Statistical Methods and Data Analysis I
Lecture 16: Are Variances Different? F-Test.

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

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

May 14, 2019

Are Variances of Two Samples Different?

- Assume we have two (or more) conceptually similar samples to analyze. It may be production numbers from different factories, sales figures from different sales teams, number of bugs in different versions of software, price history of financial instruments...
- *Another very common question*: is one sample more "stable" / "reliable" / "predictable" / "risky" than another?
- Typical assumption: the samples in question are drawn from *normally* distributed populations
  - *The assumption is very significant, and there are ways to relax it, but we will* **not** *study violations!*
  - **Important**: test for normality before proceeding — we **will** study how to do that!
- So, we have X_i for $i = 1, \ldots, n$, Y_j for $j = 1, \ldots, m$
- We can compute the samples' means and variances:

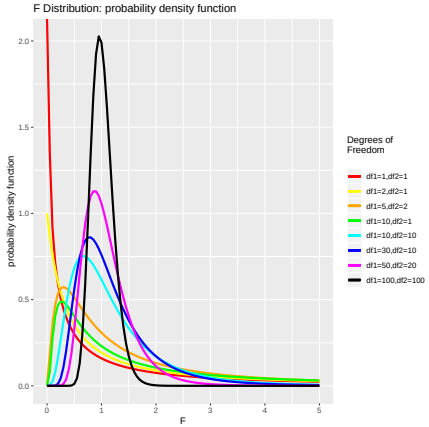
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j, \quad \sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad \sigma_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

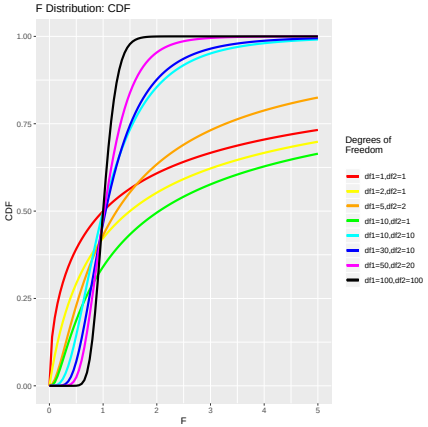
F-Test for Significantly Different Variances

- H_0: the variances of 2 samples are similar
- H_A: the variances are different

- Test statistics: $F = \frac{\sigma_1^2}{\sigma_2^2}$

- We reject H_0 if $F \ll 1$ or $F \gg 1$





F Distribution in R

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

```
df(x, df1, df2) - probability density function (PDF)
pf(q, df1, df2) - cumulative distribution function (CDF)
qf(p, df1, df2) - quantile function
rf(n, df1, df2) - random generator
```

- *Notes:*
  - There are two d.o.f. numbers, typically integers, but they don't have to be.
  - There is `pf(q,df1,df2,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.