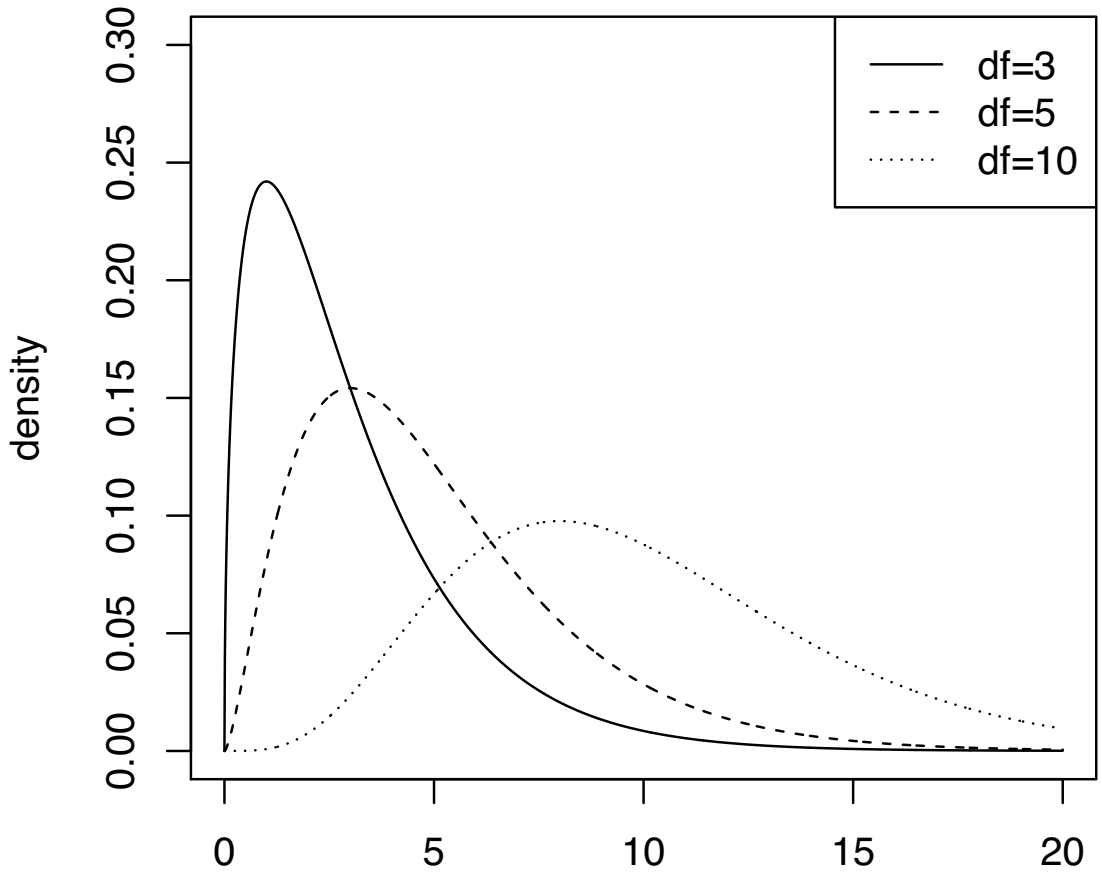
Lecture 13: Chi-Square Tests
STAT 630, Fall 2021

The χ^2 Distribution

- Let $Z \sim N(0,1)$, then Z^2 follows a chi-square distribution with 1 degree of freedom, denoted by $Z^2 \sim \chi^2_1$.
- Let $Z_1, Z_2, \cdots, Z_n$ be independent random variables each following $N(0,1)$. Let $V = \sum_{i=1}^{n} Z_i^2$. Then V follows a chi-square distribution with n degrees of freedom, denoted by $V \sim \chi^2_n$.
- Let $V \sim \chi^2_n$, then $E(V) = n$ and $Var(V) = 2n$.

Plot of χ^2_n distribution with different degrees of freedom n:



$V = Z_1^2 + Z_2^2 + \cdots + Z_{10}^2$

Example: Let $Z_1, Z_2, \cdots, Z_{10}$ be independent random variables from $N(0,1)$. Let $V = \sum_{i=1}^{10} Z_i^2$.

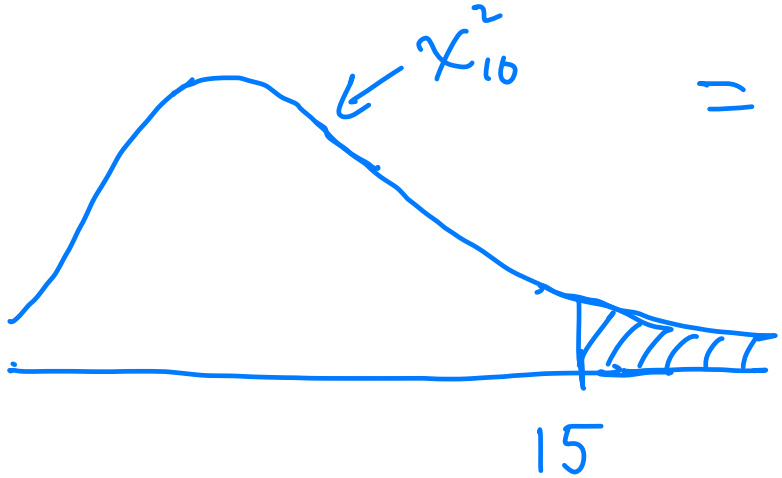
(a) What distribution does V follow? What is $E(V)$ and $Var(V)$?

$V \sim \chi^2_{10}$

$E(V) = n = 10$

$Var(V) = 2n = 20$

(b) Calculate $P(V > 15)$ $= 1 - P(V < 15)$



$= 1 - pchisq(15, df = 10)$

$= 0.132$

Chi-Square Test of Independence

A **chi-square test of independence** is used to evaluate whether or not two categorical variables, expressed as a contingency table, are independent of each other.

Example: The General Social Survey (GSS) surveyed a random sample of Americans in 2002. As part of this survey each participant was asked whether they favor or oppose the death penatly. The suvey also collected data on the education level of each participant. Does the survey provide evidence of a relationship between education level and postion on the death penalty?

H_0: Education and position on death penalty are independent
H_A: Education and position on death penalty are not independent

Contingency table of the data (table of observed counts):

	Death Penalty?		
	Favor	Oppose	Row Total
Left HS	117	72	189
HS	511	200	711
Jr Col	71	16	87
Bachelors	135	71	206
Graduate	64	50	114
Column Total	898	409	1307

Assuming that the null hypothesis is true, and the two categorical variables are independent, we can construct a table of the **expected counts**.

- n: overall total number of observations
- R_i: total for row i
- C_j: total for column j

Then the expected count in row i and column j is given by

$$E_{ij} = \frac{R_i \cdot C_j}{n}$$

Notes:

If events A and B are indep, then

$$P(A \text{ and } B) = P(A)\,P(B)$$

For example, assuming indep

$$P(\text{HS and Oppose}) = P(\text{HS})\,P(\text{Oppose})$$
$$= \frac{711}{1307} \cdot \frac{409}{1307}$$

So the expected count is given by

$$1307 \cdot \frac{711}{1307} \cdot \frac{409}{1307} = \frac{711 \cdot 409}{1307} = 222.5$$

Table of expected counts:

	Death Penalty?		
	Favor	Oppose	Row Total
Left HS	129.9	59.1	189
HS	488.5	222.5	711
Jr Col	59.8	27.2	87
Bachelors	141.5	64.5	206
Graduate	78.3	35.7	114
Column Total	898	409	1307

$$E_{21} = \frac{R_2 \cdot C_1}{n} = \frac{711 \cdot 898}{1307} = 488.5$$

$$E_{22} = \frac{R_2 \cdot C_2}{n} = \frac{711 \cdot 409}{1307} = 222.5$$

Chi-square test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$
$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed count, I is the number of rows, and J is the number of columns.

The degrees of freedom is $(I - 1)(J - 1)$.

For the example, the value of test statistic is

$$\begin{aligned}\chi^2 &= \frac{(117-129.9)^2}{129.9} + \frac{(72-59.1)^2}{59.1} + \frac{(511-488.5)^2}{488.5} + \frac{(200-222.5)^2}{222.5} \\ &+ \frac{(71-59.8)^2}{59.8} + \frac{(16-27.2)^2}{27.2} + \frac{(135-141.5)^2}{141.5} + \frac{(71-64.5)^2}{64.5} \\ &+ \frac{(64-78.3)^2}{78.3} + \frac{(50-35.7)^2}{35.7} \\ &= \boxed{23.45}\end{aligned}$$

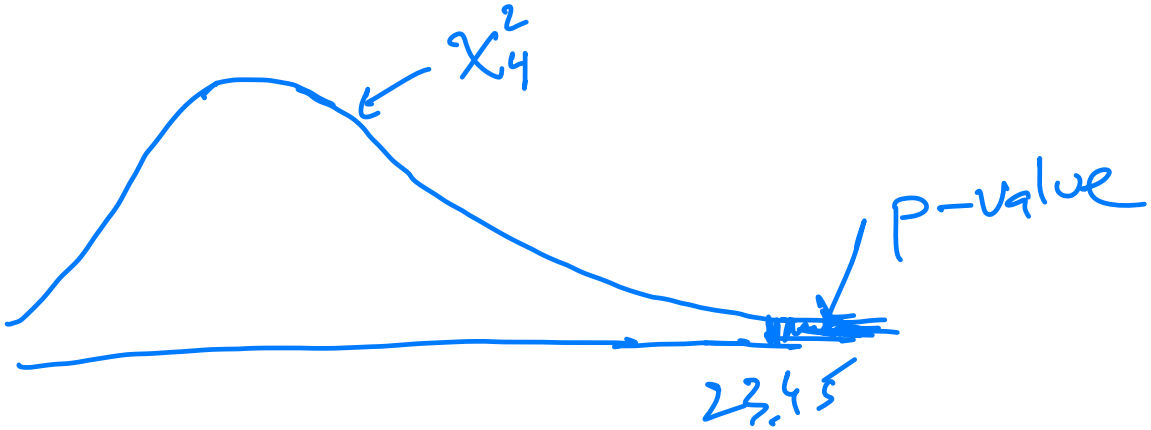
$\chi^2 = 23.45$

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

Next, we calculate the p-value using the chi-square distribution. Note the degrees of freedom is given by $(5-1)(2-1) = 4$.

$$p\text{-value} = 1 - \texttt{pchisq(23.45, df=4)} = 0.0001029$$

Since the p-value < 0.05 we reject H_0. The survey provides convincing evidence that education level and position on the death penalty are not independent.



Conditions for the chi-square test of independence:

- There are at least 5 expected counts in each cell, i.e., $E_{ij} \geq 5$.
- Independence / random sampling (so the results can be generalized to the population).

For our example, the conditions are satisfied.

Chi-Square Test of Goodness-of-Fit

A **chi-square test of goodness-of-fit** is used to evaluate whether data follow a particular probability distribution.

Example: Roll a die $n = 60$ times and count the number of times it lands on each side $j = 1, 2, \cdots, 6$.

$H_0 : p_1 = p_2 = \cdots = p_6 = 1/6$ (die is fair)
$H_A : p_j \neq 1/6$ for at least one j (die is weighted)

where p_j is the probability the die lands on side j

Table of observed and expected counts:

j	1	2	3	4	5	6
O_j	20	11	6	7	6	10
E_j	10	10	10	10	10	10
$(O_j - E_j)^2/E_j$	10	0.1	1.6	0.9	1.6	0

- O_j is the observed count for side j
- $E_j = 60 * (1/6) = 10$ is the expected count for side j

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(20-10)^2}{10} = 10$$

Chi-square test statistic:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

$$= 10 + 0.1 + 1.6 + 0.9 + 1.6 + 0 = \boxed{14.2}$$

The degrees of freedom for a chi-square test of goodness-of-fit is $k - 1 = 6 - 1 = 5$, i.e., the number of categories minus 1.

Next, we compute the p-value:

$$p\text{-value} = P(\chi^2 > 14.2) = 1 - \texttt{pchisq(14.2, df=5)} = 0.0144$$

Since the p-value < 0.05, we reject H_0 and conclude that at least one $p_j \neq 1/6$, and so the die is weighted.

In R:

```
> chisq.test(c(20, 11, 6, 7, 6, 10), p=rep(1/6,6))

Chi-squared test for given probabilities

data:  c(20, 11, 6, 7, 6, 10)
X-squared = 14.2, df = 5, p-value = 0.01439
```

Conditions for the chi-square test of goodness-of-fit:

- There are at least 5 expected counts in each cell; i.e., $E_j \geq 5$.
- Independence / random sampling (so the results can be generalized to the population).

The conditions are satisfied for the dice rolling example.

Example: Consider data from a random sample of 275 jurors in a small county. Jurors identified their racial background, as shown in the table below. We would like to determine if these jurors are racially representative of the population of registered voters in this county.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

H_0 : There is no racial bias in juror selection.
H_A : There is racial bias in juror selection.

Can also write hypotheses symbolically:

$H_0: p_1 = 0.72,\ p_2 = 0.07,\ p_3 = 0.12,\ p_4 = 0.09$

$H_A: p_1 \neq 0.72$ or $p_2 \neq 0.07$ or $p_3 \neq 0.12$ or $p_4 \neq 0.09$

Table of observed and expected counts:

Race (j)	White	Black	Hispanic	Other
O_j	205	26	25	19
E_j	198	19.25	33	24.75
$(O_j - E_j)^2/E_j$	0.247	2.367	1.939	1.336

The chi-square test statistic:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} = 0.247 + 2.367 + 1.939 + 1.336 = \boxed{5.889}$$

The degrees of freedom is $k - 1 = 4 - 1 = 3$.

$E_1 = 275(0.72) = 198$

$E_2 = 275(0.07) = 19.25$

Next, we compute the p-value:

$$p\text{-value} = P(\chi^2 > 5.889) = 1 - \texttt{pchisq(5.889, df=3)} = 0.117$$

Since the p-value > 0.05, we do not reject H_0. Therefore, the data do not provide convincing evidence of racial bias in juror selection.

Last, the conditions for the test are satisfied since the expected counts are at least 5 for each race category, i.e., $E_j \geq 5$.

In R:

```
> chisq.test(c(205, 26, 25, 19), p=c(0.72, 0.07, 0.12, 0.09))

Chi-squared test for given probabilities

data:  c(205, 26, 25, 19)
X-squared = 5.8896, df = 3, p-value = 0.1171
```