CS 2810 April 5 Day 20

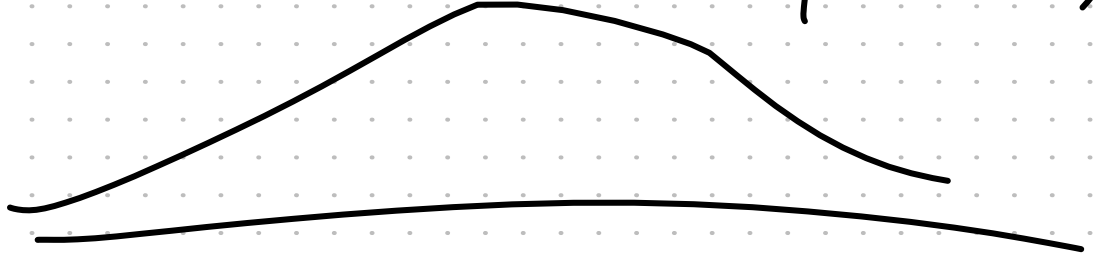
chi square "goodness of fit"

Admin:
- short class
- reschedule quiz4 to may 3? (with section 2)
    - Ell Hall AUD (8-10am)
    - please email me if you can't make it

"STANDARD" NORMAL DISTRIBUTION

$Z \sim N(\mu=0, \sigma^2=1)$

Z IS OFTEN USED AS NAME OF SAMPLE FROM THIS DISTRIBUTION



IF $X \sim N(\mu, \sigma^2)$ THEN $\frac{X-\mu}{\sigma} \sim N(\mu=0, \sigma^2=1)$

"Z-SCORE"

Let X be a normally distributed random variable with mean 7 and variance 10. Identify the linear function of X so that it has a "standard" normal distribution (mean 0 and variance 1).

$$X \sim N(\mu = 7, \sigma^2 = 10)$$

$X_0 = 14$

$$Z_0 = \frac{14-7}{\sqrt{10}} \approx \frac{7}{3.1} \approx 2.2$$

OBSERVATION IS 2 STD DEV ABOVE MEAN

OUTCOME

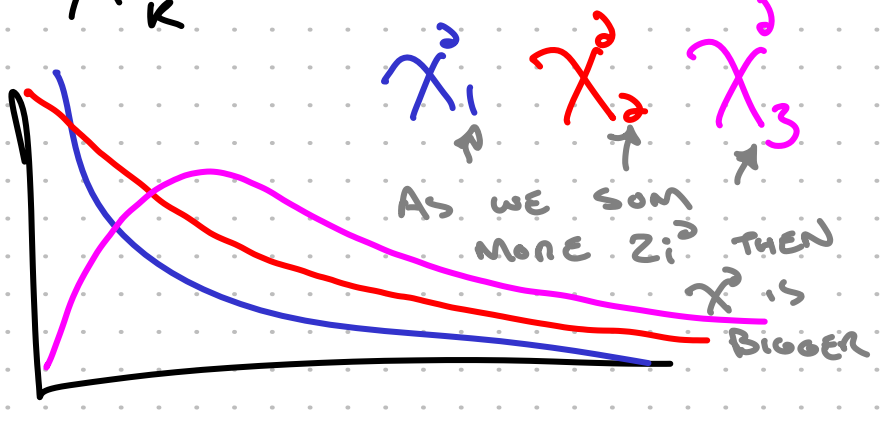
$$Z = \frac{X-\mu}{\sigma} = \frac{X-7}{\sqrt{10}} = \frac{1}{\sqrt{10}}X - \frac{7}{\sqrt{10}} \sim N(\mu = 0, \sigma^2 = 1)$$

Chi-Square Distribution

Let $Z_i \sim N(0,1)$ be iid standard normal

Chi-square is sum of k standard normal squared

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$$



CHI-SQUARE "GOODNESS OF FIT" TEST

ARE THESE OUTCOMES OF A 6-SIDED DIE FAIR?

Do these N=36 outcomes come from a fair (uniform) 6-sided die?

H_0: PROB OF EACH OUTCOME IS 1/6

H_1: NO ITS NOT

(we sort outcomes below so they're easier to work with):

1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6

8	9	5	2	7	5
O_1	O_2	O_3	O_4	O_5	O_6

O_i HOW MANY TIMES DID WE OBSERVE OUTCOME i?

FOR CHI-SQUARE TEST

H_0: OBSERVATIONS COME FROM $P(x)$

H_1: NO THEY DON'T

ASSUME

→ OBSERVATIONS INDEPENDENT

Assuming H_0 (Die is fair in this Ex) what is expected count of each outcome?

$$E_i = N \cdot p_i = 36 \cdot \frac{1}{6} = 6$$

Total outcomes (36 in this Ex)

Prob of each outcome under H_0

CONTINGENCY TABLE + χ^2 TEST STATISTIC

	1	2	3	4	5	6
OBSERVED (O_i)	8	9	5	2	7	5
EXPECTED (E_i)	6	6	6	6	6	6
$\frac{(O_i - E_i)^2}{E_i}$	$\frac{(8-6)^2}{6}$	$\frac{(9-6)^2}{6}$	$\frac{(5-6)^2}{6}$	$\frac{(2-6)^2}{6}$	$\frac{(7-6)^2}{6}$	$\frac{(5-6)^2}{6}$

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

$$\chi^2 = \frac{4}{6} + \frac{9}{6} + \frac{1}{6} + \frac{16}{6} + \frac{1}{6} + \frac{1}{6} = \frac{32}{6}$$

ICA 1:
- Does there exist a minimum or maximum chi-squared statistic?
- Describe what kind of O_i and E_i achieve this min or max chi-squared statistic.
- Which values of the chi-squared stat are most typical of the null hypothesis? Justify your response with one or two sentences.

	1	2	3	4	5	6
OBSERVED (O_i)	8	9	5	2	7	5
EXPECTED (E_i)	6	6	6	6	6	6
$\frac{(O_i - E_i)^2}{E_i}$	$\frac{4}{6}$	$\frac{9}{6}$	$\frac{1}{6}$	$\frac{16}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

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

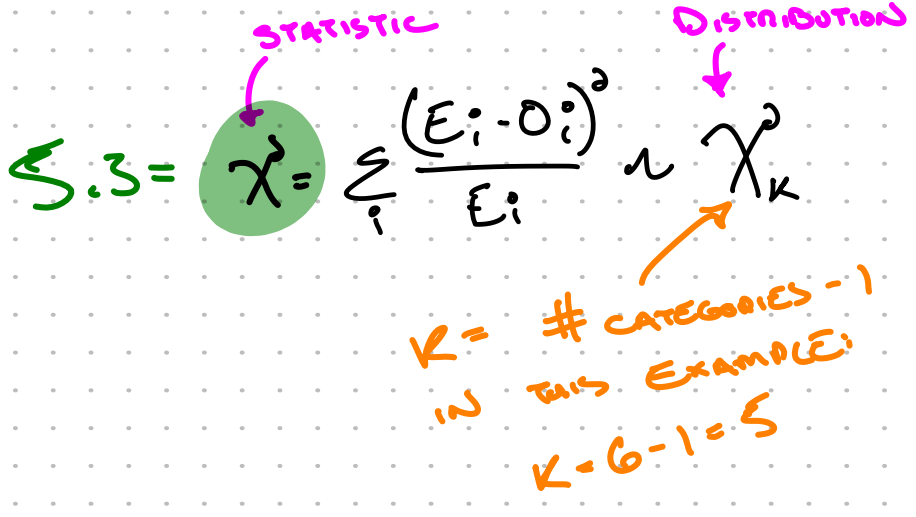
$\chi^2 \approx 5.3$

min chi square is 0. it is achieved when the observed count of each outcome equals the expected count of each outcome.

chi-square = 0 is most typical of the null hypothesis
expected counts (E_i) assume the null hypothesis

Modeling Null Hypothesis

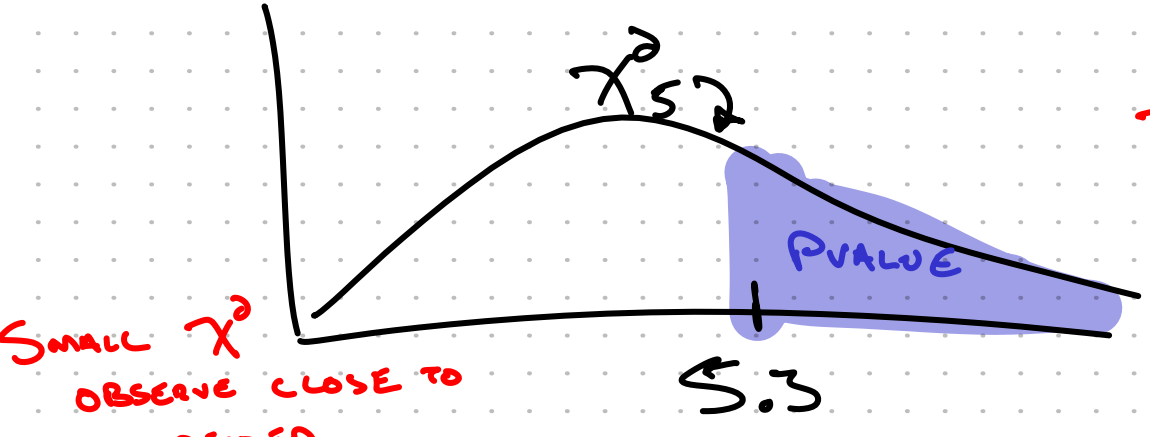
Assuming the null hypothesis (die is fair) then the chi square statistic follows a chi square distribution with k = size of sample space - 1 degrees of freedom (df=6-1=5 in this example)



Computing P-value with Chi-Squared Goodness of Fit

Remember: P-value is prob of all outcomes which are less consistent with null hypothesis

ASSUMING H_0, $\chi^2 = 5.3$ IS χ^2 DISTRIBUTED WITH DF=5

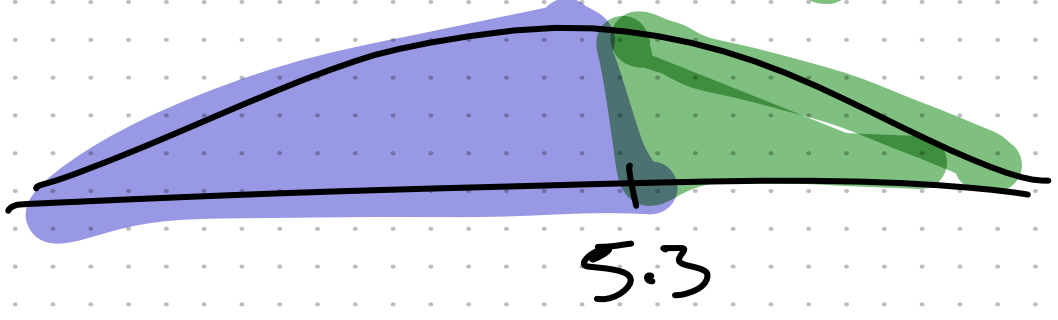


$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$P_{VALUE} =$

$1 - CDF(5.3) = .38$

$CDF(5.3)$



PVAL $> \alpha = .05$

DO NOT REJECT H_0

MAKE NO CLAIMS

PVAL $< \alpha = .05$

REJECT H_0 CLAIM H_1

"DIE IS NOT FAIR"

ICA 2

A "silly die" is supposed to roll higher outcomes more often than others:

OUTCOME	1	2	3	4	5	6
PROB	1/21	2/21	3/21	4/21	5/21	6/21

H_0: DIE IS "SILLY"

H_1: DIE IS NOT "SILLY"

If a die is observed to roll:

1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6

3 3 4 7 4 12

perform a hypothesis test to (potentially) claim the die is not "silly".

OUTCOME	1	2	3	4	5	6	
PROB	$1/21$	$2/21$	$3/21$	$4/21$	$5/21$	$6/21$	
EXPECTED E_i	$\frac{1}{21} \cdot 33$	$\frac{2}{21} \cdot 33$	$\frac{3}{21} \cdot 33$	$\frac{4}{21} \cdot 33$	$\frac{5}{21} \cdot 33$	$\frac{6}{21} \cdot 33$	
OBSERVED O_i	3	3	4	7	4	12	$3+3+4+7+4+12 = 33$ TOTAL OUTCOMES
$\chi^2 = \sum \frac{(E_i - O_i)^2}{E_i}$	1.29	.006	.108	.081	1.89	.701	

$\chi^2 = 4.08$

PVAL = 1 − CHI2.COF(4.08, DF = 5) ≈ .53

DON'T REJECT H_0

CHI-SQUARE "BINNING"

χ^2 REQUIRES A FINITE SAMPLE SPACE

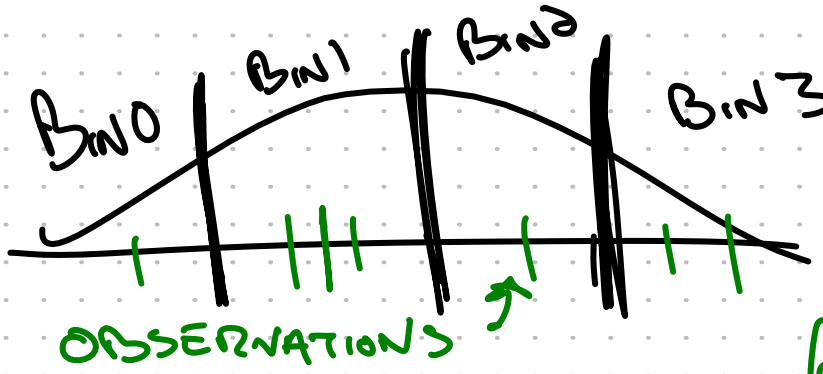
WE CAN "BIN" A DISTRIBUTION:

EXPECTED $N \cdot \int_{x_1}^{x_2} P(x)\,dx$

TOTAL OBSERVATIONS — PROB IN BIN (1)

OBSERVED ARE SPLIT INTO BINS

CHOOSING BINS IMPACTS ANALYSIS SENSITIVITY



7 TOTAL OBS

$$P(\text{BIN1}) = \int_{-\infty}^{T_0} P(x)\,dx$$

- CHOOSE BIN EDGES SO BINS HAVE EQUAL PROB

	BIN 1	BIN 2	BIN 3	BIN 4
O_i	1	3	1	2
E_i				

$$Z = \frac{\bar{X} - \bar{Y}}{S}$$

Assuming $\mu_x = \mu_y$

$$Z \sim N(0,1)$$

VALUES HERE SUGGEST $\mu_Y > \mu_X$

VALUES HERE SUGGEST $\mu_X > \mu_Y$