Lecture 15

Chi-Squared Tests

agenda

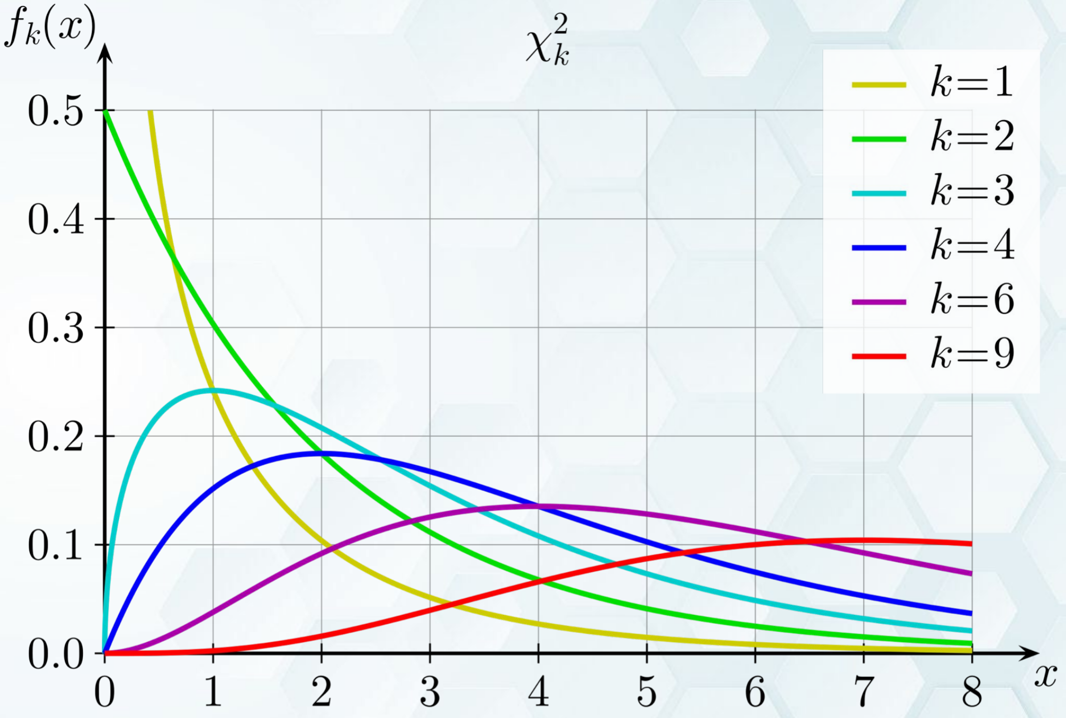
Reminders	Midterm 1 revisions (read the directions!); HW 5 due Thursday
lecture part 1	chi-squared tests for goodness of fit
lecture part 2	chi-squared tests for independence
R activity	hypothesis testing for goodness of fit and independence (`ht_chi_squared.Rmd`)

$Z \sim N(0,1) \quad Z^2 \sim \chi^2_1$

Chi-squared distribution

the chi-squared (χ^2) distribution with k degrees of freedom is the sum of squares of k independent standard normal random variables, i.e.,

$$\chi_k^2 = \sum_{i=1}^{k} Z_i^2$$



Chi-squared distribution

If $V = \sum_{i=1}^{k} Z_i^2$ then,

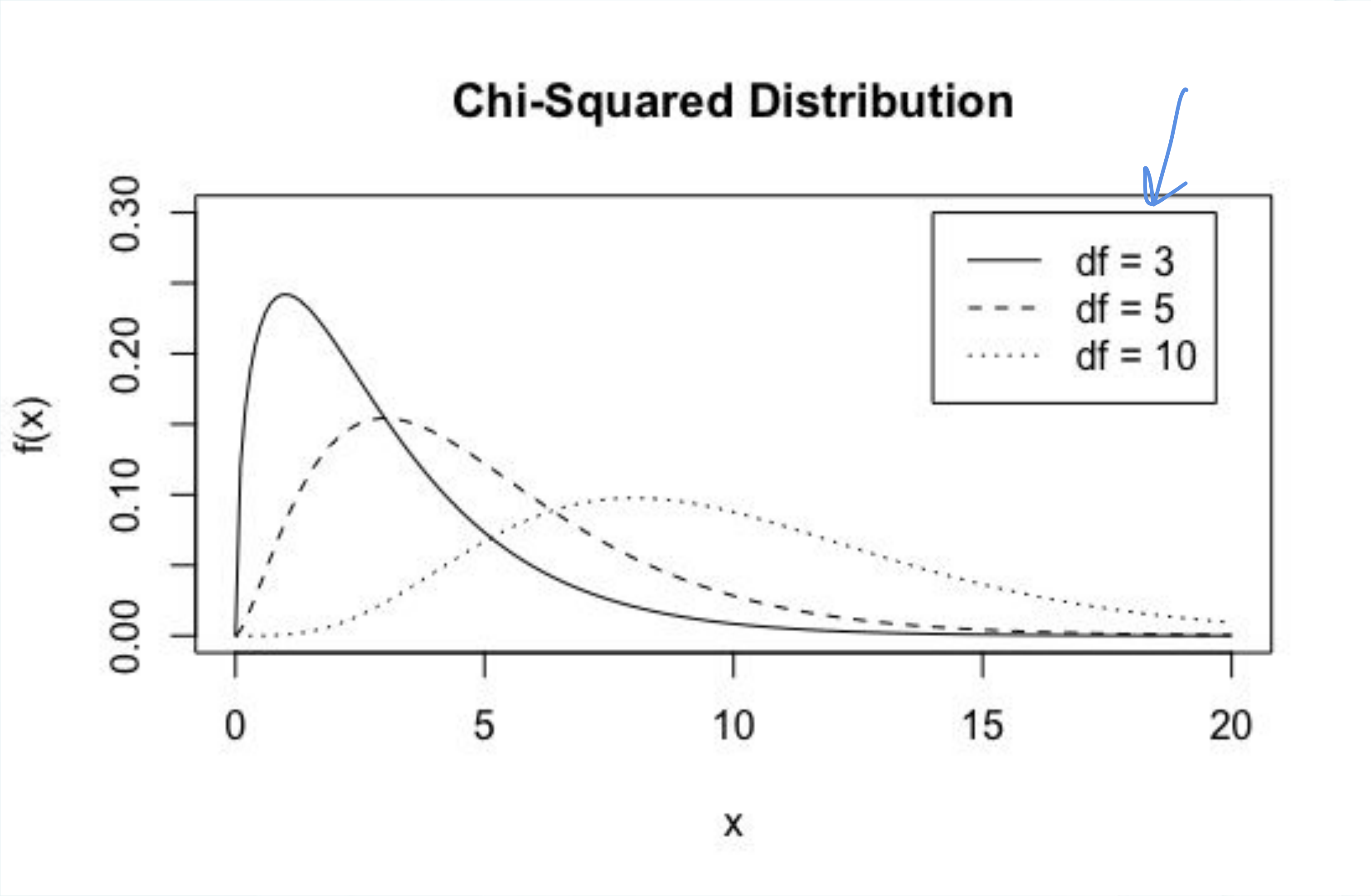
- $E[V] = k$
- $Var[V] = 2k$

$SD[V] = \sqrt{2K}$

To plot a χ^2- distribution, use the code:

```
x <- seq(0, 20, 0.1)
chi3 <- dchisq(x, df = 3)
chi5 <- dchisq(x, df = 5)
chi10 <- dchisq(x, df = 10)

plot(x, chi3, type = "l", ylim = c(0, 0.3))
lines(x, chi5, lty = 2)
lines(x, chi10, lty = 3)
```


Chi-Squared Distribution
f(x)
0.00
0.10
0.20
0.30
0
5
10
15
20
x
df = 3
df = 5
df = 10

Chi-squared probabilities

Let $V = \sum_{i=1}^{k} Z_i^2$, where k = 13 $Z \sim N(0,1)$, $V \sim \chi^2_{13}$

1) Find P(V > 10)

pchisq(10, 13, lower.tail= FALSE)

2) Find P(V $\geq$ 15) = P(V > 15)

pchisq(15, 13, lower.tail =FALSE)

Chi-squared test for Goodness of Fit

categorical var
w/ K > 2 categories

1) Hypotheses

- H_0: the data follow a particular probability distribution
- H_A: the data do not follow the H_0 dist.

2) Test conditions:
   - Independence
   - Expected counts

3) Calculate test statistic

$$\chi^2 = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}$$

4) Compute p-value

```
> pchisq(test_stat, df,
         lower.tail = FALSE)
```

*where df = # groups - 1

5) Decision and conclusion in context

Consider a standard package of milk chocolate M&Ms. There are six different colors: red, orange, yellow, green, blue and brown. Suppose that we are curious about the distribution of these colors and ask, do all six colors occur in equal proportion?

Suppose that we have a simple random sample of 600 M&M candies with the following distribution:

- 212 of the candies are blue.
- 147 of the candies are orange.
- 103 of the candies are green.
- 50 of the candies are red.
- 46 of the candies are yellow.
- 42 of the candies are brown.

Hypothesis Testing

null

H_0: proportion of all 6 m&m colors is the same

$H_0: P_{BL} = P_R = P_O = P_G = P_{BR} = P_Y = \frac{1}{6}$

alternative

H_A: at least one of the proportions is different

H_A: H_0 not true

Check Conditions

Independence

observational units are randomly selected

random sample ✓

Expected counts

expected counts are all greater than 5

$$np_R = 600 \times \frac{1}{6} = 100$$
$$np_O = 100$$
$$np_Y = 100$$
$$np_G = 100$$
$$np_{BL} = 100$$
$$np_{BR} = 100$$

≥ 5 ✓

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed - Expected})^2}{Expected}$$

$$= \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(50-100)^2}{100} + \frac{(147-100)^2}{100} + \frac{(46-100)^2}{100} + \dots$$

$$= \boxed{235.42}$$

$$\chi^2 \sim \chi^2_5$$

Calculate test statistic

$$\chi^2 = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}$$

Calculate test statistic

Compute p-value

pchisq(235.42, 5, lower.tail = FALSE)

≈ 0

Conclusion in Context

Reject the null: We have enough evidence that distribution of _____ differs from the distribution specified in the null.

Fail to reject the null: We do not have enough evidence that distribution of _____ differs from the distribution specified in the null.

We have enough evidence that the proportions of m&m colors are not all equal.

Break time!

Two-way Tables

Var 2

	j = 1	j = 2	...	j = J	Total
i = 1	n_{11}	n_{12}	...	n_{1J}	$n_{1.}$
i = 2	n_{21}	n_{22}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...
i = I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{J.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	n

var 1

n_{1+}

n_{+1}

Chi-squared test for Independence

1) Hypotheses

- H_0: var1 and var2 are independent
- H_A: var1 and var2 are not independent

associated

2) Test conditions:
   - Independence
   - Expected counts

3) Calculate test statistic

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

4) Compute p-value

```
> pchisq(test_stat, df,
             lower.tail = FALSE)
```

*where df = (# rows - 1) x (# cols - 1)

$df = (I - 1) \times (J - 1)$

5) Decision and conclusion in context

Independence

Suppose 2 events: A, B

A and B are indep. if

$$P(A \cap B) = P(A) \times P(B)$$

H_0: $p_{ij} = p_{i\cdot} \times p_{\cdot j} \quad \forall i, j$

H_A: H_0 not true

Suppose 395 people are randomly selected, and are "cross-classified" into one of eight cells, depending into which age category they fall and whether or not they support legalizing marijuana.

H_0: mj & age group indep.

Are marijuana support and age group independent?

H_A: mj & age group not indep. / assoc.

Age Groups

Marijuana Support

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

Check Conditions

Independence

observational units are randomly selected

random sample ✓

Expected counts

expected counts are all greater than 5

	j = 1	j = 2	…	j = J	Total
i = 1	n_{11}	n_{12}	…	n_{1J}	$n_{1.}$
i = 2	n_{21}	n_{22}	…	n_{2J}	$n_{2.}$
…	…	…	…	…	…
i = I	n_{I1}	n_{I2}	…	n_{IJ}	$n_{J.}$
Total	$n_{.1}$	$n_{.2}$	…	$n_{.J}$	n

Expected counts

$$E_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

Observed Counts

	18-24	25-34	35-49	50-64	Total
Yes	60	54	46	41	201
No	40	44	53	57	194
Total	100	98	99	98	395

50.89

49.87

Expected Counts

	18-24	25-34	35-49	50-64	Total
Yes	$\frac{100 \times 201}{395}$	$\frac{98 \times 201}{395}$	$\frac{99 \times 201}{395}$	$\frac{98 \times 201}{395}$	201
No	$\frac{100 \times 194}{395}$	$\frac{98 \times 194}{395}$	$\frac{99 \times 194}{395}$	$\frac{98 \times 194}{395}$	194
Total	100	98	99	98	395

Calculate expected counts

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed - Expected})^2}{Expected}$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(60-50.89)^2}{50.89} + \frac{(54-49.87)^2}{49.87} + \cdots$$

Calculate test statistic

Compute p-value

Conclusion in context

Reject the null: We have enough evidence that var1 and var2 are not independent.

Fail to reject the null: We do not have enough evidence that var1 and var2 are not independent.