (i.e., small effect)?
- depending on what we are doing, small effects may be important findings

statistical significance vs *practical* significance

when *N* is large, even a trivial difference (small effect) may be large enough to produce a statistically significant result

- e.g., menu choice:
  mean selection time of menu A is 3 seconds;
  menu B is 3.05 seconds

statistical significance does not imply that the difference is important!

- a matter of interpretation, i.e., subjective opinion
- should always report means to help others make their opinion

there are measures for effect size, regrettably they are not widely used in HCI research

single factor analysis of variance

compare means between two or more factor levels within a single factor

e.g.:

- dependent variable: typing speed (time)
- independent variable (factor): keyboard
- between subject design

also called a one-way ANOVA

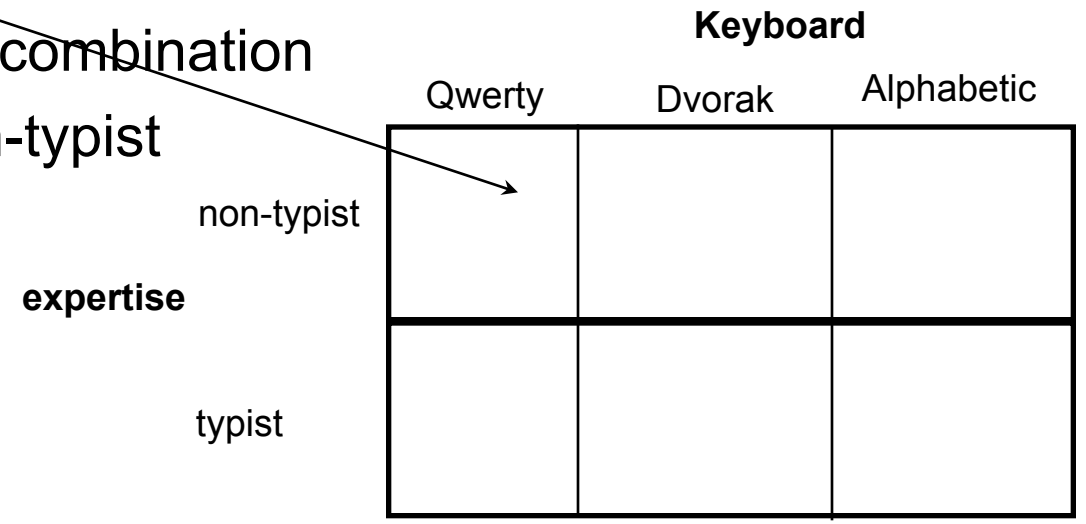
Qwerty	Alphabetic	Dvorak
S1: 25 secs S2: 29 … S20: 33	S21: 40 secs S22: 55 … S40: 33	S51: 17 secs S52: 45 … S60: 23

ANOVA terminology

- factorial design
  - cross combination of levels of one factor with levels of another
  - e.g., keyboard type (3) x expertise (2)

2-way factorial ANOVA

- cell
  - unique treatment combination
  - e.g., qwerty x non-typist



ANOVA terminology

mixed factor (called split-plot in Lazar reading)

- contains both between and within subject combinations

		Keyboard		
		Qwerty	Dvorak	Alphabetic
expertise	non-typist	*S1-20*	*S1-20*	*S1-20*
	typist	*S21-40*	*S21-40*	*S21-40*

ANOVA

compares the relationships between many factors

provides more informed results

- considers the interactions between factors
- e.g.,
  - typists type faster on Dvorak, than on alphabetic and Qwerty
  - non-typists are fastest on alphabetic

		Keyboard		
		Qwerty	Dvorak	Alphabetic
expertise	non-typist	*S1-20*	*S1-20*	*S1-20*
	typist	*S21-40*	*S21-40*	*S21-40*

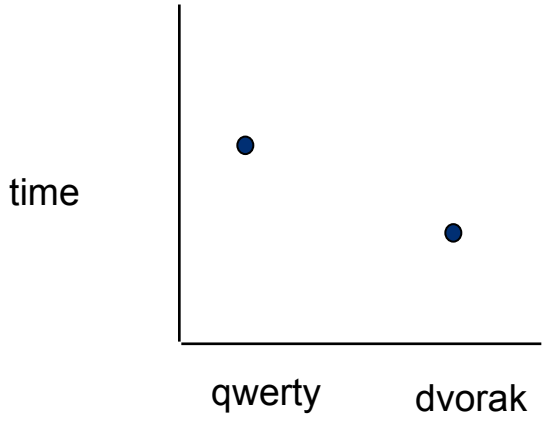
ANOVA

in reality, we can rarely look at one variable at a time

example:

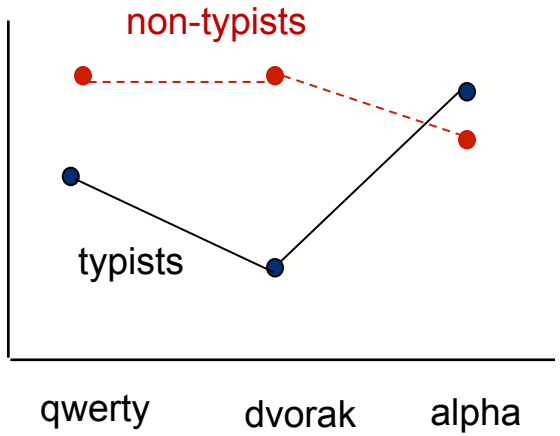
- t-test:

  subjects faster on dvorak than qwerty



- anova: keyboard x expertise

  alphabetic fastest for non-typists
  dvorak fastest for typists



ANOVA case study

WIMP (GUI) vs. HYBRID (graphical command line)

motivation:

- WIMP interfaces are slow because of the mouse
- can we create a hybrid interface that is graphical but can be fully operated through the keyboard? (sort of like a command line)
- assume that one has been designed
- how should it be evaluated?

ANOVA case study

WIMP (GUI) vs. HYBRID (graphical command line)

independent variables:

- interface: WIMP, hybrid
- expertise: novice, expert
- command parameters: zero, one, two
  - E.g., bold (zero), font ariel (one), print –copies 2 –color greyscale (two)
  - Note: zero parameter commands can be done using shortcuts keys in WIMP

dependent variables:

- performance: speed, error
- satisfaction

ANOVA case study

possible hypotheses:

H1: experts will perform better than novices (not that interesting)

H2: novices will perform better with WIMP than hybrid

H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters

2 level (interface) x
2 level (expertise) x
3 level (parameters)

mixed factor design

		WIMP	hybrid
zero	novice	S1-8	S1-8
	expert	S9-16	S9-16
one	novice	S1-8	S1-8
	expert	S9-16	S9-16
two	novice	S1-8	S1-8
	expert	S9-16	S9-16

task

assume that the task is to enter a whole series of commands, one after the other

there is an equal number of 0, 1, and 2 parameter commands used

identical commands are used in both interface conditions

statistical results: speed

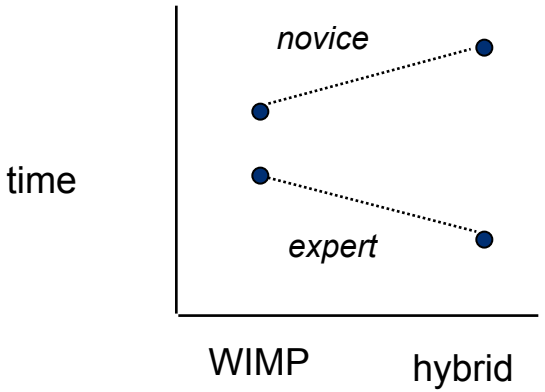
	F-ratio.	*p*	
Interface (I)	0.4		
Expertise (E)	5.5*	<0.05	main effects
Parameters (P)	31.0**	<0.01	
IxE	15.2*	<0.05	
IxP	8.0*	<0.05	interactions
ExP	5.0		
IxExP	14.1*	<0.05	

main effect: the effect of the variable collapsing across all levels of other variables in the experiment

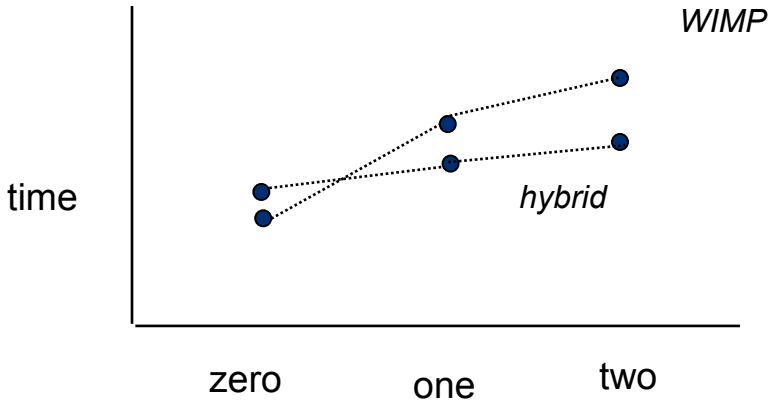
interaction effect: the effect of one variable differs depending on the level of another (other) variable(s)

statistical results: speed (time)

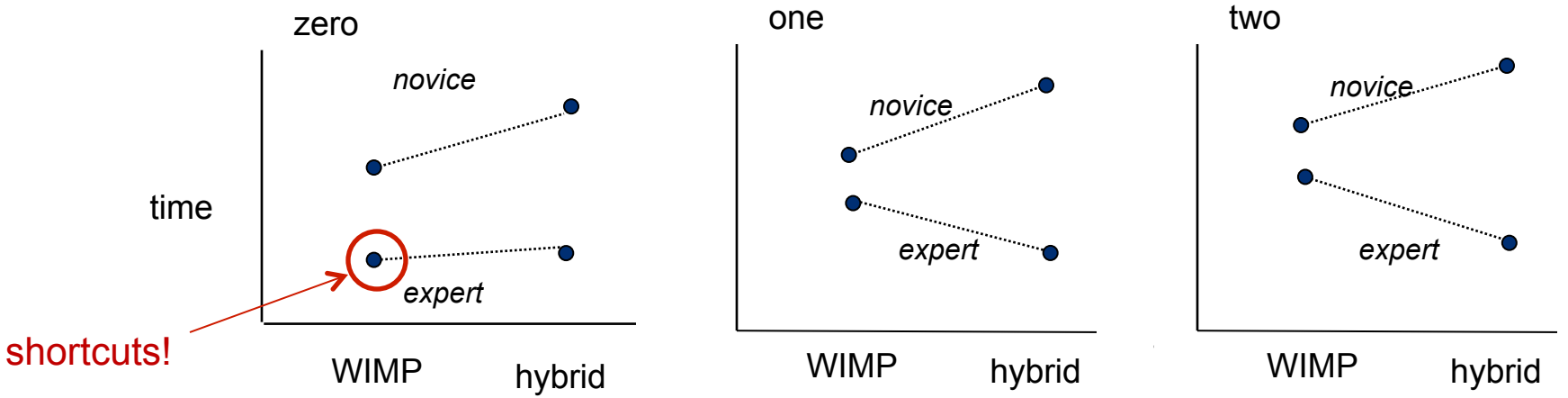
Interface x Expertise (IxE)



Interface x Parameters (IxP)



Interface x Expertise x Parameters (IxExP)



summary of results

Assuming same results for errors as speed…

H1: experts will perform better than novices (not that interesting)

Supported: main effect of expertise, showing experts better

H2: novices will perform better with WIMP than hybrid

Supported: 2-way interaction effect of interface and expertise, showing novices overall better with WIMP

H3: experts will perform better with hybrid than WIMP, but only for commands with one or more parameters

Supported: 3-way interaction effect of interface, expertise, and number of parameters, showing experts better with hybrid, but only with one and two parameters

case study conclusions

- expertise makes a big difference
- WIMP interaction should be kept for novices
- hybrid interaction should be available for experts

part I – you now know

there are many statistical methods that can be applied to different experimental designs

- t-tests
- single factor ANOVA
- factorial ANOVA (case study)

ANOVA terminology

- factors, levels, cells
- factorial design
  - between, within, mixed designs

difference between statistical and practical significance

part II

learning goals

significance levels and two types of error

- what is the difference between a type I and type II error?
- how does choice of significance levels relate to error types?
- how do I chose a significance level?

other tests: what are correlation & regression?

choice of significance levels and two types of errors

Type I error: reject the null hypothesis when it is, in fact, true ($\alpha = .05$)
Type II error: accept the null hypothesis when it is, in fact, false (β)

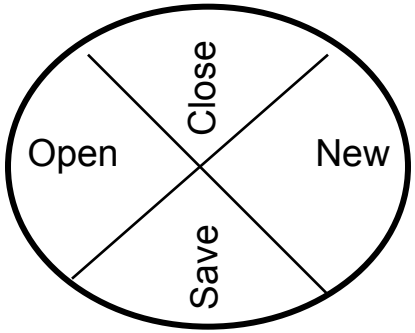
	H_0 True	H_0 False
Reject H_0	α (Type I error)	1 - β (Power)
Not Reject H_0	1 - α	β (Type II error)

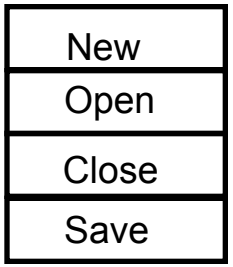
Effects of levels of significance

- very high confidence level (eg .0001) gives greater chance of Type II errors
- very low confidence level (eg .1) gives greater chance of Type I errors
- tradeoff: choice often depends on effects of result

choice of significance levels and two types of errors

H_0 There is no difference between Pie menus and traditional pop-up menus





Type I: (reject H_0, believe there is a difference, when there isn't)

- outcome?

Type II: (accept H_0, believe there is no difference, when there is)

- outcome?

choice of significance levels and two types of errors

Type I: (reject H_0, believe there is a difference, when there isn't)

- extra work developing software and having people learn a new idiom for no benefit

Type II: (accept H_0, believe there is no difference, when there is)

- use a less efficient (but already familiar) menu

Case 1: Redesigning a traditional GUI interface

Case 2: Designing a digital mapping application where experts perform extremely frequent menu selections

other tests: correlation

measures the extent to which two concepts are related

- e.g., years of university training vs tablet computer ownership per capita

how?

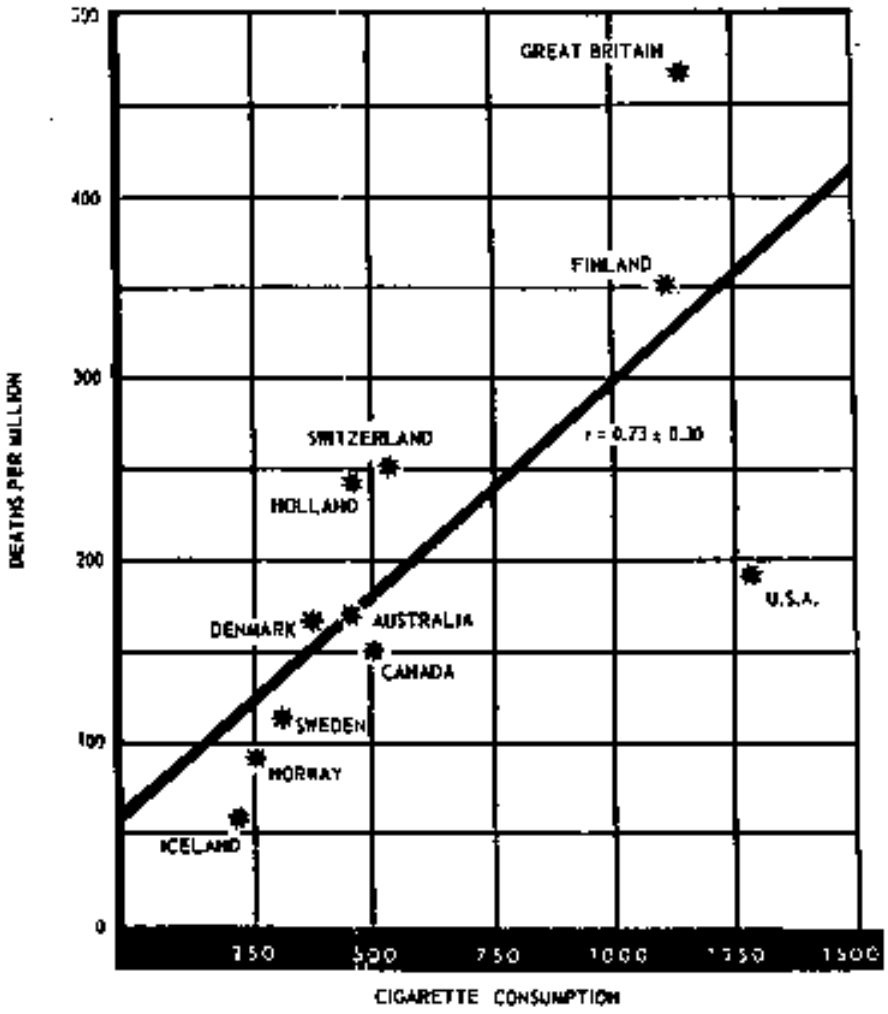
- obtain the two sets of measurements
- calculate correlation coefficient
  - +1: positively correlated
  - 0: no correlation (no relation)
  - –1: negatively correlated

dangers

- attributing causality
  - a correlation does not imply cause and effect
  - cause may be due to a third “hidden” variable related to both other variables
  - e.g., (above example) age, affluence
- drawing strong conclusion from small numbers
  - unreliable with small groups
  - be wary of accepting anything more than the direction of correlation unless you have at least 40 subjects

non-HCI sample study: cigarette consumption

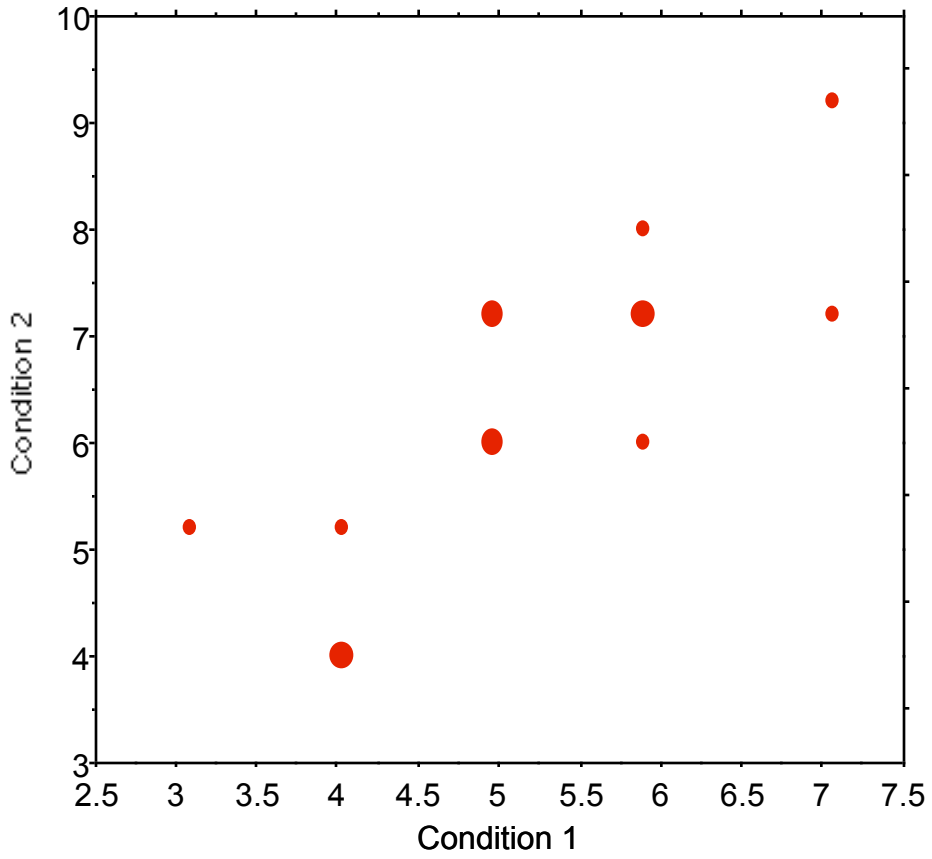
crude Male death rate for lung cancer in 1950 per capita consumption of cigarettes in 1930 in various countries



correlation

$r^2 = .668$

condition 1	condition 2
5	6
4	5
6	7
4	4
5	6
3	5
5	7
4	4
5	7
6	7
6	6
7	7
6	8
7	9



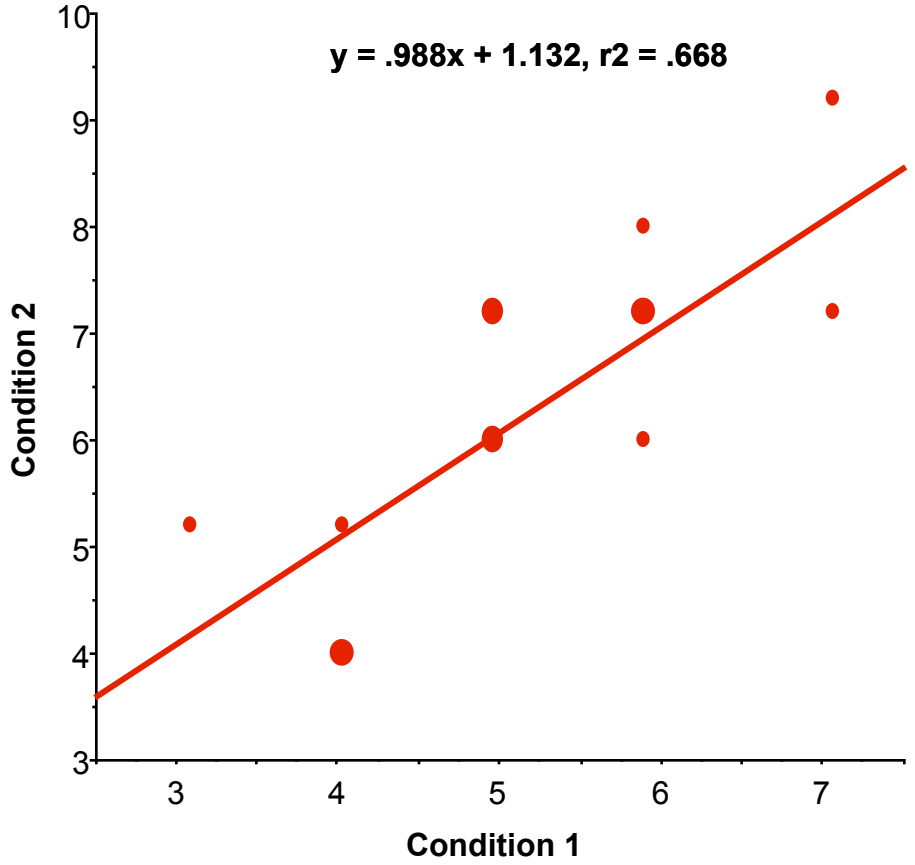
regression

calculate a line of “best fit”

use the value of one variable to predict the value of the other

- e.g., 60% of people with 3 years of university own a tablet computer

condition 1	condition 2
5	6
4	5
6	7
4	4
5	6
3	5
5	7
4	4
5	7
6	7
6	6
7	7
6	8
7	9



now you know…

significance levels and two types of error

- the difference between a type I and type II error
- how the choice of significance levels relates to error types
- how to chose a significance level based on the implications of error types

correlation and regression