Lecture 29 Compare Two Populations - Variance

BIO210 Biostatistics

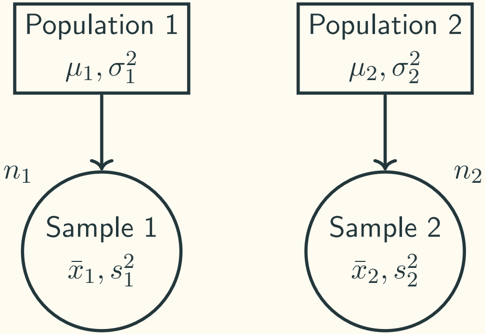
Xi Chen

Spring, 2023

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Comparing Two Variances



If Populations 1 & 2 follow normal distributions:

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \boldsymbol{\chi^2(n_1-1)} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \boldsymbol{\chi^2(n_2-1)}$$

Is σ_1^2 equal to σ_2^2 ?

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

$$\Leftrightarrow \begin{cases} H_0: & \delta = \sigma_1^2 - \sigma_2^2 = 0 \\ H_1: & \delta = \sigma_1^2 - \sigma_2^2 \neq 0 \end{cases}$$

$\boldsymbol{D = S_1^2 - S_2^2} \sim ?$ not so useful

$\mathcal{F}$-distributions

Definition

If we let $U_1 = \dfrac{(n-1)S_1^2}{\sigma_1^2}$ and $U_2 = \dfrac{(n-1)S_2^2}{\sigma_2^2}$, a more useful random variable is:

$$\boldsymbol{F} = \frac{U_1/\nu_1}{U_2/\nu_2} = \frac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2}\Big/(n_1-1)}{\dfrac{(n_2-1)S_2^2}{\sigma_2^2}\Big/(n_2-1)} = \frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \sim \boldsymbol{\mathcal{F}}(\nu_1,\nu_2)$$

$$f_{\boldsymbol{X}}(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}x+1\right)^{\frac{\nu_1+\nu_2}{2}}}$$

$\mathcal{F}$-distributions And F Scores

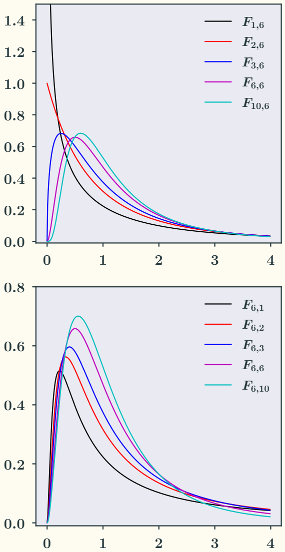
$$\boldsymbol{F} = \frac{U_1/\nu_1}{U_2/\nu_2} = \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim \boldsymbol{\mathcal{F}}(\nu_1, \nu_2)$$

We want to test the hypothesis of $\sigma_1^2 = \sigma_2^2$. Therefore, the situation is:

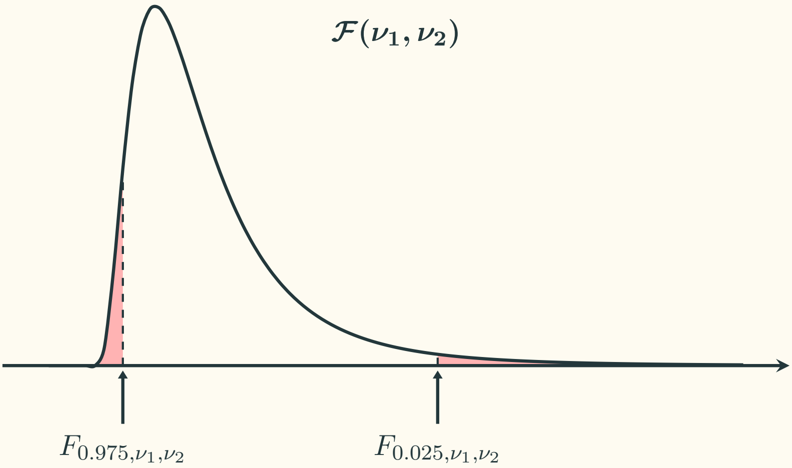
$$\begin{cases} H_0 : & \sigma_1^2 = \sigma_2^2 \\ H_1 : & \sigma_1^2 \neq \sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0 : & \dfrac{\sigma_2^2}{\sigma_1^2} = 1 \\ H_1 : & \dfrac{\sigma_2^2}{\sigma_1^2} \neq 1 \end{cases}$$

If H_0 were true, we compute the test statistic (the $\boldsymbol{F}$ score):

$$\boldsymbol{F} = \frac{U_1/\nu_1}{U_2/\nu_2} = \frac{S_1^2}{S_2^2} \sim \boldsymbol{\mathcal{F}}(\nu_1, \nu_2)$$



The $\mathcal{F}$-distribution Rejection Regions



The F-test Example

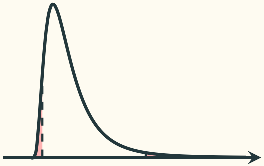
Driving: A psychologist was interested in exploring whether or not male and female college students have different driving behaviours. The particular statistical question she framed was as follows: "Is the variability in fastest speed driven by male college students different from female college students ?" The psychologist conducted a survey of a random $n_1 = 34$ male college students and a random $n_2 = 29$ female college students. Here is a descriptive summary of the results of her survey:

Male (n_1)	**Female** (n_2)
$n_1 = 34$	$n_2 = 29$
$\bar{x}_1 = 105.5$	$\bar{x}_2 = 90.0$
$s_1^2 = 404.01$	$s_2^2 = 148.84$

A Step-by-step Hypothesis Testing

1. Specify what you are comparing.
2. Formulate hypotheses
3. Check assumptions
4. Determine significance level α
5. Compute the test statistic
6. Check significance
7. Make a decision about whether to reject H_0
8. Interpret findings

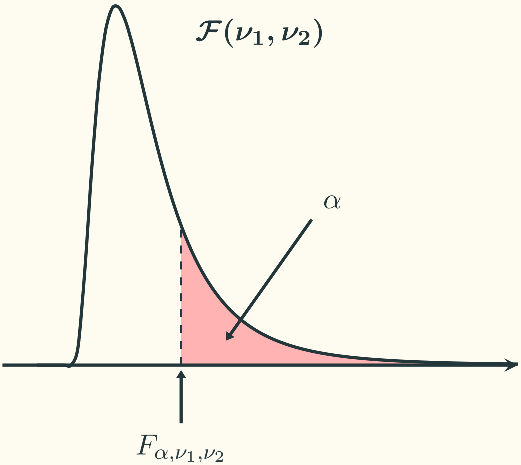
6a)



6b) Calculate the p-value.

6c) Construct $(1-\alpha) \times 100\%$ confidence interval to see if it covers the H_0 value.

Practical property of the $\mathcal{F}$-distribution



$$\mathbb{P}\left(F \geqslant F_{\alpha,\nu_1,\nu_2}\right) = \alpha$$

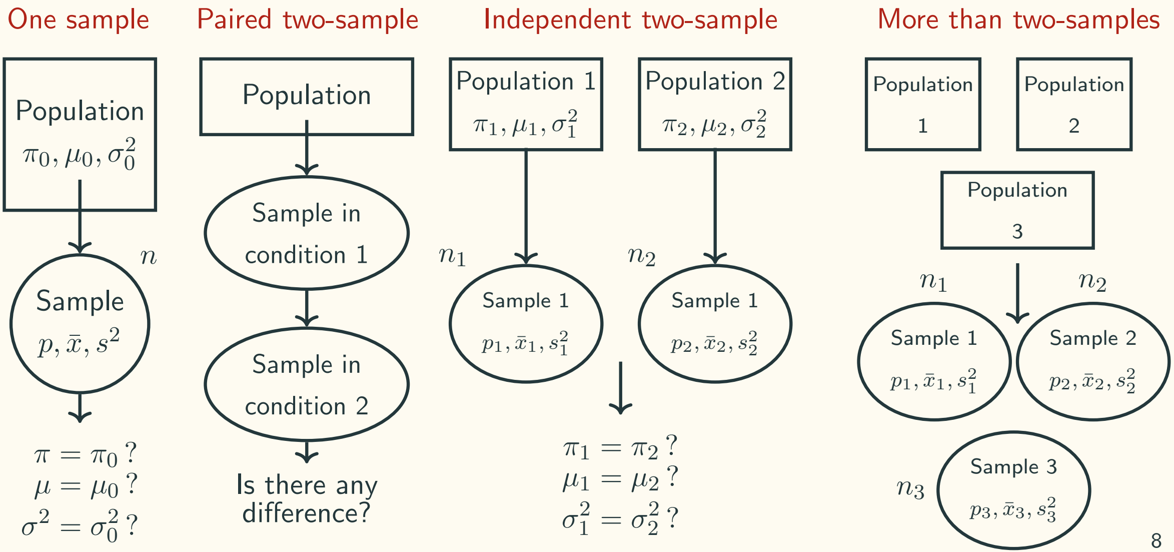
$$\mathbb{P}\left(\frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \geqslant F_{\alpha,\nu_1,\nu_2}\right) = \alpha$$

$$\mathbb{P}\left(\frac{s_2^2}{s_1^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} \leqslant \frac{1}{F_{\alpha,\nu_1,\nu_2}}\right) = \alpha$$

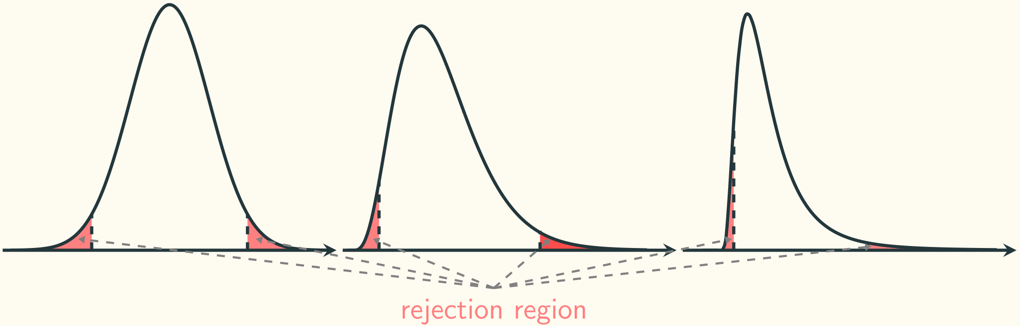
$$\mathbb{P}\left(F_{\nu_2,\nu_1} \leqslant \frac{1}{F_{\alpha,\nu_1,\nu_2}}\right) = \alpha$$

$$F_{\alpha,\nu_1,\nu_2} = \frac{1}{F_{1-\alpha,\nu_2,\nu_1}}$$

Summary of Hypothesis Testing



Foundation



- Sampling distribution of the difference/ratio of the sample proportion/mean/variance
- **Logic:** if H_0 were true, we would expect the majority of the test statistics (z, t, χ^2, F) falling into the middle area of the corresponding distribution. Therefore, the probability that the test statistic falls into the rejection regions is small. If we observe that, we reject H_0.

Interpreting The Results of Hypothesis Testing

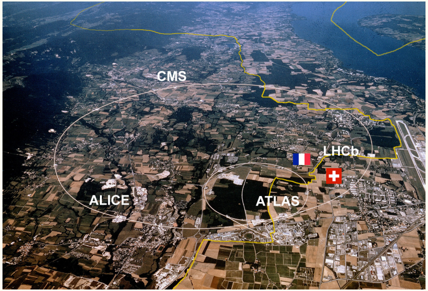
- p -value $= \mathbb{P}\,(\text{data} \mid H_0 \text{ is true})$
- Reject H_0 the data suggests that there is significant difference of ...
- Do not reject H_0: the data does not provide enough evidence to support that there is significant difference of . . .

Interpreting The Results - The Higgs Boson

The Tevatron @ The Fermi National Accelerator Laboratory

Research in March, 2012 reported here found evidence for the existence of the Higgs Boson particle. However, the evidence for the existence of the particle was not statistically significant. "We see some tantalizing evidence but not significant enough to make a stronger statement" said Rob Roser.

LHC @ CERN



Just a few months later: the CMS team from CERN: "CMS observes an excess of events at a mass of approximately 125 GeV with a statistical significance of five standard deviations (5 sigma) above background expectations. The probability of the background alone fluctuating up by this amount or more is about one in three million."

Interpreting The Results

- $p \geqslant 0.05$ does not mean H_0 is correct. You may need large sample size to detect small effect.

- Use p-values as a rule to guide behaviour **in the long run**.

\- $p < \alpha$: **Act** as if the data is not noise.

\- $p \geqslant \alpha$: Remain uncertain or **act** as if the data is noise.

* If you follow these rules, you will not make type I errors more than α of the time in the long run.

- When $p \geqslant 0.05$: think and explain. Use this to design better and progressive experiments.