Chi Squared Tests: Goodness of fit, Independence

9.1, 9.2

Today's topics

Response / Dependent

categorical data

Chi Squared Tests

- Goodness of fit (9.1)
- Independence (9.2)
- Exact p-value

~~2x~~ 61820

Data

numerical

weight ht

Categorical

0 1 2

Cat dog rabbit

zip code

χ^2 Goodness of Fit Test - Background

1 categorial var

Developed by Pearson in 1900

Tests the appropriateness of different models

Approximate test for use with large samples (large n).

~~2~~

Only appropriate when all expected cell counts are greater than 5. (Otherwise, should consider calculating *exact p-value*

20 c ← 30 D ←

χ^2 Goodness of Fit Test

10 Other ←

group k=3

Random sample of size n is classified into k categories or cells.

- Let $Y_1, Y_2, \ldots, Y_k$ denote the respective cell frequencies; $\sum_{i=1}^{k} Y_i$ @#n
- Denote cell probabilities $p_1, p_2, \ldots, p_k$.

p_{10}

→
- H_0: $p_1 = \boxed{p_{10}}$, $p_2 = \boxed{p_{20}, \ldots}$, $p_k = \boxed{p_{k0}.}$
- H_A: H_0 not true

p_{20}

numerical

1-Prop z test

$H_0: p = .2$

$H_A: p \neq .2$

$np \quad \{ np_1 , np_2 , np_3 , \ldots$

χ^2 Goodness of Fit Test

	Group 1	Group 2	…	Group k	Total
Observed Freq (O_i)	Y_1	Y_2	…	Y_k	n
Probability | H_o (p_{io})	p_{10}	p_{20}	…	p_{k0}	1
Expected Freq (E_i)	np_{10}	np_{20}	…	np_{k0}	n

always right tailed

C	D	O

df = 3-1
= 2

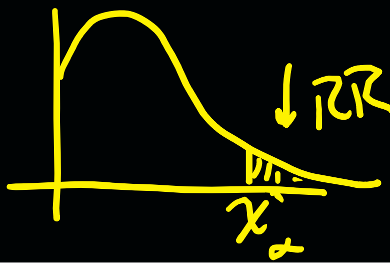
χ^2 Goodness of Fit Test

→ Test statistic:

Data

$$X^2 = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}} = \underbrace{\sum_{i=1}^{k} \frac{(Obs_i - Exp_i)^2}{Exp_i}}_{\text{test stat}} \sim \chi^2_{(k-1)}$$

Reject H_0 if $X^2 \geq \underbrace{\chi^2_{(k-1),\alpha}}_{\text{crit}}$



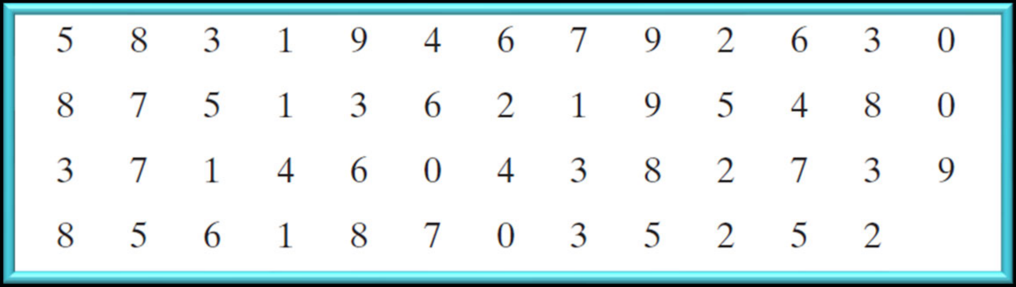
$$\sum \frac{(O-E)^2}{E}$$

5 8, 2, 9, 7, 6, 6

Random Digit Example (Goodness of Fit)

When making random numbers, people are usually reluctant to record the same or consecutive numbers in adjacent positions. Even though these true probabilities should be p_{10} = 1/10 and p_{20} = 2/10, respectively. gt8suv tests a friend's concept of a random sequence by asking them to generate a sequence of 51 *random* digits.

May0 generates the following sequence: (Ex 9.1-1)



5	8	3	1	9	4	6	7	9	2	6	3	0
8	7	5	1	3	6	2	1	9	5	4	8	0
3	7	1	4	6	0	4	3	8	2	7	3	9
8	5	6	1	8	7	0	3	5	2	5	2	

Use a goodness of fit test to make a statistical decision about whether this sequence seems to be truly random or not (at α = 0.05)

50

5	8	3	1	9	4	6	7	9	2	6	3	0
8	7	5	1	3	6	2	1	9	5	4	8	0
3	7	1	4	6	0	4	3	8	2	7	3	9
8	5	6	1	8	7	0	3	5	2	5	2	

Random Digit Example (Goodness of Fit)

$H_0 : p_1 = p_{10} = 1/10,\ p_2 = p_{20} = 2/10,\ p_3 = p_{30} = 7/10$

H_A : ?

Group 1 Same

G2 Cons

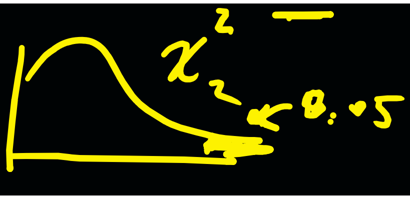
G3 other

	Group 1	Group 2	…	Group k	Total
Observed Freq (O_i)	Y_1 0	Y_2 8	…	Y_k 3 42	n 50
Probability | H_o (p_{io})	p_{10} 1/10	p_{20} 2/10	…	p_{k0} 7/10	1
Expected Freq (E_i)	np_{10} 5	np_{20} 10	…	np_{k0} 35	n

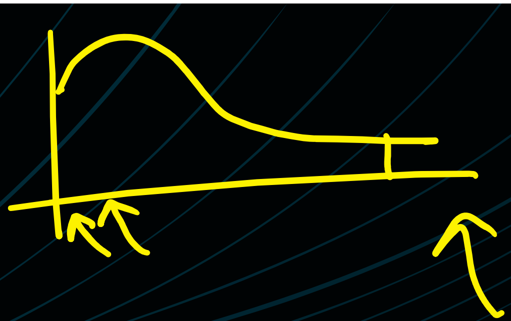
test stat ✓

$$\frac{(0-5)^2}{5}+\frac{(8-10)^2}{10}+\frac{(42-35)^2}{35}=6.8$$

$$6.8 > 5.991 = \chi^2_{0.05}(2).$$



$$\underbrace{\sum \frac{(O-E)^2}{E}}_{\text{small}} = 0$$



- Reject H_0:
- There is significant evidence at alpha = 0.05 to suggest a lack of fit. The sequence does **not** seem to be random.

* good fit

Obs	~~5~~ 4	~~10~~ 11	35
Exp	5	10	35

3 factors or levels

P_{10}

	Y	B	R
O	25	40	35
E	30	50	20

$$\sum \frac{(O-E)^2}{E} = \frac{5^2}{30} + \frac{10^2}{50} + \frac{15^2}{20} \sim \chi^2_2$$

Quick Example Goodness of fit

Claim skittles

- ~~M&M~~s: 1 2 3
- 30% yellow, 50% blue, 20% red
- You open a bag of 100 skittles ~~M&M~~s and get 25 Y, 40 B, 35 R

$H_0: p_1 = .3, \; p_2 = .5, \; p_3 = .2$

1 variable: color

Definition

Contingency table

Summarizes relationship between categorical variables by displaying frequency distribution.

e.g.

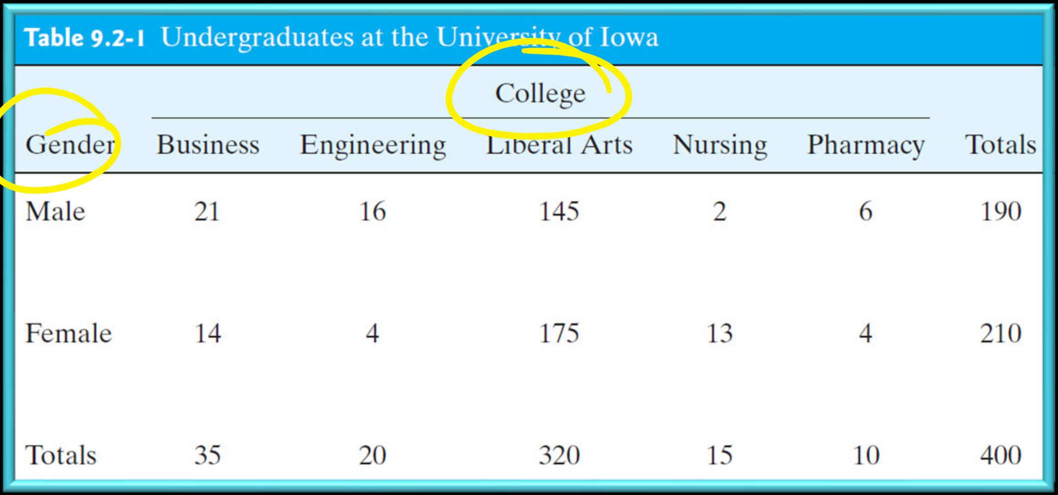


Table 9.2-1 Undergraduates at the University of Iowa

	College					
Gender	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Totals
Male	21	16	145	2	6	190
Female	14	4	175	13	4	210
Totals	35	20	320	15	10	400

χ^2 Test for Homogeneity & Independence (9.2)

same test

Tests relationship between two categorical variables.

The difference in the two names depends on how the data is collected.

χ^2 Test for Homogeneity & Independence (9.2)

χ^2 Test for **Homogeneity**: (one margin fixed)

Tests whether two or more sub-groups of a population share the same distribution of a single categorical variable. E.g., do different age groups have the same proportion of people who prefer Twitch, YouTube Live, or Zoom?

Independent random samples from r populations.

Each sample is classified into c response categories.

$\boxed{\mathrm{H}_0}$: In each category, the probabilities are equal for all r populations.

AGE

Pref	Y	M	O
T			
Y			
Z			

χ^2 Test for Homogeneity & Independence (9.2)

data collection step

χ^2 Test for **Independence**: (no margins fixed)

Tests whether two categorical variables are associated with one another in the population, e.g. age group vs video streaming platform preference.

A random sample of size n is simultaneously classified with respect to two characteristics, one has r categories and the other c categories.

H_0 : The two classifications are independent; i.e., each cell probability is the product of the row and column marginal probabilities

(9.2) notes

2 cat. variables Grade, Instructor

Homogeneity (Independence)

In each group (BooleanHypercube, Obese future), I collect a separate sample. I look at whether Instructor and grade are related

Independence

I collect a large sample of students who took 420 and ask them who they had as an instructor, as well as their grade.

notes

Why would we do a test for homogeneity?

Say you have a rare disease that you are trying to determine something about

If you do test for independence and collect a sample of size n

χ^2 Test for Homogeneity & Independence (9.2)

Test statistic (both tests):

$$X^2 = \sum_{\text{cells}} \underbrace{\frac{(O-E)^2}{E}}_{\star} \sim \chi^2_{\underbrace{(r-1)}*\underbrace{(c-1)}}$$

	Y	O	
	5		
Z	1	9	10
T	9	1	10
	10	10	20

$(2-1)\cdot(2-1)$ df $= 1$

$\frac{10 \cdot 10}{20} = 5$

#	#	F
F	F	
#	#	

O = observed cell frequency

E = $\frac{\text{row total} * \text{column total}}{\text{overall total}}$

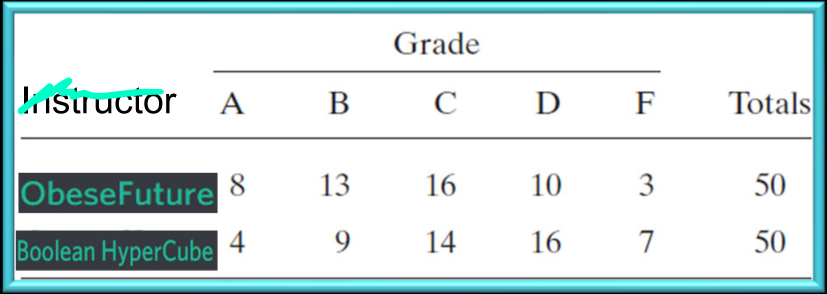
(more formally):

$$X^2 = \underbrace{\sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}} \sim \chi^2_{(r-1)(c-1)}$$

χ^2 Test of homogeneity example:

Saumdog is wondering which instructor to select for Stat 420. Albert doesn't know much so he states H_0: their grade distributions are the same.

We collect 2 separate random samples from each instructor and obtain the following data.



Instructor

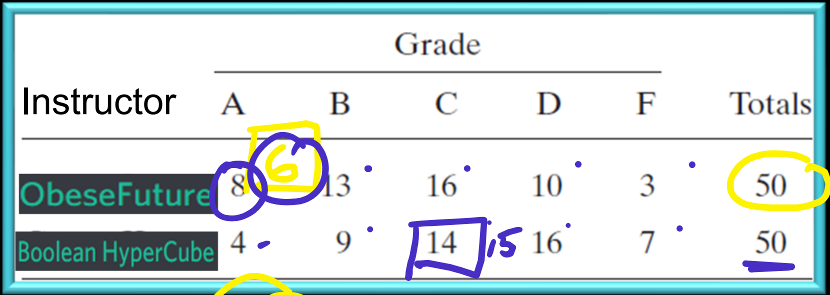
	Grade					
~~Instructor~~	A	B	C	D	F	Totals
ObeseFuture	8	13	16	10	3	50
Boolean HyperCube	4	9	14	16	7	50

Perform a χ^2 test of homogeneity at significance level 0.05 to determine if these grade distributions are similar.

$\alpha = 0.05$

obs

EXP



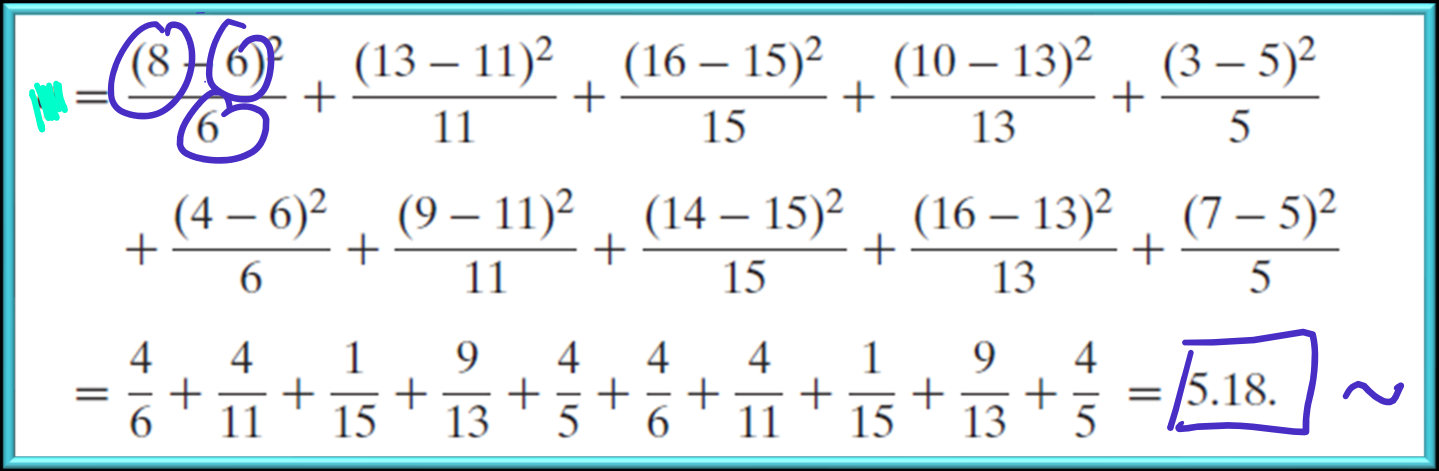
	Grade					
Instructor	A	B	C	D	F	Totals
ObeseFuture	8 6	13	16	10	3	50
Boolean HyperCube	4	9	14 15	16	7	50
	12	22	30	26	10	100

$$\frac{50 \times 12}{100} = \underline{6}$$

$$\frac{30 \times 50}{100} = 15$$

df

$(5-1)(2-1) = 4$



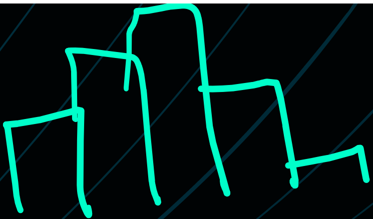
$$\chi^2 = \frac{(8-6)^2}{6} + \frac{(13-11)^2}{11} + \frac{(16-15)^2}{15} + \frac{(10-13)^2}{13} + \frac{(3-5)^2}{5}$$

$$+ \frac{(4-6)^2}{6} + \frac{(9-11)^2}{11} + \frac{(14-15)^2}{15} + \frac{(16-13)^2}{13} + \frac{(7-5)^2}{5}$$

$$= \frac{4}{6} + \frac{4}{11} + \frac{1}{15} + \frac{9}{13} + \frac{4}{5} + \frac{4}{6} + \frac{4}{11} + \frac{1}{15} + \frac{9}{13} + \frac{4}{5} = \boxed{5.18.} \sim \chi^2_4$$

test stat

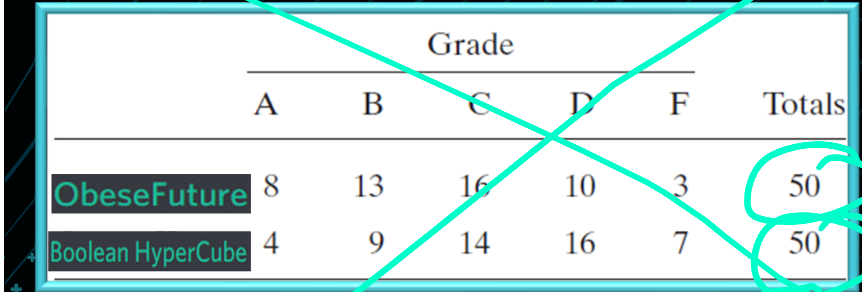
distrib



say I collect $n = 100$ studen

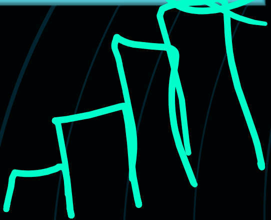
Test of independence

Homogereity ←



	Grade					
	A	B	C	D	F	Totals
ObeseFuture	8	13	16	10	3	50
Boolean HyperCube	4	9	14	16	7	50

Inft



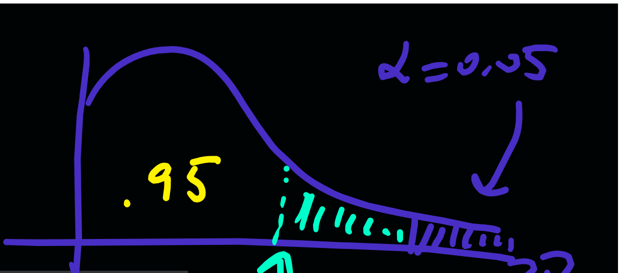
	Grade A	B	C	D	F	Tot
OF	12	10	14	10	16	60
BH	8	6	10	8	8	40
	20	16	24	18	18	100

notes

DNR H_0

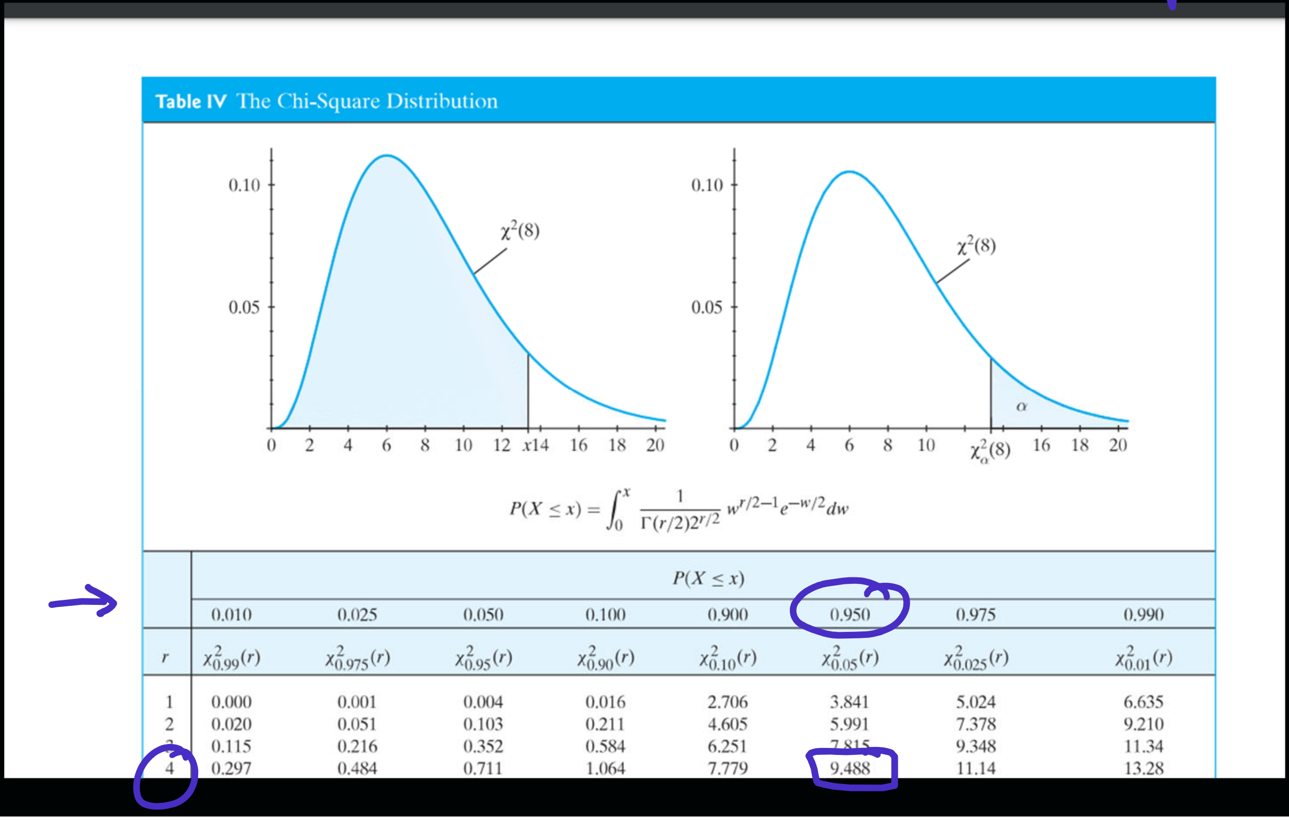
5.18

χ^2_4



5.18 t.s.

9.488 c.v.



0.10
0.05
$\chi^2(8)$
0 2 4 6 8 10 12 x14 16 18 20
0.10
0.05
$\chi^2(8)$
α
0 2 4 6 8 10 $\chi^2_\alpha(8)$ 16 18 20

Table IV The Chi-Square Distribution

$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28

Example

A random sample of 400 undergraduate students at the University of Iowa was selected, then classified by college and gender. Are these variables independent at α = 0.01?

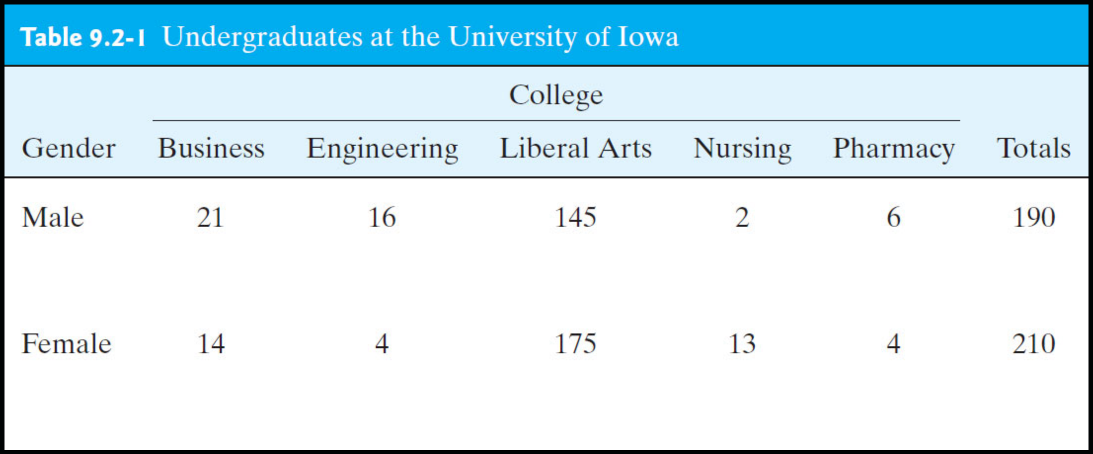


Table 9.2-1 Undergraduates at the University of Iowa

	College					
Gender	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Totals
Male	21	16	145	2	6	190
Female	14	4	175	13	4	210

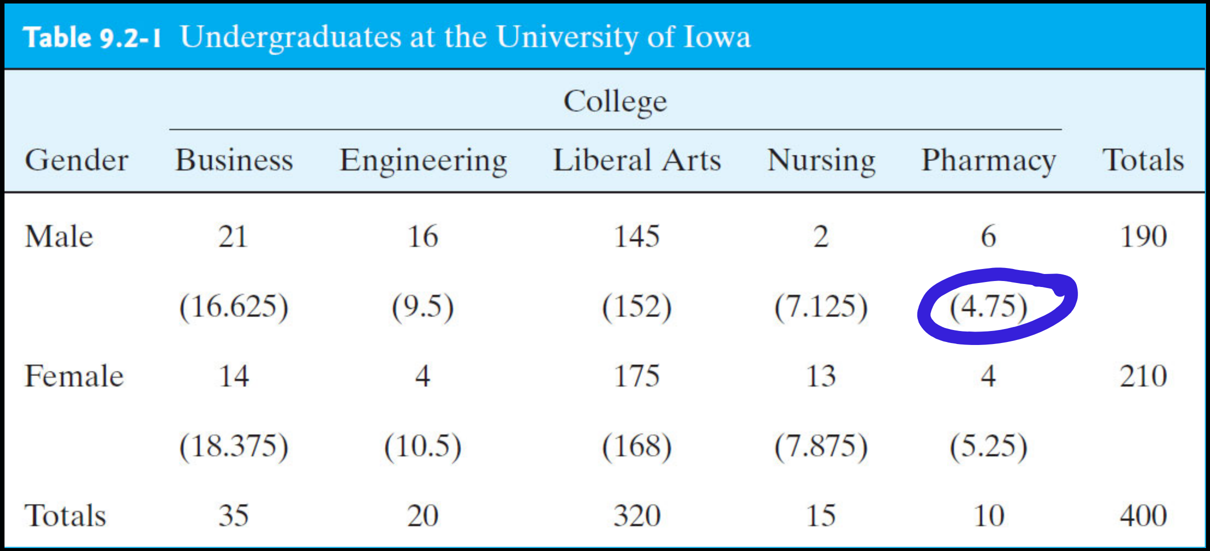
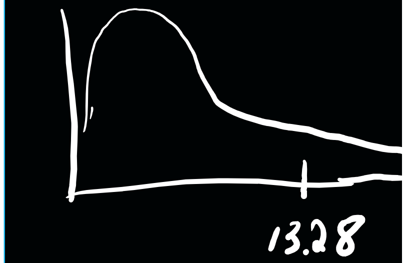
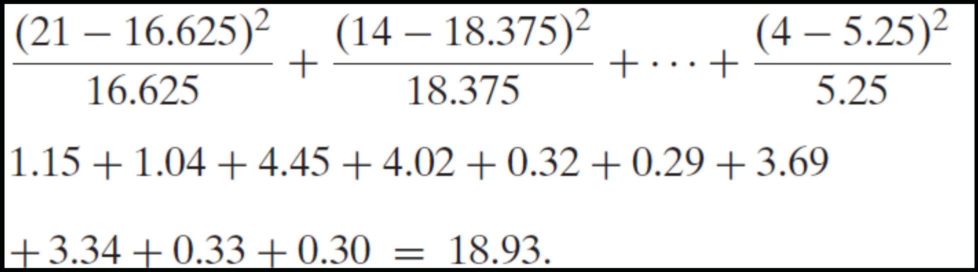


Table 9.2-1 Undergraduates at the University of Iowa

	College					
Gender	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Totals
Male	21	16	145	2	6	190
	(16.625)	(9.5)	(152)	(7.125)	(4.75)	
Female	14	4	175	13	4	210
	(18.375)	(10.5)	(168)	(7.875)	(5.25)	
Totals	35	20	320	15	10	400





$$\frac{(21-16.625)^2}{16.625}+\frac{(14-18.375)^2}{18.375}+\cdots+\frac{(4-5.25)^2}{5.25}$$

$1.15 + 1.04 + 4.45 + 4.02 + 0.32 + 0.29 + 3.69$

$+ 3.34 + 0.33 + 0.30 \; = \; 18.93.$

critical value : 13.28

Exact p-value

(not really a new definition)

Review: What is the definition of a p-value?

Calculate this probability directly, instead of using a Normal or Chi Squared approximation.

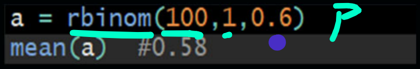
prop test

$H_0: p = 0.5$ $H_A: p > 0.5$

Type II Error

Exact p-value - Example

Luigiatl wants to test whether a coin is a fair coin or not (loaded in favor of Heads) at α = 0.05. Data from 100 flips (heads = 1, tails = 0):



```
a = rbinom(100,1,0.6)
mean(a)  #0.58
```

P

1 1 1 0 1 0 0 0 1 0 1 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 1 1 0 0 1 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 1 1 0 0 1 0 0 1 0 1 1 1 0 0 0 1 0 1 0 0 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 0 1 1 1 1 1 1 0 1 0 1 0 1

(Could find Z value and do a test that we already know) **OR**

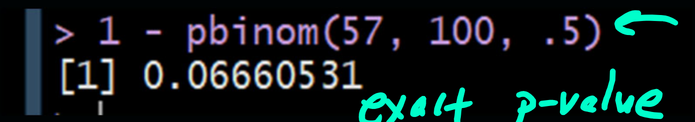
(9.1) → 58 1's
42 0's

To find exact p-value, just use the definition of the p-value

Let X = # of heads in 100 tosses Under H_0: X $\sim Binomial(100,\ 0.5)$

P-value = $P[X \geq 58 | p = 0.5]$. $= 1 - P\{X \leq 57 / p = 0.5]$

In R:



```
> 1 - pbinom(57, 100, .5)
[1] 0.06660531
```

exact p-value

notes

if $n = 10000$

approx easier

if small sample sizes,

can't use CLT, can't use χ^2

use exact p-value

notes

if α was lower 0.01

what would happen to Type II error ?

↑

$\alpha \uparrow$, $\beta \downarrow$