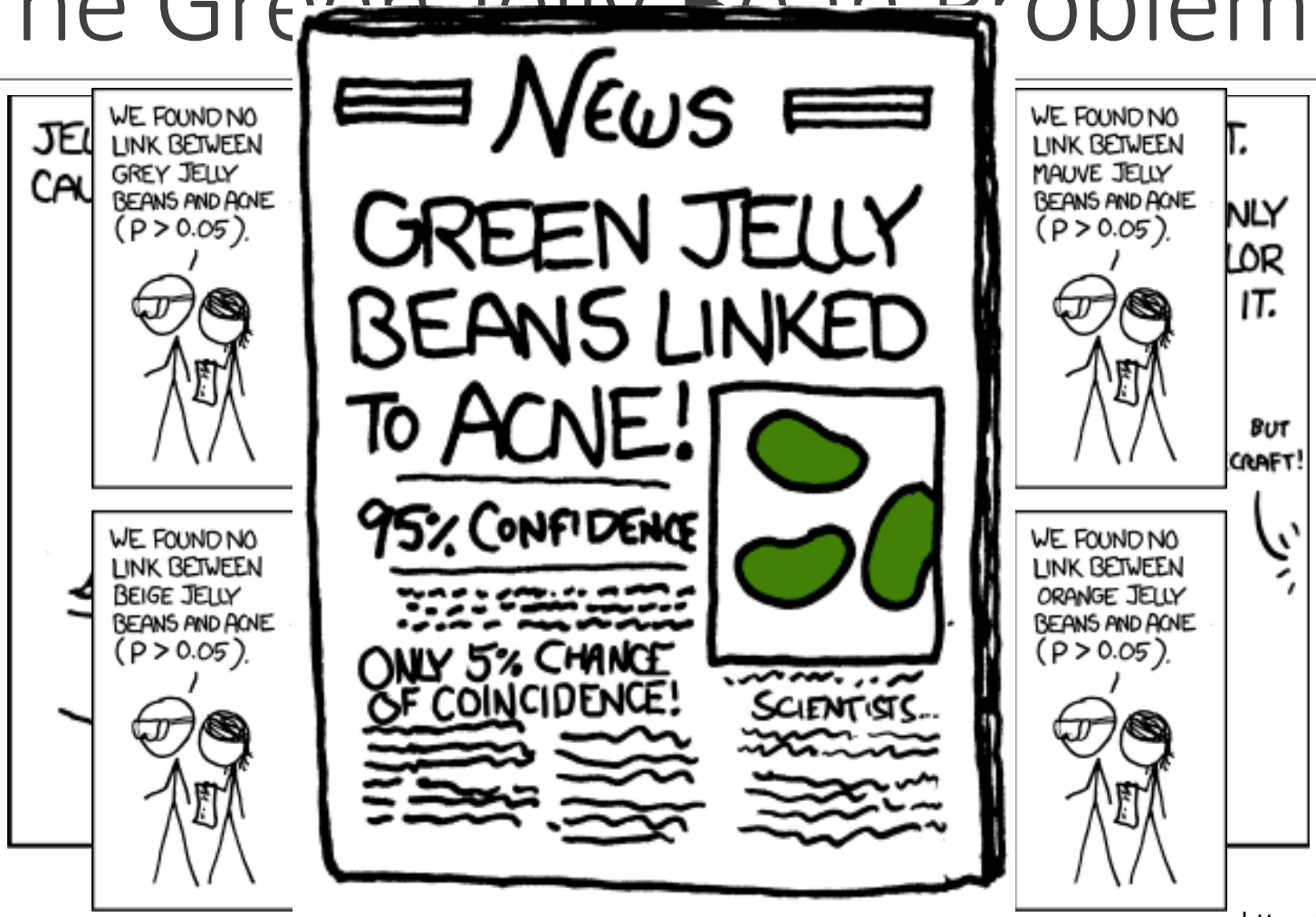
Linear Combinations and Multiple Comparisons

MULTIPLE HYPOTHESIS TESTS

TUKEY-KRAMER, DUNNETT, SCHEFFÉ, LSD, BONFERRONI, ….

(NOTE: SKIP PLANNED VS. UNPLANNED MULTIPLE COMPARISONS)

The Green Jelly Bean Problem



https://xkcd.com/882/

The Green Jelly Bean Problem & Study Designs

- Suppose we are really interested in investigating whether jelly beans **cause** acne

- We "find" a group of people, randomly split them into two groups
  - One group (treatment) who are instructed to eat lots of jelly beans
  - Other group (control) who are banned from eating jelly beans

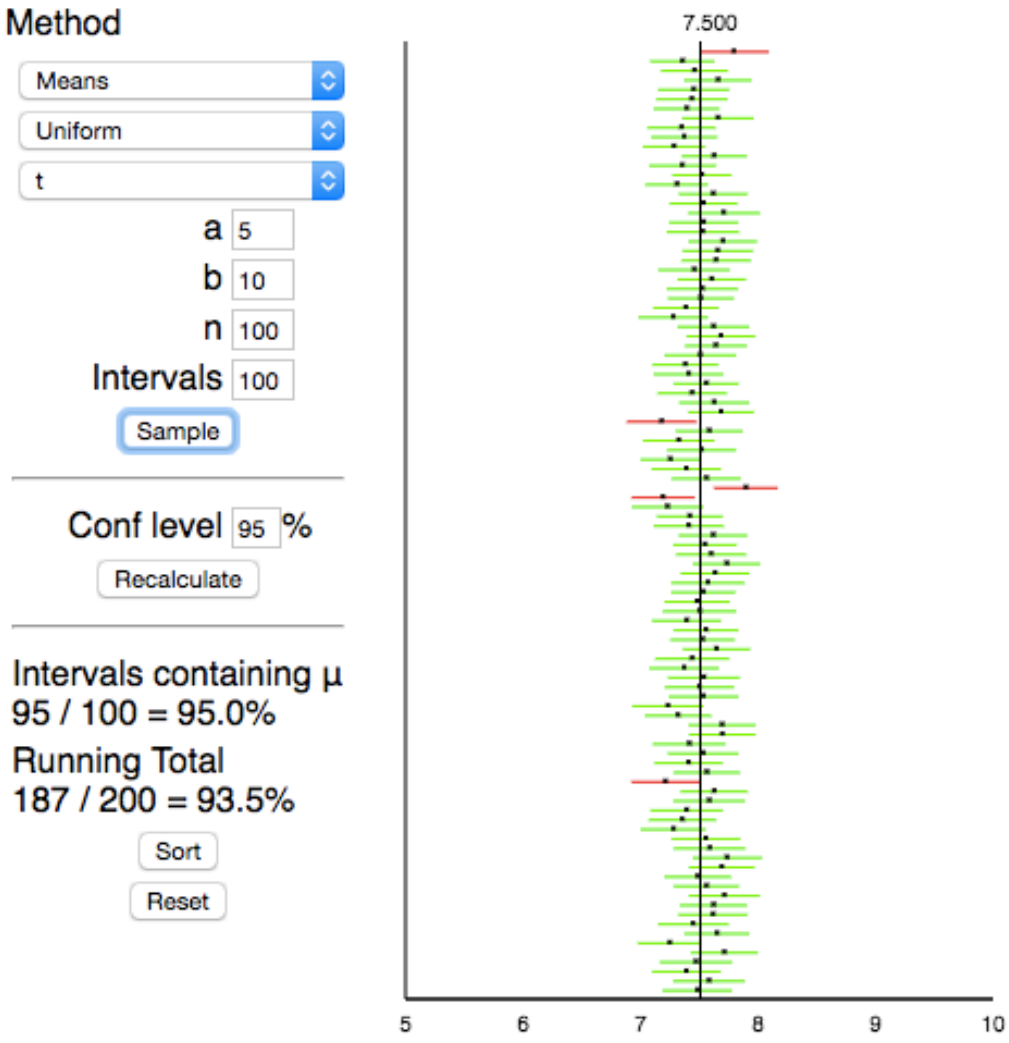
  (Ideally: neither the researchers nor subjects should know who is in which group)

- After some time you measure/compare the acne in the two groups

- If more people in the group that eat jelly beans have acne then you might think that jelly beans cause acne.

Confidence Intervals

The confidence level is the fraction of times that the true parameter is in the interval (if the assumptions are met) if we were to repeat the same experiment a large number of times



Individual confidence level *is the success rate of a procedure for constructing a single confidence interval.*

The Green Jelly Bean Problem & Multiple Comparisons

- Even if we do everything right, for a 100(1-α)% confidence interval there is still an 100α% chance we are wrong (in this case, $\alpha = 0.05$)

- The scientists in the comic repeat the experiment 20 times for 20 different colors

  (I'm ignoring the initial test for plain jelly beans)

- Each time, each confidence interval (or equivalently hypothesis test that we reject if the p-value < 0.05) has a 5% chance of incorrectly rejecting H_0: $\mu_{Treat} = \mu_{Control}$

  (This is a "false positive" or "Type I error")

The Green Jelly Bean Problem & Multiple Comparisons

Two facts about this series of experiments:

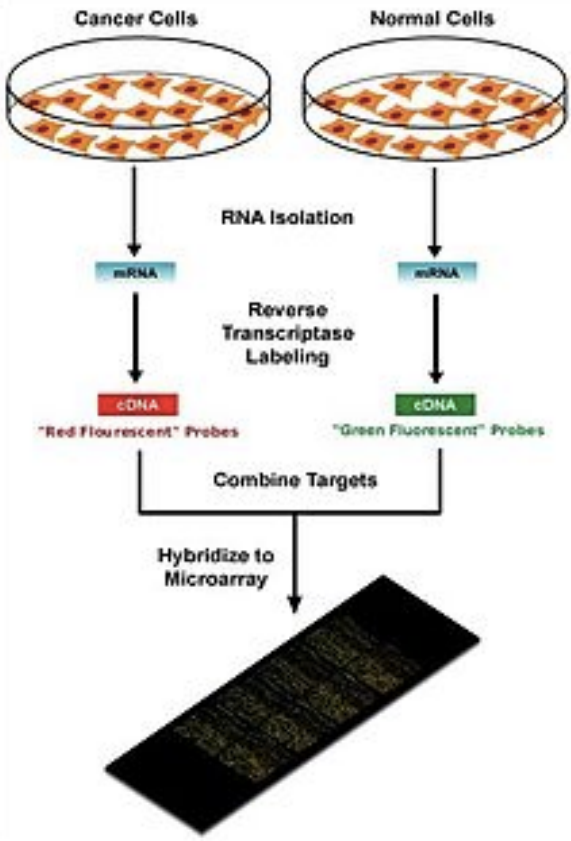
- As there are 20 experiments, each with a 5% chance of failure, we should expect that on average 1 experiment would have a false positive if all of the H_0 are true
- Additionally, suppose each test is run independently. Then:

(Prob. of at least 1 false positive) = 1 – (Prob. of no false positives)

= 1 – (Prob. of no false positives on test 1) (Prob. of no false positives on test 2)⋯ (Prob. of no false positives on test 20) = $1 - (0.95)^{20} = 0.64$

- There is a 2/3 Prob. of at least 1 false positive if all the H_0 are true!

Another Example: Genetic Testing



A common task in modern medicine is to:

- Find a group of people with some condition

(e.g. Pancreatic Cancer)

- Find another group of people without that condition
- Collect RNA samples from each subject and measure

how much specific genes of interest are expressed

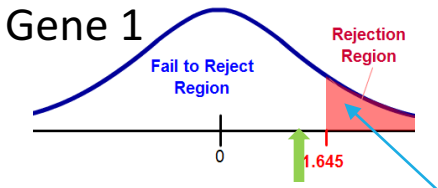
We can now run (a possibly large number) of t-tests for the mean difference in gene expression in the two groups

The inference is that these genes are **associated** with the condition

(causality is very tough: can't assign humans to cancer status..)

Genetic Testing: Single Gene

Individual confidence level *is the success rate of a procedure for constructing a single confidence interval.*



(Interpretation: the number of false positives if we were able to run a large number of identical experiments/trials)

Is there evidence that the expression of "Gene 1" is larger for the cancer group than the not cancer group?

H_0: $\mu_{cancer} = \mu_{no}$
H_A: $\mu_{cancer} > \mu_{no}$

Specify an α such that $P(rejectH_0|H_0) = \alpha$

$P(rejectH_0|H_0) = P(t_{obs} > threshold \mid H_0)$

If:

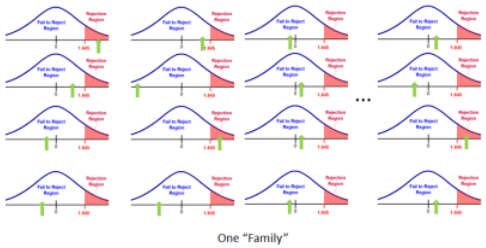
1. The pooled t-tools assumptions hold
2. $threshold = t_{\alpha,n-2}$

Then:
$P(rejectH_0|H_0) = \alpha$

(Note: α is not a p-value!)

Genetic Testing: Multiple Genes

> ***Familywise confidence level*** *is the success rate of a procedure for constructing a family of confidence intervals, where a "successful" usage is one in which all intervals in the family capture their parameters.*



For each gene, is there evidence that the expression is larger for the cancer group than the not cancer group?

H_0: $\mu_{1,cancer} = \mu_{1,no}$
H_A: $\mu_{1,cancer} > \mu_{1,no}$

…

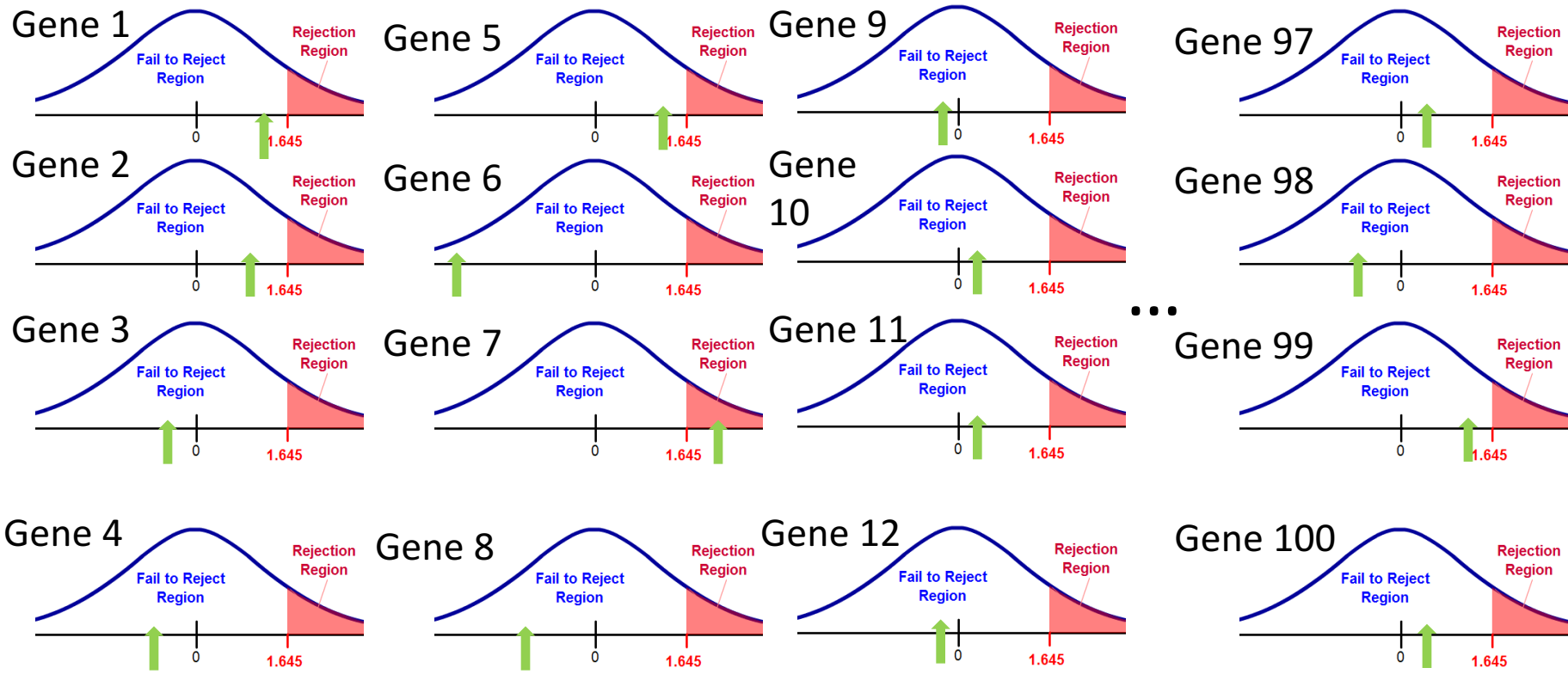
H_0: $\mu_{100,cancer} = \mu_{100,no}$
H_A: $\mu_{100,cancer} > \mu_{100,no}$

(Interpretation: the number of **families of tests** that have at least one false positive if we were able to run a large number of identical experiments/trials)

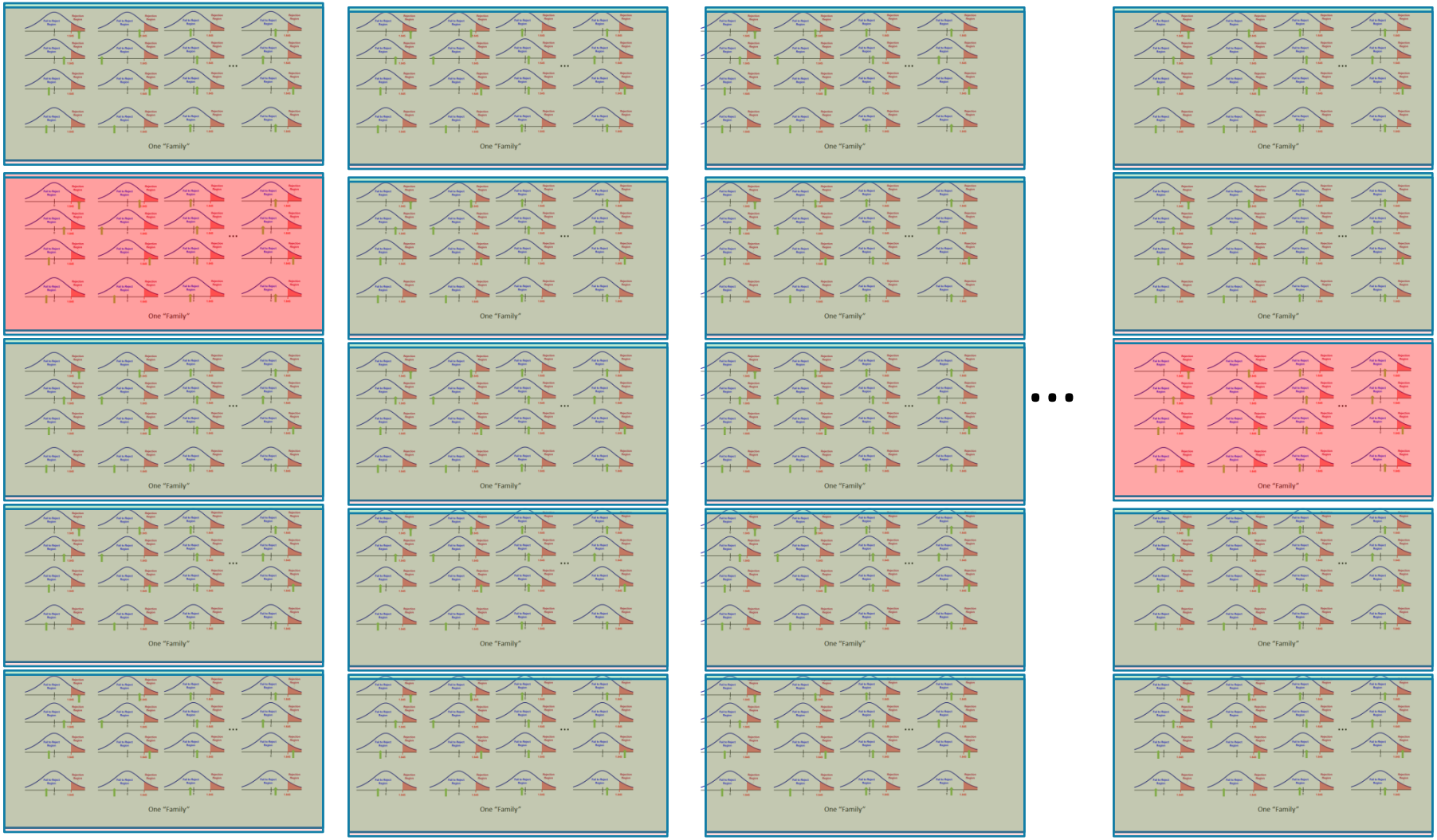
Specify an α such that $P(at\ least\ one\ test\ rejects\ H_0 | H_0) = \alpha$

Genetic Testing: Multiple Genes

> ***Individual confidence level*** *is the success rate of a procedure for constructing a single confidence interval.*



Genetic Testing: A Large # of Trials



Goal: control the number of false positives for a **family** of tests for a large # of trials)

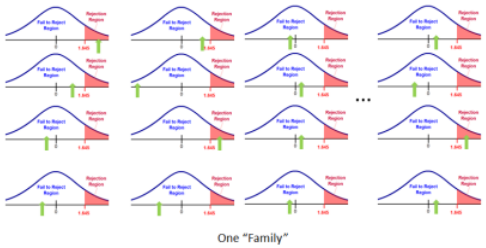
Genetic Testing: A Large # of Trials

Each “YES” is a family of 100 individual tests all failing to reject Ho Correctly. Each “No” is a family of 100 individual tests that rejected at least once Incorrectly.

YES	YES	YES	YES	YES	YES	YES		YES	YES
NO	YES	YES	YES	YES	YES	YES		YES	YES
YES	YES	YES	YES	YES	YES	YES		YES	YES
YES	YES	YES	YES	YES	YES	YES	•••	YES	YES
YES	YES	YES	YES	**NO**	YES	YES		**NO**	YES
YES	YES	YES	YES	YES	YES	YES		YES	YES
YES	YES	YES	YES	YES	YES	YES		YES	YES

Genetic Testing: Multiple Genes

Familywise confidence level *is the success rate of a procedure for constructing a family of confidence intervals, where a "successful" usage is one in which all intervals in the family capture their parameters.*



For each gene, is there evidence that the expression is larger for the cancer group than the not cancer group?

H_0: $\mu_{1,cancer} = \mu_{1,no}$ … H_0: $\mu_{100,cancer} = \mu_{100,no}$

H_A: $\mu_{1,cancer} > \mu_{1,no}$ … H_A: $\mu_{100,cancer} > \mu_{100,no}$

(Interpretation: the number of **families of tests** that have at least one false positive if we were able to run a large number of identical experiments/trials)

Specify an α such that $P(at\ least\ one\ test\ rejects\ H_0|H_0) = \alpha$

$$P(at\ least\ one\ test\ rejects\ H_0|H_0) = P(t_{1,obs} > threshold\ or\ \ldots or\ t_{100,obs} > threshold|\ H_0)$$

Choosing $threshold = t_{\alpha,n-2}$ only controls individual-wise errors

We need to choose $threshold = ?$ that controls the family-wise errors

Multiple Hypothesis Tests: Procedures for Family-wise Errors

For ease of exposition, we will detail these procedures in terms of confidence intervals. The same procedures apply to hypothesis tests via adjusting the **critical values**

Typically doesn't change (e.g.: $SE(\bar{Y}_1 - \bar{Y}_2)$)

$$\text{Interval half-width} = (\text{Multiplier}) \times (\text{Standard error}).$$

Typically gets adjusted to larger values to control for family-wise errors

Multiple Hypothesis Tests : Bonferroni

If the confidence level for each of k individual comparisons is adjusted to $100\left(1-\frac{\alpha}{k}\right)\%$, the chance that all intervals succeed simultaneously in containing the populations means is at least $100(1-\alpha)\%$

$$\text{(Multiplier)} = t_{\frac{\alpha}{2k},df}$$

$$\text{Interval half-width} = (\text{Multiplier}) \times (\text{Standard error}).$$

This is a very conservative adjustment, which will lead to wide confidence intervals

$P(at\ least\ one\ test\ rejects\ H_0|H_0) =$
$P(t_{1,obs} > threshold\ or\ \ldots or\ t_{100,obs} > threshold|\ H_0) \leq$
$P(t_{1,obs} > threshold|\ H_0) + \cdots + P(\ t_{100,obs} > threshold\ |\ H_0) = \alpha k$
→ Choose $threshold =\ t_{\alpha/k,n-2}$

Multiple Hypothesis Tests : Tukey-Kramer

Creates a "Multiplier" that ensures (under H_0) that the largest test statistic from all the comparisons would lead to rejecting H_0 $100(\alpha)\%$ of the time.

(this reference distribution is known as the "studentized range distribution")

Example: If we want a familywise confidence of 95%, we set the "Multiplier" so that we would reject 5% of the H_0 under the assumption that H_0 are true

$$\text{(Multiplier)} = \frac{q_{(1-\alpha),I,n-I}}{\sqrt{2}}$$

$$\text{Interval half-width} = (\text{Multiplier}) \times (\text{Standard error}).$$

This correction has a complicated reference distribution

It should only be used when making pairwise comparisons

(It is a very commonly used correction for this purpose)

This procedure requires a lot of assumptions (such as normality.

Multiple Hypothesis Tests : Scheffé

Creates a "Multiplier" that gives familywise control over all possible contrasts (of which, differences in means are a special case).

$$\text{(Multiplier)} = \sqrt{(I-1)F_{(1-\alpha),I-1,n-I}}$$

$$\text{Interval half-width} = (\text{Multiplier}) \times (\text{Standard error}).$$

As this procedure works for all contrasts, it is the most conservative approach and hence will have the widest confidence intervals
(generally to be used for **confidence bands** in multiple regression, not multiple groups)

Multiple Hypothesis Tests : Dunnett's Procedure

- This works in the special case that we are making many comparison to a particular group (like the placebo or control group) to different treatments $(T1, T2, \dots, TI)$

$$t_1 = \frac{\bar{Y}_C - \bar{Y}_{T1}}{SE(\bar{Y}_C - \bar{Y}_{T1})} \qquad t_2 = \frac{\bar{Y}_C - \bar{Y}_{T2}}{SE(\bar{Y}_C - \bar{Y}_{T2})} \qquad \dots \qquad t_I = \frac{\bar{Y}_C - \bar{Y}_{TI}}{SE(\bar{Y}_C - \bar{Y}_{TI})}$$

- Each of the comparisons are dependent as $\bar{Y}_C$ appears in each term, but the dependence is easier to track as the other averages only appear once

- The (Multiplier) is based on the "multivariate t distribution"

Interval half-width = (Multiplier) × (Standard error).

Multiple Hypothesis Tests : Least Significant Difference (LSD)

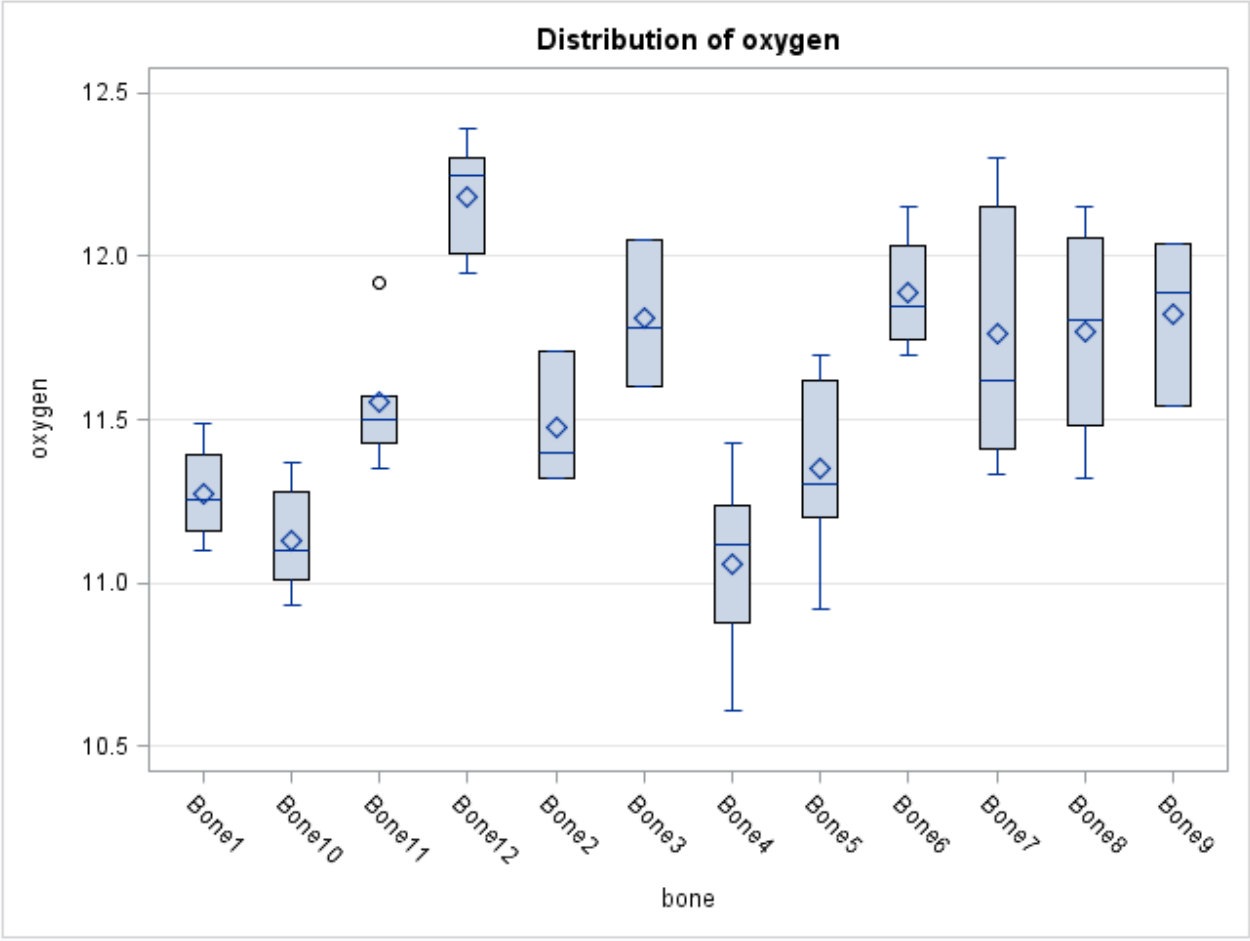
Creates a “Multiplier” that gives individual control over false positive rate

$$\text{(Multiplier)} = t_{(1-\alpha/2),df}$$

$$\text{Interval half-width} = (\text{Multiplier}) \times (\text{Standard error}).$$

This is the procedure we have been using all along. It is the most liberal approach and hence will have the narrowest confidence intervals

Example: Oxygen Levels in T-rex Bone Samples



Least Significant Difference (LSD)

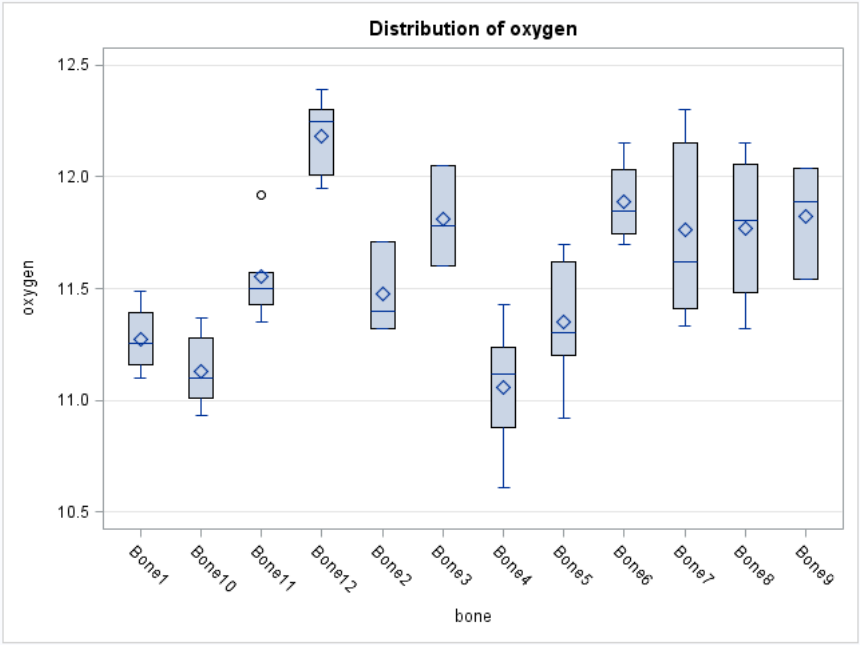
```
proc glm data = TREX;
class bone;
model oxygen = bone;
means bone / cldiff;
run;
```

Bone12 - Bone6	0.2925	-0.0770	0.6620	
Bone12 - Bone9	0.3567	-0.0456	0.7589	
Bone12 - Bone3	0.3700	-0.0322	0.7722	
Bone12 - Bone8	0.4100	0.0405	0.7795	***
Bone12 - Bone7	0.4180	0.0696	0.7664	***
Bone12 - Bone11	0.6260	0.2776	0.9744	***
Bone12 - Bone2	0.7033	0.3011	1.1056	***
Bone12 - Bone5	0.8320	0.4836	1.1804	***
Bone12 - Bone1	0.9050	0.5355	1.2745	***
Bone12 - Bone10	1.0483	0.7148	1.3819	***
Bone12 - Bone4	1.1240	0.7756	1.4724	***
Bone6 - Bone12	-0.2925	-0.6620	0.0770	

→ 62 significant differences

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	40
Error Mean Square	0.07427
Critical Value of t	2.02108



Tukey-Kramer

```
proc glm data = TREX;
class bone;
model oxygen = bone;
means bone /     cldiff        TUKEY;
run;
```

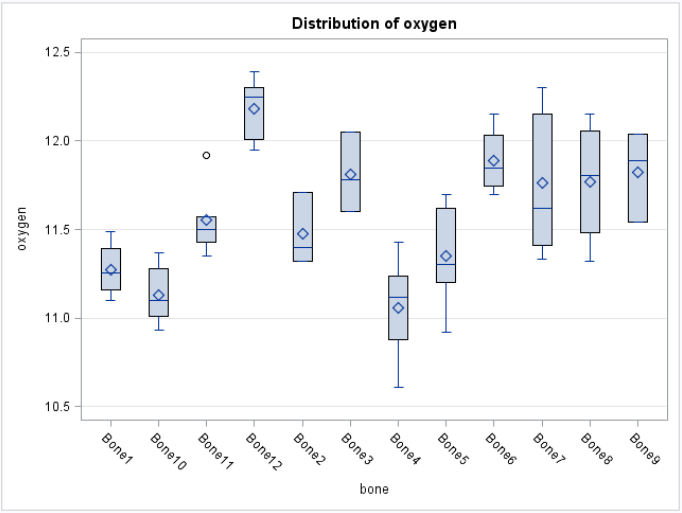
Bone10 - Bone12	-1.0483	-1.6206	-0.4761	***
Bone10 - Bone6	-0.7558	-1.3658	-0.1458	***
Bone10 - Bone9	-0.6917	-1.3599	-0.0234	***
Bone10 - Bone3	-0.6783	-1.3466	-0.0101	***
Bone10 - Bone8	-0.6383	-1.2483	-0.0283	***
Bone10 - Bone7	-0.6303	-1.2026	-0.0581	***
Bone10 - Bone11	-0.4223	-0.9946	0.1499	

→ 32 significant differences

Tukey's Studentized Range (HSD) Test for oxygen

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	40
Error Mean Square	0.07427
Critical Value of Studentized Range	4.90393



Bonferroni

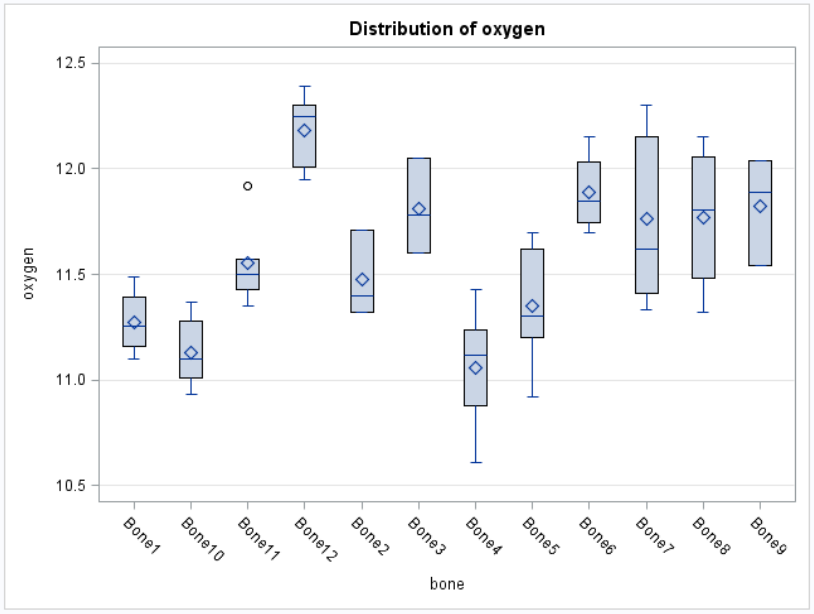
```
proc glm data = TREX;
class bone;
model oxygen = bone;
means bone /    cldiff        BON;
run;
```

Bonferroni (Dunn) t Tests for oxygen

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	40
Error Mean Square	0.07427
Critical Value of t	3.64679

Bone12 - Bone5	0.8320	0.2034	1.4606	***
Bone12 - Bone1	0.9050	0.2383	1.5717	***
Bone12 - Bone10	1.0483	0.4465	1.6501	***
Bone12 - Bone4	1.1240	0.4954	1.7526	***
Bone6 - Bone12	-0.2925	-0.9592	0.3742	
Bone6 - Bone9	0.0642	-0.6949	0.8232	
Bone6 - Bone3	0.0775	-0.6816	0.8366	



→ 22 significant differences

Example: Handicap & Capability Study

- **Goal:** How do physical handicaps affect perception of employment qualification?
- The researchers prepared 5 recorded job interviews with same actors
- The tapes differed only in the handicap of the applicant:
  - No handicap
  - One leg amputated
  - Crutches
  - Hearing Impaired
  - Wheelchair

14 people were randomly assigned to each tape to rate applicants: 0-10 pts

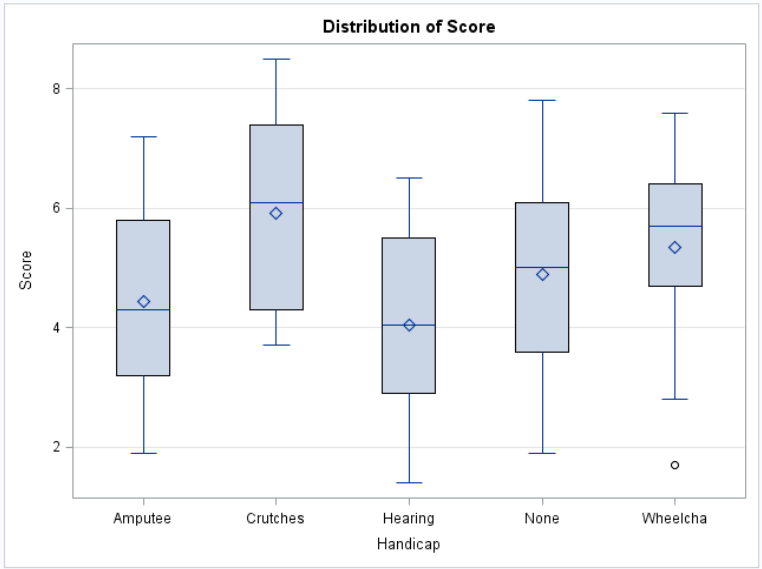
Example: Handicap & Capability Study

Do subjects systematically evaluate qualifications differently according to handicap?

If so, which handicaps are evaluated differently?

	None	Amputee	Crutches	Hearing	Wheelchair
0					
1	9	9		4	7
2	5	56		149	8
3	06	268	7	479	5
4	129	06	033	237	78
5	149	3589	18	589	03
6	17	1	0234	5	1124
7	48	2	445		246
8			5		
9					

Legend: 7 | 4 represents a score of 7.4 on the Applicant Qualification Scale.



Example: Handicap & Capability Study

Questions of interest:

1. Is there any evidence that at least one pair of mean qualification scores are different?

2. Let's say we are only interested in Amputee versus None. Test the claim the Amputee has a different mean score than the None group.

3. Now let's assume that we are interested in "any" differences. Find evidence of any differences in the means between the groups.

4. Next assume that we were interested in testing the means of the handicapped groups to the non-handicap group. Test this claim and identify any significant differences.

Handicap & Capability Study: First Question

$H_0: \mu_{None}=\mu_{Amp}=\mu_{Crutch}=\mu_{Hear}=\mu_{Wheel}=\mu$

$H_A: \mu_j \neq \mu_k$ for some j, k

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	30.5214286	7.6303571	2.86	0.0301
Error	65	173.3214286	2.6664835		
Corrected Total	69	203.8428571			

There is evidence to that there are at least two population means different (p-value of 0.0301 from a 1-way ANOVA).

Handicap & Capability Study: Second Question

Amputee versus None:

$$H_0: \mu_{Amputee} = \mu_{None}$$
$$H_A: \mu_{Amputee} \neq \mu_{None}$$

handicap	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
None		4.9000	3.8644	5.9356	1.7936	1.3003	2.8895
Amp		4.4286	3.5130	5.3441	1.5857	1.1496	2.5547
Diff (1-2)	Pooled	0.4714	-0.8438	1.7866	1.6928	1.3331	2.3199
Diff (1-2)	Satterthwaite	0.4714	-0.8447	1.7876			

```
PROC TTEST DATA = handicap ORDER = DATA;
    WHERE handicap = 'None' | handicap = 'Amp';
    CLASS handicap;
    VAR score;
RUN;
```

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	26	0.74	0.4678
Satterthwaite	Unequal	25.615	0.74	0.4679

```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    CONTRAST 'Amp vs. None Contrast' handicap 1 -1 0 0 0;
    ESTIMATE 'Amp vs. None Estimate' handicap 1 -1 0 0 0;
RUN;
```

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	13	13	1.28	0.6635

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Amp vs. None Estimate	0.47142857	0.61719220	0.76	0.4477

There is not sufficient evidence to suggest that the mean points for the amputee group is different than mean points for control group (p-value = 0.4477 from t-test contrast)

Handicap & Capability Study: Third Question

Find evidence of any differences in the means of the groups

```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    LSMEANS handicap / pdiff;
RUN;
```

(I used pdiff here as there are only 5 groups)

handicap	score LSMEAN	LSMEAN Number
None	4.90000000	1
Amp	4.42857143	2
Crut	5.92142857	3
Hear	4.05000000	4
Whee	5.34285714	5

- 5 groups → 10 (two-sided) tests
- If we use a Bonferroni adjustment, we need to adjust: $\alpha = 0.05 \rightarrow \frac{\alpha}{k} = \frac{0.05}{10} = 0.005$
- Now, only 1 significant difference:

Evidence that the crutches and hearing groups have different mean qualification rating scores

Least Squares Means for effect handicap
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: score

i/j	1	2	3	4	5
1		0.4477	0.1028	0.1732	0.4756
2	0.4477		0.0184	0.5418	0.1433
3	0.1028	0.0184		0.0035	0.3520
4	0.1732	0.5418	0.0035		0.0401
5	0.4756	0.1433	0.3520	0.0401	

Handicap & Capability Study: Third Question

Least Squares Means for effect handicap
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: score

i/j	1	2	3	4	5
1		0.4477	0.1028	0.1732	0.4756
2	0.4477		0.0184	0.5418	0.1433
3	0.1028	0.0184		0.0035	0.3520
4	0.1732	0.5418	0.0035		0.0401
5	0.4756	0.1433	0.3520	0.0401	

Least Squares Means for effect handicap
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: score

i/j	1	2	3	4	5
1		1.0000	1.0000	1.0000	1.0000
2	1.0000		0.1838	1.0000	1.0000
3	1.0000	0.1838		0.0349	1.0000
4	1.0000	1.0000	0.0349		0.4010
5	1.0000	1.0000	1.0000	0.4010	

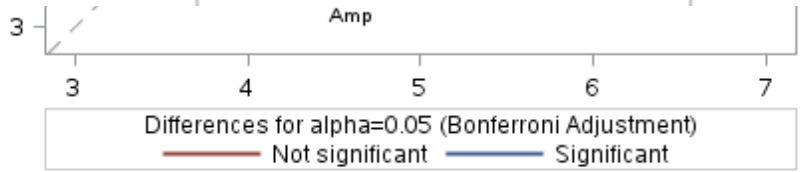
```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    LSMEANS handicap / pdiff;
RUN;
```

```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    LSMEANS handicap / pdiff ADJUST = BON CL;
RUN;
```

Handicap & Capability Study: Third Question

score Comparisons for handicap

Least Squares Means for Effect handicap				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.471429	-1.322141	2.264999
1	3	-1.021429	-2.814999	0.772141
1	4	0.850000	-0.943570	2.643570
1	5	-0.442857	-2.236427	1.350713
2	3	-1.492857	-3.286427	0.300713
2	4	0.378571	-1.414999	2.172141
2	5	-0.914286	-2.707856	0.879284
3	4	1.871429	0.077859	3.664999
3	5	0.578571	-1.214999	2.372141
4	5	-1.292857	-3.086427	0.500713



Handicap & Capability Study: Fourth Question

Is there a difference between any handicap and the control group ('None')?

```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    LSMEANS handicap / pdiff ADJUST = DUNNETT CL;
RUN;
```

handicap	score LSMEAN	H0:LSMean=Control Pr > |t|
None	4.90000000	
Amp	4.42857143	0.8597
Crut	5.92142857	0.2918
Hear	4.05000000	0.4516
Whee	5.34285714	0.8836

Least Squares Means for Effect handicap

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
2	1	-0.471429	-2.016357	1.073500
3	1	1.021429	-0.523500	2.566357
4	1	-0.850000	-2.394929	0.694929
5	1	0.442857	-1.102072	1.987786

There is insufficient evidence that there is a difference between any of the handicap groups and the mean of 'None' group. The 95% Dunnett's procedure-corrected confidence intervals all contain zero.

Multiple Comparisons in SAS

```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    MEANS handicap / CLDIFF BON;
RUN;
```

```
PROC GLM DATA = handicap ORDER=DATA;
    CLASS handicap;
    MODEL score = handicap;
    LSMEANS handicap / pdiff ADJUST = BON CL;
RUN;
```

You can either use the “means” or “lsmeans” statements.
I prefer to use ”lsmeans”.