Categorical Data Analysis

Cohen Chapters 19 & 20

For EDUC/PSY 6600

Creativity involves breaking out of established patterns in order to look at things in a different way.

--

Edward de Bono

Motivating examples

*Dr. Fisel wishes to know whether a random sample of adolescents will prefer a new of formulation of 'JUMP' softdrink over the old formulation. The **proportion** choosing the new formulation is tested against a hypothesized value of 50%.*

*Dr. Sheary hypothesizes that 1/3 of women experience increased depressive symptoms following childbirth, 1/3 experience increases in elevated mood after childbirth, and 1/3 experience no change. To evaluate this hypothesis Dr. Sheary randomly samples 100 women visiting a prenatal clinic and asks them to complete the Beck Depression Inventory. She then re-administers the BDI to each mother one week following the birth of her child. Each mother is classified into one of the 3 previously mentioned categories and **observed proportions** are compared to the **hypothesized proportions**.*

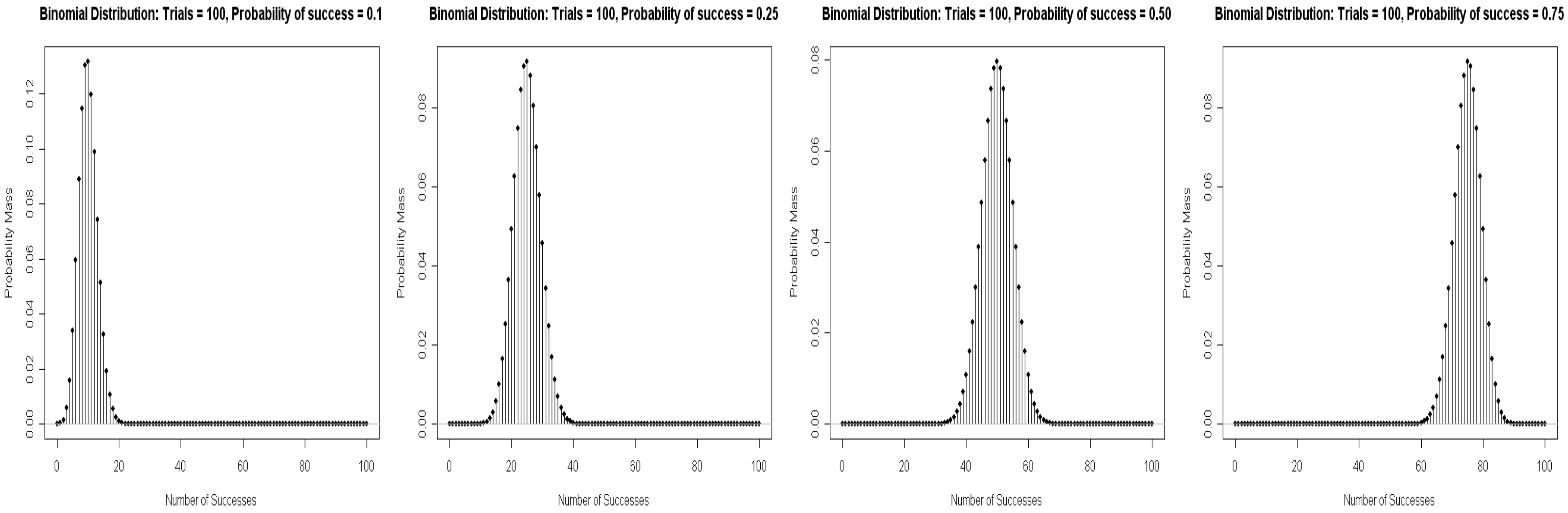
*Dr. Evanson asks a random sample of individuals whether they see both a physician and a dentist regularly (at least once per year). He compares the **distributions of these binary variables** to determine whether there is a relationship.*

Categorical Methods

- Instead of means, comparing **counts** and **proportions** within and across groups
  - E.g., # ill across different treatment groups
- Associations / dependencies among categorical variables
- Data are **nominal** or **ordinal**
- **Discrete** probability distribution
  - Number of finite values as opposed to infinite
- Each subject/event assumes 1 of 2 **mutually exclusive** values (binary or dichotomous)
  - Yes/No
  - Male/Female
  - Well/Ill

Categorical Methods

- Instead of means, comparing **counts** and **proportions** within and across groups
  - E.g., # ill across different treatment groups



or

The Binomial Distribution: EQ & coin example

$$p(X) = \frac{N!}{X!(N-X)!} P^X Q^{(N-X)}$$

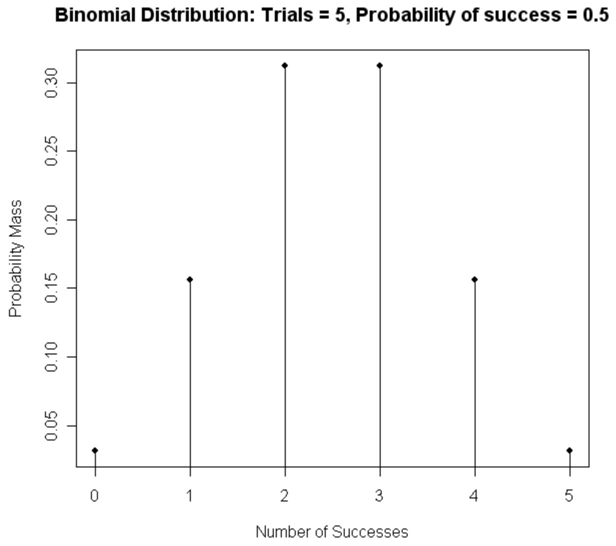
- ***N* = # events**
- ***X* = # "successes"**
- ***P* = *p*("success")**
  - **Hypothesized proportion / probability of success**
- ***Q* = *p*("failure")**
  - **Hypothesized proportion / probability of failure**
- ***P* + *Q* = 1**

- Remember: 0! = 1; $x^0 = 1$

- **(Arbitrarily) assign 1 outcome as 'success' and other as 'failure'**
- **<u>Example</u>: Probability of correctly guessing side of coin 4 out of 5 flips?**
  - **5 events, 4 successes, 1 failure**
  - ***P* = *p*(correct guess on each flip) = .50**
  - ***Q* = *p*(incorrect guess on each flip) = .50**

<u>Use equation to obtain:</u>
5 out of 5 successes = .03
4 out of 5 successes = .16
3 out of 5 successes = .31
2 out of 5 successes = .31
1 out of 5 successes = .16
0 out of 5 successes = .03

Sum of probabilities = 1.0



Sampling distribution for the binomial

- Binomial probability distribution for N = 5 events, and P = .5
- Binomial Distribution Table (exact values)
- Sampling distribution as it was derived mathematically
  - We can only reject H_o with 0 or 5 out of 5 successes (1-tailed)

Sampling Distribution

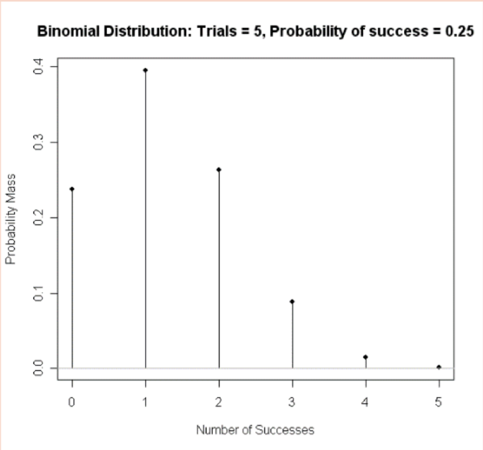
$$mean = NP$$
$$variance = NPQ$$
$$SD = \sqrt{NPQ}$$
$$SE_{MEAN} = \sqrt{\frac{PQ}{N}}$$

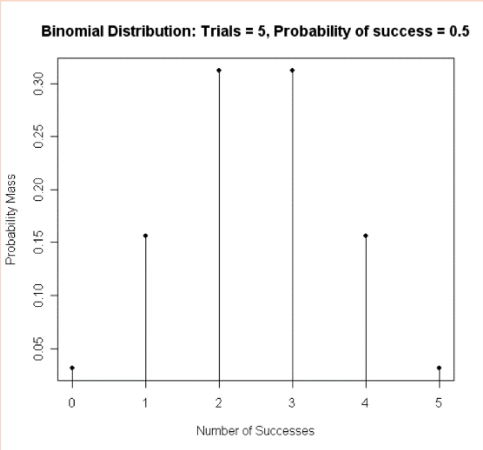
Example

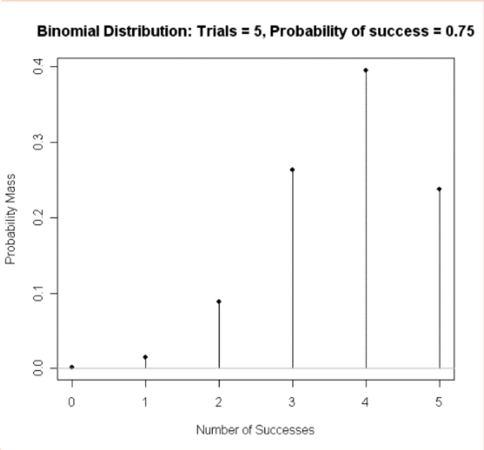
M = 5*.5 = 2.5 (See Histogram)
VAR = 5*.5*.5 = 1.25
SD = sqrt(1.25) = 1.12

Different binomial distribution for each *N*

Normal when P = .50, skewed when $P \neq$ .50
Critical value depends on: N events, X successes, P





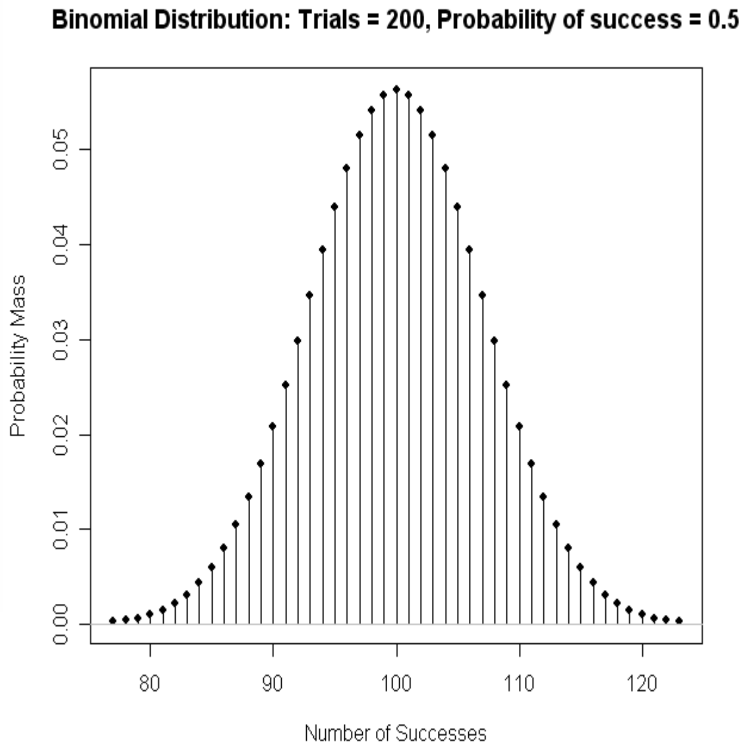


As *N* increases, binomial distribution → normal

n	X	p	n	X	p	n	X	p
1	0	.5000		1	.0176	13	0	.0001
	1	.5000		2	.0703		1	.0016
2	0	.2500		3	.1641		2	.0095
	1	.5000		4	.2461		3	.0349
	2	.2500		5	.2461		4	.0873
3	0	.1250		6	.1641		5	.1571
	1	.3750		7	.0703		6	.2095
	2	.3750		8	.0176		7	.2095
	3	.1250		9	.0020		8	.1571
4	0	.0625	10	0	.0010		9	.0873
	1	.2500		1	.0098		10	.0349
	2	.3750		2	.0439		11	.0095
	3	.2500		3	.1172		12	.0016
	4	.0625		4	.2051		13	.0001
5	0	.0312		5	.2461	14	0	.0001
	1	.1562		6	.2051		1	.0009
	2	.3125		7	.1172		2	.0056
	3	.3125		8	.0439		3	.0222
	4	.1562		9	.0098		4	.0611
	5	.0312		10	.0010		5	.1222

Table A.13
Probabilities of the Binomial Distribution for $P = .5$

"Equally Likely"
Means p = 0.5



Binomial Sign Test

- Single sample test with binary/dichotomous data
- **Proportion or % of 'successes' differ from chance?**
  - H_o: % of observations in one of two categories equals a **specified %** in population
    - H_o: Proportion of 'yes' votes = 50% in population

- Experiment: Coin flipped 10x, heads 8x
  - Is coin **biased** (Heads > .50)?
- Experiment: 10 women surveyed, 8 select perfume A
  - Is one perfume preferred **over another**?
- For both:
  - H_o: *Proportion* (X) = .50 in population
  - H_1: *Proportion* (X) ≠ .50 in population (2-tailed)

Assumptions

- Random selection of events or participants
- Mutually exclusive categories
- Probability of each outcome is same for all trials/observations of experiment

Binomial sign test: example

```
data.frame(heads = 8,
           tails = 2) %>%
  as.matrix() %>%
  as.table() %>%
  binom.test(alternative = "greater")
```

- Experiment: Coin flipped 10x, heads 8x
  - Is coin **biased** (Heads > .50)?
  - H_0: *Proportion* (X) = .50 in population
  - H_1: *Proportion* (X) ≠ .50 in population (2-tailed)

```
	Exact binomial test

data:  .
number of successes = 8, number of trials = 10, p-value = 0.05469
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4930987 1.0000000
sample estimates:
probability of success
                   0.8
```

Normal approximation to the binomial (i.e. "z-test" for a single proportion)

- **<u>What if *N* were larger, say 15?</u>**
  - Same proportions: 80% (12/15) Heads & Perfume A
  - Sum $p(12, 13, 14, 15/15) = .0178$ (1-tailed *p*-value)
- Reject H_o under both 1- and 2-tailed tests
  - 2-tailed $p = .0178 \times 2 = .0356$

<u>Experiment:</u>
Senator supports bill favoring stem cell research. However, she realizes her vote could influence whether or not her constituents endorse her bid for re-election. She decides to vote for the bill only if 50% of her constituents support this type of research. In a random survey of 200 constituents, 96 are in favor of stem cell research.

Will the senator support the bill?

- Earlier: Binomial distribution → normal distribution, as *N* → infinity
- Recommendation: Use *z*-test for single proportion when *N* is *large* (>25-30)
  - When *NP* and *NQ* are both > 10, close to normal
- H_o and H_1 are same as Binomial Test
- Test statistic:

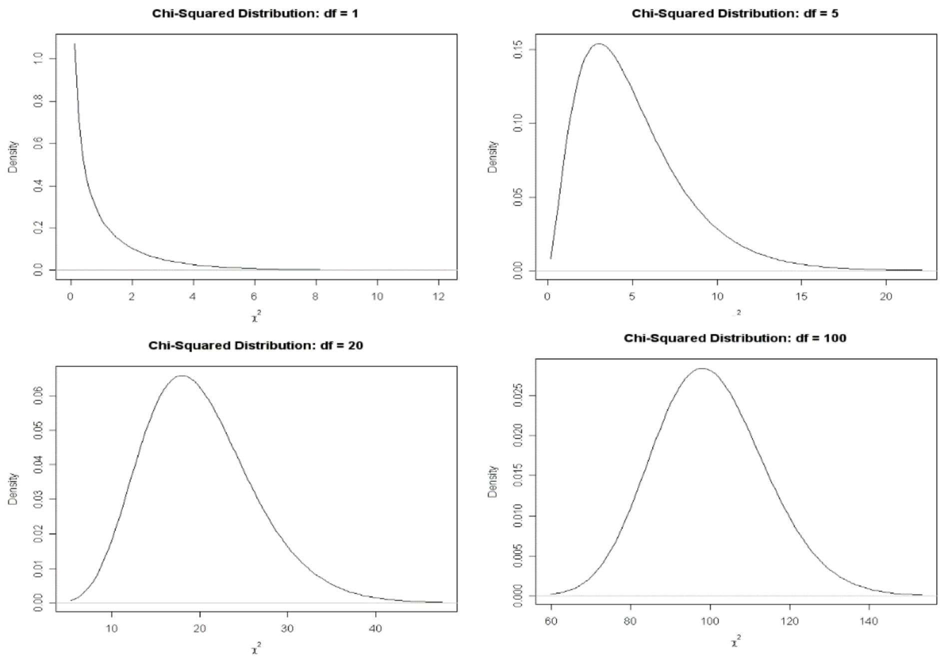
$$z = \frac{X - PN}{\sqrt{NPQ}} = \frac{p_1 - P}{\sqrt{\frac{PQ}{N}}}$$

Chi-Square (χ^2) Distribution

- Family of distributions
  - As *df* (or *k* categories) ↑
    - Distribution becomes more normal, bell-shaped
    - Mean & variance ↑
      - Mean = *df*
      - Variance = 2* *df*

- $z^2 = \chi^2$
  - Always positive, 0 to infinity
  - 1-tailed distribution

- χ^2 distribution used in many statistical tests



"GOODNESS OF FIT" Testing:

Are observed frequencies **similar** to frequencies expected by chance?

Expected frequencies

Frequencies you'd expect if H_o were true

Usually equal across categories of variable (N / k)

Can be unequal if theory dictates

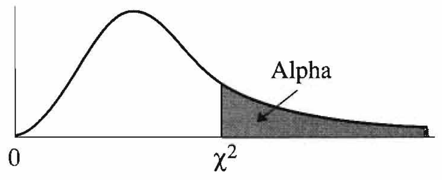
Chi-Squared: GOODNESS OF FIT Tests "GoF"

- **Hypotheses**
  - H_o: Observed = Expected frequencies in population
  - H_1: Observed ≠ Expected frequencies in population
- **General form:**
  - O = observed frequency
  - E = expected frequency

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

- If H_o were true, numerator would be small
- Denominator standardizes difference in terms of expected frequencies
- ***Aka:* Pearson or '1-way' χ^2 test**
  - 1 nominal variable
  - 2 or more categories
- If **nominal variable ONLY has 2 categories**, χ^2 GoF test:
  - Is another large sample approximation to Binomial Sign Test
  - Gives same results as z-test for single proportion as $z^2 = \chi^2$
  - Has same H_o and H_1 as binomial or z-tests
- Compare obtained χ^2 statistic to critical value based on $df = k - 1$, k = # categories

Chi-Squared: GOODNESS OF FIT Tests "GoF"



ALPHA (AREA IN THE UPPER TAIL)

df	.10	.05	.025	.01	.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.35	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.54	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.75
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.[illegible]	[illegible]	[illegible]	[illegible]
15	22.31	25.[illegible]	[illegible]	[illegible]	[illegible]
16	23.54	26.[illegible]	[illegible]	[illegible]	[illegible]
17	24.77	27.[illegible]	[illegible]	[illegible]	[illegible]

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

uencies in population
uencies in population

would be small

difference in terms of expected frequencies

χ^2 test

- If nomi
  - Is an
  - Give
  - Has
- Compar

Assumptions

Independent random sample

Mutually exclusive categories

Expected frequencies: ≥ 5 per each cell

GOODNESS OF FIT Tests – EXAMPLE: K = 2

- **Hypotheses:**
  - H_0: P = 0.50
  - Observed frequencies: 96 and 104
  - Expected frequencies: N / k =200/2 = 100df = 2 – 1 = 1

- **Test Statistic:**
- $\chi^2_{OBSERVED} =$

- **Critical Value:**
- $\chi^2_{CRIT}(\_\_\_) =$

- **Conclusion:**

Experiment:
Senator supports bill favoring stem cell research. However, she realizes her vote could influence whether or not her constituents endorse her bid for re-election. She decides to vote for the bill only if 50% of her constituents support this type of research. In a random survey of 200 constituents, 96 are in favor of stem cell research.

Will the senator support the bill?

ALWAYS USE COUNTS!!!	1 = "success"	0 = "failure"
OBSERVED (the data)	96	
EXPECTED (based on N, P, Q)		

- ***Note:***

GOODNESS OF FIT Tests – EXAMPLE: K = 2

```
data.frame(support = 96,
           not_support = 104) %>%
  as.matrix() %>%
  as.table() %>%
  chisq.test()
```

```
	Chi-squared test for given probabilities

data:  .
X-squared = 0.32, df = 1, p-value = 0.5716
```

```
exp_obs <- data.frame(support = 96,
                      not_support = 104) %>%
  as.matrix() %>%
  as.table() %>%
  chisq.test()
exp_obs$observed
exp_obs$expected
```

```
> exp_obs$observed
 96 104

> exp_obs$expected
100 100
```

Experiment:
Senator supports bill favoring stem cell research. However, she realizes her vote could influence whether or not her constituents endorse her bid for re-election. She decides to vote for the bill only if 50% of her constituents support this type of research. In a random survey of 200 constituents, 96 are in favor of stem cell research.

Will the senator support the bill?

GOODNESS OF FIT Tests – EXAMPLE: K > 2

(any number of categories within 1 variable)

ALWAYS USE COUNTS!!!

Hypotheses:

- H_0: " equally likely" (k = 6 & N = 120)
- Expected frequencies: N / k =120/6 = 20
- Observed frequencies: 20, 14, 18, 17, 22, 29 {Mon – Sat}
- $df = 6 - 1 = 5$

	M	T	W	Th	F	S
OBS	20	14	18	17	22	29
EXP						

Test Statistic:

$\chi^2_{OBSERVED} =$

Critical Value:

$\chi^2_{CRIT}(\_\_\_) =$

QUESTION:

Is there a difference in # books checked out for different days of the week?

Conclusion:

We do NOT have evidence the # of books checked out is NOT the same EVERY day

GOODNESS OF FIT Tests: **Confidence Intervals**

- ***CI*s for proportions**
  - If $k > 2$, original table converted into table with 2 cells
    - Proportion for category of interest vs proportion in **all other** categories
  - Use same formula for *z*-test for single proportion:

$$P_{obs} \pm z_{crit} \times \sqrt{\frac{P_{obs} \times Q_{obs}}{N}}$$

- **Say we wanted a *CI* for proportion of books from Saturday (29/120=0.242)**

GOODNESS OF FIT Tests: **Effect Size**

$$\chi^2_{\,Effect\ Size} = \frac{\chi^2}{N(k-1)}$$

- Ranges from 0 to 1
  - 0: Expected = Observed frequencies exactly
  - 1: Expected ≠ Observed frequencies as much as possible

GOODNESS OF FIT Tests: Post Hoc Pairwise Tests

- Like ANOVA, omnibus test, but where do differences lie?
  - 'Pinpointing the action' in contingency tables
  - Post-hoc Binomial, z-tests, or smaller 1-way χ^2 tests
    - Collapsing, ignoring levels
    - Bonferonni correction, more conservative α per comparison
  - Examining
    - Observed *vs*. expected frequencies per cell
    - Contributions to χ^2 per cell
  - Visual analysis of differences in proportions

2-way Pearson χ^2 Test of "Independence" or "Association"

- *Aka:* Contingency table, cross-tabulation, or *row* x *column (r x c)* analysis
  - > 1 nominal <u>variable</u>
- Is distribution of 1 variable *contingent* on distribution of another?
  - Is there an association or dependence between 2 categorical variables
- Extension of χ^2 Goodness of Fit Test
- **<u>Hypotheses:</u>**
  - H_o: Variables are independent in population
  - H_1: Variables are dependent in population
- Again, χ^2_{obt} is compared with χ^2_{crit} → $df = (r\text{-}1)(c\text{-}1)$

2-way Pearson χ^2 Test of "Independence" or "Association"

Same equation: Standardized squared deviations summed for all cells

$$\chi^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Different method for computing *E*

- For each cell: Multiply corresponding row and column totals (marginals), divide by *N*
  - $$E_{Cell_A} = \frac{(a+b)(a+c)}{N}$$

		Var1		
Var2	a	b	a + b
	c	d	c + d
	a + c	b + d	a + b + c + d = *N*

$$\boldsymbol{EXP_{cell} = \frac{Total_{row} \times Total_{column}}{Total_{grand}}}$$

χ^2 Test of "Independence" – Example

- **Experiment:**
- Random sample of 200 inmates are surveyed about abuse and violent criminal histories
  - Relationship between history of abuse and violent crime?

- H_0: **No association** between abuse history and violent criminal history in population of prison inmates
  - $O_{ij} = E_{ij}$ for all cells in population

- H_1: **Association** between abuse history and violent criminal history in population of prison inmates
  - $O_{ij} \neq E_{ij}$ for at least one cell in population

Observed frequencies

	Violent Crime		
Abuse	Yes	No	Row Sum
Yes	70	30	100
No	40	60	100
Column Sum	110	90	200

Expected frequencies:

Test Statistic:

APA format:

χ^2 Test of "Independence" – Example

```r
data.frame(violent_yes = c(70, 40),
           violent_no  = c(30, 60),
           row.names = c("Abuse_Yes", "Abuse_No")) %>%
  as.matrix() %>%
  as.table() %>%
  chisq.test(correct = FALSE)
```

	Violent Crime		
Abuse	**Yes**	**No**	**Row Sum**
Yes	70	30	100
No	40	60	100
Column Sum	110	90	200

```
          violent_yes violent_no
Abuse_Yes          70         30
Abuse_No           40         60
```

```
	Pearson's Chi-squared test

data:  .
X-squared = 18.182, df = 1, p-value = 2.008e-05
```

χ^2 Test of "Independence" – Example with Raw Data

	Violent Crime		
Abuse	Yes	No	Row Sum
Yes	70	30	100
No	40	60	100
Column Sum	110	90	200

```r
data %>%
  table() %>%
  chisq.test(correct = FALSE)
```

```
 ID violent abuse
 01       1     1
 02       1     0
 03       0     1
 04       1     1
 05       0     0
...     ...   ...
199       0     1
200       1     1
```

```
            violent_yes violent_no
Abuse_Yes            70         30
Abuse_No             40         60
```

```
        Pearson's Chi-squared test

data:  .
X-squared = 18.182, df = 1, p-value = 2.008e-05
```