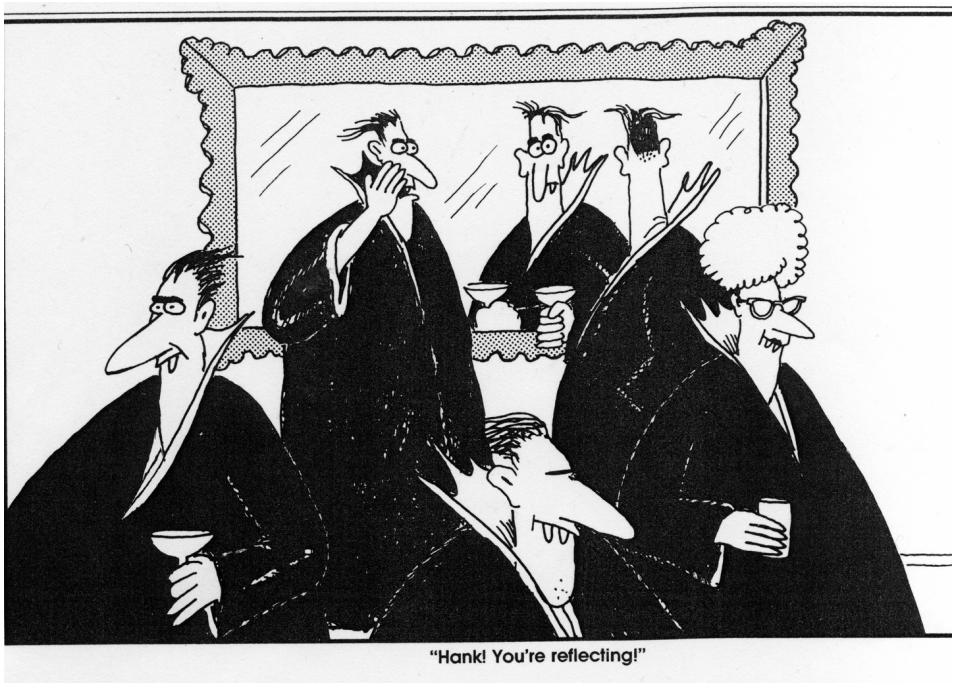
Introduction to Chi Square Analysis



"Hank! You're reflecting!"

***Wild-type* vampires do not cast a reflection in a mirror. *Mutant* vampires are known that do cast a reflection in a mirror. [They are also insensitive to sunlight and do not have to return to their coffins at dawn.]**

A vampire from a true-breeding wildtype Transylvanian line mates with a mutant (reflecting) American vampress.
They produce:
8 mutant (reflecting progeny)
2 wild-type (non-reflecting) progeny

How would you interpret these data?

Are you concerned that the progeny exhibit a 8:2 ratio rather than 5:5 ratio of dominant:recessive?

Why do we expect more deviation from expected in small sample sizes than we do in large sample sizes?

Small sample sizes – chance deviations in one direction are not necessarily cancelled out by chance deviations in the other direction

See Law of Large Numbers:

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

http://www.khanacademy.org/math/probability/random-variables-topic/random_variables_prob_dist/v/law-of-large-numbers

According to the "law", the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed (see last page of lecture)

What if the vampire family had 20 kids and the ratio was the same: 16:4 ?

Assuming that your hypothesis about the inheritance of this trait is correct, how would you calculate the probability of the 8:2 outcome or the 16:4 outcome or possibly an outcome of 85:75 or 80:60 or whatever ?

How much deviation from expected is too much deviation? How much deviation makes you concerned that your genetic interpretation is incorrect?

The chi square test will help us address these qustions

http://udel.edu/~mcdonald/statchigof.html

You have a set of expected observations that are predicted from hypothesis (pr interpretation): *in the vampire case -- dominant inheritance of a mutant phenotype controlled by a single gene; the parental mating was between a homozygous wildtype and a heterozygous mutant*

Your genetic hypothesis predicts the number (fraction) individuals in each category: *based on the genetic hypothesis (above) and Mendel's principle of segregation and random fertilization, we expect an equal number of progeny with the dominant and recessive phentoypes*

The observed numbers don't match (exactly) the predicted numbers

Is this just a chance deviation?

FIRST consider a simple experiment:

10 fertilization events for Aa X aa *OR* 10 flips of a coin

Expect 5 Aa progeny and 5 aa progeny *OR* Expect 5 heads and 5 tails or

Observed: 8 head and 2 tails Is there something wrong with the coin?

heads	tails	*numerator/1024	%
10	0	1	0.1
9	1	10	1
8	2	45	4.4
7	3	120	11.7
6	4	210	20.5
5	5	252	24.6
4	6	210	20.5
3	7	120	11.7
2	8	45	4.4
1	9	10	1
0	10	1	0.1

How often do we see deviations of 8:2 (or greater) in either direction?

From this table (using basic rules of probability and the binomial): the answer is 0.11 or 11%

http://www.khanacademy.org/math/probability/random-variables-topic/binomial_distribution/v/binomial-distribution-1use

BUT What if we flipped 100 coins and got 80 heads and 20 tails?

We could do this calculation as above, but it would be very time consuming.

The chi square test though will quickly tell us that the probability (p value) of this deviation from expected with a sample size of 100 is <0.001 or <0.1%

The primary goal of a statistical test is to determine whether an observed data set is so different from what you would expect under the null hypothesis that you should reject the null hypothesis.

Most statistical tests take the following form:

http://udel.edu/~mcdonald/statexactbin.html

1. *Collect the data.*
2. *Calculate a number, the test statistic, that measures how far the observed data deviate from the expectation under the null * hypothesis.*
3. *Use a mathematical function to estimate the probability of getting a test statistic as extreme as the one you observed, if the null hypothesis were true. This is the P-value.*

Null hypothesis: working hypothesis

The P value addresses this question:

- If the theory that generated the expected values were correct, what is the probability of observing such a discrepancy (or larger one) between observed and expected values?

- In other words, the chi square test tells us the probability that the data we've collected deviate from the expected just due to chance

1. Before proceeding with the chi square calculation, clearly state the genetic hypothesis concerning the data

This hypothesis is an interpretation of the data that gives a precise prediction about what the expected outcome of your experiment should be (have been) assuming that your hypothesis/interpretation is correct.
Examples:

- Single gene trait, two alleles with simple dominance
- Mendel's principle of segregation
- Mendel's principle of independent assortment
- Complications to dominance, lethal alleles, etc.

2. Use the rules of probability (product, sum and conditional) to make explicit predictions of the types and proportions of progeny that should be observed if your hypothesis is true

In other words, your hypothesis should give a straight-forward prediction with respect to progeny classes (genotype or phenotype) and ratios

3. ***For each class of progeny in turn, subtract the expected number from the observed number. Square this difference and divide it by the expected number.***

Note that you are to use the actual numbers of progeny, not the proportions, ratios, fractions or percentages.

4. Sum the results of the calculation described in step 3 for all classes of progeny

Chi square value =

$$\chi^2 = \Sigma \frac{[O - E]^2}{E}$$

O = observed value in a given category of progeny
E = expected value in that category (predicted by your

Σ = sum of value in each progeny category

5. Use the chi square table to determine p, which helps you assess whether the data (the χ^2 value) represent a good fit or a bad fit to the expected numbers

Table: Chi-Square Probabilities

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515

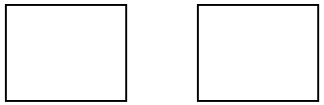
P = *probability that an equal or worse fit would occur by chance, assuming that your hypothesis is true*

degrees of freedom = number of progeny classes – 1

degrees of freedom = number of progeny classes - 1

if you have 400 progeny

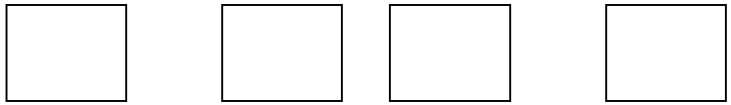
progeny class 1 | progeny class 2



if class 1 is assigned 300 progeny, then class 2 must have 100 (so that the total is 400) -- this is one degree of freedom

similar logic for multiple classes

progeny class 1 | progeny class 2 | progeny class 3 | progeny class 4



Count 400 progeny. Once numbers are assigned to class1, class 2 and class 3, then the number in class 4 must be 400- (# in classes 1+2+3)

4-1 or three degrees of freedom

Tips for χ^2 chi squaring

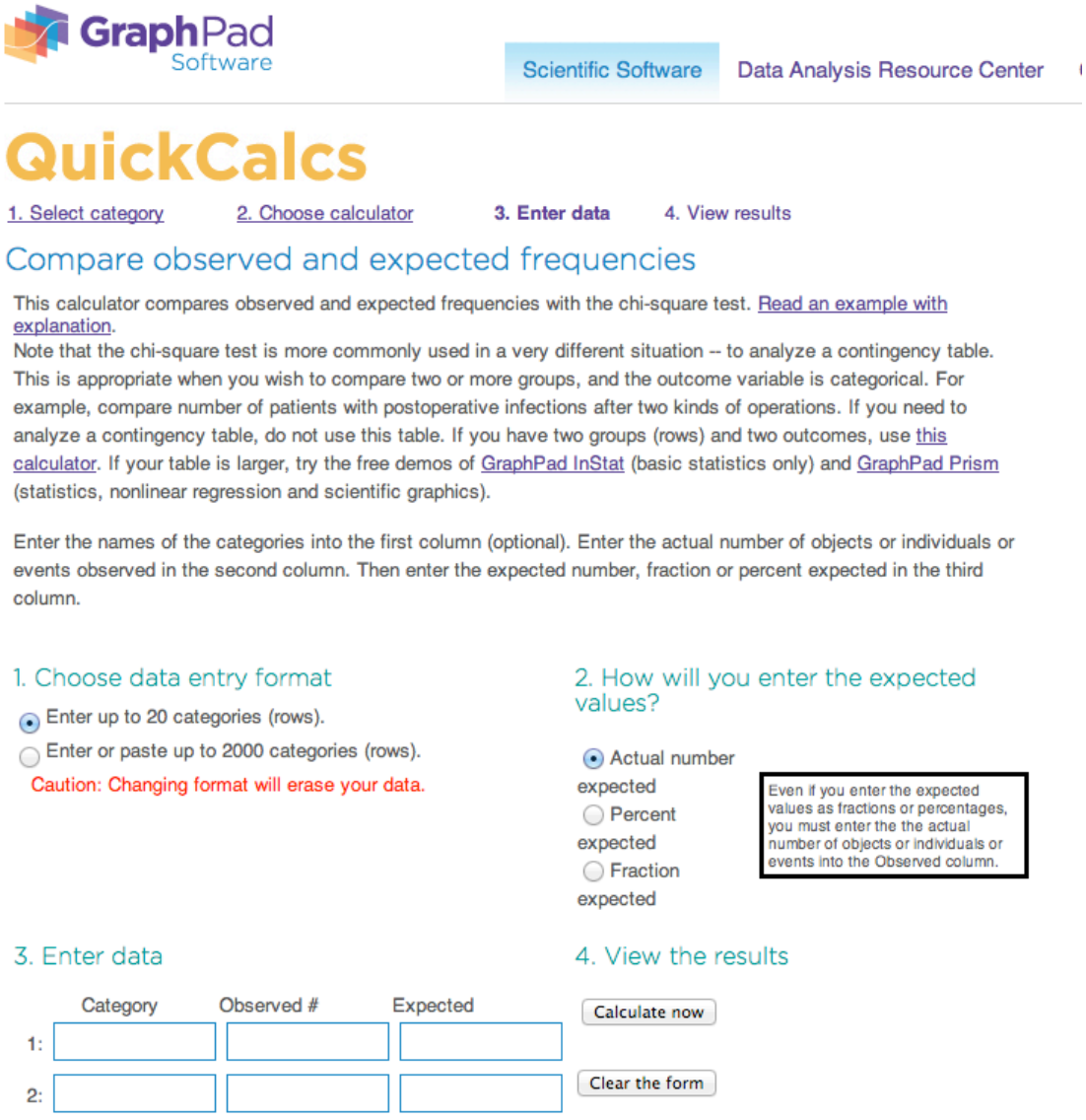
Significant decimal places:

- Express numbers in expected categories to **one decimal** place.
- Express the final chi square value to **3 decimal places**, because that is the accuracy of the table of critical values
- To avoid rounding errors, *all intermediate computations*, including the expected values should be carried out to **4 decimal places**

ALWAYS USE ACTUAL NUMBERS: NEVER USE FRACTIONS OR PERCENTAGES OR DECIMAL FRACTIONS

1. A comparison of ratios or percentages alone will never allow you to determine whether or not the *observed* data are significantly different from the *predicted* values.
2. The absolute numbers are important because they reflect the size of the experiment. The larger the sample size the closer the observed ratios or percentages can be expected to match the values predicted by the experimental hypothesis, *if the hypothesis is correct*.

http://graphpad.com/quickcalcs/chisquared1.cfm



Roller Example

A friend of your is puzzling over some data he has collected. He did a mutant screen and picked up a single worm with a dominant roller phenotype. He selfed this worm and scored the progeny, counting 28 rollers and 14 wild-type. Because of the 2:1 ratio of rollers to wild-type, he concludes that the roller allele is a lethal in the homozygous state.

You look at his data and suggest that his conclusion is premature. You suggest a simple, alternative interpretation and do a statistical analysis to test your hypothesis.

1. State your alternative interpretation
2. Show your chi-square calculation
3. State explicitly in words what your calculation tells you
4. Does your calculation *prove* your hypothesis or your friend's hypothesis? Why or why not
5. Suggest follow-up experiments

STATE A SIMPLE hypothesis (interpretation) that gives a precise prediction:

Hypothesis: *The RR genotype is not lethal. So if a Rr heterozygote is selfed, the dominant and and recessive phenotypes should be observed in the progeny in a 3:1 ratio*

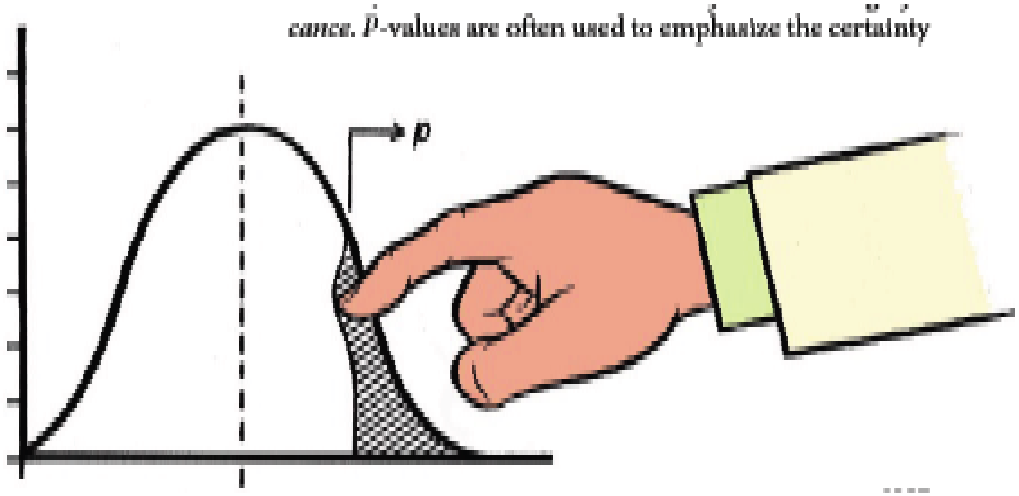
For the roller problem, the chi square value is 1.556

Does this value represent a good fit or a bad fit to the data?
df = 2-1= 1
probability from table:
$0.975 > p > 0.2$

P= probability that a particular set of observed experimental results represents a chance deviation from the values predicted by the hypothesis

$$\chi^2 = \Sigma \frac{|O - E|^2}{E}$$

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515



How do I determine if a particular p-value is significant?

If p is large, the ***observed deviation*** *from the expected results is considered* ***insignificant.***

If the probability is very low (<0.05) the ***observed deviation*** *from the expected results becomes* ***significant****.*

Which of the two interpretations is correct?

Interpretation 1: *Roller mutation is lethal in the homozygous state*

Interpretation 2: *Roller mutation is viable in the homozygous state*

The chi square test assists the investigator in accepting or rejecting a hypothesis by calculating the probability that the data are compatible with the hypothesis

<u>It can not be emphasized too strongly that any test of goodness-of-fit can only assist an investigator in making up his/her mind.</u>

It neither proves or disproves a hypothesis

The Chi square calculation does not tell you which interpretation about the roller mutant is correct: it tells you that likelihood that you would get this set of data just by chance assuming RR is viable

HOW would you proceed in determining whether the roller allele was homozygous lethal?

What additional experiments could you do?

Consider this: a similar situation but p value is <0.05.

Can we conclude that the RR genotype is lethal?

Or is it possible that RR is not lethal and something else is going on – that is, the deviation from the expected 3:1 ratio is due to, for example,

- *a flaw in the design of the experiment*
- *or a failure to execute the protocol correctly*
- *or some other issue*

Interpreting the results of a chi square analysis

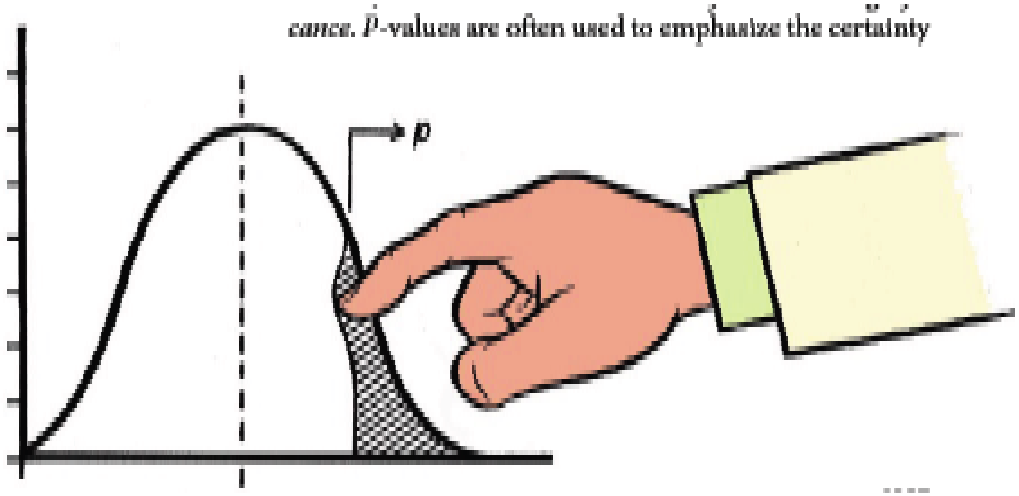
p value > 0.05

- ***Your hypothesis may be correct*** *and any differences between O and E due to chance.*
- *On the other hand, a p value > 0.05 Does NOT prove your hypothesis as competing hypotheses may also have a p value that is > 0.05……*

p value < 0.05

- ***Your hypothesis may be incorrect.*** *The difference betwee E & O is not due to chance but due to an incorrect hypothesis. If we decide to reject the hypothesis based on the chi square analysis, what do we do or ask next?*
- *On the other hand, a p value < 0.05 does NOT disprove your hypothesis.* ***Your hypothesis may be correct*** *and something else is going that results in a difference between O and E that is not due to just to chance. We're not going to throw out our hypothesis just yet but…*

What should you do next?
Where would you go from here to resolve the problems?
What could that something else be?



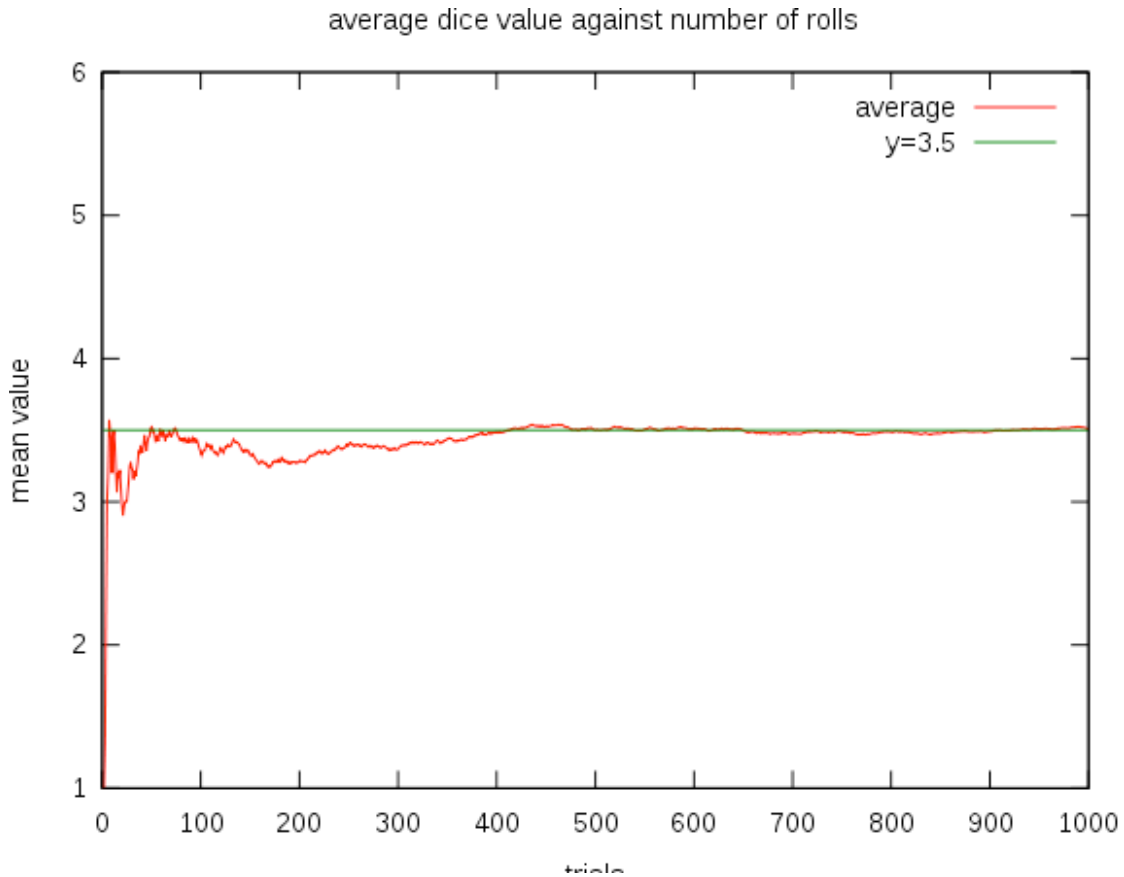
How do I determine if a particular p-value is significant?

If p is large, the ***observed deviation*** *from the expected results is considered* ***insignificant.***

If the probability is very low (<0.05) the ***observed deviation*** *from the expected results becomes* ***significant****.*

What does significant mean?

- *In statistics, a result is called statistically significant if it is unlikely to have occurred by chance*
- *The amount of evidence required to accept that an event is unlikely to have arisen by chance is known as the significance level or critical p-value: in traditional frequentist statistical hypothesis testing, the p-value is the frequency or probability with which the observed event would occur, if the null ** hypothesis were true.*
- *In statistics, a null hypothesis is a hypothesis set up to be nullified or refuted in order to support an alternative hypothesis.*
- *When used, the null hypothesis is presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.*
- *If the obtained p-value is smaller than the significance level, then the null hypothesis is rejected -- well, MAYBE*



An illustration of the law of large numbers using a particular run of rolls of a single die. As the number of rolls in this run increases, the average of the values of all the results approaches 3.5. While different runs would show a different shape over a small number of throws (at the left), over a large number of rolls (to the right) they would be extremely similar.

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