BIO 682
Nonparametric Statistics
Spring 2010

Steve Shuster

http://www4.nau.edu/shustercourses/BIO682/index.htm

Lecture 4

Binomial Test

1. Also, we can test specific hypotheses
   a. Whether the observed distribution occurred *by chance*
   a. H_o: $P_i = P_o$.
2. Or if it deviates from this distribution (i.e., *could not* have occurred by chance).
   b. H_1: $P_i \neq P_o$.

Example:

a. Suppose you are observing a lek of male sage grouse ($N_{males} = 6$)

1. 5 females will enter the lek and mate.
2. You want to figure out the probability that a male will *mate more than 2 times*.

Method:

1. Let k = the sum of counts of one class (# mated males) = 2.
2. Let N = total number of opportunities a male has to mate (total # of mated males) = 5.
3. Let p = proportion of observations in which x=1.
4. Let $q = (1 - p)$ = proportion of observations in which x=0.

Method:

5. For each mating, each males probability of mating is:

$1/N_{males} = 1/6 = p$

of not mating = $(1 - p) = 5/6 = q$.

6. The number of objects in k and in N-k is given by the equation for a binomial distribution:

The Binomial Equation

$$P[k] = \binom{N}{k} p^k q^{N-k}$$

Where,

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

The Exact Probability

1. That one of the males will mate *twice* is:

$$P_{[x=2]} = \binom{N}{k} p^k q^{N-k}$$

$$P_{[x=2]} = \frac{5!}{2!3!} (1/6)^2(5/6)^3 = .16$$

However,

2. We are really interested in finding out the degree to which this could occur *by chance*.

a. Thus, we need to find out the probability of obtaining values *as extreme or more extreme* as the observed value.

b. Or, what is the probability that a male will mate *two and fewer times*.

b. Thus, $P_{[k < 2]} = P_{[k=0]} + P_{[k=1]} + P_{[k=2]}$

So,

$$P_{[k=0]} = \frac{5!}{0!5!} (1/6)^0(5/6)^5 = .40$$

$$P_{[k=1]} = \frac{5!}{1!4!} (1/6)^1(5/6)^4 = .40$$

$$P_{[k=2]} = \frac{5!}{2!3!} (1/6)^2(5/6)^3 = .16$$

And,

P[k ≤ 2] = .96.
Thus, the chances that one male will mate *two or fewer times* = .96.

Thus, probability of mating *more than 2 times* is:

$1 - P_{[k \leq 2]} = .04$

This is unlikely to occur by chance alone with p = 0.05.

Binomial Test, Continued

Small samples (N < 35)

1. S&C present a table in the back of the book that calculates the probabilities for various values of *N* and *k* if Ho: *p* = 1/2.

TABLE D. TABLE OF PROBABILITIES ASSOCIATED WITH VALUES AS SMALL AS OBSERVED VALUES OF *x* IN THE BINOMIAL TEST*

Given in the body of this table are one-tailed probabilities under H_0 for the binomial test when $P = Q = \frac{1}{2}$. To save space, decimal points are omitted in the *p*'s.

N \ *x*	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	031	188	500	812	969	†										
6	016	109	344	656	891	984	†									
7	008	062	227	500	773	938	992	†								
8	004	035	145	363	637	855	965	996	†							
9	002	020	090	254	500	746	910	980	998	†						
10	001	011	055	172	377	623	828	945	989	999	†					
11		006	033	113	274	500	726	887	967	994	†	†				
12		003	019	073	194	387	613	806	927	981	997	†	†			
13		002	011	046	133	291	500	709	867	954	989	998	†	†		
14		001	006	029	090	212	395	605	788	910	971	994	999	†	†	
15			004	018	059	151	304	500	696	849	941	982	996	†	†	†
16			002	011	038	105	227	402	598	773	895	962	989	998	†	†
17			001	006	025	072	166	315	500	685	834	928	975	994	999	†
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

* Adapted from Table IV, B, of Walker, Helen, and Lev, J. 1953. *Statistical inference.* New York: Holt, p. 458, with the kind permission of the authors and publisher.

† 1.0 or approximately 1.0.

Binomial Test: Example

1. For N = 10 and k = 3, one-tailed probability is .172.

TABLE D. TABLE OF PROBABILITIES ASSOCIATED WITH VALUES AS SMALL AS OBSERVED VALUES OF x IN THE BINOMIAL TEST*

Given in the body of this table are one-tailed probabilities under H_0 for the binomial test when $P = Q = \frac{1}{2}$. To save space, decimal points are omitted in the p's.

N \ x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	031	188	500	812	969	†										
6	016	109	344	656	891	984	†									
7	008	062	227	500	773	938	992	†								
8	004	035	145	363	637	855	965	996	†							
9	002	020	090	254	500	746	910	980	998	†						
10	001	011	055	172	377	623	828	945	989	999	†					
11		006	033	113	274	500	726	887	967	994	†	†				
12		003	019	073	194	387	613	806	927	981	997	†	†			
13		002	011	046	133	291	500	709	867	954	989	998	†	†		
14		001	006	029	090	212	395	605	788	910	971	994	999	†	†	
15			004	018	059	151	304	500	696	849	941	982	996	†	†	†
16			002	011	038	105	227	402	598	773	895	962	989	998	†	†
17			001	006	025	072	166	315	500	685	834	928	975	994	999	†
18			001	004	015	048	119	240	407	593	760	881	952	985	996	999
19				002	010	032	084	180	324	500	676	820	916	968	990	998
20				001	006	021	058	132	252	412	588	748	868	942	979	994
21				001	004	013	039	095	192	332	500	668	808	905	961	987
22					002	008	026	067	143	262	416	584	738	857	933	974
23					001	005	017	047	105	202	339	500	661	798	895	953
24					001	003	011	032	076	154	271	419	581	729	846	924
25						002	007	022	054	115	212	345	500	655	788	885

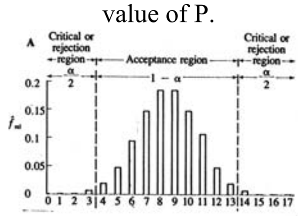
* Adapted from Table IV, B, of Walker, Helen, and Lev, J. 1953. *Statistical inference*. New York: Holt, p. 458, with the kind permission of the authors and publisher.

† 1.0 or approximately 1.0.

Binomial Test: Example

a. Note that one-tailed test is used because you are predicting *in advance* which of the values will be smaller.

b. If a two tailed test is used, you *double* the value of P.



Large samples ($N > 35$)

1. As N increases, the binomial distribution approaches a normal distribution.

a. It is then possible to use values of p, q and N to estimate z,

A parameter that estimates the probability of occurrence of observed value, x,

based on a binomial distribution.

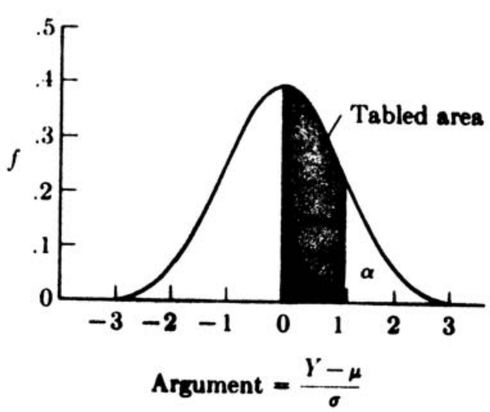
The Equation Is:

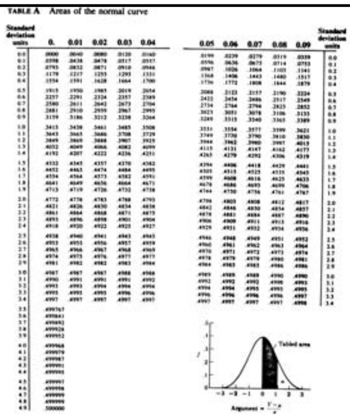
$$z = \frac{(x - \mu_x)}{\sigma_x}$$

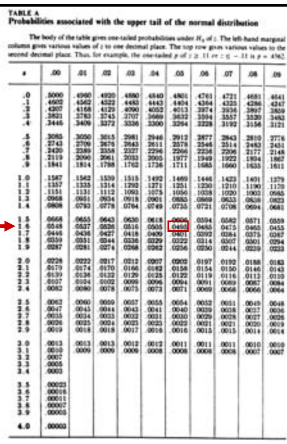
Which is equivalent to,

$$= \frac{x - Np}{(Npq)^{1/2}}$$

Look up the value of z on table of normal distribution (Appendix A in S&C).







So, the probability that a calculated value of $z = 1.65$ could have occurred by chance is, 0.0495.

Rule of Thumb:

1. If $Npq > 9$ to use this technique.
2. For example:

N	*p*	*q*	
25	0.5	0.5	6.25
50	0.5	0.5	12.5
100	0.5	0.5	25

Goodness of Fit Tests

1. Most people are familiar with goodness of fit tests (chi-square)

a. Considers situations when researcher wants to see whether observed distribution of counts fits predicted frequency for various categories.

b. X^2 equals the sum of all (squared deviations of observed values from expected)/ expected.

Chi Squared Test

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

1. Where, for *k* categories, O_i is the observed value of the i-th class, and E_i is the expected value of the i-th class.

Chi Squared Test

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

2. It is easy to see that small deviations from expected -> small X^2.

a. Significance is tested using a chi-squared distribution with df = *k*-1.

Chi Squared Test

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

3. Expected values are determined by the hypothesis.

a. Most often simply equal expected frequencies across groups (N/k).

b. This hypothesis is *intrinsic* to the data.

Expected Frequencies

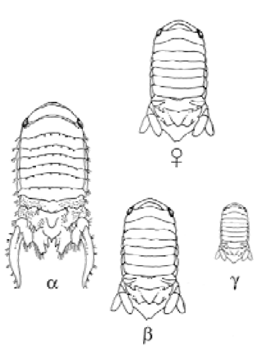
2. Also can be determined by some *extrinsic* hypothesis.

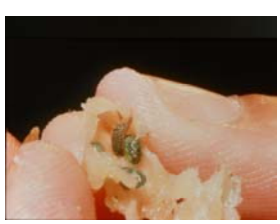
a. Mendelian inheritance:

1. 75:25, as expected in monohybrid cross with dominance.
2. 9:3:3:1 as expected with dihybrid cross.

Example

1. The frequency of male morphs in *P. sculpta*





Example

Male type:	α	β	γ	N
Expected:	33.3	33.3	33.3	100
Observed:	82	4	14	100

$X^2 = [(82-33.3)^2]/33.3 + [(4-33.3)^2]/33.3 + [(14-33.3)^2]/33.3 = 108.1$

df = 3-1 = 2;

$X^2_{[0.05, 2]} = 5.99$, P >> 0.001

Important assumptions:

1. If $k = 2$, smallest expected value should be > 5.
2. When df > 1 (i.e., $k > 2$), can't use the test if
   a. > 20% of expected frequencies are < 5.
   b. Any expected frequency is < 1.

This Is Important Because:

1. Distribution of values for X^2 test approximates the actual X^2 distribution only as expected frequencies become *large*.
2. Small cell values can be overcome by pooling cells.
3. More on pooling later.

Chi Squared Test

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

1. Where, for *k* categories, O_i is the observed value of the i-th class, and E_i is the expected value of the i-th class.

Chi Squared Test

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

2. It is easy to see that small deviations from expected -> small X^2.
   a. Significance is tested using a chi-squared distribution with df = *k*-1.

Chi Squared Test

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

3. Expected values are determined by the hypothesis.
   a. Most often simply equal expected frequencies across groups (N/k).
   b. This hypothesis is *intrinsic* to the data.

Expected Frequencies

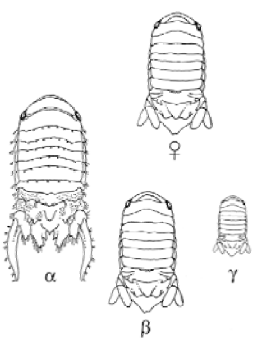
2. Also can be determined by some *extrinsic* hypothesis.

a. Mendelian inheritance:

1. 75:25, as expected in monohybrid cross with dominance.
2. 9:3:3:1 as expected with dihybrid cross.

Example

1. The frequency of male morphs in *P. sculpta*



Example

Male type:	α	β	γ	N
Expected:	33.3	33.3	33.3	100
Observed:	82	4	14	100

$X^2 = [(82-33.3)^2]/33.3 + [(4-33.3)^2]/33.3 + [(14-33.3)^2]/33.3 = 108.1$

df = 3-1 = 2;

$X^2_{[0.05, 2]} = 5.99$, P >> 0.001

Important assumptions:

1. If $k = 2$, smallest expected value should be > 5.

2. When df > 1 (i.e., $k > 2$), can't use the test if

   a. > 20% of expected frequencies are < 5.

   b. Any expected frequency is < 1.

This Is Important Because:

1. Distribution of values for X^2 test approximates the actual X^2 distribution only as expected frequencies become *large*.

2. Small cell values can be overcome by pooling cells.

3. More on pooling later.

Goodness of Fit Tests

1. Most people are familiar with goodness of fit tests (chi-square)

   a. Considers situations when researcher wants to see whether observed distribution of counts fits predicted frequency for various categories.

   b. X^2 equals the sum of all (squared deviations of observed values from expected)/ expected.

G-tests

1. Another goodness of fit test.

a. Similar to chi-squared, except that logarithms are used.

b. Makes it computationally ***simpler***, especially with complex designs.

2. Also, use of logarithms makes individual tests ***additive***

a. This permits ***partitioning of heterogeneity*** tests similar to what is possible with ANOVA.

Calculated As:

$$G = 2\sum^{a} f_i \ln\left(\frac{f_i}{\hat{f_i}}\right)$$

where: a = # of classes (k in other notation).

f_i = observed number of counts in the i-th class.

$f_{i\text{-hat}}$ = expected number of counts in the i-th class; = p_i(N)

with (a – 1) degrees of freedom.

Previous Example

Male type:	α	β	γ	N
Expected:	33.3	33.3	33.3	100
Observed:	82	4	14	100

G = 2{[82 ln (82/33.3)] + [4 ln (4/33.3)] + [14 ln (14/33.3)]}

= 2{73.9 - 8.5 -12.1}

= 106.5

$X^2_{[0.05,\,2]}$ = 5.99, P >> 0.001

Note:

Note that the value of *G* is less than the value of Chi-squared; for this test

this provides a more conservative test, because the sample size is relatively large.

Properties of G-tests

1. Tend to generate higher probability of Type I error than X^2.

a. i.e., *G* values are often higher than X^2

b. Mainly with small sample sizes,

2. This can be remedied using

Williams' Correction.

a. A method for making values of *G* more conservative.