Statistical Methods and Data Analysis I
Lecture 20: Comparing Distributions.

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

June 11, 2019

Are Two Distributions Different?

- Generalize the question from "is the distribution consistent with normal?" to "are the 2 samples drawn from the same distribution function?"
- More formally:
  - H_0: the two data sets are drawn from the same population (with a certain distribution function)
  - At a given confidence level, can we reject H_0?
- *Reminder*: one can never *prove* H_0: no practical amount of data can distinguish between two distributions that differ by one part in 10^{10} (or something...).
- Examples:
  1. Are the visible stars distributed uniformly in the sky?
  2. Are "educational patterns" (last grade attended) the same in Brooklyn and Bronx?
  3. Do 2 brands of lights have the same distribution of burn-out times?
  4. is the incidence of measles the same for first-born, second-born, etc?
- Continuous (1, 3) or binned/categorical (2, 4) data
- Either we compare one data set to a known (according to H_0) distribution (1, 4) or we wish to compare 2 data sets for which distributions are unknown (2, 3).

Comparing Distributions of Categorical/Binned Variables: χ^2 Test

- n observations in a random sample are classified into k mutually exclusive categories (or bins): x_i for $i = 1, \ldots, k$.
- H_0: the probability p_i that an observation belongs to category i (known distribution).
  - We have *expected numbers* $m_i = np_i$, where

$$\sum_{i=1}^{k} p_i = 1, \quad \sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = \sum_{i=1}^{k} x_i$$

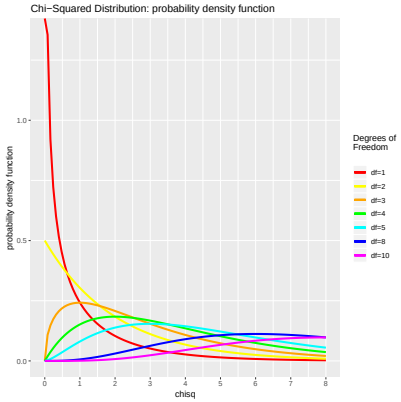
- Test statistics:

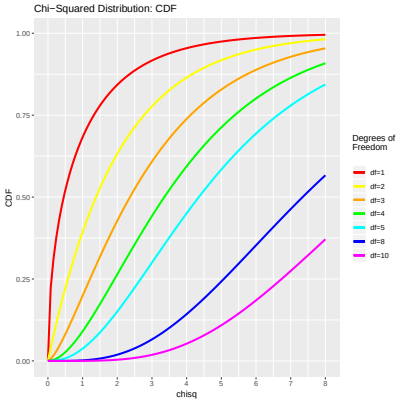
$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$$

  - Any term j with $0 = x_j = m_j$ should be omitted.
  - Any term j with $m_j = 0$ and $x_j \neq 0$ leads to *infinite* χ^2, as it should!
- A large value of χ^2 indicates that H_0 is unlikely (corresponding to small p-value)

Distribution of χ^2

- Probability that the sum of squares of k *standard normal* ($N(0,1)$) random variables is distributed according to χ^2 *distribution with k degrees of freedom*.
  - The terms in χ^2 are **not** individually normal.
  - If $k \gg 1$ or $x_i \gg 1$ then it is a good approximation.





χ^2 Distribution in R

- You don't need to be frightened any more than by t or F distributions
- R has the usual collection of function for the χ^2 distribution,too:

```
dchisq(x, df) - probability density function (PDF)
pchisq(q, df) - cumulative distribution function (CDF)
qchisq(p, df) - quantile function
rchisq(n, df) - random generator
```

- *Notes:*
  - Degree of freedom (`df`) is typically integer, but doesn't have to be.
  - There is `pchisq(q,df,lower.tail=FALSE)` for the tail distribution function (the default is `lower.tail=TRUE`).
  - There are other arguments that may be important, but we will ignore them for now.

χ^2 Test

Scenario: multiple choice exams

IDC provides a bank of past exams and says that in multiple choice questions the correct answer may be A, B, C, or D *with equal probability*. Can we test this?

- Choose 100 multiple choice questions at random.
- H_0: probabilities that A,B, C, or D are the correct answers are equal, each is 25%.

Answer	Expected	Observed
A	25	20
B	25	20
C	25	25
D	25	35

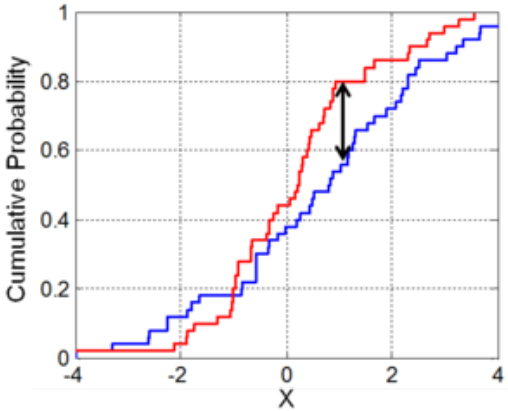
```
> chisq.test(c(20,20,25,35))

Chi-squared test for given probabilities

data:  c(20, 20, 25, 35)
X-squared = 6, df = 3, p-value = 0.1116
```

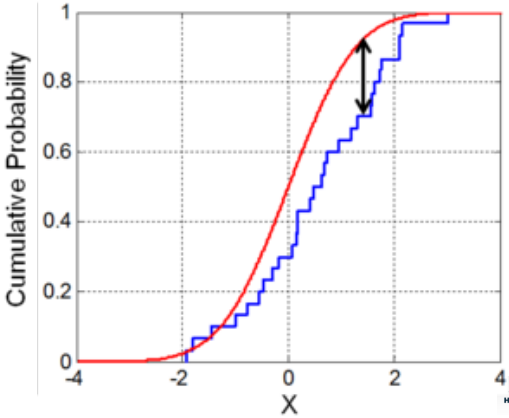
Comparing Distributions of Continuous Variables

- Suppose we have a sample of observations:
  - The list of data points can easily be converted into an empirical estimator of the *cumulative distribution function (empirical CDF)*.
  - Given n observations x_i ($i = 1, \ldots, n$) we just look at the fraction of data points to the left of x_i.
  - The estimator function is *constant* between x_i and x_{i+1} (assuming x_i are sorted) and *jumps by* $1/n$ at each x_i.
  - *The larger the sample, the closer the estimator is to the real CDF.*
- For two samples we can measure the difference between the respective estimators of CDF and decide whether the CDFs are close to eachother, thus whether the samples are drawn from the same distribution.
- Measuring the difference between the estimators:
  - compute the area between the CDFs
  - compute the mean square difference
  - *Kolmogorov-Smirnov D: the maximum value of the absolute difference between the two (estimated) CDFs — a particularly simple measure.*

Kolmogorov-Smirnov Distance

- The procedure outlined in the previous slide works for comparing 2 samples or comparing 1 sample to a known distribution.

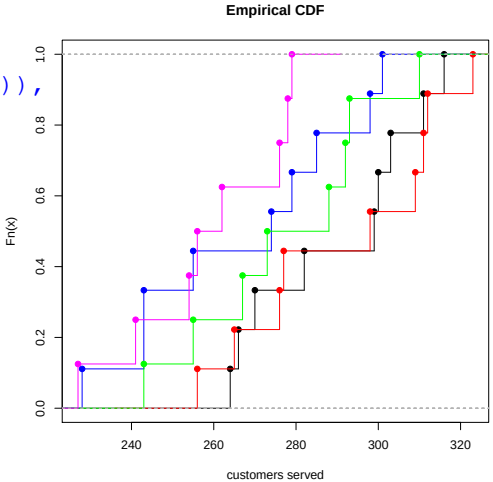




R Toolbox: Empirical CDF — `ecdf()`

- With the data collected by the data analyst at the customer service center (cf. the previous lecture):

```
> plot(ecdf(Sun),
       xlim=range(c(Sun,Mon,Tue,Wed,Thu)),
       verticals=TRUE,col="black",
       xlab="customers served",
       main="Empirical CDF")
> plot(ecdf(Mon),add=TRUE,
       verticals=TRUE,col="red")
> plot(ecdf(Tue),add=TRUE,
       verticals=TRUE,col="blue")
> plot(ecdf(Wed),add=TRUE,
       verticals=TRUE,col="green")
> plot(ecdf(Thu),add=TRUE,
       verticals=TRUE,col="cyan")
```



Kolmogorov-Smirnov Test

- H_0: the 2 distributions are the same (for 2 samples, or the single sample's distribution is the same as the theoretical one).
- What makes K-S statistic D useful is that *its distribution under H_0 can be calculated to a useful approximation*.
- *p-value*: probability to get D as large or larger than observed under H_0.

```
> ks.test(Sun,Mon)

data:  Sun and Mon
D = 0.22222, p-
value = 0.9794
alternative hypothesis:
    two-sided
```

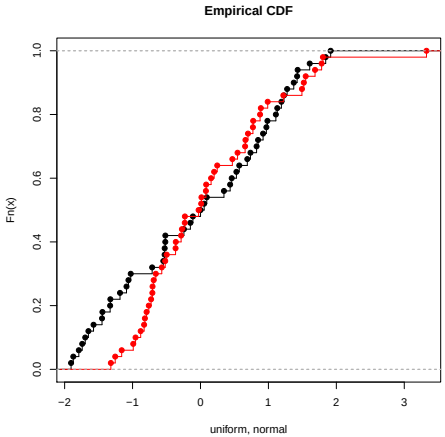
```
> ks.test(Sun,Thu)

data:  Sun and Thu
D = 0.66667, p-
value = 0.02024
alternative hypothesis:
    two-sided
```

```
> ks.test(Sun,Tue,
    alternative="less")

data:  Sun and Tue
D^- = 0.44444, p-
value = 0.169
alternative hypothesis:
    the CDF of x lies
    below that of y
```

Kolmogorov-Smirnov Test: Another Example

```
> set.seed(223344)
> y <- runif(50,-2,2)
> mean(y)
[1] -0.04003613
> sd(y)
[1] 1.15041
> x <- rnorm(50)
> ks.test(x,y)

Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.24, p-value = 0.1124
alternative hypothesis: two-sided
```

Try it with $n = 500$?



Can you tell which one is uniform and which one is normal?