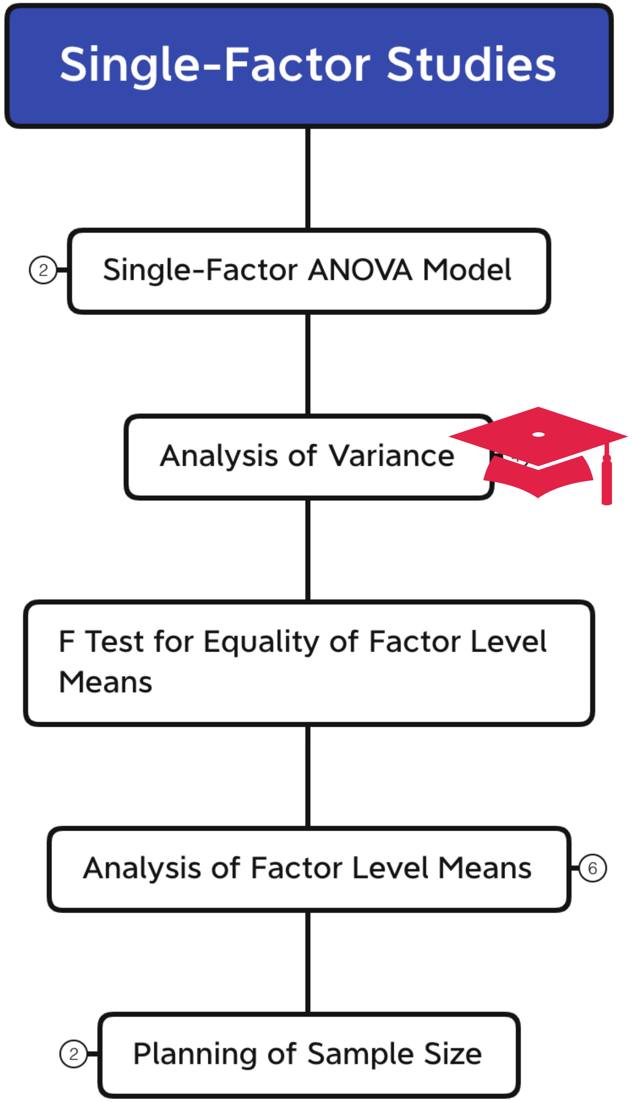
Lecture 2: Single-Factor Studies

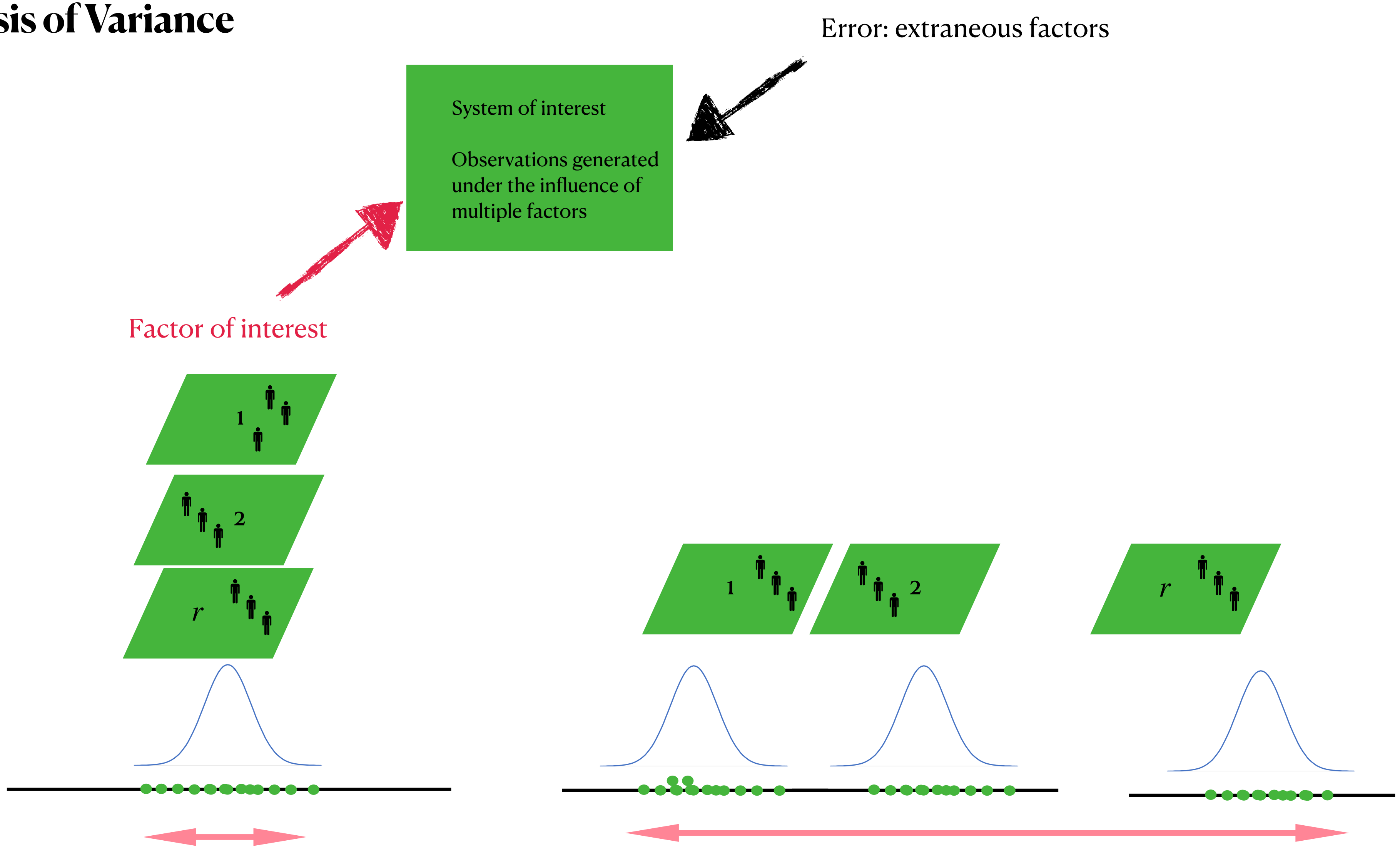
STA 106: Analysis of Variance

Suggested reading: ALSM Chapter 16 & 17

Xiner Zhou
Department of Statistics, University of California, Davis


Single-Factor Studies
Single-Factor ANOVA Model
2
Analysis of Variance
F Test for Equality of Factor Level Means
Analysis of Factor Level Means
6
Planning of Sample Size
2

Analysis of Variance



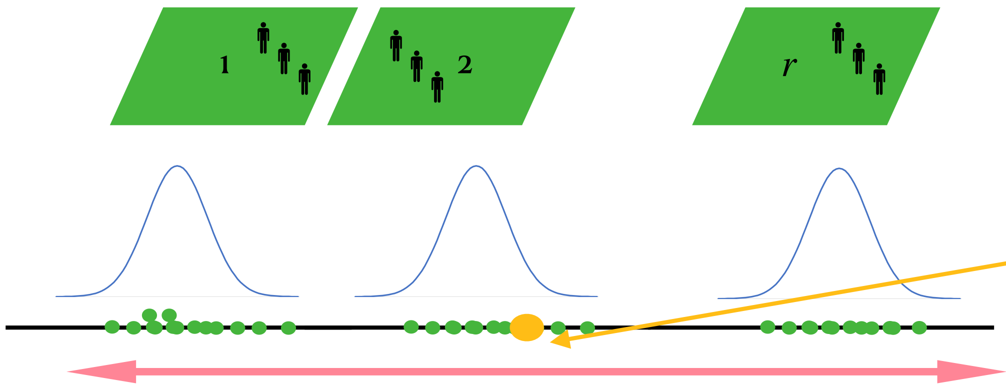
Without factor of interest, the observations have some natural variation due to other extraneous factors, i.e. "error variance"

If the factor of interest indeed has some effects on the system, then we would expect more volatility than a system without the factor

Analysis of Variance

Idea : Partition sums of squares (= variation) associated with responses $Y_{ij}'s$

Notion of "Total Variation":



$$\overline{Y_{..}} = \frac{1}{n_T} \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$$

Each observation Y_{ij} deviates from overall sample mean by $Y_{ij} - \overline{Y_{..}}$

Measure of total variation is sum of squared deviations

$$SSTO = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{..}})^2$$

SSTO: Total Sum of Squares

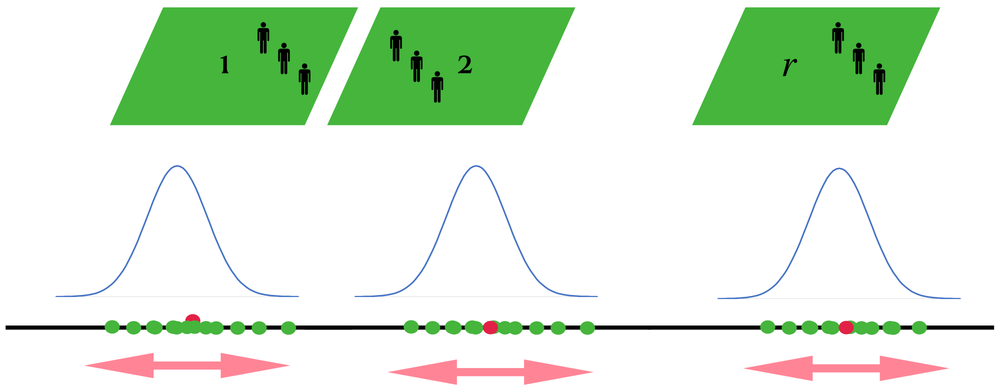
- If all $Y_{ij}'s$ are the same -> $SSTO = 0$
- If there is more variation among $Y_{ij}'s$ —> SSTO increases

SSTO: measures total variation or uncertainty observed in data, while these variation can be due to many different factors (reasons)

Analysis of Variance

Idea : Partition sums of squares (= variation) associated with responses $Y_{ij}'s$

Notion of "Variation due to error":



When we consider the factor under study, the variation within treatment is due to extraneous factors that we collectively call "error".

Each observation Y_{ij} deviates from treatment-specific sample mean by $Y_{ij} - \overline{Y_{i.}}$

Measure of variation due to error is sum of squared deviations

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i.}})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} e_{ij}^2$$

SSE: Error Sum of Squares or Residual Sum of Squares

- If there is no extraneous factors influencing the response, then we would expect all $Y_{ij}'s$ within each treatment to be exactly the same
  $Y_{ij} = \overline{Y_{i.}}$ —> SSE = 0

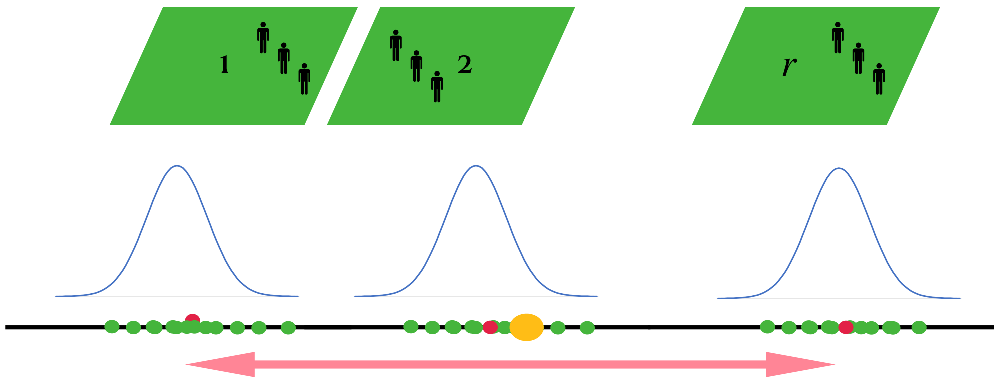
- If there are many extraneous factors influencing the response, then we would expect $Y_{ij}'s$ vary dramatically within each treatment
  Y_{ij} and $\overline{Y_{i.}}$ very different —> SSE large

SSE: measures residual variation or remaining variation observed in data, that is not due to treatment, but due to extraneous factors

Analysis of Variance

Idea : Partition sums of squares (= variation) associated with responses $Y_{ij}'s$

Notion of "Variation due to treatment":



For each observation Y_{ij}, we expect some variation for the fact that they belong to different treatment groups that have different treatment means

So we expect to see some variation in each observation Y_{ij} by an amount $\overline{Y_{i.}} - \overline{Y_{..}}$

Measure of variation due to treatment is sum of squared deviations

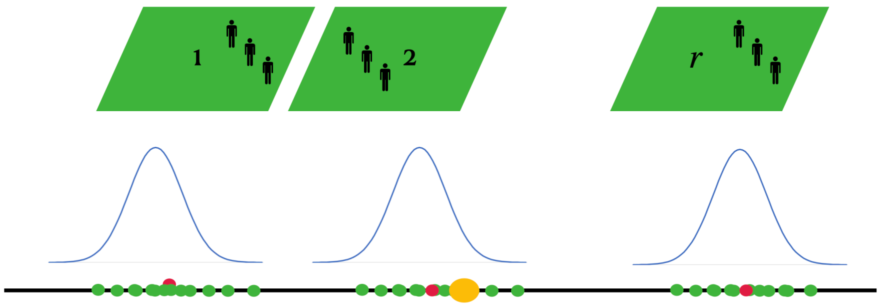
$$SSTR = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\overline{Y_{i.}} - \overline{Y_{..}})^2 = \sum_{i=1}^{r} n_i (\overline{Y_{i.}} - \overline{Y_{..}})^2$$

SSTR: Treatment Sum of Squares

- If there is no difference whatsoever in the treatment means, that is, the factor has no treatment effect at all, then we would expect SSTR to be small
- If there is huge difference in the treatment means, that is, the factor under study do make a difference, then we would expect SSTR to be large

SSTR: measures variation due to the factor under study

Analysis of Variance



$$Y_{ij} - \overline{Y}_{..} = (Y_{ij} - \overline{Y}_{i.}) + (\overline{Y}_{i.} - \overline{Y}_{..})$$

Total deviation — Deviation around estimated treatment mean — Deviation of estimated treatment mean around overall mean

$$\longrightarrow \quad (Y_{ij} - \overline{Y}_{..})^2 = (Y_{ij} - \overline{Y}_{i.})^2 + (\overline{Y}_{i.} - \overline{Y}_{..})^2 + 2(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..})$$

$$\longrightarrow \quad \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\overline{Y}_{i.} - \overline{Y}_{..})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})(\overline{Y}_{i.} - \overline{Y}_{..})$$

$$= 2\sum_{i=1}^{r}(\overline{Y}_{i.} - \overline{Y}_{..})\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)$$

Observed that: $\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right) = 0$ for all i

Convince yourself if not clear

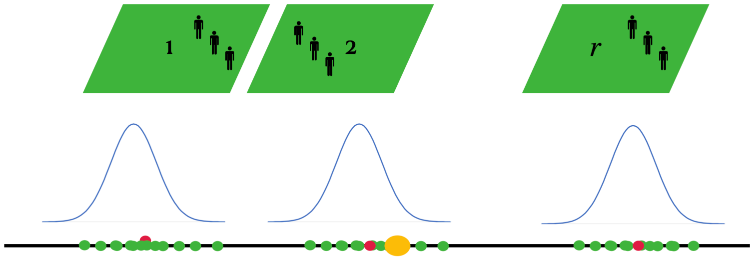
$$= 2\sum_{i=1}^{r}(\overline{Y}_{i.} - \overline{Y}_{..})\,0 = 0$$

$$\longrightarrow \quad \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{r} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

Total variation — Variation due to extraneous factors/ error — Variation due to factor or treatment under study

SSTO: Total sum of squares — SSE: Error sum of squares — SSTR: Treatment sum of squares

Analysis of Variance



$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2 + \sum_{i=1}^{r} n_i(\overline{Y}_{i.}-\overline{Y}_{..})^2$$

Total variation — Variation due to extraneous factors/ error — Variation due to factor or treatment under study

SSTO: Total sum of squares — SSE: Error sum of squares — SSTR: Treatment sum of squares

Partition of SSTO:

$$\boxed{\text{SSTO} = \text{SSTR} + \text{SSE}}$$

Error: extraneous factors

Factor of interest

System of interest

Observations generated under the influence of multiple factors

Can we separate the force that comes from the factor under study and the extraneous forces not of interest?

Yes! by analyzing and decomposing variation

- If there is no difference whatsoever in the treatment means, that is, the factor has no treatment effect at all, then we would expect SSTR to be small
- If there is huge difference in the treatment means, that is, the factor under study do make a difference, then we would expect SSTR to be large

This will be our intuition for detecting (testing) if the factor under study indeed has effect versus not!

Degrees of Freedom

Estimates of parameters can be based on different amount of "independent" information.
The number of independence pieces of information that go into the estimate of a parameter is called:
degree of freedom (d.o.f.)

Think of: dimensions of the space where an estimator lives in and allows to run free

$$\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{..}})^2}_{\text{Total variation}} = \underbrace{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i.}})^2}_{\text{Variation due to extraneous factors/ error}} + \underbrace{\sum_{i=1}^{r}n_i(\overline{Y_{i.}}-\overline{Y_{..}})^2}_{\text{Variation due to factor or treatment under study}}$$

SSTO: Total sum of squares | SSE: Error sum of squares | SSTR: Treatment sum of squares

How many independence pieces of information go into each quantity?

$Y_{ij} - \overline{Y_{..}}$

n_T pieces

But $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{..}}) = 0$

$df(SSTO) = n_T - 1$

$Y_{ij} - \overline{Y_{i.}}$

n_T pieces

But $\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i.}}) = 0$ for $i = 1\ldots r$

$df(SSE) = n_T - r$

$\overline{Y_{i.}} - \overline{Y_{..}}$

r pieces

But $\sum_{i=1}^{r} n_i((\overline{Y_{i.}}-\overline{Y_{..}}) = 0$

$df(SSTR) = r - 1$

$$df(SSTO) = df(SSE) + df(SSTR)$$

Mean Squares

$$\text{Mean Squares} = \frac{Sum\ of\ Squares}{d.f.}$$

Variation due to difference sources, but in equal comparison footing (in unit d.f.)

Treatment Mean Square:

$$\text{MSTR} = \frac{SSTR}{d.f.(SSTR)} = \frac{SSTR}{r-1}$$

Error Mean Square:

$$\text{MSE} = \frac{SSE}{d.f.(SSE)} = \frac{SSE}{n_T - r}$$

What's expected values of Mean Squares?

$$MSE = \frac{1}{n_T - r}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{i\cdot}})^2$$

$$= \frac{1}{n_T - r}\sum_{i=1}^{r}(n_i - 1)\underbrace{\frac{\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{i\cdot}})^2}{n_i - 1}}_{S_i^2\text{ : sample variance}}$$

$$= \frac{1}{n_T - r}\sum_{i=1}^{r}(n_i - 1)S_i^2$$

$$\Longrightarrow \quad E[MSE] = \frac{1}{n_T - r}\sum_{i=1}^{r}(n_i - 1)\underbrace{E[S_i^2]}_{\text{Sample variance is unbiased estimate of population variance which is error variance in this case}}$$

$$= \sigma^2$$

What's expected values of Mean Squares?

$$MSTR = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\overline{Y_{i\cdot}} - \overline{Y_{\cdot\cdot}})^2$$

$$\begin{aligned}
(\overline{Y_{i\cdot}} - \overline{Y_{\cdot\cdot}})^2 &= (\overline{Y_{i\cdot}} - \frac{\sum_{i=1}^{r} n_i \overline{Y_{i\cdot}}}{n_T})^2 \\
&= ((1-\frac{n_i}{n_T})\overline{Y_{i\cdot}} - \frac{\sum_{k\neq i} n_i \overline{Y_{k\cdot}}}{n_T})^2 \\
&= ((1-\frac{n_i}{n_T})(\overline{Y_{i\cdot}} - \mu_i + \mu_i) - \sum_{k\neq i}\frac{n_i}{n_T}(\overline{Y_{k\cdot}} - \mu_k + \mu_k))^2 \\
&= ((1-\frac{n_i}{n_T})(\overline{Y_{i\cdot}} - \mu_i) + (1-\frac{n_i}{n_T})\mu_i - \sum_{k\neq i}\frac{n_i}{n_T}(\overline{Y_{k\cdot}} - \mu_k) - \sum_{k\neq i}\frac{n_i}{n_T}\mu_k)^2
\end{aligned}$$

$$= ((1-\frac{n_i}{n_T})(\overline{Y_{i\cdot}} - \mu_i) - \sum_{k\neq i}\frac{n_i}{n_T}(\overline{Y_{k\cdot}} - \mu_k))^2 + ((1-\frac{n_i}{n_T})\mu_i - \sum_{k\neq i}\frac{n_i}{n_T}\mu_k)^2 + 2((1-\frac{n_i}{n_T})(\overline{Y_{i\cdot}} - \mu_i) - \sum_{k\neq i}\frac{n_i}{n_T}(\overline{Y_{k\cdot}} - \mu_k))((1-\frac{n_i}{n_T})\mu_i - \sum_{k\neq i}\frac{n_i}{n_T}\mu_k)$$

$$E[\,*\,] = (1-\frac{n_i}{n_T})^2\frac{\sigma^2}{n_i} + \sum_{k\neq i}(\frac{n_k}{n_T})^2\frac{\sigma^2}{n_k}$$

$$= \frac{\sigma^2}{n_T^2}\{\frac{(n_T-n_i)^2}{n_i} + \sum_{k\neq i} n_k\}$$

$$E[\,*\,] = (\mu_i - \overline{\mu_\cdot})^2$$

$$\overline{\mu_\cdot} = \sum_{i=1}^{r}\frac{n_i}{n_T}\mu_i$$

$$E[\,*\,] = 0$$

What's expected values of Mean Squares?

$$E[MSTR] = \frac{1}{r-1}\sum_{i=1}^{r} n_i \left(\frac{\sigma^2}{n_T^2}\{\frac{(n_T - n_i)^2}{n_i} + \sum_{k \neq i} n_k\} + \left(\mu_i - \overline{\mu}.\right)^2 \right)$$

$$= \frac{1}{r-1}\sum_{i=1}^{r} \frac{\sigma^2}{n_T^2}\left((n_T - n_i)^2 + n_i(n_T - n_i)\right) + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}.)^2$$

$$= \frac{1}{r-1}\sum_{i=1}^{r} \frac{\sigma^2}{n_T}\left(n_T - n_i\right) + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}.)^2$$

$$= \frac{r}{r-1}\sigma^2 - \frac{1}{r-1}\sum_{i=1}^{r} \frac{n_i}{n_T}\sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}.)^2$$

$$= \sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}.)^2$$

$$\overline{\mu}. = \frac{1}{n_T}\sum_{i=1}^{r} n_i \mu_i$$

What’s the distributions of Mean Squares?

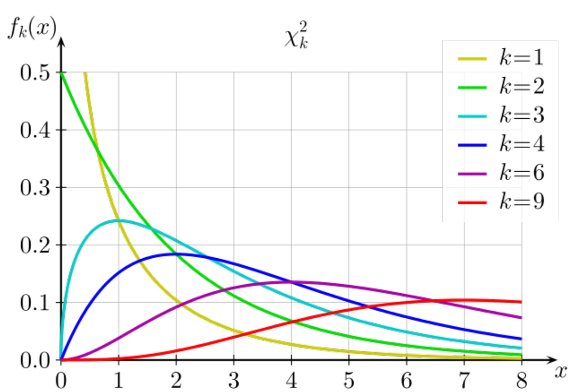
$$\frac{n_T - r}{\sigma^2} MSE = \frac{SSE}{\sigma^2} \sim \chi^2_{n_T - r}$$

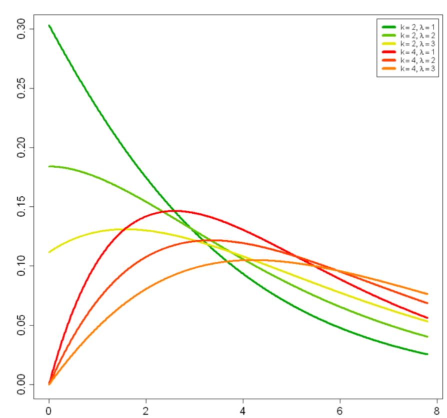
Chi-square distribution with degree of freedom $n_T - r$

$$\frac{r-1}{\sigma^2} MSTR = \frac{SSTR}{\sigma^2} \sim \chi^2_{r-1}\left(\frac{1}{\sigma^2}\sum_{i=1}^{r} n_i \left(\mu_i - \overline{\mu}.\right)^2\right)$$

Non-central Chi-square distribution with degree of freedom $r - 1$, and non-central parameter $\frac{1}{\sigma^2}\sum_{i=1}^{r} n_i \left(\mu_i - \overline{\mu}.\right)^2$

MSE and MSTR are independent random variables





ANOVA Table for Single-factor Studies

Source of Variation	SS	df	MS	E{MS}
Between treatments	$SSTR = \sum n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$r - 1$	$MSTR = \frac{SSTR}{r-1}$	$\sigma^2 + \frac{\sum n_i(\mu_i - \mu_.)^2}{r-1}$
Error (within treatments)	$SSE = \sum\sum(Y_{ij} - \bar{Y}_{i.})^2$	$n_T - r$	$MSE = \frac{SSE}{n_T - r}$	σ^2
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y}_{..})^2$	$n_T - 1$		

$$E[MSE] = \sigma^2$$

→ The expected value of MSE is the error variance

MSE is an unbiased estimate for σ^2

Thus, we use MSE as the estimate for the parameter σ^2

$$\hat{\sigma}^2 = MSE = \frac{1}{n_T - r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i.}})^2$$

ANOVA Table for Single-factor Studies

Intuition confirmed by mathematical derivation from analyzing variance:

$$E[MSTR] = \sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}.)^2$$

$$\geq$$

$$E[MSE] = \sigma^2$$

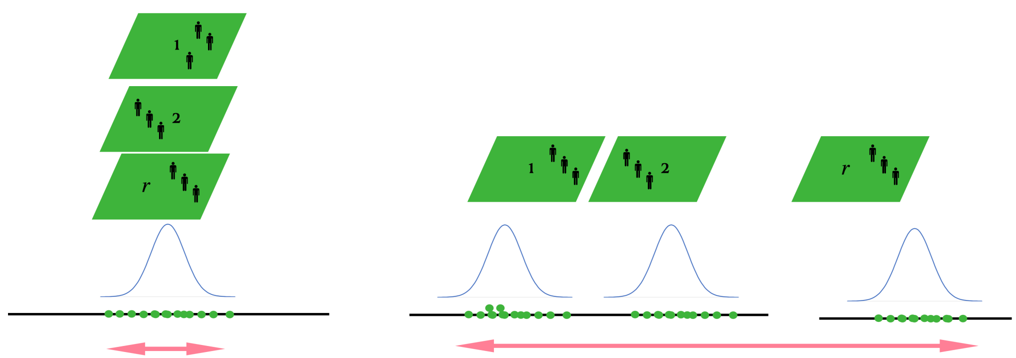
= holds if and only if when $\mu_i = \overline{\mu}.$ for all i

When all μ_i are equal: $E[MSTR] = E[MSE]$

MSTR will tend to be very close to MSE

When not all μ_i are equal:

MSTR tends to be larger than MSE,
how much larger is determined by how much different among population treatment means $\frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}.)^2$



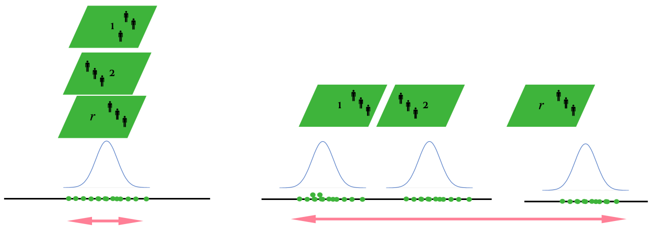
ANOVA Table for Single-factor Studies

Intuition confirmed by mathematical derivation from analyzing variance:

$$E[MSTR] = \sigma^2 + \frac{1}{r-1}\sum_{i=1}^{r} n_i(\mu_i - \overline{\mu}_{\cdot})^2$$

$$\geq$$

$$E[MSE] = \sigma^2$$



This unique property of MSE and MSTR (two sources of variance) gives us tool to construct test that can signal which situation is more likely to be true:

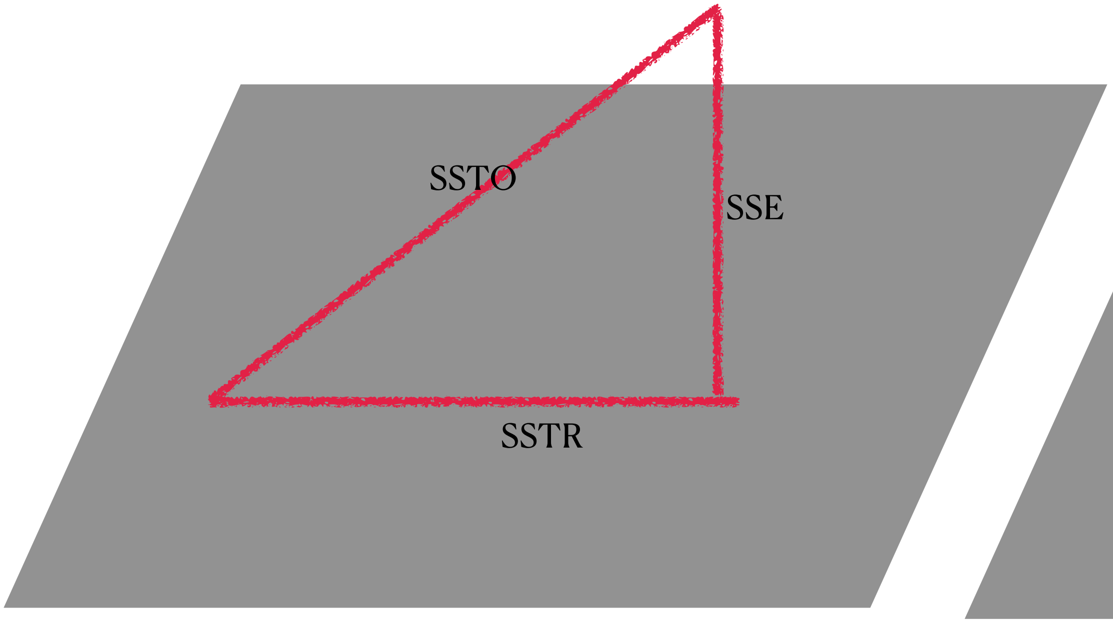
If we observe $MSTR \approx MSE$

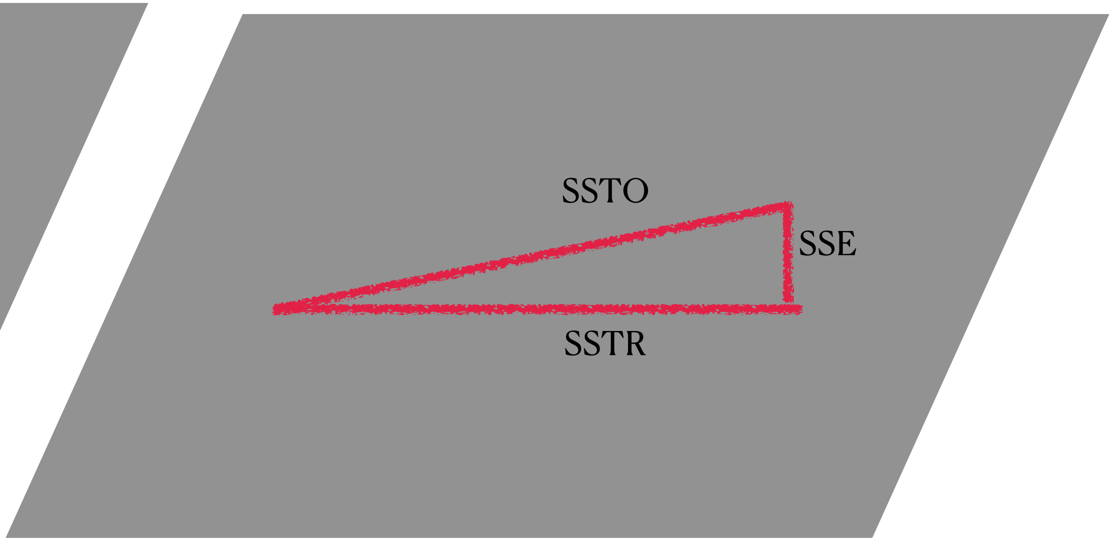
Then it signals that population treatment means are more likely to be the same,
that is, the factor does not have any effect on the response.

If we observe $MSTR >> MSE$

Then it signals that population treatment means are more likely to be not the same,
that is, the factor does not have some effect on the response.

Geometry of Decomposition of Variance:





Example

ANOVA Table

	SS	df	MS
Between treatments	672	2	336.0
Error (within treatments)	416	21	19.8
Total	1088	23	0.0