PROPORTIONS

- Distribution Relationships
  - Binomial
  - Normal
  - Poisson
  - Chi-Square
- Importance of values of *n* and *p*
- Confidence Interval
  - Two Tails
  - One Tail
- Hypotheses
  - One proportion
  - Two proportions
  - Several proportions

Binomial Distribution utilized when

- two outcomes for each trial
- probability of success same for each trial
- there are *n* trials, where *n* is a constant
- the *n* trials are independent

Many experiments can be cast into the model of a Binomial Experiment:

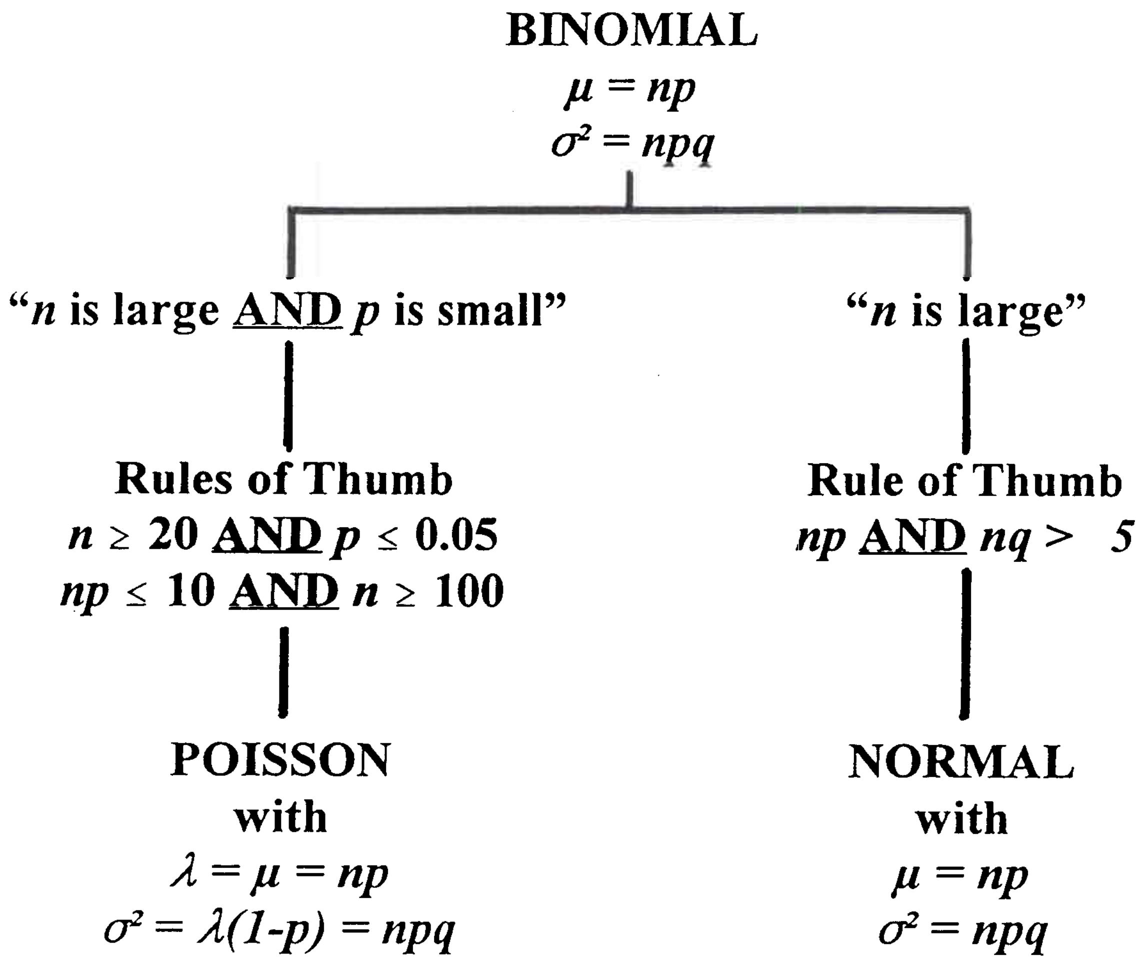
For example, we can measure the heights of a sample of people,

Define "success" as being 5'2" or less, and

Now we can count the number of "successes" (i.e., ≤ 5'2") and the number of "failures" (i.e., > 5'2")

Binomial Distribution is computationally cumbersome, and thus approximations are most frequently employed for the purpose of inference

DISTRIBUTION RELATIONSHIPS



BINOMIAL

$\mu = np$

$\sigma^2 = npq$

"n is large <u>AND</u> p is small"

Rules of Thumb

$n \geq 20$ **<u>AND</u>** $p \leq 0.05$

$np \leq 10$ **<u>AND</u>** $n \geq 100$

POISSON

with

$\lambda = \mu = np$

$\sigma^2 = \lambda(1-p) = npq$

"n is large"

Rule of Thumb

np **<u>AND</u>** $nq > 5$

NORMAL

with

$\mu = np$

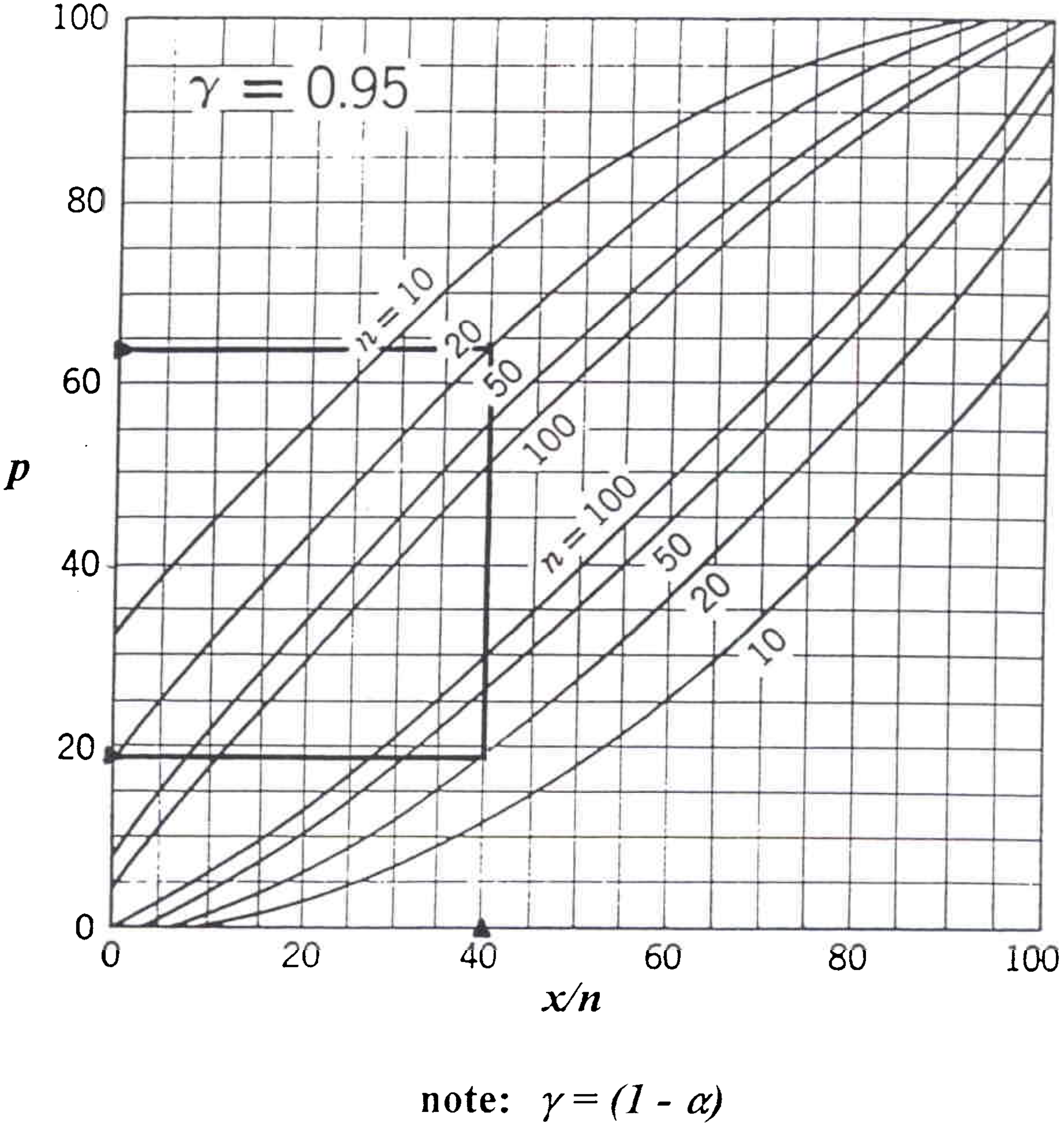
$\sigma^2 = npq$

note: $q = 1-p$

prop-4

CONFIDENCE INTERVAL
Binomial ---- Proportion

When $n = 20$ and $x = 8$, then $x/n = 40\%$ (or .40)
If $\gamma = 0.95$, then CONF $\{0.19 \leq p \leq 0.64\}$



note: $\gamma = (1 - \alpha)$

prop-5

CONFIDENCE INTERVAL
One-Tailed ---- Proportion
BINOMIAL / POISSON / CHI-SQUARE

One-tailed test used when *p* is small:

$$p < \frac{1}{2n} \chi^2_\alpha \text{ with } 2(x+1) \text{ dof}$$

Exact method involves the relationship,

$$1 - F(X; n, p) = Pr\left[F(m_1, m_2) < \left(\frac{n-x}{x+1}\right) \cdot \left(\frac{p}{1-p}\right)\right]$$

where $m_1 = 2(x+1)$ and $m_2 = 2(n-x)$

which is not amenable to simple solution.

EXAMPLE

Assuming three failures amongst 1000 items, determine a one-tailed Confidence Interval for the probability of failure of an item with α = 0.05.

$x = 3$; $n = 1000$; $\alpha = 0.05$

$dof = 2(x+1) = 2(3+1) = 8$

from the χ^2 table for $\alpha = 0.05$ with 8 dof, $\chi^2 = 15.507$

Thus $p < \frac{1}{2(1000)} \cdot 15.507 = 0.0078$

CONFIDENCE INTERVAL
Normal Approximation ---- Proportion

Let n = number of trials (i.e., sample size)

x = number of "successes" out of n trials

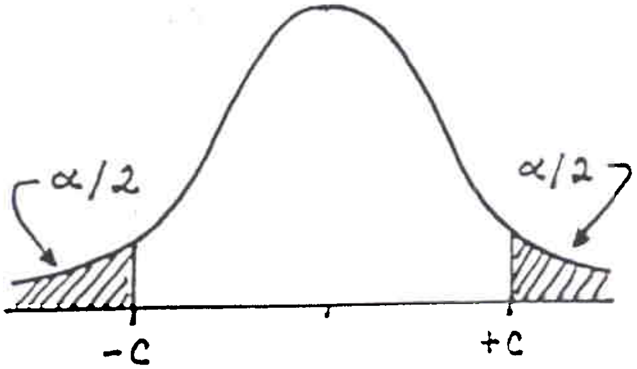
α = significance level

c = critical value = Normal distribution z-score corresponding to $\alpha/2$

note: $x/n = p$
$(n-x)/n = 1-p = q$

$k = c * \sqrt{pq/n}$

then CONF $\{(p-k) \leq \pi \leq (p+k)\}$



EXAMPLE

In a public opinion poll, 320 out of 400 persons interviewed supported their country's policy on disarmament. Establish a 95% confidence interval estimate of the percentage of persons supporting their government's stand on disarmament.

$n = 400$ $\qquad$ $p = x/n = 320/400 = 0.8$

$x = 320$ $\qquad$ $q = 1 - p = 0.2$

$\alpha = 0.05$

$c = 1.96$ (two-sided test, $\alpha/2 = 0.025$, Normal distribution table)

$$k = c * \sqrt{pq/n}$$

$$= (1.96)\sqrt{\frac{0.8(0.2)}{400}} = 1.96 * 0.02 = 0.039$$

Thus $CONF_{95\%}\{(0.8 - 0.039) \le \pi \le (0.8 + 0.039)\}$

and $CONF_{95\%}\{0.761 \le \pi \le 0.839\}$

prop-8

HYPOTHESIS TESTING

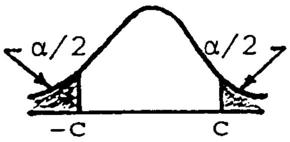
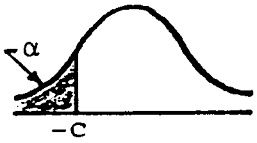
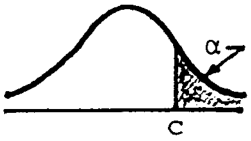
POPULATION PROPORTION - NORMAL APPROXIMATION

Variance "Known"

TEST OF HYPOTHESIS CONCERNING A SINGLE PROPORTION

(normal approximation)

Test Statistic: $z = \frac{p - \pi}{\hat{\sigma}_p}$, $\hat{\sigma}_p = \sqrt{\pi(1-\pi)/n}$

	CASE 1:	CASE 2:	CASE 3:
HYPOTHESES	H_o: $\pi = \pi_o$ H_1: $\pi \neq \pi_o$	H_o: $\pi = \pi_o$ H_1: $\pi < \pi_o$	H_o: $\pi = \pi_o$ H_1: $\pi > \pi_o$
REJECTION REGION(S)			
DECISION RULE	Reject H_o if $\lvert z \rvert > \lvert c \rvert$	Reject H_o if $z < -c$	Reject H_o if $z > c$

π_o is the hypothesized value of the population proportion. c is the critical value obtained from normal distribution tables for a particular α.

EXAMPLE

A television manufacturer claims that, on the average, 90% of his color television sets do not require any repair during the first two years of operation. The Consumer Protection Union selects a random sample of 100 sets and finds that 15 sets require some repair within the first two years of operation. If the Consumer Protection Union is willing to reject a true claim no more than 5 times in 100, will the Union reject the manufacturer's claim?

$H_0 : \pi = 0.90$
$H_1 : \pi < 0.90$

$$\left.\begin{array}{l} n = 100 \\ x = 100 - 15 = 85 \end{array}\right\} \Rightarrow p = 0.85$$

$\alpha = 0.05$; one-sided test

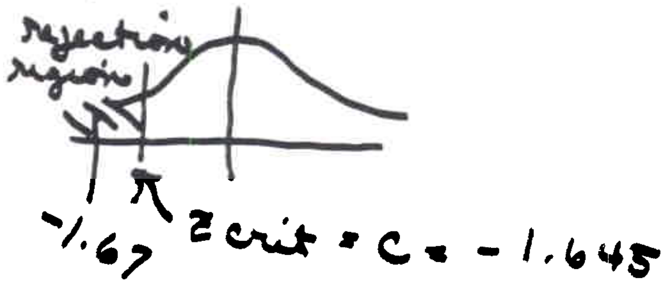
z_{crit}: $C = -1.645$ (from Normal Distribution table)

Test Statistic : $\hat{\sigma}_p = \sqrt{\pi(1-\pi)/n}$

so $\sigma_p = \{0.90(1-0.90)/100\}^{1/2} = 0.03$

thus $z_{test} = \frac{p-\pi}{\sigma_p} = \frac{0.85-0.90}{0.03} = \underline{-1.67}$

Thus we reject H_0



HYPOTHESIS TESTING

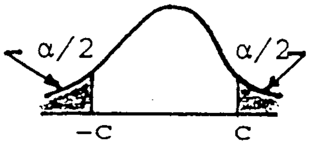
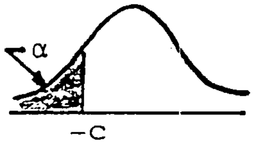
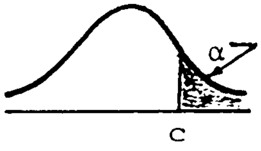
TWO POPULATION PROPORTIONS NORMAL APPROXIMATION

TEST OF HYPOTHESIS CONCERNING THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

(normal approximation)

Test statistic: $z = \dfrac{(p_1-p_2) - \overset{(\pi_1-\pi_2)}{d}}{\hat{\sigma}_d}$, $\hat{\sigma}_d = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\dfrac{n_1+n_2}{n_1 n_2}\right)}$,

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1+n_2}$$

	CASE 1:	CASE 2:	CASE 3:
HYPOTHESES	$H_o: \pi_1 - \pi_2 = d$ $H_1: \pi_1 - \pi_2 \neq d$	$H_o: \pi_1 - \pi_2 = d$ $H_1: \pi_1 - \pi_2 < d$	$H_o: \pi_1 - \pi_2 = d$ $H_1: \pi_1 - \pi_2 > d$
REJECTION REGION(S)			
DECISION RULE	Reject H_o if $\lvert z\rvert > \lvert c\rvert$	Reject H_o if $z < -c$	Reject H_o if $z > c$

EXAMPLE

Two different types of polishing solution are being evaluated for possible use in a tumble-polish operation for manufacturing interocular lenses used in the human eye following cataract surgery. Three hundred lenses were tumble-polished using the first polishing solution, and of this number 253 had no polishing-induced defects. Another 300 lenses were tumble-polished using the second polishing solution, and 196 lenses were satisfactory upon completion. At α = 0.01, is there reason to believe that the two polishing solutions differ?

$H_0 : \pi_1 - \pi_2 = 0$
$H_1 : \pi_1 - \pi_2 \neq 0$

$$\left.\begin{array}{l} n_1 = 300 \\ x_1 = 253 \end{array}\right\rangle p_1 = 0.8433 \qquad \left.\begin{array}{l} n_2 = 300 \\ x_2 = 196 \end{array}\right\rangle p_2 = 0.6533$$

$\alpha = 0.01$; two-sided test
$c = 2.575$ (from Normal Distribution Table)

<u>Test Statistic</u>: $\hat{\pi} = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{253+196}{600} = 0.7483$

$$\hat{\sigma}_d = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{n_1+n_2}{n_1 n_2}\right)} = \left\{(0.7483)(0.2517)\left(\frac{600}{90,000}\right)\right\}^{1/2}$$
$$= 0.0354$$

$$Z = \frac{(0.8433 - 0.6533) - 0}{0.0354} = 5.367$$

THUS since $(Z = 5.367) > (c = 2.575)$,
we (strongly!) Reject Ho

CHI-SQUARE ANALYSIS METHODS INVOLVING PROPORTIONS

TEST OF HYPOTHESIS OF TWO OR MORE PROPORTIONS

TEST OF VARIABLE INDEPENDENCE (BASED ON PROPORTIONS)

GOODNESS OF FIT TEST (BASED ON PROPORTIONS)

TEST OF HYPOTHESIS
OF TWO OR MORE PROPORTIONS

Given k sample sets,

$H_0: \overset{\pi_1}{p_1} = \overset{\pi_2}{p_2} = \cdots = \overset{\pi_k}{p_k}$

H_1: not all p_i are equivalent;

at least one p_i is significantly different;

the k sample sets were not all drawn from the same population

at least one sample set was drawn from a different population

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

with $(k-1)$ dof

TEST OF HYPOTHESIS OF TWO OR MORE PROPORTIONS

Samples of three kinds of materials, subjected to extreme temperature changes, produced the results shown in the following table:

	Material A	*Material B*	*Material C*	*Total*
Crumbled	41	27	22	90
Remained intact	79	53	78	210
Total	120	80	100	300

Use the 0.05 level of significance to test whether, under the stated conditions, the probability of crumbling is the same for the three kinds of materials.

SAMPLE SETS

		k_1 *Material A*	k_2 *Material B*	k_3 *Material C*	*Total*	
OUTCOMES	*Crumbled* i_1	$O_{11} = 41$ $e_{11} = 36$ $(0.30 * 120)$	$O_{12} = 27$ $e_{12} = 24$ $(0.30 * 80)$	$O_{13} = 22$ $e_{13} = 30$ $(0.30 * 100)$	90	$\frac{90}{300} = 0.30$
	Remained intact i_2	$O_{21} = 79$ $e_{21} = 84$ $(0.70 * 120)$	$O_{22} = 53$ $e_{22} = 56$ $(0.70 * 80)$	$O_{23} = 78$ $e_{23} = 70$ $(0.70 * 100)$	210	$\frac{210}{300} = 0.70$
	Total	120	80	100	300	1.00

TEST OF HYPOTHESIS OF TWO OR MORE PROPORTIONS

$H_0 : p_A = p_B = p_C$

H_1: not <u>all</u> p are equivalent

$\alpha = 0.05$

$dof = k - 1 = 3 - 1 = 2$

$\chi^2_{crit} = \chi^2(\alpha = 0.05, dof = 2) \doteq 6.0$

$$\chi^2_{test} = (41-36)^2/36 + (27-24)^2/24 + (22-30)^2/30$$
$$+ (79-84)^2/84 + (53-56)^2/56 + (78-70)^2/70$$
$$\doteq 4.58$$

Since $\chi^2_{test} < \chi^2_{crit}$, Do not reject H_0.

TEST OF VARIABLE INDEPENDENCE (BASED ON PROPORTIONS)

CHI-SQUARE CONTINGENCY TEST

Given one sample, categorized according to two characteristics (numerous attributes for each characteristic)

H_0: The characteristics are independent of one another

H_1: The characteristics are not independent of one another

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

with $(r-1)(c-1)$ dof

TEST OF VARIABLE INDEPENDENCE (BASED ON PROPORTIONS)

CHI-SQUARE CONTINGENCY TEST

To determine whether there really is a relationship between an employee's performance in the company's training program and his or her ultimate success in the job, it takes a sample of 400 cases from its very extensive files and obtains the results shown in the following table:

		Performance in training program			
		Below average	*Average*	*Above average*	*Total*
Success in job (employer's rating)	*Poor*	23	60	29	112
	Average	28	79	60	167
	Very good	9	49	63	121
	Total	60	188	152	400

Use the 0.01 level of significance to test the null hypothesis that performance in the training program and success in the job are independent.

TEST OF VARIABLE INDEPENDENCE (BASED ON PROPORTIONS)

CHI-SQUARE CONTINGENCY TEST

Performance in training program

Success in job (employer's rating)	Below average $j=1$	Average $j=2$	Above average $j=3$	Total
Poor $i=1$	$o_{11} = 23$ $e_{11} = 16.80$ $(60 * 0.2800)$	$o_{12} = 60$ $e_{12} = 52.64$ $(188 * 0.2800)$	$o_{13} = 29$ $e_{13} = 42.56$ $(152 * 0.2800)$	112 $\frac{112}{400} = 0.2800$
Average $i=2$	$o_{21} = 28$ $e_{21} = 25.05$ $(60 * 0.4175)$	$o_{22} = 79$ $e_{22} = 78.49$ $(188 * 0.4175)$	$o_{23} = 60$ $e_{23} = 63.46$ $(152 * 0.4175)$	167 $\frac{167}{400} = 0.4175$
Very good $i=3$	$o_{31} = 9$ $e_{31} = 18.15$ $(60 * 0.3025)$	$o_{32} = 49$ $e_{32} = 56.87$ $(188 * 0.3025)$	$o_{33} = 63$ $e_{33} = 45.98$ $(152 * 0.3025)$	121 $\frac{121}{400} = 0.3025$
Total	60	188	152	400

TEST OF VARIABLE INDEPENDENCE (BASED ON PROPORTIONS)

CHI-SQUARE CONTINGENCY TEST

H_0: Performance in training program and success in job are independent

H_1: Performance in training program and success in job are NOT independent

$\alpha = 0.01$

$dof = (r-1)(c-1) = (3-1)(3-1) = (2)(2) = 4$

$\chi^2_{crit} = \chi^2(\alpha = 0.01, dof = 4) \doteq 13.3$

$$\chi^2_{test} = (23-16.80)^2/16.80 + (60-52.64)^2/52.64 + (29-42.56)^2/42.56$$

$$+ (28-25.05)^2/25.05 + (79-78.49)^2/78.49 + (60-63.46)^2/63.46$$

$$+ (9-18.15)^2/18.15 + (49-56.87)^2/56.87 + (63-45.98)^2/45.98$$

$$\doteq 20.2$$

Since $\chi^2_{test} > \chi^2_{crit}$, We Reject H_0

Discrete Distributions Only

prop-20

χ^2 GOODNESS OF FIT TEST (BASED ON PROPORTIONS)

Q: At $\alpha = 0.01$, can the number of radio messages received by air traffic control during a standard time interval be modeled by a Poisson distribution with $\lambda = 4.6$?

$0.010 * 400 = 4$

Number of radio messages	# time intervals Observed frequencies O_i		Poisson probabilities	Expected frequencies e_i		k
0	3	18	0.010	4.0	22.4	1
1	15		0.046	18.4		
2	47		0.107	42.8		2
3	76		0.163	65.2		3
4	68		0.187	74.8		4
5	74		0.173	69.2		5
6	46		0.132	52.8		6
7	39		0.087	34.8		7
8	15		0.050	20.0		8
9	9		0.025	10.0		9
10	5	8	0.012	4.8	8.0	10
11	2		0.005	2.0		
12	0		0.002	0.8		
13	1		0.001	0.4		
	400			400.0		

Poisson ⇒ parameter λ ⇒ $dof = (k-r-1) = 10-1-1 = 8$

r: # parameters estimated from the data

Discrete Distributions Only

prop-21

GOODNESS OF FIT TEST (BASED ON PROPORTIONS)

From Kreyszig

Table 15.1.1. Chi-square Test of the Hypothesis that $F(x)$ is the Distribution Function of a Population from which the Sample $x_1, \cdots, x_n$ is Taken

1st step. Subdivide the x-axis into K intervals $I_1, I_2, \cdots, I_K$ such that each interval contains at least 5 values of the given sample $x_1, \cdots, x_n$. Determine the number b_j of sample values in the interval I_j ($j = 1, \cdots, K$). If a sample value lies at a common boundary point of two intervals, add 0.5 to each of the two corresponding b_j.

2nd step. Using $F(x)$, compute the probability p_j that the random variable X under consideration assumes any value in the interval I_j ($j = 1, \cdots, K$). Compute

$$e_j = np_j.$$

(This is the number of sample values theoretically expected in I_j if the hypothesis is true.)

3rd step. Compute the deviation

$$\chi_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}. \qquad (1)$$

4th step. Choose a significance level α (5%, 1%, or the like).

5th step. Determine the solution c of the equation

$$P(\chi^2 \leqq c) = 1 - \alpha$$

from the table of the chi-square distribution with $K - 1$ degrees of freedom (Table 6 in Appendix 4; cf. also Fig. 15.1.1). If $\chi_0^2 \leqq c$, do not reject the hypothesis. If $\chi_0^2 > c$, reject the hypothesis.

Discrete Distributions Only

prop-22

GOODNESS OF FIT TEST (BASED ON PROPORTIONS)

The intervals I_1 and I_K in the first step are infinite. In the case of a discrete distribution the boundary points of the intervals must not coincide with the points where $F(x)$ has jumps.

The numbers e_j in the second step should be equal to or greater than 5. If for some interval this condition is violated, then one should take a larger interval (most simply by uniting that interval with one of its neighbors). If the sample is so small that this is impossible, one should continue with the test but use the result with great care. The reason for that condition and for the appearance of the chi-square distribution in the test is a consequence of the following theorem by K. Pearson (1900).

Theorem 1. *Suppose that the hypothesis in Table* 15.1.1 *is true. Then the random variable for which* χ_0^2 *in Table* 15.1.1 *is an observed value has a distribution function which approaches the distribution function of the chi-square distribution with* $K - 1$ *degrees of freedom as n approaches infinity.*

A proof of this theorem can be found in the book by Cramér (1961), pp. 416–420; cf. Appendix 3.

If $F(x)$ *involves r unknown parameters, we may use the corresponding maximum likelihood estimates and then the chi-square distribution with* $K - r - 1$ *degrees of freedom (instead of* $K - 1$).

This rule results from a theorem by R. A. Fisher. A proof is included in the book by Cramér (1961), pp. 427–434.

χ^2 Distribution

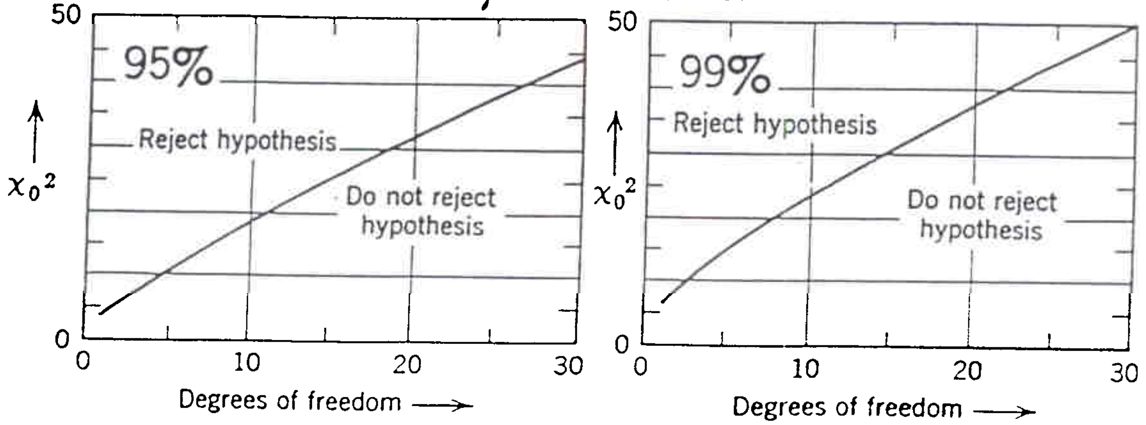


Fig. 15.1.1. Graphical representation of Table 6 in Appendix 4 ($1 - \alpha = 95\%$ and 99%)