chi-square GOF test

- "goodness of fit"
- one categorical variable, >2 levels

Dr. Mine Çetinkaya-Rundel
Duke University

Duke
UNIVERSITY

jury selection

- In a county where jury selection is supposed to be random, a civil rights group sues the county, claiming racial disparities in jury selection.
- Distribution of ethnicities of the people in the county who are eligible for jury duty (based on census results):

ethnicity	white	black	nat. amer.	asian & PI	other
%in population	80.29%	12.06%	0.79%	2.92%	3.94%

- Distribution of 2500 people who were selected for jury duty the previous year:

ethnicity	white	black	nat. amer.	asian & PI	other
observed #	1920	347	19	84	130

Image source: http://en.wikipedia.org/wiki/File:Trial_by_Jury_Usher.jpg

jury selection

The court retains you as an independent expert to assess the statistical evidence that there was discrimination. You propose to formulate the issue as an hypothesis test.

H_0 (nothing going on): People selected for jury duty are a simple random sample from the population of potential jurors. The observed counts of jurors from various race/ethnicities follow the same ethnicity distribution in the population.

H_A (something going on): People selected for jury duty are not a simple random sample from the population of potential jurors. The observed counts of jurors from various ethnicities do not follow the same race/ethnicity distribution in the population.

evaluating the hypotheses

- quantify how different the observed counts are from the expected counts
- large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- called a goodness of fit test since we're evaluating how well the observed data **fit** the expected distribution

practice

Calculate expected number of jurors from each ethnicity if in fact the jury selection is random.

ethnicity	white	black	nat. amer.	asian & PI	other	total
%in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007 +	302 +	20 +	73 +	98	? = 2500 ✔

2500×0.8029

2500×0.1206

ethnicity	white	black	nat. amer.	asian & PI	other	total
%in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	2500

observed < expected

observed > expected

Conditions for the chi-square test:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
   - each case only contributes to one cell in the table
2. ***Sample size:*** Each particular scenario (i.e. cell) must have at least 5 expected cases.

anatomy of a test statistic

general form of a test statistic

$$\frac{point\ estimate - null\ value}{SE\ of\ point\ estimate}$$

1. identifying the difference between a point estimate and an expected value if the null hypothesis were true
2. standardizing that difference using the standard error of the point estimate

chi-square statistic

when dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the chi-square (χ^2) statistic.

χ^2 **statistic:**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E}$$

O : observed

E : expected

k : number of cells

why square?

- positive standardized difference
- highly unusual differences between observed and expected will appear even more unusual

degrees of freedom

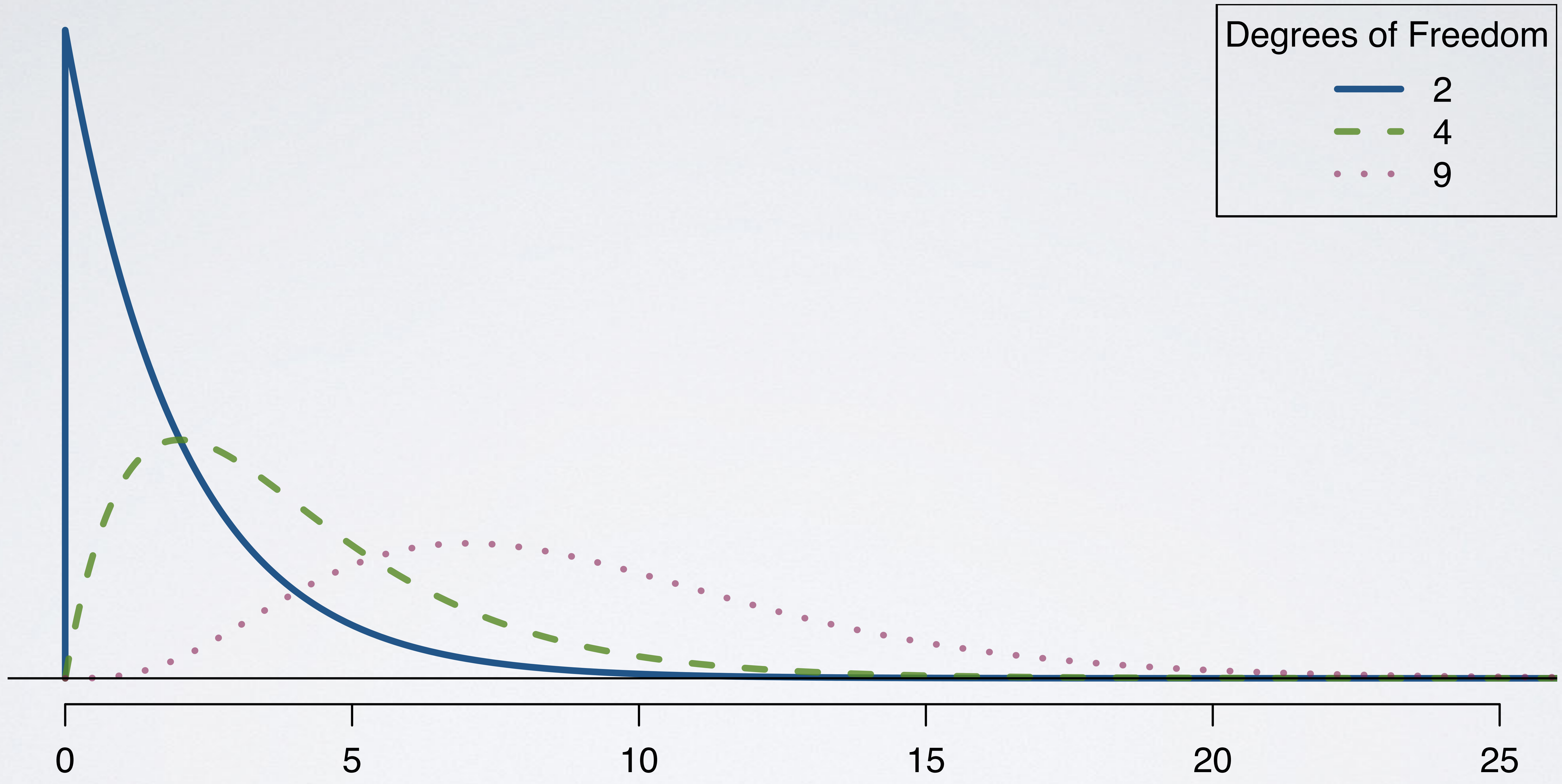
- to determine if the calculated χ^2 statistic is considered unusually high or not we need to first describe its distribution
- chi-square distribution has just one parameter:
  - degrees of freedom (df): influences the shape, center, and spread

χ^2 degrees of freedom for a goodness of fit test:

$$df = k - 1$$

k : number of cells

chi-square distribution & degrees of freedom



practice

putting it all together…

ethnicity	white	black	nat. amer.	asian & PI	other	total
%in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007	302	20	73	98	2500
observed #	1920	347	19	84	130	2500

H_0: The observed counts of jurors from various race/ethnicities follow the same ethnicity distribution in the population.

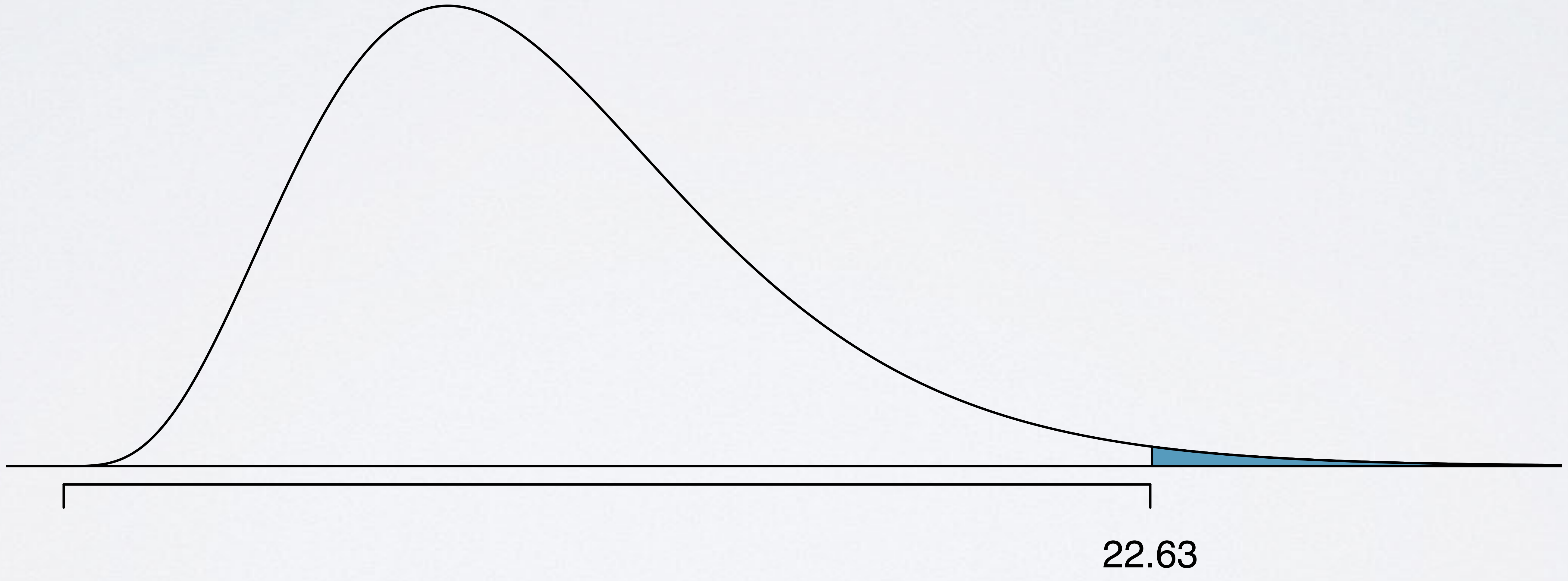
H_A: The observed counts of jurors from various ethnicities do not follow the same race/ethnicity distribution in the population.

$$\chi^2 = \frac{(1920 - 2007)^2}{2007} + \frac{(347 - 302)^2}{302} + \frac{(19 - 20)^2}{20} + \frac{(84 - 73)^2}{73} + \frac{(130 - 98)^2}{98} = 22.63$$

$$df = k - 1 = 5 - 1 = 4$$

p-value

- p-value for a chi-square test is defined as the tail area above the calculated test statistic
- because the test statistic is always positive, and a higher test statistic means a higher deviation from the null hypothesis



p-value

$\chi^2 = 22.63$ $df = 4$

using R

R

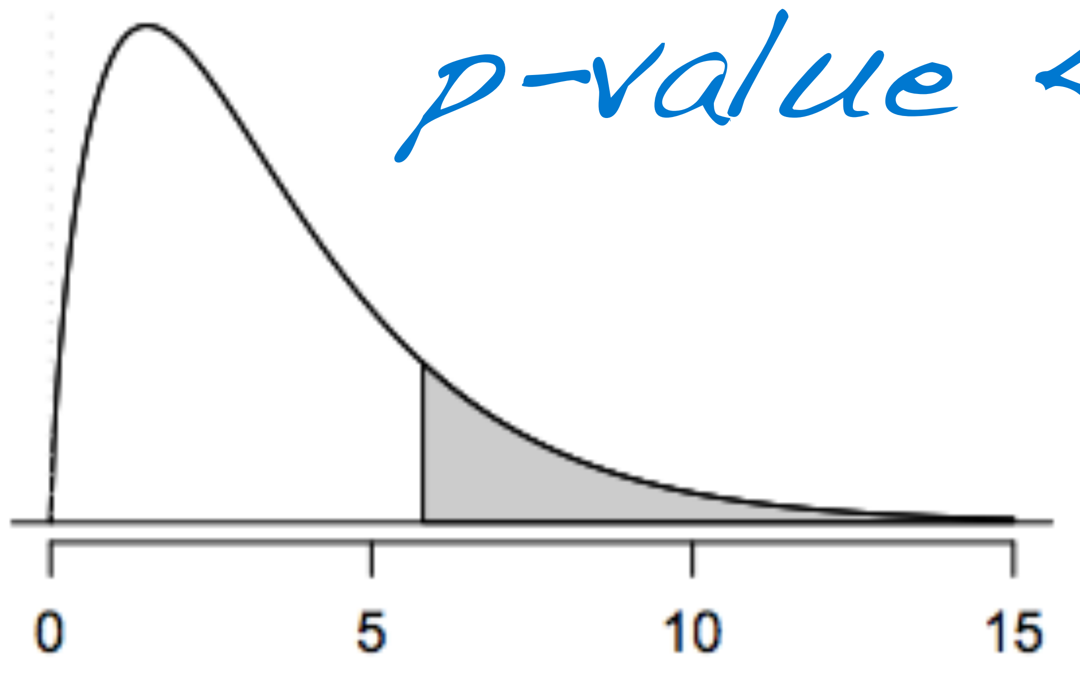
```
> pchisq(22.63, 4, lower.tail = FALSE)
[1] 0.0002
```

using the applet

http://bitly.com/dist_calc

using the table

Chi-square probability table



p-value < 0.001

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32