Mean Square for Treatments (MST) = $\dfrac{\text{SST}}{k-1}$

Mean Square for Error (MSE) = $\dfrac{\text{SSE}}{n-k}$

(handwritten) total # of observations in whole study *(arrow pointing to n)*

The ratio $\dfrac{MST}{MSE}$ **is called the ANOVA F-statistic.**

If $F = \dfrac{MST}{MSE}$ **is much bigger than 1, then the variation between groups is much bigger than the variation within groups, and we would reject**
H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ **in favor of** H_a.

Example (Table 10.3)
<u>Response</u>: **Distance a golf ball travels**
<u>4 treatments</u>: **Four different brands of ball**

$\bar{X}_1 = 250.8$, $\bar{X}_2 = 261.1$, $\bar{X}_3 = 270.0$, $\bar{X}_4 = 249.3$.

$\Rightarrow \bar{X} = 257.8$. $\longleftarrow$ *overall sample mean for whole study*

$n_1 = 10$, $n_2 = 10$, $n_3 = 10$, $n_4 = 10$. $\Rightarrow n = 40$.
Sample variances for each group:
$s_1^2 = 22.42$, $s_2^2 = 14.95$, $s_3^2 = 20.26$, $s_4^2 = 27.07$.

$$\text{SST} = 10(250.8 - 257.8)^2 + 10(261.1 - 257.8)^2$$
$$+ 10(270.0 - 257.8)^2 + 10(249.3 - 257.8)^2$$
$$= 2794.4$$

$$\text{SSE} = 9(22.42) + 9(14.95) + 9(20.26) + 9(27.07)$$
$$= 762.3$$

$$\text{MST} = \frac{\text{SST}}{k-1} = \frac{2794.4}{4-1} = 931.5$$

$$\text{MSE} = \frac{\text{SSE}}{n-k} = \frac{762.3}{40-4} = 21.175$$

$$\text{F} = \frac{\text{MST}}{\text{MSE}} = \frac{931.5}{21.175} = 43.99$$

This information is summarized in an <u>ANOVA table</u>:

b/w grps

within groups

Source	df	SS	MS	F
Treatments	$k-1$	SST	MST	MST/MSE
Error	$n-k$	SSE	MSE	
Total	$n-1$	SS(Total)		

Note that df(Total) = df(Trt) + df(Error)
and that SS(Total) = SST + SSE.

For our example, the ANOVA table is:

Source	df	SS	MS	F
Brands	3	2794.4	931.5	43.99
Error	36	762.3	21.175	
Total	39	3556.7		

Note: $\dfrac{SS(Total)}{df(Total)}$ **is simply the sample variance for the**
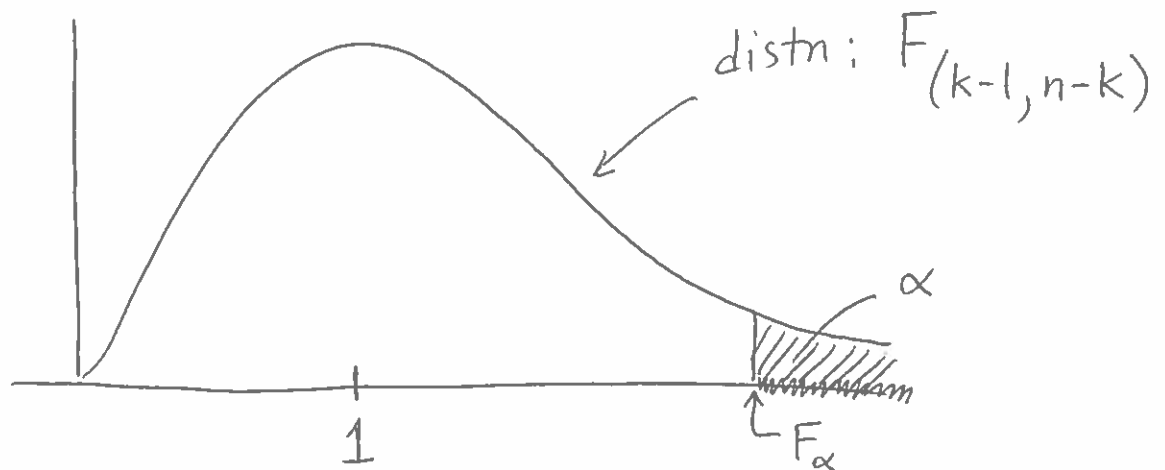
entire data set, $\dfrac{\sum(X-\bar{X})^2}{n-1}$.

In example, we can see F = 43.99 is "clearly" bigger than 1 ... but how much bigger than 1 must it be for us to reject H_0?

ANOVA F-test:
If H_0 is true and all the population means are indeed equal, then this F-statistic has an F-distribution with numerator d.f. $k-1$ and denominator d.f. $n-k$.

We would reject H_0 if our F is unusually large.
Picture:

distn: $F_{(k-1, n-k)}$

α

F_α

1

H₀: $\mu_1 = \mu_2 = \ldots = \mu_k$
Hₐ: **At least two of the treatment population means differ.**

Rejection Region: $F > F_\alpha$, where F_α based on $(k-1, n-k)$ **d.f.**

Assumptions:
• **We have random samples from the k populations.**
• **All k populations are normal.**
• **All k population variances are equal.**

Example: Perform ANOVA F-test using $\alpha = .10$.

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

H_a: At least two population means differ.

Reject H_0 if $F > F_{.10}$ (3 num. df, 36 denom. d.f.)

Table $\underline{V}$: $F_{.10} \approx 2.25$ for $(3, 36)$ d.f.

$F = 43.99 > 2.25$, so we reject H_0 and conclude that at least two of the golf ball brands have different population mean distances.

Which treatment means differ? Section 10.3 (Multiple Comparisons of Means) covers this issue.

Tukey method: Brands A and D do _not_ differ significantly in mean distance. All _other_ pairs of means differ significantly.

• **Mostly we have studied the behavior of a single random variable.**
• **Often, however, we gather data on two random variables.**
• **We wish to determine: Is there a relationship between the two r.v.'s?**
• **Can we use the values of one r.v. to predict the other r.v.?**
• **Often we assume a straight-line relationship between two variables.**
• **This is known as <u>simple linear regression</u>.**

Probabilistic vs. Deterministic Models

If there is an exact relationship between two (or more) variables that can be predicted with certainty, without any random error, this is known as a <u>deterministic relationship</u>.

Examples:

$$A = \pi r^2 \quad (\text{area of circle})$$

$Y = \text{pts scored in basketball game}$

$X_1 = \# \text{ free throws made}$

$X_2 = \# \text{ of 2-point Field goals}$

$X_3 = \# \text{ of 3-point Field goals}$
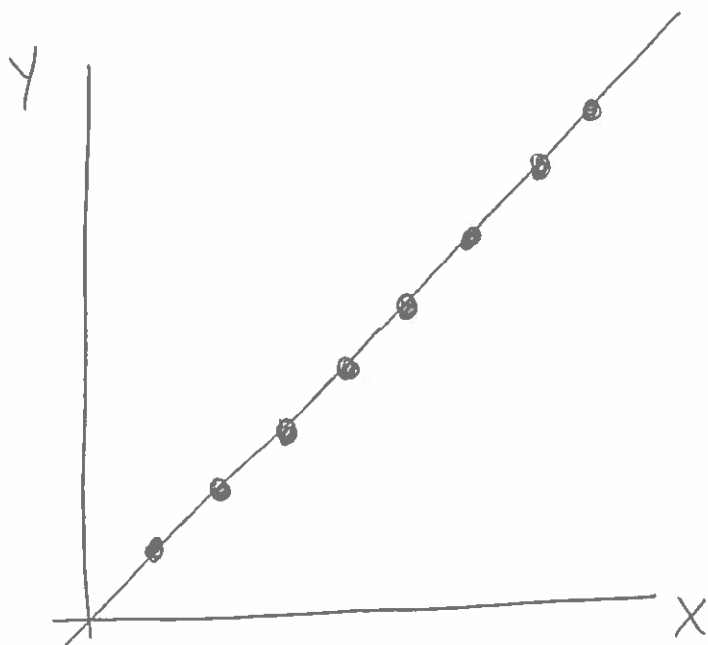
$$Y = X_1 + 2X_2 + 3X_3$$

In statistics, we usually deal with situations having random error, so <u>exact</u> predictions are not possible.

This implies a <u>probabilistic</u> relationship between the 2 variables.

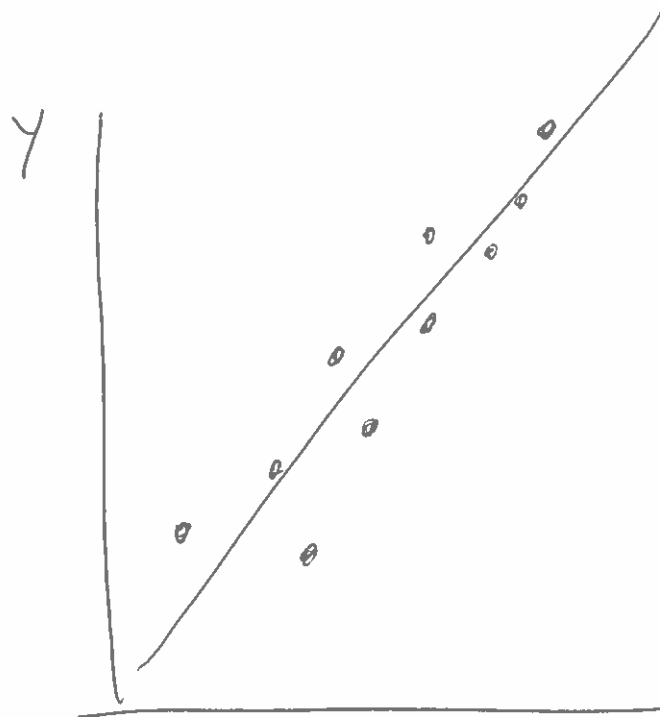**Example:** Y = breathalyzer reading
                    X = amount of alcohol consumed (fl. oz.)

<u>Possible Probabilistic Model</u>

$$Y = .003X + \text{"random error"}$$

A Deterministic Model

A Probabilistic Model