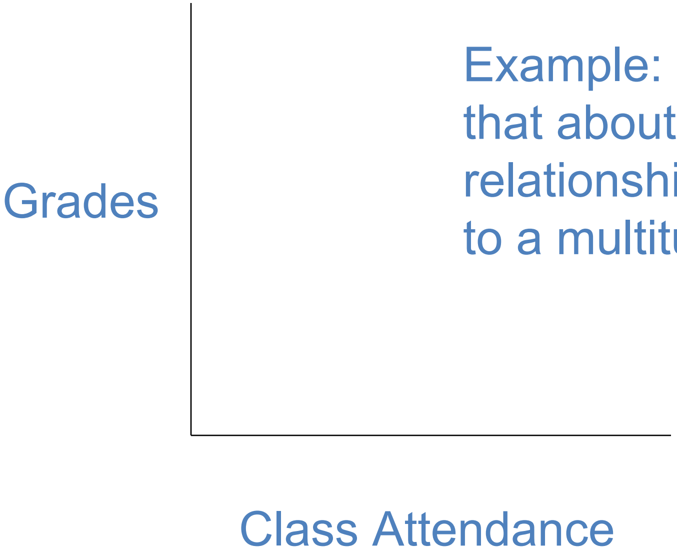
Statistics

From BSCS: Interaction of experiments and ideas, 2nd Edition. Prentice Hall, 1970 and Statistics for the Utterly Confused by Lloyd Jaisingh, McGraw-Hill, 2000

r^2…

- is the fraction of the variation in the values of y that is explained by the least-squares regression line of y on x.

Example: If $r^2 = 0.61$ in the graph to the left, this means that about 61% of one's grade is accounted for by the linear relationship with attendance. The other 39% could be due to a multitude of factors.



What is statistics?

- a branch of mathematics that provides techniques to analyze whether or not your data is significant (meaningful)
- Statistical applications are based on probability statements
- Nothing is “proved” with statistics
- Statistics are reported
- Statistics report the probability that similar results would occur if you repeated the experiment

Statistics deals with numbers

- Need to know nature of numbers collected
  - Continuous variables: type of numbers associated with measuring or weighing; any value in a continuous interval of measurement.
    - Examples:
      - Weight of students, height of plants, time to flowering
  - Discrete variables: type of numbers that are counted or categorical
    - Examples:
      - Numbers of boys, girls, insects, plants

Can you figure out…

- Which type of numbers (discrete or continuous?)
  - Numbers of persons preferring Brand X in 5 different towns
  - The weights of high school seniors
  - The lengths of oak leaves
  - The number of seeds germinating
  - 35 tall and 12 dwarf pea plants
  - Answers: all are discrete except the 2nd and 3rd examples are continuous.

Populations and Samples

- Population includes all members of a group
  - Example: all 9th grade students in America
  - Number of 9th grade students at DHS
  - No absolute number
- Sample
  - Used to make inferences about large populations
  - Samples are a selection of the population
  - Example: 1st period AP biology
- Why the need for statistics?
  - Statistics are used to describe sample populations as estimators of the corresponding population
  - Many times, finding complete information about a population is costly and time consuming. We can use samples to represent a population.

Sample Populations avoiding Bias

- Individuals in a sample population
  - Must be a fair representation of the entire pop.
  - Therefore sample members must be randomly selected (to avoid bias)
  - Example: if you were looking at strength in students: picking students from the football team would NOT be random

Is there bias?

- A cage has 1000 rats, you pick the first 20 you can catch for your experiment
- A public opinion poll is conducted using the telephone directory
- You are conducting a study of a new diabetes drug; you advertise for participants in the newspaper and TV
- All are biased: Rats-you grab the slower rats. Telephone-you call only people with a phone (wealth?) and people who are listed (responsible?). Newspaper/TV-you reach only people with newspaper (wealth/educated?) and TV(wealth?).

Statistical Computations (the Math)

- If you are using a sample population
  - Arithmetic Mean (average)

$$\overline{X} = \frac{\Sigma X}{N}$$

$$x = \{1, 2, 3, 4, 5\}; \bar{x} = 3$$

The sum of all the scores divided by the total number of scores.

  - The mean shows that ½ the members of the pop fall on either side of an estimated value: mean

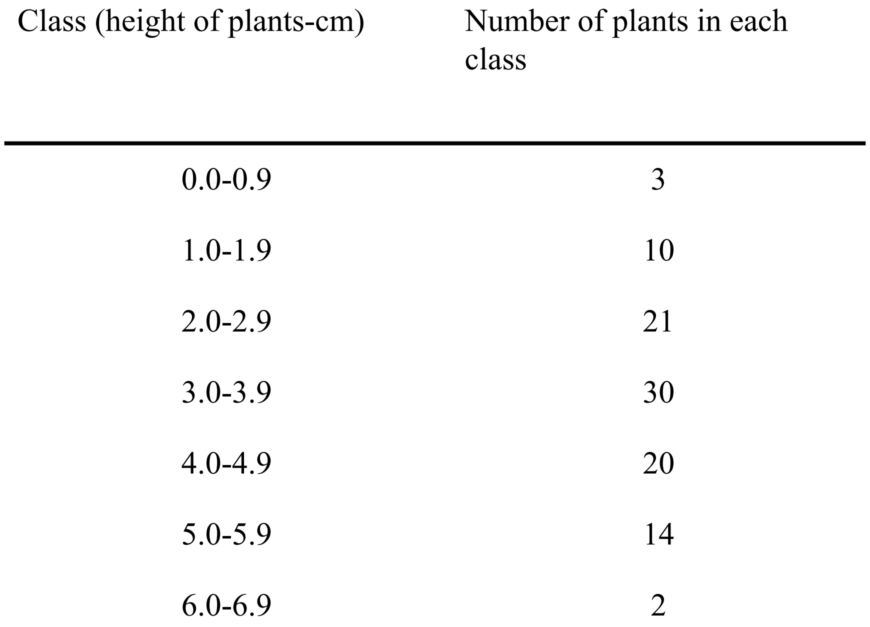
http://en.wikipedia.org/wiki/Table_of_mathematical_symbols

Distribution Chart of Heights of 100 Control Plants

Looking at profile of data: Distribution

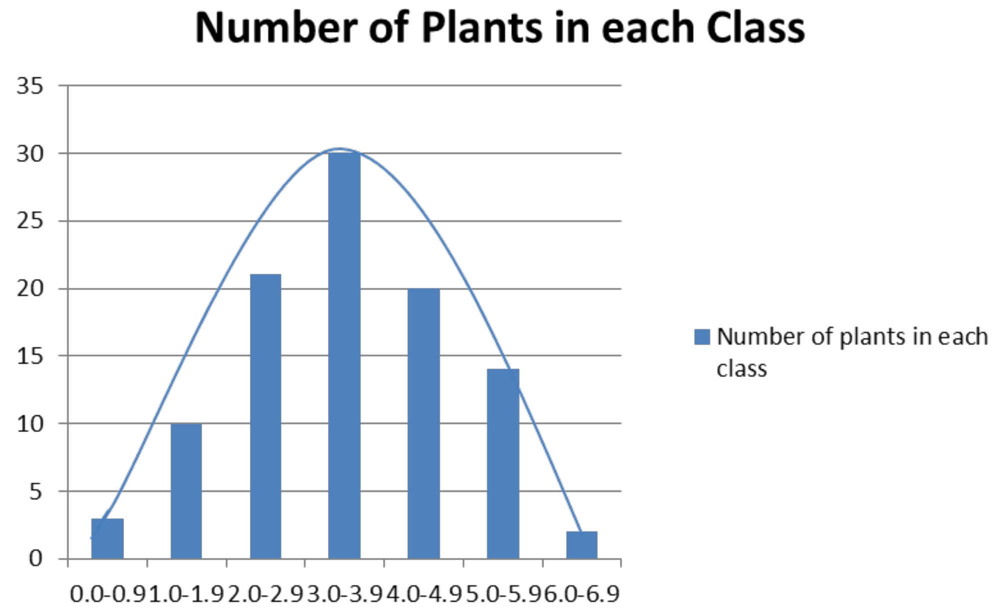
- What is the frequency of distribution, where are the data points?

Distribution Chart of Heights of 100 Control Plants



Class (height of plants-cm)	Number of plants in each class
0.0-0.9	3
1.0-1.9	10
2.0-2.9	21
3.0-3.9	30
4.0-4.9	20
5.0-5.9	14
6.0-6.9	2

Histogram-Frequency Distribution Charts



This is called a “normal” curve or a bell curve
This is an “idealized” curve and is theoretical based on an infinite number derived from a sample

Mode and Median

- Mode: most frequently seen value (if no numbers repeat then the mode = 0)
- Median: the middle number
  - If you have an odd number of data then the median is the value in the middle of the set
  - If you have an even number of data then the median is the average between the two middle values in the set.

Variance (s^2)

- Mathematically expressing the degree of variation of scores (data) from the mean
- A large variance means that the individual scores (data) of the sample deviate a lot from the mean.
- A small variance indicates the scores (data) deviate little from the mean

Standard Deviation

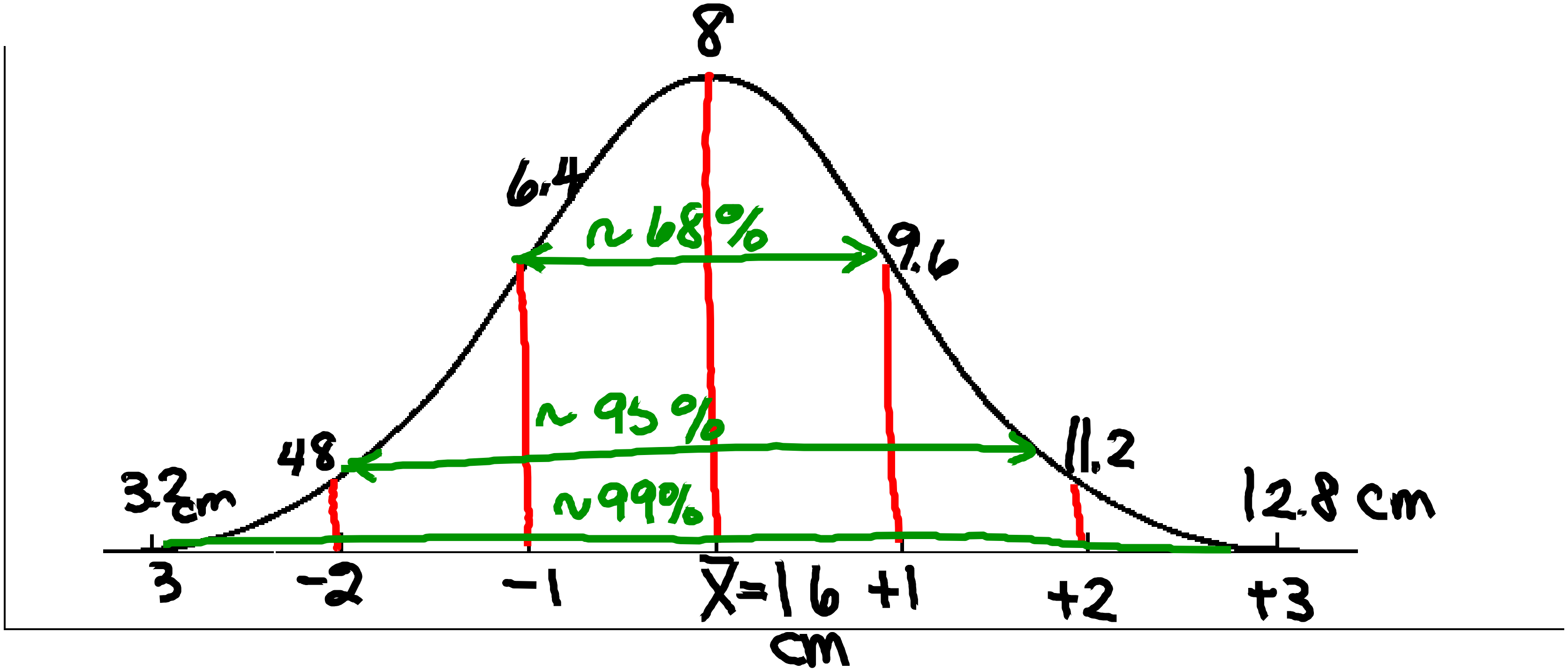
- An important statistic that is also used to measure variation in biased samples.
- S is the symbol for standard deviation
- Calculated by taking the square root of the variance
- So from the previous example of pea plants:

  The square root of 2.5 ; s=1.6
- Which means the measurements vary plus or minus +/- 1.6 cm from the mean

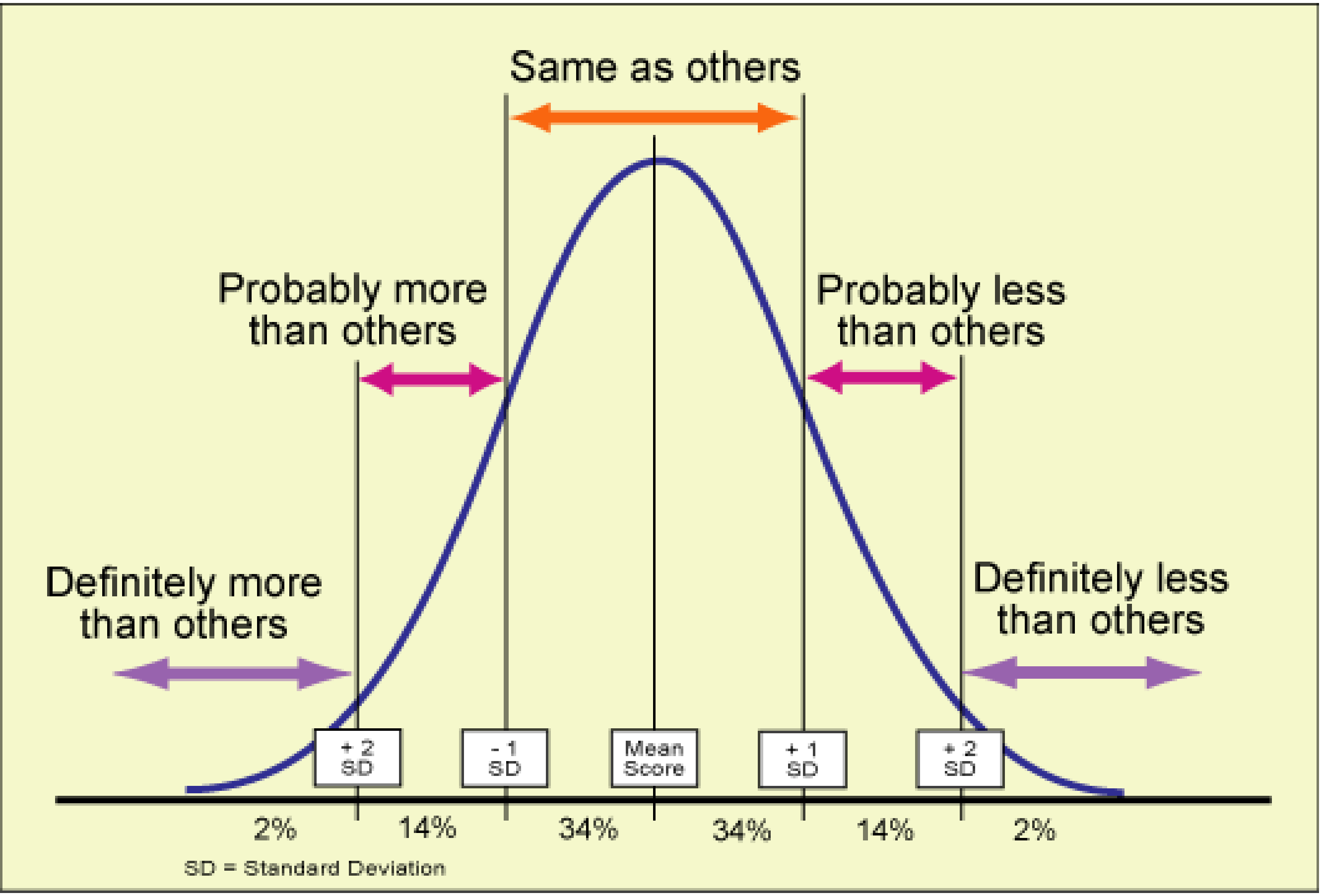
What does “S” mean?

- We can predict the probability of finding a pea plant at a predicted height… the probability of finding a pea plant above 12.8 cm or below 3.2 cm is less than 1%
- S is a valuable tool because it reveals predicted limits of finding a particular value

Pea Plant Normal Distribution Curve with Std Dev



The Normal Curve and Standard Deviation


Same as others
Probably more than others
Probably less than others
Definitely more than others
Definitely less than others
+ 2 SD
- 1 SD
Mean Score
+ 1 SD
+ 2 SD
2%
14%
34%
34%
14%
2%
SD = Standard Deviation

A normal curve:

Each vertical line is a unit of standard deviation

68% of values fall within +1 or -1 of the mean

95% of values fall within +2 & -2 units

Nearly all members (>99%) fall within 3 std dev units

http://classes.kumc.edu/sah/resources/sensory_processing/images/bell_curve.gif

Standard Error of the Sample Means AKA Standard Error

- The mean, the variance, and the std dev help estimate characteristics of the population from a single sample
- So if many samples were taken then the means of the samples would also form a normal distribution curve that would be close to the whole population.
- The larger the samples the closer the means would be to the actual value
- But that would most likely be impossible to obtain so use a simple method to compute the means of all the samples

A Simple Method for estimating standard error

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Standard error is the calculated standard deviation divided by the square root of the size, or number of the population

Standard error of the means is used to test the reliability of the data

Example… If there are 10 corn plants with a standard deviation of 0.2

Se_x = 0.2/ sq root of 10 = 0.2/3.03 = 0.006

0.006 represents one std dev in a sample of 10 plants

If there were 100 plants the standard error would drop to 0.002
Why?

Because when we take larger samples, our sample means get closer to the true mean value of the population. Thus, the distribution of the sample means would be less spread out and would have a lower standard deviation.

Probability Tests

- What to do when you are comparing two samples to each other and you want to know if there is a significant difference between both sample populations
- (example the control and the experimental setup)
- How do you know there is a difference
- How large is a “difference”?
- How do you know the “difference” was caused by a treatment and not due to “normal” sampling variation or sampling bias?

Laws of Probability

- The results of one trial of a chance event do not affect the results of later trials of the same event. p = 0.5 (a coin always has a 50:50 chance of coming up heads)
- The chance that two or more independent events will occur together is the product of their changes of occurring separately. (one outcome has nothing to do with the other)
- Example: What's the likelihood of a 3 coming up on a dice: six sides to a dice: p = 1/6
- Roll two dice with 3's p = 1/6 *1/6= 1/36 which means there's a 35/36 chance of rolling something else…
- Note probabilities must equal 1.0

Laws of Probability (continued)

- The probability that either of two or more mutually exclusive events will occur is the sum of their probabilities (only one can happen at a time).
- Example: What is the probability of rolling a total of either 2 or 12?
- Probability of rolling a 2 means a 1 on each of the dice; therefore $p = 1/6*1/6 = 1/36$
- Probability of rolling a 12 means a 6 and a 6 on each of the dice; therefore $p = 1/36$
- *So the likelihood of rolling either is* $1/36+1/36 = 2/36$ *or* $1/18$

The Use of the Null Hypothesis

- Is the difference in two sample populations due to chance or a real statistical difference?
- The null hypothesis assumes that there will be no “difference” or no “change” or no “effect” of the experimental treatment.
- If treatment A is no better than treatment B then the null hypothesis is supported.
- If there is a significant difference between A and B then the null hypothesis is rejected...

T-test or Chi Square? Testing the validity of the null hypothesis

- Use the T-test (also called Student’s T-test) if using continuous variables from a normally distributed sample populations (ex. Height)
- Use the Chi Square (X^2) if using discrete variables (if you are evaluating the differences between experimental data and expected or hypothetical data)… Example: genetics experiments, expected distribution of organisms.

T-test

- T-test determines the probability that the null hypothesis concerning the means of two small samples is correct
- The probability that two samples are representative of a single population (supporting null hypothesis) OR two different populations (rejecting null hypothesis)

STUDENT’S T TEST

The student’s t test is a statistical method that is used to see if to sets of data differ significantly.

The method assumes that the results follow the normal distribution (also called student's t-distribution) if the null hypothesis is true.

This null hypothesis will usually stipulate that there is no significant difference between the means of the two data sets.

It is best used to try and determine whether there is a difference between two independent sample groups. For the test to be applicable, the sample groups must be completely independent, and it is best used when the sample size is too small to use more advanced methods.

Before using this type of test it is essential to plot the sample data from he two samples and make sure that it has a reasonably normal distribution, or the student’s t test will not be suitable.

It is also desirable to randomly assign samples to the groups, wherever possible.

Read more: http://www.experiment-resources.com/students-t-test.html#ixzz0OIl72cbi

http://www.experiment-resources.com/students-t-test.html

EXAMPLE

You might be trying to determine if there is a significant difference in test scores between two groups of children taught by different methods.

The null hypothesis might state that there is no significant difference in the mean test scores of the two sample groups and that any difference down to chance.

The student's t test can then be used to try and disprove the null hypothesis.

RESTRICTIONS

The two sample groups being tested must have a reasonably normal distribution.

If the distribution is skewed, then the student's t test is likely to throw up misleading results.

The distribution should have only one mean peak (mode) near the center of the group.

If the data does not adhere to the above parameters, then either a large data sample is needed or, preferably, a more complex form of data analysis should be used.

Read more: http://www.experiment-resources.com/students-t-test.html#ixzz0OlllZOPZ

http://www.experiment-resources.com/students-t-test.html

RESULTS

The student's t test can let you know if there is a significant difference in the means of the two sample groups and disprove the null hypothesis. Like all statistical tests, it **cannot prove** anything, as there is always a chance of experimental error occurring.

But the test can support a hypothesis. However, it is still useful for measuring small sample populations and determining if there is a **significant difference** between the groups.

by Martyn Shuttleworth (2008).

Read more: http://www.experiment-resources.com/students-t-test.html#ixzz0OlmGvVWD

http://www.experiment-resources.com/students-t-test.html

Use t-test to determine whether or not sample population A and B came from the same or different population

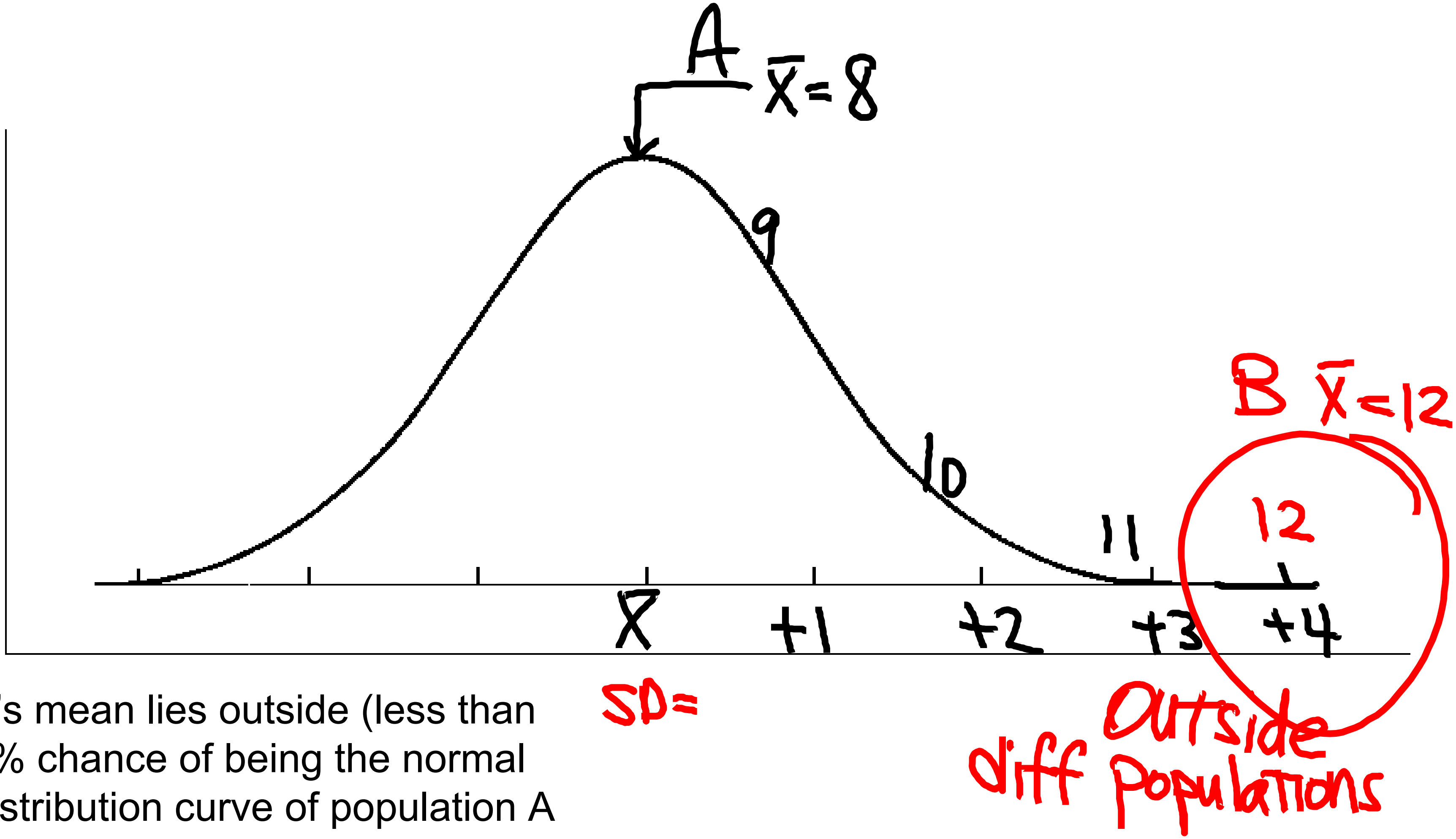
t = x1-x2 / sx1-sx2

x1 (bar x) = mean of A ; x2 (bar x) = mean of B
sx1 = std error of A; sx2 = std error of B

Example: Sample A mean =8
Sample B mean =12
Std error of difference of populations =1

12-8/1 = 4 std deviation units

Comparison of A and B



B's mean lies outside (less than 1% chance of being the normal distribution curve of population A

Reject Null Hypothesis

Online calculators:

http://www.physics.csbsju.edu/stats/t-test_bulk_form.html
online calculates for you… and a box plot also
http://www.graphpad.com/quickcalcs/ttest1.cfm

The *t* statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

X_1 - mean sample 1 n_1 - # in Sample 1

$\bar{X}_2$ = mean sample 2 n_2 # in sample 2

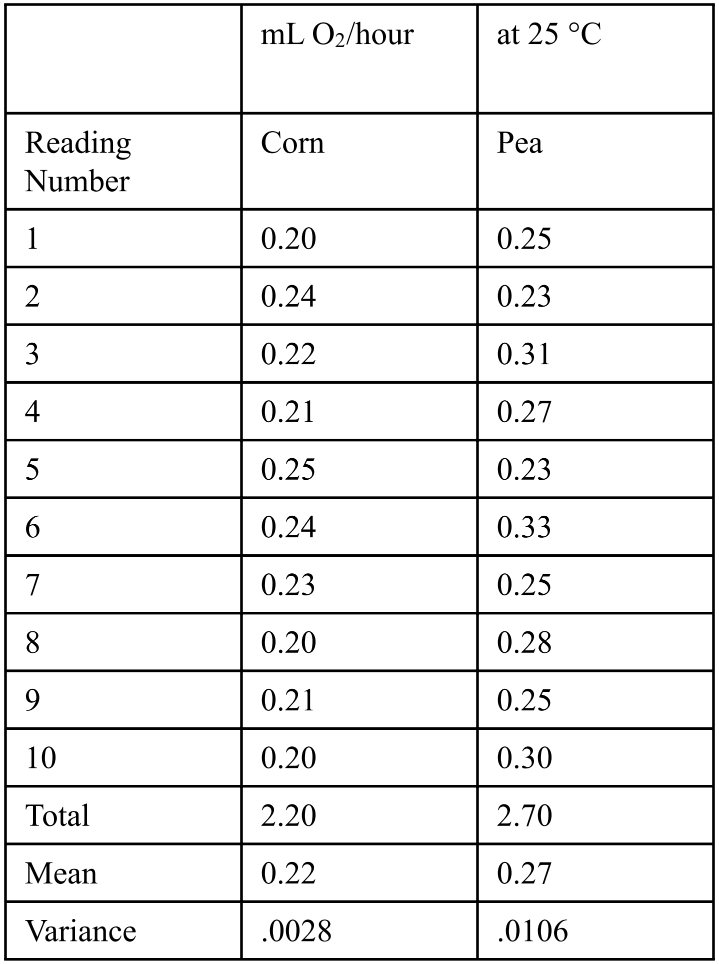
s_1^2 - variance of sample 1

s_2^2 - variance of sample 2

If samples are equal in size simplified formula ($n_1 = n_2 = n$)

$$t = \frac{\bar{X} - \bar{X}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

Amount of O_2 Used by Germinating Seeds of Corn and Pea Plants

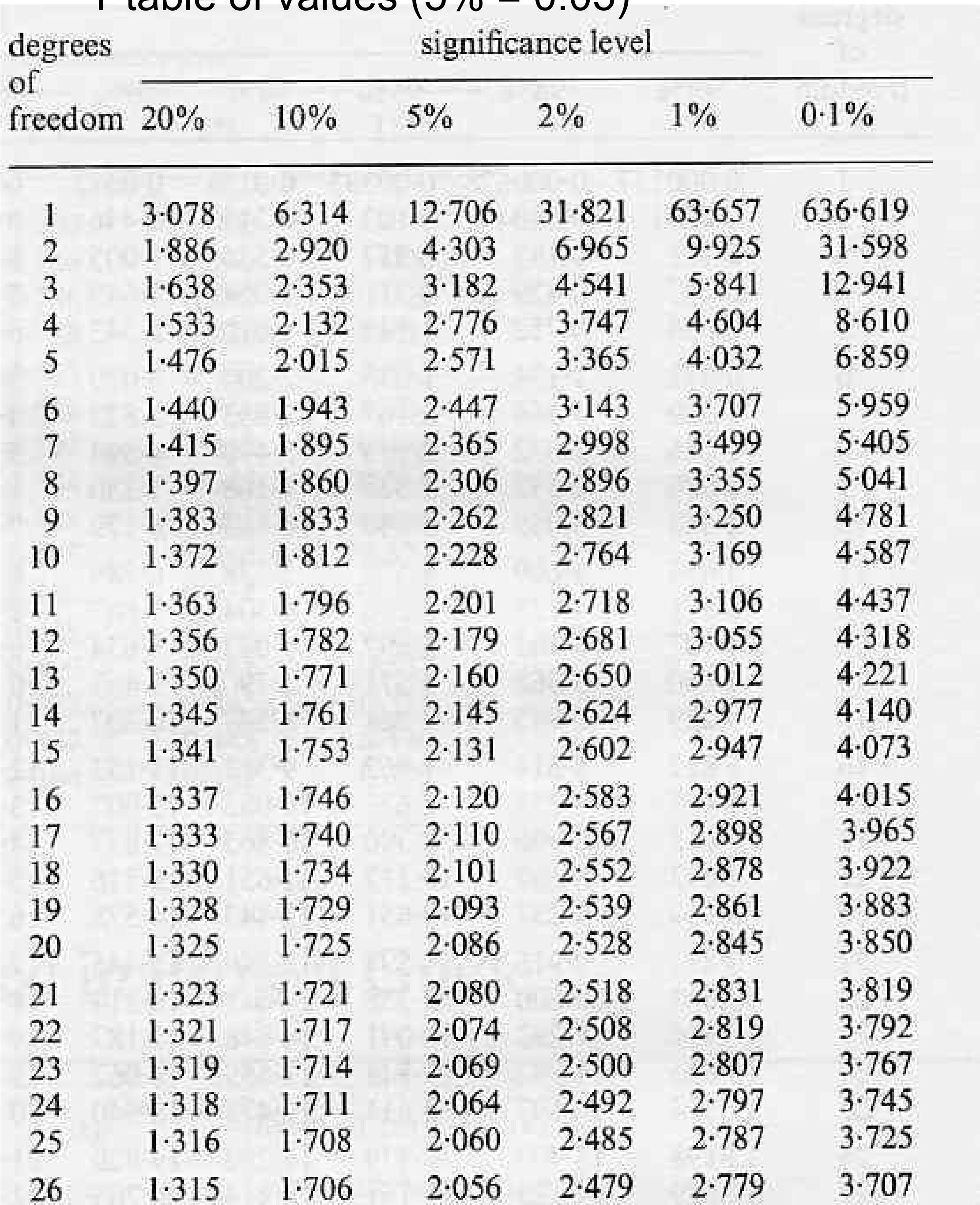


	mL O_2/hour	at 25 °C
Reading Number	Corn	Pea
1	0.20	0.25
2	0.24	0.23
3	0.22	0.31
4	0.21	0.27
5	0.25	0.23
6	0.24	0.33
7	0.23	0.25
8	0.20	0.28
9	0.21	0.25
10	0.20	0.30
Total	2.20	2.70
Mean	0.22	0.27
Variance	.0028	.0106

H_o = null hypothesis if the t value is larger than the chart value (the yellow regions) then reject the null hypothesis and accept the H_A that there is a difference between the means of the two groups… there is a significant difference between the treatment group and the control group.

http://www2.cedarcrest.edu/academic/bio/hale/biostat/session19links/nachocurve2tail.jpg

T table of values (5% = 0.05)



degrees of freedom	significance level 20%	10%	5%	2%	1%	0·1%
1	3·078	6·314	12·706	31·821	63·657	636·619
2	1·886	2·920	4·303	6·965	9·925	31·598
3	1·638	2·353	3·182	4·541	5·841	12·941
4	1·533	2·132	2·776	3·747	4·604	8·610
5	1·476	2·015	2·571	3·365	4·032	6·859
6	1·440	1·943	2·447	3·143	3·707	5·959
7	1·415	1·895	2·365	2·998	3·499	5·405
8	1·397	1·860	2·306	2·896	3·355	5·041
9	1·383	1·833	2·262	2·821	3·250	4·781
10	1·372	1·812	2·228	2·764	3·169	4·587
11	1·363	1·796	2·201	2·718	3·106	4·437
12	1·356	1·782	2·179	2·681	3·055	4·318
13	1·350	1·771	2·160	2·650	3·012	4·221
14	1·345	1·761	2·145	2·624	2·977	4·140
15	1·341	1·753	2·131	2·602	2·947	4·073
16	1·337	1·746	2·120	2·583	2·921	4·015
17	1·333	1·740	2·110	2·567	2·898	3·965
18	1·330	1·734	2·101	2·552	2·878	3·922
19	1·328	1·729	2·093	2·539	2·861	3·883
20	1·325	1·725	2·086	2·528	2·845	3·850
21	1·323	1·721	2·080	2·518	2·831	3·819
22	1·321	1·717	2·074	2·508	2·819	3·792
23	1·319	1·714	2·069	2·500	2·807	3·767
24	1·318	1·711	2·064	2·492	2·797	3·745
25	1·316	1·708	2·060	2·485	2·787	3·725
26	1·315	1·706	2·056	2·479	2·779	3·707

For example:
For 10 degrees of freedom (2N-2)
The chart value to compare your t value to is 2.228

If your calculated t value is between
+2.228 and -2.228
Then accept the null hypothesis the mean are similar

If your t value falls outside
+2.228 and -2.228
(larger than 2.228 or smaller than -2.228)
Fail to reject the null hypothesis (accept the alternative hypothesis) there is a significant difference.

So if the mean of the corn = 0.22 and the mean of the peas =0.27
The variance (s^2)of the corn is 0.000311 and the peas is .001178.
Each sample population is equal to ten.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

Then:
0.22-0.27 / √ (.000311+.001178)/10

-0.05/ √ 0.001489/10
-0.05/ √ .0001489
(ignore negative sign)
t= 4.10
Df = 2N-2 = 2(10) -2=18
Chart value =2.102
Value is higher than t-value… reject the null hypothesis there is a difference in the means.

The “z” test
-used if your population samples are greater than 30
-Also used for normally distributed populations with continuous variables

-formula: note: “σ” (sigma) is used instead of the letter “s”

z= mean of pop #1 – mean of pop #2/
√ of variance of pop #1/n1 + variance of pop#2/n2

Also note that if you only had the standard deviation you can square that value and substitute for variance

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

Z table (sample table with 3 probabilities

α	Zα (one tail)	Zα/2 (two tails)
0.1	1.28	1.64
0.05	1.645	1.96
0.01	2.33	2.576

Use a one-tail test to show that sample mean A is significantly greater than (or less than) sample mean B. Use a two-tail test to show a significant difference (either greater than Or less than) between sample mean A and sample mean B.

Z table use:

α = alpha (the probability of) 10%, 5% and 1 %

Z α: z alpha refers to the normal distribution curve is on one side only of the curve "one tail" can be left of the mean or right of the mean. Also your null hypothesis is either expected to be greater or less than your experimental or alternative hypothesis

Z α/2 = z alpha 2: refers to an experiment where your null hypothesis predicts no difference between the means of the control or the experimental hypothesis (no difference expected). Your alternative hypothesis is looking for a significant difference

Example z-test

- You are looking at two methods of learning geometry proofs, one teacher uses method 1, the other teacher uses method 2, they use a test to compare success.
- Teacher 1; has 75 students; mean =85; stdev=3
- Teacher 2: has 60 students; mean =83; stdev=

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

= (85-83)/√3^2/75 + 2^2/60

= 2/0.4321 = 4.629

Example continued

Z= 4.6291

Ho = null hypothesis would be Method 1 is not better than method 2

HA = alternative hypothesis would be that Method 1 is better than method 2

This is a one tailed z test (since the null hypothesis doesn't predict that there will be no difference)

So for the probability of 0.05 (5% significance or 95% confidence) that Method one is not better than method 2 … that chart value = Zα 1.645

So 4.629 is greater than the 1.645 (the null hypothesis states that method 1 would not be better and the value had to be less than 1.645; it is not less therefore reject the null hypothesis and indeed method 1 is better

Z table (sample table with 3 probabilities)

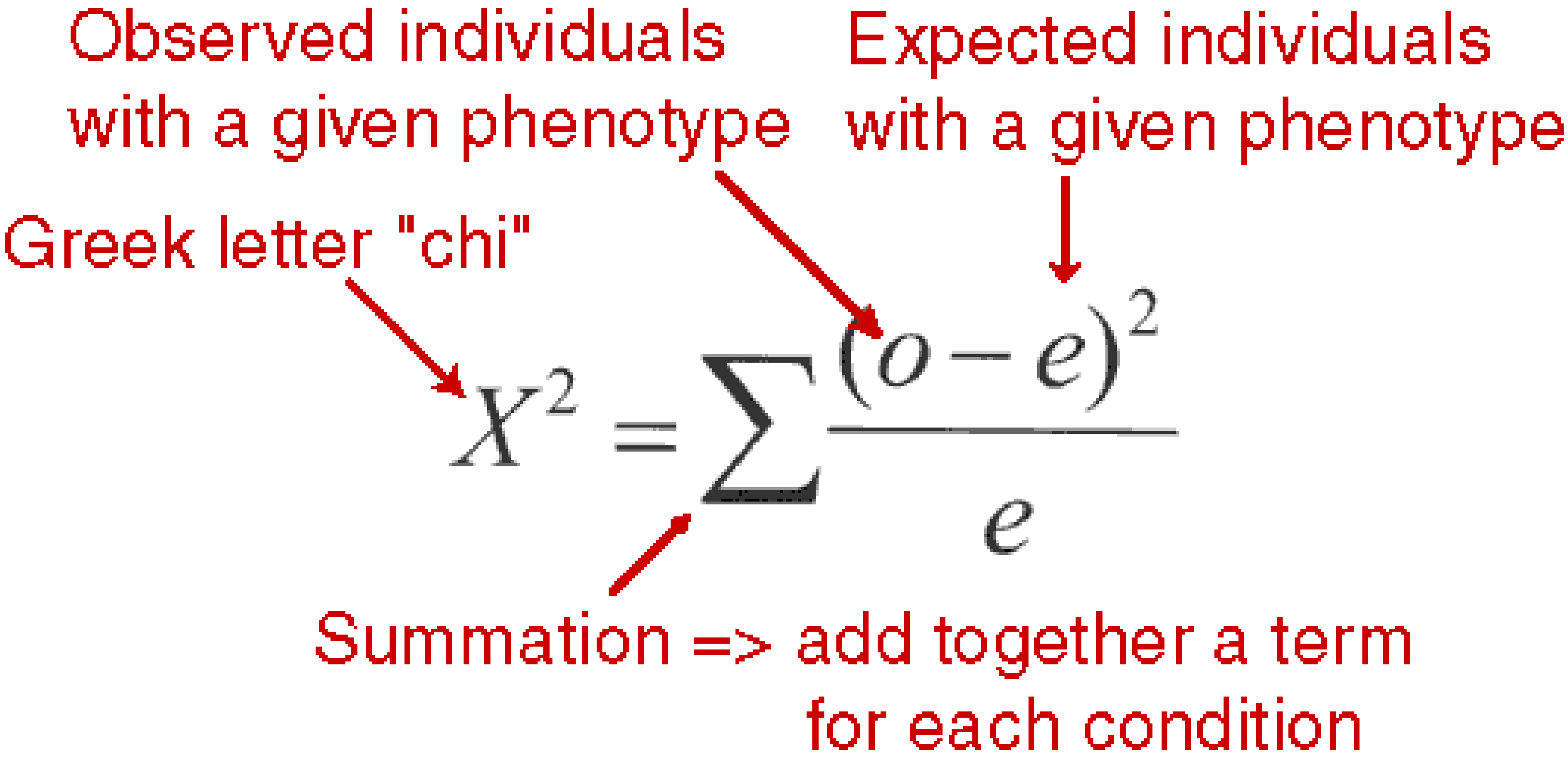
α	Zα (one tail)	Zα/2 (two tails)
0.1	1.28	1.64
0.05	1.645	1.96
0.01	2.33	2.576

Chi square

- Used with discrete values
- Phenotypes, choice chambers, etc.
- Not used with continuous variables (like height… use t-test for samples less than 30 and z-test for samples greater than 30)
- O= observed values
- E= expected values

$$X^2 = \sum \frac{(o-e)^2}{e}$$

http://www.jspearson.com/Science/chiSquare.html http://course1.winona.edu/sberg/Equation/chi-squ2.gif



Observed individuals with a given phenotype

Expected individuals with a given phenotype

Greek letter "chi"

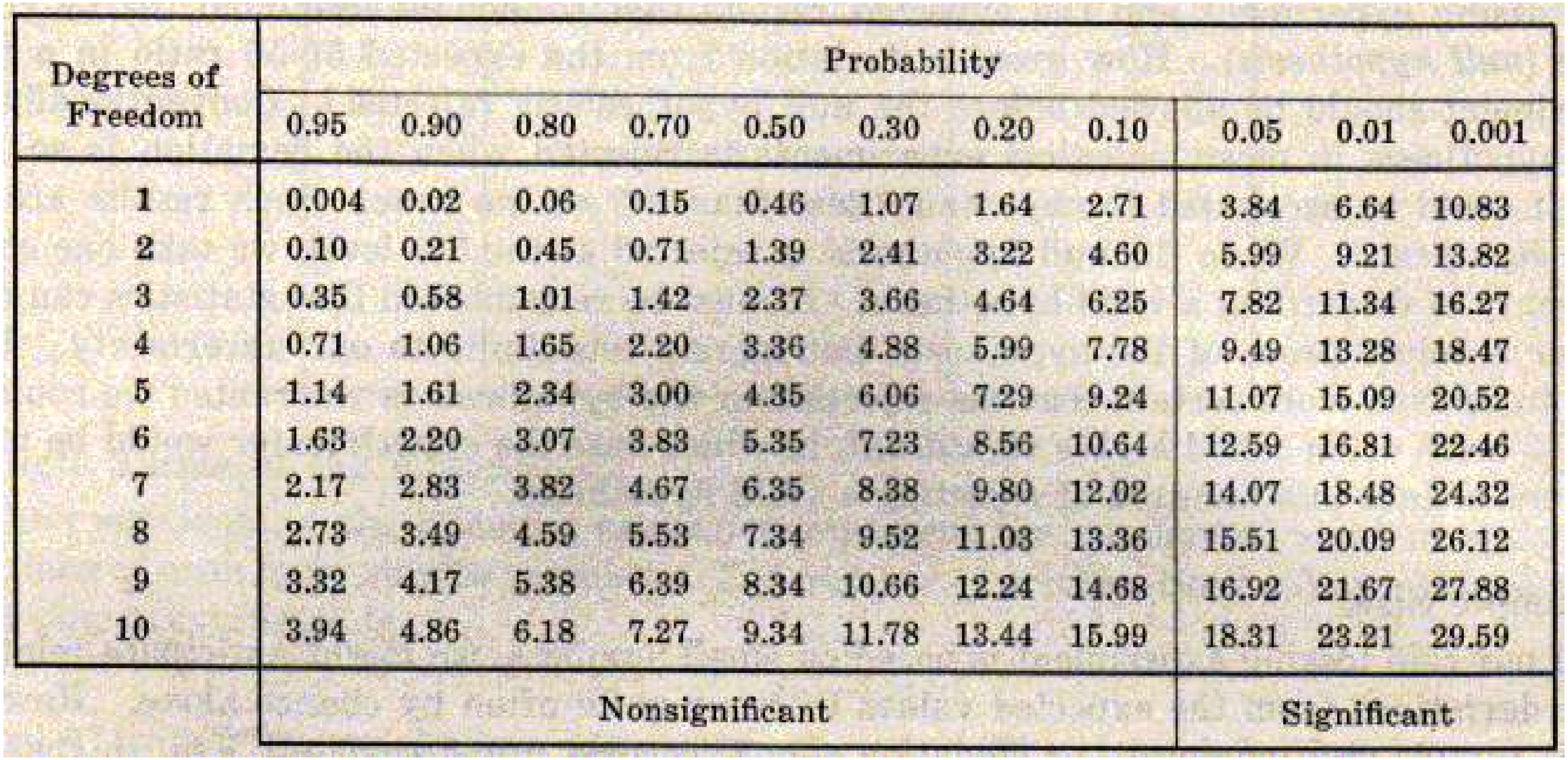
$$X^2 = \sum \frac{(o-e)^2}{e}$$

Summation => add together a term for each condition

Interpreting a chi square

- Calculate degrees of freedom
- # of events, trials, phenotypes -1
- Example 2 phenotypes-1 =1
- Generally use the column labeled 0.05 (which means there is a 95% chance that any difference between what you expected and what you observed is within accepted random chance.
- Any value calculated that is larger means you reject your null hypothesis and there is a difference between observed and expect values.

How to use a chi square chart



Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

http://faculty.southwest.tn.edu/jiwilliams/probab2.gif

Chi-Square

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Example:

$$\chi^2 = \frac{(615-600)^2}{600} , \frac{(385-400)^2}{400}$$

$$= \frac{15^2}{600} + \frac{15^2}{400} = \frac{225}{600} + \frac{225}{400} =$$

$$= 0\ 375 + 0.562 = 0.937$$

degrees of freedom = N-1 = 2 samples - 1
= 1