

## Statistics

**Constant** – a number that is not changing (according to user's wishes)

- '*Kavua*'

**Variable** – a unit that changes according to what people plug in (i.e. x)

- *Mishtane*

**Dependant variable** variable that changes dependant on another number in the equation

- *Mishtane talui*

**Independent variable** a variable that does not necessarily change depending on other variables

- *Mishtane bilti talui*

Example:  $Y=f(X)$

Dependent: Y

Independent: X

**Qualitative**- something relating to a quality of something – not #'s

**Quantitative**- something relating to #

**Continuous** – a variable that has subunits: i.e. 160cm has subunits (i.e. 160.5)

- *Ratzif*

**Discrete** – a variable w/o subunits: i.e. # of people in class – there can be 13 people in the class but there can't be 13.5 people in the class

- *Badid*

**Infinite**-einsofi

**Finite** -sofi

**Population** the full group being tested on

**Sample** the random group from the population chosen to collect data

**Graphs** tables that portray #'s in an arranged order

**Histograms** -frequencies are shown in form of bars

**Polygram** –a line graph

## **Oct. 31 tutorial**

-Statistics deals w/ the ways to describe and analyze empirical facts

-deals with cases of uncertainty/cases of partial uncertainty

-statistics divides into *descriptive* (teurit) and *conclusive* (hiskit) areas

Descriptive- the way to lay out the input/intaken data →SD/mean/frequency/etc

Conclusive – drawing conclusions on populations based on samples of that population

-generalizing a rule from data

→need descriptive in order to get conclusive.

### **Sample**

Sampling needs to be done correctly in order for it to truly represent the population

There are 2 basic criteria for good sampling

1) the individuals involved in collecting data have to be chosen ***RANDOMLY***

2) representative of all the population (i.e. has to account for ***ALL*** segments of the population)

***measurement*** – ‘madad’ – a statistic which can be used to draw conclusions from, that would be applicable to the society.

***Parameter*** – data which was collected from all the population – written w/ latin letters

***Statistic*** – collected from sample of population – written w/ English letters

### **Scales** – ‘Sulamot’

***Important:*** -In *scales*, the variables are *categories*; therefore,  $a \neq b$

#### 1) ***Nominal scale***

A variable which is not preferred over another → just tells you a fact/just defines categories

→ i.e. ID # → just assigns a fact to a variable → a person to a number.

→ the # has no consequence

#### 2) ***Ordinal*** – ‘Dirugi’

-the variables are in some sort of order.

Example: I like the following order:

#1 Coke

#2 Pepsi

#3 R.C.

-the # is not an existing # → you cannot measure preference numerically

→ the issue is not the difference but the order!

Example: If I want to accept the top 5 students, I do not care the difference b/w them but their ranking

-in any given *ordinal scale* has more than 5/7 variables, it is called an ‘***Improved scale***’  
‘**Improved Scale**’ is considered a ***Interval graph***

#### 3) ***Interval***

***student #1 get 90%***

***student #2 gets 80%***

-90% is 10% more than 80%

-but 90% is not necessarily double of 45%

-difference is set, but not the ratio → because 0 is not absolute → you can't know exactly 0% → if you get 0, it does not necessarily mean that you know *nothing*

-the difference b/w 45 and passing is different from 80 and 90!

-in any given *ordinal scale* has more than 5/7 variables, it is called an '**Improved scale**' '**Improved Scale**' is considered a *Interval graph*

#### 4) *Ratio*

**Ruler #1** = 4 meters

**Ruler #2** = 2 meters

**Ratio** = 2:1

Ruler # 2 is x2 Ruler #1

In nominal/Ordinal, # nor comparable, since no absolute 0  
→ here, since there is an absolute 0, the 2 variables are comparable

Scales→	NOMINAL	ORDINAL	INTERVAL	RATIO
<b>IDENTITY</b> a≠b	√	√	√	√
<b>ORDER</b> What is more/less	X	√	√	√
<b>DEFINITE DIFFERENCE</b> a-b	X	X	√	√
<b>RATIO</b> a/b	X	X	X	√
<b>TRANSFER</b> (manipulation)	Yes, as long as a≠b		Only in linear i.e. when you manipulate both variables	Multiplication of division

--

Some psych. think psych is ordinal

most think there's measurable difference to diff. People -proportional

-when giving an questionnaire, some psych like to give linear graph – shows more accurately difference b/w +/- feelings

### Distribution

#### IQ of 6 imaginary people

124 118 123 109 117 114

-to analyze info, we need to distribute it (usually in chart)

<b>Class interval</b>	<b>F (Frequency)</b>	<b>CF (Cumulative Frequency)</b>
125-129	0	6
120-124	2	6
115-119	2	4
110-114	1	2
105-109	1	1

**Q)** H.M. intervals do I need

**A)** Absolutely, no more than 20 –Recommended: no more than 10

**\*Preferable intervals:** 5;10;20;50;100

**\*on chart, put #s in descending order**

<b>Class interval</b>	<b>Real Interval/ Exact interval</b>
125-129	124.5-129.5
120-124	119.5-124.5
115-119	114.5-119.5
110-114	109.5-114.5
105-109	104.5-109.5
Discrete	Continuous

-Depending on what's being measured, there might be a subunit. In case you can't measure the subunit, the intervals in the 'real interval' column apply

### 2 kinds of experiments

**1)Experimental:** researchers manipulate the independent variable and compares it's effects on dependant variables

Example:

->If researchers want to see effect on LSD on mood, he gives 3 groups  
LSD/Prozac/nothing to see how different independent variables effect group (dep. variable)

**2) Observational/correlation**

→When researchers can't decide which is indep/dep variable

→Experiment: If women have lower IQ scores, we do not know if the ind/dep variable is gender/environment influence etc

**Nov. 13, 2000**

In a cumulative graph, the line never goes down, since the variable is cumulative: the variable is really the addition of all the variables up to and including the new one.

**$\Sigma$ -Sigma**

I=n

$$\Sigma X_i$$

i=1

The I is the variable you will add. Therefore i=1 is the first variable  
The N represents the last variable that you will add

**Example**

i=4

$$\Sigma X_i$$

i=1

x1=4 x2=5 x3=6 x4=10	Therefore, i=4 $\Sigma X_i$ 4+5+6+10=25 i=1
-------------------------------	--

$$\Sigma c = nc$$

c is a constant

**Example**

i=5

$$\Sigma 4$$

i=1

$$4+4+4+4+4$$

(5x4, because, the diff. b/w bottom and top variables is 5)

$$\Sigma cX_i = c \Sigma X_i$$

**Exercise**

$$X_1=1$$

$X_2=2$ $X_3=3$  $Y_1=7$ $Y_2=8$ $Y_3=10$ $C=1$	
$\Sigma(x-c)$	$(1-1) + (2-1) + (3-1)$ =3
$\Sigma(CY_1)$	$(1 \times 7) + (1 \times 8) + (1 \times 10)$ =25
$\Sigma XY$	$(1 \times 7) + (2 \times 8) + (3 \times 10)$ = 53
$(\Sigma X)(\Sigma Y)$	$(1+2+3) \times (7+8+10)$ = 6x25 =150
$\Sigma X^2$	$1^2 + 2^2 + 3^2$ =1+4+9 =14
$(\Sigma X)^2$	$(1+2+3)^2$ =6 <sup>2</sup> =36

Day	Journeys (x)	Food (y)
Sunday	10	100
Mon.	10	80
Tues	20	100

#### Rule #1

n

$$\Sigma(X_i + Y_i) = \Sigma X_i + \Sigma Y_i$$

I=1

#### Rule #2

$$\Sigma(X_i - Y_i) = \Sigma X_i - \Sigma Y_i$$

#### Rule #3

N

n

$$\Sigma C x_i = C \Sigma X_i$$

I=1                      I=1

**Example:** if above sum was in US\$ which is 4 shekels; the constant is 4  
 $(40 \times 10) + (4 \times 10) + (4 \times 20) = 160$   
 $= 4(10 + 10 + 20)$

**Rule #4**

n

$$\sum_{i=1}^n C_i = nC$$

I=1

**Example** (there are 3 days)  $N=3$ ;  $c=4$   
 $\rightarrow 4 \times 4 \times 4 = 64$

**Rule #5**

n

$$\sum_{i=1}^n X_i Y_i \neq \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$$

I=1

**Rule #6**

N

$$\sum_{i=1}^N X_i / Y_i \neq \sum_{i=1}^N X_i / \sum_{i=1}^N Y_i$$

I=1

**Rule #7**

N

$$\sum_{i=1}^N (x_i)^2 \neq (\sum_{i=1}^N x_i)^2$$

I=1

**Rule #8**

N

$$\sum_{i=1}^N (C + X_i) \neq C + \sum_{i=1}^N X_i$$

I=1

**Double Sigma**

$\rightarrow$  tells what way you add the variables

if  $\sum X_j$  is b/f  $\sum X_i$ , you do the sideways; if I is b/f j, you do the addition downwards

I↓ J→	1	2	3	...k
1	100	110	90	
2	82	93	115	

3				
4				
...N				

-k= end of groups

-n – end of the # of participant w/I each group

k n

$$\sum_{j=1}^k \sum_{I=1}^n X_{ji} = X_{11} + X_{12} \dots X_{1k} + X_{21} \dots X_{2k}$$

### **Central tendencies:**

-ways to describe the middle:

1)mean

2)median

3)mode

### **MEAN**

Average = mean

Two formulas to write the mean

1)

$$\frac{\sum X}{n}$$

2)

$$\frac{\sum f_i X_i}{\sum f_i}$$

### **Explanation of formula 2**

i.e. variables: 2,4,4,6

$X_i$	$f_i$
2	1
4	2
6	1



-->  $\sum f_i = N$

-->

$$\frac{\sum f_i X_i}{\sum f_i}$$

$\bar{X}$	-average of a sample -called mu symbol: $\mu$
-----------	---

\*If you take off average from each of the # and then square them respectively it will get you the smallest sum. Any other manipulation to the #'s b/f squaring will get a higher total sum of the individual #'s.

i.e. 70;80;90 → average=80

$$(70-80)^2 = 100$$

$$(80-80)^2 = 0$$

$$(90-80)^2 = 100$$

$$100+0+100=200$$

### **Median**

The # that splits the series of #'s in  $\frac{1}{2}$

i.e. 1;3;6;7;8;9;10

#### **In case of even # of #'s**

70;70;↓85;90

→the median would be half way b/w 2<sup>nd</sup> and 3<sup>rd</sup> #  
(in this case, 77.5)

In case where there are an even # of # and the majority of them (aspecially along the  $\frac{1}{2}$  way mark is 1 #

i.e.

70;70;70;90

We take the exact values

70	69.5-70.5
70	69.5-70.5

70	69.5-70.5
90	89.5-90.5

-Median is 2/3 of the way through the identical # (b/w the 2<sup>nd</sup> and 3<sup>rd</sup> 70)  
→ median is 2/3 of 69.5→70.5  
→median=70.17

### Example 2

69	59.5-60.5
70	69.5-70.5
70	69.5-70.5
70	69.5-70.5

-Median is 1/3 of the way through the identical # (b/w the 1<sup>st</sup> and 2<sup>nd</sup> 70)  
→ median is 1/3 of 69.5→70.5  
→median=69.67

**Signs for Median:** Md/Mdn/Mn (mn could also be mean)  
→though nothing official – researcher must define it b/f using it →no consensus

### Mode

Most common #

70;90;90 →mode=90

**Bimodal** – 2 modes

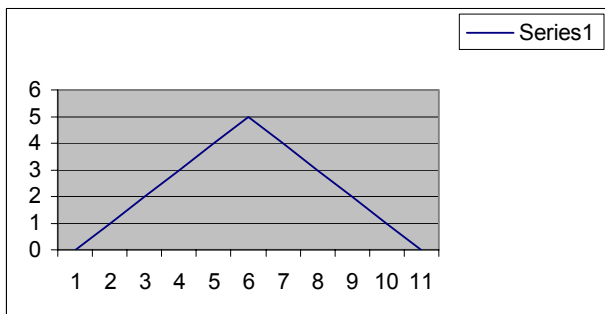
70;70;91;91;92

modes =70;91

-There are cases w/o a mode!

### Normal distribution chart - Gauss

[will appear on test]



-only in his normative graph is median/mean/mode the same

When advertising for a job for a teller, (say, at a bank)  
→ not honest advertising to give average of all workers of the bank  
→ the mean will be higher than the tellers' average pay

Honest advertising:

- 1) the mean of the tellers
- 2) the mode of all the workers

Negative skew – a case of more variables on the negative side than on the positive side

Positive skew – a case of more variables on the positive extreme

\*average is influenced more by extremes than median

### **Range**

#### **Case 1**

70

80 → mean/average/mode

90

#### **Case 2**

79

80 → mean/average/mode

82

Range in case 1 = 20

Range in case 2 = 2

### **Tutorial Nov. 14, 2000**

2 ways of organizing data:

1) tables

2) graphs

### **Frequency tables**

A coin is tossed 10 times: H=heads, T=tails

HT

TH

HH

TH

HT

Value of coins show:	(f)requency	Cf/F (small cf or	P (proportional	Cp Cumulative
-------------------------	-------------	----------------------	--------------------	------------------

		capital F)	frequency)	proportional frequency
Tails	4	4	4/10 =40%	4/10 = 40%
Heads	6	10	6/10 =60%	10/10 =100%

### **Proportional Frequency**

- percentage/average
- f/n (frequency over total #)
- allows you to deal w. # in relation to all the research
- also, we do not have to deal w/ large #.
- allows us to compare various measurements (i.e. 20 toss vs. 50 toss of coin)

### **Cumulative Proportional Frequency**

CF/N

--

- Continuous table is slightly different
- we will put # in interval class

### **Example**

<b><u>Time of 300 runners running 100m (in seconds)</u></b>		
<b><u>Exact limits</u></b>	<b><u>Running time</u></b>	<b><u>Frequency</u></b>
16-17.999...	16-17	50
14-15.999...	14-15	100
12-13.999...	12-13	100
10-11.999...	10-11	50

### **Title of a chart**

- Has to be accurate
- who is the population
- # of the population
- what is being tested (what is the variable?)
- other variables if possible

### **Chart**

- has to include all variables
- all variables need to be in a category
- not more than 1 category

### **Rules of making a table**

- researcher has to decide how to present:

**1) # and range of category**

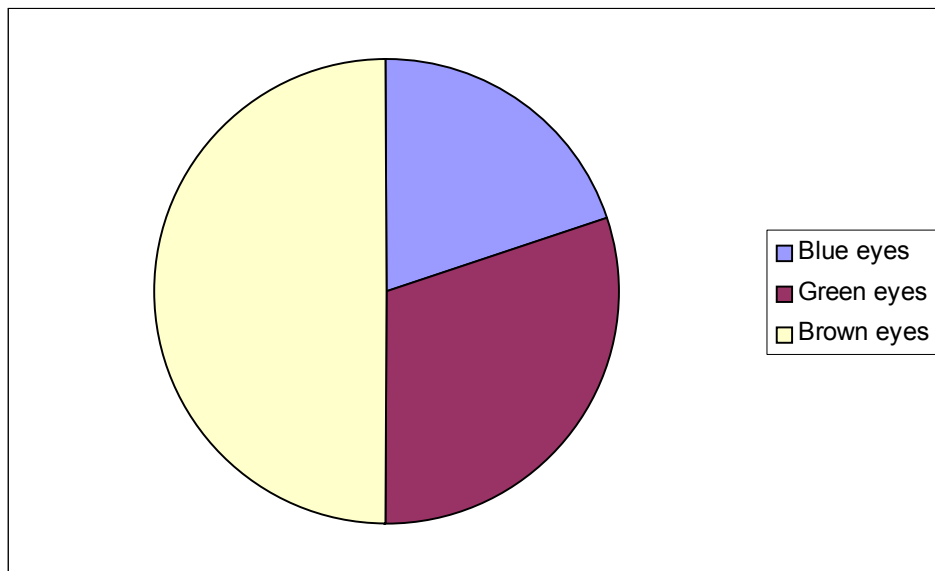
**Example from above chart (runners)**

<b>Researcher has # of categories</b>	8/4=2 (range divided by # of categories = size of category)
Range/#of categories = size of category	
<b>Researcher has range of categories</b>	8/2=4
Range of research/range of category = # of categories	8=range of total data 2 = range of category 4 = # of categories
<b>Write top value of categories</b>	
Write in the correct column the right top number of that class interval's range	
<b>Write bottom value of categories</b>	
1 unit about top # of category below	
<b>Putting in the exact limits</b>	
-Or writing them at the bottom of the chart as a definition	

## **Graphs**

### **PIE**

- Nominal graph (qualitative)
- reflects 100% of all categories
- divided proportionally



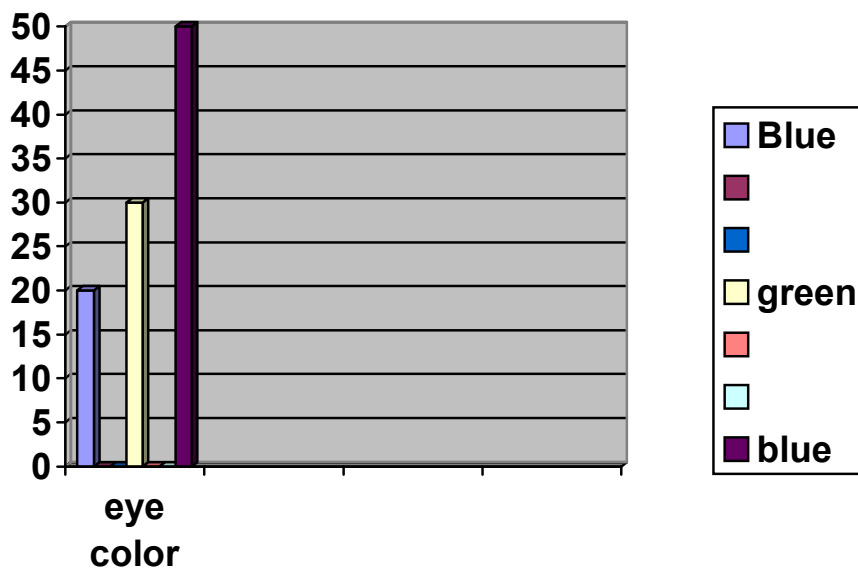
50% -Brown Eyes  
30% -Green Eyes  
20% - Blue Eyes

To make own pie-chart

$$360 \times 20 / 100 = 72$$

The angle in the blue eye's slice is 72 degrees

Bar Chart



Please note the space b/w bars: it is b/c this graph is not continuous

Please note: if you want to connect bars w/ lines, it has to be w/ a dotted line, since those variables are really discrete

Histogram

-continuous

Please note: the area of the bar reflects frequency

→ area of bar = proportional to # of cases (frequency)

Example:

	<u>Exact value</u>	<u>Frequency</u>	<u>d</u>
0-4	0-4.99...	20	$20/5 = 4$
5-9	5-9.99...	30	$30/5 = 6$
10-19	10-19.99...	40	$40/10 = 4$

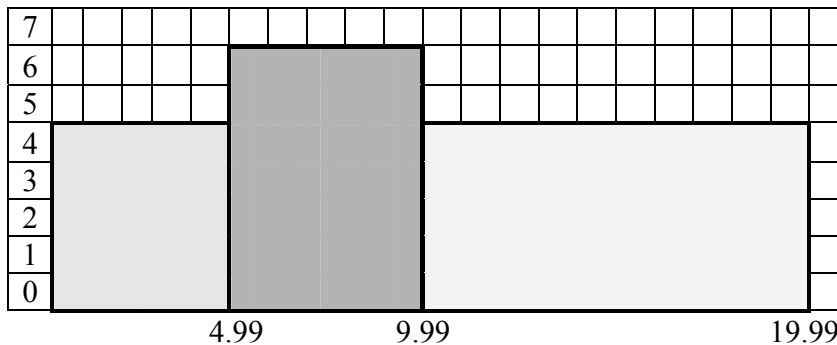
## D

$$F/I = d$$

F = frequency;

I = class frequency;

D = takes into account **both** height and width of bar in a histogram



first box = 20 units (5x4)

**Polygram:** -if you want to draw a line in the histogram to show the change, it has to be a continuous line (unlike the discrete charts, where the line is dotted)

\*Line has to be straight!

\*has to hit the box at the horizontal midpoint

**midpoint** – the unit half way b/w largest and lowest # within a range

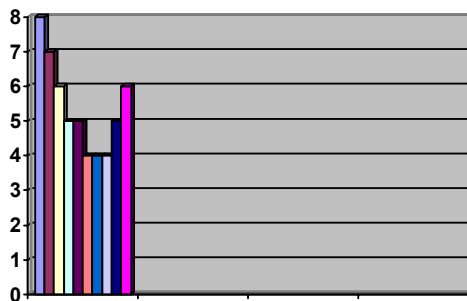
## Polygon

A graph w/ a straight line to show values, as opposed to boxes

## Curves

-a continuous chart where each frequency variable is a small line, which is adjacent to the next line

-->if you connect all the tops of the lines, you get a curved-

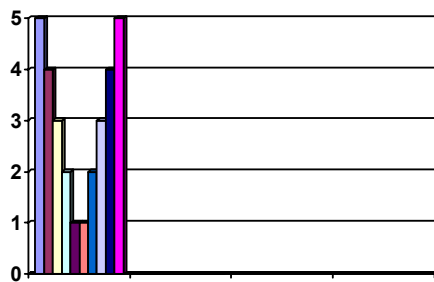


1) unison distribution/blocks


2)Gauss (normative)

-see above

3)



### A-symmetric distribution

1)asymmetric distribution with a positive skew:

-->a gauss chart w/ longer positive extreme

2)asymmetric distribution w/ negative skew

-->A gauss chart w/ longer negative extreme

### Variance

→have to know those formula by heart!!!!

Difference of Sample	Difference of population
$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$	$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n}$
Xi= midpoin of class interval, when class interval is a range of #'s	

$$\bar{x} = \frac{\sum x}{n}$$

### S=Standard of Variation (a.k.a. SD)

Size of unit of difference in the Normative distribution graph

**Example:** 70;80;90



$$\bar{x}=80$$

$$S^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

$$70-80 = -10$$

$$80-80 = 0$$

$$90-80 = 10$$

$$S^2 = \frac{\sum (-10)^2 + (0)^2 + (10)^2}{3-1}$$

$$S^2 = 200/2$$

$$S^2 = 100$$

$$S = 10$$

--

### **Example:**

7;8;x

$\bar{x}=8$

x=?

x must be 9 → in order to make mean 8, the last variable must be 9

→ in a normal distribution, the samples could be along the range as long as the last one has to be such that the normative distribution is kept

*[Irrelevant until later]*

### **Normal distribution**

-in a normal distribution,  $\pm 1$  degree of variation represents 34% of population of each side of the mean.

1=34%

2=13%

3=%2

-one can add degrees by extending the line of the graph at its end.

Example

$X/\sigma$ (SD)	Area (Cumulative percentage of total variable of the normative distribution)	Ordinate → height- irrelevant for our purposes. Ordinate's peak= mean
00	00	0.3989
1	0.3413	0.2420

2	0.4772	0.0540
3	0.4987	0.0044

$X - \bar{x} = 0$  from average = average.

→ 1 from average = 2 degree of variation

### **Tutorial, Nov. 28**

Researcher

#### **central measurement**

-a measurement that describes the characteristics all the included X values

-mode/mean/median

- usually, average gives the most accurate reflector of average– least amount of variance from the variable  
→ problem: affected a lot by extremes
- When describing income of all the staff of a company, average is not accurate, since it influenced by few extremes → i.e. executives  
→ the best is median

i.e. average – gives you a general idea about the ballpark range of the measurements

-describes a general tendency

Madad netiya merkazit

X	F
4	2
5	4
6	1
7	3
8	1
9	1
10	1

#### **Mode**

The most common variable.

#### **Accepted abbreviation:**

Mo/<sup>x</sup>

→ Usually describes nominal variables

#### **You can find out the mode by:**

- 1) Counting
- 2) Putting data into frequency chart, and mention the variable that happens to have most frequency

- When value of all variables are identical –no mode
- When even # of variables, the mode is average of 2 middle #'s
- Bi/trinomials – 2/3 modes

### **Median**

-Median – value of a variable where  $\frac{1}{2}$  of the variables are above it and  $\frac{1}{2}$  are below it

- $\sum |X_i - Md|$  is the lowest results (as opposed to  $\rightarrow \sum |X_i - \bar{X}|$ )
- Median is not influenced by extreme variables  $\rightarrow$  only influenced by # of variables
- Appropriate for Ordinal/Interval/Ratio

### **Abbreviation**

Md/~x

### **How do we find it?**

Formula:  $\frac{n+1}{2}$

$\rightarrow$  When there is an even # of variables = median is average of 2 middle variables

### **Mean**

Sum of all variables divided by # of variables

### **Formula to find out the mean**

$\frac{\sum X_i}{N}$

To find out mean of following chart

X	F
4	2
5	4
6	1
7	3
8	1
9	1
10	1

$\bar{X} = \frac{\sum X_i F_i}{\sum f_i}$

$\rightarrow$  If the range of each class interval is more than one, multiply the midpoint of the category as opposed to the actual Class interval, in case of the single # as the class interval.

### Symbols

-/x=population

-μ =population

**Average** – only works for interval/ratio scale

### Tendencies of Average

- $C+X_i = C+/X$
- $CxX_i = Cx/X$

### Mode/Mean/Median's effect on graphs

**Normal distribution** = mean/mode/median are all in the middle

**Symmetric distribution** (2 peaks at both ends of graph and a low-point in the middle)

→ /X and Md are in the middle, Mode is at both peaks

### Order of influence by extremes:

- 1) Mean
- 2) Median
- 3) Mode

Madad pizur

Standard of variance –

Madad pizur yachasi

Madas kesher

### Ways to differentiate b/w normal distribution charts

**1) Range** – difference b/w height lowest values of distribution

→  $R = X_{\max} - X_{\min}$

### Example:

-range = 1-9

→  $R = 9 - 1$

→  $R = 8$

### 2) Mean deviance

-wants to find how far people are from mean

→ no practical usage for us

$\Sigma |X - \bar{X}|$

n

→ no usage to us

3)

**Variance**

-Better version of mean deviance

$$\frac{\sum(X-\bar{X})^2}{N}$$

**Symbols**

$S^2$  - difference within a sample

$\sigma^2$  - difference within a population

N= population

N = sample

**Standard deviation**

-Same thing as variance, just not squared

**Symbols**

Population:  $\sigma$

Sample: S or SD

**Formula:**

$$\sigma = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

→the answer is a unit of deviation from mean

**Other versions of formula – working formula**

$$\sigma = \sqrt{\frac{\sum Xi^2 - \frac{(\sum Xi)^2}{N}}{N}}$$

$$\sigma = \sqrt{\frac{\sum FiXi^2 - \frac{(\sum FiXi)^2}{N}}{N}}$$

$$\sigma = \frac{1}{N} \sqrt{N \sum Xi^2 - (\sum Xi)^2}$$

**Calculator**

-instead of adding the variables normally, use  $m^+$  and then press the appropriate function (SD, etc)

Population button= n+1

Sample button: n

### **Constant rules**

- 1) A constant added does not change the SD or the variable → only the mean
- 2) A constant multiplied stretches out the graph  
→ if you multiply by a constant, S is also multiplied by 3  
→  $S^2$  is multiplied by  $C^2$

$$\frac{S}{C} = \frac{S^2}{C^2}$$

--

### **Class, December 4, 2000**

In IQ, average is 100, SD is 15

### **Z score / standard deviation**

-Defines How Many SD from average a given variable is

Formula
$Z = \frac{X - \bar{X}}{SD}$

### **Example #1**

-if a person has an IQ of 85 (we want to find out his Z score - how many SD he is from average)

$$Z = \frac{85 - 100}{15}$$

$$Z = -1$$

\*Therefore, he has a z score of -1

### **Example #2**

$$\bar{X} = 5$$

$$SD = 1.5$$

$$x = 7$$

$$z = \frac{x - \bar{X}}{SD}$$

$$z = \frac{7 - 5}{1.5}$$

$$z = 2/1.5$$

Z-1.33

-->look up 1.33 in SD chart

-->91% have less than him

$X/\sigma$  = difference of average

(please note the script X; the script x means:  $(x - \bar{x})$ )

### **Probability**

$P(A_i) = \frac{\text{success}}{\text{Success} + \text{no success}}$

$P$  = probability

$A_i$  = what is the variable

### **Example:**

I throw a coin once and get a head

### **Formula**

Head

Head + tail

$\frac{1}{1+1}$

Probability =

$\frac{1}{2}$

### **Rule #1**

$0 \leq P(A_i) \leq 1$

### **Rule #2**

$\sum P(A_i) = 1$

### **Rule #3**

$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

-->What is the probability of  $A_1$  and/or  $A_2$ s

$Z$  = # of SD above/below average

$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

(3RD RULE OF PROBABILITIES)

$\cup$  = union

$\cap$  = intersection/overlap

### **Example #1**

-Person #1 has 4 balls

-Person #2 has 4 balls

-but one of the balls are shared by both!

### **Therefore:**

$$P(A \cup B) = P(4) + P(4) - P(1)$$

-->therefore, there are 7 balls

### **Example #2**

$P(A) = 0.12$  (people studying at BIU)

$P(B) = 0.11$  (people studying at Haifa)

$$P(A \cup B) = P(0.12) + P(0.11) - P(0)$$

-->  $= 0.24$  (24%) of population is studying at either BIU or Haifa

[there is no overlap of people who study at both  $\rightarrow$  overlap  $= 0$ ]

$$P(A_1 \cup A_2) = P(a_1) + P(a_2) - (P(a_1 \cap A_2))$$

### **Independent events**

i.e.

$P(A) = 0.12$  (i.e. the % of people of the general population who are in BIU)

$P(B) = 0.05$  (i.e. the % of people of the general population who eat Falafels)

### **Formulas for independent events:**

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

The answer for this formula is the overlap. you plug that into the original probability formula.

### **Dependant events**

-events that 1 event is dependant on the other (1 event changes the probability of the another event

i.e. H.M. people use umbrellas when it is raining vs. H.H. people buy umbrellas

i.e. probability of taking a specific card out of a deck with out returning it b/f next event

\***Note:** you can't get an overlap bigger than the smaller variable



### Formula for dependant events

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

$P(A/B)$  = on the condition that A took place, what is the probability of B?

### Bay's formula

-for a case where it is given that a certain event already took place

→ we want to know the probability of an event dependent on this past event

[cases where the givens show that simple multiplication of probabilities is not the right answer]

### Steps To Solve

- 1) Draw chart
- 2) Fill out unknowns using the givens
- 3) Apply #'s into Bay's formula

### Example of Bay's formula:

Male population of uni = 0.6

Males of university who are married = 0.2

Female singles in Uni. = 0.3 ( $\bar{A} \cap B$ ) → where  $\bar{A}$  overlaps /B

### Steps 1

	Married (B)	Single (/B)	Total
Male (A)	0.2		0.6
Female ( $\bar{A}$ )		0.3	
Total			1

### Step 2

	Married (B)	Single (/B)	Total
Male (A)	0.2	0.4	0.6
Female ( $\bar{A}$ )	0.1	0.3	0.4
Total	0.3	0.7	1

### Step 3

$$P(\bar{A} // B) = \frac{P(\bar{A} \cap B)}{P(\bar{A})}$$

$$= 0.4 / 0.7$$

signs for Bay's: given that...., what is...

variance:  $S^2 = NPQ$

SD:  $= \sqrt{S^2}$

Average:

### Z-score

→ you can plug in bay's theorem answer into Z-score

-in this case, it is standard to plug in the lower class interval for  $X_i$

**Scenarios:**

**Overlap results is variables are:**

- 1) mutually exclusive = 0
- 2) Independent =  $p(a) \times P(b)$
- 3) Dependant = whatever answer doesn't correspond to #1 or #2

\* When the events are mutually exclusive, the  $P(A \cap B) = 0 \rightarrow$  there is 0 overlap!

**Overlap**

**Formula**

$$P(A \cap B)$$

- in overlaps where the events are independent of each other

$\rightarrow$  1 does not affect the other

$\rightarrow$  they might overlap

-in dependant events that has no common denominator  $\rightarrow$  NO overlap!

**Example:**

-When wanting to choose a specific card several times, and returning it after each event

$\rightarrow$  independent

-if the specific card is not returned  $\rightarrow$  dependant [on the previous event]  $\rightarrow$  no overlap

**Permutation**

$$n! = n(n-1)(n-2)(n-3) \dots 1$$

**Example:**

Students A B C D E

you can place them in any order.

-This formula gives you the # of orders it can possibly have

ABCDE

ADBEC

ACDEB

etc.

-in permutations, the order of variables matter!

- $n$  = # objects/experiments (# of variables)
- $r$  = # of successes (# of variables I want to find)

### **Example**

Variables: a, b, c

-  $n=3$

-  $r=2$

ab

ba

ac

ca

bc

cb

### **Formula**

$$\frac{N!}{(n-1)!}$$

$$=3!/(3-2)!$$

$$=6/1$$

=6 permutations

### **Combinations**

-in combinations, the order of variables is irrelevant

### **Example**

-in variables a, b, c

ab

ac

bc

### **Formula**

$$\frac{N!}{R!(n-r)!}$$

$$=6/2 \times 1$$

$$=3$$

-in permutations, the order of variables matter!

-in combinations, the order of variables is irrelevant

### **Binomial**

-a formula to find the probability of a specific combination (**Note:** not a permutation but a *combination*)

- Probability has no memory

P = % of successes

Q = % of non-success

N = number of events

R = # out of N I am interested in (what are the chances of 3 out of 10 cars being.... R=3)

$$P(r,n,p) = c(n,r) P^r \times q^{n-r}$$

**Combination formula**  $[C(n,r)] =$

$$\frac{N!}{R!(n-r)!}$$

### **Rule #1**

-2 possibilities

### **Rule #2**

-Independent (probability doesn't change)

### **Formula**

$$\binom{N}{R} \times P^r \times (1-P)^{n-r}$$

### **Example #1**

-n=3

-r=2

$$= \binom{3}{2} \times \left(\frac{1}{2}\right)^2 \times \left(1 - \frac{1}{2}\right)^{3-2}$$

$$= 3 \times \frac{1}{4} \times \frac{1}{2}$$

$$= \frac{3}{8}$$

### **Example #2**

-Probability of 1 king out of a deck of a king twice

n=2 (I did the experiment twice)

r=1 (# of successes)

$$= \frac{1}{2} \times \left(\frac{4}{52}\right)^1 \times \left(1 - \frac{4}{52}\right)^{2-1}$$

$$= 2 \times \frac{4}{52} \times \frac{48}{52}$$

$$= \frac{8}{52} \times \frac{48}{52}$$

$$=2/13 \times 12/13$$

$$=24/169$$

### **Tutorial, December 12, 2000**

#### **Z-Score**

**Z-score** –the standardized score which places a variable in relationship of the  $\bar{x}$  of group

- Z-score tells us the distance from the average in terms of SD

-Useful when comparing 2 different distributions i.e. marks of 2 classes

#### **Formula**

$$Z_{xi} = \frac{X_i - \bar{X}}{SD}$$

→to compare 2 diff. distributions or a person and a dis

<i>Intro to psych</i>	<i>Intro to statistics</i>
$\bar{x}=85$	$\bar{X}=88$
SD=6	SD=4
$X_i=86$	$X_i=88$
$Z_{xi} = \frac{86-85}{6} = 1/6$	$Z_{xi} = \frac{88-88}{4} = 0$
Even though 88 is more than 86, 86 has a higher z-score	

#### **Times when Z-score is useful**

2 distributions are w/I the same scale, but different  $\bar{x}$  or distribution

Difference of measure: length vs. color

#### **2 Aspects of Z-Score**

- 1) Absolute size of Z-score [how far from  $\bar{x}$ ]
- 2) What sign does Z-score have (+ or -)

- When  $Z > 0 \rightarrow X_i > \bar{x}$
- When  $Z < 0 \rightarrow X_i < \bar{x}$
- When  $Z = 0 \rightarrow X_i = \bar{x}$

-When we convert all the scores (variables) into SD, it b/c a normative distribution

The average of the Z-score =  $\bar{Z}$

$$S_z = 1$$

$$S_z^2 = 1$$

#### **Aspects of normative distribution**

- 1) Has bell shape
- 2) The distribution is symmetrical  $\rightarrow \bar{x} = md = mo \rightarrow$  Are all in middle of chart

- 3) The area within the chart = 1 or 100%
- 4) Theoretically,  $-\infty = +\infty$
- 5) Normal distribution has 2 parameters
  - $\mu$  = /z of population
  - $\sigma$  - SD of population
  - only has +values → doesn't have negative values
- various normative distribution charts differ from each other in sd/ $\sigma$  or both
- 6) It is standard to change the variables into SD and therefore it becomes a normative distribution chart
- 7) In normative chart area is % allotted to that frequency of variables

SD	Area
$\pm 1$	34%
$\pm 2$	13.59%
$\pm 3$	2%

Example $\mu = 100$ $SD = 10$ $X_i = 115$
$Z = \frac{X_i - \mu}{SD}$ $Z = \frac{115 - 100}{10}$ $= 15/10$ $= 1.5$ <p>→ look at chart: 0.433</p> $50\% + 43.3\% = 93.3\%$

### **Example of Z score:**

20,000 sign up for uni. 2000 get accepted  
 the  $\mu$  of all the students = 81  
 $SD = 5$

10% get accepted → look up in chart: 40% ( those above average who didn't get accepted)  
 $= 1.28$

$$1.28 = \frac{X_i - \mu}{SD}$$

$$5 \times 1.28 = X_i - 81$$

$$5.4 = X_i - 81$$

$$81+6.4=X_i$$

$$X_i=87.4$$

Minimum acceptance % is 87.4

### **Probability**

**Sample space** = S = the group of ALL the possibilities of the experiment

**Event** – a subgroup of the sample/population

-***Event's symbol*** = a capital letter. I.e. A

- if I want a specific outcome to happen =A
- If it did not happen [all other possibilities] =/A

### **Sure event**

-an event that will inevitably happen

→i.e. during birth, it is a sure event that it will be a baby

### **Probability**

-a # b/w 0 and 1, which is related to each event within a sample space

→probability is not a function of % but of a #!!!!

-Called P(A)

- $P(A) > 0$

### **Rules of probability**

#### **Rule #1**

Probability of a sure event =1

#### **Rule #2**

$$P(/A) = 1 - P(A)$$

=sum of probability of all other possibilities =1-the probability of that one event

#### **Rule #3**

#### **Formula for probability**

-With a sample w/ a uniform probability, the probability is its relative occurrence within a sample

$$P(A) = \frac{n(A)}{N(S)}$$

### **Example**

-  $\frac{n(A)}{N(S)} \rightarrow$  # of kings in a deck

-  $N(S) \rightarrow \# \text{ of cards}$

### Union

What are the chance of 2 events happening  
 $\rightarrow$  what are the chances of A and/or B to happen

### Formula

P

### Tutorial, Jan 9, 2001

$$P(r,n,p) = c(n,r) P^r \times q^{n-r}$$

Combination formula  $[C(n,r)] =$

$$\frac{N!}{R!(n-r)!}$$

### Expectancy

Way to compare a range of answers into a binomial

$\epsilon$  – *expectancy*

$\rightarrow$  helps us find the average

$$\bar{x} = \sum p(X_i) \times X_i$$

### For Binomial

$$\epsilon(x) = np$$

### Example of expectancy:

Probability of car working on a morning = 0.6

- 1) What is the probability that on every day of the week, his car will not work?
- 2) What is the estimated times in a year that his car will work?

### Answer:

1)

$$P = 0.6$$

$$Q = 0.4$$

$$N = 7$$

$$R = 0$$

$$\frac{7!}{0!7!} \times (0.6)^0 \times (0.4)^{7-0}$$

2)

$$\epsilon(x) = 365(0.6)$$

$$\epsilon(x) = 219$$



### **Rule #1**

When N is bigger than 10, and the bigger it is, the closer the distribution is to binomial distribution

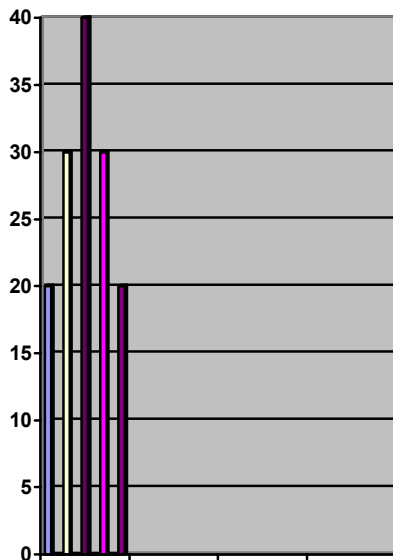
### **Rule #2**

The closer p/q are closer to 0.5, the closer to being normative the distribution is.

**Binomial distribution:** a distribution looking like a normative distribution. But the distribution can't be normal if variable are discrete!!! But it can be similar to one.

$$M = \mu = np$$

### **Example of a binomial distribution**



$$Z = \frac{\bar{X} - \mu}{\sigma(\text{sd})}$$

### **Example:**

Out of 100 kids, finding 60 boys

P=0.5

Q=0.5

100 !

$$60!(100-60)! \times (0.5)^{60} \times (0.5)^{100-60}$$

### **Large Binomial R's**

When there is a large range of R's, binomial is too painstaking.

→compare it to normative distribution

**2 criteria to make binomial similar to normative distribution**

1)  $P \approx \frac{1}{2}$

2)

- $NP > 10$
- $NQ > 10$

<b><u>Average:</u></b> NP	<b><u>Standard deviation</u></b> $SD = \sqrt{NPQ}$
<b><u>Formula:</u></b> <ul style="list-style-type: none"> <li>• Find average</li> <li>• Find SD</li> <li>• Plug into z-score</li> <li>• Find %'s in chart</li> </ul>	

**Example**

**Question:**

-40 matches in a box.

-chances the match is broken: 2/5

-what are the chances that in a box of 40m 20-25 are broken?

**Answer**

<b><u>Givens</u></b> N=40 P=0.4 R=20-25	
<b><u>Average</u></b> $\mu = np$ $\mu = 40 \times 0.4$ $\mu = 16$	<b><u>SD</u></b> $\sqrt{NPQ}$ $= \sqrt{40 \times 0.4 \times 0.6}$ $= \sqrt{16 \times 0.6}$ $= 2.098$
<b><u>Z-score for top of the range</u></b> $Z = \frac{25 - 16}{2.098}$ $Z = 2.9$ $= 49.81\%$	<b><u>Z-score for bottom of the range</u></b> $Z = \frac{20 - 16}{2.098}$ $Z = 1.29$ $= 40.15\%$
<b><u>Final Answer</u></b> $49.81\% - 40.15\% = 9.66\%$	

### **N!A!B!C!**

In cases when the question is not the chances of an order (n) but merely order of n, given the grouping of categories

### **Class, Jan 8, 2001**

**Positive correlation:** When X goes up, y goes up

**Negative correlation:** When X goes up, Y goes down

**No correlation:** no connection b/w X and Y

### **Class Jan 15, 20001**

**Range of correlation:**  $-1 \leq R \leq 1$

### **Formula for correlation**

$$R = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

### **Other way to say it:**

$$\frac{\text{Covariance } (x,y)}{S^2XS^2y}$$

\***Note:** correlation does not define causation

### **Example**

#### **Step 1**

	<u>X</u> (high school mark)	<u>Y</u> (SAT Scores)
	8	600
	9	700
	10	800
<u>/X</u>	9	700

#### **Step 2**

-put into  $(x-\bar{x})$

	<u>X</u> (high school mark)		<u>Y</u> (SAT Scores)	
	8-9	=-1	600-700	=-100
	9-9	=0	700-700	=0
	10-9	=1	800-700	=100

#### **Step 3**

$$R = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

$$\sqrt{\sum(x-\bar{x}) \sum(y-\bar{y})}$$

$$R = \frac{(-1)(-100) + 0 + 1(100)}{\sqrt{(-1^2 + 0 + 1^2) \times (-100)^2 + 0 + (100)^2}}$$

$$R = \frac{100 + 100}{200}$$

$$R = 1$$

Answer

Correlation is 1

Example

If  $r=0.5$ ,  $R^2=0.25$

Tutorial, Jan 17, 2001

Correlations

- how 2 variables affect each other
- i.e. relation b/w success and fear on a test.

1) Range of correlation:  $=R$

$$= -1 \leq R \leq 1$$

2) Signs of correlation:

- 1) + -positive relationship b/w the variables  
→(if x goes up, so does Y; if C goes up, so does Y
- 2)- -negative relationship b/w variables  
→if x goes up, Y goes down. If Y goes up, Y goes down
- 3) 0 –no **linear** relationship

curvilinear:

- a curve correlation
- i.e. at first, anxiety helps one succeed at a test.  
→too much or too little will affect the test results negatively

3) Value of correlation

- the bigger the absolute value of correlation is, the more significant the correlation is

3) The fact that a correlation is present does not dictate causation

Pearson correlation –for interval/ratio

Spearman correlation –for ordinal/interval scales

Pie correlation –for Nominal (not relevant for us)

Pearson correlation

2 conditions:

- for interval or ratio scales
- for a linear correlation b/w 2 variables

$$r = \frac{N\sum XY - \sum X \sum Y}{[\sum X^2 - (\sum N)^2][\sum Y^2 - (\sum Y)^2]}$$

**Example –see handout, January 16**

-correlation b/w father/son

- 1) linear transformation of variables does not change the correlation b/w them  
→ i.e. adding/multiplying by a constant
- 2) Pearson correlation: interval and onwards.  
→ but can't compare the correlations: 0.2 is not ½ of 0.4

**Covariables**

$$R = \frac{\sum (x - \bar{x})(y - \bar{y})}{N}$$

=cov (x,y)

r=coefficient

→ how well (correlation) the 2 variables go together

**Spearman**

-Good for ordinal or interval correlations

$$R_s = 1 - \frac{6\sum d^2}{N(n^2 - 1)}$$

D=difference in the rank x and y, where each one variable represents a different placement w/I the order of variables

**Example of Spearman**

-Beauty contest – 4 women were ranked for beauty

Contestant	Judge 1's ranking	Judge 2's ranking	D (diff in ranking)
A	3	2	1
B	1	1	0
C	2	3	-1
D	4	4	0

$$1 - \frac{6(2)}{60}$$

$$= 1 - 12/60$$

$$1-0.2$$

$$=0.9$$

**Note:** answer of spearman = difference in judges answer  
 -->  $D^2 = 0$  = the judges ranked same

-see handout –jan 16, 2001

### **Class, Jan 22, 2001**

**Variance:** = how diff. from the mean they are = spread out

**R<sup>2</sup>**

$$(Y - \bar{Y}) = (Y - Y_p) + (Y_p - \bar{Y})$$

↓

↓

↓

A

B

C

$Y_p$  = probability of Y

**A** = difference b/w variable and average →  $(x - \bar{x})$  or B+C

**B** = difference b/w  $X_i$  and the correlation/regression line

**C** = difference b/w regression line and average

**Explained variance-** the variance from the average that you can account for  
i.e. b/c of socioeconomic level/education/genes

**Important:** the closer to the line you are = more explained variance

→ bigger variance from  $\bar{x}$  → less explained variance

### **Regression**

-given that a correlation was established, I can try to figure out 1 variable, once the other variable is given.

X = the predictor

Y = the predicted

### **2 conditions for regressions**

-The total sum of distances of all variables to the average (the line running through the correlation graph) = 0

-the total sum squared of the variables are minimal

### **Regression line**

-the name of that line running through the correlation

→ represents correlation

-if all variables (dots) are exactly on the line, the correlation is 1

-the line represents the ideal – the correlation of 1.  
→ deviations from this line is the correlation

-smaller distances b/w variables and line:  
→ Better predictions  
→ Bigger correlation

$Y^l$  = the predicted y

### **Formula for line in the regression**

$$Y^l = a + bx$$

(*note*: for practical reasons, the a and the b have been changed from the original formula for lines  $y = ax + b$ )

### **Variables**

A = the point in which the line hits the y.

B = regression coefficient – the slope of the correlation chart

*Note*: the b's sign (+ or -) is always indicative of the correlation's sign

### **Formulas needed for regression**

$$B = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b = \frac{\text{cov}(x, y)}{S^2 X}$$

$$B = R \times \frac{SY}{SX}$$

$$A = \bar{y} - b/\bar{x}$$

### **To find a:**

$$A = \bar{Y} - b/\bar{X}$$

→ (average of all the Y's minus b times average of all the X's)

### **Exercise**

#### **Given info**

- BA marks ( $\bar{y}$ ) =  $\bar{y} = 90$
- SAT =  $\bar{X} = 100$
- $R = +0.8$
- $B = 0.68$

$$-A = \bar{y} - b/\bar{x}$$

$$-A = 90 - 0.68 \times 100$$

$$-A = 90 - 68$$

$$-A=22$$

**Note:** the diff. b/w slope and correlation

### **Plugging in the answer into usage**

$$Y^l = a + bx$$

$$Y^l = 22 + 0.68X$$

→ X is the specific variable.

### **Example:**

If a person's high school mark (x) is 98. what is his *predicted* mark in his BA (y)

- plug into  $Y^l = a + bx$

$$Y^l = 22 + 0.68X$$

$$Y^l = 22 + 0.68(98)$$

→ Therefore, his *predicted* mark is 88.64

$$R^2 = S_{Y^l} / S_y = \sum (y^l - y)^2$$

→ incomplete?

### **Standard error of estimate (of the prediction)** → after the fact

-Standard deviation of the line

→ how far the individual variables are from the line

To find standard deviation of a line

$S_{yx}$  == sd of a line xy

$$S_{xy} = S_y \times \sqrt{1 - r^2}$$

R = correlation

### **Example #1**

If  $r=1$

$$S_{xy} = S_y \times \sqrt{1 - 1^2}$$

$S_{xy}=0$  → perfect correlation –no error of estimate.

### **Example #2**

If  $r=0$  → no correlation

$$S_{xy} = S_y \times \sqrt{1 - 0^2}$$

$$S_{xy} = S_y$$

→ prob! = no prediction! The of the distribution from the mean of the variable is equal to the total distributions of all the variables from the mean

### **Radar**



### Decision theory

-relation b/w the truth and what is reported

	<u>Truth:</u>	
<u>Reported:</u>	Plane	bird
Bird	X (not good)	✓
Plane	✓	X(?)

### When comparing 2 normative distributions:

- if only 5% of group 1 place w/I range of group 2, it could be said that 'they are better/more'.
- If more than 5% of group 1 is w/I group 2, it is assumed that the difference is not sharp enough to assume that group 2 is more/better than group 1

-this 5% is called the '*critical value*'

-'*Omdam*' –estimate

-'*bilti-mut*'e' –no biases to either side → always in the middle.

-sample's average is a form of unbiased estimate of the population's average

→ ***Sampling error*** = the errors that sometimes occur w/ the diff. b/w  $\bar{x}$  (sample average) and population average ( $\mu$ )

-b/c of *sampling error*, we can't solely rely on it. We need to make a distribution of all  $\bar{x}$  of all the samples.

→ called *sampling distribution*?

### Sampling distribution

Note: in *sampling distribution*, *SD* is called *Standard Error. (SE)*

### Conditions:

- population is normal
- sample is over 30 or 100
- etc.?

### Results:

- average of sample distribution is  $\mu$
- Standard error:  $\sigma/\sqrt{n}$

--

### 2 kinds of errors

***Alpha ( $\alpha$ ) error*** –error related to sample: I chose the smarter student/not enough students samples → up to a 5% error.

- **deals w/ the right skew of group 1**

- Alpha: I say that there is a difference w/ in reality, there is no difference

**Beta ( $\beta$ ) Error-** when, because of the 5% technicality, I declare that group 2 is really not really more better than group 1, when it really is better.

- **deals w/ the left skew of group 2**

- saying that there is no difference when there really is a difference

Tutorial; Feb. 27, 2001

**General distribution:**

- 1) Explained differentiation  $\rightarrow r^2 \rightarrow$  the correlation b/w the dep. and indep. variables
- 2) unexplained differentiation.  $\rightarrow$  the # in the dep. variable that is diff. from the prediction made by the indep. Variable
  - $1-r^2$ 
    - Unexplained differentiation –difference b/w predicted (regression line) and actual value of variable

$R^2$  =explained variance

--

**Conclusive statistics** –trying to apply a conclusion made about a sample of the population on the general population.

**Measures:**

**Statistic** –a measure relating to a sample of a population  
 $\rightarrow$  slightly diff, according to which sample we use

**Parameters** -a measure relating to ALL the population  
 $\rightarrow$  always constant w/I a group

**2 things we want to do:**

- Estimate of parameter from sample
- Finding the parameter, using the statistics, when you have an hypothesis.

**Distribution of the sample**

-the distribution of a mean  $\rightarrow$  a distribution of all the means within a certain range for n.s  
 -a theoretical distribution

parameter mean = 102.6 =  $\mu$   
 SD = 13.29 =  $\sigma$

### **Formula**

$\mu - \sigma = \text{sampling error}$

--

### **Standard error of distribution**

-the SD of an *sampling distribution*

→ the diff. b/w sample average and population average

***Sampling distribution***: a distribution chart we make out of *averages of distributions* and not variables.

### **Conditions for sampling distributions to be normative distribution:**

- 1) if the distribution of the population is normative, so is the sampling distribution
  - 2) sample has to be big enough.
    - over 100
    - after 30, one can say it is '*almost-normative*'
  - 3) if you take an infinite # of samples of the population, the distribution will be normative
- has to be one or the other. I.e. even if pop. is not a normative distribution, the averages roughly are

### **Central border axiom**

-if you take all the samples from the population and figure out their average, the closer you get to all the population (i.e. the bigger the sample is):

- The sample distribution will b/c increasingly normative.
- average =  $\mu$  of population
- SD of this distribution:  $\sigma$  (sd of population) /  $\sqrt{n}$

### **Formula**

$\sigma / x = \sigma / \sqrt{n}$

### **Example**

$\mu = 170$

$\sigma = 20$

$n = 25$

$\sigma / x = \sigma / \sqrt{n}$

$20 / \sqrt{25}$

$20 / 5 = 4$

4 = the SD of this sample distribution

-if n of population is 100, and we want to figure it out:

$$\sigma / x = \sigma / \sqrt{n} =$$

$$20 / \sqrt{100}$$

$$20 / 10 = 2$$

SD of population: 2

-the larger N is, the smaller the SD is.

**Question:**

Does my sample  $\bar{x}$  reflect the average of all the population, or is there a small diff, based on sampling error?

**Answer: we need to go through some statistical manipulation**

**Example:**

Population SAT ( $\mu$ ) = 500

SAT course alumni  $\bar{x}$  = 670

**Question:**

Is the alumni average reflective of the course, or is it just simply a sampling error spin-off of the population?

**Stages:**

**Stage 1**

**Statistical hypothesis**

→ says some estimation of a parameter of a population (as opposed to as sample)  
→ 2 opposing phrases, opposing each other, where we have to see which one is true or not.

→ i.e. is 1 really different/better than 2 or not?

- $H_0$  → a known/accepted fact → we assume that unless proven otherwise, we assume everyone is equal.  $\bar{x} \leq \mu$
- $H_1$  → usually the 'claim' → the 'chidush' that the examiner is trying to prove  
 $\bar{x} > \mu$

→ we only measure  $H_0$  → if it is not correct, we assume  $H_1$  is correct.

1-ended (tailed) hypothesis: I assume 1 thing will happen

2-ended (tailed) hypothesis: either this or than could happen → I am not sure of the results

-the average ( $\bar{x}$ ) of the people of that saw movie x is diff from population's average ( $\mu$ )

**2 ended**

either  $\bar{x} = \mu$

or  $\bar{x} \neq \mu$

## Stage 2

-collecting/checking empirical/statistical facts

→this is where you go an gather info (and actually get the averages for the samples/population

→we will see how big the diff. b/w populations/samples

## Stage 3

What is the probability (p) to get the result of stage #2 randomly?

→what are the chances that the samples of stage #2 are random?

→is 670 a sampling error or really reflective of its population?

→figure out z-score.

-if z-score % is small, then the likelihood for this to be random is smaller.

## Stage 4

-so we refute H<sub>0</sub>?

→alpha: the amount of risk I am willing to take.

→defined by researcher b/f study

→standard in psychology =5% answer to the z-score.

## Tutorial march 13, 2001

-all conclusions in psychology are based on probability

<u>Reality</u> →	H <sub>0</sub>	H <sub>1</sub>
<u>Researcher's hypothesis:</u>		
<i>H<sub>0</sub> is correct</i>	<i>Hit</i>	<i>Alpha mistake</i> =we refute H <sub>0</sub> even though it is in reality correct  →Researcher claims that H <sub>1</sub> is better, even though it really isn't
<i>H<sub>1</sub> is correct</i>	<i>Beta mistake</i> =we refute H <sub>1</sub> even though it is in reality correct  →Researcher claims that H <sub>1</sub> is really a slightly extreme sample of the H <sub>0</sub> population, even though in	<i>Hit</i>  -beta/H <sub>0</sub> is correct

	reality, $H_1$ is a sample from a different/better population	
--	---	--

### $H_0$

-all the area except alpha (5%)

→when we refute  $H_0$  →we have a chance for an *alpha* mistake

### $H_1$

-all area in the distribution of the sample except beta (the 5% of this sample that are w/I alpha area)

→when we refute  $H_1$  →we have a chance for a *beta* mistake

### Alpha:

2 approaches: -if I got 1%

- 1) 5% chance to make a mistake, since that is the range that we are looking at
- 2) 1% chance to make a mistake, because that is where my result is

### What characterized alpha/beta?

#### Alpha

Is decided by researcher b/f the research paper

#### Beta

#### Has 3 factors:

1)Size of gap b/h  $H_1/H_0$

- the more overlap =less gap →more chances for bigger *beta* mistake
- less overlap =more gap →less chances for bigger *beta* mistake

#### To see the gap:

- Averages of the distributions
- The standard error of sample
  - $\sigma / x = \sigma / \sqrt{n} =$ 
    - also influenced by # of participants in sample
    - bigger n =SD is smaller →beta is smaller
    - bigger =bigger SD →beta is bigger

2)There is a reverse b/w alpha and beta

→alpha is bigger→beta is smaller

3)*power* of the test

1-beta

-Class, March 12, 2001

**Z-Test** –the formula to see if the group is better/worse than the population

→i.e. where they land in relationship to each-other

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

→for sample compared to a population

As opposed to:

$$Z_{xi} = \frac{X - \bar{X}}{SD}$$

→to compare 2 diff. distributions or a person and a dis

**Example**

$$\bar{x}=105$$

$$n=100$$

$$\sigma=15$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\frac{105-100}{15/\sqrt{100}}$$

=

$$\frac{5}{15/10}$$

$$=5/1.5$$

$$=3.33333$$

z score =3.3333 →look it up in table

→if the z score for that is less than 5%[critical point], it is better (in this case, it is)

**Standards**

-We split up the 5% critical value into 2 (2.5%)

-and put each at one end of the table, since we assume that our group might be better or worse than the general population's normative distribution

→so we split up the allotted 5% into 2 and put it at both ends of distribution

<b><u>Null hypothesis</u></b>	<b><u>1-tailed</u></b>
$H_o : \bar{X} = \mu$	$H_o : \bar{X} \leq \mu$
$H_a : \bar{X} \neq \mu$	$H_a : \bar{X} > \mu$

A=alternative	
-We assume, in null hypothesis that our population ( $\mu$ ) is not equal to the population ( $\mu$ ) →better or worse →if difference is not beyond critical value, we say that $H_0$ is irrefutable	

**1-tailed** – 5% on top end of distribution

**2-tailed** –2.5% on each side

### **Pros and cons**

#### **2-tailed system:**

##### **Pros**

-we will account for all the possible results of the comparison  
→**i.e.** whether  $H_1$  is better or worse than  $H_0$

##### **Cons**

-if I get 3% for the better → I will statistically never have found out it is better, even though it is

#### **1-tailed system**

##### **Pros:**

-you have rational/clear direction on what direction that the hypothesis will go

### **Central Limit/border theorem**

-if we take the average of all sample groups taken from any random distribution: the closer we get to infinite @ of samples, the closer we get to normative distribution

→called: **standard error of the mean**

### **Confidence interval**

-when you don't have the  $\mu$  of the population; you need to look for the  $\mu$  using the Z-test

$$\bar{X} - 1.96 S_{\bar{X}} \leq \mu \leq \bar{X} + 1.96 S_{\bar{X}}$$

(1.96 = z score for 5%)

$$S_{\bar{X}} = s/\sqrt{n}$$

### **Example**

$s=10$

$n=25$

$\bar{X}=40$



### Formula

$$\mu = \bar{X} \pm Z_{\alpha} S(\sigma)/\sqrt{n}$$

→ first do it for both + and -

$$10/5 = 2$$

$$= 40 - 1.96(2) \leq \mu \leq 40 + 1.96(2)$$

$$= 40 - 3.92 \leq \mu \leq 40 + 3.92$$

$$= 36.08 \leq \mu \leq 43.92$$

### Note:

- **More precision:** smaller alpha (1.96 b/c bigger) → smaller range → bigger chance for mistake
- **Bigger sample:** more N = smaller range → smaller chance for mistake

### Power

$$\text{Power} = 1 - \beta$$

-critical point and onwards

→ power = the probability that I will refute  $H_0$

→ the standard: 80%

→ the further the  $\bar{x}$ 's are from each other, the greater the powers

### 3 factors in power

- 1) **N** – more N = more power
- 2) **Alpha** – the bigger the alpha is, the bigger the power is  
-i.e. the higher % for error (either 1% or 5%), the more chance for a mistake
- 3) **The mean** -  $\bar{x}$

→ **Significant results:** a result w/I alpha → we accept  $H_0$

→ **Non-significant results:** we don't refute  $H_0$

→ sometimes just called '**NS**'

### T-test

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

→ similar to Z-test

→ S instead of  $\sigma$

**Note:** S in T-score means something else

### The difference b/w T and Z tests:

- 1) T relates to a sample/Z related to a population

- 2) T test doesn't have a SD → when you don't know the SD, use the T-test
- a. B/c you only have the variance of the population, but not the SD of the population b/c you do not know the population

- T used ***Degrees of freedom***  
→ has its own 'T-score' table
- t average = 0
- normative distribution
- but flatter/wider than normal distribution
- the more df, the closer it looks like z score
- every df has its own t-score

### **Degrees of Freedom**

-the number of components that are free to change  
-there are n-1 degrees of freedom

→ the last variable has to even out the rest in order to make the average the average of the population

- at the end, we've got to look up t score in its chart  
  
→ since the book only gives us degrees of freedom in intervals of 10, we ***ALWAYS*** round ***down***. ***I.e.*** 39 gets rounded down to 30 and ***not*** to 40!!!  
→ that gives us the critical point

-once I plug in my values into the formula and get the t-score, I plug it into the T table. I ***ALWAYS*** round down – even if it is 0.69 – I round down to 0.60 and not to 0.70!!!

### **T-test for one average.**

#### **Example question**

-it is known that kids help their friends 16 times a week. it is also known that it is a normative distribution

- a researcher estimates that kids involved in learning centers will help their friends in a different way than not
  - he sampled 7 kids from learning centers

### **His results**

24  
23  
22  
18  
17  
16  
20

### **Step 1** –formulating the estimations

$$H_0: \mu=16$$

$$H_1 : \mu \neq 16$$

### **Stage 2-Level of certainty and conditions of acceptance**

- $\alpha = 0.05$
- 2-tailed test
- df
  - $= n-1$ 
    - $= 7-1$ 
      - $= 6$
- $t = \pm 2.447$ 
  - Area of acceptance  $-2.447 < t < +2.447$
  - Area of refutation  $t < -2.447$  or  $t > +2.447$

### **Stage 3**

$$T_{/x} = \frac{\bar{X} - \mu_{/x}}{\frac{\sigma^{\wedge}_s}{\sqrt{n}}}$$

$$\sigma^{\wedge}_s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

<b><u>Variable</u></b>	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
24	$24 - 20 = 4$	$4^2 = 16$
23	3	9
22	2	4
20	0	0
18	-2	4
17	-3	9
16	-4	16
		=58
<b><i>Average=20</i></b>		

$$\sigma^{\wedge}_S = \sqrt{58/19}$$

$$T_{/x} = 20 -$$

**\*\***

## tutorial, mar 20, 2001

### Mistakes

Alpha –when  $H_1$  is better, but I don't say so

Beta –I dismiss  $H_1$  mistakenly  $\rightarrow H_1$  is really better.

Power  $1=\beta$  –probability that  $H_1$  is correct

$Z_c$  – critical Z –also called *critical value*

### Example:

-average of 1 years old kids =100 words

-SD=18

$\mu=100$

$\sigma=18$

-researcher claims that speaking to kid in-uterus increases vocabulary

### Experiment result

$N=81$

-at age 3  $\bar{x}=105$

--

5) what are the chances for a mistake

### 2 tailed mistake

-almost the same as 1 tailed

-uses same # as b/f

1)  $H_0=H_1 \rightarrow$  as opposed to either equal or  $>$  in 1 tailed

2)  $H_0=H_1 \rightarrow$  not  $H_0 < H_1$

$\rightarrow$  we need  $Z_c$  for this

$H_0$  acceptance rate:  $-1.96 \leq Z \leq 1.96$

$H_1$  acceptance rate:  $-1.96 \geq Z \geq 1.96$

### Formula

$$Z_{/x} = \frac{\bar{x} - (\mu_{/x})}{\sigma_{/x}}$$

$$S_{/x} = (\text{SD of a sample}) = \sigma / \sqrt{n}$$

$\rightarrow$  formula is the same as b/c

-in our case, it is the same as b/f =2.5

$\rightarrow$  we still refute  $H_0$

### Stages/process of experiment

1) wording of hypothesis:

$H_1 =$  knew more ( $\mu < \bar{x}$ )

$H_0$  =babies didn't know more ( $\mu \geq H_0$ )

2) establishing the level of certainty

**Standard** = Alpha =0.05 –unless otherwise states  
-->in this case, it is 1-tailed

**$Z_c$**  =critical Z

-the point that divides  $H_0$  into 2 sides -->the line that alpha represents

-->by looking up alpha % in the z-score chart

- In 1 tailed =1 critical Z
- In 2-tailed =2 critical Z
  - In 2-tailed – they are equal in size, but opposite signs

### **Results**

- higher than z score =  $H_0$  is refuted
- lower than z-score =  $H_0$  is accepted

**Note:** in all of this process,  $H_0$  is refuted or accepted  
-->not  $H_1$

### **Statistical computation**

#### **Formula**

$$Z_{/x} = \frac{\bar{x} - \mu_{/x}}{\sigma_{/x}}$$

$$\frac{105-100}{\sigma/\sqrt{n}}$$

$$\frac{105-100}{18/\sqrt{81}}$$

$$\frac{105-100}{2}$$

$$=5/2$$
$$=2.5$$

### **3) conclusion**

**i.e.** is  $H_0$  refuted or not>

$$2.5 > 1.9$$

--> $H_0$  is refuted

- also, write out in full the conclusion

### **Class –Mar 26 2001**

#### **T-tests**

-in reality, you do not know SD of a population/sample that you're experimenting. That is why you need to *estimate* it. This is called  $S$

#### **T test factors**

##### **Assumptions**

- 1) Normative distribution
- 2)  $N_1 = N_2$  (2 groups are the same N)
- 3) Homoskedasticity  $S_1^2 = S_2^2$   
→ see down for details

#### **2 kinds of T-Tests**

T tests for:

- 1) dependant tests
- 2) independent tests

#### **Dependant**

\*\*

#### **Independent**

\*\*

i.e. comparing:

- Where I live [city vs. village] (independent)
- Intelligence

→ I find that City IQ is 103 and village IQ is 102

The question reminds: since the diff. is so small, can I assume that they are the same population, or does this, seemingly insignificant diff. really make them a diff. population?

#### **2 independent tests**

-i.e. is teaching statistics class on Monday better than Tuesday (or vice versa)?

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{x1-x2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

#### **In this test:**

$$H_1 =$$

- $\mu_1 = \mu_2$
- $\mu_1 - \mu_2 = 0$

**This formula gets you  $S^2$  from the two  $s$ 's (of the two independent samples/tests)**

$$S^2 = \frac{S_1^2(N-1) + S_2^2(N-1)}{N_1 + N_2 - 2}$$

### **Matched tests**

- When you want to match 2 similar groups, except 1 criteria
  - Examples:
    - before/after.
    - Smokers in one neighborhood [w/ same history] vs. non smokers in that neighborhood w/ that same history]

### **T-test Formula**

$$T = D / S_d$$

-D = difference

**Reminder:**  $D = \sum d / N$

Formula to find  $S_d$

$$S_d = \frac{\sqrt{\sum (D - D)^2}}{\sqrt{N-1}}$$

### **Example**

<b><i>Before</i></b>	<b><i>After</i></b>	<b><i>Difference</i></b>	<b><i>(D-D)^2</i> (for the <math>S_d</math>)</b>
8	7	+1	$(+1 - (-1))^2 = 4$
9	10	-1	$(-1 - (-1))^2 = 0$
8	8	0	$(0+1)^2$
6	10	+4	$(-4 - (-1))^2 = 9$

$$T = D / S_d$$

$$D = -1$$

$$S_d = S_d / \sqrt{n}$$

Formula to find  $S_d$

$$S_d = \frac{\sqrt{\sum (D - D)^2}}{\sqrt{N-1}}$$

$$= \sqrt{14/4 - 1}$$

$$= \sqrt{14/3}$$

**To check homoskedasticity: [see top for t-test conditions]**

$$F = S_1^2 / S_2^2$$

**Important:** the higher S is always on top!!!!

-the answer is looked up in the F chart. The top # is for 1% the bottom is for 5%.

[ Homoskedasticity  $S_1^2 = S_2^2$  ]

**Tutorial –mar 27, 2001**

- The averages of the all the sampling combinations' averages gives us the actual average ( $\mu$ )

→ not so for the SD → can't find it from sampling → need a formula

→ **confidence interval** – to know  $\mu$  (average of population → I can take 1 sample and define a certain range around which  $\mu$  is likely to fall, to a certain percentage (usually 95%))

- chances to be right: 1-alpha

**Formula**

$$-p\{L_1 \leq 0 \leq L_2\} = 1 - \alpha$$

L=limits of confidence

- usually, we say that the confidence level is 95%
  - we say that we are 95% sure that  $\mu$  is w/I this limit

-when the distribution is taller/wider → diff. placing of the confidence interval

$$S_{/x} = \sigma / \sqrt{n}$$

$$Z_{\alpha/2} = 1.96$$

**Example:**

$$/x = 1500$$

$$n = 400$$

$$\sigma = 300$$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

**Formula**

$$/X - Z(\alpha/2) [1.96] \sigma / \sqrt{n} [S_{/x}] \leq \mu \leq /X + 1.96 S_{/x}$$

(1.96 = z score for 5%/2)

$$S_{/x} = s / \sqrt{n}$$



$$\{1500 - 1.96 \times 300/20 \leq \mu \leq 1500 + 1.96 \times 300/20\} = 95\%$$

**Example:**

$$\bar{x} = 115$$

$$n = 400$$

$$\sigma = 15$$

$$\alpha = 0.01$$

$$\alpha/2 = 0.005 \rightarrow 2.58$$

**Note:**

- the bigger the sample is, for the same confidence interval (say, 95%), will b/c smaller

**Example**

**Case** We know the range of confidence, but we don't know how big the sample has to be.  $\rightarrow$  the more people  $\rightarrow$  the more sure I am

- sometimes, I want a certain amount of participants, in order to get a specific range
  - what n do I need for this?

$$\bar{x}_1 - \bar{x}_2 = 40$$

$$n = ?$$

$$\sigma = 15$$

$$\alpha = 0.01$$

$$\alpha/2 = 0.005 \rightarrow 2.58$$

**Original Formula**

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

-Everything we will do is a recombination of this formula

- now –we have to isolate n to find out what n is

**Original Formula**

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$Z \times \sigma/\sqrt{n} = \bar{x} - \mu$$

$\bar{x} - \mu$  will hence be called L/2 [l=length of interval. L/2 is the range of one side of the distribution]

$$\frac{1}{\sqrt{n}} = \frac{L/2}{Z \times \sigma}$$

$$\sqrt{n} = \frac{Z \times \sigma}{L/2}$$

$$N = \frac{(Z \times \sigma)^2}{a (L/2)^2}$$

--

Dana  
6969-517

Prof. Koslowsky -12-1

--

### **Tutorial, April 17, 2001**

<b><u>Sample</u></b>	<b><u>Population</u></b>
$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$ $= \hat{S}^2$	$\sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$

### **Review of T-test**

#### **Independent T-test**

-to compare averages of 2 samples that are independent  
→ i.e. 2 people

#### **Example:**

-New principal in a school claims that the SAT scores are diff. than the one in a diff. highschool.

-inspector checked 18 students from the first high school. W/

$$S^2 = 290$$

$$\bar{X} = 85$$

$$N = 18$$

Other highschool

$$N = 12$$

$$S^2 = 340$$

$$\bar{x} = 80$$

#### **Conditions for T-test**

- normative distributive

- when we don't know the variance, we use t test. If we do, then we use the z-score
- $\bar{d}$  = the diff b/w each all the  $\bar{X}_1$  and  $\bar{X}_2$ 's
  - that distribution is the distribution of the  $\bar{x}$  of differences.
    - If there is no diff, then the  $\bar{x}$  of this distribution is 0

### Steps

- 1) Phrase the hypothesis
  - a.  $H_0 : \mu_1 = \mu_2$
  - b.  $H_1 : \mu_1 \neq \mu_2$
- 2) Declaring the level of significance:
  - Refutation/acceptance level: 0.05%
  - 2 tailed test
  - $df = n_1 + n_2 - 2$
- 3) Statistical computation

$$\frac{\bar{X}_1 - \bar{X}_2}{S/\bar{d}}$$

Formula to find  $S_{\bar{d}}$

$$S_{\bar{d}} = \sqrt{\frac{\sum (D - \bar{D})^2}{N-1}}$$

### T-test Formula

$$T = D / S_d$$

-D = difference

**Reminder:**  $D = \sum d / N$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2 / D}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

$$\sqrt{S^2 / D} = \sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$$

To find  $S^2$ :

$$S^2 = \frac{S_1^2(N_1 - 1) + S_2^2(N_2 - 1)}{N_1 + N_2 - 2}$$

4)

85-80

6.76

### Dependant T-tests

-a Diff b/w dependant and independent

- 1) how  $D$  is computed
- 2) I use  $d \rightarrow$  there is 1 average, since here I am comparing pairs [and therefore, I use  $n-1$ ]
  - a. vs. independent t-tests, where we related to all the observed variables (from both groups) but got 2 d averages ( $n-2$ )

A.  $H_0 : \mu_1 = \mu_2$

B.  $H_1 : \mu_1 \neq \mu_2$

<i>Before</i>	<i>After</i>	<i>Difference</i>	$(D - \bar{D})^2$ (for the $S_d$ )
8	7	+1	$(+1 - (-1))^2 = 4$
9	10	-1	$(-1 - (-1))^2 = 0$
8	8	0	$(0 + 1)^2$
6	10	+4	$(-4 - (-1))^2 = 9$

### Formula for dependant t-score

$$T = D / S_d$$

-D = difference

**Reminder:**  $D = \sum d / N$

Formula to find  $S_d$

$$S_d = \sqrt{\frac{\sum (D - D)^2}{N - 1}}$$

**April 23, 2001**

***Non-parametrical test:*** \*\*

→ when you can't establish a parameter

→ nominal/ordinal

→ i.e. chi-square

→ therefore, you can't use things like mean in chi-square tests

**Chi-square**

***Chi-square:*** is the observed (o) diff or similar to the estimated (E) distribution

- defined by the DF
- the larger the DF, the bigger the critical point is (as opposed to T, which is the opposite)
- Unsymmetrical → left skew
  - but gets closer to symmetrical as it gets larger
- Chi-square is always 1-tailed (5%)
  - B/c in Chi-square, the chart really shows diff. b/w 2 samples. The more left, the less diff. b/w the 2 samples/population

***Chi-square:*** is the observed (o) diff or similar to the estimated (E) distribution

-If I have a probability and I want to compare it to an actual result

**i.e.** if I throw a coin 100 times; I expect to get 50 head/ 50 tails

→ if I get 55 heads?

$$\chi^2 = \sum \frac{(O_1 - E_1)^2}{E_1}$$

E = expected

O = observed

**→ then plug # into Chi-square** → the cross-section of DF and the 5% critical threshold

→if the answer of the formula is bigger than the critical value (the # that appears in the cross-section of the DF and the 5%) →H<sub>0</sub> is accepted

→if answer of formula is *smaller* than the cross-section, than H<sub>0</sub> is refuted

**Note: in Chi**

**Df**=categories -1

→as opposed to t, where DF is # of participants (n) -1

→i.e. in dice, there are 6 categories, so the DF = 5

**Results**

H<sub>0</sub> O=E

H<sub>1</sub> O≠E

**Example:**

We got 55 heads

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

$$= \frac{(55 - 50)^2}{50} + \frac{(45 - 50)^2}{50}$$

$$= 25/50 + 25/50 = 1$$

-look up 1 in the *chi-square* chart

**Example of CHI**

In a class, there were 6 men and 20 females.

→is this a representative population of the gender split of 50-50%

→we expected 13 males/13 females

**Stage 1 = hypothesis**

H<sub>0</sub> e = o

H<sub>1</sub> e ≠ o

**Stage 2 –Deciding the level of significance/refutation**

-σ=0.05

-The test is 2-tailed but the actual verification is 1-tailed

→no point in checking in the left side, since Chi graph merely shows the diff.

→the more right, the more diff.

$$DF = j - 1 = 2 - 1$$

$$\chi^2_c = 3.84$$

- acceptance of  $H_0$   $3.84 \geq \chi^2$
- Refutation of  $H_0$ :  $3.84 < \chi^2$

### Stage 3

O	E	(O-E) <sup>2</sup>							
6	13	49							
20	13	49							

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

$$\frac{(6-13)^2}{13} + \frac{(20-13)^2}{13}$$

$$49/13 + 49/13 = 7.54$$

→ check the table

### Chi test of independence

-the connection b/w 2 variables, where at least one of them is in the nominal

→ this test will check if there is a connection b/w the 2 groups

→ compare the expected and the observed

$$\chi^2_{(j-1)(k-1)} = \frac{(O_i - E_i)^2}{E_i}$$

### Goodness of fit

-is there a dependence/relationship b/w the 2 variables?

→ i.e. gender and subject

	Economics	Psychology	Computers	Total
Male	50 (e=47)	40 (e=50)	50 (e=33)	140
Female	50 (e=53)	80 (e=70)	30 (e=47)	160
Total	100	120	80	300

→ I check it through the CHI-square formula

→ the bigger the gap b/w expected and observed, the more I have to say that they are dependant. If expected = observed, then there are no external factors.

**Now the question is:** we know the observed, but what is the expected?

### Answer:

Total of row times total of column divided by total of all rows/columns

i.e. to find the expected number of male psychology students

$\frac{140 \text{ (total males)} \times 100 \text{ (total psychology students)}}{300 \text{ (total of all)}}$

$\frac{14000}{300}$   
 $\frac{140}{3}$

**Expected male psychologist students** = 46.667

-now we want to know if the 2 populations are *really* diff.  
→ we need to find if it passed the critical value

**To get degrees of freedom:**

$(R-1)(C-1)$

R = # of rows; C = columns

→ then plug in results into the Chi-square formula  
→ then look it up in the chi-square chart (also taking into account the degrees of freedom)

**Chi-squared:**

- not normative
- chi-square is a frequency [vs. t/z score which is a measure. (has a mean/SD)]

**example**

-a researcher claims that there is a correlation b/w location of residence and love of the sea

-he gets 310.

-for each participant, he gives the residence options: city/village/kibbutz

→ the sea options are love/hate

	Loves sea	Hates sea	
City	65	38	
Kibbutz	57	58	
Village	37	55	
Total			310

**Stage 1 stating of hypothesis**

$H_0: e = 0$

$H_1: e \neq 0$

**Stage 2**

$Df = (R-1)(C-1)$

$3-1 \times 2-1 = 2$



→look up in chart:

- df=2 and 5%
  - $\chi^2_c = 5.99$

1) First of all, plug in the totals.

→assumption: in both columns and rows: both expected and observed add up to the same total.

2)  $E = R \times C / n$

	Loves sea	Hates sea	
<b>City</b>	65 (E= $R \times C / n$ ) (E= $159 \times 103 / 310$ ) E=52.83	38 (E=50.17)	103
<b>Kibbutz</b>	57 (E=58.9)	58 (E=56)	115
<b>Village</b>	37 (E=47.1)	55 (E=44.8)	92
<b>Total</b>	159	151	310

Then , plug in the formula:

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

Answer =10.26

→look up in chart =we refute  $H_0$  =there is a diff.

#### **When you have a 2X2 box (the 2 categories both have 2 components)**

-a shortcut formula, though the observed/expected chart seen above could also be used  
-df is always 1

	C <sub>1</sub>	C <sub>2</sub>	
R <sub>1</sub>	A	B	A+B
R <sub>2</sub>	C	D	C+D
	A+C	B+D	N

$$\chi^2 = n \frac{(AD - BC)^2}{(A+B)(A+C)(C+D)(B+D)}$$

#### **Example**

A researcher claims that there is a diff b/w way of learning A and B

	Success C <sub>1</sub>	Failure C <sub>2</sub>	
A	50	10	60

	70	30	100
--	----	----	-----

$$\chi^2 = \frac{100(50 \times 20 - 20 \times 10)}{60 \times 40 \times 30 \times 70}$$

### Goodness of fit

- **If there are only 2 categories:**
  - In the expected, we need at least 10
    - Observed # is irrelevant
- **More than 2 categories**
  - at least 5 in expected

### Independent:

- at least 10 in the expected in a 2x2 box
- 5 in the expected in more than a 2x2 box

-when O and E are given in %, and not in # of people (frequency) then you multiply the % by n

### F test –

- *Levine*
- **Analysis of variance -(nituach shonut)**

F is always a factor

F =  $\frac{\text{variance b/w groups}}{\text{Variance w/i group}}$

- each of one has a df
- the more df, the more symmetrical the f-distribution looks
- the more df, there are, the critical f b/c smaller
- 
- → opposite to chi-square
- Possible range for  $\infty \geq f \geq 0$
- 

-in f chart – 2 numbers – the top # is for 5% and the bottom is for alpha of 1%

→ it is always 1 tailed, since this distribution is right-skewed

-I want the numerator (B) to be larger than the denominator (W) → so the diff. b/w populations is bigger than the random ones w/i each population

### Questions

What is the diff. b/w Mitgam/Dgima

\*\*

### May 6 – Stats makeup class

**1 tailed test:** when you know where  $H_1$  is supposed to fall

**2-tailed test:** when you don't know for sure where  $H_1$  will fall:

- **Meni:** you always need 2-tailed, since you need to always have that option open.

### T-test to check the certainty of Pearson

-we want to see if  $r$  (correlation) is big enough.

### Example:

In the whole population, it is known that there is no relationship b/w shoe-size and body weight

→ an anthropologist claims that in the Zulu tribe, there exists a correlation b/w shoe-size and weight

→ he checks a sample of 20 and comes up with a correlation of 0.60 ( $r=0.60$ )

→ now, we want to know if  $r=0.60$  is significant

→ so we're going to do a t-test in order to see whether  $r$  is significant

$H_0$  = no connection

$H_1$  = yes connection

**Note:** here,  $H_0$  and  $H_1$  is not a diff. population/sample, but rather a difference in correlation

$N=20$

$R=0.60$

### Stage 1 –statistical hypothesis

→ says some estimation of a parameter of a population (as opposed to as sample)

$\ell=Roe$  = parameter (measure of a population) of a population (as opposed to sample - statistic)

$Roe=0$

$H_0$  = no connection

$H_1$  =yes connection

### **Stage 2 –Deciding the level of significance**

-Declaring refutation %

- $\sigma = 0.05$
- 2 tailed
- $DF = n-2 = 18$
- $T_c = 2.101$ 
  - $-2.101 \leq T \leq 2.101$
  - areas of refutation:  $T > 2.101$  and  $T < -2.101$

### **Stage 3 -Statistical computations**

$$T(n-2) = r \times \frac{\sqrt{(n-2)}}{\sqrt{1-r^2}}$$

=3.18

→  $H_0$  is refuted

#5

14 # 4

11

16 #8

19 →but not #5

**Descriptive statistics:** stats that describe the variables, such as mean/mode/SD/DF/frequency

-unless stated, it is 2-tailed

-know 1.96 =5%

-the formula for SD of population/variance

-unless alpha is stated, stated, it is 5%

-division of normative distribution: 34,13,2

### **Analysis of variance**

-checks estimates in relations b/w diff. in variances

-diff in population averages

→called *variance tests*

- 1 directional: -there is 1 independent variable

→ or 1 directional

→ does ethnicity affect anxiety

- 2-directional: there are 2 independent variable:  
→ i.e. does gender and ethnicity effect anxiety

i.e. I have 3 grade 5 math teachers. Marks of the 3 teachers:

$A_1$   $A_2$   $A_3$

9 5 10

7 6 9

9 5 9

7 6 10

-**Note:** depending on the experiment we might want to refute **OR** accept  $H_0$

**Example**

$A_1$	$A_2$	$A_3$	$A_4$
1	2	4	6
1	1	3	6
2	3	2	5
4	6	3	3
--	--	--	--
2	3	3	5 (/X)

-average of the averages

$$=13/4$$

$$=3.25$$

**Note:**

-pfirst participant is called:  $X_{11}$

-second one (the one below him in the first group) is called  $X_{21}$

$X_{\bullet 1}$  → the dot relates to all in that category → all I's in J #1

$X_{1\bullet}$  → all I = 1 across groups (Js)

$X_{\bullet\bullet}$  → all I and all J (i.e. /X)

$X_{IJ}$

→ I = individual

→ J = group

**Formula**

$$S^2_B$$

$$S^2_w$$

B = between

W= within

Note: 2 kinds of variance:

- '**Between**' variance –between groups
- '**Within**' variance –within each group

**Question of researcher:** why did everyone not get the average?

**Answer:** explained variance, such as 'environment/genes/etc.'

- if B is bigger than W, then there is population diff.
- If W is bigger than B, then you know that the diff. (*variance*) is individual and not population-differences.

→therefore, to establish a significant diff. b/w populations, then:

$$\frac{S^2_B}{S^2_w} > 1$$

Source	Ss (sum of squares between groups)	DF	MS (mean square)  SS/DF	F
<b>B</b> (between)		K-1  (K =# of groups)		
<b>W</b> (within)		K(n-1) -# of group x number in each group minus 1  =12		
<b>T</b> (total)	W+B	N-1		

$$\sum \sum (X_{IJ} - \bar{X})^2 = \sum \sum (X_{IJ} - \bar{X}_I)^2 + \sum n_j (\bar{X}_j - \bar{X})^2$$

-Total                  Within                  Between

J=column

I –row/individual

**MSB =SSB/DF<sub>b</sub>**

$$SSB = \sum n_j (\bar{X}_j - \bar{X})^2$$

$$DF = k-1$$

$$\underline{\text{MSW} = \text{SSW} / \text{DF}}$$

$$\text{SSW} = \sum \sum (X_{ij} - \bar{X}_i)^2$$

$$\underline{\text{MST} = \text{SST} / \text{DF}_t}$$

$$\text{SST} = \sum \sum (X_{ij} - \bar{X})^2$$

$$\text{SST} = \text{SSB} + \text{SSW}$$

$$\text{DFT} = \text{DFB} + \text{SDW}$$

### Example

The researcher wants to see diff. lighting affects on plants

-12 plans randomly divided into 3 groups

A<sub>1</sub> –regular light conditions

A<sub>2</sub> – in natural dark

A<sub>3</sub> –in artificial light

A<sub>4</sub> –artificial dark

	<u>A<sub>1</sub></u>	<u>A<sub>2</sub></u>	<u>A<sub>3</sub></u>	<u>A<sub>4</sub></u>
<b>1</b>	1	2	4	6
<b>2</b>	1	1	3	6
<b>3</b>	2	3	2	5
<b>4</b>	4	6	3	3
<b><u>Total's</u></b>	2	3	3	5
<b><u>average</u></b>				

-average of A<sub>1-4</sub> = 3.25

$(1-3.25)^2 + (2-3.25)^2 \rightarrow \text{etc...}$

-some of the diff. is b/c of their groups (explained variance)

-called **effect** (B)

-other diff. if b/c individual diff.

→the researchers say that this diff. is a **mistake** (W) →we can't control it.

-variance b/w groups (w)[ the numerator] is Mistake+effect

Mean variance between = MSB

Mean variance within = MSW

**Ss** =sum of squares between groups

SS/DF = MS

SSB/DF =MSB  
SSW/DF =MSW

<u>Source</u>	<u>SS</u>	<u>DF</u>	<u>MS</u>	<u>F</u>	<u>P</u>
<u>B</u>	SSB	n-1 =3	SSB/k-1	MSB/MSW	P<0.05
<u>W</u>	SSW	K(n-1)	SSW/N-K		
<u>T</u>	SST		SST/n-1		

H<sub>0</sub> = all averages are equal

H<sub>1</sub> =if even 1 is not equal, than H1 is true

-the answer is looked up in the F chart. The top # is for 1% the bottom is for 5%.

==

Sheffe\*\*

-post-hoc: after the fact

-once I found that there is a significant diff. I want to know which one is the diff. one out of one of the group.

-Only works when the result is significant!!

-You take out 2 columns randomly and compare them, using a formula

-I know that the diff. b/w the top and bottom one is significant b/c the result is significant.

-if the answer is above critical value x k-1 →then that is the diff. isn't significant

→if the answer is lower, than it is significant

→there could be several significantly diff. variables

### 2-way Analysis of variance (2-way Anova (analysis of variance))

-2 factors

→i.e. if there is a diff. of mean of groups when there is 2 independent\ factors to 1 dependant factors

Example:

-An experimenter wanted to know which kinds of psychotherapies are appropriate for which disorders.

-he turns to dynamic and cognitive therapists to give info about 2 kinds of patients: neurosis and personal problems

-for each patient, the subjective improvement was measured in a range of 0-10

-subjective view –dependant

-kind of therapy: independent



	<i>Neurotic</i>	<i>Personality problems</i>	<i>Average</i>
<i>Dynamic</i>	9	5	7
<i>Cognitive</i>	6	6	6
<i>Total:</i>	7.5	5.5	6.5

**Effect:** the diff. b/w means

1) **Main effect** –the influence of the independent on the dependant beyond the other independent variables (i.e. the diff. b/w row average and columns)

→i.e. the contrast b/w therapy and disorder, regardless of the kind of disorder or therapy

- **kind of therapy:** is there improvement in those who went to cognitive vs. those who went to dynamic? (regardless of the kind of disorder)

→**note:** it gives is a relations, but not causality

→you merely check the total to see which is bigger, the cognitive or dynamic

- **kinds of disorders**

→i.e. run a t-test b/w the columns

2) **Interaction effect**

- things that we didn't expect
- The interaction (diff.) of the 2 independent variables to each other and their effects on the dependent variable

- Example: is there a diff in differences b/w the rows/ columns (Depends on what you're comparing). If there is no diff., there is no interaction

→Is there on a differential effect of the therapy on the subjective feeling, as a function in the kind of therapy/disorder

→is there something special?

-i.e. if you chart the progress of the 2 variables whatever you're looking at, any 2 that are not parallel is an interaction

-Question on test: Experiment 2X3X4 =there are 3 factors. One of them has 2 variables, one has 3 and one has 4.

### **Formula for 2 way ANOVA**

$$\frac{\sum \sum (X_{irc} - \bar{X}_{...})^2}{\sum \sum (X_{irc} - \bar{X}_{rc})^2} = nC \sum (\bar{X}_{r.} - \bar{X}_{...})^2 + nR \sum (\bar{X}_{.c} - \bar{X}_{...})^2 + n \sum \sum (\bar{X}_{rc} - \bar{X}_{r.} - \bar{X}_{.c} + \bar{X}_{...})^2$$

	SS	DF	MS	F
<b>Rows</b>				
<b>Columns</b>				

Interaction				
Within				
Total				
<b><u>Explanation for Formula for 2 way ANOVA</u></b> $\sum \sum \sum (X_{irc} - \bar{X}_{...})^2 = nC \sum (\bar{X}_{.r} - \bar{X}_{...})^2 + nR \sum (\bar{c}_{..c} - \bar{x}_{...})^2 + n \sum \sum (\bar{X}_{rc} - \bar{X}_{.r} - \bar{X}_{..c} + \bar{X}_{...})^2 + \sum \sum \sum (X_{irc} - \bar{X}_{rc})^2$ n=the actual # in the variable. Not the # of variables **				

-analysis of variance w/ 2 variables

**Example of 2-way ANOVA**

	A <sub>1</sub>	A <sub>2</sub>	A <sub>3</sub>
B <sub>1</sub>	2 4	3 5	1 5
B <sub>2</sub>	1 1	4 7	7 4

**Note:**

-except Chi-square and spearman, the assumption is that the distribution is normal.

In 1-way Anova: SST = SSW+SSB

In 2-way Anova: SST = SSW+ SSB (SS<sub>int</sub> +SS<sub>r</sub> +SS<sub>c</sub>) interaction + column +row)

- n → I (number of variable)
- R → r (# of row)
- C → c (# of column)
- rci
  - rci –111 =2 in the box above
  - rci –121 =3
  - /X1.. →average of row

--

**SSW**

$$\sum \sum X_{rci} - \bar{x}_{rc}.)^2$$

\*\*

--

Fc	Fr	Fint

Source	Ss	Df	Ms	F	P
Row	SSr	r-1	SSr/r-1	MSR/MsW	
Column	SSc	c-1	SSc/c-1	MSC/MSW	
Interaction	Ssint	(c-1)(r-1)	SSint/r-1c-2	MSint/MSW	
	SSw	B-(rc)	SSw/n-rc		
Total	SST	n-1			

**Example:**

Researcher wanted to know which therapy is best for which people:

*Therapy:* Dynamic/cognitive

*Problem:* neurotic/personal

	Neurotic	Personality	Average
Dynamic	9	5	7
Cognitive	6	6	6
Total	7.5	5.5	6.5

**Stage 1 phrasing of hypothesis**

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

\*\*

**stage 2**

alpha is 5% -since no other info was given

2-tailed test, yet in practice it is 1-tailed due to its structure

**stage 3**

DFr = r-1

DFc = c-1

Dfint = n-rc

DFw =

F =