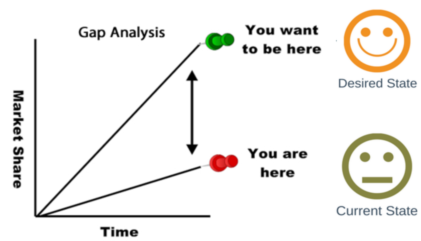
Gap Analysis and Statistical Hypothesis Testing



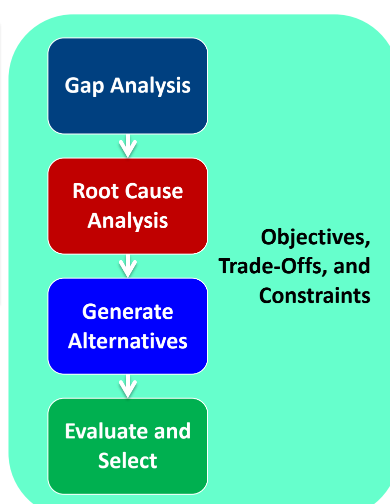
From BUS 462 (Business Analytics) at SFU
Amin Milani Fard - Fall 2018

Some slides are from Michael Brydon

1

Gap Analysis

Gap Analysis
- Is there a problem?
- What is the gap between expected performance and actual performance?



2

Statistical Variability

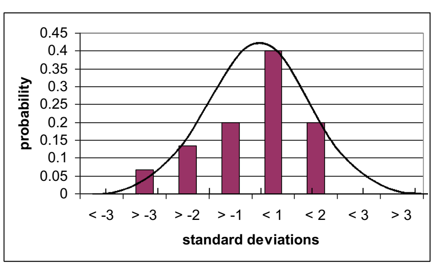
Store	Sales
1	$ 8,617.63
2	$ 10,249.81
3	$ 11,905.92
4	$ 11,021.00
5	$ 12,193.02
6	$ 6,387.60
7	$ 11,851.34
8	$ 13,442.58
9	$ 7,975.78
10	$ 9,787.87
11	$ 10,958.20
12	$ 13,326.81
13	$ 13,120.29
14	$ 9,751.82
15	$ 10,925.57

Sagging sales: Chance or incompetence?

- consider the sales data
- is there a "problem" with Store 6?
- what about Store 9?

3

Variable Processes



$$f(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)}$$

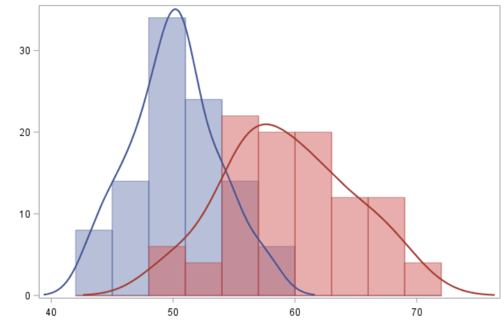
- The store data results from a process with "natural" variation
  - Normally distributed data with μ = $10,000 and σ = $2000
  - $P(s<6387) = 1.5\%$
  - $P(s<7975) = 8.5\%$
- Errors:
  - Type I: asserting a gap exists when the difference is due to chance
  - Type II: asserting no gap exists when one actually does

4

4

t-test for Gap Analysis

- Two samples of data (A and B)
- Variability
- Hypothesis: the two samples have the **same sample mean**



$$t\ statistic = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{\mathrm{var}_A}{n_A} + \frac{\mathrm{var}_B}{n_B}}}$$

5

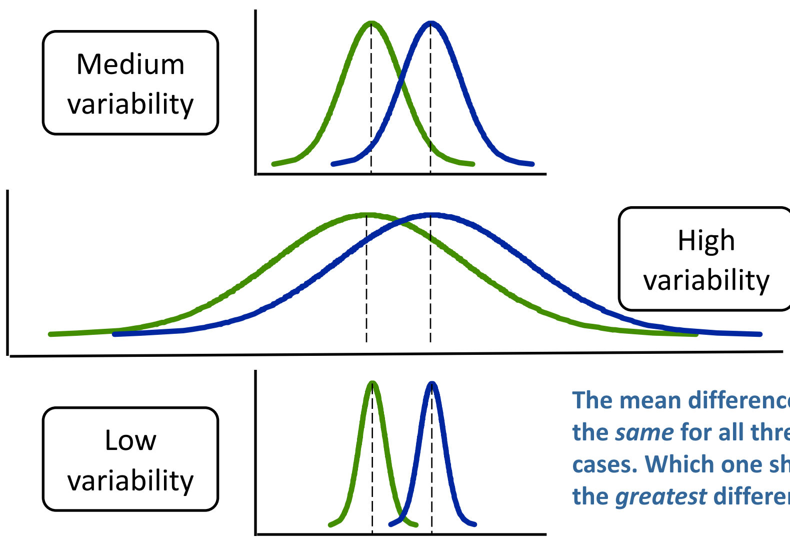
t-test for Gap Analysis

- Tests whether the means of two groups are *statistically* different from each other

- If 2 means are statistically different, then the samples are likely to be drawn from 2 different populations, i.e. they really are different.

6

What Does *Difference* Mean?

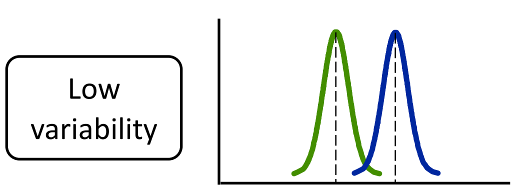


The mean difference is the *same* for all three cases. Which one shows the *greatest* difference?

7

What Does *Difference* Mean?

- A statistical difference is a function of the _difference between means_ relative to the _variability_.
- A small difference between means with large variability could be due to _chance_. Like a *signal-to-noise* ratio.

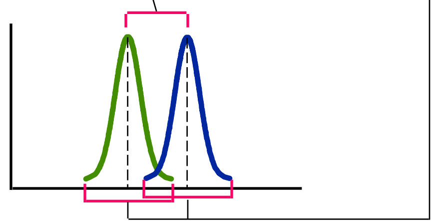


Which one shows the *greatest* difference?

8

What Do We Estimate?

$$\frac{\text{Signal}}{\text{Noise}} = \frac{\text{Difference between group means}}{\text{Variability of groups}}$$



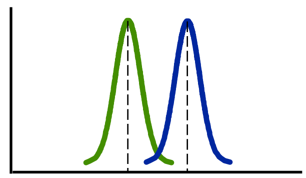
9

What Do We Estimate?

$$\frac{\text{Signal}}{\text{Noise}} = \frac{\text{Difference between group means}}{\text{Variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

= t-value



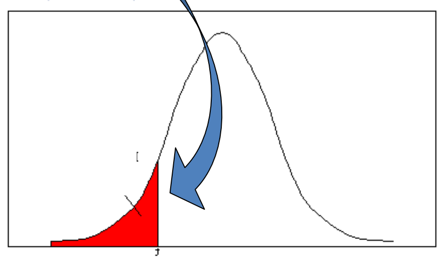
10

Bank Data *t*-test

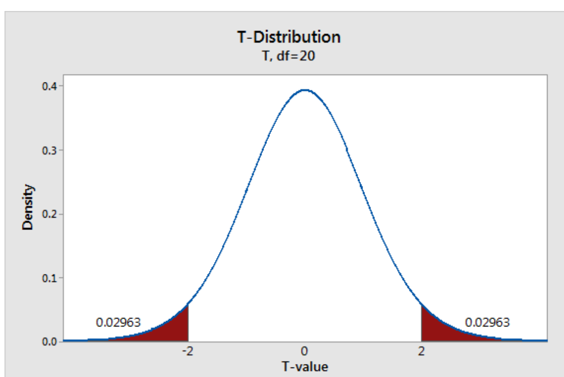
t-Test: Two-Sample Assuming Unequal Variances		
	Female	*Male*
Mean	37.20993	45.50544
Variance	45.03573	251.0076
Observations	140	68
Hypothesized Mean Difference	0	
df	79	
t Stat	-4.14105	
P(T<=t) one-tail	4.3E-05	
t Critical one-tail	1.664371	
P(T<=t) two-tail	8.59E-05	
t Critical two-tail	1.	

P value indicates the probability that two samples come from the same population

$$t\ statistic = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{\mathrm{var}_A}{n_A} + \frac{\mathrm{var}_B}{n_B}}}$$



11



The probability distribution plot indicates that each of the two shaded regions has a probability of 0.02963—for a total of 0.05926.

This graph shows that t-values fall within these areas almost 6% of the time when the null hypothesis is true (P value)

12

12

T-test Summary

Use when:
- Sampled data is continuous
- Samples are representative and large (>30)

Three types

Paired:

1. data point in sample A matches a data point in sample B

Unpaired:

2. Assuming equal variance
3. Assuming unequal variance

13

Continuous vs Categorical Samples

Continuous Data

- Numbers
- Summary measures
  - Sales
  - Opinion scores
  - Number of crimes

Categorical Data

- Frequencies of category membership
  - Polls (Candidate A or Candidate B?)
  - Purchase/Non-Purchase
  - Pay/Default

14

Recall: Gender Discrimination

- The charge is that its female employees receive substantially smaller salaries than the firm's male employees
- If we **censor** management positions, the salary gap disappears…

A	B	C	D	E	F	G	H	I
Bank salary data								
Employee	EducLev	JobGrade	YrHired	YrBorn	Gender	YrsPrior	PCJob	Salary
1	3	1	92	69	Male	1	No	32
2	1	1	81	57	Female	1	No	39.1
3	1	1	83	60	Female	0	No	33.2
4	2	1	87	55	Female	7	No	30.6
5	3	1	92	67	Male	0	No	29
6	3	1	92	71	Female	0	No	30.5
7	3	1	91	68	Female	0	No	30
8	3	1	87	62	Male	2	No	27
9	1	1	91	33	Female	0	No	34
10	3	1	86	64	Female	0	No	29.5
206	5	6	63	33	Male	0	No	88
207	5	6	60	36	Male	0	No	94
208	5	6	62	33	Female	0	No	30

15

Categorical Data

Is there a gap between the **expected** level of female participation in management and the **observed** level?

Count of Employee	Column Labels		
Row Labels	Mgmt	Non-Mgmt	Grand Total
Female	10	130	140
Male	25	43	68
Grand Total	**35**	**173**	**208**

16

Statistical Dependence and Independence

		Preferred Brand of Shoes			
		Reebok	Nike	Adidas	
Grade in Course	A	2	3	3	8
	B	6	6	8	20
	C	2	3	4	9
		10	12	15	37

$P(A \cap B) = P(A) \cdot P(B)$

$P(\text{Grade} = \text{A} \cap \text{Reebok}) = P(\text{Grade} = \text{A}) \cdot P(\text{Reebok})$

$= \frac{8}{37}(\frac{10}{37}) = 0.058$

$E(\text{Grade} = \text{A} \cap \text{Reebok}) = P(\text{Grade} = \text{A} \cap \text{Reebok})n = 0.058(37) = 2.16$

Assumption of independence seems okay

17

Statistical Dependence and Independence

		Experience with Statistics			
		Low	Med	High	
Grade in Course	A	0	1	7	8
	B	2	10	8	20
	C	8	1	0	9
		10	12	15	37

$P(A \cap B) = P(A) \cdot P(B)$

$P(\text{Grade} = \text{A} \cap \text{Low Stats}) = P(\text{Grade} = \text{A}) \cdot P(\text{Low Stats})$

$= \frac{8}{37}(\frac{10}{37}) = 0.058$

$E(\text{Grade} = \text{A} \cap \text{Low Stats}) = P(\text{Grade} = \text{A} \cap \text{Low Stats})n = 0.058(37) = 2.16$

Assumption of independence gives a poor prediction

18

There is a difference but is that significant?

		Experience with Statistics			
		Low	Med	High	
Grade in Course	A	0	1	7	8
	B	2	10	8	20
	C	8	1	0	9
		10	12	15	37

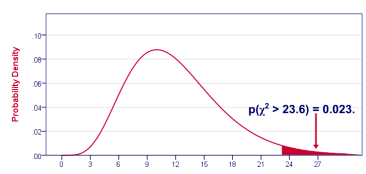
$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(\text{Grade} = \text{A} \cap \text{Low Stats}) = P(\text{Grade} = \text{A}) \cdot P(\text{Low Stats})$$
$$= \frac{8}{37}\left(\frac{10}{37}\right) = 0.058$$
$$E(\text{Grade} = \text{A} \cap \text{Low Stats}) = P(\text{Grade} = \text{A} \cap \text{Low Stats})n = 0.058(37) = 2.16$$

19

Assessing Independence

- χ^2 test of independence:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$



Larger value -> smaller prob of independence

- Assuming variables are independent (**expected**):

$$P(A \cap B) = P(A) \times P(B)$$

- Our data provides **observed** joint frequency

- Key idea: variables may be *dependent* in many different ways, but they can be *independent* in only one way

20

Parametric Test	Goal of the test	Non-Parametric Test	Goal of the test
Two Sample T-Test	To see if two samples have identical population means	Wilcoxon Rank-Sum Test	To see if two samples have identical population medians
One Sample T-Test	To test a hypothesis about the mean of the population a sample was taken from	Wilcoxon Signed Ranks Test	To test a hypothesis about the median of the population a sample was taken from
Chi-Squared Test for Goodness of Fit	To see if a sample fits a theoretical distribution, such as the normal curve	Kolmogorov-Smirnov Test	To see if a sample could have come from a certain distribution
ANOVA	To see if two or more sample means are significantly different	Kruskal-Wallis Test	To test if two or more sample medians are significantly different

A parametric test requires a parametric assumption, such as normality. A nonparametric test does not rely on such parametric assumptions.

21