Chapter 12: Inference on Categorical Data

x	$P(x)$
0	0.33
1	0.33
3	0.34

- $\sum P(x) = 1$
- $0 \le P(x) \le 1$

Review: X, x

- Discrete Random Variables: Probability Distributions, Expected Value, Binomial Distribution
  - 6.1 Discrete Random Variables and their Probability Distributions
  - 6.1 Expected Value of Probability Distribution
    - $E(X) = \mu = \sum [x \cdot P(x)]$ key "long run behavior"
  - 6.2 Binomial Probability Distribution
    - 4 requirements: (1) fixed trials n (2) trials independent (3) there are only two (BI) outcomes possible (4) probability of success is constant.
    - Expected Value: $\boxed{E(X) = \mu = n \cdot p}$ $\left(\sum E_i = 1\right)$

Examples
- flip coin 10 times
- having 3 kids

Carnival game

x	$P(x)$
-1	0.5
-5	0.2
10	0.2
20	0.1

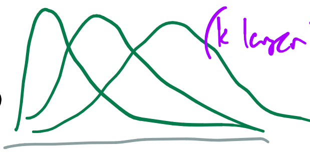
Dice

x	$P(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

notation: p = prob. success
$q = 1 - p$ = prob. failure

- Chi-Squared Distribution
  - 9.3 Chi-Squared Distribution
    - Shape: not symmetric, skewed (R)
    - Depends on $df = k$
    - Can you find probability under χ^2-distribution? What about critical values?
      ↳ yes, $\chi^2cdf(a, b, df)$ ↳ $Inv\chi^2$? No! Table!

χ^2



Section 12.1: Goodness-of-Fit Test

In this section, we study a procedure to test hypotheses about a probability distribution. For example, you might want to test if a dice is fair with each side having probability $p = \frac{1}{6}$.

Def A **goodness-of-fit test** is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.
→ sample data ↳ theoretical dist.

The M&M company claims that the distribution of plain M&M candies in a bag is 23% blue, 23% orange, 15% green, 12% red, 15% yellow, and 12% brown. Even though this is their claim, do you think this represents the true proportions of color distribution in all of the M&M bags? ***How would we check?***

$p_{Blue} = p_1$ $p_{Orange} = p_2$ etc....

claimed prop

(Blue) Orange Green Red Yellow Brown

H_0: $p_1 = 0.23$, $p_2 = 0.23$, $p_3 = 0.15$, $p_4 = 0.12$, $p_5 = 0.15$, $p_6 = 0.12$

H_A: At least one of the proportions differs from the claimed proportion.

What to Compare and How to Compare It

Expected Counts (E_i)	Observed Counts (O_i)
The number in each category we would expect to see if H_0 is true. * Two ways of calculating the expected counts: 1. If the expected counts are EQUAL, then $\boxed{E = \frac{n}{k}}$ • where n = total observed • where k = # of categories 2. If the expected counts are not equal, then calculate using $\boxed{E_i = \mu_i = n_i \cdot p_i}$ where $i = 1, 2 \dots k$ for each category	Observe how many in your **sample** are in each category. **This information will be given.*

*If the observation and experiment counts are *"close"*, then Fail to Reject H_0
(keep H_0)

LOGIC of HT

If the observation and experiment counts *"far apart"* (BIG DIFFERENCE), then Reject H_0

Ex 1: Finding the expected counts.

(a) A single die is rolled 45 times with the following results. Assuming that the die is fair and all outcomes are equally likely, find the expected frequency E for each empty cell.

↳ E are all same!

Given

Outcome	**1**	**2**	**3**	**4**	**5**	**6**
Observed O_i	13	6	12	9	3	2
Expected E_i	7.5	7.5	7.5	7.5	7.5	7.5

$\mu = E = \frac{n}{k} = \frac{\text{total observations}}{\#\text{categories}} = \frac{45}{6} = 7.5$

other formula: $E = n \cdot p = 45\left(\frac{1}{6}\right) = 7.5$ (fair)

(b) Jon works as an usher at a theatre. The theatre has 1000 seats that are accessed through five entrances. Each guest should use the entrance that's marked on their ticket. Entrances A and B should each have 30% of the guests using these entrances. Entrance C should have 20% of the guests using its entrance. Entrances D and E should each have 10% of the guests using these entrances. Find the expected frequency for each E for each entrance.

Given

Entrance	**A**	**B**	**C**	**D**	**E**
Observed O_i	398	202	205	87	108
Expected E_i	300	300	200	100	100

Note: $\sum E_i = n$

$(i=1,2)$ $E_i = n_i p_i = 1000 \cdot 0.3 = 300$

$(i=3)$ $E_i = n_i p_i = 1000 \cdot (0.20) = 200$

$E_i = n_i \cdot p_i$ $(i=3,4)$ $= 1000 \cdot (0.10) = 100$

Steps for Hypothesis Test for Goodness-of-Fit	
What to Find… • Number of categories, k • Expected Counts, E_i	***Check Requirements*** • The data has to be randomly selected. • The sample data consist of frequency counts for each of the different categories. (none missing) • For each category, the expected frequency is at least 5.
Step 1: *Hypotheses* $H_0: p_1 = p_2 = \cdots = p_k$ $(=\#)$ (EQUAL OUTCOMES – DICE) H_A: *At least one of the propbabilities is different from the others* ($p_i \neq \#$) OR $H_0: p_1 = \#\ \ p_2 = \#, \ldots, p_k = \#$ (Potentially different – M&Ms) H_A: *At least one of the propbabilities is different the claimed distribution*	***Step 2:*** *Level of Significance* α $\alpha = P(\text{Type I Error})$ if not given, assume 0.05
Step 3: *Test Statistic*  χ^2-dist $$\boxed{\chi_0^2 = \sum \frac{(O-E)^2}{E}}$$ **ALWAYS** RIGHT **-TAILED TEST!** **Note:** To compute the test statistic you will need to use lists on your calculator!	

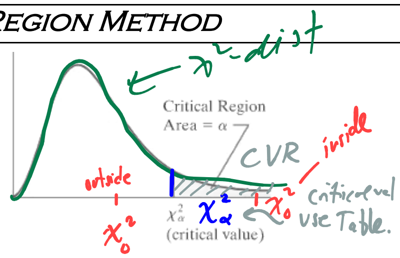
Calc: $L1 = O_i$, $L2 = E_i$, $L3 = \frac{(L1 - L2)^2}{L2}$

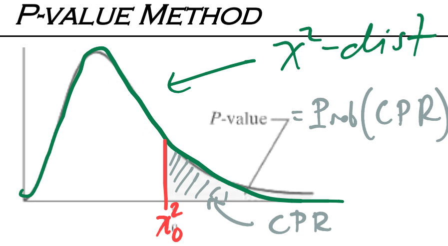
Test Statistic $\chi_0^2 = \text{sum}(L3)$

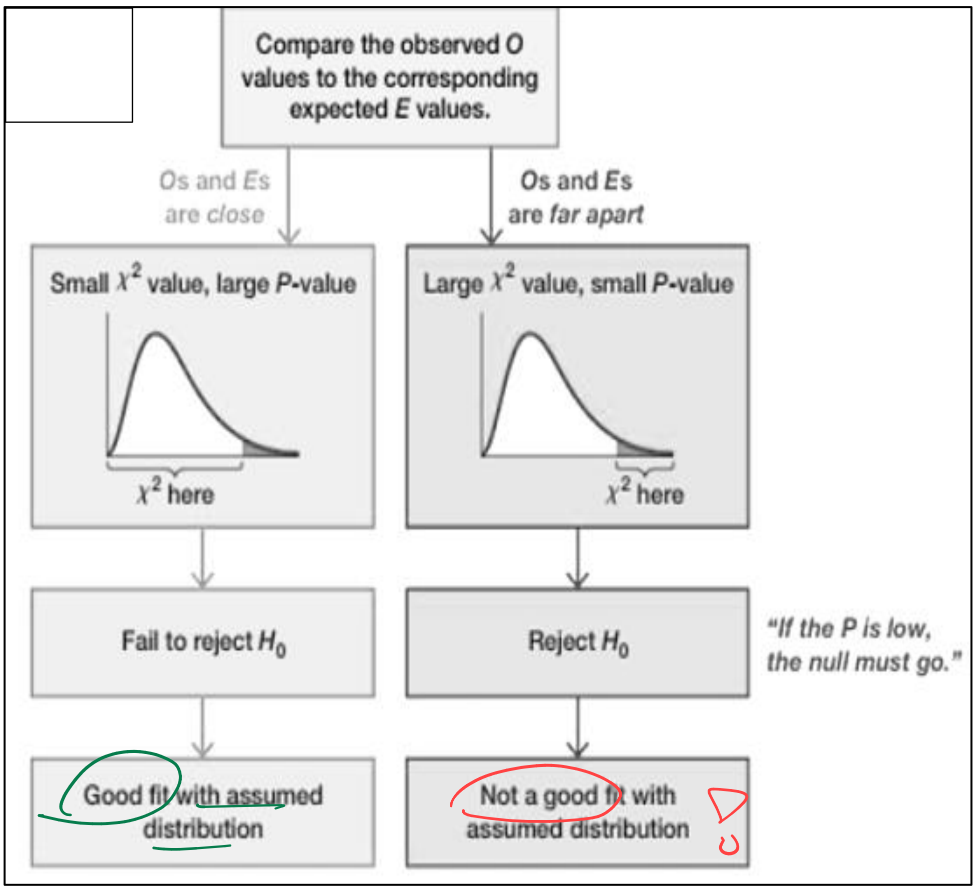
Stat > Edit > enter: L1 = O, L2 = E, L3 = (L1-L2)^2/L2, then χ_0^2 = sum(L3)

k = # of categories

Steps for Hypothesis Test for Goodness-of-Fit (Cont.)	
Step 4: *Find a Critical Value or P-Value to check either using the Critical Value Method or P-Value Method.*	
CRITICAL REGION METHOD	*P-VALUE METHOD*
* Table VIII $df = k - 1$ Reject H_0 if χ^{2*} lies in the critical region Fail to Reject H_0 if χ^{2*} doesn't lie in the critical region	$df = k - 1$ Reject H_0 if P-value $\leq \alpha$ Fail to Reject H_0 if P-value $> \alpha$
Step 5: *Make a decision and draw a conclusion.*	







GRAPHING CALCULATOR (TI-83 OR 84)

Instructions:

STAT $\Rightarrow$ TESTS $\Rightarrow$ D: $\chi^2 GOF$ -Test

$k = 6 = \#$ of colors

$L3 = \frac{(L1 - L2)^2}{L2}$

Ex 2: The M&M company claims that the distribution of plain M&M candies in a bag is 23% blue, 23% orange, 15% green, 12% red, 15% yellow, and 12% brown. Suppose we took a simple random sample of 400 M&Ms from the populations of all M&Ms. The results are shown below:

	COLOR	Blue p_1	Orange p_2	Green p_3	Red p_4	Yellow p_5	Brown p_6	
O_i:	FREQUENCY	53	66	38	96	88	59	L1
E_i:	EXPECTED	92	92	60	48	60	48	L2
		$400 \cdot 0.23$	$400 \cdot .23$	$400 \cdot .15$	$400 \cdot .12$	$400 \cdot 0.15$	$400 \cdot 0.12$	

Find $E_i = n\, p_i$

Is the proportion of each color different than the claim of the M&M's manufacturer? → HT

↳ many proportions

Null and Alternative Hypothesis actual claimed

$H_0: p_1 = 0.23,\ p_2 = 0.23,\ p_3 = 0.15,\ p_4 = 0.12,\ p_5 = 0.15,\ p_6 = 0.12$

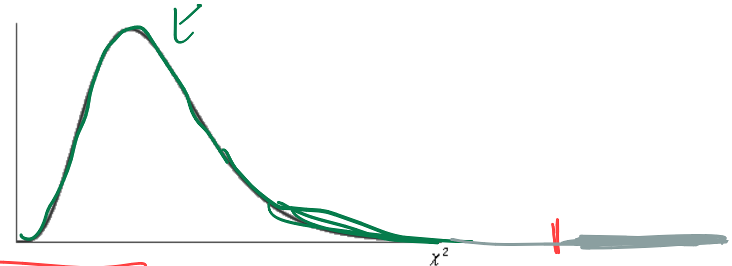
H_A: At least one proportion differs from the claimed proportions.

Test Statistic

$$\chi_0^2 = \sum \frac{(O-E)^2}{E} = \text{sum}(L3) = \boxed{95.53}$$

$L1 = O$
$L2 = E$
$L3 = (L1 - L2)^2 / L2$

1) 2nd + stats
2) Math
3) sum(



P-value/~~Critical Region~~

$$P\text{-Val} = P(\chi^2 > 95.53)$$

$$= \chi^2\text{cdf}(\text{low}, \text{high}, \underset{k-1}{df})$$

$$= \chi^2\text{cdf}(95.53, 1E99, 5) = \boxed{4.62 \times 10^{-19}} = 0.000000000000000000462 = 0+$$

Decision about Null Hypothesis

$\alpha = 0.05$

$P = 0.00\ldots0462$

$P < \alpha \rightarrow$ P low, Null go! $\boxed{\text{Reject } H_0}$

Conclusion

"There is enough statistical evidence to support the claim that at least one proportion of M&M colors differs from the manufacturer's claimed proportion!"

((Calc))

$k=5$ (# categories)

Ex 3: A company sells their products exclusively by mail. The company's management wants to find out if the number of orders received at the company's office on each of the five days of the week is the same. The company took a random sample of 400 orders received during a four-week period. The following table lists the frequency distribution for these orders by the day of the week.

n

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Orders	92	71	65	83	89
Expected Number	400·(1/5) 80	80	80	80	80

$400(1/5)$

$E = n \cdot p$

HT

Test the claim that the orders are evenly distributed over the five days of the week. Use $\alpha = .025$ LOS

Null and Alternative Hypothesis

$H_0: p_1 = p_2 = p_3 = p_4 = p_5$ "the proportion of orders is the same"

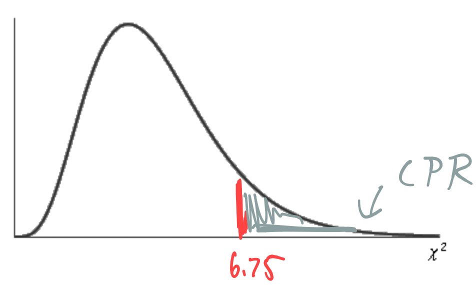
H_A: At least one is different.

Test Statistic

$\boxed{\chi_0^2 = 6.75}$

P-value/Critical Region

$\boxed{P = 0.150}$



Decision about Null Hypothesis

$\alpha = 0.025$

$P = 0.150$

$P > \alpha \longrightarrow$ P high, Null fly! $\boxed{\text{Fail to Reject } H_0}$

Conclusion

"There isn't enough statistical evidence to support the claim that at least one day of the week receives a different # of orders."