The entire set of m tests is called a "family". A "family" is "any logical collection of inferences for which it is meaningful to take into account some combined measure of error" or a set of tests where you wish to prevent "data dredging" (e.g. the spurious correlations in 342) or to "ensure a correct 'overall' decision in the collection of tests".

We'll discuss two error properties / metrics for a family of tests.

The first is called "familywise error rate" (FWER) defined as:

$$FWER := P(V > 0) \leq FWER_0$$ ← this is the level of control that I choose e.g. 5%

If you can show that FWER <= FWER_0 for any m_0 <= m subset of the m tests, this is calle "strong FWER control". We won't study it. If you can show that FWER <= FWER_0 for m_0 = m then this called "weak FWER control" which we will study. If m_0 = m....

Decision

Truth		Retain H_0	Reject H_0	
	H_0	U	V	m_0
	H_1	0	0	0
		F	R	m

$$\Rightarrow V = R \Rightarrow FWER = P(R > 0)$$

Our goal is weak FWER control under the most general settings.

$R_1 = 1$ if H_{0_1} is rejected, $R_1 = 0$ if H_{0_1} is retained

$R_2 = 1$ '' H_{0_2} '' , $R_2 = 0$ '' H_{0_2} '' ...

$\vdots$

$R_m = 1$ '' H_{0_m} '' , $R_m = 0$ '' H_{0_m} '' ...

$$FWER = P(R > 0) = P(R_1 = 1 \cup R_2 = 1 \cup \ldots \cup R_m = 1) \leq \sum_{i=1}^{m} P(R_i = 1) = m\alpha$$

recall from Math 241, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
the principle of inclusion-exclusion:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum P(A_i) - \sum P(A_i \cap A_j) + \sum P(A_i \cap A_j \cap A_k) - + - + \ldots$$

and from here you can prove "Boole's Inequality:"

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \sum P(A_i)$$

$$\Longrightarrow FWER \leq FWER_0 \Rightarrow m\alpha = FWER_0 \Rightarrow \alpha = \frac{FWER_0}{m}$$

this is called the Bonferroni correction (1936)

$\Rightarrow$ a p-value for an individual test must be less than FWER_0 / m. Equivalently, you can multiple the p-vales by m / FWER_0 and compare each to alpha = 5%.

$$p_{val} \leq \frac{FWER_0}{m} = \alpha \Rightarrow \underbrace{\frac{m}{FWER_0} p_{val}}_{\alpha} \leq \alpha.$$ Adjusted p-values

e.g. if m = 30, FWER_0 = 5% => alpha = FWER_0 / m = 0.00167

The obvious problem with this correction is... it gives you really bad power! Because it is ultra-conservative.

We can do a bit better if we assume the tests are independent.
Then, $R_1, R_2, \ldots, R_m \overset{iid}{\sim} Bern(\alpha)$. $\Rightarrow R \sim Bin(m, \alpha)$

$$FWER = P(R > 0) = 1 - P(R = 0) = 1 - (1 - \alpha)^m \leq FWER_0$$

$$\Rightarrow 1 - FWER_0 = (1 - \alpha)^m \Rightarrow 1 - \alpha = (1 - FWER_0)^{\frac{1}{m}}$$

$$\Rightarrow \alpha = 1 - (1 - FWER_0)^{\frac{1}{m}}$$ Dunn-Sidak Correction (1967)

e.g. if m = 30, FWER_0 = 5% =>
alpha = 1 - (95%)^(1/30) = 0.00171 > .00167 (the Bonferroni)
Thus, you get slightly higher power with Sidak.

$$1 - (1 - x)^{\frac{1}{c}} \approx \frac{x}{c}$$ (First order Taylor Series)

There are other methods e.g. the "Holm step-down" procedure (1979) but we won't study it because it is similar to the Simes procedure (1986) which we talk about now. Bonferroni and Sidak never looked at the p-values and there's a lot of information there. Remember, Fisher created the p-value to gauge the "strength" of a rejection. Rejecting with a p-value of 0.00001 is much stronger than rejecting with a p-value of 0.01. Holm and Simes used this. For the m tests, you get p-values p_1, p_2, ..., p_m but don't retain/reject anything yet!!! Order them from smallest to largest:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$$ (Order Statistics)

$p_{(1)}$: min pval; $p_{(m)}$: max pval

"linear step-up"

Then locate the following: $a_* := \max\left\{a : p_{(a)} \leq \frac{a}{m} FWER_0\right\}$

or let $a_* = 0$ if max doesn't exist.

Then set $\alpha = \frac{a_*}{m} FWER_0$ $\in \{$Bonferroni, ..., α (naive)$\}$

You can prove that this gives you weak FWER control. It is rare that this is not more powerful than Bonferroni / Sidak.

By construction you reject all tests up to the a_*th test (if the tests are in order of p-value). Then you retain all the other m - a_* tests.

The problem with FWER in general is maybe it's too conservative. What if you want to trade some false rejections for more power? Let's consider another metric of familywise control (not FWER), called "False Discovery Rate" (FDR). First, define the "False Discovery Proportion" (FDP),

$$FDP := \begin{cases} \frac{V}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases},$$ the random proportion of rejections that are Type I errors

$$FDR := E[FDP],$$ the expected proportion of rejections that are Type I errors.

Now we wish to control FDR so we want:
FDR <= FDR_0, a constant you set. For example if FDR_0 = 5% and I run m tests and get 100 rejections, then I expect <= 5 of the rejections to be Type I errors and >= 95 of the rejections to be real discoveries.

Note: if m = m_0 then FWER = FDR. Proof

$$m = m_0 \Rightarrow V = R \Rightarrow FDP = \begin{cases} 1 & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} = Bern(P(R > 0))$$

$$\Rightarrow FDR = E[FDP] = P(R > 0) = FWER$$

---

not on test

Note: the FDR procedure is more powerful than the FWER procedure

$$\mathbb{1}_{V \geq 1} \geq \frac{V}{R}$$ if $V = 0 \Rightarrow 0 \geq 0$ ✓

$V = 1 \Rightarrow 1 \geq 1$ or $\frac{1}{2}$ or $\frac{1}{3}$ ... $\forall R$

$V \geq 1 \Rightarrow 1 \geq 1$ or $\frac{V}{V+1}$ or ... $\forall R$

$$\Rightarrow E[\mathbb{1}_{V \geq 1}] \geq E\left[\frac{V}{R}\right]$$

$$P(V \geq 1) \geq FDR$$

$$\| $$

$$FWER \geq FDR$$ ✓

Benjamini and Hochberg (1995) proved the Simes procedure controls FDR for any m_0 subset of the m tests. In fact
FDR = m_0 / m FDR_0 <= FDR_0 thus for a small m_0 (which you don't observe), the FDR control is much better than FDR_0.