§10.1: Single Factor ANOVA (Introduction)

Background: Random variable X is sampled N times
- samples are divided into "blocks" according to values of a "factor".

[**Notation:** Use $f_1, f_2, \ldots$ or just f for "factor values"
#factors $= k$.]

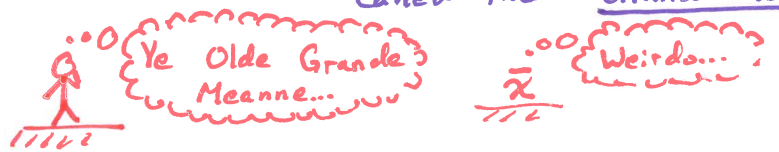
Notation: $X_i^{(f)}$ = sample i with factor f

$\mu^{(f)} = E[X^{(f)}]$ mean of factor f

$\bar{X}^{(f)} = \bar{x}^{(f)}$ sample mean of factor f

$\bar{\bar{X}} = \bar{\bar{x}}$ sample mean of all data

↳ Called the "Grand Mean"



Independent random variables $X^{(f_1)}, X^{(f_2)}, \ldots X^{(f_k)}$ (the "blocks" for each "factor value") with same variance.

Each variable is sampled n_f times

$X^{(f_1)} \to X_1^{(f_1)}, X_2^{(f_1)}, \ldots, X_{n_1}^{(f_1)}$ ← (#samples $= n_1$)

$X^{(f_2)} \to X_1^{(f_2)}, X_2^{(f_2)}, \ldots, X_{n_2}^{(f_2)}$ ← (#samples $= n_2$)

$\vdots$

Total #samples $n_1 + n_2 + \cdots + n_k = N$

Note: Either assume each $X^{(f)}$ is Normal or that $n_f > 30$.

(There is also a version which does not assume equal variance... but we won't discuss it.)

Hypothesis Test

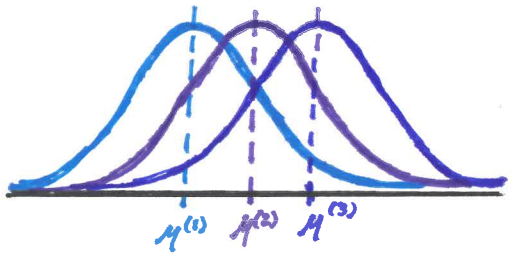
H_0: "Factor doesn't matter"
↳ All $\mu^{(f)}$ are equal, so all $X^{(f)}$ are just X.

H_A: At least two factors are different
↳ Two $\mu^{(f)}$ are unequal (at least)

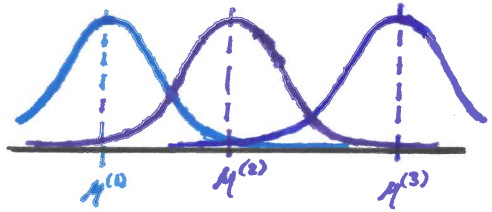
[If there are only two factor values then we would simply do a (pooled variance) two-sample t-Test. In general we could try to do a t-Test for each pair of factor values... but we would have "α-amplification" — performing many t-Tests gives high chance of Type I error. (see §10.3)]

We will use the F-distribution to test H_0 without also checking which factors have different means.

↳ Compare variance within factors to variance between factors.



Maybe all from same distribution X.



Definitely not from same distribution X.

The name "Analysis of Variance" comes from using test comparing variance within & between factors.

Recall: If $\left\{ \begin{matrix} X \sim \chi^2(n) \\ Y \sim \chi^2(m) \end{matrix} \right\}$ are "Chi-squared" r.v.'s (for example: sample var.)

then $\frac{X/n}{Y/m} \sim F(n, m)$

↳ F with "numerator d.f. = n" "denominator d.f. = m"

Idea: Use $F = \frac{\text{"sample variance between factors"}}{\text{"sample variance within factors"}}$

Total Variance

Define "Total Sum of Squares" as

$$SS_T = \sum_i^{'} (x_i^{(f_1)} - \bar{x})^2 + \cdots + \sum_i^{'} (x_i^{(f_k)} - \bar{x})^2$$

$$= \sum_{i,f}^{'} (x_i^{(f)} - \bar{x})^2$$

Σ Compare all samples to "Grand Mean"

Define "Total Mean Square" as

$$MS_T = SS_T / (N-1)$$

"Grand Sample Variance"

Note: This is just s^2 - the sample variance of X (ignore division into factors)

"Total Variance" breaks into two parts:

- "Factor Variance" ← Variance between factors
- "Residual Variance" ← Variance within factors

Factor Variance

Define "Factor Sum of Squares" as

$$SS_F = n_1 \cdot (\bar{x}^{(f_1)} - \bar{x})^2 + \cdots + n_k (\bar{x}^{(f_k)} - \bar{x})^2$$

and "Factor Mean Square" as

$$MS_F = SS_F / (k-1)$$

Note: This is basically $s^2_{\bar{x}}$ "sample variance of means."

Residual Variance

This is often called "Sum of Squared Error"

Define "Residual Sum of Squares" as

$$SS_E = \sum_i^1 (x_i^{(f_1)} - \bar{x}^{(f_1)})^2 + \cdots + \sum_i^1 (x_i^{(f_k)} - \bar{x}^{(f_k)})^2$$

$$= (n_1 - 1)\, s_1^2 + \cdots + (n_k - 1)\, s_k^2$$

(s_1^2: sample variance of first factor; s_k^2: sample variance of last factor)

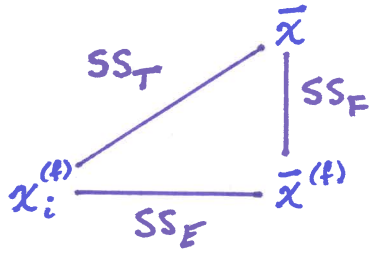
and "Residual Mean Square" as "Mean Squared Error"

$$MS_E = \frac{SS_E}{N-k} = \frac{SSE}{(n_1-1)+\cdots+(n_k-1)}$$

Note: This is basically "mean of sample variances"

Since variances add like Pythagorean thm, you can show that

$$SS_T = SS_F + SS_E$$



In practice this is used to compute SS_E:

$$SS_E = SS_T - SS_F$$

(SS_T: Easy to compute. (Relatively))

Summary so far:

- "Total Mean Square" MS_T = (sample variance of X ignore (factors))
- "Factor Mean Square" MS_F = (sample variance of factor means)
- "Residual Mean Square" MS_E = (mean of factor sample variances)

[All of the scaling & denominators were chosen so that variances would match and results would be χ^2.]

Thm:

$$SS_T \sim \chi^2(N-1)$$ ← Total #samples − 1

$$SS_F \sim \chi^2(k-1)$$ ← #factors − 1

$$SS_E \sim \chi^2(N-k)$$ ← Total #samples − #factors

Thus

$$\frac{SS_F/k-1}{SS_E/N-k} = \frac{MS_F}{MS_E} \sim F(k-1, N-k)$$

"

$$\frac{\text{Sample Variance of Factor Means}}{\text{Mean of Factor Sample Variances}} \sim F(\#\text{factors} - 1, \#\text{samples} - \#\text{fact.})$$

"

Hypothesis Test: H_0: All $\mu^{(k)}$ are equal

↳ All factors have same distribution

Test Statistic:

$$F = \frac{MS_F}{MS_E} \sim F(k-1, N-k)$$

p-value:

$$1 - pf\left(\frac{MS_F}{MS_E}, k-1, N-k\right)$$

Note: This should always be a right tail test.

The F-test above is "Analysis of Variance"

Note: If H_0 is true then $MS_F \approx MS_E$

↳ So $F \approx 1$

If H_0 is false then $MS_F > MS_E$

↳ So $F > 1$

Larger F ⟹ Smaller p-value

Note: If there are only two factors then

$F(2-1, N-2) = t(N-2)$ (F-test is pooled var. t-Test)

In general,

Larger MS_F (Big variance between sample means of factors)

Smaller MS_E (Small sample variance within factors)

Larger n_f (Bigger sample sizes for each factor)

⟹ Larger F & Smaller p-value

When you ask a computer to perform an ANOVA the output will be an ANOVA table which records the p-value as well as a lot of other data

Example:

	DF ("degrees of freedom")	SS ("sum of squares")	MS ("mean square")	F-value	p-value
Factor	5	1815	363.0	3.0	0.0184
Error	54	6534	121.0		
Total	59	8349	141.5		

Note: $MS = \frac{SS}{DF}$, $F\text{-val} = \frac{MSF}{MSE}$

Looking at the "degrees of freedom" column of the table, we can see that there were $60 = 59+1$ total samples divided among $6 = 5+1$ factors.

Total variance: $s^2 = 141.5$

⑤ §10.1 Addendum: "Effect Size" (η^2)

Sometimes ANOVA tables will also list another value:

The "Effect Size" is

$$\eta^2 = \frac{SS_F}{SS_T} = \frac{SS_F}{SS_F + SS_E}$$

↑ Greek letter "eta" for Effect

Idea: η^2 = proportion of total variance coming from the spread of factors

η^2 is useful for "rule of thumb" calculations where you don't have access to something that can compute p-values of F-statistics.

(This is an unlikely situation in our current age when anyone's cell phone can compute p-values...)

Basic rule:

- If $\eta^2 > 0.01$ then there is small effect
- If $\eta^2 > 0.06$ then there is medium effect
- If $\eta^2 > 0.14$ then there is large effect

Effect size is about more than just p-value

↳ it also involves how far apart the different factor distributions are

Note: Knowing that $\bar{X}^{(f_1)}$ & $\bar{X}^{(f_2)}$ have different means might not be that interesting if $\mu^{(f_1)} - \mu^{(f_2)} = .001$

↳ i.e. $\mu^{(f_1)} \neq \mu^{(f_2)}$ but $\mu^{(f_1)} \approx \mu^{(f_2)}$

"Large effect" means not only that p-value is small, but also that the difference between means is significant.