

Statistics 431:

Statistical Inference

**Lecture 9: Two-sample variance problems and
one-way ANalysis Of VAriance (ANOVA)**

Inference about variance parameters: basics

- Up to now, all our two-sample inferences have concerned the mean parameters μ_1 and μ_2 . As for the variances σ_1^2 and σ_2^2 , we
 - assumed we knew their values, or
 - substituted (S_1^2, S_2^2) for (σ_1^2, σ_2^2) .
- We also considered making the assumption $\sigma_1^2 = \sigma_2^2 = \sigma^2$, in which case we
 - assumed we knew the common variance σ^2 , or
 - substituted the pooled variance estimate S_p^2 for σ^2 .
- If using pooled methods is important to us, we may want to consider testing the hypothesis that $\sigma_1^2 = \sigma_2^2$, or forming a CI for σ_1^2/σ_2^2 (not $\sigma_1^2 - \sigma_2^2$; the reason will be given later).
- For testing, we will need a test statistic with a known null distrn, as usual.
- For a CI, we will need a pivot, as usual.
- It turns out we require a new family of distrns—not the $N(0, 1)$ distrn or t_n family we have encountered so far.

The F_{ν_1, ν_2} distribution

- Let U be a χ^2 rv with ν_1 degrees of freedom.
- Let V be a χ^2 rv with ν_2 degrees of freedom, *independent* of U .
- Then the random variable

$$W = \frac{U/\nu_1}{V/\nu_2}$$

has a known distrn, called the “ F distrn with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom,” and denoted $F_{\nu_1; \nu_2}$.

- “Known” means there is a formula for the pdf, and there are tables for the cdf, lower quantiles, and upper quantiles. But it’s easy to find a lower quantile given an upper quantile (and vice-versa), because

$$F_{1-\alpha; \nu_1; \nu_2} = 1/F_{\alpha; \nu_1; \nu_2} .$$

- It’s awkward to use printed tables of F probabilities and quantiles, because the family of F distrns is indexed by two parameters. For many years, practicing statisticians have turned to statistical computing packages to look up these numbers (indeed, to look up the cumulative distrn or quantile function of any distrn).

Two variance parameters: inference

- $X_1, \dots, X_m \sim N(\mu_1, \sigma_1^2)$.
 $Y_1, \dots, Y_n \sim N(\mu_2, \sigma_2^2)$. As usual for unpaired data, all the X_i 's are independent of all the Y_i 's.
- Writing S_1^2 for the sample variance of the X_i 's and S_2^2 for the sample variance of the Y_i 's (as usual), we see that

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has the $F_{m-1;n-1}$ distrn. (The duplicate notation is conventional.)

- This is because $(m-1)S_1^2/\sigma_1^2 \sim \chi_{m-1}^2$, and $(n-1)S_2^2/\sigma_2^2 \sim \chi_{n-1}^2$, and they are independent as random variables (why?).
- The intuition behind the procedure to test $H_0 : \sigma_1^2 = \sigma_2^2$ is: under this null hypothesis, the ratio $F = S_1^2/S_2^2$ ought to be near 1, because the top and bottom are just two independent estimates of the same number. Furthermore, the null distrn of the rv F is $F_{m-1;n-1}$, so we can compute critical values.

- The formal testing procedure is then as follows:
 - $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_A : \sigma_1^2 \neq \sigma_2^2 \Rightarrow$ reject when $F = S_1^2/S_2^2 > F_{\alpha/2;m-1;n-1}$ or $F < F_{1-\alpha/2;m-1;n-1}$
 - $H_0 : \sigma_1^2 \leq \sigma_2^2$ vs $H_A : \sigma_1^2 > \sigma_2^2 \Rightarrow$ reject when $F > F_{\alpha;m-1;n-1}$
 - $H_0 : \sigma_1^2 \geq \sigma_2^2$ vs $H_A : \sigma_1^2 < \sigma_2^2 \Rightarrow$ reject when $F < F_{1-\alpha;m-1;n-1}$
- Notice this test works no matter what the values of μ_1 and μ_2 are, just like our test of $H_0 : \mu_1 - \mu_2 = \Delta_0$ worked for any values of (σ_1^2, σ_2^2) .
- But there is a big difference: μ_1 and μ_2 don't even matter when deriving F . Recall that for large-sample tests about $\mu_1 - \mu_2$, we had to substitute sample variances in for σ_1^2 and σ_2^2 to get the test statistic T .
- The reason we use the test statistic $F = S_1^2/S_2^2$ rather than $S_1^2 - S_2^2$ is simple: the null distrn of the former is known and easy to work with; the null distrn of the latter does not have a “closed form.”
- To form a CI for σ_1/σ_2 , treat F as a pivot:

$$P \left(F_{1-\alpha/2;v_1;v_2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{\alpha/2;v_1;v_2} \right) = 1 - \alpha .$$

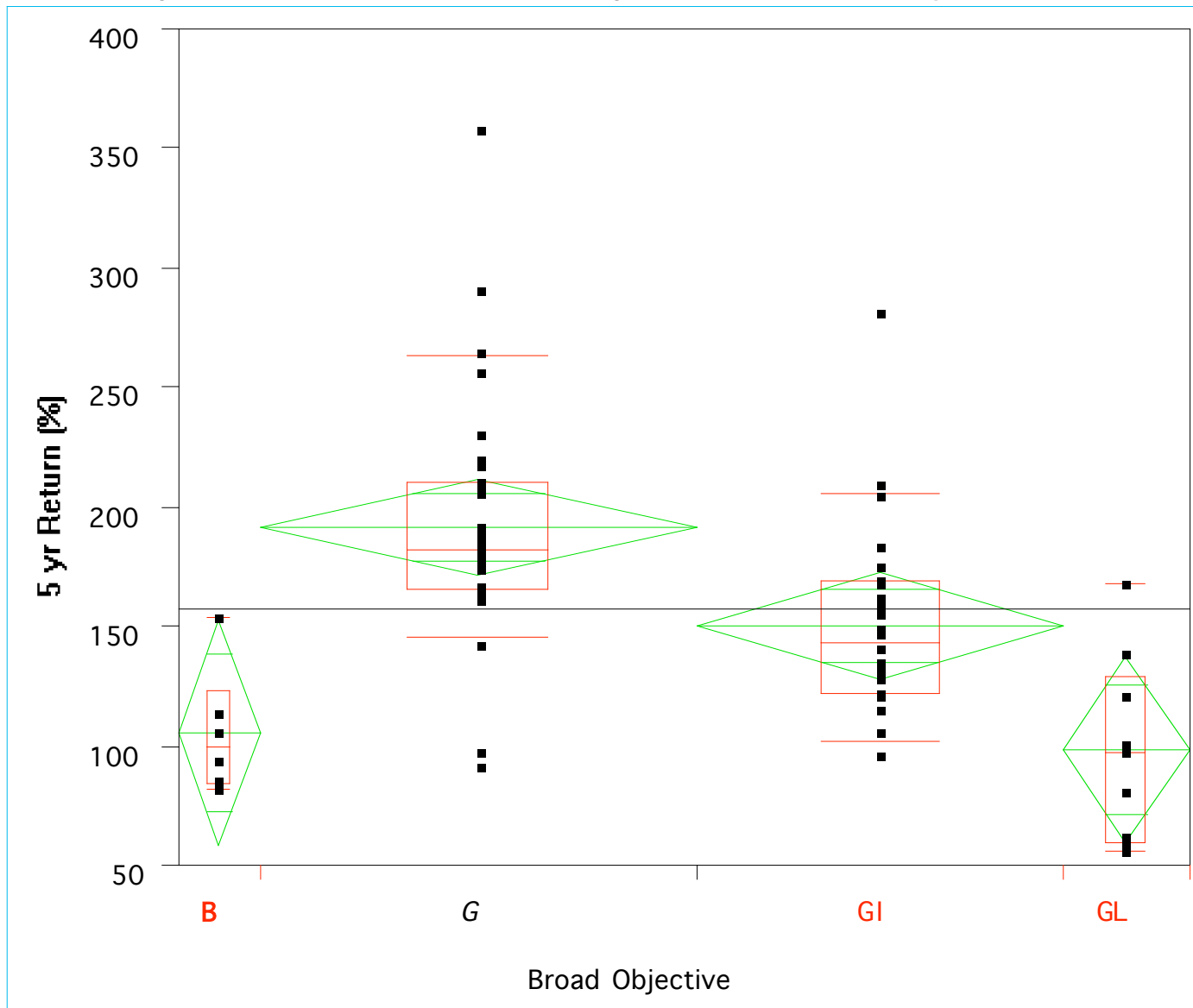
You should be able to get the CI from this.

ANalysis Of VAriance (ANOVA): basics

- So far we have looked at making inferences about one population and about two populations (including two treatments applied in randomized experiments).
- Often there are many populations we'd like to compare, or many treatments we'd like to apply. We need an inference procedure for this setting too; it's called ANOVA.
- Examples:
 - how much does average annual income vary across the four age brackets 18-25, 26-35, 36-50, and 51+ ?
 - how long do similar cancer patients live on average under five different treatment regimes?
 - do average hourly sales on the Amazon.com web site differ according to which of 10 versions of the home page are displayed?
- We want statistical tests that compare the population means, and confidence sets (not intervals) that have high probability of containing all of the population means.

Grouped data: example

5 yr Return (%) By Broad Objective



ANOVA terminology

- The *response variable* Y (continuous) is the random variable of interest, whose distrn varies by group. In the examples, the response variable is
 - (A) dollars earned per year
 - (B) years lived after the initiation of cancer treatment
 - (C) dollars of sales per hour
- The *factor* is the characteristic that distinguishes the populations / treatments / groups. Because there is only one factor, this is *one-way* ANOVA. In the examples, the factor is
 - (A) age bracket, (B) treatment type, (C) home page version
- The *levels* of the factor are the different values it takes on: one level per population / treatment / group. In the examples, the levels are
 - (A) age bracket 1 (18-25), age bracket 2 (26-35), age bracket 3 (36-50), age bracket 4 (51+)
 - (B) cancer treatment 1, . . . , cancer treatment 5
 - (C) home page version 1, . . . , home page version 10

ANOVA: conventional notation

- the number of levels is denoted I .
- the mean of the i th population / treatment distrn is denoted $\mu_i, i = 1, \dots, I$.
- the number of observations drawn from the i th population or treatment distrn is J . For now, J does not depend on i : we have the same number of observations at each level of the factor.
- the j th observation from the i th population or treatment distrn is Y_{ij} . the indexing is $Y_{\text{group, obs in group}}$. (The book uses X , not Y : X is non-standard.)
- a dot (\cdot) means “sum over the index.” for example, $Y_{i\cdot} = \sum_{j=1}^J Y_{ij}$, and $\bar{Y}_{i\cdot} = (1/J) \sum_{j=1}^J Y_{ij}$. Notice $\bar{Y}_{i\cdot}$ is the sample mean of obsvns from level i .
- the *grand mean*, and the sample variance of the i th distrn, are

$$\bar{Y}_{\cdot\cdot} = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J Y_{ij} \quad \text{and} \quad S_i^2 = \frac{1}{J-1} \sum_{j=1}^J (Y_{ij} - \bar{Y}_{i\cdot})^2 .$$

ANOVA: hypothesis testing

- To start with, we're interested in the null hypothesis that all the population means are equal, $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$.
- The alternative hypothesis is H_A : at least two of the μ_i 's differ. H_A contains every possible configuration of values for the means, *except* configurations where all the means are the same.
- The null hypothesis is a *straw man*: rejecting H_0 allows us to conclude that something is different about the means, somewhere among the levels of the factor. Rejecting H_0 does not tell us which means differ, or by how much: that requires further analysis.
- To conduct the test, we will make some strong assumptions: each population distrn i is $N(\mu_i, \sigma^2)$. Notice we are assuming the variance is the same at every level of the factor (this is called *homoscedasticity*). Equal variances allow us to use a pooling procedure.
- We can form a single normal quantile plot to check these assumptions (how?).

The ANOVA test statistic

- If H_0 is true, then the entire data set of Y_{ij} 's is just one big sample from the $N(\mu, \sigma^2)$ distrn, because $\mu_1 = \mu_2 = \dots = \mu_I = \mu$ under H_0 . In this case we can estimate σ^2 in two different ways:
 1. look at the spread of the group means around the grand mean
 2. look at the spread of each group's observations around that group's mean
- Intuitively, if the population means are the same, then
 1. the observed spread of the group means will be about σ^2/J (why?)
 2. the observed spread of each group's observations will be about σ^2 (why?)
- So, under H_0 , two reasonable estimates of σ^2 are
 1. $J \times$ sample SD of $\bar{Y}_{1.}, \dots, \bar{Y}_{I.}$
 2. equal-weighted average (pooling) of S_1^2, \dots, S_I^2

- Define the *mean square for treatment* (also known as *mean square for model*)

$$\text{MSTr} = J \cdot \frac{1}{I - 1} \sum_{i=1}^I (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

- Define the *mean square for error*

$$\text{MSE} = \frac{1}{I} \sum_{i=1}^I s_i^2$$

- Then the test statistic for $H_0 : \mu_1 = \mu_2 = \dots = \mu_I$ vs $H_A : "H_0 \text{ is false}"$ is

$$F = \frac{\text{MSTr}}{\text{MSE}} .$$

- This explains the name “analysis of variance”: we’re interested in whether the group means differ, but we test for a difference by estimating (“analyzing”) the common variance parameter in two different ways.